

UC San Diego

UC San Diego Electronic Theses and Dissertations

Title

Beyond the Search Bar: Augmenting Discovery, Synthesis & Creativity By Mining Unstructured User-Generated Context

Permalink

<https://escholarship.org/uc/item/1z75f2k4>

Author

Palani, Srishti

Publication Date

2024

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA SAN DIEGO

**Beyond the Search Bar: Augmenting Discovery, Synthesis & Creativity By Mining
Unstructured User-Generated Context**

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Cognitive Science

by

Srishti Palani

Committee in charge:

Steven P. Dow, Chair

James D. Hollan

Scott Klemmer

Julian McAuley

Daniel M. Russell

2024

Copyright
Srishti Palani, 2024
All rights reserved.

The dissertation of Srishti Palani is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

University of California San Diego

2024

DEDICATION

To Kartik, the best big brother and coolest mentor, whose unwavering support and ingenious sense of humor guided me throughout my journey.

TABLE OF CONTENTS

	Dissertation Approval Page	iii
	Dedication	iv
	Table of Contents	v
	List of Figures	xiii
	List of Tables	xviii
	Acknowledgements	xix
	Vita	xxiv
	Abstract of the Dissertation	xxvi
Chapter 1	Introduction	1
	1.1 Observing Information Seeking & Sensemaking Workflows Longitudinally	4
	1.2 Getting Started With Information Exploration	6
	1.3 Symbiotically Supporting Information Exploration and Syn- thesis	7
	1.4 Augmenting Creative Workflows	9
	1.5 Thesis Statement	11
Chapter 2	Observing Information Seeking & Sensemaking Workflows Longi- tudinally	13
	2.1 Introduction	14
	2.2 Related Work	18
	2.2.1 Information Seeking, Sensemaking and Creativity in Knowledge Work	18
	2.2.2 Work Patterns, Information Needs, Challenges Dur- ing Creative Work	20
	2.2.3 Methods To Study Web Search	23
	2.3 Method	24
	2.3.1 Participants	24
	2.3.2 Browser Extension for Data Collection and Visualiza- tion	25
	2.3.3 Procedure	27
	2.3.4 Measures	28
	2.4 Findings	29

2.4.1	Participants allocate more time to work during the early and late project stages while taking longer breaks early on that progressively shorten as the project advances.	30
2.4.2	Creative activities take place in a non-linear, iterative manner across project stages	33
2.4.3	Participants actively search and synthesize online information across all creative activities	33
2.4.4	Artifacts generated can encode rich contextual information	36
2.4.5	Participants have distinctive information needs during each creative activity	37
2.4.6	Each creative activity and phase of work session presents unique search and sensemaking challenges, and participants envision how future tools could help	39
2.5	Discussion	44
2.5.1	Insights on Creative Work Patterns	44
2.5.2	Insights on Search and Sensemaking Patterns, Information Needs and User Challenges	46
2.5.3	Limitations and Future Work	48
2.6	Conclusion	49
2.7	Acknowledgements	50

I Getting Started With Information Exploration 51

Chapter 3	The "Active Search" Hypothesis: Characterizing Search Behavior and Challenges When Starting to Explore Information and Frame Problems	52
3.1	Introduction	53
3.2	Method	55
3.2.1	Participants	55
3.2.2	Procedure	56
3.2.3	Measures	56
3.3	Results	58
3.3.1	What information goals do searchers have during early-stage design?	59
3.3.2	What search strategies emerged to meet these information goals?	60
3.3.3	How do web search behaviors affect creative learning outcomes?	61
3.4	Discussion	63
3.5	Conclusion	65

	3.6 Acknowledgements	66
Chapter 4	CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery	67
	4.1 Introduction	68
	4.2 Related Work	70
	4.2.1 Note-taking helps individuals store, learn, and share information during exploratory search	71
	4.2.2 Query assistance methods in exploratory search	72
	4.2.3 Leveraging information in user-generated documents to provide query assistance	73
	4.3 CoNotate	74
	4.3.1 User Interface	75
	4.3.2 System Architecture	77
	4.4 Evaluation	81
	4.4.1 Conditions	81
	4.4.2 Search Tasks	82
	4.4.3 Participants	84
	4.4.4 Procedure	84
	4.4.5 Measures	85
	4.5 Results	86
	4.5.1 Effects on search behavior: Notes-based query assistance encourages more active searching	87
	4.5.2 Effects on learning: Notes-based query assistance promotes knowledge discovery	89
	4.5.3 Effects on user preferences: Participants preferred notes-based suggestions over <i>baseline</i> suggestions	91
	4.6 Discussion	91
	4.6.1 How does notes-based query assistance support exploration and knowledge discovery?	92
	4.6.2 Study Limitations	94
	4.6.3 Future Work	95
	4.7 Conclusion	97
	4.8 Acknowledgements	98

II Symbiotically Supporting Information Exploration and Synthesis 99

Chapter 5	InterWeave: Presenting Search Suggestions Within User's Evolving Sensemaking Structures Promotes Information Exploration and Synthesis	100
	5.1 Introduction	101

5.2	Related Work	106
5.2.1	Exploratory Information Seeking	106
5.2.2	Integrating Search and Sensemaking	108
5.2.3	Presenting Search Suggestions	110
5.3	InterWeave	111
5.3.1	User Challenges & Design Goals	111
5.3.2	InterWeave Interface	113
5.3.3	System Architecture	114
5.3.4	Implementation	118
5.4	Study: Where to place suggestions?	119
5.4.1	Conditions	119
5.4.2	Participants	121
5.4.3	Task	121
5.4.4	Procedure	122
5.4.5	Measures	123
5.5	Results	126
5.5.1	InterWeave encourages active searching	127
5.5.2	InterWeave assists sensemaking	128
5.5.3	InterWeave enhances knowledge gain	129
5.5.4	Participants preferred InterWeave's in context presentation of suggestions	130
5.5.5	Wizard's insights on automating the process of inferring context and placing suggestions	134
5.6	Discussion	136
5.6.1	How can in context placement of search suggestions affect exploration and learning?	137
5.6.2	Limitations and Future Work	140
5.7	Conclusion	142
5.8	Acknowledgements	143
Chapter 6	Relatedly: Scaffolding Literature Reviews With Existing Related Work Sections	144
6.1	Introduction	145
6.2	Related Work	149
6.2.1	How Scholars Conduct Literature Reviews	149
6.2.2	Tools for Supporting Literature Review	151
6.3	Formative Study & Design Goals	153
6.3.1	Formative User Study Method	153
6.3.2	User Experience when Reviewing Literature	154
6.3.3	Design Goals	156
6.4	The Relatedly System	157
6.4.1	Example User Scenario	158
6.4.2	System Features	159

6.4.3	Automatic Section Heading Generation	164
6.4.4	Implementation Details	168
6.5	User Evaluation Study Design	169
6.5.1	Experimental Setup	170
6.5.2	Study Procedure	173
6.5.3	Measures	173
6.6	Findings	175
6.6.1	Higher Quality Synthesized Outlines	175
6.6.2	Paper- vs Topic-Centric Exploration	177
6.6.3	Participants Preferred Relatedly	179
6.6.4	Volunteered Use in the Wild	181
6.7	Discussion	183
6.7.1	Limitations and Future Work	185
6.8	Conclusion	188
6.9	Acknowledgements	189

III Augmenting Creative Workflows 190

Chapter 7	The Practitioner Perspective on Creativity Support Tool Adoption	191
7.1	Introduction	193
7.2	Related Work	195
7.2.1	Designing and Evaluating Creativity Support Tools .	195
7.2.2	Theoretical Background On Technology Adoption .	197
7.3	Method	199
7.3.1	YouTube Videos	199
7.3.2	Semi-Structured Interviews	200
7.3.3	Analysis of Videos and Interview Data	201
7.3.4	Survey	202
7.4	Results	203
7.4.1	Tools' Features and Functionality	204
7.4.2	Integration with Existing Workflow	208
7.4.3	Tools' Performance	210
7.4.4	User Interface and Experience	213
7.4.5	Level of Support	216
7.4.6	Financial Costs	219
7.4.7	Emotional Connection	219
7.4.8	Exploration and Discovery of Creativity Support Tools	220
7.5	Discussion: Ties Between Framework and Literature	223
7.5.1	Features/Functionality	225
7.5.2	Integration with Current Workflow	225
7.5.3	Performance	226
7.5.4	User Interface and Experience	227

	7.5.5	Level of Support	228
	7.5.6	Financial Costs	228
	7.5.7	Emotional Connection	229
	7.5.8	Exploration and Discovery	230
7.6		Limitations and Future Work	230
7.7		Conclusion	232
7.8		Acknowledgements	233
Chapter 8		Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI	234
	8.1	Introduction	235
	8.2	Related Work	237
	8.2.1	The Creative Process and Creativity Support Tools	237
	8.2.2	Interacting with GenAI During Creative Processes	238
	8.3	Method	240
	8.3.1	Semi-Structured Interviews	241
	8.3.2	YouTube Videos	242
	8.3.3	Analysis of Videos and Interview Data	242
	8.3.4	Survey	243
	8.4	Results	244
	8.4.1	Perceived Roles When Working with Generative AI	244
	8.4.2	Trade-offs: Benefits and Challenges of Creating with Generative AI	248
	8.4.3	Evolving the Creative Process: Project- and Artifact- Level Orchestrations	252
	8.5	Discussion	258
	8.5.1	Findings, Observations & Ties to Prior Literature	259
	8.5.2	From Observations to Insights: Design Priorities and Opportunities for Future CSTs	261
	8.5.3	Limitations & Future Work	264
	8.6	Conclusion	265
	8.7	Acknowledgements	265
Chapter 9		Amethyst: Enabling Affordances for Specifying and Referring to User-Generated Context Fosters Creativity Human-Centered Orchestration of GenAI	266
	9.1	Introduction	267
	9.2	Related Work	270
	9.2.1	The creative process is just as important as the outcome	270
	9.2.2	Today's GenAI-based Creativity Support Tools assist with individual tasks, not the entire process	271
	9.3	Amethyst System	273
	9.3.1	Example User Scenario	274

	9.3.2	System Features	277
	9.3.3	Implementation Details	284
	9.4	User Evaluation Study Design	286
	9.4.1	Tasks	286
	9.4.2	Baseline	287
	9.4.3	Participants	287
	9.4.4	Study Procedure	288
	9.4.5	Measures	289
	9.5	Findings	290
	9.5.1	Better Creative Outcomes Achieved When Using <i>Amethyst</i>	290
	9.5.2	RQ3: How does the <i>Amethyst</i> help orchestrate GenAI during the creative process in human-centered ways .	292
	9.5.3	Participants' perception of <i>Amethyst's</i> and GenAI's effects on their creative work	297
	9.6	Discussion & Future Work	298
	9.6.1	Supporting <i>Both</i> Creative Tasks & The Process Using GenAI	299
	9.6.2	Context-Aware Interactions With GenAI During The Creative Process	300
	9.6.3	Modelling Interactions with GenAI To Be More Em- pathetic and Social	301
	9.6.4	Beyond Interacting with GenAI as a Creativity Sup- port Tool	302
	9.7	Conclusion	303
	9.8	Acknowledgements	304
Chapter 10		Conclusion & Future Work	306
	10.1	Future Research Agenda	307
	10.1.1	Understanding and Designing for "Good Friction" in Interaction Mechanisms	308
	10.1.2	Leveraging Collaborative Context to Guide Data Ex- ploration, Sensemaking, and Creative Insights	309
	10.1.3	Evaluation Metrics Based On Human Cognition and Social Dynamics During Information Workflows . .	310
	10.2	Closing Remarks	311
Appendix A		Appendix of Chapter 7: Relatedly	313
Appendix B		Appendix of Chapter 9: Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI	316
	B.1	Semi-Structured Interview Guide	317

Appendix C	Appendix of Chapter 10: Amethyst	320
	C.1 LLM prompts	321
	C.2 Context Prompt	321
	C.3 Generic prompt	322
	C.4 Master prompt	324
	C.5 Create persona	326
	C.6 Critique and Reflection prompt	327
	C.7 Todo prompt	329
	C.8 Do Task prompt	330

LIST OF FIGURES

Figure 1.1:	This thesis contributes the following	11
Figure 2.1:	Participants spanned many different professions and worked on projects related to a range of creative goals over different time periods.	25
Figure 2.2:	An illustrative screenshot showcasing the zoomable visualization designed for a participant to reflect on their weekly search and sensemaking patterns	26
Figure 2.3:	Participants underwent a brief screening, followed by a one-hour orientation session with the research team	27
Figure 2.4:	Work session length in minutes across the entire project for two participants	31
Figure 2.5:	Average percentage of work sessions spent on each creative activity mapped to early, middle, and late stages of the project	34
Figure 2.6:	Average time spent (minutes) by a participant on searching and synthesizing online during each creative activity	35
Figure 2.7:	Average number of queries issued (left), number of webpages opened (middle), and number of words changed in the document (right) per participant during creative activities	35
Figure 4.1:	The <i>CoNotate</i> Environment: including (a) Default Chrome Search Interface, augmented with (b) Suggestions Bar with six query suggestions and (c) the Note Taking Interface. The system supports additional interactions including (d) scrolling query suggestions, (e) resizing note-taking interface, and (f) highlighting, dragging and dropping web page content into notes.	75
Figure 4.2:	Architecture of the <i>CoNotate</i> system, a browser extension that parses a user’s notes and search terms in order to offer context-relevant query suggestions.	77
Figure 4.3:	Noun phrases extraction from SERPs and notes (bold text). The <i>gap_phrases</i> are <i>SERP_phrases</i> not in <i>notes_phrases</i>	79
Figure 4.4:	The Baseline Environment: The Default Chrome Search Interface, augmented with (a) Suggestions Bar with query autocompletion suggestions	82
Figure 4.5:	Averages (and standard deviation) of searchers’ level of agreement to these statements on a scale of 2 (Strongly Agree) to -2 (Strongly Disagree) for Baseline and <i>CoNotate</i> suggestions.	90

Figure 5.1:	InterWeave’s user interface augments (a) a search browser with (b) a sensemaking workspace where contextual search suggestions are presented at up to four levels within user’s evolving sensemaking structure at the (c) title, (d) cluster, (e) cross-clusters, and (f) individual note levels	104
Figure 5.2:	While many search systems recommend search queries, InterWeave goes further by inferring the user’s sensemaking structures, formulating context-aware query suggestions and then weaving suggestions back into the sensemaking workspace	108
Figure 5.3:	InterWeave’s system architecture which leverages NLP algorithms and a wizard to present contextual suggestions within the searcher’s emergent sensemaking representations.	114
Figure 5.4:	Wizard’s interface when choosing and placing search suggestions in the emerging sensemaking structure	115
Figure 5.5:	The Baseline Condition lists suggestions outside the user’s Sensemaking Workspace	120
Figure 5.6:	Examples of notes taken by InterWeave participants	127
Figure 5.7:	Examples of notes taken by Baseline participants	127
Figure 5.8:	InterWeave participants issued significantly more queries, particularly the suggestions compared to Baseline participants. However, they typed similar number of queries.	128
Figure 5.9:	InterWeave participants gathered significantly more information and exhibited broader and deeper sensemaking in their sensemaking workspace, while visiting similar number of websites, compared to Baseline participants	129
Figure 5.10:	InterWeave participants reported a significantly greater increase in knowledge, discovered more domain-specific terms, and idea units compared to Baseline participants.	129
Figure 5.11:	InterWeave participants agreed significantly more to the statements about the presentation of query suggestions being helpful compared to Baseline participants	131
Figure 5.12:	InterWeave participants felt they had better transparency around how the suggestions were being generated.	132
Figure 5.13:	Searchers’ level of agreement to these statements on a scale of 2 (Strongly Agree) to -2 (Strongly Disagree) for Baseline and InterWeave suggestions	135
Figure 6.1:	The <i>Relatedly</i> system presents users with related work paragraphs from prior work on a topic and scaffolds the paragraph exploration experience with features for reading, prioritization, and progress tracking	146

Figure 6.2:	To read more on the subtopic discussed in a specific paragraph in the Overview View (Fig. 6.1), this Similar Paragraphs View allows users to explore other paragraphs of that same subtopic that cited the same or similar references.	157
Figure 6.3:	Human-evaluation comparing model-generated and author-written section headings	166
Figure 6.4:	The Baseline condition that emulates common scholarly search engines (left)	170
Figure 6.5:	Participants generated learning outline summaries after 20 minutes of literature review each with the two systems. The summaries were rated by experts on 3 criteria using 5-point Likert-scales for agreement (5 indicated strong agreement)	175
Figure 6.6:	Participants interacted with significantly more paragraphs when using Relatedly vs the Baseline system	177
Figure 6.7:	Participants' level of agreement to how well Relatedly and Baseline supported their literature review process. For each statement, we report the percentage of likert responses and results from paired wilcoxon signed rank tests, with z and p values.	180
Figure 6.8:	Background and usage of participants who volunteered long-term use. Participants self-reported their job title, reason for using Relatedly in their research workflow, hours of work, hours using Relatedly, # queries, and # of relevant papers curated to read	181
Figure 7.1:	Visual abstract of this paper's investigation of what creative practitioners value when adopting creativity support tools – summarizing the key contributions: C1. Empirical Observations, C2. Creative Practitioners' Value Framework, C3. Mapping values to design principles and theories in literature	192
Figure 7.2:	Overview of creative practitioners' value categories	204
Figure 7.3:	Features and Functionality values	205
Figure 7.4:	Integration with current workflow values	208
Figure 7.5:	Performance values	210
Figure 7.6:	User Interface and Experience values	213
Figure 7.7:	Level of support values	217
Figure 7.8:	Financial costs of CSTs	218
Figure 7.9:	Emotional connection with CSTs	219
Figure 7.10:	How practitioners discover and explore CSTs	221
Figure 7.11:	Creative Practitioners' values and how they fit within existing literature across systems and creativity support tools research, and technology acceptance and adoption theories	224
Figure 8.1:	Creative practitioners perform a range of activities with GenAI over the course of the creative process	244

Figure 8.2:	Practitioner-mentioned metaphors for perceived roles, categorized by creative agency and perspective (people vs. AI), highlighting human-centric roles with higher agency and AI-centric roles as tools	245
Figure 8.3:	Social dynamics such as agency and empathy between the creative practitioner and AI models shift across the creative process	247
Figure 8.4:	Challenges derived from thematic analysis across videos, interviews, and survey responses showing how frequently something was mentioned, and the coverage across sources.	248
Figure 8.5:	Benefits derived from thematic analysis across videos, interviews, and survey responses showing how frequently something was mentioned, the coverage across sources	250
Figure 8.6:	The creative process is evolving to be iterations between project- and artifact-level orchestrations.	252
Figure 8.7:	Creative process is evolving to be iterations between project- and artifact-level orchestrations	253
Figure 9.1:	Amethyst is a smart notebook designed to facilitate the creative process in an integrated manner	274
Figure 9.2:	Users can provide a high-level objective, which the system can then break down into actionable tasks. Also, all following GenAI outputs will use this as additional context	278
Figure 9.3:	"/" (Forward Slash) Commands: Users can access various creative operations by typing "/" to open the drop-down menu and selecting the desired operation.	279
Figure 9.4:	This Summarize component is an example of a "/" operation	280
Figure 9.5:	In-line prompting	281
Figure 9.6:	Me page: Users can edit this text file to define how they would like to be seen by the system externalizing personal context such as their current emotional state and design preferences	282
Figure 9.7:	Example of a persona page	283
Figure 9.8:	Transparency lens: Hovering over the generate button of each "/" operation component displays what context it is grounded on so that the users know what information it has access to for transparency and in case they want to change it.	283
Figure 9.9:	Critique component output	284
Figure 9.10:	<i>Amethyst's</i> components	285
Figure 9.11:	Participants generated significantly more creative outcomes when using <i>Amethyst</i> vs. the baseline when rated by blind-to-condition experts on Novelty, Feasibility, and Value of ideas, on a 5-point Likert-scales for agreement (5 indicated strong agreement).	291
Figure 9.12:	Participants found <i>Amethyst's</i> goal decomposition and task management features helpful in defining their fuzzy creative goals	292

Figure 9.13: Participants often used "/" Commands to orchestrate creative operations during their process.	293
Figure 9.14: Participants found it useful to ground their creative operations in relevant context explicitly. They mostly referred to relevant context using "@"	295
Figure 9.15: Participants found it helpful to use simulated expert personas to modulate GenAI outputs and appreciated Amethyst's empathetic responses based on the user's emotional state.	296
Figure A.1: During the formative user study, participants were given access to a prototype system where they could search topics and it would return topic-relevant paragraphs from related work sections across multiple paper	315
Figure B.1: Overview of Videos Analyzed	318
Figure B.2: Overview of Participants Interviewed	319

LIST OF TABLES

Table 3.1:	Depth of Learning Measures corresponding to the Understand, Analyze and Evaluate Cognitive Learning Levels of Anderson and Krathwohl’s Taxonomy of Learning	58
Table 3.2:	A correlation analysis found a significant positive relationship between higher gain in facts stated and issuing more, longer, more diverse queries, and opening the more web.	62
Table 3.3:	More active and diverse searching relates to deeper learning and more well-defined problems.	63
Table 4.1:	Averages (and standard deviation) for key search metrics. Participants issued significantly more queries, particularly by clicking on the suggestions, and typed fewer manual queries when using <i>CoNotate</i> than when using <i>Baseline</i> system. *statistically significant at $p < 0.05$ level	87
Table 4.2:	Averages (and standard deviation) for key information gain metrics for the <i>Baseline</i> and <i>CoNotate</i> system.	89
Table 6.1:	ROUGE scores for both models on the test split of descriptive section headings	166
Table 9.1:	<i>Amethyst</i> ”/” commands or operations.	305
Table A.1:	Example model-generated headings for paragraphs with long and descriptive author-written titles side-by-side.	314
Table A.2:	Example model-generated headings for paragraphs with short and generic author-written titles side-by-side.	314
Table A.3:	There were no significant differences in task workloads when using Relatedly vs Baseline suggesting improved performance with similar precieved workload	315

ACKNOWLEDGEMENTS

I am grateful to have benefitted from the support and guidance from incredible mentors, colleagues, friends, and family. So to everyone in my Ph.D. journey, and in life — this is for you.

First, to my incredible advisor, Steven, who always cared deeply, provided insightful critique and encouraged me to pursue opportunities I would never have considered myself: thank you. Steven has been nothing short of an exemplary advisor and truly a fantastic mentor. Through brainstorming ideas, late nights editing my writing, supernatural responsiveness, and an incredible eye for detail, his relentless support has not only helped me solve problems but identify important problems to solve.

Jim, Scott, Haijun, Philip, Stephen, and Don – thank you for inspiring me with all that you do to not only be a better HCI researcher but also a caring mentor and community leader. Also, thank you to the incredible CogSci faculty — especially Seana, Marta, Doug, and Federico — your passion for your research and CogSci is truly infectious and made me fall in love with cognitive science, coming in as a computer scientist and neuroscientist. The pandemic’s endless days were brightened by Julian’s engaging machine learning classes on Twitch, making grasping these complex topics both enjoyable and accessible.

Throughout my Ph.D., I have had the wonderful opportunity to be mentored by some of the smartest and kindest people during four formative research internships at Microsoft, Autodesk, the Allen Institute for AI, and through the Google PhD Fellowship. They gave me the freedom and flexibility to choose problems that were both widely important and foundational to my thesis, and instilled in me a confidence in my research. I’m forever thankful to Merrie took a chance on me as a freshly-graduated undergrad, overviewed all the wonders of HCI research and methodology, and has continued to awe and inspire me ever since with her incredible mind, generous heart, endless passion for her work and

vision for the community. To be inspired by reading all of Gonzalo and Joseph's papers, and then finally working with them over the summers was an absolute dream. Gonzalo – thank you for always reminding me to never hesitate to shoot for the moon and aim big. Joseph – thank you for working to widening my view of how to make an impact. David, thank you for teaching me the art of visual storytelling, and Fraser for always being positive and supportive of our crazy ideas. Adam, Jonathan, Amy, Jina – thank you for all your help developing my wonderings into well-defined projects. Thank you, Dan, for always guiding my work to have real-world impact with insightful questions and thoughtful feedback.

It has been an honor to be part of the uniquely enriching communities of UCSD Design Lab and Cognitive Science. This thesis could not have been conceived and developed anywhere else. Anytime I had a question or needed to learn a new skill — from learning eye tracking to Python, filing IRBs to TAing, someone from our community was always there to help. CogSci's rotations, foundation courses, and second- and third-year classes provided unique opportunities for me to grow and learn about this fascinatingly interdisciplinary field that I didn't know much about. Despite the pandemic and moving our lab across campus, friendships and mentorships within the CogSci community have only strengthened! It was so heartwarming to see so many of you at my defense. To my fellow labmates – Tone, Lu, Jude, Annapurna, Grace, Jane, Sangho, and Bryan, thank you for your incredible daily support and enthusiasm. Thank you for all the unforgettable moments hiking, chilling by beach bonfires, hot tubbing, exploring the San Diego Zoo, celebrating Diwali, and so much more! A huge shoutout to Thanh and the CogSci staff, who always made the impossible possible – from supporting my dreams of teaching a course as Instructor of Record to advocating with Grad Div for release funding on time and working with me to design the CogSci open house website with so much care. Also, thank you so much Nivardo, for always going above and beyond to make sure everything

is organized and running smoothly. Jason, Austin, Andrew, Yingyi and Sheldon – thank you for being incredible interns and collaborators. You made our research come to life! Ailie, Tricia, Vineet, Sean, Ian – thank you for paving the way ahead and for always being real with me.

Besides this incredible academic support system, I need to give a final special thank you to my family:

Thank you, Kanishka, for growing with me this entire way, supporting me through the worst times, and celebrating with me in the best times. You make me a better person and remind me of what is important in life, and I can't wait to continue this adventure together. And for Churro for always bringing the light into our lives.

My parents who always prioritized my education, thank you for giving me the opportunity, freedom, and patience to find and follow my passions. My dad who inspired my love for math and engineering – raising me by explaining the engineering of everything from the insides of common household appliances to supersonic aircrafts and helicopters. My mom who taught me to appreciate the arts and humanities – raising me to appreciate the beauty shakespeare, languages and art. My brother Kartik has always been my biggest inspiration and supporter – thank you for always being my pillar of strength. Thank you for always leading by example, being there through thick and thin and for your unconditional love. Your support and encouragement have given me the courage to chase my dreams and embrace every opportunity. And Amrita who models what it means to not only be an incredibly organized engineer, but also never-endingly kind and generous. Thank you for everything you have done to help me get to where I am today.

Chapter 2, in part, is currently being prepared for submission for publication of the material. Srishti Palani and Steven P. Dow. The dissertation author was the primary investigator and author of this material.

Chapter 3, in part, includes portions of material as it appears in *The "Active Search"*

Hypothesis: How Search Strategies Relate to Creative Learning by Srishti Palani, Zijian Ding, Stephen MacNeil, Steven P. Dow in the Proceedings of 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval Online (CHIIR'21). The dissertation author was the primary investigator and author of this material.

Chapter 4, in part, includes portions of material as it appears in *CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery* by Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, Steven P. Dow in Proceedings of the 2021 ACM CHI Conference on Human Factors in Computing Systems (CHI'21). The dissertation author was the primary investigator and author of this material.

Chapter 5, in part, includes portions of material as it appears in *InterWeave: Embedding Query Suggestions within Searcher's Sense-making Structures Promotes Active Searching and Knowledge Discovery* by Srishti Palani, Yingyi Zhou, Sheldon Zhu, Steven P. Dow in Proceedings of the 2022 ACM Symposium on User Interface Software and Technology (UIST'22). The dissertation author was the primary investigator and author of this material.

Chapter 6, in part, includes portions of material as it appears in *Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections* by Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, Joseph Chee Chang in Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems (CHI'23). The dissertation author was the primary investigator and author of this material.

Chapter 7, in part, includes portions of material as it appears in *"I don't want to feel like I'm working in a 1960s factory": The Practitioner Perspective on Creativity Support Tool Adoption* by Srishti Palani, David Ledo, Fraser Anderson, George Fitzmaurice in Proceedings of the 2022 ACM CHI Conference on Human Factors in Computing Systems (CHI'22). The dissertation author was the primary investigator and author of this material.

Chapter 8, in part, includes portions of material as it appears in *Evolving Mental Models, Workflows and Opportunities: A Study of Creative Practitioners Interacting with Generative AI* by Srishti Palani and Gonzalo Ramos in Proceedings of the 2024 ACM Conference on Creativity and Cognition (C&C'24). The dissertation author was the primary investigator and author of this material.

Chapter 9, is currently being prepared for submission for publication of the material. Srishti Palani and Gonzalo Ramos. The dissertation author was the primary investigator and author of this material.

VITA

- 2018 B. A. *summa cum laude* in Computer Science, Psychology, with specialization in Cognitive Neuroscience, Mount Holyoke College, Massachusetts
- 2021 M. S. in Cognitive Science, University of California San Diego
- 2024 Ph. D. in Cognitive Science, University of California San Diego

PUBLICATIONS

Srishti Palani, Gonzalo Ramos. Evolving Mental Models, Workflows and Opportunities: A Study of Creative Practitioners Interacting with Generative AI, To appear in *Proceedings of ACM Conference on Creativity and Cognition (C&C'24)*, June 24-26, 2024, Chicago, IL

Xiaotong 'Tone' Xu, Srishti Palani, Steven P. Dow. Idea-Centric Search: Challenges and Opportunities for Web Search During Creative Ideation, To appear in *Proceedings of ACM Conference on Creativity and Cognition (C&C'24)*, June 24-26, 2024, Chicago, IL

Jude Rayan, Nicole Gong, Dhruv Kanetkar, Yuewen Yang, Srishti Palani, Steven Dow and Haijun Xia. Random Access Ideas: Exploring the Potential for LLM-Powered Conversational Prompts for Real-Time Collaborative Ideation. To appear in *Proceedings of ACM Conference on Creativity and Cognition (C&C'24)*, June 24-26, 2024, Chicago, IL

Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, Joseph Chee Chang. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI'23)* April 23-28, 2023, Hamburg, Germany

Sangho Suh, Bryan Wang, Srishti Palani, Haijun Xai. Sensecape: Enabling Multi-level Exploration and Sensemaking with Large Language Models. In *Proceedings of UIST'23: ACM Symposium on User Interface Software and Technology (UIST'23)*. San Francisco, USA

Kyle Lo, Joseph Chee Chang, Andrew Head, Amy X. Zhang,... Srishti Palani,.. Daniel S. Weld. The Semantic Reader Project: Augmenting Scholarly Documents through AI-Powered Interactive Reading Interfaces. *arXiv preprint arXiv:2303.14334*

Srishti Palani, Yingyi Zhou, Sheldon Zhu, Steven Dow. Embedding Query Suggestions within Searcher's Sense-making Structures Promotes Active Searching and Knowledge Discovery. In *Proceedings of UIST'22: ACM Symposium on User Interface Software and Technology (UIST)*. Oct 29-Nov 2, 2022, Bend, Oregon, USA

Srishti Palani, David Ledo, Fraser Anderson, George Fitzmaurice. "I don't want to feel like I'm working in a 1960s factory": The Practitioner Perspective on Creativity Support Tool Adoption. In *Proceedings of CHI'22. ACM CHI Conference on Human Factors in Computing Systems (CHI'22)*. April 20-May 6, 2022, New Orleans, USA.

Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, Steven Dow. CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery. In *Proceedings of CHI'21: ACM CHI Conference on Human Factors in Computing Systems (CHI'21)*. May 8-13, 2021, Virtual Conference, Japan

Matin Yarmand, Srishti Palani, Scott Klemmer. Adjacent Display of Relevant Discussion Resolves Confusion when Learning Online. In *Proceedings of the ACM of Human-Computer Interaction 5(CSCW'21) (2021) November 2021, Virtual Conference*.

Srishti Palani, Zijian Ding, Stephen MacNeil, Steven Dow. The "Active Search" Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of CHIIR'21: ACM SIGIR Conference on Human Information Interaction and Retrieval Online (CHIIR'21)*. March 14-19, 2021, Virtual Conference, Canberra, Australia

Srishti Palani, Adam Fourney, Shane Williams, Kevin Larson, Irina Spiridonova, and Meredith Ringel Morris. 2020. An Eye Tracking Study of Web Search by People With and Without Dyslexia. In *Proceedings of SIGIR'20: ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'20)*, July 25-30, 2020, Virtual Event, China.

ABSTRACT OF THE DISSERTATION

**Beyond the Search Bar: Augmenting Discovery, Synthesis & Creativity By Mining
Unstructured User-Generated Context**

by

Srishti Palani

Doctor of Philosophy in Cognitive Science

University of California San Diego, 2024

Steven P. Dow, Chair

Searching and exploring online is a part of our everyday lives – shaping how we learn, work and innovate. However, today, people are still drowning in information, with few mechanisms for managing or synthesizing large volumes of disparate information. It is a struggle to find the right information or identify relevant unknown unknowns for those who lack knowledge of a particular domain or well-defined goals. Even experts juggle dozens of disparate information silos spread out across different apps, websites, and work sessions. This is cognitively overwhelming and time-consuming, preventing people from developing a comprehensive understanding, gaining deep insights, and achieving their creative potential. This is especially true in complex creative information work

like scientific research, founding a startup or innovating to protect the public during a pandemic.

As the Web paradigm evolves to include Generative AI models and beyond, we are experiencing a shift in how we search, learn, work and create. With this transformation in human-AI interaction, it is important to investigate how we might present the user with the right information in the right context, the right representation, and at the right time. This thesis explores this in the context of cognitively complex information work (such as knowledge discovery, synthesis, and creativity). It presents two types of contributions: (1) Empirical studies that further our understanding of how people explore, make sense of, and create using information on the Web. The studies follow a mixed-methods approach, combining large-scale and longitudinal quantitative data analysis with in-depth qualitative inquiry. (2) Computational and interaction techniques that augment these cognitive processes by seamlessly integrating knowledge from the Web into the user's work context.

Each study observes user behavior, challenges, and strategies at different stages of information exploration, sensemaking, and creative processes. Each system introduces an approach for inferring contextual signals from user-generated artifacts. For example, such as *CoNotate* mines an individual's unstructured artifacts for knowledge gaps and patterns to make query suggestions, *InterWeave* analyzes and presents suggestions in the user's evolving sensemaking structures to present suggestions, *Relatedly* mines existing knowledge structures on the web from previous users to present dynamic topic overviews, and *Amethyst* enables users with affordances to specify and refer to personal, project-level, and external contexts. User evaluation studies demonstrate how these techniques, mining rich contextual signals from work done during cognitive processes, can promote information exploration, synthesis, and creativity.

Chapter 1

Introduction

Whether conducting scientific research, innovating a new product, or developing effective public policy — online search and exploration are integral to how we learn, work, and innovate. During cognitively complex information work like this, people need to explore, find, read, extract meaning, identify connections and gain creative insight across various sources. The Web was originally envisioned as a “cognitive boost to empower intelligence” [254, 64]. However, today, we are drowning in an ever-rising sea of information [366]. It is hard to articulate complex and fuzzy information goals [40, 65, 316], discover insights [36] and make breakthroughs beyond our narrow perspectives [52] while making sense of information across a fragmented ecosystem of resources, apps, and work sessions [38, 74].

While this information work is a time-consuming and cognitively overwhelming process, it can also be cognitively rewarding. Searching for information and making sense of multiple information sources can help us feel more confident in our understanding of a topic [55, 57], and learn search as a skill [286]. People often take notes when making sense of found information [107]. Taking notes involves manipulating information by summarizing, paraphrasing, and mapping. This engagement can help cognitively encode and gain a deeper understanding of the information [225, 206, 275, 163]. When taking notes, searchers often select and record relevant concepts [280, 420], process-related information (e.g. queries, links, etc.), and their own interpretations [107, 74, 420]. This suggests that the purpose of notes goes beyond just helping people record and process information but also serves to synthesize low-level raw data into high-level meaning, ideas and decisions [231, 357, 34, 163]. Furthermore, creating this artifact of their thinking and sense-making makes it easier for searchers to share knowledge and collaborate with others [420, 137, 176, 232, 163]. Organizing and structuring information in different ways can help us see interesting connections and have creative insights [37, 134]. Furthermore, planning, monitoring and evaluating artifacts generated during this process helps build

meta-cognitive skills [135, 135] Therefore, instead of automating these processes, we build on foundational visions of human-AI interaction research (like [132, 253, 180, 190]) to design and implement interactive systems in which humans and AI work together.

To build systems that support these cognitively complex processes in human-centred ways, we must first observe how people behave when working on these processes in the real world, what challenges they face and the strategies they use. Prior research in Interactive Information Retrieval (IIR) and Human-Computer Interaction (HCI) has mostly observed these searchers in the context of learning in lab studies with controlled tasks [106, 105] and developed learning-based measures for such tasks [444, 447]. Building on this research, **this thesis contributes empirical studies advancing our understanding of how people think, learn, and create, leveraging online information and generative AI in the real world.** Here, I follow a mixed-methods approach that combines longitudinal and large-scale quantitative data analysis with in-depth qualitative inquiry.

Today, web search engines and chat-based LLMs are the primary mechanisms by which people seek information online. When exploring a new domain through Web search, people often struggle to articulate queries because they lack domain-specific language and well-defined informational goals [439]. Current web search engines attempt to assist people with query formulation by leveraging search log data to detect user intents and context [52, 406]. General-purpose search engines, like Google and Bing, recommend queries to help people fulfil their information needs quicker by predicting query formulations (e.g., auto-complete), resolving ambiguity (e.g. people also ask), showing what other people searched in this area (e.g., related searches) [31, 203, 270, 79, 144, 336]. Search systems have also explored different ways of presenting query suggestions, usually in lists and visualizations [219, 323, 412, 44, 73, 387, 455] separated from the tools where the user synthesizes and works with information. Therefore, the user must determine which suggestion is most relevant to their situation, follow the information

scent to find the resource, extract information, and adapt it to their context. This is exacerbated by having to consult and identify connections across multiple information sources to develop a comprehensive, nuanced understanding of a topic. This process is getting progressively harder with the exponential growth of information on the Web [141, 51, 210, 421], and the increasingly interdisciplinary nature of information work required to solve today's problems [422, 309]. And even further exacerbated by needing to go across the many work sessions and stages of thinking it takes to complete a project [74, 212].

As the Web paradigm evolves to include foundational AI models and beyond, we stand at a pivotal opportunity to define the next generation of tools used for searching and working with information. With this transformation in human-AI interaction, it is important to investigate how we might present the user with the right information in the right context, the right representation, and at the right time. **Towards this, this thesis contributes novel algorithmic and interaction techniques that seamlessly integrate knowledge from the Web into user's work contexts to promote information discovery, synthesis, and creative insight.** These approaches are implemented as tools and evaluated in lab studies and real-world deployments.

This thesis presents these contributions of empirical studies and interactive systems in the following parts:

1.1 Observing Information Seeking & Sensemaking Workflows Longitudinally

Chapter 2: To understand and identify opportunities to augment people's workflows as they work on cognitively complex information goals, we conducted a longitudinal observational study. We collected and analyzed application logs from search and work

documents (e.g., Google Docs, Notion Workspaces, Overleaf documents) of 15 creative information workers (including startup founders, researchers, policy advisors, journalists, and novelists) throughout their project lifecycles (1 - 6 months long, avg. 2.5 months long).

We developed a novel method to collect data in a way that provides the participant transparency and control around what data is being collected while also enabling reflection on their work patterns by generating real-time semantic-zoomable data visualizations. Among other findings, we observed that participants use search across all the iterative stages of their work, from discovering relevant information and defining project scope to generating and developing ideas into narratives that can be communicated to the world. It was interesting to note that search plays a role in ideation and project scoping, which are thought to be just mental processes. We also found that artifacts generated along their iterative creative journeys can encode rich contextual information. This includes the user's goals, what they already know about a topic (or what they are missing), how they feel, their design preferences, how far they have come in their project, how they link what they know to what they are finding, and how they structure their thoughts. This way in which people exhibit distributed cognition— by externalizing thoughts in work documents — provides an opening for supporting complex creative knowledge work.

1.2 Getting Started With Information Exploration

Chapter 3 The "Active Search" Hypothesis: Characterizing Search Behavior and Challenges When Starting to Explore Information in Creative Projects

To investigate how people use web search to learn about a new domain and frame their thinking about an open problem, we collected and analyzed search log and self-report data from 34 students in a project-based design class. Participants reported struggling with scoping broad, ill-defined information goals into queries, learning domain-specific language, and assessing the usefulness of information. Analysis found that more active and diverse search behavior (i.e. issuing more frequent and diverse queries, and opening more webpages) related to more progress in early-stage design (i.e. gathering more facts, articulating more insights, and developing better problem frames). These findings imply that search behavior and strategies exhibited during exploratory creative information goals differ from those seen during the simple lookup searches that search engines are currently optimized for. Based on these findings, we discuss implications for designing search tools to support peoples' creative processes.

Chapter 4 CoNotate: Supporting Articulation of Exploratory Information Goals By Mining User-Generated Content

Towards addressing the challenges observed and expanding this contextual understanding of a user during exploratory searches, we introduce a novel system, CoNotate. CoNotate offers query suggestions based on analyzing the searcher's notes and previous searches for patterns and gaps in information. To evaluate this approach, we conducted a within-subjects study where participants (n=38) conducted exploratory searches using a

baseline system (standard web search) and the CoNotate system. The CoNotate approach helped searchers issue significantly more queries and discover more terminology than standard web search. This work demonstrates how search can leverage user-generated content to help people get started when exploring complex, multi-faceted information spaces.

1.3 Symbiotically Supporting Information Exploration and Synthesis

Chapter 5 InterWeave: Presenting Search Suggestions Within Evolving Schema in User-Generated Content Promotes Information Search and Synthesis

Exploring and synthesizing information into knowledge can be slow and cognitively demanding due to a disconnect between search tools and the workspaces where people make sense of and work on found information. In this chapter, we explore how might we integrate contextual query suggestions within a person's sensemaking environment. Building on CoNotate, we developed *InterWeave*, a prototype that leverages a human wizard to generate contextual search guidance and to place the suggestions within the emergent structure of a searcher's notes. To investigate how weaving suggestions into the emerging structures in their sensemaking workspace affects a user's search and sensemaking behavior, we ran a between-subjects study (n=34) where we compared *InterWeave's* in-context placement with a conventional list of query suggestions. *InterWeave's* approach not only promoted active searching, information gathering and knowledge discovery but also helped participants keep track of new suggestions and connect newly discovered information to existing knowledge, in comparison to presenting suggestions

as a separate list.

Chapter 6 Relatedly: Scaffolding Information Exploration and Synthesis With Existing Web Content and Structure

Today, it is still a struggle to quickly get a comprehensive understanding of an evolving multi-faceted topic (like research around COVID-19 in 2020 or how to mitigate misinformation). This requires time to read, extract meaning, and identify connections across various sources, which is becoming more challenging due to the web's exponential growth. We explored this challenge in the domain of scientific discovery through our project, *Relatedly*. Our approach is informed by the observation that the Web offers not only information on various subjects but also insight into how to effectively structure knowledge for human consumption. An example of such structure is Wikipedia's detailed table of contents, subtitles, and in-line references. Similarly, scientific articles have specific sections with headers and in-line citations, though this is scoped to support a single paper. When the user queries a topic in *Relatedly*, the system generates a list of subtopics (like a Table of Contents) and provides a descriptive summary for each. It does so by leveraging related work paragraphs from papers on the topic. It scaffolds exploring and making sense of this topic using features such as dynamic re-ranking and highlighting to spotlight unexplored dissimilar information and low-lighting redundant information. Using *Relatedly*, scholars explored twice as many scientific papers and subtopics and generated more coherent, insightful, and comprehensive topic outlines as compared to using a standard paper list within the same time. This system illustrates opportunities for leveraging prior effort (i.e., existing content and structure in related work section paragraphs written by previous researchers) to scaffold new users' discovery and synthesis journeys.

1.4 Augmenting Creative Workflows

Chapter 7: The Practitioner Perspective on Creativity Support Tool

Adoption

Most cognitively complex processes span multiple operations across multiple tools. With the rapid development of new tools, creative practitioners (e.g., designers, artists, architects) have to constantly explore and adopt new tools into their practice. While HCI research has focused on developing novel creativity support tools, little is known about creative practitioner's values when exploring and adopting these tools. This is an important gap to address, to help us present the user with the right capability at the right context and right time in their workflow. So, we collect and analyze 23 videos, 13 interviews, and 105 survey responses of creative practitioners reflecting on their values to derive a value framework. We find that practitioners value the tools' functionality, integration into their current workflow, performance, user interface and experience, learning support, costs and emotional connection, in that order. They largely discover tools through personal recommendations. To help unify and encourage reflection from the wider community of CST stakeholders (e.g., systems creators, researchers, marketers, and educators), we situate the framework within existing research on systems, creativity support tools and technology adoption.

Chapter 8: Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI

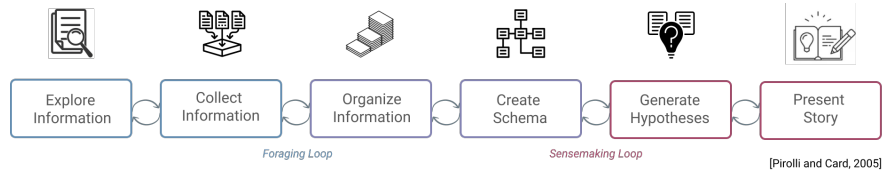
Generative AI (GenAI) models are transforming information work and creative practice by producing text, imagery, and other media that can rival human output given well-crafted prompts. Today, most existing sensemaking and creativity support tools only

leverage these models to support individual tasks, not focusing on the broader cognitive process that can include planning, exploration, ideation, reflection, and refinement. To identify challenges and opportunities for supporting the cognitive process, we interviewed ten practitioners who had successfully completed a creative project using GenAI. We find that they have trouble aligning model outputs with user intents, managing context, and operating across fragmented tool ecosystems while prioritizing their need for creative control and empathetic system responses. Our observations let us derive a set of factors that capture practitioners' perceived roles, challenges, benefits, and interaction patterns when creating with GenAI. Our insights serve to encourage reflection from the wider community of Creativity Support Tools and GenAI stakeholders, such as systems creators, researchers, and educators, on how to develop systems that meet the needs of creatives in human-centered ways, we propose design opportunities and priorities based on these factors.

Chapter 9 Amethyst: Enabling Affordances for Specifying and Referring to User-Generated Context Fosters Creativity and Human-Centered Orchestration of GenAI

Based on the insights and design guidelines derived from the above two studies, we implement *Amethyst*, a smart notebook that aims to address the above challenges by leveraging GenAI to support cognitively-complex processes involving sensemaking and creativity in an integrated, context-aware manner. *Amethyst*'s features include supporting goal decomposition, grounding prompts to specific contextual information, modulating GenAI output through simulated expert personas, and providing a range of in-line, nonblocking operations. We evaluated the potential of *Amethyst* to support the creative process through a within-subjects user study ($n = 12$), comparing *Amethyst* to a baseline

condition of standard tools such as web search, LLM-based chat, and digital notebooks. We find that participants generated more novel, feasible and creative ideas and preferred using *Amethyst* as it helped interact with GenAI in a more integrated, empathetic and context-aware manner.



I. Observing Information Seeking & Sensemaking Workflows Longitudinally

Ch. 3. A Longitudinal Study of Information Seeking & Sensemaking

II. Information Exploration

Ch. 4. The "Active Search" Hypothesis [CHIIR'21]

Ch. 5. CoNotate [CHI'21]

III. Synthesis

Ch. 6. InterWeave [UIST'22]

Ch. 7. Relatedly [CHI'23]

IV. Creativity

Ch. 8. Practitioner Perspective on Creativity Support Tool Adoption [CHI'22]

Ch. 9. Evolving Roles & Workflows with GenAI [C&C'24]

Ch. 10. Orchid

Figure 1.1: This thesis contributes the following: (i) Empirical studies advancing our understanding of how people think, learn, and create, leveraging online information and generative AI in the real world, and (ii) Interaction techniques and algorithms that seamlessly integrate knowledge from the Web and GenAI into user's work contexts to promote information discovery, synthesis, and creative insight. These approaches are implemented as tools and evaluated in lab studies and real-world deployments.

1.5 Thesis Statement

Together, these user studies, systems and their evaluations support my thesis statement:

Mining rich contextual signals from cognitively complex work can help intelligent systems scaffold information exploration, synthesis and creativity.

Each study observes how people work at different parts of the information exploration (ch. 3), sensemaking (ch. 6, 2) and creative process (ch. 8, 7). Each system introduces an approach for inferring contextual signals from work patterns: mining an individual's

unstructured artifacts for knowledge gaps and patterns in *CoNotate* (ch. 4), emerging sensemaking structures in *InterWeave* (ch. 5), existing knowledge structures on the Web from previous users in *Relatedly* (ch. 6), and presenting users with affordances to specify and refer to relevant personal, project-level and external contexts in *Amethyst* (ch. 9). These are evaluated in user evaluation studies that find that the context-aware systems' approaches promote information exploration, synthesis and creativity.

Chapter 10 presents a discussion of the challenges that remain and open questions prompted by this research. This chapter explores the future of building systems that balance automation with other cognitive and social goals, such as learning, critical thinking, creativity, and collaboration. It explores ways to improve the applications and generalizability of context-aware interaction mechanisms introduced by the systems in this dissertation, to allow them to scale to other cognitive tasks that might also require "good friction" in interactions. It also looks into how to build on these interaction mechanisms to address the challenges of collaborative information work. Additionally, this chapter discusses how future work could use the empirical insights gained about user behavior and human cognition from the studies in this dissertation to derive more human-centered evaluation metrics for intelligent systems. This dissertation demonstrates the potential of distilling and integrating the immense knowledge on the Web within the context of everyone's workflows to help augment cognitively complex work.

Chapter 2

Observing Information Seeking & Sensemaking Workflows Longitudinally

Searching and exploring online is integral to information work. To shape the future of information search and synthesis tools in a more contextual and human-centered manner, we must understand how people search and synthesize information over the course of such projects. In this paper, we collected and analyzed search activity logs, work document activities, and self-report data from 15 knowledge workers over their one to six-month-long projects. We developed a novel experimental protocol and browser extension to observe their natural behavior while preserving privacy. Our findings provide insights into patterns of knowledge work, search and sensemaking behavior, information needs, and user challenges across creative activities and temporal stages of a project.

2.1 Introduction

Knowledge workers make up 30-50% of the global workforce, and this percentage has been growing steadily in today's information age [287, 373]. The nature of Knowledge work often requires people to search the Web and make sense of information to produce some creative outcome (e.g., scientific papers, policies, startups, news articles, etc.) [462]. This goes beyond merely looking up facts, navigating to webpages, or making a purchase [54, 276]. Creative knowledge work involves grappling with complex and exploratory information goals across diverse sources, work sessions, and project phases [29]. With the rapid advancements in Large Language Models (LLMs), we stand at a pivotal moment in the evolution of Web and AI technologies where we can re-imagine such systems to better support complex creative work. This starts with understanding how people probe and interact with online information throughout creative work.

Prior work has started to shed light on how people search online during complex creative projects. A 2019 survey study [461] found that people use web searches across a range of creative domains, such as the arts, writing, cooking, and technical projects.

Most prior research on complex creative search has taken place over brief periods of time, typically using simulated tasks in a lab setting [204, 447, 316]. However, a 2020 diary study [462] captured qualitative insights on how people search when working on creative projects over a two-week period and discovered that creatives preferred different information resources for various purposes at each creative stage. For instance, people often use images to help support ideation, Q&A sites to find tips and recommendations from other creators, and social media to collect feedback on their creative projects. While prior work hints at the complexity of search and sensemaking practices during creative work, understanding the full, rich context requires more quantitative and qualitative data collected over a longer period of time.

To better shape the future of information search and synthesis tools in more contextual and human-centered ways, we must zoom out of what a searcher types into the search bar and clicks on. Instead, we must understand the knowledge worker's context around the nature of creative work, when they search what and why, when and how they synthesize information in different ways, the challenges they face, and how they want tools to better support their creative process. Therefore, we observe 15 real-world knowledge workers over the course of their projects. These observations ranged from 1-6 months long (avg. 2.5 months) and used a mixed-methods approach to collect rich self-report perspectives combined with activity logs of search and sensemaking behavior. This study builds on prior knowledge by extending the time scale of analysis (i.e., months rather than weeks or hours), level of data richness (i.e., quantitative and qualitative data), *and* sources of data (i.e., work documents and search activity). To observe participants' natural in-situ behavior longitudinally, we developed a novel experimental protocol and browser extension that logs participants' search activity and sensemaking work in their associated work document. Participants could use the tracked document for various tasks, including taking notes, organizing information, and synthesizing insights into

writing. The data was collected in a manner that protects privacy, values transparency, and preserves the participant's agency. To collect qualitative data, we asked participants to fill out an online survey every week where they could view an activity visualization and retrospectively reflect on their search and synthesis behaviors. Over time, this protocol and visualization method proved to be a stable way of not only collecting data but also providing opportunities for the participants to reflect on and share perspectives about their own work patterns. To get a rich understanding of how people search and synthesize information within different units of analysis, we partitioned the data into different time scales: by overall projects, by project stages (early, mid, and late in the overall timeline), and by creative activities (discovering insights and research, defining project goals, generating new ideas, refining and implementing ideas, and communicating ideas and artifacts). In this study, we investigate the following **research questions**:

- **RQ1:** How much time do participants spend engaged in and away from information search and synthesis work over the course of a creative project?
- **RQ2:** How do the different creative activities unfold over the timeline of a creative project?
- **RQ3:** How do participants search and synthesize during each creative activity and each project stage?
- **RQ4:** What information needs do participants want to fulfill during each creative activity?
- **RQ5:** What challenges do participants face during each creative activity and phase of a work session? And what kind of support do they want systems to provide to address these challenges?

Quantitative analysis of web search and work document activity logs and qualitative

analysis of weekly survey responses reflecting on this activity validate prior self-report findings and give us additional insights into the nature of knowledge work. First, we find that most participants exhibited a *double peak* in productivity – spending more time during early and later sessions of the project but showing a lull in activity during the middle of the project. They also took longer breaks between work sessions early on, and these breaks progressively shortened as the project advanced. Second, we can find quantitative evidence demonstrating that creative processes are non-linear and iterative in nature. For instance, we find that while the activity of discovering insights largely takes place earlier in the process, participants continue to discover new insights even in the mid and later stages of the project.

Third, participants actively search and synthesize information across all creative activities – including stages generally assumed to be offline or mental processes such as defining and scoping their project, generating new ideas, and refining and implementing ideas. Fourth, delving deeper into why they search, we add to previous research on the distinctive types of information needs workers have during each creative activity. For instance, we find that participants search for examples of other finished or drafts of in-progress projects to help define their own projects and generate new ideas. Fifth, we list the unique search and sensemaking challenges each creative activity and phase of a work session presents and how participants envision future tools helping address them.

We also found that artifacts generated along their messy iterative creative journeys can encode rich contextual information. This includes the user’s goals, what they already know about a topic (or what they are missing), how they feel, their design preferences, how far they have come in their project, how they link what they know to what they are finding, and how they structure their thoughts.

Overall, this study makes the following contributions:

- **Empirical longitudinal observations of search and synthesis behavior of knowl-**

edge workers engaged in complex creative projects: that validate prior findings with quantitative analysis and suggest practitioners have different information needs, work patterns, and challenges depending on one's stage in the project or creative activity engaged in.

- **A novel mixed-methods approach and web browser extension that collects and visualizes web search and sensemaking activity:** that logs in a manner that provides transparency around what data is being collected and gives participants control over what data to share with the researchers, while also enabling real-time reflection on their own behavior patterns.

2.2 Related Work

This research builds on prior work to capture and understand web search and synthesis behavior, and to study how knowledge workers engage in search for creative work.

2.2.1 Information Seeking, Sensemaking and Creativity in Knowledge Work

Most people use web search to look up facts or to get timely information to complete some other task. But people increasingly use the Web to explore, learn, and do more complex information synthesis for more open-ended goals. For example, academics reviewing literature, designers exploring which tool to use, startup founders performing market analysis, or individuals exploring, learning and making decisions like where to vacation. *Exploratory searches* involve multiple iterations and return sets of information that require cognitive processing and interpretation and often require the information seeker to spend time scanning/viewing, comparing, critically assessing and making

qualitative judgments before being integrated into personal and professional knowledge bases [439, 29]. Searches done during knowledge work typically involve a combination of exploratory and focused information retrieval activities [123, 439].

During the exploratory knowledge discovery process, people engage in sensemaking activities as they move through the information space. They take notes, gather information, and create representations to organize information to free their mind from having to recall everything [420, 280, 256], and from having to mentally synthesize all the information [225, 206, 275, 163]. This process of encoding information into external representations to answer complex, task-specific questions is called *Sensemaking* [358, 357]. People search for information by interacting with search results, web pages and other information sources. As they process this information, they collect and curate relevant and promising information by clipping and extracting information from web pages. Then, they organize it into structures, haphazardly at first and later systematically into a schema. Schemas are representations of the knowledge and understanding gained during the exploration and sensemaking process. Schemas can be essay outlines, comparative pros and cons lists, concept maps, etc. The searcher continues the sensemaking process until they have developed a concrete, well-tested schema. Schemas or sensemaking structures can change slightly to assimilate new information or significantly to accommodate new paradigms and perspectives. As the searcher develops a more concrete and polished schema, they progress to a state where it can be presented in a narrative that makes sense - for example, in an essay or article [358, 329].

Creativity is generally defined as producing "something original and worthwhile" [393]. This can be generating novel and appropriate ideas, processes, or artifacts in knowledge work [393]. The creative process, though defined in various ways across disciplines, is generally seen as a journey encompassing problem discovery, idea development, and final delivery [101, 307]. Web search is one of the most commonly used tools during

this creative process [381]. However, only recently have we started studying how people search online during creative tasks [346, 461, 89, 316]. While prior work hints at the complexity of search and sensemaking practices during creative work, understanding the full, rich context requires more comprehensive data collected over a longer period of time.

2.2.2 Work Patterns, Information Needs, Challenges During Creative Work

While the HCI and Information Retrieval research communities have gathered insights on exploratory and creative, most of the prior work has only observed individuals during a short period of time in controlled lab studies [204, 447, 444] or gathered data from over a large group of people [397]. We aim to build on work done so far to add rich qualitative and quantitative data-driven insights from data collected longitudinally throughout a knowledge worker's real-world project. In this section, we organize prior research along with our three research questions: investigating work patterns around time spent and browser interactions, information needs, and user challenges during creative knowledge work.

Work Patterns:

In previous research efforts, scholars have explored the relationship between search behavior and learning outcomes, focusing predominantly on laboratory studies. The *Search-as-Learning community* has contributed by developing tasks and measures based on Anderson and Krathwohl's Taxonomy of Learning, which identifies six cognitive processes: remember, understand, apply, analyze, evaluate, and create [237]. Jansen et al. [204] found that search tasks at the apply and analyze levels required more effort in

querying and result exploration than tasks at other levels. Conversely, Wu et al. [447] discovered that search interaction increased with higher levels of cognitive learning, as indicated by time on task, the number of queries, results clicked, and URLs visited. More recent work has started extending this exploration into the realm of creative work beyond learning [316]. They find that engaging in more active and diverse search behavior, characterized by frequent and varied queries and exploring a greater number of web pages, was associated with greater progress in the early stages of design, resulting in the accumulation of facts, insights, and refined problem frames. In 2000, [418, 416] studied students' problem stages in writing research proposals, connecting them to changes in search tactics, term choices, and relevance assessments. Both studies highlighted the interconnectedness of task performance stages with information types, search tactics, and relevance judgments, although the applicability of these models to the present WWW3 landscape remains uncertain. Furthermore, these studies focus on creative work as primarily one type of activity instead of iteratively working through a range of creative activities, including discovering insights and research, defining project goals, generating new ideas, refining and implementing ideas, and communicating ideas and artifacts (as described by [101, 307, 393]). Our study builds on these insights to add how workers spent time engaged in and away from searching and sensemaking work over the course of a project and maps how different creative activities unfold over the project's timeline.

Information Needs:

Zhang et al.'s 2019 survey study found that people use web search across a range of creative domains such as the arts, writing, cooking, and technical projects [461], and across creative stages like creating ideas, combining ideas, executing plans, not discovery and definition and communication of ideas. This research also found that people searched for different resources and tools depending on the creative stage of the

project. For example, users in the discovery stage are likely to use search engines, while those creating ideas may lean more on image galleries and social media [461]. Zhang et al.'s 2020 diary study [462] conducted over the course of a two-week period built on these results and found that during creative work people search for procedural information, domain information, tips/opinions/recommendations, information about specific topics, and inspiring or motivating information. This study updates and builds on this knowledge by understanding the evolution of information needs in the context of different creative activities in a project.

Search and Sensemaking Challenges:

In the 1980s and early 2000s, Kuhlthau's Information Search Process model provided valuable observations by interviewing secondary school students throughout an extensive research assignment. This model revealed a common trend in more complex information-seeking tasks, where feelings of uncertainty tended to rise before gradually diminishing during the focus formulation and construction stages of the process [239, 240]. This rise in uncertainty was frequently unexpected and caused apprehension and confusion in some searchers to the point of obstructing the task. Recent studies have expanded on this, indicating that participants encounter challenges related to uncertainty even earlier in the process, particularly when scoping broad and ill-defined information needs into queries, as well as when assessing the usefulness of information [316]. A week-long diary study of daily challenges faced by information workers finds interruptions and task-switching challenges and highlights the limitations of existing software in supporting the resumption of complex, long-term projects [112]. Our study builds on these insights by distilling distinctive challenges experienced during each creative activity (discovering insights and research, defining project goals, generating new ideas, refining and implementing ideas, and communicating ideas and artifacts) and phase of a work session (beginning, during,

ending of a work session and in-between two sessions). We go a step further and gather user insights on how to build tools to address these challenges.

2.2.3 Methods To Study Web Search

Researchers have employed a variety of methods to study web search and sensemaking patterns. These methods include analyzing search engine and web browser logs (e.g., [215, 405, 438]), gathering self-report data through surveys, interviews, or diary studies with end-users (e.g., [296, 462]), and recruiting participants for controlled tasks (e.g., [204, 447, 444]). However, as is the case with any methodology, there are trade-offs to consider. Logs can provide in-situ data from a large user base but may lack qualitative depth. Self-report data, while valuable, may exhibit gaps or inconsistencies compared to observed behavior. Additionally, controlled, in-lab task performance may exhibit unexpected differences from natural search behavior. To record a user's interactions with search browser during a search session, IR and HCI researchers have developed systems for activity logging that record queries issued and web pages visited over time, click depths, mouse trails and movements, eye fixations and saccades, dwell times, key-presses, etc. [285, 399, 318, 433, 43].

This study also extends prior methods by extending the time scale of analysis (i.e., months rather than weeks or hours), level of data richness (i.e., quantitative and qualitative data), *and* observing multiple sources of information (i.e., search engine and work document logs). To log interactions with the search browser and work documents when working on a long-term project across multiple work sessions, we develop a custom web browser extension that logs in a privacy-protecting, transparency-preserving manner, which gives participants control over what data to share with the researchers, while also enabling real-time reflection on their own behavior patterns. Our study adds rich data and builds on this prior work by triangulating a mixed-methods approach by logging

activity with privacy controls and structuring self-reports using the participants' data as a reflective prompt.

2.3 Method

We conducted a longitudinal study to investigate the information needs, search and sensemaking behavioral patterns, and challenges faced by a range of knowledge workers throughout a creative project. As data for this investigation, we collected search logs and activity history from document where participants took notes and synthesized information into a creative outcome. Participants submitted self-report reflections through a digital survey every week through a digital survey every week throughout their projects, which lasted between one to six months long. We applied mixed methods to analyze these data at different time scales (by session, by stage and by project) to understand how the information needs, search strategies and challenges played out over time.

2.3.1 Participants

We chose purposeful sampling [48] as a recruitment strategy, mixing direct contacts as well as recruitment at a large public university. We recruited a diverse mix of participants across different practices, ages, organizations, genders, and locations. We recruited 15 participants (eight female, seven male, average age 29.8 years) across six creative fields, including scientific research, product design, data visualization, product management, machine learning engineering, and policy-making (see Figure B.1). They were also based in 10 different locations across the US, Germany, and India.

PID	Profession	Project	Months
1	Policy Advisor	Economic policy changes in India and Sri Lanka in response to Ukraine war	1
2	Startup Founder	Rugged, portable 3D printer that is capable of being used in harsh environments	4
3	Startup Founder	Rugged, portable 3D printer that is capable of being used in harsh environments	4
4	Startup Founder	Cut plastic pollution with seaweed-based alternatives for retailers and consumer goods	6
5	Data Journalist	Data visualizations following the 2023 Berlin elections	1
6	Technology Consultant	Policy brief on cybersecurity education and training at a technology company	1
7	Product Manager	Researching how to integrate new technology into an existing data engineering pipeline	1
8	Machine Learning Engineer	Building a scientific paper classification system based on paper metadata and co-citation networks	6
9	Public Health Research	Researching how globalization affects disease propagation - a case study of Ebola in N.Africa	1
10	MD-PhD Researcher	Research the Role of Mul1, a Mitochondrial Localized E3 Ligase, in the Heart	1
11	Immersive Technology Researcher	Training for doctors and systems to make diagnoses by contouring medical images	1
12	NLP Research Scientist	NLP methods for evaluating generative algorithms in a human-centric way	3
13	Cognitive Science Researcher	Writing a research paper on healthcare worker's information workflows	3
14	PhD Student	Designing a tool to support content creation	1
15	Post-Doc Researcher	Researching how to improve design feedback providing systems in human-centered ways	4

Figure 2.1: Participants spanned many different professions and worked on projects related to a range of creative goals over different time periods.

2.3.2 Browser Extension for Data Collection and Visualization

To collect data longitudinally in a manner that protects privacy, values transparency, and preserves the participant's agency, we built a custom Chromium browser extension. Participants could easily view logged data by clicking on the browser extension's homepage where they had full control over data collection, including the ability to start and stop logging, delete collected data points, and share data with the researchers. Partic-

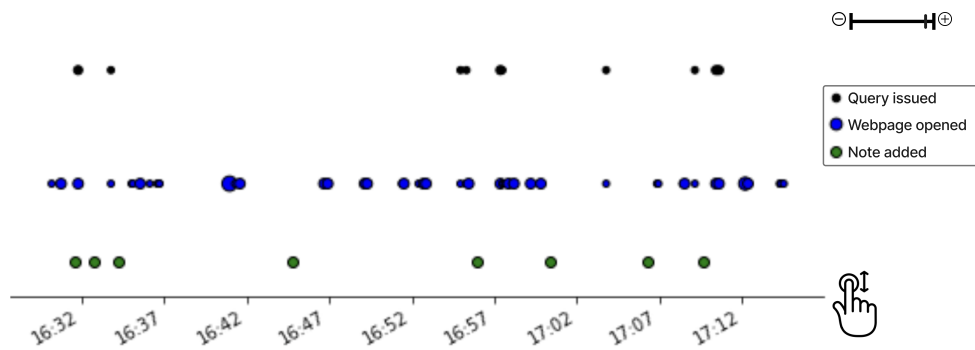


Figure 2.2: An illustrative screenshot showcasing the zoomable visualization designed for a participant to reflect on their weekly search and sensemaking patterns. This screenshot shows activity for a single hour-long work session, and plots queries issued, web pages opened and notes added to the work document.

Participants provided us with a URL for the key work document used for note-taking and sensemaking during their project (i.e., a notion workspace, overleaf document, google doc, etc.). The extension monitored if their work document was open and active. If the work document was active and the log system was currently turned off, the system would send participants a notification reminding them to turn on logging. Additionally, to avoid unnecessary data collection, the extension would stop logging when it detected inactivity in the work document tab for more than 20 minutes, notifying the participant about the logging status.

The extension prioritized the privacy and security of data through encrypted communication and maintaining the same encrypted ID across sessions. To enable participants to reflect on their own work patterns, the real-time logged data can be seen in a tabular view or a zoom-able time series visualization (see Figure 2.2), which offers comprehensive views of Web search activity over multiple days, weeks and months. We ensured that it was a zoomable visualization to prevent occlusion of data points when there are periods of lots of activity and some periods of no activity. For user convenience, participants could log in using their Google accounts, integrating the extension seamlessly into their

daily workflow.

In terms of implementation, the browser extension was built using React JS framework, and the visualizations were generated in real-time using the d3.js library with the d3-timeline package. Activity log data was stored in real-time to a Firebase database. The open-source code for the custom browser extension can be found here (to add when de-anonymized).

2.3.3 Procedure

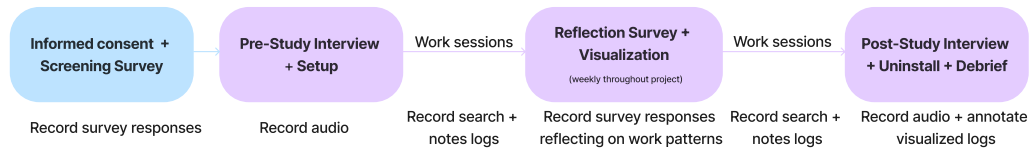


Figure 2.3: Participants underwent a brief screening, followed by a one-hour orientation session with the research team. Weekly surveys tracked their search and sensemaking activities, and a post-study interview and debrief concluded the study, including uninstalling the extension.

As part of the recruitment process, participants answered an informed consent form, and brief screening and demographics questionnaire that collected information about age, gender, occupation, and the creative project that we would observe, its timeline and how they thought they might use online information resources during the project. As noted above, we excluded participants who were under the age of 18, whose projects were too ill-defined or too long for us to follow i.e., more than 6 months, and/or if it did not require complex searching and sensemaking across sources, sessions and stages. To get participants setup at the beginning of their projects, the research team met them for one hour to obtain informed consent, to review study procedures, and to walk them through how to use the browser extension and weekly surveys to participate in the study.

After the participant received training and felt confident in how to monitor and edit their logged data, they could start searching and working on their creative projects.

To collect qualitative perspectives, the research team shared a survey link every week asking participants to reflect on their own work patterns and data, verify the data collected, delete any unnecessary data, and submit it to our database. To support reflection, participants viewed a visualization within our custom browser extension (see Figure 2.2, and details in the section 2.3). To gain insights into their work patterns. We also sent out regular reminders and messages to keep participants engaged throughout the study. The institution's ethics review board approved the recruitment process.

The reflection survey questions included: a semantically zoomable data visualization of data collected from the previous week (see Figure 2.2 for example), and for each work session it included questions (1) asking them to map each work session to a creative activity, (2) what were the information needs sought, (3) challenges encountered in the overall work session, (4) challenges encountered at the beginning, during, ending of this work session and between work sessions, (5) tool support envisioned to overcome experienced challenges.

When they were close to finishing up their projects, participants indicated that they were done in an email and we set up a post-study interview to reflect on their overall process, challenges, and strategies; we also helped participants uninstall the extension and thanked them for their participation.

2.3.4 Measures

To observe and analyze differences in search and sensemaking patterns, we collected the timestamp and content of search queries, opened web pages, and edits to their work document. These data may provide insights into the users' information retrieval habits, the extent of their engagement with external sources, and the evolution of their

understanding and creative products as they interact with the information.

To observe information needs, search and sensemaking strategies, and challenges across the project, we analyzed qualitative self-report data. The audio recordings underwent an intelligent transcription, removing pauses and filler words and doing minor grammar adjustments. The subsequent analysis encompassed open coding, where data were initially categorized without predefined labels, followed by thematic clustering using affinity mapping [48] to uncover overarching themes and patterns within the dataset. First, two coders independently coded two randomly chosen participants' data through open coding. Then, these two discussed the emerging themes and agreed on a common vocabulary. Once similar codes and themes were identified across the two participants' data with no significant discrepancies, the two researchers finalized the coding scheme and shifted to a focused coding approach. To ensure inter-rater reliability [359], we compared the independent coders' results. There was a 75.12% to 98.82% agreement level across all code categories. Given the moderate to high agreement, one of the coders independently coded the remaining participants based on the agreed coding scheme.

2.4 Findings

With the rapid advancements in Large Language Models (LLMs), we stand at a pivotal moment in the evolution of Web and AI technologies where we can re-imagine such systems to better support creative knowledge work. In this study, we unpack the nature of search and sensemaking work around creative projects to inform how we build tools to better support this complex process. In this section, we report the findings from our longitudinal study observing 15 creative practitioners over the course of their projects, which investigated:

- RQ1: How do participants spend their time engaged in and away from search and

sensemaking work over a creative project?

- RQ2: How do different creative activities take place across the stages of a project?
- RQ3: How do participants spend their time searching and sensemaking during each creative activity and project stage?
- RQ4: What information needs do participants want to fulfill during each creative activity?
- RQ5: What challenges do participants face during each creative activity and phase of working on a session? And what kind of support do they want systems to provide to address these challenges?

2.4.1 Participants allocate more time to work during the early and late project stages while taking longer breaks early on that progressively shorten as the project advances.

To understand **how people spend time searching and sensemaking over the course of their creative projects** we first analyze the total time spent actively searching and working during each work session. A *work session* is defined as the time between the start and stop of logging. On average, the length of a work session was 118.3 mins, and participants had an average of about 92 work sessions. Then, to gain insight into the evolution of time allocation across the various stages of a creative project, we analyze each participant's total work sessions by dividing them into three equal parts, referred to as "project stages": early, mid, and late. To compare the average time spent working across project stages (early, mid, and late), we used a one-way between-subjects ANOVA, and there was a significant difference level ($F(2, 15) = 4.95, p = 0.01^*$).

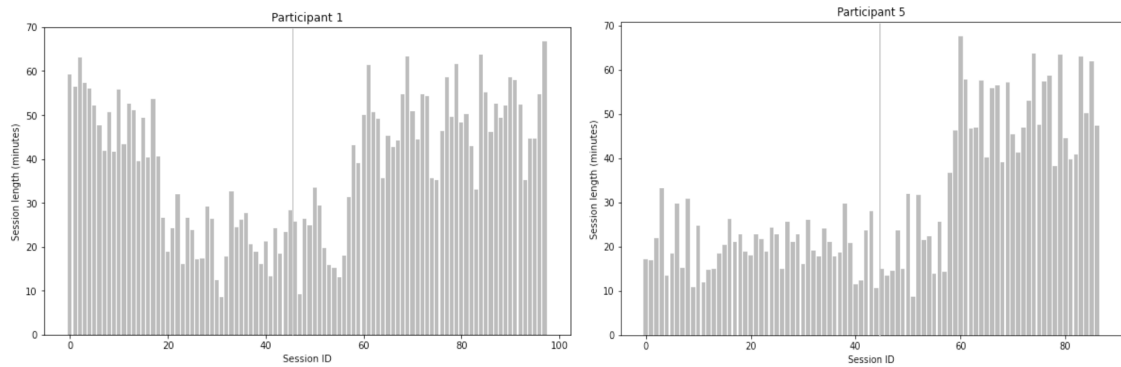


Figure 2.4: Work session length in minutes across the entire project for two participants. Participants typically exhibited one of two work patterns illustrated by P1 (left) and P5 (right): Double peak shows spikes of activity in the early and late sessions of the project, but a lull in the middle, and late Peak shows a spike in work activity in the later stages, after a steady amount of work done up to that point.

When we visualize how time was spent by *each participant* across the course of their individual projects, we see two patterns emerging: Most participants exhibited a **Double Peak** of activity in the earlier and later stages of the project but had a lull in the amount of time spent per session in the middle of the project. This might suggest empirical evidence for the “messy middle” when creators engage in conversations and deep thinking that might not have been captured by the activity logging system (for example, see Figure 2.4 (left)). The other trend was that of a **Late Peak** in work activity after a steady amount of work done at the beginning and middle stages of the project (for example, see Figure 2.4 (right)).

In the post-study interview, P01, a participant who exhibits the Double Peak, reflected on the lull *“During this phase, we held extensive discussions with experts, stakeholders, and our team, exploring economic nuances, assessing policy impacts, weighing pros and cons, and engaging in numerous brainstorming sessions. These activities formed the foundation for the eventual policy recommendations.”* Similarly, P03, also a Double Peak worker, explained the lull in the middle as *“At first, fueled by excitement, I delved into research, market analysis, and product development, driven by the thrill of something*

new. Learning as much as possible, I later engaged industry experts and stakeholders to refine our business model with a customer-centric approach. Once everything was in place, I worked tirelessly to execute and deliver.”

On the other hand, P09, a participant who exhibited a Late Peak, reflected on their work pattern, *”Initially, I diligently gathered and analyzed data, conducted literature reviews, and laid the groundwork for my white paper. In the final stretch, I raced to finalize my findings, refine arguments, and collaborate with colleagues. It was a productive burst of activity to ensure a high-quality white paper.”* Similarly, P05 said, *”I first focused on building a solid foundation for our data visualizations, emphasizing accuracy and relevance. In the project’s final stages, my efforts intensified, involving refining visuals, integrating the latest election data, and ensuring our graphics communicated the most current information effectively.”*

To investigate **how people spent their time when not actively working on the project**, We compared the time spent between work sessions across stages of the project, using a one-way ANOVA to find significant differences between the early, mid, and late stages in the project ($F(2, 15) = 5.58, p = 0.01^*$). We found that participants took longer breaks in the earlier sessions than in the later sessions. Further, to understand **how the gaps before a session affected how they spent their time in the work session**, we do a correlation analysis and find that in sessions after longer breaks, there is more searching ($r = 0.40, p = 0.01^*$) and lesser time actively synthesizing information in the working document ($r = -0.16, p = 0.01^*$). This might indicate that creative workers re-orient to their previous creative activity by focusing on searching for more information rather than synthesizing the information they already had. P13 said, *”After a long break, where I’ve been thinking about the topic deeply and discussing it with collaborators, I often have many open questions or new ideas that I want to whet, so I dive into searching for information.”* P10 reasoned about this as, *”After a long break, I usually find myself*

mostly searching to refresh my perspective or fill in any gaps. The initial wave isn't about putting the information into place just yet, but more about finding where I left off and gathering any new insights I may have missed."

2.4.2 Creative activities take place in a non-linear, iterative manner across project stages

To understand **how different creative activities play out across the stages of a project**, we used data from the participants' weekly self-reflection surveys where they categorized what they did in each work session as a creative activity: discovering insights and research, defining project goals, generating new ideas, refining and implementing ideas or communicating ideas and artifacts. Additionally, we split all the work sessions in each projects' into early, mid and late work sessions or project stages. This lets us map each different creative activity to project stages.

Averaging the percentage of work sessions spent on each creative activity across participants, we find that Discovering Insights occurs mostly in the early work sessions of the project but does not stop and continue until the end of the project. The stages of Defining the Project and Generating Ideas happen throughout but peak in the mid-sessions of the project. Lastly, the stages of implementing ideas and communicating artifacts happen throughout the project but peak in the later work sessions. This illustrates the non-linear, iterative nature of creative work (see Figure 2.5).

2.4.3 Participants actively search and synthesize online information across all creative activities

To delve deeper into understanding how participants spent their time searching and synthesizing online information during the different creative activities and stages of a

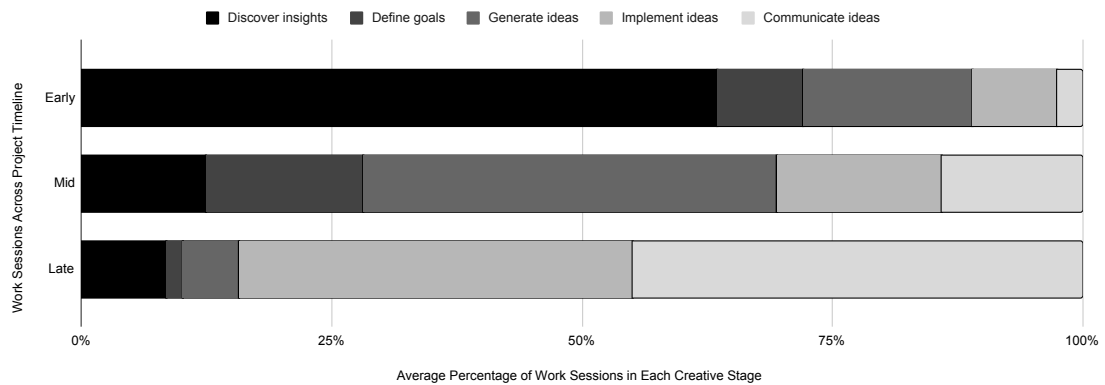


Figure 2.5: Average percentage of work sessions spent on each creative activity mapped to early, middle, and late stages of the project. Notice the non-linear and iterative nature of creative work. For instance, how participants continue to discover new information across time.

project, we analyze the participants’ search and work document logs.

To understand how participants spent their time searching and sensemaking during different creative activities, we sum the total time spent actively searching for information and the total time spent actively working in the document during the work sessions categorized in each creative activity (see Figure 2.6). We find that participants actively search and synthesize online information across all creative activities of their project. When Discovering Insights, participants spent more time searching than synthesizing. However, when Defining Goals, Generating, Developing and Communicating Ideas participants spent more time synthesizing information in their work document than searching. To statistically compare and contrast time spent searching vs synthesizing across each creative activity, we conducted a two-way ANOVA test and Tukey’s post-hoc test. There was a significant main effect of time spent on search being significantly different across the creative activities ($F(4, 15) = 3.34, p = 0.03^*$). There was another significant main effect of time spent sensemaking changing across the creative activities ($F(4, 15) = 2.53, p = 0.04^*$). Additionally, the interaction between time spent searching and sensemaking is significant ($F(4, 15) = 7.34, p = 0.02^*$) and the post-hoc revealed that

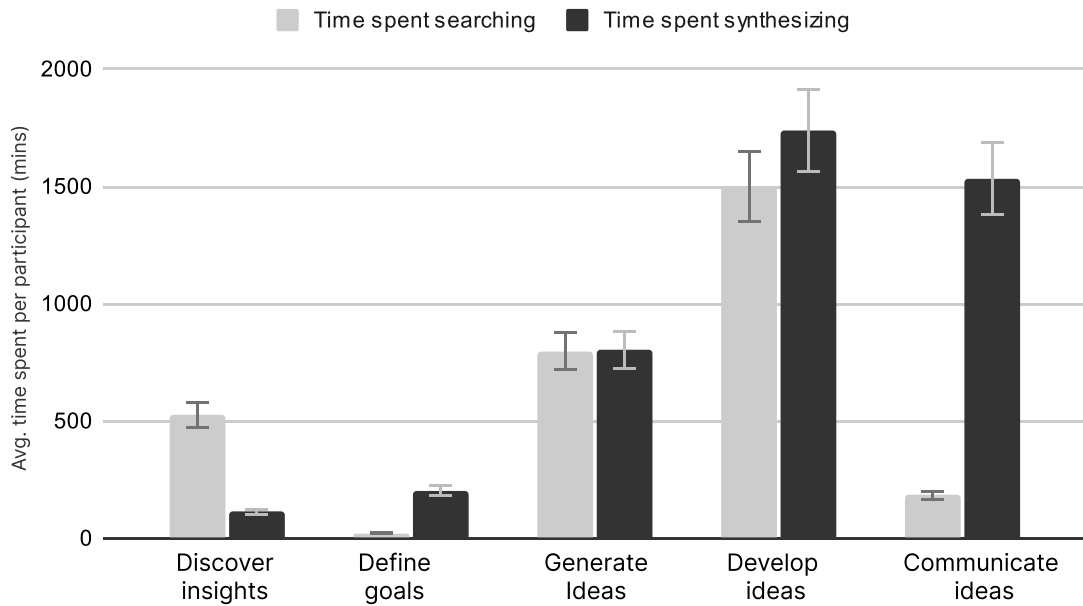


Figure 2.6: Average time spent (minutes) by a participant on searching and synthesizing online during each creative activity. When Discovering Insights, participants spent more time searching than synthesizing. Conversely, participants spent more time synthesizing information in their work document than searching as they worked to generate, implement, refine, and communicate ideas.

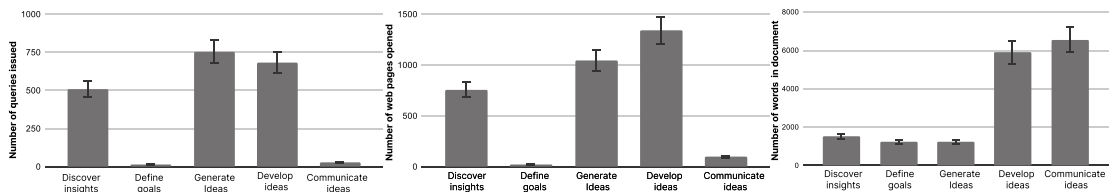


Figure 2.7: Average number of queries issued (left), number of webpages opened (middle), and number of words changed in the document (right) per participant during creative activities. The data show participants spent most of their search and sensemaking activity when implementing and refining their ideas. However, we also lot of search queries being issued, and webpages opened when discovering insights.

the mean difference was between time spent search and synthesizing information during three creative activities: Discovering Insights, Defining the Project, and Communicating ideas ($p = 0.04^*$).

To investigate participants' interactions with online information during each

creative activity, we look at the three main interactions with the browser: number of search queries issued, webpages opened and change in the number of words in the work document across the creative activities (see Figure 2.7). These interactions also reflect the same user behavior patterns as above. When Discovering insights, we also see many search queries being issued and webpages opened. Participants show most of their search activity when Implementing and Refining their ideas, and most of their words are added to the document when working to Communicate the idea in the final stage. A MANOVA test and Tukey's post-hoc test showed significant differences in search and sensemaking behavior across all stages, except between implementing ideas and communicating them.

2.4.4 Artifacts generated can encode rich contextual information

Here, we present the different types of artifacts users generate and the additional contextual signals they can encode about the user. Each participant had a document in which they tracked the overall objectives of the project and what needed to be done to complete it. Some were more structured than others. For example, one semi-structured one was the MD-PhD researcher trying to keep track of all the to-do items to complete the project in a single document. Another more structured example was the policy advisor keeping detailed notes for each stage of the project, its scope, schedule, resources, etc. Some were shared documents with other project stakeholders, whereas others were just meant for the individual themselves. These project planning documents can encode what are the project's goals and how far the user has come in their project.

All participants had a document for working with and making sense of the information they collected. The notes included background information about the topic, key concepts, specific details, useful information sources, and information to help with the broader work task. Key phrases in work documents and notes taken can reflect what they already know about a topic (or what they are missing) and could even encode patterns about how

their topic knowledge shifts and grows over time.

Notes were primarily structured in lists, and reflected a combination of linear note-taking strategies and grouping by information source or topical themes. The spatial organization of information within these documents can help us infer how the participant links what they know to what they are finding, how they structure their thoughts, and how this emerging schema can grow and shift over time.

Looking at the edit histories of these documents could give us insight into otherwise tacit knowledge about how each individual likes to work, what they prefer, and their implicit process for doing their work stage-by-stage.

2.4.5 Participants have distinctive information needs during each creative activity

To understand why participants searched during each creative activity, we thematically analyzed their responses to weekly self-reflection surveys which asked them to list their information needs during each work session, and categorize each work session as part of a creative activity.

When *discovering insights and research*, participants reported wanting to gather relevant information to form a broad overview and gain deep insights into their chosen topic. Interestingly, they did not limit their searches to just textual information but also looked for information across a range of modalities including data visualizations, news articles, research papers, YouTube videos, etc. For example, P01 a policy advisor had information needs like "Comparative study of rice export policies across countries", "Find USDA Rice Data", or "understand and evaluate India rice export restrictions and global food market effects data". Startup founder, P03 had information needs like "conducting market research through customer reviews about 3D printers", "Common problems with 3D printing machines + harsh environments", "Best 3D printers and alternatives".

When ***defining project goals*** participants searched for examples of projects, and particularly how they were scoped. Examples include reading policy briefs and particularly reading the policy recommendation statement (P01, P01), or papers' research questions (P01, P01), design briefs (P01, P01) and product specifications documents (P01, P01).

When ***generating new ideas*** to find inspiration, participants searched for examples finished projects or work in progress across different media including scholarly search engines, conference proceedings, dribbble, codepen, pinterest, github, etc. For example, the ML engineer P08 searched through paperswithcode, github, kaggle and codepen. Startup founder P04 looked for "plastic alternatives for consumer goods" in academic journals, industry reports and whitepapers, and blogs.

When ***implementing and refining ideas*** participants looked for procedural information like text and video tutorials, domain-specific language or expertise to address specific target audiences, or implement specific styles. For example, P08 the ML engineer, "*searched for coding guidelines and tutorials specific to Python web development to ensure the technique I was implementing met industry standards.*" Similarly, P03 said "*To create a marketing campaign using SEO, I followed YouTube tutorials on optimizing e-commerce websites.*"

When ***communicating refined ideas and artifacts***, participants searched for the most well-defined information needs here, compared to other stages. They often were re-finding information they had already found before, or looking up specific information like spellings, synonyms, code samples, and specific colors, images, materials, etc.. For example, P13 when writing their paper searched for 'another way to say tech savvy', 'viable definition', 'undoubtedly synonym'. Similarly, P12 "searched GitHub for code samples that execute the web crawlers more efficiently". Additionally, participants were interested in finding perspectives of users or domain experts on similar products to get initial feedback on their ideas to get initial feedback on their ideas. For example, when

crafting their pitch deck startup founder P02 said, "*similar pitch decks and how much funding and what feedback these pitches received from investors on ProductHunt and Pitches.*" Participants also searched for similar finished examples to see how to improve on their final deliverable. As technology consultant P06 said, "*I revisited my earlier policy recommendations as inspiration on how to refine it further*".

2.4.6 Each creative activity and phase of work session presents unique search and sensemaking challenges, and participants envision how future tools could help

To gain insights into the challenges participants encountered and the tool support they sought to overcome these challenges during each creative activity,

we thematically analyzed participants responses to weekly self-reflection surveys which asked them to list their challenges during each work session, and design opportunities for tools to address these challenges. Participants were also asked to categorize each work session as part of a creative activity.

When *discovering insights and research*, participants struggled to convert abstract goals into specific search queries. P03 stated, "*I knew I wanted to innovate in the sustainable packaging space but I didn't know the right terms to search for, particularly as I am new to this field.*" Another challenge was the overwhelming amount of redundant information, as P08 explained, "*The web is a sea of information - it feels like every article is just rehashing the same points, I found it hard to prioritize which ones to read.*" Similarly, P13 said, "*I can barely distinguish what's identical to what I've previously read making it hard to prioritize.*"

To help overcome this, the system could lower the barrier to cold start problem and help articulate fuzzy goals by giving the user either "*A comprehensive list of keywords*

that (they) should check out” (P12), “Recommend several gold standard papers in this area.” (P14), or offer a visual representation summarizing user’s query or topic. The system should also suggest related queries and topics for further exploration, encouraging users to constructively elaborate their search goal.

When ***defining the project***, achieving a comprehensive enough understanding of the topic to come up with a fresh or effective definition was challenging. P05 said, *“I thought I had a clear vision of my project goals, but when I started searching the web, there was a lot of information and a lot of it was conflicting. So, it’s hard to come up with a good angle without understanding this labyrinth of information”*. Another challenge was that it was emotionally a struggle letting go of previous work when the project was re-defined or re-scoped. P03 expressed frustration, saying, *“When the project vision evolved and we pivoted, I felt like my hours of searching on the original angle were wasted.”* To address these challenges, participants suggested *“topic maps or knowledge graphs depicting inter-connectedness of ideas explored and worked on so far” (P05), and “version controls or rollback features to help painlessly and safely let go of previous work, and help do more trial and error” (P13).*

During the ***idea generation stage***, participants found that search results, though relevant, were not stimulating or diverse enough. P11 noted, *“I was trying to create a novel concept but often the search results were just providing me with what’s already out there, and it’s hard to keep coming upon the same redundant information, instead of finding what the gaps are or exciting aspects of each are.”* P03 said, *“The problem with these search results is that they’re often relevant but not inspiring-ly diverse.”* To mitigate this, participants suggested *“prioritize both novelty and diversity of search results” (P06). “Sometimes, presenting information not just from well-known sources, but also highlighting less popular yet valuable resources and perspectives can spark creativity.” (P11)*

In the *implementation and refinement phase*, the main challenge was tracking the evolution of topics and ideas. One user voiced his struggle, saying, *"It becomes hard to keep tabs on all idea adaptations and topic modifications across the project."* To help with this, as ideas materialize, systems should *'monitor the evolution of topics and ideas keeping users informed of any updates or changes in their domain'*(P05). Alerts or push notifications can be used to draw users' attention to these changes, ensuring their ideas align with the most recent developments.

When *communication of ideas and artifacts*, participants talked about redundancy across different mediums as a challenge. P07 said, *"Replicating the same design in PowerPoint, then in Word, then in our project app was a drag."* Additionally, scheduling breaks and work times proved problematic, with P06 user stating, *"Since we communicated globally, scheduling suitable times for in-depth discussion and rest was no easy task."* P15 also talked about how *"Deciding when to rest and when to work is challenging"*. To avoid redundancy across mediums and promote effective time management, participants suggested providing *"text summarization, generation and rephrasing"* (P04) that help users to communicate their work more efficiently.

To gain insights into the challenges participants encountered and the tool support sought to overcome these challenges at various phases of a work session,

we conducted a thematic analysis of their responses to weekly self-reflection surveys which asked them to list their challenges phased and tool support wanted to address these challenges, during each phase of work session (beginning, during and ending of a work session, and between two work sessions).

Beginning of a work session: Participants noted considerable difficulty resuming mental context at the beginning of each session. P14 exclaimed, *"Getting started is always the hardest part, I have to rebuild my focus from scratch"*, while P12 shared

frustration with recalling previous matters, stating, *"Every time I begin a new session, it's as if I am starting the project all over again. Remembering where I left off is really tough"*.

To help with resuming a work session, participants suggest that systems should provide functionality that facilitates context resumption. P11 said, *"Returning me to the state I was in at the end of the previous session – helping me recall where I left off or even just popping up the right windows in the right layout, including being at the point in the paper and notes doc that I was previously at"*. P10 said, *"Briefly help me to summarize tasks I have done and the to-dos."*

During a work session: The challenge in this stage was largely around prioritizing tasks and scheduling breaks. P09 noted, *"I find it hard to decide what task deserves my immediate attention and what can wait. It feels like a constant juggling act."* Scheduling breaks was a common struggle, with P15 expressing, *"It's overwhelming trying to balance work and rest. Figuring out when to be productive and when to rest can sometimes be as taxing as the work itself"*.

Participants suggest that their work environment could implement advanced task management features to mitigate these challenges. These features may include a smart task scheduler that learns user behavior and suggests optimal time for intensive tasks, and priority indicating marks that help users visualize and prioritize their tasks. An intelligent notification system could provide meaningful reminders for scheduling breaks based on user activities. For example, P12 wanted the system to help with task decomposition *"help me focus possibly through helping create a more directed sub-goal"*. Similarly, P04 wanted the system to help with directing focus even during a task *"automatically highlighting important and relevant points."* And it could even help direct focus between tasks *"keeping track of the work that I've done so far, and what to do items it checks of for me"* (P14).

Ending of a work session: Towards the end of sessions, most participants reported feeling a deep emotional connection to the project, so much so that their self-worth was impacted. P12 admitted, *"It's not just about being productive, my feelings of accomplishment and worthiness are tied to how much progress I've made"*, while P02 mentioned, *"I often end up feeling stressed and burnt out, even when I made considerable progress, simply because it feels like it's never enough."* A secondary challenge was having to terminate sessions abruptly due to a variety of factors. Examples ranged from boredom to hunger, mental exhaustion or events unrelated to the project. P02 shared, *"There's a lot that disrupts my work sessions - hunger, fatigue, you name it. Occasionally, it's simply a case of losing interest"*. P07 commented, *"Work emergencies often cut my sessions short, and so do personal issues like taking care of kids and family. Trying to navigate between everything and maintain productivity is extremely challenging"*.

To help with this, participants suggested that the system could incorporate features that support recommending effective termination points. Providing a brief summary of the user's activities during the session can enhance a sense of progress. Furthermore, a functionality that enables saving the current status of work and setting reminders or tasks for the next session can help users end their sessions in a controlled manner. For example, P14 said, *"Summarize what I have already done and then show it to me when I resume the working session next time."*. Similarly, P15 said, *"Helping me save the current state/progress in a way that it's easy to return, helping summarize what I've learned/done in a way that I can quickly glance and feel like I made some progress"*.

Between work sessions of actively working on something, participants wanted the search and sensemaking system to continue making progress on the project goal. P10 wanted the search system to, *"Make sure I don't miss anything important – finding relevant papers to my searches that I might not have come across; continuing to search in the background when I've moved on to other parts of the research process."* P11 wanted

something similar in a search and synthesis system "*monitor the latest new publications and inform me of new relevant papers.*". They wanted the synthesis of information to also continue. P13 said, "*it would be great if someone could finish synthesizing and cleaning up the notes I took and information I collected into neat, useful representations.*". The search interfaces or sensemaking environments could provide richer structures that can better reflect users' mental models, making it easier for them to (re-) orient themselves in the information space and resume work.

2.5 Discussion

This study and its findings unpack the nature of search and sensemaking done during creative knowledge work at the level of creative activities, project stage in the timeline, and phases of a work session. In this section, we discuss insights gained about work patterns, search and sensemaking patterns, information needs, challenges and how these insights can help inform the design of information-seeking and synthesis tools in more creative- and human-centered ways.

2.5.1 Insights on Creative Work Patterns

The average length of a working session observed in our study, an average of 118.3 min per session across 92 work sessions, is considerably higher than the prior studies on web search and sensemaking [316, 447, 204]. This supports prior reports with quantitative data that creative tasks are inherently complex, often requiring extended periods of deep work [101, 428, 307].

First, we find that most participants exhibited a *double peak* in productivity – spending more time during early and later sessions of the project, but showing a lull in activity during the middle of the project (Figure 2.4). The 'double peak' phenomenon echoes

the "messy middle" concept discussed in design research [41], when creators engage in conversations and offline thinking that might not have been captured by our activity logging system. The 'late peak' pattern could be because of the "deadline effect" [103], which asserts that individuals work more intensively when faced with an impending deadline. If it is the deadline effect, it is still interesting to see the use of search so actively in this state, alongside working in the document. While these worker patterns have been discussed in self-reports, it has not been quantitatively observed and analyzed using both logs and self-reports by participants.

They also took longer breaks between work sessions early on, and these breaks progressively shortened as the project advanced. Perhaps discovery-related activities can make steady progress despite less frequent touch points, while synthesis requires more sustained focus. It might also suggest that people work hard and long on their projects before a deadline. A correlation analysis finds the length of a time gap between work sessions correlates with the ratio of time spent searching vs. working in their document. This might indicate that creative workers re-orient to their previous activity by focusing on searching for more information, rather than synthesizing information already collected.

Second, we can find quantitative evidence demonstrating that creative processes are non-linear and iterative in nature (as suggested by practitioner reports [101, 307, 393]). For instance, we find that while the activity of discovering insights largely takes place earlier in the process, participants continue to discover new insights even in mid and later stages of the project (see Figure 2.5).

2.5.2 Insights on Search and Sensemaking Patterns, Information Needs and User Challenges

Log analysis finds that participants actively search and synthesize information across all creative activities – including activities generally assumed to be offline or mental processes such as defining and scoping their project, generating new ideas, and refining and implementing ideas.

Delving deeper into why they search during each creativity activity, we analyzed their survey responses reporting their information needs. When discovering insights and research, participants reported wanting to gather relevant information to form a broad overview and gain deep insights into their chosen topic. Interestingly, they did not limit their searches to just textual information, but also looked for information across a range of modalities including data visualizations, news articles, research papers, YouTube videos, etc. When defining project goals, participants searched for examples of finished artifacts, particularly how they were scoped and what lessons could be applied to their current design scenario. Similarly, when generating new ideas to find inspiration, participants searched for examples of finished projects or works-in-progress to help refine or pivot their ideas. When implementing and refining ideas, participants looked for procedural information like text and video tutorials, domain-specific language or expertise to address specific target audiences, or implement specific styles. When communicating refined ideas and artifacts, participants tended to have well-defined information goals. Often their goal was to re-locate information they had already found before or search for specific guidance on how to communicate their creative work.

To understand how we might improve their experience when searching and synthesizing information, we analyzed survey responses to the questions: what challenges they faced in each work session, and how they want future tools to address these chal-

lenges. First, we analyze these at the level of different creative activities, then at the level of different phases of a work session. When discovering insights, our findings align with existing research highlighting the difficulty of articulating vague information needs during the project's early phases [240, 316, 302]. The challenge arises from grappling with uncertainty in the absence of clear information paths or concrete search queries. When Defining Project, participants struggled to attain a comprehensive understanding of the topic, emphasizing the ongoing challenges in navigating and defining project scope. While the challenges of redefining and pivoting a project are conventionally acknowledged, our study adds nuance by identifying the emotional struggle associated with letting go of previous work. During Idea Generation, participants find it challenging to come across diverse and novel information, with the outcome significantly influenced by the information presented by search engines. When Implementing Ideas, participants identified monitoring the project's evolution and development as a significant challenge. When Communicating Ideas, managing the avoidance of duplication across mediums and effective time management for work and breaks proved to be daunting for participants.

Analyzing user challenges at the level of different phases of a work session: beginning, middle, ending and between two work sessions, participants reported distinct challenges. At the beginning of each session, there is considerable difficulty in resuming mental context. During a session, challenges included prioritizing tasks and scheduling breaks effectively. Towards the end of sessions, an emotional connection to the project progress became pronounced. A secondary challenge emerged as participants had to terminate sessions abruptly. In the intervals between work sessions, participants expressed a desire for the search and sensemaking system to continue progress on the project goal. This involved conducting searches toward fulfilling larger information goals, and presenting the findings to the creator upon resumption of work.

2.5.3 Limitations and Future Work

This study has limitations as it tries to balance ecological validity and the need to analyze data to understand behavior. Here, we discuss their potential impact, how we tried to address the limitations and propose future work.

First, to preserve participant privacy and agency over what data is collected, while the browser extension's logging mechanism automatically detects whether they are working on the project, it requires the participant to start and stop logging. This means that we could have missed data points that could add to our understanding. To analyze the collected user behavior, we needed to operationally define units of observation, such as work sessions and project stages. Work sessions are the times when their work document is open and active and the participant remembered to turn on logging. To avoid unnecessary data collection, the extension would stop logging when it detected inactivity in the work document tab for more than 20 minutes and let the participant edit the logs to remove data points. We hope that by triangulating data collected across not only application logs but also self-reports of work behavior, we can mitigate the loss of insights.

Second, our focus on logging information exclusively from the search browser and designated work document is a deliberate choice, but it limits our view of the broader array of tools and collaborative elements present in knowledge workers' creative workflows. The intricate context within this ecosystem holds valuable insights into search and sensemaking behavior during creative work, suggesting a need for future studies to explore and understand this rich context.

Next, our study involved a relatively small sample of 15 participants observed over approximately 2.5 months, engaging in diverse projects of varying complexities and scopes. Recognizing that this sample may not be fully representative of all knowledge workers and creative domains, and acknowledging individual differences, we propose

future research to recruit larger and more diverse samples or to extend data collection periods. This would enable a more nuanced understanding of creative projects across different contexts.

Last, the dynamic nature of evolving technologies poses a temporal constraint on the validity of our results, given that the study was conducted between 2021-2022. As web search technologies and the landscape of work continue to evolve, it is imperative for future research to revisit and update our understanding of work practices surrounding knowledge creation. Despite these limitations, the mixed-method approach and apparatus of this study may provide valuable insights for future investigations. Future research is required to overcome these limitations. Overall, these findings from this longitudinal study of web search and sensemaking during creative work reveal process insights from a range of different practitioners.

2.6 Conclusion

To better shape the future of information search and synthesis tools in more contextual and human-centered ways, we must understand the knowledge worker's context, when they search what and why, when and how they synthesize information in different ways, the challenges they face, and how they want tools to better support their creative process. In this study, we observe 15 real-world knowledge workers over the course of their projects ranging from 1-6 months long (avg. 2.5 months). This study builds on prior knowledge by extending the time scale of analysis (i.e., months rather than weeks or hours), level of data richness (i.e., quantitative and qualitative data), *and* sources of data (i.e., work documents and search activity). To observe participants' natural in-situ behavior longitudinally, we developed a novel experimental protocol and browser extension that logs participants in a privacy-preserving and promotes reflection on work

behavior. Analysis of web search and work document activity logs and weekly survey responses reflecting on this activity gives us insights into the nature of creative knowledge work and adds context around what participants searched and synthesized in their work documents.

2.7 Acknowledgements

This chapter, in part, is currently being prepared for submission for publication of the material. Srishti Palani and Steven P. Dow. The dissertation author was the primary investigator and author of this material.

Part I

Getting Started With Information

Exploration

Chapter 3

The "Active Search" Hypothesis: Characterizing Search Behavior and Challenges When Starting to Explore Information and Frame Problems

While research shows that web search plays a role throughout the creative process, less is known about how people use web search to learn and frame their thinking about an open problem. People need web search to gather information about a problem area, but this can also influence the rest of the creative process. To understand how web search affects early-stage design, we collected and analyzed search logs and self-report data from 34 students in a project-based design class. Participants reported struggling with scoping broad, ill-defined information goals into queries, learning domain-specific terms, and assessing the usefulness of found information. Analysis found that more active and diverse search behavior (i.e. issuing more frequent and diverse queries, and opening more webpages) related to more progress in early-stage design (i.e. gathering more facts, articulating more insights, and developing better problem frames). Based on these findings, we discuss implications for designing search tools to support peoples' creative processes.

3.1 Introduction

Large, complex challenges – like keeping the public safe during a pandemic, dealing with climate change, and enabling equitable access to public transportation – often require problem solvers to form a broad and deep understanding of facts, constraints, and existing solutions [348, 129, 162]. This process of gathering information and discovering insights can have a significant impact on how designers approach or "frame" a problem [125, 236, 368].

Prior research has revealed the importance of web search throughout the creative design process, including to find existing solutions, search for inspiration, learn how to use prototyping tools [158, 148, 461, 462]. Based on a recent diary and survey study [462], researchers have found that people search to support a range of creative tasks across

different domains, such as academic writing, cooking, design (e.g. visual, architectural, etc.) [461, 182, 281, 456]. Searchers use specific information resources (e.g. images, videos) strategically to support different stages of the creative process [462]. However, less is known about how specific web search behaviors influence early-stage design and problem framing. This study builds on prior self-report studies to understand how search behavior relates to learning and problem framing, by gathering search log data to observe in-situ search behavior and survey data to gain qualitative insight into the meaning of quantitative results.

Prior work, by the Search-as-Learning community, has developed tasks [447, 204] and measures [444, 18] using the cognitive learning dimension of Anderson and Krathwohl's Taxonomy of Learning [238] a well-known education resource. In this taxonomy, six types of cognitive processes are identified: *remember*, *understand*, *apply*, *analyze*, *evaluate* and *create*. Early stages of a design process require the designer to learn about a new domain, and involve all these types of learning: from recalling facts (remember) to synthesizing information to discover insights (understand, analyze, evaluate), and asking questions and posing problems in a fruitful and radical ways, generating ideas (create), etc. [102, 368, 124, 321]. Rieh et al. [346] conceptualizes this comprehensive task as "*creative learning*". While there has been some prior work [157, 213, 447] to understand the relationship between search behavior and learning outcomes, these studies focus on learning rather than creative tasks. This study builds on this prior work to investigate the following research questions:

RQ1: What does search log and self-report data reveal about the information goals, challenges experienced and strategies employed by searchers during early-stage design?

RQ2: How do web search behaviors relate to creative learning outcomes (such as gathering facts, discovering of insights, and framing the problem)?

To address these research questions, we observed 34 students in a project-based

design classroom as they searched the web for 30 minutes during their early-stage design process. Our analysis of search log data, together with participants' self-reports about their experiences found that: participants have cognitively-complex information goals – for example, to understand domain-specific language and context, to find patterns and design constraints, and seek inspiration to generate ideas. Participants reported challenges with scoping broad, ill-defined information needs into queries, and assessing the usefulness of information. Analysing the search logs found that *active searching* (i.e. issuing more, longer, and more diverse queries and opening more webpages) relates to higher creative learning gains (such as gathering more facts, discovering more insights and developing more well-defined problem frames). Based on these findings, we discuss implications for designing search tools to better support peoples' creative learning process.

3.2 Method

3.2.1 Participants

34 undergraduate students from a project-based design course were recruited, and received 1% extra course credit for their participation. The study was conducted over a period of the first two days in the first week of classes. Participants had a diverse range of prior design experience ($\mu = 2.19, \sigma = 1.07$). Participants had only a little prior knowledge about the topic ($\mu = 1.38, \sigma = 0.72$, on a scale of $1=no\ knowledge\ at\ all$, $5=know\ a\ lot$) All participants reported extensive prior experience with web search, in general everyone reported searching multiple times a day, and using web search for more than five years. 32 of them reported using Google, and 2 using Bing and Baidu as their primary search engine. All of them used Google Chrome as their primary search browser

3.2.2 Procedure

At the start of the study, the researcher explained the study procedure and guided participants through how to install and use a Chrome browser plugin (<https://tinyurl.com/HistoryMaster>) to collect search logs. All search log data was automatically anonymized upon collection. A pre-task survey captured participants' web search experience, prior design experience, prior domain knowledge, and information seeking goals. As the main task, participants had 30 minutes to search and take notes on one of four topics being studied in a project-based design course (Refer to Supplementary Materials at <https://tinyurl.com/SearchTasks> for Search Tasks). The breadth and depth of these four multi-faceted topics provides a good opportunity to study web search in the context of early-stage design.

Before and after the task, participants were required to summarize what they knew about the topic in 3-5 sentences or 200-words, and write a problem statement that could be the focus of a quarter-long project. Additionally, in the post-task survey students were asked to report any challenges faced and strategies used when using web search during this early-stage exploratory creative design task. The study lasted 60 minutes: 30 for web search and 30 minutes for filling out the pre-and post-task surveys.

3.2.3 Measures

Qualitative Insights from Surveys

Survey questions about information seeking goals, challenges faced and strategies used were analyzed using a grounded-theory approach to thematic analysis.

Web Search Logs

To understand search behavior, from the search logs we calculated the following measures for each participant: (i) *Number of queries issued*; (ii) *Length of query* (i.e.

average number of terms per query); (iii) *Diversity of query* (i.e. number of unique query terms, stemmed to reduce the query terms to their respective base forms without affixes); (iv) *Number of unique web pages opened*.

Creative Learning Metrics

To measure creative learning outcomes, we calculated the following measures for each participant:

(i) **Change in Number of Declared Facts (Remember)** is measured by the change in number of distinct facts per statement between pre- and post-task summaries. To reliably measure number of facts, we randomly selected 20% of all 34 pre- and post-summaries for four raters to independently count facts. To account for agreement between four raters we calculated the Fleiss' κ . The raters had an inter-rater reliability of 0.74 Fleiss' κ . The rest was coded by one of the raters. (ii) **Depth of Learning (Understand, Analyze, Evaluate)** is measured by three metrics proposed by Wilson and Wilson [444] (refer to Table 3.1).

(iii) **Degree of Problem Definition (Create)** To understand the process of moving from an ill-defined to a well-defined problem scope (i.e. from level 1: Problem Discovery to level 5: Problem Definition) we adopt Abdulla and Crammond's Problem Finding Hierarchy [15]. *Level 1: Problem Discovery*: the problem statement is very ill-defined or defined very similar to the given problem. There is no relevant information and no insight to build on. *Level 2: Problem Formulation*: the problem statement is yet to specify the problem, however, there is enough information that they could discover insights from. *Level 3: Problem Construction*: the problem statement includes some background information, but the problem finder needs to further evaluate the information to specify a well-informed and well-reasoned problem. *Level 4: Problem Identification*: the problem statement includes some information and has some preliminary insights

Table 3.1: Depth of Learning Measures corresponding to the Understand, Analyze and Evaluate Cognitive Learning Levels of Anderson and Krathwohl’s Taxonomy of Learning. [238, 444] Fleiss’ κ is significant at $p < 0.05$.

Metric	Defintion	Fleiss’ Kappa
Quality of Facts (Understand)	Usefulness of recalled facts (0-3, where 0: irrelevant or useless facts, 3 :facts demonstrate technical understanding)	0.64
Interpretation of Facts (Analyze)	Synthesis of facts to draw conclusions (0-2, where 0: simply listing facts with no further interpretation, 2 : finding patterns across multiple facts)	0.58
Critique of Facts (Evaluate)	Evaluation of facts to raise questions, identify outliers and inconsistencies (0-1, where 0: true, 1 : false)	0.74

about the problem; however, there is no specific problem identified or the problem identified is still rather vague. *Level 5: Problem Definition:* (most well-defined stage of problem finding) the problem statement identifies as specific, well-informed and well-reasoned problem. When classifying the pre- and post-task problem statements, the raters had an agreement of 0.69 Fleiss’s κ

3.3 Results

10 participants chose to work on the topic of the Last Mile problem; 10 on Safe Roadways; 8 on Equitable Access and 6 on Autonomous Vehicles. Since there were no significant differences in search behavior across the topics, for the remainder of this chapter we do not differentiate between search task topics, and treat them as independent trials of the study.

3.3.1 What information goals do searchers have during early-stage design?

Participants reported using web search to fulfill the following information goals: (i) **To get an overview of the information space:** 22 participants mentioned wanting to know key concepts and terminology in their chosen topic and related topic areas. As P39 wrote, they searched *"to learn more about related topics and potential avenues to go down; for basic understanding of concepts"*. 14 participants mentioned wanting to know more about the *"history and current practices to get the context and background"* (P16). 4 participants mentioned wanting to search for perspectives of users and experts. As P99 stated, *"search forums for solutions to see if other users or experts have encountered similar issues... to collect related images and concepts."* (ii) **To discover design patterns and criteria:** 27 participants mentioned that they used web search to analyze and evaluate found information by trying to determine patterns and critique how different pieces relate to one another through differentiating, organizing, and attributing. For example, P12 stated that they use web search to, *"compare and check ideas to come up with a criteria and give me some direction"*, and as PO3 stated, *"to look for counterpoints or alternatives"*. (iii) **To seek inspiration and generate ideas:** 21 participants mentioned using web search to get inspiration to plan to or to generate ideas. For example, P48 says that they search, *"to find design inspiration when I am starting a design. ... to check and compare existing solutions."* Similarly, P115 states that they - *"seek design inspiration from what others have done as well as find resources to make the design possible."*

3.3.2 What search strategies emerged to meet these information goals?

Results from the post-task survey enrich our understanding of challenges faced and strategies employed by searchers to fulfill the above-mentioned information goals. 29 participants reported struggling to formulate their informational needs as a query. For example, P9 stated, *"I didn't know what to search for... I don't know what I don't know and what I'm missing out on. I have this FOMO [Fear of Missing out] like feeling"*. To try to better articulate their information need in an effective query, participants talked about issuing multiple queries in quick succession in an iterative manner. For example, P8 issued their first query *"congestion"* followed by the queries: *"Car congestion san diego"*, *"car congestion san diego map"*, *"car congestion san diego hotspots"*, *"car congestion san diego hotspots map"*, *"road congestion solutions san diego"*, *"environmental-friendly road congestion solutions san diego"*. Commenting on this P8 said that their strategy was to *"start searches broad and then add terms to narrow down by adding terms"*. This strategy was used by other participants to specify contexts (e.g. P8's *"san diego"*), information sources (e.g. P8's *"map"*), or other constraints (e.g. P8's *"environmental-friendly"*). Another strategy to help articulate information needs as queries was to ask natural language questions – like P6 stated, *"I didn't know how to phrase it as a search. I just searched the way I would tell it to my friend and hoped something interesting came up"* (for the query: *"why unsafe driving behavior on the rise?"*). These natural language questions, and adding terms to specify the query tend to make these queries longer than keyword queries [17].

12 participants reported challenges learning domain-specific terminology. P16 illustrates this in their use of the term *"hub"* *"I didn't know what the term for a bus or train station generally was in urban design. Now that I found it in an article, it makes*

it so much easier to search” P16 used the term *”hub”* across 7 consequent queries. Additionally, 10 participants reported challenges assessing the usefulness of information. P19 discusses their strategy to assess usefulness of search results *”I now know this is a reliable source since a lot of articles refer to it”*. Similarly, P18 also said, *”it occurs as the top result across multiple queries so it must be relevant and trustworthy”*. Participants also reported *”opening webpages in new tabs for reading later”* (almost like a *”bookmarking”* strategy).

3.3.3 How do web search behaviors affect creative learning outcomes?

From the search log data we defined key searching behaviors: the number, depth, diversity of queries and number of webpages opened. Doing more of these search behaviors were indicative of more *”active searching behavior”*.

Active, Diverse Searching Behavior Correlate with Learning More Facts

To analyze how search behavior relates to the change in number of declared facts post- compared to pre-search, we performed correlation analyses (see Table 3.2). We found a significant correlation where searchers who saw the greatest increase in number of declared facts also tended to have issued more, longer, more diverse queries, and opened more web pages.

Active Searching Behavior Relates to Articulating Deeper Insights and More Well-Defined Problem Statements

To understand the relationship between search behavior and the depth of learning measures, we performed ordinal logistic regression analyses. First, we explore the rela-

tionship between search behavior and how well searchers "understand" the information space (as measured by Quality of Facts) [444]. Searchers who issued more queries, and opened more web pages had significantly higher increases in quality of facts mentioned post- rather than pre-task. There were no significant differences in the length or diversity of queries issued with respect to the change in quality of facts reported (see Table 3.3(a)). Second, we explored how search behavior corresponds to how participants "analyze" the information space (as measured by Interpretation of Facts) [444]. We found that searchers who issued more, longer and more diverse queries had significantly higher increases in their interpretation scores post- rather than pre-task. There was no significant difference in change of interpretation scores with respect to the number of web pages opened (see Table 3.3(b)).

Third, we explored how search behavior relates to how well participants "evaluate" the information space (Critique of Facts) [444]. Searchers who issued longer and more diverse queries had significantly higher increases in critique scores, post- rather than pre-task. There were no significant differences in change of critique scores with respect to the number of queries issued and web pages opened (see Table 3.3(c)).

Last, searchers who issued more, longer and more diverse queries also have significantly better defined problems post-search than pre-search. There is no significant difference in problem definition level with respect to the number of webpages opened (refer to Table 3.3(d) for details). All analyses were corrected against effects of multiple

Table 3.2: A correlation analysis found a significant positive relationship between higher gain in facts stated and issuing more, longer, more diverse queries, and opening the more web. * significant at $p < 0.05$, ** significant at $p < 0.01$

Measure	<i>r</i>	<i>p</i>
Number of Queries	0.69	0.01**
Length of Queries	0.38	0.03*
Diversity of Query	0.18	0.04*
Number of Webpages	0.65	0.02*

Table 3.3: More active and diverse searching relates to deeper learning and more well-defined problems. Ordinal Logistic regression analyses results * $p < 0.05$, ** $p < 0.01$

Measure	Odds Ratio	p
(a) Quality of Facts		
Number of Queries	1.28	<0.01**
Length of Queries	0.20	1.20
Diversity of Query	1.09	0.08
Number of Webpages	1.04	0.03*
(b) Interpretation of Facts		
Number of Queries	1.71	<0.01**
Length of Queries	1.23	0.03*
Diversity of Query	1.11	0.03*
Number of Webpages	0.75	0.17
(c) Critique of Facts		
Number of Queries	1.09	0.09
Length of Queries	1.42	0.04*
Diversity of Query	1.64	0.02*
Number of Webpages	0.96	0.89
(d) Degree of Problem Definition		
Number of Queries	1.27	<0.01**
Length of Queries	1.04	0.03*
Diversity of Query	1.10	0.02*
Number of Webpages	0.75	1.20

comparisons using Bonferroni correction.

Overall, we observe that more active, diverse searching (as shown by issuing more, longer, and diverse queries, and opening more web pages) generally corresponds to higher increases in creative learning outcomes (such as gathering more facts, articulating deeper insights, and framing more well-defined problems).

3.4 Discussion

This study sheds light on how people use the web to search for information during early-stage design to help them frame a problem. By analyzing search log and survey data, we find that searchers have cognitively-complex information goals during early-

stage design – goals that go beyond the recall and lookup tasks that current search tools are optimized to fulfill (ref. 3.1) [462, 346, 276, 439]. Designers' information goals include learning about key concepts and terminology, history and current practices, and the perspectives of users and experts to get an overview of the information space. This exploratory behavior is consistent with research that describes how designers explore a problem space to find a problem [156] and then iteratively re-framing that problem by discovering and integrating new information [368, 124, 321]. This divergent exploration and convergent synthesis is a hallmark of the design process [162] and is exemplified in the Design Thinking model [409]. Additional information goals included wanting to discover design patterns and criteria, and seek inspiration to generate hypotheses and ideas. Trying to surface patterns and previously unknown connections can be an effective technique to also generate ideas through creative combination (i.e. coming up with something new by combining two concepts/ideas) [70, 441, 389, 133] and analogical reasoning (i.e. the process of making connections through examples) [158, 164, 426].

Searchers reported challenges scoping broad, ill-defined information needs into queries, learning domain-specific language, and assessing the usefulness of information. These challenges are related to those faced by design novices [95, 156, 124, 162, 367]. Active, diverse searching strategies (such as issuing more, longer, and diverse queries, and opening more web pages) related to higher creative learning gains (such as more, better quality facts and insights, and more well-defined problems). This strategy of iteratively probing the information space is similar to the design-thinking strategy of using prototyping as a method for actively probing and getting feedback from the design space and user community [368, 124, 321, 162]. By issuing more frequent and diverse queries and opening more webpages, searchers might have had more opportunities to learn terms to better articulate their queries, explore a subset of the information space and develop their relevance judgement criteria. Promoting discovery of domain-specific

language and query diversity should be an important design goal for future search tools. They should build on work like [26, 158, 84, 83], to guide novice designers to articulate more diverse queries and support discovery of domain-specific language, hypothesis and idea generation.

We observed significant relationships between active searching behavior and creative learning outcomes, however, we cannot make any causal claims. It could be that these relationships are a better reflection of the searchers' skills or aptitude, rather than a function of specific strategies. Future work needs to conduct a large-scale analysis of naturalistic search behavior during design to further test the hypotheses generated by this short paper. Furthermore, the study has limitations as it tried to balance ecological validity with experimental control. For instance, we controlled all participants' search sessions to be 30 minutes long to help us compare between participants. However, this might be different from the individuals' usual searching behavior. Since complex search tasks such as exploratory search are often carried out over multiple sessions and devices [276, 439, 461], we encouraged participants to continue searching beyond this session, and to think of this as their first search session. Future research is required to overcome these limitations, build out and test suggested design implications of these findings.

3.5 Conclusion

By collecting and analyzing search logs and self-report data from 34 students in a project-based design course, this study provides insights into the information goals, challenges faced, and strategies used by people when using web search during early-stage creative processes. This study applies measures from information science and creativity research to operationalize creative learning outcomes in early-stage design. We learned that active, diverse searching (as shown by issuing more, longer, and diverse queries and

opening more web pages) relates to higher gains in creative learning outcomes (such as breadth, depth of learning and problem framing). We conclude by reflecting on our findings to propose design implications for search systems to support cognitively complex tasks such as creative learning during design.

3.6 Acknowledgements

This chapter in part, includes portions of material as it appears in *The "Active Search" Hypothesis: How Search Strategies Relate to Creative Learning* by Srishti Palani, Zijian Ding, Stephen MacNeil, Steven P. Dow. in the Proceedings of 2021 ACM SIGIR Conference on Human Information Interaction and Retrieval Online (CHIIR'21). The dissertation author was the primary investigator and author of this material.

Chapter 4

CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery

When exploring a new domain through web search, people often struggle to articulate queries because they lack domain-specific language and well-defined informational goals. Perhaps search tools rely too much on the query to understand what a searcher wants. Towards expanding this contextual understanding of a user during exploratory search, we introduce a novel system, CoNotate, which offers query suggestions based on analyzing the searcher’s notes and previous searches for patterns and gaps in information. To evaluate this approach, we conducted a within-subjects study where participants ($n=38$) conducted exploratory searches using a baseline system (standard web search) and the CoNotate system. The CoNotate approach helped searchers issue significantly more queries, and discover more terminology than standard web search. This work demonstrates how search can leverage user-generated content to help people get started when exploring complex, multi-faceted information spaces.

4.1 Introduction

Web search provides a powerful way to browse, learn and discover new information about topics of interest from the largest repository of knowledge – the World Wide Web. However, people who lack knowledge of a particular domain or well-defined goals generally struggle to articulate useful search terms: They have not yet learned domain-specific language that could help them translate their fuzzy goals into concrete queries [276, 439, 29]. During this exploratory knowledge discovery process, people often also take notes to help process, store, and share information [74, 294, 107, 420].

Current search engines attempt to assist people with query formulation by leveraging search logs data to detect user intents and context [390, 445, 222]. Others have explored the potential of recommending queries based on prior searches from others (e.g. People Also Search, Related Searches) [31, 203] and presenting search trails (i.e. interactive

visualizations of how previous searchers explore an information space) [44, 73, 387, 455]. Existing query assistance approaches aim to predict query formulations (e.g. auto-complete), resolve ambiguity (e.g. people also ask), and help searchers find information quicker [270, 79, 144, 336]. While these query assistance strategies may be helpful in many circumstances, they lack a rich understanding of searchers' goals, interests, or gaps in knowledge [72, 346, 439]. Our research investigates whether we can generate contextual insight for query suggestions by leveraging how searchers naturally capture and synthesize information.

The personal notes that people take during an exploratory search task can encode rich contextual information about a users' goals [231, 29, 357], and current state of learning [107, 420, 225, 280]. Notes can provide a signal of what a user finds relevant across multiple sessions and information sources [406, 107, 74, 295]. Further, notes potentially reflect the searchers' current understanding of a domain, or vice versa, the gaps in knowledge. Mining personal notes for these richer insights could improve query assistance, especially for more exploratory tasks where users grapple with vast amounts of information, by arming searchers with domain-relevant language and indicating what else there is to learn. Therefore, this project explores the following research questions:

- **RQ1:** How does leveraging notes to inform query suggestions affect query formulation and search behavior compared to standard web search?
- **RQ2:** How do notes-based query suggestions affect knowledge discovery and learning behavior compared to standard web search?
- **RQ3:** How do notes-based query suggestions affect the perceived value of query suggestions compared to standard web search?

To explore the potential for integrating note-taking and searching, we developed the *CoNotate* system (Fig. 4.1) to provide query suggestions based on patterns and gaps

in the searchers' notes and previous searches. The system mines the users' notes and prior searches to implicitly infer relevant and potentially undiscovered information and recommend them as query suggestions. We conducted a within-subjects study ($n=38$) where participants were asked to search on two different multi-faceted exploratory topics. Participants used *CoNotate* for one topic and then used the Baseline system (standard web search) for the other. Analysis shows that the *CoNotate* approach helped searchers issue significantly more queries and discover more domain-specific terminology than standard web search. Also, participants reported preferring notes-based suggestions over standard query suggestions, particularly to help them discover new terms and concepts related to the topic.

This work makes the following contributions:

- *CoNotate* system: a browser-based tool that integrates note-taking and searching to recommend contextualized query suggestions for exploring broad multi-faceted information spaces,
- Empirical insights from a within-subjects experiment that suggest leveraging rich context from the knowledge development process to inform search assistance can encourage active searching and knowledge discovery.

4.2 Related Work

This research builds on prior work in the CHI community related to note-taking, and query formulation assistance during exploratory search. We extend the prior work by leveraging contextual information captured in notes to recommend queries.

4.2.1 Note-taking helps individuals store, learn, and share information during exploratory search

Searchers explore the web to learn more about a topic of interest, and often take notes to help store, process and share found information [294, 420, 107, 225, 280, 256]. The note-taker records relevant information, freeing their mind from having to recall everything. As a searcher consults multiple sources and explore unfamiliar domains across multiple sessions, it is hard to hold all the information in memory to satisfy their information goal. Note-taking bridges the gap of carrying information learned in one session or information source to the next [280, 294, 256].

Taking notes involves manipulating information by summarizing, paraphrasing, and mapping. This engagement can help cognitively encode and gain a deeper understanding of the information [225, 206, 275, 163]. When taking notes, searchers often select and record relevant concepts [280, 420], process-related information (e.g. queries, links, etc.), and their own interpretations [107, 74, 420]. This suggests that the purpose of notes goes beyond just helping people record and process information, but also serves to synthesize low-level raw data into high-level meaning, ideas and decisions [231, 357, 34, 163]. Furthermore, creating this artifact of their thinking and sense-making makes it easier for searchers to share knowledge and collaborate with others [420, 137, 176, 232, 163].

Existing information gathering tools support (i) capturing information (eg. Web clippers for Google Keep ¹, EverNote ², Information Scraps [42]), (ii) structuring notes to help people make sense of information (eg. Knowledge Accelerator [176], Scatter/Gather [110], Web Summaries [111], SearchLens [84]), and (iii) using notes to help individuals or collaborators resume a search session (eg. SearchBar [294], CheatSheet [425], SearchTogether [297]). However, previous tools have not explored how

¹Google Keep Chrome Extension: <https://chrome.google.com/webstore/detail/google-keep-chrome-extens/lpcaedmchfhocbbapmcbpinfpghiddi?hl=en>

²EverNote Web Clipper: <https://evernote.com/features/webclipper>

search tools can leverage the rich contextual information from notes to guide querying and searching. CoNotate builds on previous work by leveraging the rich contextual information captured in notes and previous searches to suggest queries and reflexively further the search process.

4.2.2 Query assistance methods in exploratory search

The Human Computer Interaction and Information Retrieval communities have explored different approaches for query assistance to help people search more effectively [211, 223, 244, 17, 365, 437]. Silvestri [385] provides an overview of the different query assistance techniques which include: (i) *Query auto-completions* (e.g. autocomplete or auto-correct: These aim to help a user complete their query formulation by mining searcher's history and similarities across users³; (ii) *Session-based search*: This considers similarity between query terms, clicked documents, or sequences of queries in a session to improve the accuracy of suggested queries, links and snippets of information [270, 79, 144, 336]; (iii) *Related query recommendations* (e.g. Related Searches): These suggestions aim to help the user explore the information space [31, 203]; (iv) *"People Also Ask"*: These suggest questions asked by other users who issued the same query, in order to help specify their information needs [23, 353]. These computational approaches aim to better understand a user's information needs by predicting search formulations, resolving ambiguity, or getting searchers to specific information quicker. Standard evaluation metrics such as the precision, recall, and relevance of search results retrieved by the query aim to help people find specific information faster, but they do not necessarily help a user broadly explore a domain [347, 417, 439].

Hsieh-Yee [192] found that when working with less familiar topics, subjects were

³Google Auto-complete Suggestions: <https://blog.google/products/search/how-google-autocomplete-works-search/>

more likely to consult a thesaurus for term suggestions. Vakkari [415] found that as subjects learned more about the topics they began to use a wider and more specific search vocabulary. Niu and Kelly [303] found that participants with lower search experience used more suggestions and they used more suggestions when searching for more difficult topics. Jansen and McNeese [205] found that users adopted query suggestions in 71% of instances where it was requested. This prior work shows that people find value in query suggestions and that they might be particularly useful when users are unfamiliar with topics.

Computational algorithms for query assistance are limited in their ability to infer context and intent as they only have access to users' search logs and webpage metadata (e.g. title, authors, date of creation, etc.) [32]. To capture additional user context, previous work has explored how to leverage *user-generated metadata* (e.g. descriptive keywords, annotations, etc.) [214, 84, 209, 332, 185]. While helpful in guiding exploration, these approaches are limited by users' abilities to effectively articulate their needs and interests (which is especially hard during exploratory search when searchers lack familiarity with the domain). Therefore, to improve query assistance and provide adaptive recommendations to aid knowledge discovery, we need to be able to effectively *leverage both user-generated information and computational methods*. While the above-mentioned methods consider search as an isolated activity, CoNotate leverages the rich contextual information in the notes taken and previous searches to suggest queries and reflexively further the search process.

4.2.3 Leveraging information in user-generated documents to provide query assistance

Prior work has proposed using user-created artifacts (e.g. written papers, emails or notes) to implicitly get feedback on whether a search result is relevant to the user

[224, 153, 406], and measure what people learned during search [107, 74, 347]. Notes include relevant information and ideas spanning across documents, information sources (e.g. online documents, friends, books), and can monitor the note-taker's progress through the information space [107, 74, 256]. However, to the best of our knowledge, prior work has not explored the value of mining contextual information in notes to guide query formulation.

Modern text-editing software (e.g. Google Docs, Microsoft Word) includes the ability to select words/phrases in the document and issue them as queries. This allows users to search without leaving the document. Furthermore, Teevan et al. [406] re-rank search results to help users find information quicker by implicitly inferring interests from user-generated documents and emails. While this prior work moves in the direction of integrating search and note-taking, these methods still rely on the user to identify and articulate their information need as queries, and do not recommend queries that help them further explore knowledge gaps. CoNotate leverages the notes taken by a searcher, a context-rich activity beneficial to both the user and the computer, with their searching activity to recommend query suggestions based on patterns and gaps in searcher's notes.

4.3 CoNotate

CoNotate is a browser extension that aims to help users articulate queries related to their information goals during exploratory search tasks by leveraging a user's notes to computationally offer query suggestions. This section describes *CoNotate*'s user interface and system architecture.

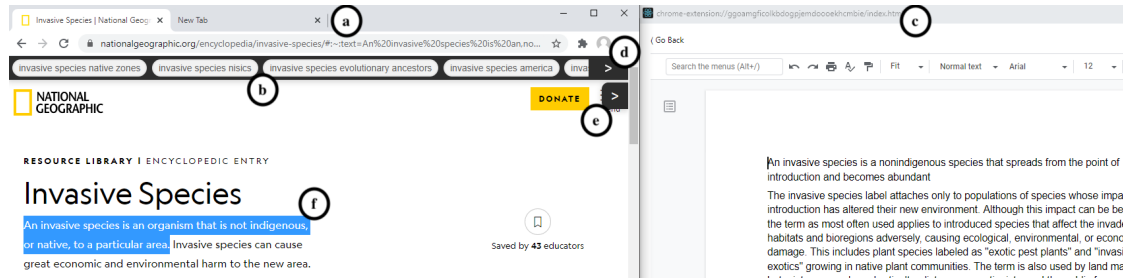


Figure 4.1: The *CoNotate* Environment: including (a) Default Chrome Search Interface, augmented with (b) Suggestions Bar with six query suggestions and (c) the Note Taking Interface. The system supports additional interactions including (d) scrolling query suggestions, (e) resizing note-taking interface, and (f) highlighting, dragging and dropping web page content into notes.

4.3.1 User Interface

The *CoNotate* interface (see Fig. 4.1 or watch supplementary video for a demo) augments the standard Chrome browser with two main components:

1. **Suggestions Bar:** As the user issues queries and takes notes, the Suggestions Bar updates with query suggestions (Fig. 4.1b). It displays a row of six randomly-ordered query suggestions as buttons (Refer to Section 4.3.2 for details on the suggestions). Clicking on a suggestion issues it as a new query, and displays search results in the chrome browser window. When the Chrome browser window is not full screen, some of the query suggestions are hidden. Clicking the **scroll markers** (see Fig. 4.1d) in the Suggestions bar reveals any hidden query suggestions.
2. **Note-taking Interface:** a window that allows users to create a new notes documents and take free-form notes (Fig. 4.1c). The note-taking window by default takes up 50% of the screen. Users can resize the notes window (to take up 10% or 50% of screen space) by clicking the **toggle arrow** (Fig. 4.1e). The notes documents created are modified Google docs. We chose to use Google Docs since they are widely used for note-taking and offer the basic tools for adding and modifying text. To maximize space in the user interface, we hide the default Google Docs

menu such that only the document and toolbar are presented to the user. Searchers can take notes by either typing them in, or by copying in content and links from the browser window (either by using Ctrl + F, or **highlighting, dragging and dropping** as seen in Fig. 4.1f). Our note-taking interface was designed to allow flexible, idiosyncratic note-taking styles since individuals structure notes very differently. Crescenzi et al. 2019 [107] found that searchers organize information in lists, outlines, matrices, and tables. Therefore, we enabled tools in the Google Doc to allow these. Since they [225, 107] also found that people organize their notes chronologically and linearly, sometimes grouping information by information source or topical themes, we let searchers type and modify text as they would naturally.

Front-End Implementation Notes: The front-end of the *CoNotate* prototype is a browser extension developed using Javascript alongside the ReactJS framework. The Suggestion bar is part of an injected content script which renders every new page load. The same content script is also responsible for scraping data from each url visited for the search logs and text analyses. The Note-taking Interface has three main components: (i) a master container which allows document creation and contains the other two components; (ii) an iFrame containing the Google Doc (ii) a DOM Watcher that monitors updates to the notes document and triggers updates to the Suggestions bar. All of these scripts communicate with each other through a background script which makes API requests to our server, stores our persistent data, and sends search log data to a Firebase database at the end of each study session.

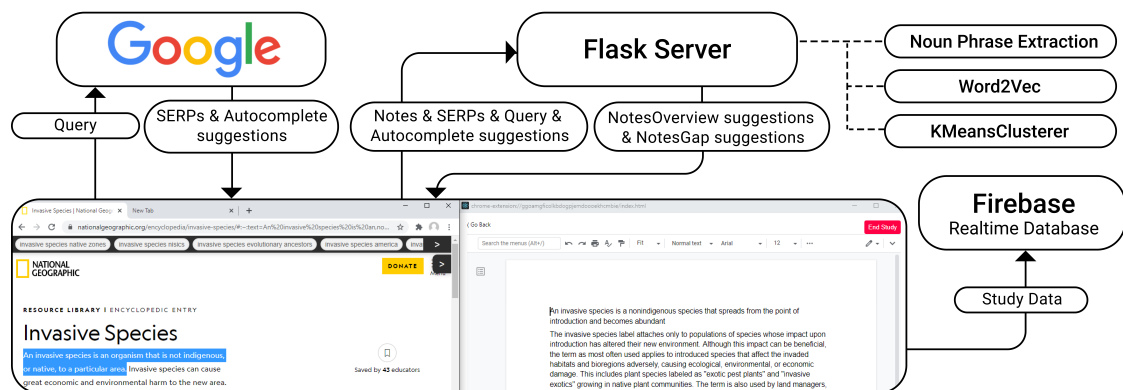


Figure 4.2: Architecture of the *CoNotate* system, a browser extension that parses a user’s notes and search terms in order to offer context-relevant query suggestions.

4.3.2 System Architecture

CoNotate outputs two types of suggestions:

1. **NotesOverview Suggestions:** these suggestions aim to present opportunities to dig deeper into phrases/concepts the user has already mentioned in the notes document.
2. **NotesGap Suggestions:** these suggestions aim to present opportunities to expand the area of exploration by suggesting phrases/concepts that are mentioned in the top 10 search engine result pages (SERPs), but are missing from the user’s notes documents.

Step 1: Detecting Contextual Information

To implicitly infer user interests we mine user-generated notes. We consider the notes taken by the user during the search task as a snapshot of what they have explored so far and found interesting [231, 357, 29]. Furthermore, to mine and surface additional opportunities for exploration, we mine the top 10 SERPs of the issued query. Since searchers do not usually go beyond the first SERP, these terms have potentially unknown information and interests [195, 90]. Search result diversification algorithms ensure that

SERPs cover multiple possible subtopics and intents for the given query, and minimize redundancy across retrieved documents for each subtopic or intent [335, 457, 410].

First, we *extract all the noun phrases* from the notes, and titles and snippets from the top 10 SERPs of the issued queries. Since there are a lot of individual differences in note-taking [107, 42, 295, 74], and we wanted *CoNotate* to work across individual differences we do not consider the notes' structures and extract noun-phrases. Moreover, noun-phrases include persons, locations, organizations (i.e. named entities), as well as values, characteristics and emotions (not named entities). Therefore, they preserve more meaningful and contextual information than named entities. Before extracting noun-phrases we pre-processed the data to remove special characters, HTML tags, and punctuation.

Then, we create sets of extracted noun-phrases from both the Notes (*notes_phrases*), and all titles and snippets in the top 10 SERPs (*SERP_phrases*). To get a mutually exclusive set of suggestions unique from the query's autocomplete suggestions, which many search engines offer already as part of the default search experience, we compare and exclude the query autocomplete suggestions from the *notes_phrases* and *SERP_phrases*.

Since the NotesOverview suggestions aim to present opportunities to dig deeper into familiar phrases and concepts that the user had already mentioned in the notes document – we considered only the *notes_phrases* set. On the other hand, since the NotesGap suggestions aim to present opportunities to expand exploration by suggesting phrases/concepts mentioned in the SERPs but missing from the notes documents, we calculate the *set difference* between *SERP_phrases* and *notes_phrases* and create a new set called *gap_phrases*. These are the *SERP_phrases* not in *notes_phrases* (See Fig. 4.3).

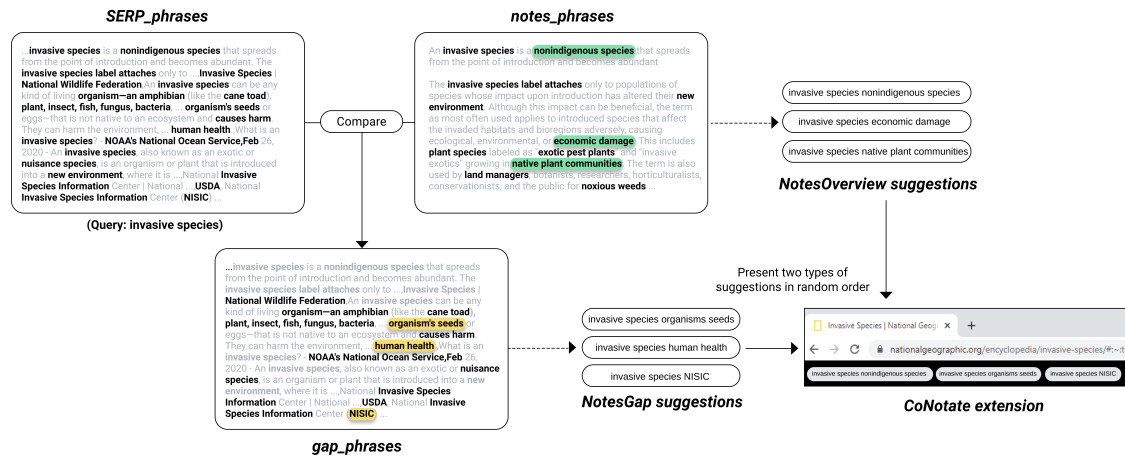


Figure 4.3: Noun phrases extraction from SERPs and notes (bold text). The *gap_phrases* are *SERP_phrases* not in *notes_phrases*. After word embedding, clustering and labeling, six noun phrases, highlighted in green and yellow respectively, are added to the original query as *NotesOverview suggestions* and *NotesGap suggestions* and presented to the user in random order.

Step 2: Creating Semantic Vector Representations

To contextualize and determine the semantic relationships between phrases, we create a semantic vector representation for each set of phrases (*notes_phrases*, *gap_phrases*). These vector representations are *CoNotate*'s "context" of what is covered in the notes document, and what is missing from the notes, but could be related.

Once we have two semantically meaningful vector spaces (as trained by Word2vec [291]), we want to compute a mutually exclusive set of phrases that represents each set. To do this, we run a *k-means clustering algorithm* using cosine similarity on the vector space. Since the size of *notes_phrases* set is smaller than the *gap_phrases* set, especially at the beginning of a search session, we set the clustering algorithm to $k=4$ for the *notes_phrases* and $k=8$ for the *gap_phrases*. We experimented with different numbers of clusters for each vector space and this appeared to be an optimal cutoff to get a representative set of phrases. When there are few notes taken or when a query is too specific, higher k values generate an overlapping, repetitive set of clusters. On the other hand, lower k values create a set which is not representative of the diversity of phrases in the notes and SERPs.

Step 3: Choosing Suggestions to Present

We choose to present six suggestions (three of each type: NotesOverview and Notes-Gap) so that it proactively presents the users with enough options, yet does not overwhelm them. The six suggestions are presented in a random order in the Suggestions Bar in the *CoNotate* interface. Before presenting the suggestions, we conduct a “final check” to guarantee the list of suggestions was mutually exclusive from each other as well as from the issued query. We did this by calculating the pairwise cosine similarity between each of the NotesOverview suggestions and NotesGap Suggestions against the original query. We only present the query suggestion if it had less than 0.4 cosine similarity with the original query. We tested this algorithm on user-generated data from pilot studies to set a custom threshold value for the similarity score such that it recommends a set of minimally overlapping set of terms. We chose to present the suggestions as query expansion terms concatenated to the end of the issued query since the suggestions were specific or related to the currently issued query. Participants can issue variants of these suggestions by issuing it and then editing it in the search bar.

Back-end Implementation Notes The backend of *CoNotate* was implemented in Python, using TextBlob [266] for noun phrase extraction, gensim [338] for the the word semantic model, and NLTK [45], and sklearn [322] for the k-means clustering. We experimented with other libraries and variants, however, these seemed to work best for our particular usecase. The Flask Python framework was used for our HTTP server. All events in the search and notes logs (such as new queries, webpages, notes, etc.) were logged to Firebase in a JSON format (see Fig. 4.2, or refer to the code ⁴ for implementation details). All communications between the server, database and users’ browser are encrypted and anonymized by creating anonymous session and Firebase IDs.

⁴CoNotate Source Code Repository: <https://github.com/creativecolab/CHI2021-CoNotate>.

4.4 Evaluation

A comparative within-subjects experiment ($n=38$) investigated the following research questions:

- **RQ1:** How does leveraging notes to inform query suggestions affect query formulation and search behavior compared to standard web search?
- **RQ2:** How do notes-based query suggestions affect knowledge discovery and learning behavior compared to standard web search?
- **RQ3:** How do notes-based query suggestions affect the perceived value of query suggestions compared to standard web search?

4.4.1 Conditions

To control for individual differences in searching, note-taking, and learning behavior, we compared the *CoNotate* system to a *Baseline* (standard web search & note-taking interface, see Fig. 4.4) in a *within-subjects* study (i.e. each participant experienced both the systems using different search topics). We counterbalanced topics and conditions to reduce order effects.

To ensure parity we only changed the types of query suggestions presented across *CoNotate* and *Baseline* systems, and kept all other system features the same. When using the *Baseline* interface, participants could issue queries using the Suggestions Bar that presented the the standard search engine *auto-complete suggestions*. In order to maintain parity across conditions, we moved the autocomplete suggestions to the Suggestions Bar and disabled the character-by-character responsiveness as it aims to predict users' query before they finish typing. This allowed *Baseline* users to access query suggestions not only when typing queries but also as they search and take notes (Fig. 4.4a). They

could also issue queries using features standard to the search results page, such as *People Also Ask* and *People Also Search* (Fig. 4.4b, c), and *Related Searches* (Fig. 4.4d). They could also choose to not use any of the query formulation assistance, but manually type the queries in the search engine. To isolate the effects of just *CoNotate*'s query suggestions, we removed the query suggestion features found in standard web search from the *CoNotate* system. When using the *CoNotate* system, participants could issue queries: using suggestions provided in the Suggestions Bar (randomized order of NotesOverview and NotesGap suggestions); or manually typing the queries.

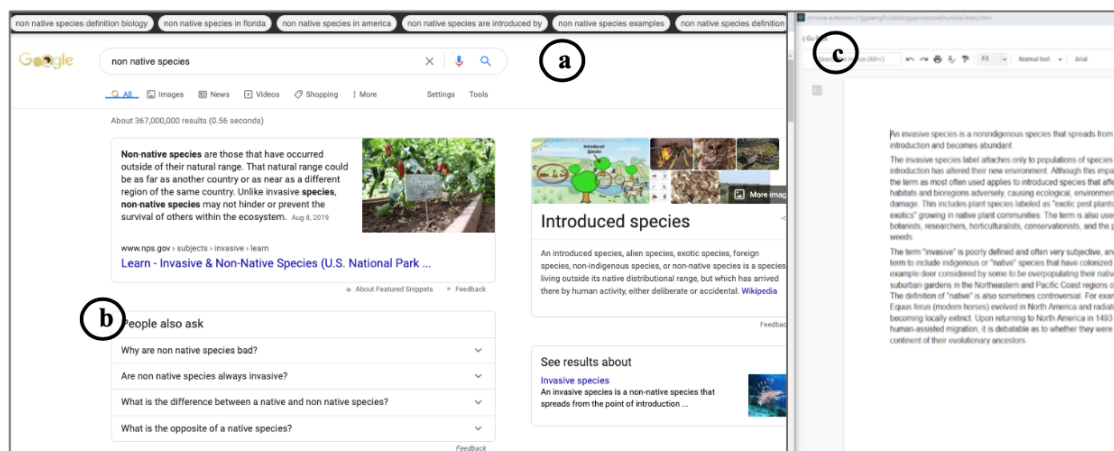


Figure 4.4: The Baseline Environment: The Default Chrome Search Interface, augmented with (a) Suggestions Bar with query autocompletion suggestions. The standard query assistance features remain like (b) People Also ask Related Searches. (c) Note-taking Interface is to the left of the Search Interface. Picture cropped excluding some search results

4.4.2 Search Tasks

Participants were given a simulated work task scenario [50, 107] to help contextualize their searching and note-taking task. We chose two topics with relatively large and complex information spaces and where the average person has relatively limited knowledge coming into the task. This effectively simulated a work scenario where participants

would need to take notes in order to synthesize major themes for the topic. Participants read the following task:

”Imagine that you are a journalist writing an article for an online magazine. As part of that process, your editor asked you to do research for an article on the following topic:

[One of two search task topics: Impact of Non-Native Species OR Impact of Technology on Mental Health]

Today, your editor would like you to do initial research to get a broad overview of the topic. Your goal should be to identify as many terms/concepts and perspectives related to the topic as you can find by searching and gathering information on the internet. Use the notes document displayed on the right-window to take notes that would be helpful to yourself to resume work on this task in the future.”

As part of the within-subjects study design for evaluating user behavior across the two conditions, each participant worked on the above task twice (i.e. once for each condition). To prevent carryover effects in learning, each participant completed the task on the two topics listed below. To avoid order effects, participants were counterbalanced such that they saw a different topic with each condition.

- ***Impact of Non-native Species:*** Non-native species are a species living outside its native distributional range, but which have arrived there by human activity, either deliberate or accidental. Non-native species can have various effects on the local ecosystem. While they are often seen as a detriment to local environments, they can be beneficial as well. Your editor asked you to write an article about the benefits and consequences of non-native species.
- ***Impact of Technology on Mental Health:*** While technology has opened up opportunities to benefit mental health, there are significant risks and unintended consequences too. There are many factors that affect mental health (i.e. our emotional, psychological, and social well-being). Your editor asked you to write an article about the benefits and consequences of technology use on mental health.

4.4.3 Participants

38 participants (22 female) were recruited through online advertisements (on SONA, a university-based participant pool) and e-mails to remotely-enrolled students at a university. All studies were conducted remotely over a video conference call. As incentive for participating in the 90-minute study, participants received extra credits that could be used to fulfill lab study requirements in their classes. As part of the recruitment process, participants answered a brief screening questionnaire that gathered demographics and information about their search and note-taking behavior. We excluded participants under age 18 ($M = 20.1$ years) and not enrolled at the university. The exclusion criteria, together with the entire study procedure were approved by our institution's IRB.

All participants reported that they use Google as their primary search engine, and use search engines multiple times a day. Some reported using Bing or Baidu when they were overseas. All participants reported taking notes multiple times per week. All of them reported taking notes on paper, as well as using a variety of applications to take notes on their computer: 32 primarily use Google docs, 4 participants use Evernote, and 2 reported using combination of other applications (e.g. Notion, Notability on iPad, Microsoft Word). Before searching, participants were asked to rate their level of knowledge about each topic on a scale of 1 (*not knowledgeable at all*) to 5 (*extremely knowledgeable*). Participants reported having little to no prior knowledge about the two topics covered by the search task: Impact of Non-native Species ($M = 1.02$); and on the Impact of Technology on Mental Health ($M = 1.52$).

4.4.4 Procedure

The experimenter reviewed the study procedure and then walked the participant through setting up the browser extension and how to use the browser extension for

searching and note-taking (see Fig. 4.1). All participants signed an informed consent to agree to recording the screen and audio, and sharing their search logs and notes documents with us. The study then proceeded as follows: Participants were asked to complete the two search tasks, with a maximum of 20 minutes to complete each task. During the 20 minutes of using the interface, participants could issue queries, view pages, and take notes. Before and after each search task, participants filled out a questionnaire which assessed their knowledge level on the topic by asking them to (1) Self-rate their knowledge level on a scale of 1-5 (where 1=*not knowledgeable at all*, 5=*extremely knowledgeable*); (2) List out all known topic-related key terms/concepts.

In addition to the knowledge questions, the post-task questions also asked them to rate their level of agreement with statements about the helpfulness of query suggestions (e.g. "*query suggestions helped me discover new terms and concepts related to the topic.*", all statements in Table 4). Furthermore, after the post-task questionnaire, to gain insight into the participant's thought processes, participants were asked to perform a retrospective think-aloud (for a maximum of 10 minutes) as they scrubbed through a screen-recording of them doing the task. They were prompted to reflect on how and why they issued each query, and if the query suggestions helped during this process. At the end of the study, once they had experienced both types of suggestions (*Baseline* and *CoNotate*), they were asked if they had a preference for the first or the second version of suggestions encountered.

4.4.5 Measures

First, to observe and analyze differences in participants' search behavior, we logged all interactions with the search engine. To understand the differences in use of query suggestions, we measured: (i) *Number of queries issued*; (ii) *Number of times query suggestions were used*; (iii) *Total number of query suggestions presented to the searchers*

during the session.

Second, to measure learning as information gain, we examine the change in knowledge level between the pre- and post-surveys: (i) *Change in Self-rated knowledge*; (ii) *Change in number of domain-specific terms listed* For each topic, we had a standard glossary of terminology that fit the domain specifications. For topic: Impact of Non-Native Species we referred to the Wikipedia Glossary of Invasion Biology Terms ⁵. For topic Impact of Technology on Mental Health we referred to the National Institute of Mental Health's Glossary of Digital Media Use and Mental Health ⁶. We counted the number of unique domain-specific terms that were covered in the pre- and post-task questions that asked them to list all known terms in this topic. (iii) *Number of open questions at the end of task*: We asked them to list out the questions or queries they might want to explore further if they had more time

Third, to understand the perceived value of query suggestions we monitored statements made in their retrospective think-alouds. Also, in the post-task survey questions, we asked participants to rate their level of agreement to statements (see Table 3) on a scale of 1-5, where 1=*strongly disagree* and 5=*strongly agree*.

4.5 Results

In this section we report the findings of our user study that investigated how integrating notes to inform query suggestions might affect query formulation, breadth of exploration, and user perception of query suggestions compared to standard web search.

⁵Wikipedia Glossary of Invasion Biology Terms: https://en.wikipedia.org/wiki/Glossary_of_invasion_biology_terms

⁶National Institute of Mental Health's Glossary of Digital Media Use and Mental Health: <https://www.nimh.nih.gov/health/topics/schizophrenia/raise/glossary.shtml>

Table 4.1: Averages (and standard deviation) for key search metrics. Participants issued significantly more queries, particularly by clicking on the suggestions, and typed fewer manual queries when using *CoNotate* than when using *Baseline* system. *statistically significant at $p < 0.05$ level .

Query Formulation Measure	<i>Baseline</i>	<i>CoNotate</i>	<i>p</i>	F_{37}
Number of Queries Issued	4.71 (2.88)	6.12 (3.03)	0.02*	5.57
Number of Typed Queries	4.16 (0.32)	2.17 (0.32)	0.00*	18.48
Number of Suggestions Issued	2.28 (2.46)	4.27 (3.29)	0.02*	6.16
Number of Queries issued from Suggestions Bar	1.85 (0.43)	3.80 (0.43)	0.00*	10.16

4.5.1 Effects on search behavior: Notes-based query assistance encourages more active searching

Participants issued more queries, particularly by clicking on the suggestions, and typed fewer manual queries when using *CoNotate* rather than the *Baseline* system (See Table 4.1 for details).

This could be either due to the content of the query suggestions or the quantity of query suggestions. *CoNotate* updates its query suggestions list every time the notes document was updated or a new query issued. On the other hand, the *Baseline* system only updates query suggestions when the user issues a new query. When analyzed using a paired-samples t-test, the difference in the number of query suggestions participants saw in *Baseline* ($M = 56.17, SD = 42.34$) and *CoNotate* ($M = 76.86, SD = 40.28$) was not statistically significant ($t_{37}=2.78, p=0.24$). Since we did not know how many queries participants would issue or the changes to notes, it was hard to control for the number of query suggestions shown experimentally. Therefore, we control for this as a co-variate during analysis. To examine just the *effect of the content of query suggestions on querying behavior*, we controlled for the number of query suggestions participants potentially saw in each system for all our analyses. We performed two-way repeated measures ANCOVA to examine the effect of using *Baseline* vs *CoNotate*, and the two topics on each of our measures of querying behavior. Based on post-hoc analyses using Tukey’s

HSD and Bonferroni correction for multiple comparisons, we found that participants issued significantly more queries and typed fewer queries when using *CoNotate* rather than the *Baseline* system (See Table 4.1 for details).

Since the positioning of query suggestions varied across the two search systems, we conducted two separate analyses to examine if this had an effect on querying behavior. In the *CoNotate* system, query suggestions are presented only in the Suggestions Bar augmenting the default chrome browser window. In the *Baseline* system (standard web search), query auto-completions of the issued query were shown in the Suggestions Bar. However, if issued queries had other query suggestion features (e.g. People Also Ask, Related Searches), they appeared as they naturally would on the SERP of the issued query. We examined if this had any effect on querying behavior by performing the same two-way repeated measures ANCOVA test as above for two separate dependent variables: (i) number of query suggestions issued overall, and (ii) number of query suggestions issued from just the Suggestions Bar. In both cases, we found that participants issued more query suggestions when using *CoNotate* rather than the *Baseline* system (See Table 4.1 for details).

There were no statistically significant differences between the topics and no significant interaction effects between topics and system used for querying behavior across all the query formulation measures. This suggests that the search task topic did not have an effect on querying behavior or on the use of the two search systems.

To observe if there were differences across the type of query suggestion used, we conducted a chi-square (χ^2) test between the types of query suggestions when using *CoNotate*: there was no significant difference in how often participants used NotesOverview suggestions ($n=52$) vs NotesGap suggestions ($n=69$) ($\chi^2(1,37) = 1.42, p=0.03$).

As participants explored the topics using *CoNotate* we saw some interesting patterns of behavior emerge. Even participants who manually typed out their queries reported

finding the query suggestions helpful. In retrospective think-alouds after using *CoNotate*, 18 participants explicitly said that they scrolled through the suggestions bar to identify relevant terminology and concepts in the topic and used them as inspiration before typing out their own query. When prompted to reflect on it, participants said typing out the query allowed them to restructure the query suggestion. “*While the recommended suggestions had some useful terms, they had grammatically-incorrect structure*” (P43). P18 described how *CoNotate* helped them explore the connection between the suggestions:

“I kept seeing a suggestion for ‘invasive species restaurants’. So I decided to click ... and found this really interesting connection that restaurants ... are putting invasive species on the menu as a way to curb their spread... And since I read about over-fishing and had seen the suggestion for ‘invasive species climate change’ I wanted to search on both to see if there were any interesting connections there...”

This suggests that people might be using the query suggestions in unexpected ways, for example finding novel connections between suggested terms/concepts.

4.5.2 Effects on learning: Notes-based query assistance promotes knowledge discovery

In terms of information gain, we found that participants reported a greater increase in knowledge and discovered more domain-specific terminology when using *CoNotate* versus using standard web search.

Table 4.2: Averages (and standard deviation) for key information gain metrics for the *Baseline* and *CoNotate* system. *statistically significant at $p < 0.05$ level. *CoNotate* led to greater increase in self-rated knowledge and terminology than the *Baseline* approach.

Measure of Information Gain	<i>Baseline</i>	<i>CoNotate</i>	<i>p</i>	F_{37}
Change in self-rated knowledge level	0.94 (0.89)	1.43 (0.83)	0.04*	4.39
Change in number of domain-specific terms	0.36 (0.96)	1.97 (1.42)	0.03*	8.03
Number of webpages opened	1.68 (1.33)	1.97 (1.73)	0.08	3.13

Statement	Baseline	CoNotate	<i>p</i>	χ^2
Query suggestions helped me discover new terms and concepts related to the topic	0.34 (0.88)	1.61 (0.99)	0.02*	2.20
Query suggestions helped me identify the most appropriate keywords or phrases for the information needed	0.29 (0.84)	0.90 (0.39)	0.14	1.88
Query Suggestions helped me narrow or broaden my search to retrieve the appropriate quantity of information	0.18 (0.83)	0.91 (0.42)	0.08	1.64
Query suggestions helped organize my notes	-0.16 (0.97)	-0.22 (0.93)	0.75	0.24
Query suggestions helped me think deeply (i.e. discover new connections between pieces of information in the topic)	0.32 (0.93)	0.93 (0.36)	0.91	0.11
Query suggestions inspired me to ask questions	0.37 (0.97)	1.00 (0.47)	0.06	1.97
Query suggestions helped me reflect on what I had learnt so far	0.32 (0.84)	1.02 (0.47)	0.05	2.13

Figure 4.5: Averages (and standard deviation) of searchers' level of agreement to these statements on a scale of 2 (Strongly Agree) to -2 (Strongly Disagree) for Baseline and CoNotate suggestions. * are significant differences at $p=0.05$ level. Participants reported higher agreement to the statement "Query suggestions helped me discover new terms and concepts" after using CoNotate than after using Baseline

This could be because *CoNotate* encourages more active searching (as discussed above) or because the query suggestions are more helpful. To tease apart these confounding effects, we performed a 2-way repeated measures ANCOVA with condition (*Baseline* or *CoNotate*) and topic (non-native species or mental health), with two covariates for the number of query suggestions presented and number of queries issued. On performing post-hoc analyses using Tukey's HSD and using a Bonferroni correction for multiple comparisons, we found that participants reported a significantly higher increase in self-rated knowledge and to a larger number of domain-specific terms listed on the post- rather than the pre-survey (see Table 4.2 for details). However, there was no significant difference in the number of web pages opened across both conditions. We found no statistically significant differences between the topics and no significant interaction effects between topics and system used for querying behavior. This suggests that the search task topic did not have an effect on information gained when using either of the two search systems.

4.5.3 Effects on user preferences: Participants preferred notes-based suggestions over *baseline* suggestions

To understand how searchers perceived the value of query suggestions, at the end of the study, participants were asked to rate their preference for one of the two systems they had used. 23 participants reported preferring the *CoNotate* for these broad, multi-faceted exploratory tasks, while 13 participants preferred the *Baseline* system, and 2 had no preference. Those who had no preference reported not using the suggestions bar because they spent their time mostly reading and taking notes, rather than issuing queries.

The post-task questionnaire asked each participant to rate their level of agreement on a Likert scale (2=*Strongly agree*, 0=*Neutral or Did not use Query Suggestions*, -2=*Strongly Disagree*) with the statements in Table ???. To check if there were any statistically significant differences between searchers' perceived value of *Baseline* and *CoNotate* suggestions, we ran Friedman tests, along with post hoc analysis using a Bonferroni correction applied. After using *CoNotate* system, participants agreed significantly more strongly to the statement: "*Query suggestions helped me discover new terms and concepts*". Similarly, after using *CoNotate*, participants agreed marginally more with the statements "*Query suggestions helped me reflect on what I had learned so far*" ($p=0.05^*$); and "*Query suggestions inspired me to ask questions*" ($p=0.06$) compared to using the *Baseline* suggestions (refer to Table 4.5 for details). There were no other significant differences (Table 4.5) across the two search task topics or conditions.

4.6 Discussion

This paper presents a novel system, *CoNotate*, that integrates note-taking and searching to recommend contextualized query-suggestions to help explore broad multi-faceted information spaces. To evaluate this approach, we conducted a within-subjects study

where participants ($n=38$) conducted exploratory searches using both a baseline system (standard web search) and the *CoNotate* system. The *CoNotate* approach helped searchers to issue significantly more queries and discover more terminology than standard web search. Also, participants reported preferring using *CoNotate* suggestions over standard web query suggestions.

4.6.1 How does notes-based query assistance support exploration and knowledge discovery?

The *CoNotate* approach appears to encourage more active searching. When using *CoNotate*, participants issued significantly more queries— particularly through the use of the Suggestions Bar — than when using Baseline search. *CoNotate* users also typed fewer queries. Digging deeper, there could be multiple explanations for these behavioral differences. *CoNotate* users not only got query suggestions based on the contents of their notes, they also got them more frequently. *CoNotate* updates its query suggestions list every time the notes document was updated or a new query issued. On the other hand, the *Baseline* interface only updates query suggestions when the user issues a new query. To tease this apart, we considered the number of query suggestions presented to the searchers as a co-variate in all our analyses. Even when controlling for the number of suggestions, participants issued more query suggestions in *CoNotate* than in the Baseline system. This suggests that it is, indeed, the content, the actual words behind the query suggestions that promotes more active searching.

In terms of information gain, we found that participants reported a greater increase in knowledge and discovered more domain-specific terminology, when using *CoNotate* versus using standard web search. This could be because *CoNotate* encourages active searching (as discussed above) or directly because of the content of the query suggestions. To tease these confounding effects apart, when analyzing information gain measures, we

controlled for not only the number of query suggestions presented, but also the number of queries issued. Controlling for these, we still found a significantly greater increase in self-rated knowledge level, number of domain-specific terms, and number of web pages opened in the *CoNotate* system than in the Baseline. This suggests that notes-based query assistance promotes knowledge discovery, particularly domain-specific terminology and information sources. This could be because participants discover new domain-specific terms and sub-topics without even having to open web pages. This would align with previous work that visualizes the topical overview of search results [323, 242]. On the other hand, it could be that notes-based query suggestions also led searchers to find more useful information sources where they discovered these domain-specific terms. Previous work has explored the role of query suggestions in creating information scent (i.e. the proximal cues from which searchers perceive the value of distal information sources) [327, 186, 219, 222]. Since *CoNotate* is able to review what has already been covered in the notes — and look ahead at 100 result snippets across 10 SERPs to glean what has not been covered in notes — it could create a more contextualized trail of information which in turn helps with knowledge discovery.

During the retrospective think-aloud interviews, at least 18 out of 38 participants reported making new interesting connections in the *CoNotate* system. This could be because *CoNotate* query suggestions presented related phrases representing concepts, entities, or perspectives next to each other as query expansions. This parallel presentation could stimulate conceptual blending (i.e. the process of making connections between concepts) or analogical reasoning (i.e. the process of making connections through analogy) [441, 134]. Even in the Baseline system, query suggestions were usually presented as query expansions (e.g. in query auto-completions in the Suggestions Bar, and Related Searches). However, the *CoNotate* suggestions might suggest more heterogeneous and diverse phrases next to each other which could stimulate creative

combination when exploring one query or across queries. Other research seeks to help people break out of filter bubbles in personalized search. Prior work shows that highlighting suspicious sentences [450] and disputed topics in the search results rankings [451] are perceived as useful during credibility assessment. Since NotesGap suggestions explicitly and persistently suggest phrases that are not covered in one’s notes, but are still related, they could be helping people step out and diversify exploration [34]. However, further research is needed to explore how the diversity and heterogeneity of suggestions affects exploration, and creativity.

4.6.2 Study Limitations

As the study tried to balance ecological validity with experimental control, we limited the task time to only a 20-minute session. While this controlled the amount of time taken for each task across participants, exploratory searches often take multiple sessions of searching and note-taking [276, 29, 439], and this might have altered the searching and note-taking behavior during exploratory search [229].

We recruited searchers only above the age of 18 from a university. This recruitment method biased us to a population with a certain level of technical literacy. Also, all participants reported having little to no prior knowledge about the two domains covered by the search task (refer section 4.3). Therefore, our sample size was biased towards a lower domain expertise which impacts search behavior [187, 435]. In future studies we could employ different recruitment and sampling methods to reduce these biases.

In addition to using user-generated notes for the CoNotate algorithm to implicitly infer users’ patterns and gaps, the notes also present an opportunity to experimentally measure each user’s level of knowledge over time (as proposed by prior work [107, 346, 444, 123]). However, due to a logging error not all notes got logged to the the Firebase database with the rest of the study data. Therefore, we leave it to future work to explore how notes

taken over time can be analysed to measure learning over that time period.

4.6.3 Future Work

Many design decisions were motivated by our particular use case of exploring new multi-faceted domains through search. Currently, *CoNotate* shows only six query suggestions in the Suggestions bar and randomly orders it. Future systems could dynamically assess learning from notes, and calculate measures such as relevance, novelty, and diversity to dynamically change the number and order of suggestions presented. Also, *CoNotate* currently presents suggestions as query expansions. This form of query reformulation adds phrases to the issued query, just like query auto-completions and Related Searches, to further refine the previously issued query [344, 222]. Since novices usually start out with high-level goals instead of specific queries [344, 194, 259, 94], this design decision might have helped them further specify their informational goals, find new connections, and therefore issue more queries. However, at least four participants mentioned that they would prefer more natural language, better grammatically-structured queries, such as the People Also Ask feature in standard web search. Future work should consider how different presentations of query suggestions influence search behavior and exploration (e.g. [222]). Furthermore, the text analyses performed to detect patterns and gaps in notes and previous searches do not consider the context or structure of the phrases. We leave it to future work to explore how additional information signified by the context and structure of notes can inform the text analyses algorithms.

While proactive support can be beneficial, a key challenge is providing assistance without being too disruptive [282, 423, 448]. Since newbies to a topic may not even realize when they need help, *CoNotate* proactively provides suggestions every time the user issues a new query or edits the notes document. Some participants found this distracting. Improving the contextual understanding of the searcher's workflow [74, 107]

and assessing their current knowledge level [347, 417, 157] could help inform both the timing and content of query suggestions.

CoNotate requires searchers to write notes in order to suggest relevant phrases. When the user has not yet added anything to their notes document, no query suggestions are provided. Mining user data and notes history is, by definition, retrospective (i.e., it describes what the user has already done). In contrast, search is often prospective (i.e., looking for something the user has not explored yet). To overcome this cold-start problem, *CoNotate* suggestions could be combined with standard query assistance features (such as autocomplete, People Also Ask, and Related Searches) to offer relevant queries when the user first begins to search, and transition to notes-based queries as notes accumulate.

In the post-survey, we asked participants to rank their agreement with the statement: "*Query suggestions helped me structure my notes better*". Participants, on average, disagreed in both the Baseline and *CoNotate* versions. While our system focuses on leveraging notes to guide querying and exploration, prior work has already explored how to help searchers quickly gather relevant information [42, 256] and make sense of it (both individually and collaboratively) [176, 110, 333, 111, 163]. Future work could explore integrating these systems with *CoNotate* to build a holistic system that integrates the knowledge development workflow more closely with the search process. This system could help with quick capture, sense-making, re-finding and sharing of information [294, 297], as well as adaptively guiding querying and exploration. Future work should examine how these alternatives might change people's behavior and workflows over time.

This project's major insight is that given a source of user context (e.g., notes), search systems can highlight patterns and gaps to guide exploration of a broad multi-faceted information space. Beyond notes taken, search systems could glean contextual information from other artifacts (e.g. documents, emails, annotations etc.) that give insight into a user's goals that could guide exploration and knowledge discovery. *CoNotate*

demonstrates this approach using text; however, it could be modified to mine other types of content. Extending beyond this domain, different exploratory activities (e.g. programming, exploratory data analysis, visual design) may benefit from mining other types of content (e.g. code, images, video). For example, it could mine the searchers' Jupyter notebook and data-set to infer what variables, connections have already been explored, and recommend insightful connections that are yet to be explored. It could build on current systems that already recommend contextual help to debug programming errors (like Unakite [260]) or present example code (like Blueprint [53]) to further guide exploration of that data and information space. Similarly, for visual design search systems could mine artifacts like mood-boards and large idea galleries (like Behance or Pinterest) to detect patterns and gaps in exploration of the design space. A challenge for future work is to convert video, code, image data to textual data that can be parsed by a search system [145, 425]. We leave it to future work to overcome these limitations, build and test out suggested design improvements from these findings.

4.7 Conclusion

This paper introduces an approach that integrates contextual information present in note-taking and search systems to recommend query suggestions. CoNotate, our prototype system, shows that by detecting patterns and gaps in a user's notes and the SERPs of issued queries, we can inform query suggestions in a flexible, domain-general way. A comparative user study demonstrated that notes-based query suggestions helped people explore broad multi-faceted information spaces by promoting active querying, and discovery of domain-specific terminology and information sources. Future work should investigate these challenges and examine how contextual help affects workflows in the real world through a longitudinal study. This work brings us one step closer to

leveraging the wisdom of the Web for contextualized knowledge discovery and learning.

4.8 Acknowledgements

This chapter in part, includes portions of material as it appears in *CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery* by Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, Steven P. Dow in Proceedings of the 2021 ACM CHI Conference on Human Factors in Computing Systems (CHI'21). The dissertation author was the primary investigator and author of this material.

Part II

Symbiotically Supporting Information

Exploration and Synthesis

Chapter 5

InterWeave: Presenting Search

Suggestions Within User's Evolving

Sensemaking Structures Promotes

Information Exploration and Synthesis

Web search is increasingly used to satisfy complex, exploratory information goals. Exploring and synthesizing information into knowledge can be slow and cognitively demanding due to a disconnect between search tools and sense-making workspaces. Our work explores how we might integrate contextual query suggestions within a person's sensemaking environment. We developed InterWeave a prototype that leverages a human wizard to generate contextual search guidance and to place the suggestions within the emergent structure of a searcher's notes. To investigate how weaving suggestions into the sensemaking workspace affects a user's search and sensemaking behavior, we ran a between-subjects study (n=34) where we compare InterWeave's in context placement with a conventional list of query suggestions. InterWeave's approach not only promoted active searching, information gathering and knowledge discovery, but also helped participants keep track of new suggestions and connect newly discovered information to existing knowledge, in comparison to presenting suggestions as a separate list. These results point to directions for future work to interweave contextual and natural search guidance into everyday work.

5.1 Introduction

People increasingly use web search to learn and work online. When searching the Web to address complex, exploratory information goals – such as academics reviewing literature, policymakers researching policy briefs, lawyers engaged in case discovery, startup founders performing market analysis, or individuals learning how to take care of a loved one – people not only look up facts, they also read, collect articles and take notes to make sense of the information space. However, exploratory information-seeking is often arduous and difficult. The user must first articulate a search query to fulfil their information goals. This can be especially challenging in new areas where people often

lack domain knowledge to know what to ask, let alone how to ask it [439, 346, 276]. Then, once the user finds useful information, they must switch their attention back and forth between the resource and sensemaking applications – like note-taking tools – where they collect, annotate, and synthesize information from multiple queries, sources, and sessions. Furthermore, to make progress on exploratory, complex projects, users must synthesize and make connections between newly discovered information and their existing knowledge about the topic [439, 358, 74]. The work required to synthesize information while continuing to discover new resources can be time-consuming and cognitively demanding.

To help alleviate some of these challenges around exploratory search, search engine developers and researchers have devoted much attention to developing and fine-tuning search recommendation and suggestion algorithms. For example, current search engines attempt to assist with query formulation such as: *Auto-completions* to help people type queries quicker, *People Also Ask* to clarify the information need, or *Related Searches* to explore related topics [31, 203, 270, 79, 144, 336]. Researchers have also explored presenting search guidance in representations such as hierarchical lists [69], concept maps [362, 71, 323], lists of stacked bar charts [412] and trails [436, 44]. While evaluations of these systems show evidence of supporting active search processes, they often create a representation space that is independent of the searcher’s own representation of the information space [436, 69, 362, 326]. This forces searchers to reconcile the two representations or to adopt the representation provided by the system (e.g., using the category space from Topic-Relevance maps). This also forces users to switch back and forth between the query suggestions lists and their own work to check for updates. This context switching is not only distracting and cognitively demanding, it also makes it hard to discover updated suggestions and integrate new information into the sensemaking workspace. Our work explores how we might integrate contextual search suggestions

within a person's sensemaking environment.

Prior work has shown that integrating guidance with the user's work context can make it easier to seek help for learning and creative production [317, 143, 148, 147, 170, 283]. Modern text-editing software (e.g. Google Docs, Microsoft Word) includes the ability to select phrases in the document and issue them as queries. Personalized search systems go further by recommending suggestions based on user-generated content. For example, Teevan et al. [406] re-rank search results to help users find information quicker by implicitly inferring interests from user-generated documents and emails. More recent systems such as CoNotate [317] and ForSense [334] demonstrate how search systems can offer search and sense-making suggestions based on analyzing the searcher's notes and previous searches for patterns and gaps in information. While this approach helps make query suggestions more relevant, these suggestions are typically presented as a list separate from the user's work context. Therefore, users still need to context switch back and forth between their search tool and sensemaking workspace.

Recent work has also demonstrated the benefits of presenting search suggestions within the workspace where the information is used. This has been particularly explored in the context of computer programming [148, 147, 170] where embedding software tutorials [147, 170, 93, 160, 196] and discussion topics [283] reduces the need for context switching and supports active learning. It is unclear whether contextual placement of search query suggestions also provides an advantage for free-form, unstructured activities like note-taking.

To explore the potential of weaving query suggestions directly into a user's emergent synthesis of a knowledge space, we developed a wizard-of-oz prototype [128] called *InterWeave* as a web browser extension that piggybacks [169] on top of the online whiteboarding platform Miro (<https://miro.com>). *InterWeave* embeds search suggestions within the emerging representation of a searcher's sensemaking structures (Figure 5.1).

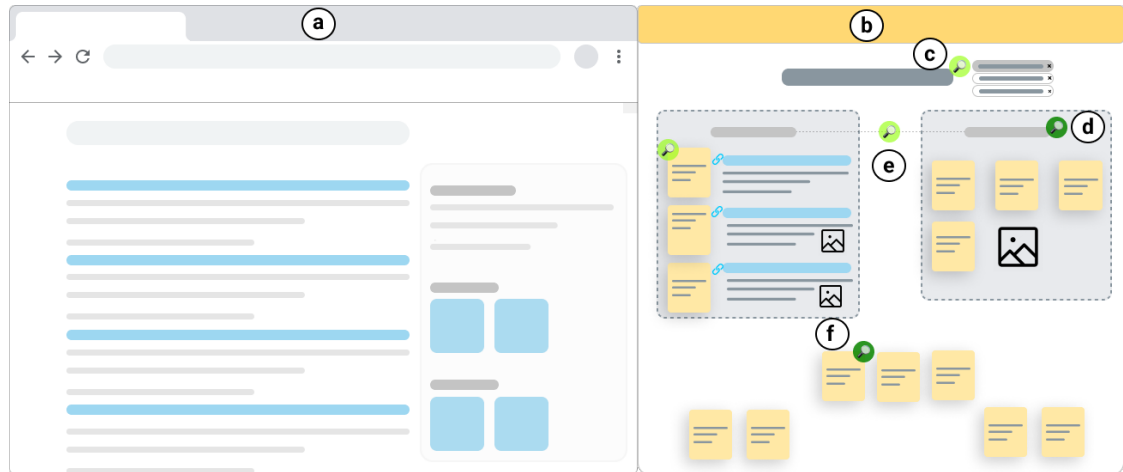


Figure 5.1: InterWeave’s user interface augments (a) a search browser with (b) a sensemaking workspace where contextual search suggestions are presented at up to four levels within user’s evolving sensemaking structure at the (c) title, (d) cluster, (e) cross-clusters, and (f) individual note levels

Different types of suggestions appear (1) on the document title, (2) around clusters of similar information (3) across multiple diverse clusters and (4) on individual units of information. InterWeave was built as a wizard-of-oz prototype where a confederate observes how users search and add content to notes. The wizard paid attention to the content and structure of the searcher’s notes and previous searches, in order to infer relevant and potentially undiscovered information. The wizard then has the ability to recommend pre-assembled query suggestions at the appropriate level of the emergent sensemaking structure. The context-aware search suggestions appear seamlessly integrated into the user’s representation of information.

To evaluate how in context placement of suggestions affects search, sensemaking, and learning behaviors, we conducted a between-subjects study (n=34) where we compare InterWeave’s placement of suggestions with a conventional list of query suggestions. Participants search the web on an exploratory topic (e.g. future of space travel or environmental impacts of COVID-19 pandemic), while they also collect information, take notes, and synthesize their knowledge within the digital whiteboard space. Participants

were randomly assigned to either *InterWeave* or a baseline system which lists the same suggestions outside the user's sensemaking context. The baseline condition attempts to simulate the placement of suggestions on general-purpose search engines (e.g. Google, Bing) while controlling for the content, quantity and timing of query suggestions.

Our analysis shows that, compared to seeing a list of query suggestions in the web browser, *InterWeave* participants issued significantly more queries, discovered more domain-specific terms and concepts, gathered more information and made connections across subtopics towards a more holistic understanding of the topic. Also, participants reported that the *InterWeave* suggestions were more easy to discover, led to greater information gain, and helped them connect new information to information already gathered. These results provide directions for future work to interweave contextual and natural search guidance into everyday work. This chapter offers the following contributions:

1. We conceptualize the potential of inferring a user's emergent sensemaking structures in order to present query recommendations weaved into a note-taking and synthesis workspace.
2. We created a prototype, *InterWeave*, that leverages a human wizard to present contextual search guidance on a digital whiteboard and weaved into the emergent structure of searchers' notes.
3. We conducted an evaluation study that demonstrates the *InterWeave* approach not only promoted active searching, information gathering, and knowledge discovery, but also helped participants keep track of new suggestions and connect newly discovered information to existing knowledge, in comparison to positioning suggestions as a list.

5.2 Related Work

This research builds on prior work related to information foraging and sensemaking assistance during complex, exploratory work.

5.2.1 Exploratory Information Seeking

Most people use web search to look up facts or to get timely information to complete some other task. But people increasingly use the Web to explore, learn and do more complex information synthesis for more open-ended goals. For example, academics reviewing literature, designers exploring which tool to use, startup founders performing market analysis, or individuals exploring, learning and making decisions like where to vacation. *Exploratory searches* involve multiple iterations and return sets of information that require cognitive processing and interpretation and often require the information seeker to spend time scanning/viewing, comparing, critically assessing and making qualitative judgments before being integrated into personal and professional knowledge bases [276, 439]. The search task does not exist in isolation from the surrounding task context. Not only does the context influence the performance of the task, but it also affects what action should be taken with the found information. Given the strong relationship between exploratory search and information use and information understanding, it is likely that these searches will involve engagement with multiple applications in the user's information workflow.

People engaged in exploratory searches are generally: unfamiliar with the domain of their goal (i.e., need to learn about the topic in order to understand how to achieve their goal); unsure about the ways to achieve their goals (either the technology or the process); and/or even unsure about their goals [439]. There may also be periods of heightened uncertainty and confusion as people try to articulate their information

needs, discover new information and assimilate knowledge to make sense and acquire meaning. Exploratory search can give rise to feelings of doubt, confusion, frustration, and anxiety [240]. The complexity and uncertainty of exploratory search leads to a nonlinear, dynamic process involving a tacking back and forth between deduction and induction [63]. It involves balancing divergent thinking with the convergence of ideas [140]. The processes of exploring and working with information are critical for building connections, discovery, and creativity. These processes rely on the effective provision, processing, and manipulation of information at all stages of an exploratory search and information work. As the information need evolves, the searcher's ability to articulate query statements and identify relevant information increases based on their improved level of problem comprehension [40, 439]. Furthermore, the creativity, innovation, and knowledge discovery that is often necessary as part of exploratory searches requires traveling beyond what is known by the user – exploratory search involves lateral thinking, and serendipitous connections [35, 142].

Systems such as the Relation Browser [277], Phlat [109] and mSpace explorer [369] try to support exploratory search by dynamically updating presentation of search results in real-time during the session. Other systems, such as [179, 243] employ categorization or clustering of search suggestions and results. To determine how well systems support exploratory search activities, they must be evaluated in terms of their ability to facilitate key elements of search exploration such as helping users obtain new insights, assisting learning, etc. [439]. Therefore, in our evaluation study we not only measure search activities, but also information gathering, sensemaking and learning activities. InterWeave aims to build on this prior work by leveraging search context to support exploratory search, particularly query formulation, learning and understanding.

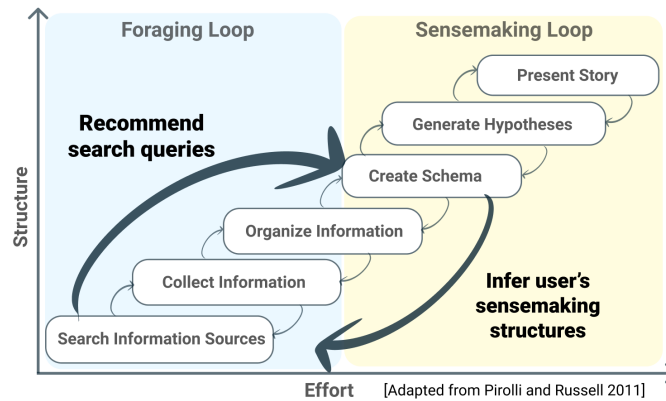


Figure 5.2: While many search systems recommend search queries, InterWeave goes further by inferring the user’s sensemaking structures, formulating context-aware query suggestions and then weaving suggestions back into the sensemaking workspace

5.2.2 Integrating Search and Sensemaking

During the exploratory knowledge discovery process, people are constantly engaged in sensemaking activities as they move through the information space. They take notes, gather information, and create representations to organize information to free their mind from having to recall everything [420, 280, 256], and from having to mentally synthesize all the information [225, 206, 275, 163]. This process of encoding information into external representations to answer complex, task-specific questions is referred to as *sensemaking* [358, 357].

Figure 5.2 (adapted from [328, 330, 357]) illustrates how foraging and sensemaking activities can be organized and iterated through during knowledge work. During the foraging loop, people search for information by interacting with search results, web-pages and other information sources. As they process this information read, they collect and curate relevant and promising information by clipping and extracting information from web pages. Then, they start organizing it into structures, haphazardly at first and later systematically into a schema. *Schema* are representations of the knowledge and understanding gained during the exploration and sensemaking process. Schema can be

essay outlines, comparative pros and cons lists, concept maps, etc. The searcher continues the sensemaking process until they have developed a concrete, well-tested schema. Schema or sensemaking structures can change slightly to assimilate new information, or significantly to accommodate new paradigms and perspectives [326, 358]. As the searcher develops more concrete and polished schema, they progress to a state where it can be presented in a narrative that makes sense - for example in an essay or article.

Prior work has focused on designing tools help with quickly moving information from the information foraging loop to the sensemaking loop (refer to Figure 5.2) [327, 328, 357, 358]. For example, there are several research and industry tools to support active reading while searching using highlighting and note-taking [356, 355, 107], collecting information by bookmarking and clipping web content [42, 176]), curating and organizing collected web content in a way that helps make sense of information [430, 261, 110, 84], re-finding information or resuming search sessions [294, 425, 143].

However, there has been relatively little work done to support query formulation and the foraging loop based on the searcher's context-rich sensemaking. Recent work has started to explore this opportunity of leveraging user-generated content and sensemaking to support search. For example, InkSeine [185], Google Docs and Microsoft Word allow people to issue words and annotations in their notes as queries. However, these methods still rely on the user to identify and articulate their information need as queries, and do not guide the searcher to further explore their knowledge gaps. Research systems like CoNotate build on this and offer query suggestions based on analyzing the searcher's notes and previous searches for patterns and gaps in any multi-faceted information space [317]. Similarly, ForSense suggests parts of web pages to clip and cluster based on what information the user has previously clipped and gathered [334]. InterWeave builds on these systems that leverage not only the content of the user's sensemaking, but also embeds contextual suggestions in the user's evolving schema and sensemaking

knowledge structures.

5.2.3 Presenting Search Suggestions

Current search engines support query formulation with assistance such as: *Auto-completions* to help type queries quicker, *People Also Ask* to help clarify the information need, or *Related Searches* to help explore related topics [31, 203, 270, 79, 144, 336]. Research systems designed to support search have also explored different ways of presenting query suggestions. For example, *Search Trails* visualizes how previous searchers explore an information space [44, 73, 387, 455]. *ScentBar* [412] visualizes to what extent valuable information remains to be collected from the search results of individual queries. *SParQS* [219] helps searchers understand inter-query relationships by presenting query suggestions into automatically generated categories. *Topic-Relevance Map* [323] visualizes a topical overview of the search result space as keywords with respect to relevance and topical similarity. These search tools can be cognitively overwhelming because they require the searcher to not only articulate their ill-defined information goals as queries initially, but also reconcile the two representations or to adopt the representation provided by the system. Also, they have to constantly switch back and forth between the suggestions lists and their work to check for updates.

In the related field of software learning, research has shown that presenting resources, such as relevant software videos [148, 147], tutorials [148, 170], and discussion fora [453, 283], in context reduces the need for context switching and supports active learning [170, 145]. Similarly, other systems embed resource suggestions such as reusable examples [384, 53], executable operations [146] which helps people more easily integrate these into their tasks. In this chapter, we introduce InterWeave, a system that presents query suggestions within the searcher's evolving sensemaking context and structure, and evaluate whether it makes sense to weave work-aware suggestions into the sensemaking

workspace or to present them as a separate list, as most general-purpose search systems currently do.

5.3 InterWeave

InterWeave is a web-browser extension and a wizarded prototype that presents contextual search suggestions within the user's evolving sensemaking representations. In this section, we first describe the user challenges that inspired our design goals, then we provide details on the system's user interface and its implementation.

5.3.1 User Challenges & Design Goals

Inspired by the extensive prior work done by the HCI and IR communities to document the user challenges when searching the web to address complex, exploratory information goals, we identified our design goals. These are the user challenges we aimed to address:

- It is cognitively overwhelming and time consuming to **switch attention back and forth between the search browser and sensemaking applications** – like note-taking tools – where people collect, annotate, and synthesize information from multiple queries, sources, and sessions [148, 355, 356, 74].
- When exploring a new domain through web search, people often **struggle to articulate queries** because they lack domain-specific language and well-defined informational goals. [439, 29]
- When encountered new information during an exploratory search session, people often **struggle to synthesize** and make connections between newly-discovered information and their existing knowledge about the topic [439, 74, 29]

Based on these user insights from prior work, we present InterWeave's key goals and design principles:

- **Integrated with Sensemaking Workspace:** to support quick connections between newly-discovered information and their existing knowledge about the topic the suggestions should be well-integrated and adapt to the users' sensemaking externalized in their sensemaking workspace. The system should present timely and limited options for search that arrange spatially within notes in their sensemaking workspace.
- **Context-aware:** The suggestions should be relevant and connected to what the searcher currently knows, however, it should still push them to learn about information that is a certain extent beyond their current level of knowledge.
- **Discoverable:** the searcher's should be able easily find and interact with the suggestions
- **Easy-to-learn:** the user interface should have a smooth learning curve and build on existing tools they use.
- **Domain-general:** The suggestions should not be domain-specific, and adapt to provide contextual guidance regardless of the searcher's domain or topic. This system should work across any topic or domain.
- **Natural note-taking:** We ensured the interactions within the sensemaking workspace were based on studies of note-taking during search. Our note-taking interface was designed to allow flexible, idiosyncratic note-taking styles since individuals structure notes very differently [107].

5.3.2 InterWeave Interface

To investigate how the presentation of search suggestions affects search, sensemaking, and learning behavior, we wanted to build a system that just slightly modifies the search and sensemaking tools that users might already use. Therefore, we designed InterWeave as a Chrome browser extension that is integrated with with Miro (<https://miro.com>), a general-purpose digital whiteboard. Chromium-based browsers (e.g. Google Chrome, Firefox, Microsoft Edge) make up 80% of the world's market search browser market share [8]. Miro is used widely used by 20 million users, and more than 100,000 enterprise clients [5, 6]

InterWeave shows a digital whiteboard space for notetaking and sensemaking (Figure 5.1b) on the right of any Chrome browser on the left (Figure 5.1a). Each window defaults to 50% of the user's screen, but can be re-positioned and sized as desired. Miro offers the basic tools for adding and modifying text, images, videos, etc. and users may use the infinite 2D space to spatially arrange their notes. Users can take notes either by typing, adding sticky notes or dragging and dropping in links, images, videos, etc. from the browser. When users want to explicitly relate two pieces of content, they can draw a line between them. When they want to form a cluster, they can use the cluster tool to draw an outline box around the content they want to cluster. Clusters usually indicate semantic similarity or conceptual relatedness [107, 17].

Suggestions appear as green search icons within the searcher's emerging sensemaking structure. Different types of suggestions appear (1) on the document title (Figure 5.1c), (2) around clusters of similar information (Figure 5.1d) (3) across clusters (Figure 2e) and (4) individual units of information on note-cards (Figure 5.1f). Clicking on any green search suggestion icon opens a list of suggestions at that location (Figure 5.1c). Dark green icons indicate that there are new query suggestions at that location (Figure 5.1d, 5.1f). Light green indicate that all the query suggestions at that location have been

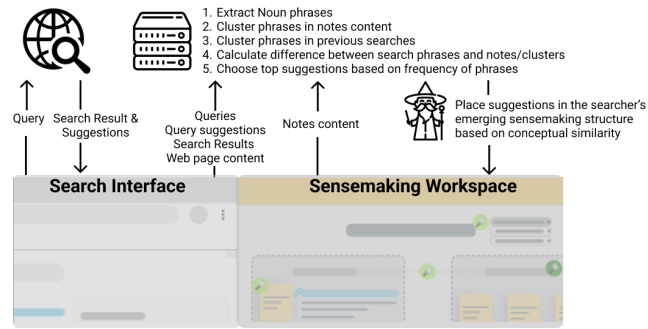


Figure 5.3: InterWeave’s system architecture which leverages NLP algorithms and a wizard to present contextual suggestions within the searcher’s emergent sensemaking representations.

previously viewed (Figure 5.1c, 5.1e). Clicking on a suggestion in the list at a location issues the suggestion text as a new query and displays search results in the web browser.

To add additional context cues, the suggestion text is appended with the text at the corresponding location in the sensemaking structure. For example, title-level suggestions append the document title to the suggestion text before issuing it as a query. Similarly, the cluster-level suggestions add the cluster-title text to the suggestion text and the cross-cluster-level suggestions append the corresponding clusters’ title texts to the suggestion text when issuing it as a query. For the notes-level suggestions, the notes’ content is appended to the suggestion text when issuing it as a query. However, if the note on which a suggestion is placed has more than 10 words, then the document title is appended instead.

5.3.3 System Architecture

Infer searcher’s current knowledge level

(NLP) First, to implicitly infer the searcher’s current knowledge level, the system’s NLP algorithm mines the searcher’s sensemaking workspace for noun-phrases at regular intervals and creates a dictionary called *sensemaking_pphrases*. The system considers these

to be a snapshot of what they have explored so far and found interesting [231].

Generating queries that guide the searcher to new areas of knowledge

(NLP) To surface additional opportunities for exploration, the system also mines the content of the top 100 Search Engine Results Pages (SERPs) of each issued query and websites visited for noun-phrases from the titles and snippets to create a dictionary called *SERP – phrases*. Since the suggestions aim to present opportunities to expand exploration by suggesting phrases/concepts mentioned in the SERPs but missing from the sensemaking workspace, we calculate the difference between *SERP – phrases* and *sensemaking – phrases* and create a new dictionary called *gap – phrases*, which is ordered based on the number of times each phrase occurs in the SERPs. For every significant change to the notes (>50 characters) or each new query issued, the system can only present three new suggestions to avoid overwhelming the searcher with too many suggestions while still providing proactive guidance. The top three *gap – phrases* are chosen to be sent to the wizard as search suggestions.

Placing the suggestions with respect to emerging sensemaking structures

(Wizard-of-Oz) Then, the wizard selects where to place these suggestions within the searcher’s emerging sensemaking structure. We decided to use a wizard-of-oz

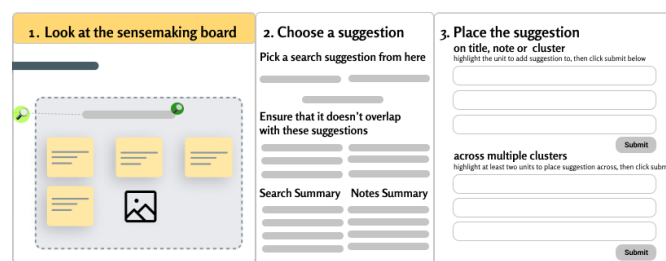


Figure 5.4: Wizard’s interface when choosing and placing search suggestions in the emerging sensemaking structure

approach to quickly prototype how the presentation of query suggestions would affect search, sensemaking and online learning behavior. The wizard places the suggestion at a particular 2D location in the searcher's information hierarchy based on the *conceptually similarity* to what is already in the emergent sensemaking structure at a particular location.

The wizard used the following heuristics for choosing between four options to place query suggestions::

- The *title-level suggestions* aim to present opportunities to expand exploration by suggesting phrases/concepts that are entirely missing from the notes, and conceptually far from the phrases mentioned in clusters and note cards, but still related to the topic. The wizard checks the phrases on the board at the cluster and note card level to ensure there is little overlap with themes there before presenting title-level suggestions. For example, say the board has clusters about "air pollution", "water pollution", and the wizard sees suggestions such as "heritage conservation", "global warming", "restaurants", the wizard will present "heritage conservation" and "restaurants" at the title as these are conceptually far and missing from the searcher's notes.
- The *cluster-level suggestions* aim to present opportunities to dig deeper into the information mentioned within a cluster of notes and other clusters of notes in the sensemaking work-space. The wizard considers conceptual similarity between the suggestions and the phrases in this particular cluster to suggest conceptually similar, but missing concepts from the cluster. Extending the example from above, suppose the cluster is about "air pollution" and wizard sees suggestions for "heritage conservation", "global warming", "restaurants", then the wizard will suggest "global warming" on the cluster as this is conceptually similar to "air pollution" but is not already included in the cluster.

- The *cross-cluster suggestions* aim to present opportunities to learn more about the concepts/phrases at the intersection of more than one cluster. Therefore, if a suggestion is not mentioned on the board, but is conceptually similar to more than one cluster, the wizard will choose to present this at the intersection of the conceptually-similar clusters. Say the board has clusters about "soil pollution", "water pollution", and the wizard sees suggestions such as "heritage conservation", "global warming" and "farming", the wizard will present "farming" on a line connecting the "soil pollution" and "water pollution" clusters as this is conceptually similar and relevant to both clusters.
- The individual note-level suggestions aim to present opportunities to dig deeper into the information mentioned on a particular notes unit. The wizard considers conceptual similarity between suggestions and the phrases on this particular note-card to suggest similar, but missing concepts on this card. For example, if the note card is about "ozone spikes" and the wizard sees suggestions such as "CO2 emissions", "climate change", "restaurants", the wizard will suggest "CO2 emissions" on the note-level as that is conceptually similar to "ozone spikes", but is not mentioned in the note-card.

The system presents a set of suggestions that is mutually exclusive and unique from a general-purpose search engine's suggestions (e.g. Google's suggestions). Before presenting the searcher with the suggestions, the wizard compares and excludes the general-purpose search engine's query suggestions which have been scraped and presented as a list to the wizard (Figure 5.4 (top of panel 2)).

For the purpose of this prototype, the wizard determines conceptual similarity by taking into account the following factors: (i) lexicographic similarity (i.e. overlapping words e.g. "air quality" and "air pollution"); (ii) semantic similarity (i.e. relationships between concepts/phrases often calculated using domain-specific ontologies e.g. "car" is

similar to "bus" and related to "road" and "driving"); (iii) and structural similarity (i.e. words that co-occur in the same part of the document, e.g. "air pollution" and "tourism" could occur under the same heading in an article suggesting they are conceptually related).

The wizard was a member of the research team that spent six weeks learning and training up on each study topic and gaining expertise. Also, they had prepared a sheet summarizing their knowledge on each topic to help aid them in placing each suggestion in real-time. Since the wizard had gained knowledge in each area and was assisted by NLP algorithms that summarize the searcher's activities, it is easier for the wizard, compared to current state-of-the-art information retrieval and machine learning algorithms, to determine conceptual similarity of query suggestions in real time and place the query suggestions within the searcher's emerging sense-making structures. The system is mostly automated and the wizard's task of placing NLP algorithm generated suggestions within the sensemaking structure based on conceptual similarity heuristics is assisted by clear instructions, and information sheets the wizard created during their six weeks of research to summarize their knowledge. We discuss the limitations of this approach further in the §6.2 of the Discussion section.

5.3.4 Implementation

InterWeave is a chromium-based web browser extension that employs Google Chrome javascript APIs for the front-end, a Flask Python framework as a web socket server. In the server, we process the natural language content from the websites, SERPs and notes documents using BeautifulSoup4 [7] for parsing, TextBlob [266] for noun phrase extraction, NLTK [45] and sklearn [322] for k-means clustering. We bridged the browser to the sensemaking workspace by developing a Miro web plugin using the Miro REST APIs [9]. The wizard saw, chose and placed suggestions on the users' boards also using a separate Miro web plugin.

During the experiment, we logged all interactions with the search browser and the sensemaking workspace to a Realtime Firebase database [3]. To ensure privacy during data collection, we automatically anonymized and encrypted all data by creating anonymous session and Firebase IDs. Please refer to the open-source code in the supplementary materials or linked here ¹ for implementation details.

5.4 Study: Where to place suggestions?

While presenting query suggestions within the searcher’s emerging sensemaking structure might help searchers quickly explore the information contextualize the suggestions in their work, make suggestion easier to discover, and reduce the need for context switching between the browser and their notes to integrate learner knowledge, these can also be distracting, cognitively overwhelming and confusing. To investigate how the presentation of search suggestions impacts search, sensemaking and learning behavior, we conducted a between-subjects experiment. 34 participants were asked to search the Web, gather, take notes on, and synthesize information on a given topic. We collected usage logs of each participant’s interaction with the search browser and sensemaking workspace, as well as self-report data about their perception of the search suggestions’ content and presentation.

5.4.1 Conditions

Participants were randomly assigned to search and make sense of a topic using either *InterWeave* or the baseline system which lists the same suggestions outside the user’s sensemaking context. The baseline condition (Figure 5.5) augments the traditional web browser interface (a) with (b) a list of contextual search suggestions and (c) a

¹<https://github.com/creativecolab/IntegratedSearch>

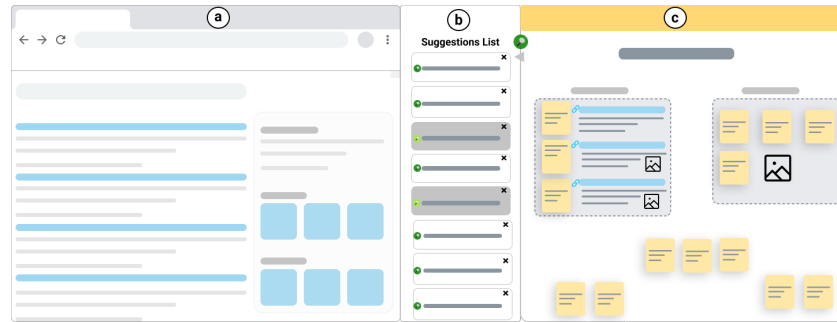


Figure 5.5: The Baseline Condition lists suggestions outside the user's Sensemaking Workspace

sensemaking workspace where people can take free-form notes. This condition tries to simulate the lists in which we see search suggestions on general-purpose search engines (e.g. Google, Bing) while controlling for potential effects suggestions' content, quantity and timing. This let us distinguish the effect of access to suggestions per se from the effect of presenting suggestions in a context-aware manner. To ensure parity across conditions, we only changed where the search suggestions were presented, and kept all other system features the same.

This list gets updated based on patterns and gaps in the searcher's searches and note-taking. This list does not disappear when the searcher navigates to a new webpage (unlike the current query suggestions which are only offered on the search results page). This list can be minimized by clicking the search icon at the top. When there are new suggestions the Suggestions list icon glows green. If a searcher has already seen all the search suggestions in the list, the green fades away. Clicking on a suggestion issues the suggestion text with the topic append as a new query and displays the search results in the Search Interface. Suggestions that have been issued have a grey background.

Lastly, so as to not bias the wizard, the wizard does not know whether the searcher is seeing the InterWeave or other experimental interface. They only see a mirrored version of the searcher's board, with the search suggestions as placed in the InterWeave interface.

5.4.2 Participants

We recruited 34 participants (21 female, 1 non-binary; average age 23.69) through online advertisements (on Prolific, an online diverse world-wide participant pool), and e-mails to remotely-enrolled students at a university. All studies were conducted remotely over a video conference call because of a pandemic. As incentive for participating in the 90-minute study, participants received \$15 or equivalent gift card. Our institution's ethics review board approved all recruitment materials and entire study procedure.

When asked about their background using search tools, all participants reported that they use search engines for look up searches multiple times a day. 14 of them reported performing exploratory searches at least once a week, 13 said multiple times a week and 7 said daily. 23 self-reported as proficient in search, 11 as experts. When asked about their background using sensemaking tools, 22 participants reported taking digital notes multiple times per week, 12 said daily. When asked about how frequently they mind map, 11 said never, 12 said multiple times per week, and 11 said daily. 13 reported being competent at digital note-taking, 12 reported being proficient and 9 self-reported as experts. When asked about their experience with research, 10 reported being competent, 12 as proficient and 12 as experts.

5.4.3 Task

To help situate their searching and sensemaking [50], participants were given a prompt:

"Imagine that you are a journalist writing an article for an online magazine. As part of that process, your editor asked you to do research for an article on the following topic:

[One of two search task topics: Environmental Impacts of COVID-19 OR Future of Space Travel]

Today, your editor would like you to do initial research to get a broad overview of the topic. Your goal should be to identify as many terms,

concepts and perspectives related to the topic as you can find by searching and gathering information on the internet. Use the sensemaking canvas displayed on the right-window to gain a broad and deep understanding of the topic.”

Participants were randomly assigned to one of these two topics:

1. ***Environmental Impacts of COVID-19:*** The recent pandemic has brought about unprecedented changes in our daily lives, requiring us to adopt habits and measures, such as wearing surgical masks, that may be new to many. These new changes have various unintended environmental consequences. At this stage, your editor asked you to collect information about the environmental impacts of COVID-19 as the first step before writing an article about it.
2. ***Future of Space Travel:*** Several billionaires have dedicated projects investing in space travel. More specifically, private companies are emerging as new actors in the future of space travel. At this stage, your editor asked you to collect information about factors affecting the future of space travel as the first step before writing an article.

We chose these two task topics as they are relatively large and complex information spaces and the average person has relatively limited knowledge coming into the task. This effectively simulated a work scenario where participants would need to search and take notes in order to explore and synthesize their topic knowledge.

5.4.4 Procedure

Participants were randomly assigned to one of the two task topics to search and make sense of using one of the two interface conditions (InterWeave or Baseline). Participants answered a pre-task questionnaire which asked questions about their prior knowledge-

level on the topic, and watched an 10-minute long video that presented the main features of the system (see Supplementary Videos) before the task.

Then, participants were asked to search the Web, collect, take notes on, and synthesize information on their task topic for 45 minutes. During the 45 minutes of using the interface, participants could use the system to issue queries, view pages, and take notes, as they naturally would. Next, participants answered a post-task questionnaire which asked questions about their knowledge-level on the topic after their search session; and discuss their perception of the query suggestions' content, presentation and their interpretation of how the suggestions were generated.

Lastly, to gain insight into the participant's thought processes, participants were asked to perform a retrospective think-aloud (for a maximum of 10 minutes) as they scrubbed through a screen-recording of them doing the task. They were prompted to reflect on how and why they issued each query, added information to the board, etc. and how the query suggestions and their presentation affected their process.

5.4.5 Measures

To observe and analyze the differences in search, sensemaking and learning patterns across searchers who saw the suggestions placed within and outside their sensemaking structures, we measure the following:

Search Behavior Measures

From the search logs we measured: *Number of queries issued; Number of query suggestions issued; Number of queries typed; Total number of query suggestions presented to the searchers during the session; Number of webpages opened.*

Sensemaking Behavior Measures

To observe patterns in their information gathering and sensemaking behavior, we logged interactions with their sensemaking work-space. The sensemaking measures are based on the Sensemaking Model by Pirolli and Card (Figure 2, [328]) and prior work [414, 202, 444]. Information gathered is the second step in the model and therefore we measure the quantity of information gathered (as number of words) in the sensemaking workspace as a measure of sensemaking [414, 444]. The third and fourth steps in the model are organizing information and creating schema, respectively. The sensemaking workspace supported organization and schematization of information by forming clusters, drawing connections between notes or labeling the cluster titles. Therefore, we measure the number of connections as Breadth of Sensemaking, and the average number of words within each cluster as Depth of Sensemaking.

Learning Measures

To measure learning as information gain, we examine the change in knowledge level between the pre- and post-surveys:

(i) *Change in Self-rated knowledge* where the participants were asked to rate how knowledgeable they were on the topic on a scale of 1-5, where higher is more knowledgeable, before and after searching.

(ii) *Change in number of domain-specific terms listed*: We asked participants to “Please list any terms/concepts/phrases you currently know about this topic” pre- and post- search task. We calculated learning as the difference between the number of unique domain-specific terms listed both pre- and post-task by each participant. Free recall of domain specific terms and our operational definition of information gain have been used consistently by the search-as-learning and IR communities to measure learning [414, 346]. To clean the data of not domain-specific words, a domain expert curated a

standard glossary of terminology based on gathering participants' responses to this pre- and post-task question, and removing generic terms.

(iii) *Change in number of idea units listed*: Most prior work involves asking participants to demonstrate what they have learned by producing a written summary and measuring the change in number of recalled facts or ideas [444, 346, 414]. We choose not to use a quiz format to measure learning: during open-ended exploratory tasks, users traverse and discover information from a much larger unconstrained space of information on the web. Even a reasonably long quiz would limit the areas of knowledge that could be tested. Therefore, we asked participants to “*Please summarize what you know about this topic*” both before and after the task. Change in the number of facts has been used as a learning measure by the search as learning communities [414], however since participant's statements were not always facts but sometimes ideas or opinions, we calculated learning as the change in the number of unique idea units written about pre- and post-task by each participant. Two raters coded the number of idea units in each participants' short write-up based on gathering participants' responses to this question, and their knowledge (IRR = 0.93 Cohen's Kappa).

To understand quantitative differences in search, sensemaking and learning behaviors across the Interface conditions (InterWeave vs Baseline) and topics (Environmental Impacts of COVID-19 and Future of Space Travel), we performed two-way ANOVA tests, followed by post-hoc two-way Tukey's HSD pairwise test in case of significance ($p < 0.05$).

Self-Reported Perceived Value of Suggestions'

To understand the perceived value of the presentation of query suggestions within or outside the sensemaking structures, in the post-task survey questions, we asked participants' to rate their level of agreement to the statements about their perceptions

of the suggestions' content, placement, and their interpretation of how the suggestions were generated (all statements in section 5.3). Here, participants rated their level of agreement with each of these statements on a scale of 1-5 where 1=strongly disagree and 5=strongly agree. We also thematically analyzed the transcripts of their post-task reflective think-aloud interviews. Here two researchers identified themes based on an open coding session of the transcripts in a grounded theory manner to develop a coding schema. Then, the two researchers coded all the transcripts closely on the coding schema. There was an inter-rater reliability of 0.85 Cohen's Kappa between the two raters.

5.5 Results

During the task of searching and taking notes to explore and synthesize knowledge on their assigned topic, participants, on average, issued 16.3 queries, 10.1 suggestions and typed 9.3 queries, per session. They visited 13.6 websites, gathered 280.9 words into their notes, on average. Figures 5.6 and 5.7 show a few example sensemaking workspaces of InterWeave and Baseline participants, respectively. When comparing the responses to pre- and post-questionnaires, participants on average reported an increase in their topic knowledge, learning 5.6 new domain-specific terms/concepts on average.

We found no statistically significant differences between the task topics and no significant interaction effects between topics and interface condition used across all search, information gathering, sensemaking and learning measures. In this section, we report the findings of the study, beginning with how the presentation of search suggestions within vs outside the sensemaking context affects search, and then respectively how it impacted information gathering, sensemaking and learning behavior.

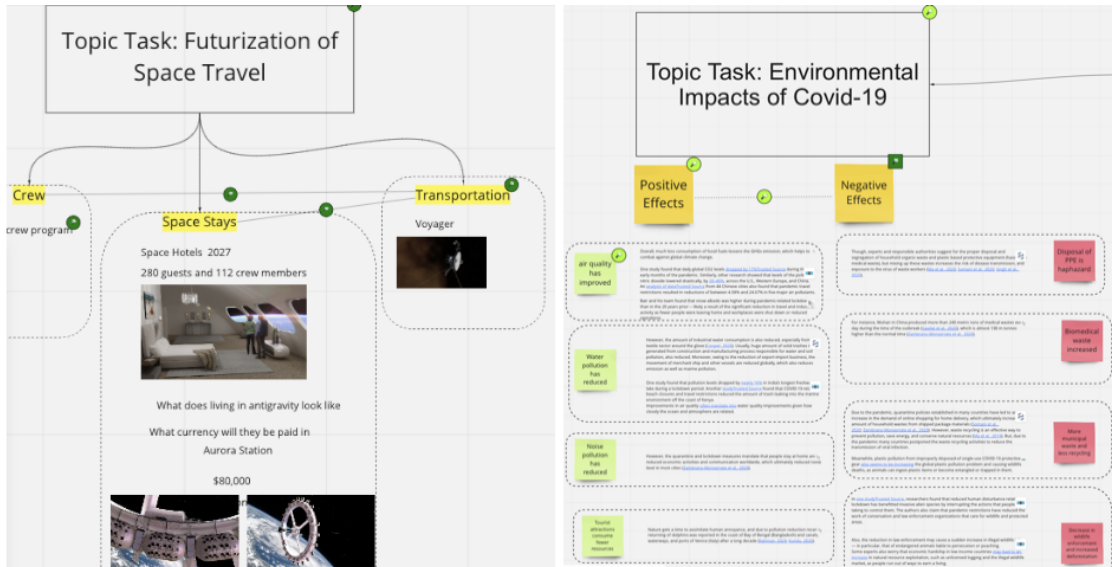


Figure 5.6: Examples of notes taken by InterWeave participants. Note the suggestions embedded within the participants' evolving sensemaking structure as green icons.

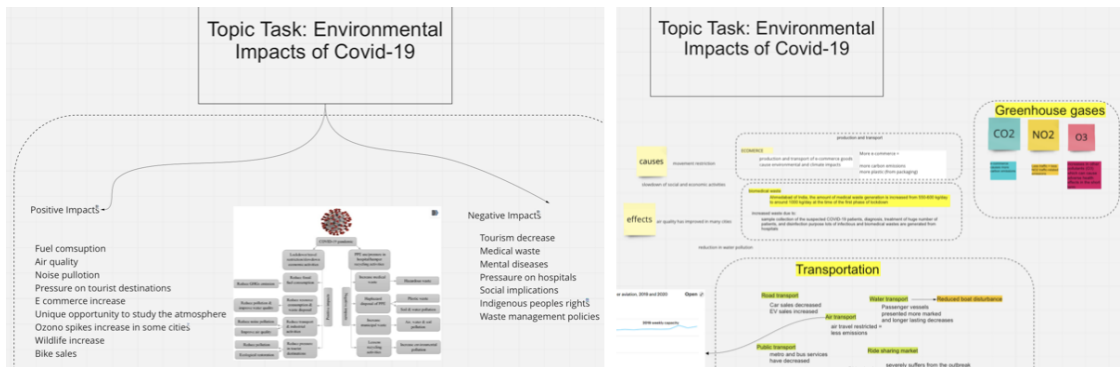


Figure 5.7: Examples of notes taken by Baseline participants

5.5.1 InterWeave encourages active searching

InterWeave participants averaged 22.5 queries each, while Baseline participants averaged significantly fewer queries at 14.8 queries ($F_{33}=1.79, p=0.04^*$). Of these queries issues, InterWeave participants issued 12.5 suggestions on average whereas Baseline participants issued significantly fewer suggestions i.e. 6.9 ($F_{33}=2.65, p=0.01^*$). However, there was no significant difference in the number of queries typed out across Baseline and InterWeave participants ($F_{33}=0.55, p=0.29$) (Figure 5.8).

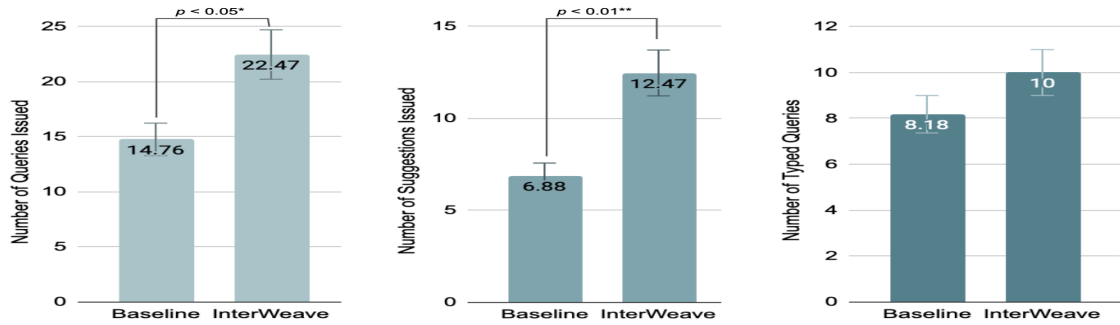


Figure 5.8: InterWeave participants issued significantly more queries, particularly the suggestions compared to Baseline participants. However, they typed similar number of queries.

To observe if there were differences across the type of query suggestion used in the InterWeave condition, we conducted a chi-square test (χ^2) between the types of query suggestions. Participants issued notes-level the most ($M = 3.2, SD = 3.66$), then cluster-level suggestions ($M = 1.8, SD = 0.21$), and then cross-cluster level ($M = 1.3, SD = 1.45$) and lastly title-Level ($M = 1.0, SD = 1.50$). Participants issued significantly more note-level suggestions and cluster-level suggestions than the other types ($\chi^2(1,33) = 1.42, p = 0.03^*$).

5.5.2 InterWeave assists sensemaking

There is no significant difference across the number of webpages opened per query issued across InterWeave and Baseline participants ($F_{33} = -1.39, p = 0.09$). However, InterWeave participants gathered nearly double the information per query issued ($M = 405.5, SD = 388.63$ words) compared to Baseline participants ($M = 219.4, SD = 183.50$ words, $F_{33} = 1.79, p = 0.04^*$) (Figure 5.9). This implies that participants got more information out of visiting similar number of websites.

InterWeave participants exhibited significantly broader sensemaking ($M = 13.2, SD = 7.49$ connections) than Baseline participants ($M = 8.5, SD = 7.70$ connections, $F_{33} = 1.80, p = 0.04^*$) as they created more connections across gathered information (including cluster

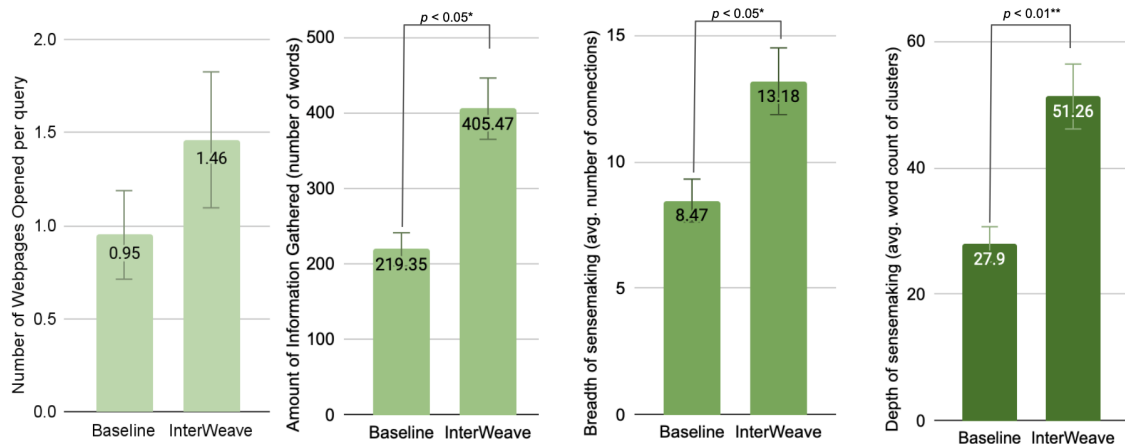


Figure 5.9: InterWeave participants gathered significantly more information and exhibited broader and deeper sensemaking in their sensemaking workspace, while visiting similar number of websites, compared to Baseline participants

titles, cluster groups, connection lines). Similarly, InterWeave participants also tended to develop deeper sense by writing more within each cluster ($M=51.3$, $SD = 42.08$ avg. words per cluster) compared to the Baseline participants ($M=27.9$, $SD = 23.90$ avg. words per cluster, $F_{33}=2.33$, $p=0.01^*$) (Figure 5.9).

5.5.3 InterWeave enhances knowledge gain

InterWeave participants reported a significantly greater increase in knowledge ($M=1.88$, $SD=0.83$) compared to Baseline participants ($M=1.1$, $SD = 0.81$, $F_{33}=2.23$, $p=0.03^*$).

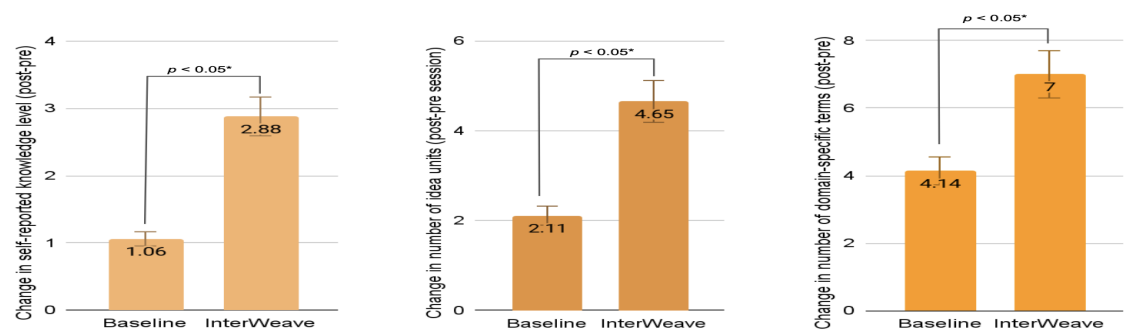


Figure 5.10: InterWeave participants reported a significantly greater increase in knowledge, discovered more domain-specific terms, and idea units compared to Baseline participants.

When analyzing their answers to their topic knowledge pre and post-task, we found that InterWeave participants discovered significantly more domain-specific terms ($M=7.0$, $SD=4.78$), compared to Baseline participants ($M=4.1$, $SD=3.06$, $F_{33}=2.45$, $p=0.02^*$). Similarly, they also discovered significantly more idea units ($M=4.7$, $SD=1.55$), compared to Baseline participants ($M=2.1$, $SD=1.35$, $F_{33}=2.02$, $p=0.02^*$) (Figure 5.10).

5.5.4 Participants preferred InterWeave's in context presentation of suggestions

To understand how searchers perceive the value of query suggestions, we asked participants to rate their level of agreement to the statements about their perceptions of the suggestions' placement, their interpretation of how the suggestions were generated, and the content of the suggestions (on a scale of 1-5 where 1=*strongly disagree* and 5=*strongly agree*, in the graphs lighter colors indicates more agreement) in the post-task survey. To check if there were any statistically significant differences between participants' perceived value of Baseline and InterWeave suggestions, we ran Friedman tests, along with post hoc analysis using a Bonferroni correction applied on their ratings for each statement.

Placement of suggestions

InterWeave participants agreed significantly more to the statements about the presentation of query suggestions being helpful compared to Baseline participants: "*Suggestions were positioned in a manner that was easily discoverable*", "*Placement of suggestions helped me connect new information to gathered information*" and "*Placement of suggestions helped me discover information faster*" (Figure 5.11). In the retrospective think-aloud, InterWeave participants P15 said, "*I liked that the suggestions were right*

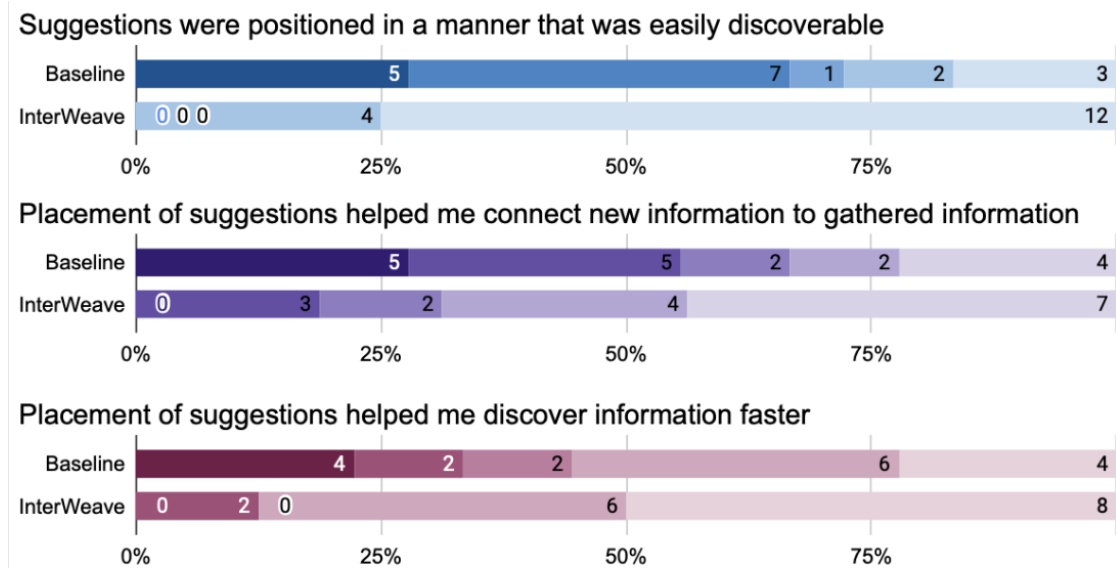


Figure 5.11: InterWeave participants agreed significantly more to the statements about the presentation of query suggestions being helpful compared to Baseline participants

next to the components that they were building on. That made it clear what the suggestions were relating to.” Similarly, another InterWeave participant P24 said, “I was easily able to see the connections between my notes and what I searched for.”

Meanwhile, many participants in the Baseline condition (nine out of 17) believed that suggestions could have been more helpful. Out of these nine, five participants attributed this dissatisfaction to the placement of the suggestions. Specifically, they thought that it was difficult to see how suggestions relate to the notes taken on the board. Baseline participants said: “I wouldn’t say that the suggestions were very discoverable... Also the fact that it is presented as a list makes it less interesting in terms of connections... it was not easy to directly transfer them in my mindmap.” (P20) ; “I think it would be nice to see how certain queries were connected to what I already had on the Miro board, since there were times where I wondered whether any of the queries were relevant to what I’m looking at. Like The Wolf Amendment was suggested to me, but I wasn’t sure what it related to... I thought it was a cool amendment related to wolves or something, definitely not space related” (P4) Therefore, the presentation of suggestions in vs out of

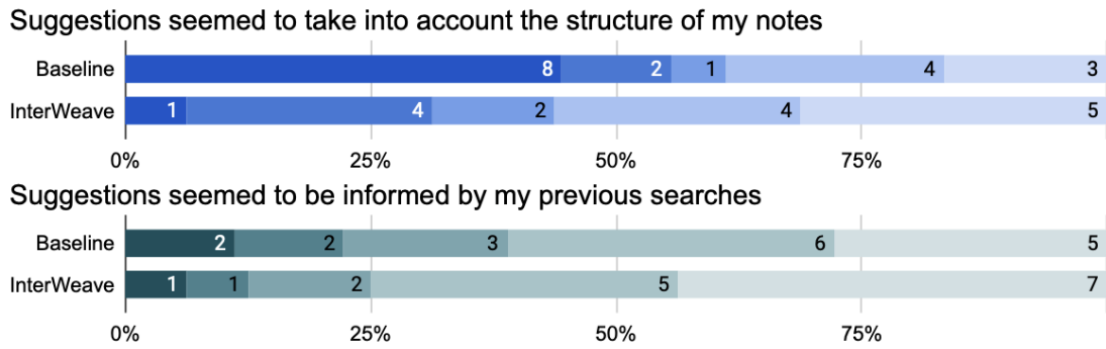


Figure 5.12: InterWeave participants felt they had better transparency around how the suggestions were being generated.

context affected the participants' perceptions and value of the suggestions.

Interpretation of suggestions

When asked about how they thought the suggestions were generated, InterWeave participants seemed to have better transparency around how the suggestions were being generated (Figure 5.12). They agreed significantly more to the statements: *"Suggestions seemed to take into account the structure of my notes"* and *"Suggestions seemed to be informed by my previous searches"*. This implies that they were able to glean the context of the suggestions and what data was being used to generate these suggestions based on their interactions with the suggestions in the sensemaking workspace. In the post-task retrospective think-aloud, InterWeave participant P24 said, *"It was really helpful and grounded the suggestions in my notes. So I was easily able to see the connections between my notes and what I searched for."* Similarly, 12 out of the 17 InterWeave participants mentioned found the suggestions helpful and reasoned that the query suggestions were relevant to their search and sense-making process.

Content of suggestions

When asked about their perceived values of the suggestions, InterWeave participants agreed significantly more to the statements *"Suggestions helped me ...": "reflect on what I had learnt so far", "organize and structure my notes better", and "discover new connections across gathered information"* (Figure 5.13). There was no significant difference across the interface conditions for the statements *"Suggestions helped me...": "better articulate my information goals", "ask new questions"*. Lastly, InterWeave participants disagreed significantly more to the statement: *"Suggestions helped me narrow my search to retrieve the right quantity of information"*. Generally, when we asked participants why they used the query suggestions, the common answer was that it helped open up new routes of research and expanded the topic domain. As InterWeave participant P20 suggested, they often used the query suggestions when they *"get stuck in [their] flow or to search for branches for my clusters."* InterWeave participant P17 also *"thought [the query suggestions] were very useful in expediting the creation of new clusters and also connecting them."* Other than providing new perspectives and insight into the topic, two participants specifically mentioned that InterWeave provided unique queries that the popular search engine did not. *"Very helpful in showing me different avenues to explore and were different from the google related searches I usually search."* (P30) ; *"They suggested topics that Google did not suggest."* (P7) These comments underscore the appeal and potential benefits of uniquely tailored search suggestions that popular search engines are not currently sufficiently implementing.

Baseline participants raised several pain points concerning the query suggestions. There were many instances in which participants felt that they were too distracting or overwhelming. Some thought the suggestions were *"way too detailed and I did not want to get that deep"* (P6). Others found the suggestions distracting and irrelevant. For example, P18 mentioned how they *"distracted [their] thought process because then*

[they] tried to reason how these suggestions came to be and what connections they had to the topic at hand.”

On the other hand, although InterWeave participants thought the query suggestions provided were semantically related to a part of the user’s sensemaking structure, they were not always aligned with their thought process which ultimately hindered their workflow. P28 talks about about the suggestions *”were really useful in directing me to explore different parts of this larger more abstract research topic. . . It was really useful to see that they appended parts of my notes to clarify the query suggestions. Sometimes this was not so helpful because the terms appended were not relevant to what I was doing then, but it might be useful as I explore further so I want to bookmark or save these for later.”* This indicates that not only do suggestions need to be presented in context, they also need to be presented in a timely manner that aligns with the searcher’s train of thought and workflow.

5.5.5 Wizard’s insights on automating the process of inferring context and placing suggestions

Our goal was to evaluate an interaction approach and explore where to best present suggestions with respect to the user’s sensemaking and work. To understand this aspect, we employed the wizard-of-oz prototyping technique [128] to develop and evaluate the InterWeave interaction techniques. We gained many insights about not only the effects of presenting suggestions in this manner, but also about what it would entail to develop such a context-aware system. Based on discussions with the human wizard who placed the suggestions within the user’s evolving sensemaking structures, we learned that the main challenges were:

(1) **Timeliness of suggestions:** The wizard reported that it was at times challenging to prioritize when to provide which suggestions at a particular location. They said *”at*

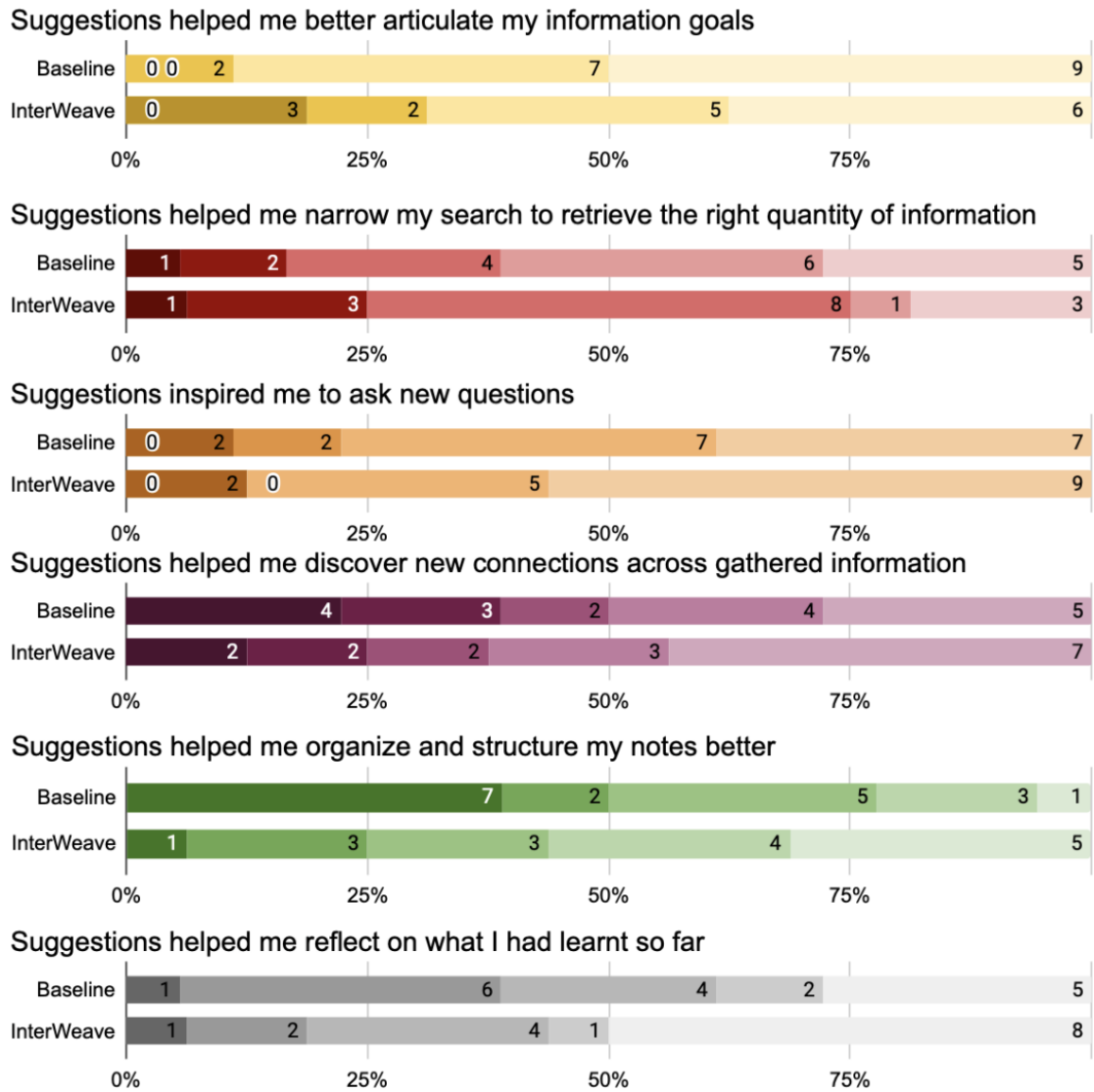


Figure 5.13: Searchers' level of agreement to these statements on a scale of 2 (Strongly Agree) to -2 (Strongly Disagree) for Baseline and InterWeave suggestions. Lighter colors indicate higher level of agreement.

times it was difficult to be on the same wavelength with the user". While proactively placing suggestions at a location can be beneficial to the user, the challenge is providing assistance without being too disruptive to the user's workflow. To maintain experimental control, the wizard placed the three suggestions across the board after every major edit or query issued. However, in a future automated system that builds on this work, the system might only show suggestions where and when a user requests it, allowing them to moderate when they request help and how it affects with their workflow.

(2) **Cross-cluster query suggestions:** To provide useful cross-cluster suggestions, an automated system must effectively model the topic space of each cluster of information [202] and the topic overall. The wizard discussed how these suggestions required extensive research, preparation, and abstract-level thinking and therefore, hypothesized that for an automated system, this task might be difficult because it hinges on high-level decision-making.

(3) **Assessing usefulness of suggestion:** The wizard wondered if the users were able to understand why a suggestion had been provided at a particular location. They worried that *"a seemingly irrelevant query suggestion may disincentivize participants to initiate the search."* To help assess relevance and usefulness of suggestions and further integrate the search and sensemaking environments this future system could allow users to preview the search results of a suggestion or highlight relevant website clippings from issuing the suggestion (like [334, 458, 178]).

5.6 Discussion

Complex, exploratory information work can be slow, tedious and cognitively demanding. It can be hard to articulate ill-defined information goals into specific queries, synthesize new information with prior knowledge, and select optimal exploration strate-

gies as people might be unaware of better alternatives. Our work in this paper seeks to reduce the cognitive load through an intelligent system that symbiotically guides a user towards fulfilling information goals during exploratory search and sensemaking. This paper presents a novel approach, InterWeave, which infers a user's information goals from the structure of notes taken and presents query recommendations weaved into the context of their emergent sensemaking.

5.6.1 How can in context placement of search suggestions affect exploration and learning?

Our analysis finds that InterWeave participants issued more search queries, particularly using the suggestions provided compared to baseline participants (Figure 5.8). When asked about their perceived value of these suggestions, InterWeave participants agreed significantly more to the statements "*Suggestions helped me ...*": "*reflect on what I had learnt so far*", "*organize and structure my notes better*", and "*discover new connections across gathered information*"; and disagreed significantly more "*Suggestions helped me narrow my search to retrieve the right quantity of information*" (Figure 5.13). Generally, when we asked participants why they used the query suggestions, the common answer was that it helped open up new routes of research and expanded the topic domain. InterWeave provided unique queries that popular search engines usually did not. This highlights the potential synergy in which an intelligent system, such as InterWeave, can help enhance and speed up the user's search and sensemaking process.

InterWeave participants issued more suggestions offered at the individual notes-level and the cluster-level than the cross-cluster or topic-level suggestions. This might suggest some level of a *Goldilocks effect* where people pay attention to suggestions that are neither too broad and nor too deep. The notes-level and cluster-level suggestions might broaden their exploration just enough, while still keeping the exploration focused. This

preference for semantically- and structurally- near suggestions is similar to a phenomenon studied in creativity research: people are more likely to hit an impasse when presented with semantically far ideas during brainstorming [82, 81, 80]. As such, presenting query suggestions at the title level may need more context than those presented at the cluster and notes level. "Far" recommendations need more context and informational cues to understand how they relate. Since this type of suggestion deliberately goes beyond the informational structures currently present in a user's notes, it might be less essential for these suggestions to be placed directly in the notes. It is worthwhile to investigate ways to make the connections between the queries and notes more concrete and clear at the title level.

Although InterWeave participants thought the query suggestions provided were semantically related to a part of their sensemaking structure, the guidance was not always aligned with their thought process which some participants found distracting. This concern was highlighted not only by the participants, but also by the wizard. Therefore, future work must build on this contextual presentation of search suggestions to also perhaps match the timeliness in which the query suggestions are presented at any particular location of work.

In terms of sensemaking behavior, InterWeave participants gathered significantly more information in their sensemaking workspace, and demonstrated broader and deeper sensemaking, even with no significant difference in the number of webpages visited, compared to baseline participants (Figure 5.9). This implies that presenting the suggestions within the evolving sensemaking structure, helps glean more information from a similar number of webpages. InterWeave participants might have read more of the websites they opened, because they were primed to how the suggestion that opened the website and thus the information on the website was directly connected to their notes. This might be affected by the availability heuristic, which is a mental shortcut where people often

form connections, here of usefulness, between things that co-occur or seen in the same place together [411, 273, 86]. Previous work has explored the role of query suggestions in creating information scent (i.e. the proximal cues from which searchers perceive the value of distal information sources) [327, 186, 219, 222]. As InterWeave suggestions present the user with gaps in their knowledge directly next to the parts of what they already know, it is creating a more contextualized trail of information which in turn helps with assessing usefulness and relevance of suggestions and information found on SERPs and websites.

Correspondingly, InterWeave participants also reported a significantly greater gain in knowledge, discovered more domain-specific terms and idea units compared to baseline participants (Figure 5.10). The enhanced sensemaking and knowledge gain seen in InterWeave participants might be related to schema theory which states that explicitly linking new information to the knowledge and schema that learners already possess can help learners integrate the new information into their schema [326, 328].

When talking about the perceived values and challenges around the presentation of suggestions, participants mentioned that they preferred InterWeave's in context presentation of suggestions compared to the Baseline's in terms of its content, placement (Figure 5.11) and their interpretation of why the suggestion was being provided. Particularly, InterWeave participants seemed to have better transparency around how the suggestions were being generated (Figure 5.12). As many machine learning papers in the contemporary zeitgeist have shown – the explainability and transparency of recommender systems and algorithms is critical [311, 388]. Presenting suggestions within the context of the where the suggestion might be used might help users demystify what signals recommender system algorithms take in as input, and how they might be being processed to provide recommendations.

5.6.2 Limitations and Future Work

As we primarily wanted to study the interaction mechanism of where do users see query suggestions – in or out of their work context – we decided to prototype InterWeave and Baseline conditions using a wizard-of-oz prototyping technique that leveraged natural language processing algorithms to provide real-time, suggestions positioned with respect to the users’ knowledge and work structures. As there are many individual differences across how people make sense and work on complex, exploratory information goals, this prototyping technique enabled us to quickly test and gain insights about this interaction mechanism without committing to extensive coding and development. However, the wizard-of-oz prototyping approach limits the replicability of this system because it depends on the wizard’s knowledge on a topic. The wizard in our study spent six weeks researching a topic to gain enough topic expertise to know whether two terms, concepts or subtopics were conceptually related or not. To help with reproducibility, we have linked the sheets they generated to outline their topic knowledge as part of the supplementary materials linked here: ². Based on the findings and participant feedback we have summarized in this paper, future work can translate the InterWeave wizard-of-oz algorithm based on searcher’s actions, and our operational definition of *conceptual similarity* into a completely automated process for providing query suggestions. Here, conceptual similarity can be calculated based on wizard’s heuristics for placing query suggestions using new state-of-the art complex language models such as Bidirectional Encoder Representations from Transformers (BERT) [120]), and general-purpose ontologies like ConceptNet [391] or even leveraging the structure of websites like Wikipedia (<https://en.wikipedia.org>).

The current prototype is a Chrome browser extension and Miro plugin. However, people take notes and make sense of information across a variety of tools and applications.

²<https://tinyurl.com/InterWeaveUIST22>

Now that we have shown the benefits of presenting query suggestions within work context, we leave it to future work to integrate these suggestions across various different note-taking, sensemaking and information work platforms (e.g. Word documents, Google Docs, emails, etc.).

Self-reported measures of learning are common in the CHIIR and search as learning community, however, self-report data may have gaps or inconsistencies with actual observed behavior and might be affected by cognitive biases such as the Dunning-Kruger effect [130] where people with limited knowledge or competence in a given intellectual topic greatly overestimate their own knowledge or competence in that topic relative to objective criteria or to the performance of their peers or of people in general. To mitigate the impact of this measure, we also measured learning by asking participants to recall terms, concepts and facts, and write a summary of what they knew about the topic before and after the search task. However, written summary measure can be affected by memory biases, and co-variables such as the summary length [444]. To control for these factors, we asked participants to write no more than 500 words, and to write the summary immediately after their search session and they could consult their notes taken in their sensemaking workspace.

Another limitation of the controlled lab study was that we controlled the time of exploratory search and sensemaking to only 45 minutes. However, complex, exploratory information work often span multiple sessions over multiple days [439, 276]. This controlled timed experiment might have affected the searcher's normal searching, sensemaking and learning behavior [229]. It is important to understand users search and sensemaking practices in the wild and study how presenting suggestions in vs out of context affects search, sensemaking and learning behaviors over the longer, natural course of users' information workflows. We intend to make all the code from this project open-source and accessible so that future work can conduct longitudinal studies in the

wild.

The current prototype pushes suggestions proactively to all locations across the board. While proactively presenting suggestions can be beneficial, participants also reported being distracted from their train of thought at times [423, 448]. To prevent this InterWeave not only needs to be aware of the content and structure of the users' notes, but also where they are in their overall information foraging and sensemaking workflow. Future work could use additional signals to better time offering query suggestions during complex, exploratory information work.

Modern knowledge work is often collaborative, and while collaboration has its benefits, effectively coordinating work in a team can be challenging. Collaborators must spend time dividing and assigning search goals and tasks, locating, sharing, and synthesizing information to create a shared mental model [370, 75]. Challenges may include repeated work done across collaborators, and confusions about process and results [114, 370, 74]. InterWeave presents an interesting first step in alleviating some of these challenges for individual information workers. This highlights an interesting opportunity to build tools to promote collaborative knowledge discovery and reducing sensemaking coordination costs by recommending queries based on each collaborator's prior experience, searches, contribution to a shared document in future work.

5.7 Conclusion

In this paper, we present a novel interaction mechanism, InterWeave, that leverages patterns and gaps in a searcher's sensemaking structures to present query recommendations weaved into their evolving work context. To evaluate how this interaction mechanism affects users' search, sensemaking and learning activities, we implemented this system as a web browser extension using NLP algorithms and wizard-of-oz tech-

niques. A between-subjects user study ($n=34$) found that InterWeave’s approach not only promoted active querying, more information gathering, broader and deeper sensemaking and discovery of domain-specific terms and concepts but also helped participants keep track of suggestions and connect newly discovered information to existing knowledge when compared to presenting suggestions as a list separated from the sensemaking context. As the information work becomes increasingly complex, the ability to ask questions and explore easily and naturally is becoming especially important. This work brings us one step closer to the vision of leveraging people’s natural information-searching and sensemaking activities as relevant contexts for scaffolding knowledge discovery and online learning.

5.8 Acknowledgements

This chapter in part, includes portions of material as it appears in *InterWeave: Embedding Query Suggestions within Searcher’s Sense-making Structures Promotes Active Searching and Knowledge Discovery* by Srishti Palani, Yingyi Zhou, Sheldon Zhu, Steven P. Dow in Proceedings of the 2022 ACM Symposium on User Interface Software and Technology (UIST’22). The dissertation author was the primary investigator and author of this material.

Chapter 6

Relatedly: Scaffolding Literature

Reviews With Existing Related Work

Sections

Scholars who want to research a scientific topic must take time to read, extract meaning, and identify connections across many papers. As scientific literature grows, this becomes increasingly challenging. Meanwhile, authors summarize prior research in papers' related work sections, though this is scoped to support a single paper. A formative study found that while reading multiple related work paragraphs helps overview a topic, it is hard to navigate overlapping and diverging references and research foci. In this work, we design a system, *Relatedly*, that scaffolds exploring and reading multiple related work paragraphs on a topic, with features including dynamic re-ranking and highlighting to spotlight unexplored dissimilar information, auto-generated descriptive paragraph headings, and low-lighting of redundant information. From a within-subjects user study ($n=15$), we found that scholars generate more coherent, insightful, and comprehensive topic outlines using *Relatedly* compared to a baseline paper list.

6.1 Introduction

Scientific discovery and innovation rely upon scholars to have a rich understanding of prior work, which they achieve through reviewing the literature, extracting meaning, and identifying connections across many papers with large amounts of ambiguous domain-specific information [234, 460]. This process is getting progressively harder with the exponential growth of scientific publications [141, 51, 210, 421] and the increasingly interdisciplinary nature of science [422, 309]. Unfortunately, current approaches such as reading survey papers or using textual or visual search engines are limited in terms of sensemaking support or timeliness. For example, survey papers present a broad overview of a research topic with coherent research themes and carefully synthesized descriptions [59, 234]. But since they require significant manual effort to compile, survey papers are not always available on all topics and can quickly become outdated as new research

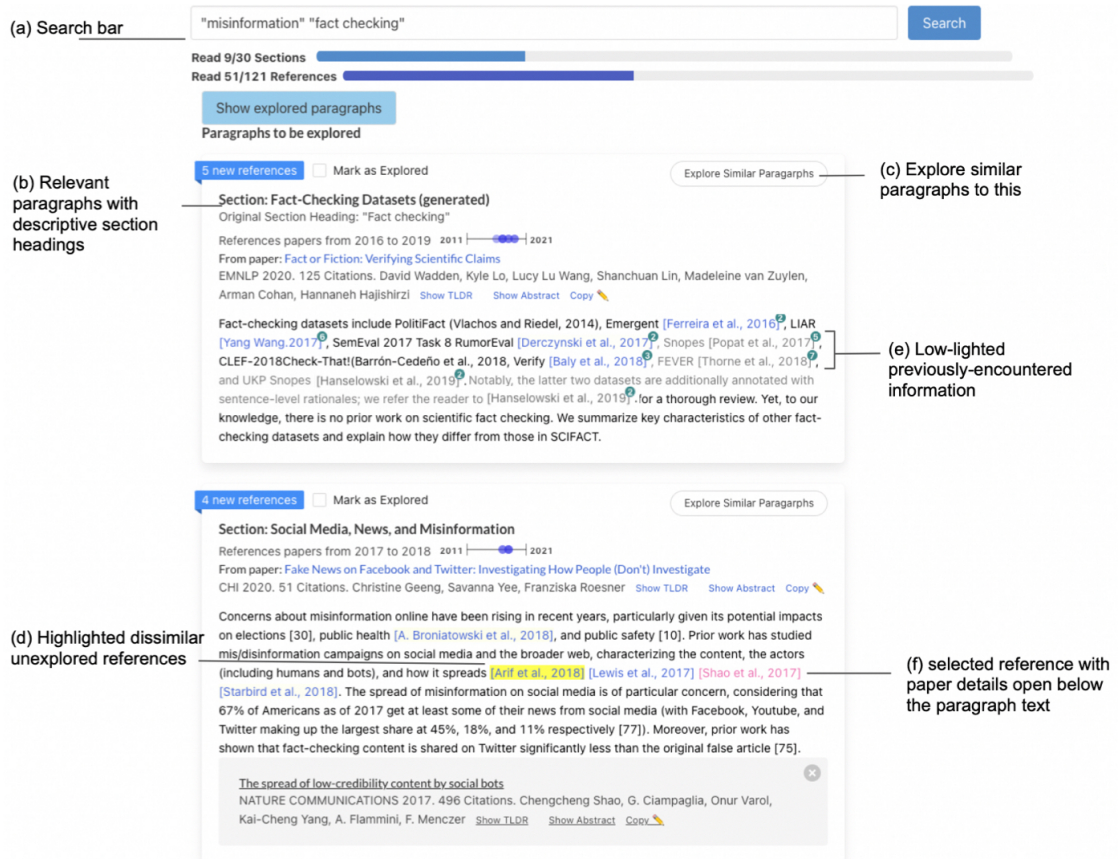


Figure 6.1: The *Relatedly* system presents users with related work paragraphs from prior work on a topic and scaffolds the paragraph exploration experience with features for reading, prioritization, and progress tracking. Here, the **Overview View** shows paragraphs relevant to the high-level query topic and ranked by diversity, so the top results show a wide range of subtopics. A user interested in learning more about one of the paragraph’s subtopics could click on the “Explore Similar Paragraphs” button, which would take them to the **Similar Paragraphs View** in Fig 6.2.

emerges. To address this, scholars also frequently rely on automatic approaches to help explore literature such as scholarly search engines, including Google Scholar¹ and Semantic Scholar². These tools can be effective in looking up papers relevant to a query but do not present higher level themes that connect multiple papers. Other tools use visualization to connect and cluster papers [230] using metrics based on citations or semantic embedding vectors, such as Connected Papers.³ However, it can be hard for users to comprehend the underlying meaning of complex graphs and clusters as automatic clusters often conflate multiple dimensions [179]. As a result, when timely survey papers are not available, scholars still need to examine many individual papers and try to figure out the latent themes and connections between them to conduct literature reviews [328].

Meanwhile, authors of scholarly papers also go through a similar process of exploring and summarizing prior research whenever they need to write the related work sections of their papers. While a related work section provides up-to-date and well-synthesized summaries of prior work [404, 407], because they are scoped to support a single paper they often do not provide a comprehensive overview of the topic like survey papers. However, this issue could potentially be mitigated if readers are presented with *multiple* related work sections about a topic from different papers so that they gain broader coverage and different perspectives of the space from multiple authors.

To investigate this opportunity further, we first built a text search engine over a set of related work sections extracted from many papers, and used it to conduct a formative interview study with 10 scholars. We asked scholars about their reactions to and challenges with exploring related work sections when compared to their current practices. We found that while participants preferred reading related work sections, they had difficulty prioritizing and tracking their reading, given that different related work

¹scholar.google.com

²www.semanticscholar.org

³www.connectedpapers.com

sections have both overlapping and diverging references and foci.

Motivated by insights from the interviews, we designed *Relatedly*, a novel system for scaffolded exploration of literature that leverages related work sections to provide a synthesis of a broad topic. As shown in Figure 6.1, when the user queries a topic in *Relatedly* (a), the system retrieves relevant paragraphs from different papers' related work sections along with their section headings (b) to help users gain a quick overview of disparate research threads. In cases where a paragraph does not have a descriptive section heading, *Relatedly* automatically generates one. Users can also drill-down on a subtopic by exploring similar paragraphs for a given paragraph (c, followed by Figure 6.2). To support users in prioritizing and tracking their reading, as a user is exploring related work sections in *Relatedly*, it tracks which paragraphs and references the user has read and then dynamically re-ranks the remaining paragraphs and highlights unexplored references that diverge from the user's history (d). *Relatedly* also low-lights sentences that refer to papers that have been cited in already-seen paragraphs (e).

We conducted a within-subjects study ($n = 15$) to evaluate *Relatedly* where participants were asked to explore literature on two scientific topics, with the ultimate goal of producing an outline of a survey paper on each topic using *Relatedly* in one condition and using a baseline system that returns a list of papers in another. We find that participants produced better quality outlines when using *Relatedly* versus in the baseline condition, as rated by topic experts who were blind to the conditions. System logs reveal that users of *Relatedly* interacted with significantly more information (both paragraphs and papers) than in the baseline condition, despite having the same amount of time for each condition and access to the same set of papers. Participants also self-reported that they preferred to explore related work sections using *Relatedly* rather than explore a list of papers to conduct literature review.

In summary, this work makes the following contributions:

- A novel approach to discovering and systematically reviewing literature on a scientific topic by reading and exploring relevant related work sections extracted from many papers.
- Results from a formative user study ($n = 10$) outlining current literature review practices and user challenges with this approach.
- The *Relatedly* system, which scaffolds related work paragraph exploration with reading, prioritization, and progress tracking features.
- Empirical insights from a within-subjects study with 15 participants that finds that scaffolded exploration of related work sections promotes literature discovery and synthesis.

6.2 Related Work

Our work builds on prior work studying how scientists explore and review literature, and tools built to support these complex exploratory processes.

6.2.1 How Scholars Conduct Literature Reviews

Literature review helps scholars identify patterns and gaps in prior research in order to find opportunities, determine rationale for a new investigation, and situate research goals within the literature [404]. Reviews detail both known research and open research questions in this topic. A high quality literature review comprehensively includes all the main themes and sub-themes found in a chosen topic of study, from both classic foundational work and recent studies to demonstrate an in-depth understanding of the topic at hand [115, 234, 207]. To achieve these goals, scholars must take time to comprehensively explore a topic and read many individual papers. However, the sensemaking

process of trying to get an overview of a field from reading individual papers can be time-consuming and cognitively overwhelming [59, 407, 354]. For example, it can be hard for users to diversify their readings to quickly identify different threads of research. The overwhelming number of individual papers and redundant information scholars need to go through often leads to information overload [292].

One way scholars have addressed this is to write survey papers for different research topics [115, 234, 207]. Yet with the exponential increase in publishing rates, survey papers are often unavailable [141, 51, 210, 421], and even when they are, they quickly get outdated as newer research emerges.

Meanwhile, in most scientific papers, authors summarize and draw connections across multiple papers to situate their own work in related work sections [407, 404]. Each paragraph in these related work sections adds context and structure to individual papers referenced. For example, the related work section of a paper on misinformation might group a set of referenced papers into a paragraph with a title of “How misinformation affects public health”, and another set of papers might be grouped under “How misinformation spreads on social media”. However, related work sections only focus on a paper’s specific point of view and do not attempt to exhaustively overview all the themes and sub-themes in the broader topic. For example, the above paper about misinformation might focus its related work section on health misinformation on social media because that is what is relevant but lack coverage of other work related to misinformation, such as, say, computational techniques for detecting misinformation. Therefore, scholars hoping to gain a broader picture of literature on a topic would likely need to read multiple related work sections across multiple papers. This task is what the Relatedly system is attempting to scaffold.

Information foraging theory [327] provides some pointers on how to go about this task. During complex exploratory tasks, people switch between *exploring* different information

patches and *exploiting* a discovered patch to optimize information gain. They rely on various cues, or “information scent”, in the information environment to assess whether a source is promising for gaining information. We take inspiration from information foraging theory to provide information scent cues in Relatedly such as displaying how much new information the user can learn about by reading each paragraph. Also, to support switching from exploring to exploiting, Relatedly allows a user to dive in to view similar paragraphs given a paragraph; this enables them to gain a deeper understanding of a sub-topic from different perspectives.

6.2.2 Tools for Supporting Literature Review

One of the most common tools scholars rely on today for literature review is scholarly search engines [407], such as Google Scholar¹ and Semantic Scholar². These can be very effective in helping users look up individual papers relevant to a query. However, to gain deeper understanding of a research area, such as during literature reviews, scholars often need to synthesize information across individual papers. This effortful and time consuming process of making sense of connections between papers and uncovering the different nuanced research themes within a larger topic is largely left to the users with minimal support [276, 358]. For example, when exploring papers from a search results list, it can be hard for users to prioritize their readings, keep track of information scattered across multiple papers, or have a sense of their overall progress within the unfamiliar information space.

Faceted search interfaces allow users to navigate search results by applying multiple filters across categories [179]. Categorizing provides coherent and mostly complete labels. However, manual categorization takes time and effort and is hard to keep updated. Automatic categorization is typically based on metadata [179]. For example, Google Scholar supports filtering paper results by *time of publication* and *relevance*, among

others. Similarly, Semantic Scholar presents *'fields of study'*, *'publication types'*, etc., as facets by which papers can be filtered. But metadata is not always available. Also, these labels are often too general and don't provide meaningful insight into the topic or domain.

Visual clustering systems attempt an alternative approach to help scholars discover relationships between papers. For example, given a seed paper, Connected Papers³ utilizes the citation graph to find clusters of other relevant papers. Research systems like PaperQuest [331] and Apolo [88] visualize citation relationships between a set of papers as input, with support to overcome information overload by progressively revealing further related papers given a source paper and its citations. However, prior research in clustering search interfaces has also pointed to how automatically generated clusters can be incoherent and difficult for users to understand because they often conflate multiple dimensions [179]. Specifically, visual paper clustering approaches often show edges between similar papers but do not describe their semantic relationships [179]. They also show clusters of similar papers but lack high-level descriptions of the underlying themes [179]. As a result, scholar still need to examine individual papers to determine the meaning of each automatically generated clusters and how different papers relate to one another [88].

Automatic summarization techniques like Multi-Document Summarization [121] and Metro Maps of Science [371] add explanations to otherwise complex and hard to understand citation graphs. However, these explanations are not always accurate, coherent, or comprehensive. On the other hand, manual (e.g., Threddy [216]) and crowd-powered systems (e.g., Knowledge Accelerator [176, 85, 218], Crowdlines [272]) help provide more coherent, comprehensive, and accurate summaries of topic spaces, but take time and effort to generate.

Relatedly sidesteps the issue of generating high quality connections and clusters by

building on the significant effort that related work authors already expend to construct these for their paper. The main challenge then becomes about exploring multiple papers' overlapping clusters and differing perspectives on how papers connect to each other.

6.3 Formative Study & Design Goals

To understand user challenges and strategies when exploring and making sense of related work paragraphs on a topic, we conducted a formative interview study with scholars.

6.3.1 Formative User Study Method

We conducted semi-structured interviews with 10 people who have experience searching for, reading, and writing scientific literature for more than three years (5 male and 5 female, average age of 27.5 years). One had completed their doctoral degree, while five had completed a master's, and four had completed their bachelor's degree. In terms of job titles, we had: one post-doctoral researcher, one research assistant, one research scientist and the remaining seven were doctoral researchers. Five reported using scholarly web search multiple times a day, four reported doing this at least multiple times a week, and one said rarer than every week. Eight had experience conducting systematic literature reviews for three or more years, one reported doing this for two years, and one for one year. Participants came from diverse domains: neuroscience, geography, biomedical sciences, human-computer interaction, natural language processing, AR/VR design, and wearable computing. They reported mostly using scholarly search engines and paper lists for exploring literature, including Google Scholar, Semantic Scholar, and domain-specific conference proceedings, journals, and organizations (e.g., ACL for NLP or CHI'23 proceedings for HCI). They used a wide range of applications for reading and

writing scientific literature.

We asked participants about their workflows for conducting literature review and about any challenges they experience. Then, we gave them 20 minutes to explore and read a set of related work paragraphs extracted from multiple research papers on a topic. The paragraphs were displayed in a list on a simple text search engine interface. To contextualize their exploration, we gave them a simulated task [50] of conducting initial research to get a broad overview of the topic of “misinformation, fake news and fact-checking”, towards the ultimate goal of writing a literature review. To get insight into their user experience, participants were asked to think-aloud as they explored the list of paragraphs. Afterward, they were asked about their experience reading multiple paragraphs instead of papers, the challenges surrounding this, and strategies they used to overcome these challenges. We then presented them with alternative mock-up designs that augment the related work paragraphs with highlighting of references and terms that are unexplored and low-lighting of redundant citations.

Interviews were conducted remotely over video calls by the first author and lasted around 45 minutes. They were recorded and then transcribed using an auto-transcription service. Then, the first author went through the transcripts and coded them for themes using an open coding approach [87]. Through multiple iterations along with periodic discussions with the rest of the research team, we identified the user challenges and subsequently the design goals for our approach.

6.3.2 User Experience when Reviewing Literature

Current Literature Review Workflows and Challenges

When asked about their current workflows, all participants (10/10) mentioned using scholarly search engines to discover relevant papers on a topic. Some (5/10) mentioned

socially gathering a list of papers from collaborators, advisors, or social media such as Twitter. All of them mentioned reading papers one by one to extract meaning, 3/10 mentioned annotating the PDF documents with notes and highlights, and 2/10 mentioned saving these papers to a bibliography manager (e.g., Zotero, Mendeley).

When asked about challenges, 10/10 mentioned that it was hard to make connections across papers. 8/10 participants mentioned challenges with unknown unknowns ranging from not knowing the right search keywords that would lead to the right papers to not knowing what all the latent subtopics were within the topic of interests: *“I often don’t know which keywords/domain-specific language to search to get to the right literature”*, *“Even if people refer you to a shortlist of papers, it’s hard to get an overview of the topic and it feels like I might be being myopic and might have blind spots”*. 7/10 mentioned that it was hard to keep track of what they had read before: *“hard to keep track of many different research threads, points of view and see the bigger picture.”* 5/10 discussed challenges prioritizing what to read first : *“When I see so many papers in results, I get overwhelmed and open them up in tabs. But then I don’t know which to read first so they will just stay open in these tabs”*.

Preference for Exploring Related Work Sections and Challenges

When asked about their experience reading multiple paragraphs to get a topic overview, 9/10 participants preferred reading related work paragraphs to papers. Some positive reactions discussed getting a broad overview of the topic: *“Reading even a few paragraphs equips me quickly with the relevant vocabulary, references and takeaways from the topic”*, *“I’m able to see the different threads of research immediately”*. Others talked about the value of the text summarizing the referenced papers: *“I like the the additional explanation around the references, so I can understand the context and decide quickly whether I want to open it up to read more or not”*. Some also indirectly referenced

how they already use papers to help find other papers to read, and how extracting related work sections and their citations streamlines the process: *“Definitely more helpful than reading PDFs and doing the ritual of opening PDFs, reading introductions, and the paper, going back and forth between references in the bibliography and the paper to identify which papers might be useful”*.

However, there were also challenges with reading multiple paragraphs to get a topic overview. Some participants desired prioritization and navigation support to know what to read next: *“hard to prioritize which order to read these in”* (5/10), *“it is unclear what the similarities and differences between paragraphs are”* (7/10), *“want to know which are the most important or central papers summarized in this paragraph”* (3/10). Participants also wanted support for tracking their exploration: *“want to keep track of what [paper and paragraph] has been read vs not”* (2/10), *“hard to assess how much more there is to read on this topic”* (2/10).

When probed about how they would like to prioritize which paragraphs and papers to read, participants mentioned that they wanted to prioritize paragraphs with high coverage, sourced from papers that cover both recent and fundamental work, highly cited and ideally survey papers, and that minimally discuss the paper’s own work. Also, in terms of ranking, most (8/10) discussed how the first few paragraphs they read should map out the diversity of subtopics, and (6/10) said similar paragraphs should not be on the same page.

6.3.3 Design Goals

Motivated by the findings above, we list our core design goals:

- [D1] Support users in inferring higher-level meaningful organization of topics and dive deeper into subtopics
- [D2] Enable users to fluidly prioritize and explore similarities and differences between

related work paragraphs

[D3] Help users keep track of paragraphs and references they have explored

6.4 The Relatedly System

The screenshot displays the 'Relatedly System' interface. At the top, a search bar contains the query '"misinformation" "fact checking"' and a 'Search' button. Below the search bar, there are progress indicators: 'Read 9/30 Sections' and 'Read 52/121 References'. A 'Show explored paragraphs' button is visible. The main content area is divided into several sections:

- Selected Paragraph:** A blue box highlights a paragraph from the section 'Social Media, News, and Misinformation'. It includes a '4 new references' indicator and a 'See All Paragraphs' button. The paragraph text discusses concerns about misinformation online and its impact on elections, public health, and public safety, citing various research papers.
- Other Sections from the Same Paper:** A list of related sections is shown, including 'Investigative Strategies', 'Motivation', 'Understanding and Curation of Own Social Feeds', 'BACKGROUND AND RELATED WORK', 'Social Media, News, and Misinformation', and 'How People Interact With Misinformation'.
- 3 new references:** A section titled 'The Blurry Boundaries of Online Journalism' with a 'Mark as Explored' checkbox. It includes a timeline from 2015 to 2021 and a paragraph discussing the immediacy of social media and its impact on journalism.
- 2 new references:** A section titled 'The Dilemma of Fact-Checking' with a 'Mark as Explored' checkbox. It includes a timeline from 2013 to 2020 and a paragraph discussing the confidence of participants in fact-checking services.
- Mark as Explored:** A section titled 'Fact-checking Claims' with a 'Mark as Explored' checkbox. It includes a timeline from 2011 to 2018 and a paragraph discussing research on automatic fact-checking of claims and rumors.


Figure 6.2: To read more on the subtopic discussed in a specific paragraph in the **Overview View** (Fig. 6.1), this **Similar Paragraphs View** allows users to explore other paragraphs of that same subtopic that cited the same or similar references.

Guided by the insights from our formative study, we developed Relatedly, a novel approach to literature review that helps users achieve a broader and more insightful overview of a research topic. In this section we will describe the system through an example user scenario, a walk-through of the main features of the system, an explanation and evaluation of our automatic section heading generation pipeline, and the implementation details of the system as a whole.

6.4.1 Example User Scenario

Consider a junior computer science researcher interested in getting a broad understanding of a research area with which she is unfamiliar—*misinformation and fact-checking*. Not knowing what the important subtopics are, she starts by conducting a literature review using a common scholarly search engine and searches for the phrases: “misinformation”, “fact-checking”. However, even though all the papers in the search results look relevant, it is difficult for her to see the higher level themes and how individual papers relate to each other by only looking at the paper titles and search snippets.

Feeling overwhelmed, she switches to Relatedly with the same query, and the system returns a list of related work paragraphs relevant to the query in the *Overview View* (Fig.6.1). Wanting to get an overview, she skims through the section headings and quickly learns different subtopics, such as *Fact Checking Datasets*, *Social Media*, *News*, and *Misinformation*, and *Fake News Detection Techniques*. The section headings allow her to skim through the Overview View to get a sense of the different high-level research foci. As authors often structure their related works section into relevant subsections based on themes, these titles can help describe the gist of the paragraph’s focus.

As she becomes interested in the subtopic of *Social Media*, *News*, and *Misinformation*, she starts to read the related work paragraph that has it as a section heading. She clicks some of the references to see their metadata, including title, abstract, TLDR [68], authors, publication year, conference, and citation count, and she collects some of the ones she wants to read later by clicking “Copy” . Noticing the current paragraph was published four years ago, she clicks on the “*Explore Similar Paragraphs*” button. In the *Similar Paragraphs View* (Fig. 6.2), the system brings up other paragraphs from the search result that are also about *Social Media*, *News*, and *Misinformation*. As she skims the similar paragraphs, she reads about how other author summarized prior work about this particular subtopic across paragraphs (e.g., *The Blurry Boundaries of Online Journalism*,

The Dilemma of Fact Checking, and *Fact Checking Claims*) extracted from different source papers. She starts to understand connections between multiple referenced papers and concepts discussed in this subtopic. She starts to feel like she is getting a more holistic and well-rounded understanding of this subtopic. She continues reading other paragraphs including ones that were published more recently to comprehensively explore this subtopic.

She then returns to the Overview View to explore new subtopics. She notices that the paragraphs she is shown have changed as a result of her exploration thus far. Paragraphs have been dynamically re-ranked to prioritize ones with more unexplored and dissimilar references. She also notices that some paragraphs have sentences low-lighted that reference papers corresponding to ones she has already explored. Paragraphs also now sometimes have certain inline citations highlighted that point to unexplored references that are semantically different from those she explored before. She skips over some paragraphs with many sentences low-lighted and focuses on a paragraph with multiple highlights. Lastly, she checks the progress bar to keep track of what proportion of the entire set of related work paragraphs and references has she explored thus far.

6.4.2 System Features

We organize the description of the system features according to our three main design goals from the formative study.

[D1] Infer Topic Overview + Drill-Down to Subtopics

Given a search query, Relatedly presents a list of related work paragraphs relevant to the query in the *Overview View* (Figure 6.1). Each paragraph is representative of a different subtopic and is ranked to cover a broad range of subtopics. To drill-down to a subtopic covered by a paragraph (e.g., *Fact Checking Claims*), a user can click

on the “*Explore Similar Paragraphs*” button on the top right corner of this paragraph card. This brings up the *Similar Paragraphs View* with the similar paragraphs in the right column, and pins the selected paragraph to the left column (Figure 6.2). Below are specific features to enable topic overviews and drill-down to subtopics.

Diversity ranking of paragraphs: To help users see high-level organization in the Overview View, Relatedly first retrieves the most relevant paragraphs based on the standard BM25 [349] scoring,⁴ and then re-ranks the retrieved paragraphs using the Maximal Marginal Relevance (MMR) technique[76] to balance query-relevance with information-novelty in the top results. While the original MMR technique relied on text similarity to measure the information novelty given a document, here we use the number of unexplored references in each paragraph to approximate its information-novelty. The goal is to re-rank the paragraphs such that the top paragraphs jointly contain the most number of unique and unexplored references and present a wide range of diverse subtopics, while accounting for their relevance to the query term and number of references. This ranking was determined based on the participants’ responses to how they would like to rank paragraphs in the formative user study. The ranking score for paragraph at rank i is as follows:

$$MMR_i = \arg \max_i \{ BM25_i [\lambda |Refs_i| - (1 - \lambda) |((\cup_{j=1}^{i-1} Refs_j) \cup Refs_{exp}) \cap Refs_i|] \}$$

In the equation, $BM25_i$ is the relevance score based on the paragraph text and the query term, $Refs_i$ is the set of references in the paragraph ranked at position i , $Refs_{exp}$ is the set of references already explored by the user, and λ is a hyper-parameter for adjusting the penalty for containing references that already appeared in previously ranked paragraphs.⁵

Descriptive paragraph headings: Each paragraph in the Overview View comes with

⁴We showed the top 30 paragraphs by default to control for the length of the user study.

⁵Based on play-testing during development, we set λ to 0.3 to give a moderate advantage to paragraphs containing more references and a high penalty for containing references already covered by higher-ranked paragraphs to diversify topic coverage of the top results.

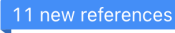
a descriptive title to describe the gist of the paragraph’s focus and serve as a subtopic for the user. As authors often structure their related works section into relevant subsections based on themes, for most paragraphs, these are extracted from the section headings of the related work sections. For paragraphs for which authors had only written generic or short section headings that contained less information (e.g. “Related Work” or “Fact-Checking”), or for paragraphs with no section headings, Relatedly generates descriptive titles using a BART-based [251] model (described in more detail in §6.4.3).

Similar paragraphs given a paragraph: In the Similar Paragraphs View, a list of similar paragraphs are shown for a selected paragraph. We determine paragraph similarity by whether they reference either the same papers as the selected paragraph (primary sort-order), or are semantically similar papers (secondary sort-order, using a threshold on the Euclidean distances between their SPECTER paper embeddings [99]).

Reading the paper behind a given paragraph: In the Similar Paragraphs View, underneath the selected paragraph, the user can access all the other sections from the same paper, including other portions of the related work section, in order to gain more context behind the paragraph.


[D2] Prioritize and explore similarities and differences across related work paragraphs

In the formative study, participants expressed that it was “hard to prioritize in which order to read the paragraphs”. Thus, Relatedly presents a number of features to support prioritizing and reading diverse unexplored information.

Unexplored references count badge: Both our formative interviews and prior work pointed to wanting to prioritize paragraphs that had the highest unread information first. To aid with this, all paragraphs have an unexplored references count badge (like ) that conveys the number of unique unexplored number of papers dis-


cussed in this paragraph. This number dynamically updates as the user interacts with more references across the paragraphs. The ranking algorithm prioritizes and ranks paragraphs with more unread references higher in the Overview View.

Highlighting of dissimilar unexplored references: To further facilitate prioritizing unexplored novel information and address the need to “*identify similarities and differences between papers*”, Relatedly highlights dissimilar unexplored references (like [Harrington.2012] [L. Holbert et al., 2007]). As the user clicks and reads references and paragraphs, some references get highlighted yellow indicating that these papers are semantically different to other papers interacted with so far (calculated using a threshold on the Euclidean distances between their SPECTER paper embeddings [99]). These references are highlighted on a yellow gradient, where the brighter the yellow, the highlighted paper is more different than the most similar papers interacted with so far.

Reference timeline visualization: Another heuristic that users wanted to use was to prioritize paragraphs that covered both recent and fundamental prior research on the topic. To help triage this, all paragraphs have a reference timeline visualization that visualizing the time-range of papers referenced in this paragraph (like 2011  2021). Here, each semi-transparent blue dot is a referenced paper publication year. As the min and max of the timeline signify the earliest and latest papers referenced across all paragraphs, users can use this to prioritize reading paragraphs that reference recent papers or more fundamental older papers. This feature was based on the formative study participants’ saying they wanted to triage reading priority based on the recency of papers referenced in the paragraphs.

Citation frequency badges: Participants in the formative interviews mentioned wanting to prioritize parts of a paragraph, particularly wanting to know which paper to prioritize when there are multiple papers cited for a claim or in a paragraph. Relatedly offers Citation frequency badges that aim to indicate how central or important a referenced

paper is to a topic. These green tags with a number refer to the number of times this paper has been referenced in these result paragraphs (like [Popat et al., 2017]⁴). If more than one of the paragraphs returned in the Overview View referenced it, it means that multiple authors discuss this paper, therefore it might be central or important to this topic.

Self-reference icons: Participants in the formative interviews mentioned wanting to de-prioritize parts of a paragraph where the authors were situating their own work in the background. To aid this Relatedly identifies which parts of the paragraph refer to the paper's work and signal this to the reader with an  icon.

[D3] Keeping Track Exploration Progress over Papers and Paragraphs

In addition to wanting to prioritize dissimilar unexplored information, users mentioned that it was challenging to track which papers or paragraphs have been read versus not. Relatedly provides a number of features to give users a sense of their progress in covering content while minimizing redundancy.

Low-lighting previously encountered information Relatedly low-lights previously encountered information by graying out the entire sentence in a paragraph. If a user has clicked on a referenced paper in a paragraph and it is referenced in another unseen paragraph, the reference and the corresponding text will be low-lighted there too to indicate that they have previously encountered this information (like

was receiving treatment in hospital was circulated on Twitter, and was subsequently retweeted many times [L. Rubin. 2017]. Moreover fake news across the spectrum, even the most innocuous satire, has arguably)

Mark paragraphs as explored: Similarly, at the paragraph level, once a paragraph is “*marked as explored*”, it is removed from the Overview View. To access these explored paragraphs, a user can click on the “*Show explored paragraphs*” button at the top left of the page. Also, every time a paragraph is marked as explored, all of its references are considered as “encountered” and there is a dynamic re-ranking of the paragraphs list based on the number of unexplored papers in them, how dissimilar the paragraph is to

what has been read, and the relevance to the topic queried (formula 1, described in a previous section).

Progress bars: As paragraphs get marked as explored, the paragraph progress bar at the top of the page is updated (like

Read 3/30 Sections



...). As the user clicks on refer-

ences, the paper progress bar at the top of the page gets updated as well and conveys that the user has read n out of the total number of unique references across the paragraphs returned (like

Read 61/123 References



). The

unexplored reference count badges across paragraphs also get updated. This remains persistent across queries, so as a user issues new queries and if they have read any of the papers or paragraphs before, these would be tracked in the progress bar too. This feature is designed to help address the user challenge that it is difficult to keep track of information read over papers and subtopics explored.

6.4.3 Automatic Section Heading Generation

Relatedly leverages section headings of related work sections to provide users a quick overview over different research threads. One challenge here is that not all authors create descriptive subsection headings. In these cases, showing the generic higher-level section heading (e.g., “Related Work” or “Fake News”) for the query of *fake news* provides little value to the users. To address this, we developed an automated heading generation model, trained on heuristically identified descriptive section headings, and applied it to paragraphs that did not have descriptive headings written by the authors of the source paper.

Method

We experimented with two popular transformer-based sequence-to-sequence models for heading generation: (i) BART [251], and (ii) T5 [337]. These pre-trained models have become de-facto starting points to develop various text generation models due to their strong performance and ability to adapt to different tasks. To further train these models for scientific heading generation, we use a set of heuristics to gather paragraphs that contain descriptive titles from our dataset. These heuristics include filtering out all titles that are single acronyms, shorter than three words or contain generic terms.⁶ This strict filtering favors precision over recall in order to reduce the amount of noise in the filtered dataset for training and testing. Our final dataset consists of 23,957 paragraphs and their titles, which we further randomly divide into 80% training, 10% validation and 10% test splits. We train the large variants of both BART and T5 on this training split for 10 epochs and use loss on the validation split to select the best-performing model checkpoint. Table 6.1 presents an evaluation of both models on our held-out test split using the ROUGE metric [255], which measures title quality via n-gram overlap between generated and human-written titles. Based on these scores, the two models seem to generate similar-quality titles, and we sampled and examined a small subset of the generated headings to compare the two models and found that the BART model tends to produce headings that are more detailed and descriptive. Therefore, we then used the BART model to generate headings for paragraphs outside of the filtered set that did not have descriptive author-generated headings (39,186 in total or around 62% of the entire dataset) for a more rigorous human evaluation detailed in the next subsection.

⁶literature review, background, limitations, future work, conclusion, discussion, related work, results.

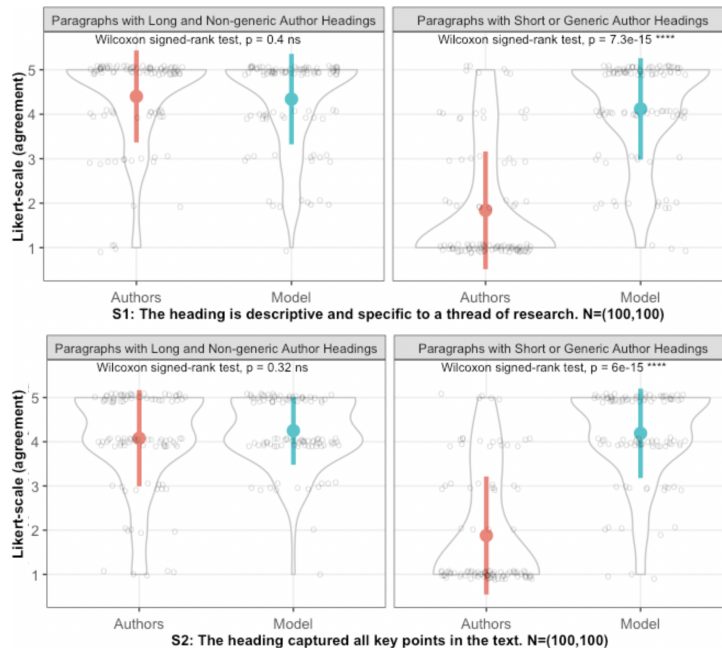


Figure 6.3: Human-evaluation comparing model-generated and author-written section headings. Results suggest that model-generated headings were of comparable quality when the authors had written long and non-generic headings and were of significantly higher quality when the authors did not.

Table 6.1: ROUGE scores for both models on the test split of descriptive section headings. ROUGE-1, -2, and -L measure unigram, bigram, and longest subsequence overlap between generated and author-written titles, respectively.

Model	ROUGE-1	ROUGE-2	ROUGE-L
BART	30.7	14.2	28.8
T5	31.0	14.3	29.0

Human Evaluation

We conducted a human evaluation of the BART-based heading generation model. This is important because automatic metrics, such as ROUGE, do not always correlate well with human perception. We manually rated four sets of headings: both author- and model-generated headings for both paragraphs with descriptive headings but not included in the training (the test split) and paragraphs without descriptive headings. Two statements were rated for 5-point agreement: *The heading is descriptive and specific to a thread of research* (S1) and *The heading captured all key points in the paragraph* (S2). S1 aimed to measure the quality and specificity of the headings, and S2 aimed to measure how well they represent the paragraph text.

To ensure rating quality, two of the authors went through two rounds of redundant rating of 40 randomly sampled headings per round (10 from each set). After two rounds of comparison and discussion to calibrate rating standards, inter-annotator agreements based on Krippendorff’s alpha reached 0.90 and 0.74 for S1 and S2, respectively.⁷ The authors then proceeded to rate 400 headings (100 from each set, paired) without redundancy. During the rating process the authors were blind to the condition each heading was sampled from to avoid bias.

As shown in Fig. 6.3, our model was able to generate headings of comparable quality to long and descriptive titles written by the authors (S1: $p=0.40$; S2: $p=0.32$; $n = [100, 100]$, Wilcoxon signed-rank tests), and significantly higher quality headings when descriptive headings were not available from the authors (S1: $p<0.001^{***}$; S2: $p<0.001^{***}$; $n = [100, 100]$, Wilcoxon signed-rank tests). In addition to these ratings, we also quickly screened all model-generated titles (200 in total) for repetition and hallucinations, i.e. mentions of concepts not present in the paragraph. We do this as a sanity check since generation models are often prone to these issues, especially

⁷Agreements from round 1 were 0.66 and 0.49 for S1 and S2, respectively.

hallucinations [208]. Based on our screening, we found that our model rarely suffers from these issues - only 5/200 (2.5%) titles have repetition and 9/200 titles (4.5%) have hallucination. This may partly be due to the fact that generated headlines are fairly short, which offers less scope for repetition and hallucination to creep in. Given these promising results, in the system, we showed model generated titles whenever a paragraph did not have a descriptive author-written heading (Fig. 1). Some examples of model-generated headings side-by-side with their corresponding author-written headings are presented in the Appendix (Table A.1 and A.2).

6.4.4 Implementation Details

Relatedly is built as a standard web application. The front-end was written in approximately 3,500 lines of TypeScript using the ReactJS framework. The back-end is implemented in approximately 1,500 lines of Python and SQL code. We used Flask for HTTP server framework and PostgreSQL database for both dataset access and behavior logging for the user studies. We used the Whoosh⁸ Python library, which implements the standard BM25 document retrieval algorithm [349], to support full-text search of the paragraphs. For the evaluation study, all interactions with the system (such as new queries, papers and paragraphs read, etc.) are logged to a database in a JSON format (refer to supplementary materials). All communications between the server, database and users' browser are encrypted and anonymized by creating anonymous session and user IDs.

Dataset

To test the Relatedly approach, we gathered a dataset of full papers from five HCI and NLP conferences (ACL, EMNLP, UIST, CSCW, CHI) published between 2016-2021

⁸<https://github.com/mchaput/whoosh>

from S2ORC, a large open-source corpus of 81.1M English-language academic papers spanning many academic disciplines [264]. These topics were selected out of convenience so that the authors could evaluate the usefulness of the system during development, and also for recruiting participants who are likely more engaged with this topic during our user studies. To find paragraphs that summarize multiple prior studies, for each paper, we extracted all paragraphs that contained three or more references along with their section titles. Since a related work section would typically reference its source paper, which can seem out of context when read independently, we used a simple word list to resolve self-referencing phrases (e.g., *in this paper*, *our approach*, *our system*, ..., etc.) to the source paper. In the end, this dataset contained 63,144 paragraphs extracted from 11,382 papers. Approximately 49,975 paragraphs were from related work sections and the remaining paragraphs were mostly extracted from introduction and discussion sections. The inline references were resolved by S2ORC [264] to their metadata including authors, citation count, abstract, and TLDs [67]. This also allowed *Relatedly* to reformat the reference text into APA format (i.e., the first author’s last name and the publication year) in the system so that the same references have the same surface form across paragraphs.

6.5 User Evaluation Study Design

The design of *Relatedly* changes the common literature review process of exploring individual papers (e.g., from a search engine), to exploring paragraphs describing multiple papers on a topic. To investigate its effects, we conducted a within-subjects experiment with 15 participants conducting literature reviews comparing *Relatedly* to a standard paper search engine as baseline. During the study, participants used the assigned system to explore the literature with the goal of creating an outline and notes for writing a survey paper on assigned topics. This allowed us to capture what participants had learned

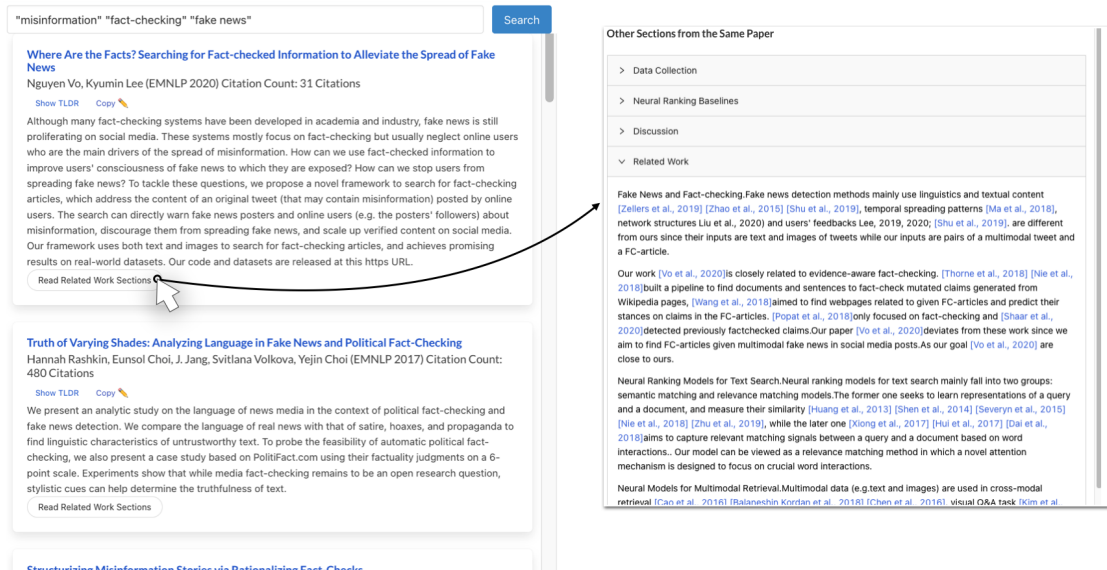


Figure 6.4: The Baseline condition that emulates common scholarly search engines (left). In addition, the “Read Related Work Sections” buttons allow users to read the paper’s sections with paragraphs that contained three or more references with lowered-interaction costs (right).

during the tasks. After the study, we analyzed the behavior logs to understand how they utilized each system, and rate the quality of their outlines to see which process allowed participants to gain a better overview of the literature.

6.5.1 Experimental Setup

We compared *Relatedly* to a Baseline system simulating standard scholarly search engines as an within-subject condition. During the study, each participant used both systems to conduct literature reviews on two different topics. To control for individual differences and learning behavior, we counterbalanced topics and conditions to reduce order effects.

The Baseline Conditions

The baseline condition was a standard BM25 search engine that returned a list of papers from our dataset that mentioned the query term in their titles or abstracts (Fig. 6.4). For each paper in the search results users can access its metadata including the title, authors, venue, publication year, abstract, and a TLDR summary[68].⁹ To lower the interaction costs of using the Baseline condition, users can click on a “Read Related Work Sections” button to access the section headings and paragraphs that contained three or more references (Fig 6.4). This ensures that 1) participants have access to the same data in both conditions, and that 2) the interaction cost of accessing them is low in the Baseline. Similar to the *Relatedly* condition, participants could also click a “copy” button to copy a paper title and paste into their outlines.

Tasks

To contextualize their exploration and sensemaking, participants were given a simulated work task scenario [50] to conduct initial research to get a broad overview of the topic towards the ultimate goal of writing a survey article:

Imagine that you are surveying and summarizing scientific work in HCI and NLP on the topic of:

[One of two task topics: Misinformation, Fake News and Fact Checking OR Crowdsourcing]

Today, please do an initial research to get a broad overview of the topic. Your goal should be to get a broad overview of this topic and identify as many terms, concepts and perspectives related to the topic as you can find by searching and gathering information on this search engine. During the task, write an outline in the notes document provided to you such that it would help you resume work on this task in the future. This may include planning out all the sections of your paper, recording important papers and research you find, etc.

⁹These are 1-2 sentence summaries also available on popular scholarly search engines.

As part of the within-subjects study design for evaluating user behavior across the two conditions, each participant worked on the above task twice (i.e. once for each condition). To prevent carryover effects in learning, each participant completed the task on the two topics listed below. To avoid order effects, the systems were counterbalanced such that they saw a different topic with each condition.

- **Misinformation, Fake News and Fact Checking:** The internet makes it easy for billions of people to access information with a few simple keystrokes. However, it also makes it easy to spread false information, which can have disastrous effects on both individuals and society as a whole. Research in HCI and NLP has focused on detection methods, their use and impacts. Research the impacts of fake news and the methods being developed to combat it.
- **Crowdsourcing:** Crowdsourcing involves a large group of dispersed participants contributing to a task. As we move towards a new future of work with digital platforms for crowdsourcing, research in HCI and NLP has focused on the different methods of crowdsourcing and the applications across different domains. Research the methods and applications of crowdsourcing.

The chosen task topics are relatively large, complex, multi-faceted information spaces where the average person has relatively limited knowledge coming into the task. They are also fairly interdisciplinary tasks so that even if we do get people with domain expertise, there's more they can learn in this area. Also, these topics are well-represented in our dataset – which is papers for HCI and NLP conferences.

Participants

15 participants were recruited from research labs across three universities (8 identified as female and the rest as male; age: 19-32. $M=25.88$, $SD=3.72$). All studies were

conducted remotely over video calls. Compensation was \$45 USD for the 90 minute study. The participants were mostly research scientists, post-docs, and graduate students engaging in research activities.

6.5.2 Study Procedure

Before the study appointment, participants were sent the informed consent form and asked to fill out demographic information. During their study appointment, each participants went through two literature review tasks where the order and the combination of tasks and system assignments were counterbalanced. Each of the two tasks lasted 20 minutes. During the 20 minutes, participants freely interact with the system and create their learning outline on a Google Doc while thinking outloud about their experiences [250, 198]. Before starting each task, participants watched a short tutorial on each system and were given 5 minutes to explore the system using the test topic of “sensemaking”.

6.5.3 Measures

Quality of Learning Outlines

Our primary measure focused on how well *Relatedly* supports literature reviews compared to the baseline by analyzing the learning outline participants wrote in the lab study. For this, we used topic experts to examined and rate each of the Google Doc outlines while being blind to which condition they came from. We defined an expert as someone who has obtained a doctoral degree focusing on the topic, and had multiple years and publications in the field. Two of the authors matched these criteria for the tasks used in the study, blind to condition. The experts counted the number of research themes participants added to their outlines (a proxy for *comprehensiveness*), and rated the outlines for the following aspects on a five-point scale (higher is better):

- *Coherence*: The category structures make sense and the papers and subcategories in the them fit.
- *Insightfulness*: The categories were insightful and captured important research threads in the space.
- *Level of Detail*: The categories contain rich details of relevant subtopics and papers.

These criteria were inspired by literature on human evaluation of clustering (e.g. [459]) and NLP evaluation criteria for automatically generated outlines (e.g. [154]) The two experts went through two rounds of rating and discussion to calibrate the number of themes and their final scores. The sum of these scores is then used to calculate overall quality of the outline.

Behavior Log Analysis

Using the behavior logs from both conditions, we measured how frequently each participant interacted with both the systems at both the paragraphs and papers level. For example, references clicked on, paper titles copied, or paragraphs explored. In addition, we also examined how frequently participants interacted with features only available in the *Relatedly* condition, such as reference low- and high-lights, citation frequency badges, and using the Overview View and Similar Paragraphs View.

Qualitative Insights and Perceived Values

In order to gain deeper understanding of the challenges and benefits of using the two systems, we transcribed participants' think-aloud recordings during the tasks. The first author then went through the transcripts in two passes using an open coding approach [87]. Through discussions with the rest of the research team, we identified common themes in participants' experiences. Additionally, we also conducted a post-task survey

where we asked participants to rate a set of statements around system values (Table 3) using a 5-point Likert-scale for agreement.

6.6 Findings

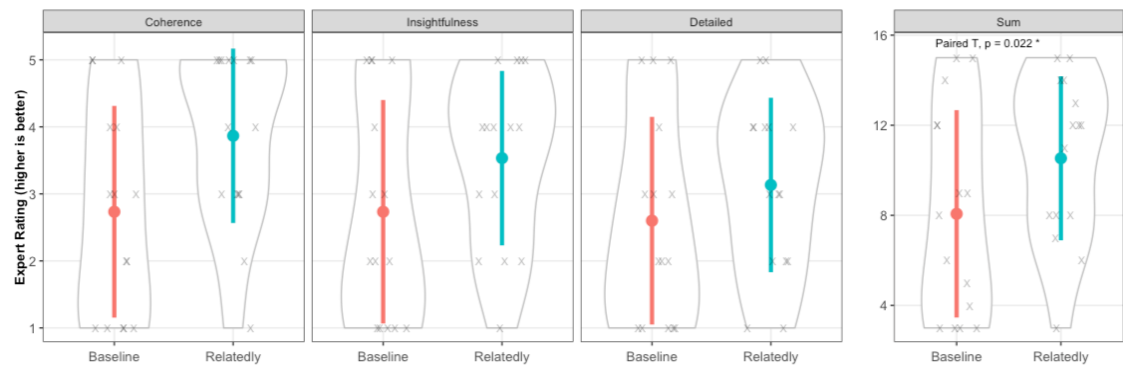


Figure 6.5: Participants generated learning outline summaries after 20 minutes of literature review each with the two systems. The summaries were rated by experts on 3 criteria using 5-point Likert-scales for agreement (5 indicated strong agreement). The experts were blind to which condition each outline came from during rating. A MANOVA and a Wald-type test were used to correct for multiple comparisons and found a statistically significant difference ($manovaF = 7.78, p = 0.02^*$, $Wald\chi^2 = 13.18, p = 0.004^{**}$) between the conditions on the combined dependent variables of Coherent, Detailed, and Insightful.

In this section, we combine results from our three measurements described in the previous section to give a holistic view of the costs and benefits of using *Relatedly* when compared to the Baseline condition, which simulates standard scholarly search engines.

6.6.1 Higher Quality Synthesized Outlines

Based on the sum of 5-point expert ratings on the three aspects, participants wrote significantly *better quality outlines* when using *Relatedly* compared to the Baseline ($M = 10.53, SD = 3.64$ vs $M = 8.07, SD = 4.61; t = 2.58, p = 0.02$ out of 15, Fig. 6.8). To correct for multiple comparisons, we ran a Wald-type test and a MANOVA with

repeated measures and found a significant difference between the two systems on the combined measures of *Coherence*, *Insightfulness*, and *Detailed* ($manovaF = 7.78, p = 0.02^*$, $Wald\chi^2 = 13.18, p = 0.004^{**}$; Fig 6.8).¹⁰ Breaking this quality score down, participants wrote significantly more coherent ($M = 3.87, SD = 1.30$, out of 5) and insightful ($M = 3.53, SD = 1.30$, out of 5) outlines when using Relatedly compared to when using Baseline (Coherence: $M = 2.73, SD = 1.58$ out of 5; Insightfulness: $M = 2.73, SD = 1.67$ out of 5). Participants also wrote more detailed outlines when using Relatedly ($M = 3.13, SD = 1.30$ out of 5) compared to when they used Baseline ($M = 2.60, SD = 1.55$ out of 5).

Qualitative analysis of the think-aloud recordings revealed participants' exploration strategies when using *Relatedly*. Most commonly, 13 out of the 15 participants talked about using a "breadth-first approach for exploring different paragraphs and topics," and 8 specifically mentioned using the Unexplored Reference Count Badge (11 new references) at the paragraph level to prioritize. For example, P09 used the Overview View to quickly capture diverse topics and relevant papers in their outline, taking advantage of how *Relatedly* ranked the paragraphs dynamically to maximize marginal novelty [76]:

"I spent most of my time in the all paragraphs view looking at the various summaries. – I like that they are in [the] order of most unread references to fewest, so it felt like going in-order made sense. As I found new topics, I jotted them down – sometimes as a short summary, sometimes I included entire quotes of the overall message, and then copied over the related paper to go back to reference if I needed." (P09).

All participants switched between the Overview View and Similar Paragraph View for broad overview and drill-down into different subtopics. Furthermore, when using Relatedly, participants explored significantly more paragraphs in the Overview View ($M = 5.80, SD = 8.00$) than in the Similar Paragraphs View ($M = 1.80, SD = 1.13, t =$

¹⁰R package: *MANOVA.RM: Resampling-Based Analysis of Multivariate Data and Repeated Measures Designs*

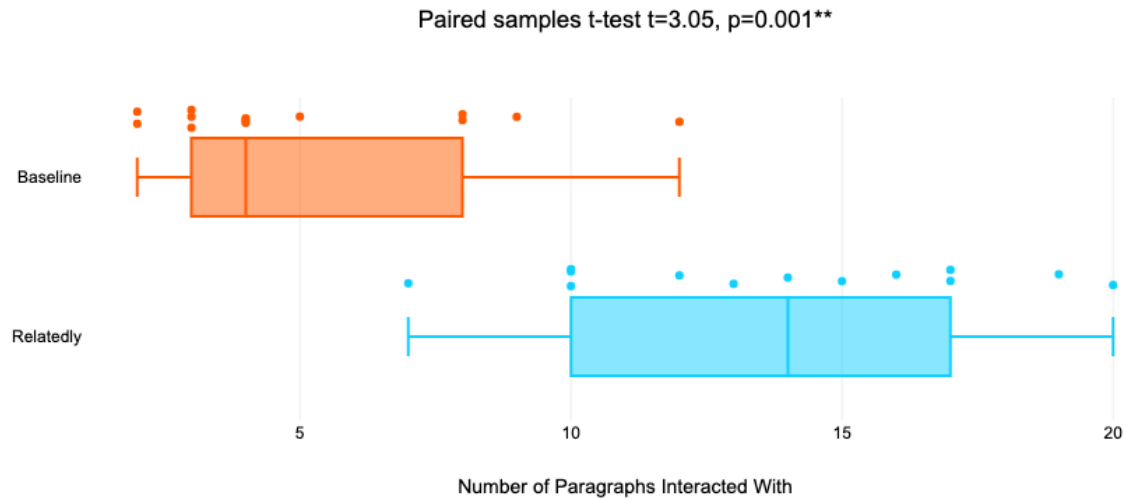


Figure 6.6: Participants interacted with significantly more paragraphs when using Relatedly vs the Baseline system

2.55, $p = 0.02^*$). Two participants specifically mentioned switching between the two views to control both the depth and breath of their explorations:

“I’d describe my searching as like a limited-depth depth first search strategy. I start at a high level idea, try to find everything related to that topic up to some effort level. Similar Paragraphs [View] was a great feature for this! Then, I switch to another sub topic [in the Overview View] and repeat.”
(P17).

These system log analysis and qualitative insights suggested that participants had more control over their exploration when using *Relatedly*, fluidly switching between exploring many diverse subtopics and drill-down to specific subtopics, and are potential explanations for how they collected more subtopics and generated higher quality overview learning outlines.

6.6.2 Paper- vs Topic-Centric Exploration

While participants in both conditions had access to papers’ related work paragraphs, we found that when using *Relatedly*, participants interacted with more than twice as many paragraphs compared to when they were using the Baseline (Relatedly: $M =$

16.85, $SD = 7.66$ vs $M = 7.21$, $SD = 5.70$, $t = 4.40$, $p = 0.001^{**}$). Participants described their exploration strategies were centered around individual papers instead of paragraphs when using the Baseline. Most commonly, nine (/15) mentioned relying mostly on the abstracts and TLDR summaries to decide which papers to read: “*Go paperwise, skim through abstract - read more if it’s interesting or relevant*” (P04). Participants also mentioned using other signals such as the citation counts (3/15), spotting unfamiliar terms in the abstracts to find categories to add to their outlines (3/15), and reading related work sections (4/15).

In addition to differences in exploration strategies between the two conditions, we also found participants actively used the paragraph reading support features when using *Relatedly* based on both behavioral and qualitative data. On average, 38.73% of citations clicked were low-lighted previously read references, 37.84% of citations clicked were highlighted and 10.35% of citations clicked had citations frequency badges. Based on qualitative data, participants utilized the reading support cues to both prioritize and de-prioritize their reading activities. 12/15 participants talked about relying on the yellow highlighted references to prioritize their reading:

“oh I see yellow, which means that there is something different, and the brighter the yellow, it looks more attractive to me. I want to see papers that disagree to the papers I’ve read so far” (P04).

In addition, 11/15 participants paid attention to the citation frequency tags to find important papers on the topic:

“I will look for the citation with high numbers, because those tend to be popular and more classic or fundamental.” (P03).

For deprioritization, 12/15 participants talked about their use of low-lighted references:

“when I start to read a paragraph and I see grayed out text and references, it is very helpful to see that I have read about this paper before in another paragraph. I’m going to click on this to verify which paper this is and then

read how this paragraph is discussing it. Oh it seems to be a slightly different take on this paper's contributions so I'll copy this paper. It seems like there's some discussion around it." (P09).

Finally, participants mentioned using the descriptive section headings to explore topics when using *Relatedly*. Ten participants explicitly mentioned how the section titles indicate topics that are useful for organizing knowledge:

"These [titles] are roughly the broad categories for which I would look for, and I'm first going through the titles and adding the unique ones to my notes ... I will now start looking into specific things similar to this subtopic title" (P14).

6.6.3 Participants Preferred Relatedly

After using both systems, participants were asked about their preferences and 13 / 15 participants said they preferred *Relatedly*'s approach of reading related work paragraphs on the topic to the Baseline's approach of reading papers to review literature. While the two participants who preferred the Baseline mostly described it as a familiar interface, participants who preferred *Relatedly* reported much richer explanations: (1) provides a good structure and organization to an otherwise unstructured complex exploratory task: *"I can have a relatively clear path to explore. I can use what other researchers have summarized so I don't need to start from zero."* (P05); (2) helps understand connections between multiple papers: *"For literature review, it is important to see connections between cite works... I find the Related work sections helpful for this, not so much for the paper list."* (P14); (3) gives the right amount of relevant context around papers: *"Gave a lot more context and resulted in fewer papers being read – had to open fewer pdfs."* (P13); and (4) helps track progress and prioritize what order to read things in: *"can help me to keep track of my pace of learning about the topic."* (P03).

In a post survey about participants' opinions about the two systems, participants thought the *Relatedly* system helped them significantly more than the Baseline for "find

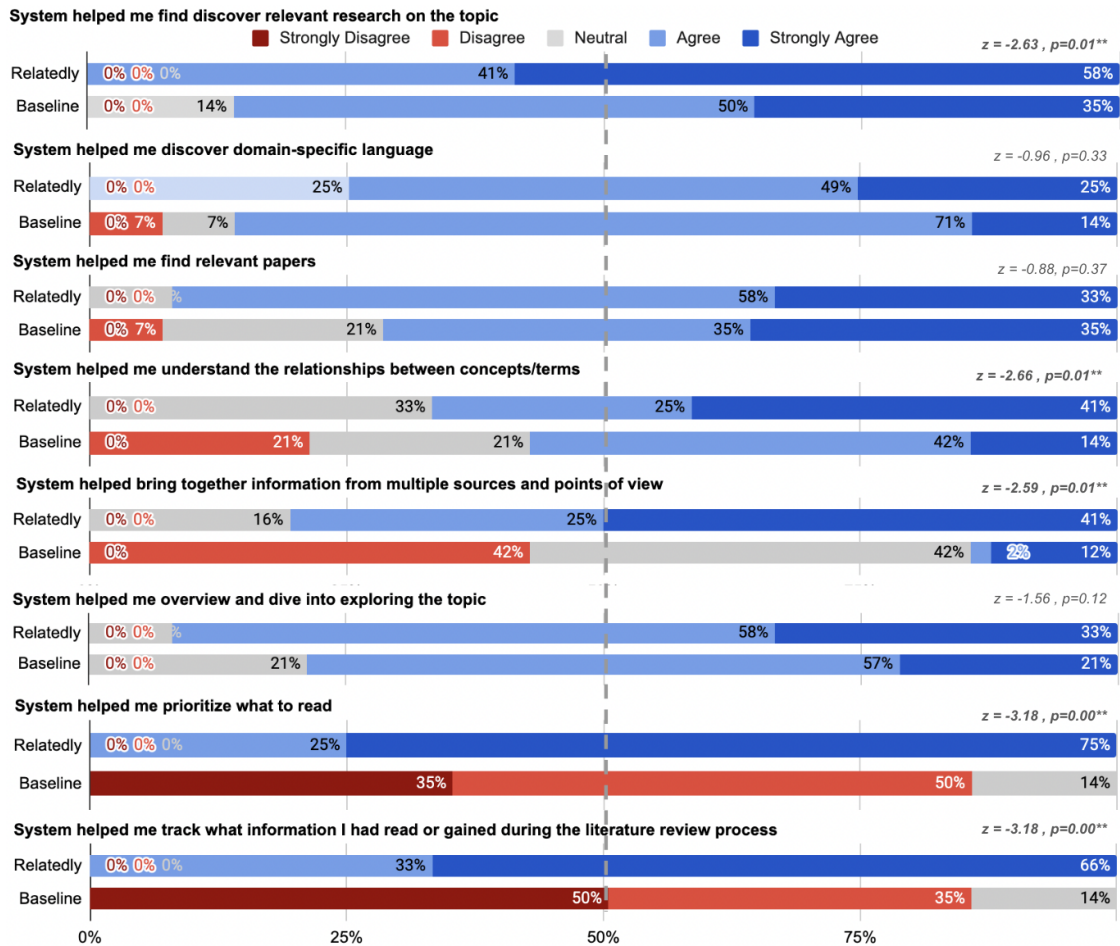


Figure 6.7: Participants' level of agreement to how well Relatedly and Baseline supported their literature review process. For each statement, we report the percentage of likert responses and results from paired wilcoxon signed rank tests, with z and p values.

relevant research on the topic”, ”understand relationships between terms/concepts”, ”bring together information from multiple sources and points of view”, ”prioritize what to read” and ”keep track of information gained or read during the literature review process” (Wilcoxon signed-rank tests for all statements reported in Fig. 6.7).

Lastly, to measure the perceived workload of using the two systems, we used the NASA TLX questionnaire, and found no significant differences (Appendix Table A.3). This result suggests that participants did not perceive higher workload when using *Relatedly* even though it consisted of significantly more features, and that they were able to utilize these features to synthesize learning outlines of significantly higher quality.

6.6.4 Volunteered Use in the Wild

PID	Job Title	Reason for using Relatedly during their research workflow	Topic of Interest	Hours using Relatedly/ Total Hours Working	# of queries	# of papers curated
P07	Research Assistant	Finding and curating relevant papers into an annotated reading list	Smartphone accessibility techniques for people with motor impairments	4/4	10	30
P09	Post-Doctoral Researcher	Researching related work to identify gaps	Cognitive and design theories on feedback	25/36	14	102
P11	PhD student	Starting a new project in an unknown research topic	Creativity support tools, for 3D prototyping	10/10	4	98
P13	Machine Learning Researcher	Checking for unknown papers when writing a paper on a known topic	Multi-document summarization techniques	5/5	3	51
P15	Research Scientist	Checking for unknown papers when writing a paper on a known topic	Use of AR/VR techniques in health and education	1.5/4	11	23

Figure 6.8: Background and usage of participants who volunteered long-term use. Participants self-reported their job title, reason for using Relatedly in their research workflow, hours of work, hours using Relatedly, # queries, and # of relevant papers curated to read

One interesting but unplanned observation was that five of the participants were planning to conduct literature reviews for their upcoming paper submissions and expressed interest in using Relatedly after the study had concluded. We saw this as an opportunity to

better understand how *Relatedly* performs for real-world tasks over a prolonged period of time. Therefore, we continued to allow them access to *Relatedly* from their computers to conduct their own tasks, and scheduled interviews with them after two weeks to learn about their user experience. The first author then open coded the transcripts of these interviews to identify common themes. All participants were Ph.D. students across three large research universities with an average age of 25.7 years. Two identified as male and three as female. Table ?? shows an overview of their real-world tasks and their engagements with *Relatedly*.

These scholars expressed their preferences for reading related work sections in *Relatedly* over reading individual papers to “*learn about a topic*” (P09), and “*verify whether I have covered all the important research threads, papers and perspectives when writing my paper*” (P13) , “*discover what are the central papers on this topic*” (P15). They also mentioned that it is still helpful to use traditional scholarly search engines to look up specific papers or author, and saw the two approaches as complementary to each other. P13 summarized their experience by comparing it to their previous literature review process:

“there’s a learning curve to getting used to the shift from reading papers individually one by one to reading paragraphs instead, but once you get used to this paradigm, it’s easier to explore the topic this way.” (P13)

Most other benefits mentioned in the post-interviews echoed benefits uncovered in the lab study. Additionally, participants reported using *Relatedly* with other apps such as: scholarly search engines and PDF readers to read the papers thoroughly; and reference managers like zotero , note-taking apps like notion, documents to attach notes and curate papers for later use. We see this as a promising signal that *Relatedly* is also able to support real-world tasks over a prolonged period of time, and the fact that participants from Study 1 volunteered to use *Relatedly* after the study under no obligations nor compensation suggested that it provided real-world value to at least the five scholars.

6.7 Discussion

This chapter illustrates opportunities of leveraging prior effort (i.e., content and structure of existing related work section paragraphs) to scaffold the users in discovery and synthesis for literature reviews. Motivated by a formative study that identified challenges in this approach, we design a novel system, Relatedly, which provides scaffolding features such as auto-generated descriptive paragraph titles, high and low-lighting paragraph text to facilitate reading overlapping and novel information, re-ranking paragraphs to maximize subtopic breadth while allowing users to drill-down to specific subtopics. Here, we draw connections between observations from the user study and theory to further contextualize our findings.

Educational psychologists [19, 352] have found that giving students multiple explanations or different representations of the same topic helps them overcome the weaknesses of any particular explanation or representation and better make sense of a complex scientific topic [20] and problem solve [104]. In this context, reading multiple similar paragraphs written by different authors who might frame the topic slightly differently might help users understand and synthesize information topic better. Based on the behavior logging, we indeed found that participants explored significantly more paragraphs when using Relatedly than when using the Baseline, which could potentially explain why they wrote significantly higher quality learning outlines when compared to the Baseline.

Information foraging theory suggests that people have a tendency to switch between information patches in order to maximize the amount and breadth of information gained during exploration [327]. Relatedly's Overview and Similar Paragraphs Views were designed to facilitate this strategy when exploring multiple related work paragraphs extracted from multiple papers with overlapping and dissimilar information. During the think-alouds and in the post-task surveys, participants described using the two views to

to fluidly alternate between exploring diverse subtopics and exploiting reading details about a specific subtopic, which could potentially explain why they were able to write down significantly more themes/subtopics in their learning outlines when compared to the Baseline.

Finally, participants thought Relatedly was more helpful than the Baseline for literature reviews. Participants in the evaluation study agreed significantly more to the statements: the system helped me “find relevant research on the topic”, “understand relationships between terms/concepts”, “bring together information from multiple sources and points of view”, “prioritize what to read” and “keep track of information gained or read during the literature review process” compared to the Baseline system. Overall 13 out of 15 participants preferred using Relatedly compared to the Baseline for literature review, and five adopted Relatedly to support their upcoming paper submissions after the study had concluded. Considering this continued usage was volunteered without obligations nor compensations, we see this as a promising indication that Relatedly was able to provide real-world benefits when participants used it to support their own literature review tasks.

We believe that any system that uses algorithmic approaches to help user manage their attentions (recommendations, search, summarization), should be aware of and account for potential model errors misleading users, implicit biases, and echo chamber effects in their designs. When developing Relatedly, we also aimed to mitigate these potential risks. For example, we were careful about the quality of Relatedly’s generated section headings and conducted an additional human evaluation in addition to the standard automatic evaluation techniques in NLP (ROUGE). Several of Relatedly’s UI features also aimed to combat these risks. For example, the progress bars encourage users to cover more papers and paragraphs instead of feeling satisfied with what they had already explored. The benefits of this increased exploration can help offset possible changes in the distribution

of papers explored. Further, while most prior recommender systems help users in finding documents *similar* to what they have already explored, Relatedly, in contrast, actively re-ranks documents and highlight sentences to encourage users to prioritize exploring information *most dissimilar* to what they have already explored.

Relatedly's main insight is that given multiple different texts on a particular topic, it scaffolds the reader's exploration by helping prioritize new dissimilar information and de-prioritizes redundant information. Relatedly demonstrates this approach using scientific texts and related work sections, however, this approach could be applied in other domains such as policy makers reviewing policy literature for policy briefs, or lawyers researching prior cases to identify patterns, legal precedent and make a case, or voters tracking important issues across multiple politicians' platforms or news articles, or programmers trying to navigate different discussion fora or Jupyter notebooks for solutions for a bug. We envision an augmented reading experience which supports getting a broad overview of different perspective or themes, lets you prioritize what to read and track what you have read so far across information sources on any topic on the internet. We leave these promising research avenues for future work to explore.

6.7.1 Limitations and Future Work

Many design decisions were influenced by our focus on literature review of scientific topics by scholars. One assumption we made was that scientists write high quality related work sections in their paper that can provide more benefits to users than looking at individual papers and synthesizing them. For this, we selected papers from leading venues in HCI and NLP to construct our dataset. Future work on this approach that wishes to expand the coverage to all scientific publications would require additional support for finer-grained user control over the sources that the related work paragraphs are extracted from. For example, allowing users to curate a set of venues or authors that

they trust. Alternatively, future research could expand on NLP techniques for assessing writing quality [290, 402] to automatically rate the quality of related work sections [404] and incorporate them into the ranking algorithm. Another future direction for further improving Relatedly is to analyze the importance of each reference in the context of the paragraph they were mentioned. For example, NLP techniques such as [419] could be used to estimate the level of influence of each references in a paragraph, so that Relatedly could mitigate the effects of bulk and passing citations [188, 226, 189].

When designing our user study, we also considered citation graph visualization tools as an additional baseline condition. However, literature review tasks can be difficult to study [241], because they can be mentally taxing and time consuming for participants due to their exploratory nature [241]. To keep our study realistic, we wanted to keep participants engaged with longer literature review sessions, while keeping the whole procedure under 90-minutes to prevent fatigue. Therefore, we chose the Baseline condition which simulates the most-popular literature review strategy for most users (i.e., scholarly search engines and reading papers individually). Future work can build on our study by including prior visualization systems as a baseline to compare Relatedly against to help us further understand the costs and benefits of Relatedly.

In the formative interviews, we mostly interviewed PhD students who are junior scholars. Since Relatedly aims to help people jumpstart their lit review process by broadly overviewing and finding relevant papers in an unknown topic, we focused on understanding the needs of junior researchers. While three of the participants in the formative study were full-time, post-graduate researchers, we focused on junior researchers for the formative study because they tend to face more challenges and need more support with the literature review process [138, 199, 113, 403] compared to senior scholars. Senior scholars are more likely to rely on wider social connections to support paper recommendations and have richer adjacent domain knowledge to draw

from [113, 138]. Existing research systems support senior researchers' literature searches by recommending papers based on social signals such as who they have collaborated or interacted with before [217, 218], and papers, authors, institutions, venues of work that they have read or curated [220].

To avoid potential unintended consequences such as plagiarism, Relatedly's design aims to highlight the provenance of ideas and encourage correct referencing practice by prioritizing author information at the top of paragraphs and attaching author information when users copy over references.

One potential obstacle to broader adoption of this approach is licensing and access to scientific documents. Specifically, not all scholarly papers can be freely accessed and searched by anyone on the internet. On the other hand, recent trends in promoting *open science* [288] and efforts such as the S2ORC dataset [264], arXiv.org [289, 159], and the Open Science Foundation¹¹, and making older articles accessible using technology such as OCR, GROBID [265], VILA [374], point to a promising future where scholars can take fuller advantage of each others prior research effort, enabling new technologies and interactions such as Relatedly.

Currently, Relatedly is designed for scholarly users. However, an interesting future direction could be supporting lay-people to make exploring scientific information more accessible. For example, if an individual wanted to overview scientific literature on vaccines to determine whether or not to get vaccinated or if they want to overview literature to apply scientific research as a startup product. The opportunity here is that seeing different perspectives from authors of different papers describing a research topic and each other's work has the potential of avoiding lay-people overly trusting a single piece of evidence [136]. In this case, a future version of Relatedly could help not only link unfamiliar terminology to definitions [177] or summarize paragraphs in plain language

¹¹Open Science Foundation: <https://www.cos.io/products/osf>

[28], but more importantly also surface agreements and disagreements between prior works and their levels of uncertainty while helping users build confidence and trust about their learning could be especially important [136].

Knowledge work and literature reviews usually involve exploring multiple topics by issuing multiple queries [138, 316]. This is evidenced by the results of the participants who volunteered to use Relatedly for their real-world tasks. While our user evaluation lab study observed how Relatedly helped participants explore information on a single topic and query, an exciting avenue for future work is investigating how users shift their exploration over multiple topics and queries. Recent work, like [317, 320], help people exploring a new domain articulate queries when they lack domain-specific language and well-defined informational goals. We leave it to future work to extend Relatedly’s approach to better support exploring scientific literature over multiple queries and topic shifts.

6.8 Conclusion

In this chapter we explore a novel approach for supporting literature review workflows—instead of focusing on making sense of individual papers one-by-one to understand a topic, *Relatedly* guided users to explore different subtopic areas using many related work section paragraphs extracted across multiple papers. The idea here is that by leveraging prior scientific efforts of authors conducting literature reviews to write their related work sections, we can improve other researchers’ literature review process. To address how paragraphs extracted from different documents might cover both similar and distinct topics, *Relatedly* also provides reading support cues and information gain tracking features to facilitate users in reading many related work paragraphs to cover a broad overview of different topics more efficiently. A comparative user study demonstrated

that Relatedly’s approach to literature review helped scholars synthesize information on the topic in a broader, more coherent and insightful manner. This might have been because Relatedly’s reading support features scaffold discovering and interacting with more paragraphs and papers, which helps explore broad multifaceted information spaces. Additionally, participants discovered more domain-specific terms when using Relatedly and preferred using it over the Baseline. We believe the Relatedly approach brings us one step closer to leveraging information and structure available on the web to support knowledge exploration and synthesis.

6.9 Acknowledgements

This chapter in part, includes portions of material as it appears in *Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections* by Srishti Palani, Aakanksha Naik, Doug Downey, Amy X. Zhang, Jonathan Bragg, Joseph Chee Chang in Proceedings of the 2023 ACM CHI Conference on Human Factors in Computing Systems (CHI’23). The dissertation author was the primary investigator and author of this material.

Part III

Augmenting Creative Workflows

Chapter 7

The Practitioner Perspective on Creativity Support Tool Adoption

With the rapid development of creativity support tools, creative practitioners (e.g., designers, artists, architects) have to constantly explore and adopt new tools into their practice. While HCI research has focused on developing novel creativity support tools, little is known about creative practitioner’s values when exploring and adopting these tools. We collect and analyze 23 videos, 13 interviews, and 105 survey responses of creative practitioners reflecting on their values to derive a value framework. We find that practitioners value the tools’ functionality, integration into their current workflow, performance, user interface and experience, learning support, costs and emotional connection, in that order. They largely discover tools through personal recommendations. To help unify and encourage reflection from the wider community of CST stakeholders (e.g., systems creators, researchers, marketers, educators), we situate the framework within existing research on systems, creativity support tools and technology adoption.



Figure 7.1: Visual abstract of this paper’s investigation of what creative practitioners value when adopting creativity support tools – summarizing the key contributions: C1. Empirical Observations, C2. Creative Practitioners’ Value Framework, C3. Mapping values to design principles and theories in literature

7.1 Introduction

Creative practitioners (e.g. professional designers, software developers, artists, architects, film-makers, etc.) harness digital technology to achieve their goals and augment their creative potential. This use of digital technologies to support creative practice – Creativity Support Tools (CSTs, e.g. AutoCAD and Illustrator), have been studied for decades and is considered a “*grand challenge*” in HCI research [381, 150]. Frich et al. [150] define a CST as “*technology that runs on one or more digital systems, encompasses one or more creativity-focused features, and is employed to positively influence users of varying expertise in one or more distinct phases of the creative process.*”

Over the past couple of decades, *creative domains*, i.e., industries that conceive products and services [351, 100], such as design, software development, architecture, and film, and entertainment, have grown both in industry and research [381, 150, 340]. In this rapidly evolving landscape, it has become imperative for creative practitioners to constantly explore CSTs, and decide whether to adopt new tools or abandon current ones. However, little is known about creative practitioners’ values when choosing and exploring tools.

HCI research has developed several novel tools to stimulate creative thinking and support design processes (e.g., [151, 150, 149, 116, 228]). However, most of these prototype CSTs exist in a lab setting – few explorations are carried out for tools in-the-wild, over a long period of time [150, 340, 342, 380]. To address this issue, Frich et al. [150] suggest “*shifting our efforts to studying in-vivo use of creativity support tools, not just the ones we build ourselves, but the ones that most creative practitioners employ in practice*”. This premise motivates our research questions:

RQ1: *What do creative practitioners value when adopting CSTs?*

RQ2: *How do creative practitioners discover and explore new CSTs?*

To address these questions, we analyzed 13 interviews and 23 YouTube videos of creative practitioners reflecting on their values when adopting CSTs. We synthesize the findings from this analysis in a conceptual framework of values held by creative practitioners when deciding whether to adopt a new CST. Then, to contextualize and verify identified trends in values with a larger population of creative practitioners, we surveyed 105 creative practitioners and asked them to rate and rank each of the values in the framework.

This investigation uncovers that creative practitioners care about multiple factors: CST's features and functionality, integration with existing workflow, performance, interface and user experience, support, financial cost, and even the emotional connection with the tool. Delving into the subcategories, the highest-rated values were a CSTs' reliability in performance and ease of use. This chapter makes the following contributions (Figure 7.1):

C1. Empirical observations from creative practitioners [§7.4]. The analysis of YouTube videos, interviews and survey responses, adds the creative practitioners' perspective to existing CST developer, educator or researcher-centric perspectives described in literature.

C2. Creative Practitioners' Value Framework [§7.4]. A conceptual framework of creative practitioners' values for discovery and adoption of CSTs as shaped by C1.

C3. Unified mapping of practitioners' values to design principles and theories in literature [§7.5] We connect our proposed framework to principles in the existing literature to encourage reflection and innovation from CST stakeholders (e.g. systems creators, researchers, marketers, educators).

To contextualize these contributions, we describe existing design heuristics in HCI systems, CST research, and theories of tool adoption [§7.2]. The research methods [§7.3] detail how we collected and analyzed video, interview data, and survey data. The

definitions of our framework, along with the empirical observation and numerical data are outlined [§7.4] before they can be tied back together to the foundational literature [§7.5]. We conclude by discussing the limitations and the avenues for future work [§7.6].

7.2 Related Work

To frame practitioners' values, one must consider elements of both technology, as well as usage preferences and practitioner needs. This work builds on HCI systems and CST design and evaluation; and social science theories of technology acceptance and adoption.

7.2.1 Designing and Evaluating Creativity Support Tools

As a sub-field of HCI research, studies of CSTs formally began two decades years ago, when Shneiderman alluded to computers' potential to become tools that enhance human creativity [379, 380]. CST research has developed tools for many stages, such as making discoveries or inventions from information gathering [271, 317], hypothesis and idea generation [383], and initial production [116, 148], to refinement [228], validation [149], and dissemination [380, 150].

Note how the term "*tool*" is tied to the Human-Centered perspective on what is used to accomplish a task, ranging from applications (e.g., Figma), toolkits (e.g., D3 to visualize data), and programming languages (e.g., C#), as opposed to individual commands (e.g., undo, copy) or a tool's *features* (e.g., using a brush inside an application).

The HCI and creativity research communities have proposed quantitative and qualitative approaches to evaluate the usefulness of CSTs. One quantitative measure is the Creativity Support Index [77, 92], a general-purpose survey to gauge a novel CST's effectiveness. Other methods include co-design workshops [127], physiological re-

sponses (e.g., galvanic skin responses, EEG) [78], and self-report in post-study reflective think-alouds and surveys[364, 431].

CST research also follows design principles proposed by HCI systems research. Myers outlines that systems should facilitate: (i) *Path of Least Resistance* (i.e., leading users towards doing the right things, and away from doing the wrong things) and (ii) *Predictability* (i.e., alignment with the user's mental model), (iii) "*Low Thresholds, High Ceilings, and Wide Walls*" (i.e. that tools should be easy for novices to get started, yet provide ambitious functionality that experts need and provide a wide range of functionality with underlying services). Olsen [310] outlines similar concepts: (i) *Generality* (i.e., the ability for a tool to generalize across situations, tasks and users), (ii) *Reduce solution viscosity* (i.e. reducing the effort required to iterate on many possible solutions), (iii) *Enabling Expressive Leverage* such that a designer can accomplish more by expressing less, (iv) *Facilitating Expressive Match* (i.e., mapping how close the means for expressing design choices are to the problem being solved), (v) *Power in combination* (i.e., supporting combinations of more basic building blocks through: (a) *Inductive combination* (i.e., combining features within one tool to accomplish larger, more complex goals), or (b) *Simplifying Interconnection* (i.e., all components/features of the tool should work with each other within and across other tools). Similarly, Cognitive Dimensions of Notation [166, 47] was also used to reflect on systems, though its usage in the literature has decreased in favour of Olsen's framework likely given the high overlap [245].

Similar design principles and heuristics are outlined in CST research as well. For example, Resnick et al. [342] echo Myer's [301] design principle of "*low thresholds, high ceilings and wide walls*". Resnick et al. also proposed additional principles: (ii) *support many paths and many styles*, (iii) *support collaboration*, (iv) *support open interchange*, (v) *make it as simple as possible*, (vi) *choose black boxes of explorability carefully*. Additional perspectives informed by developers and HCI researchers include: (vii) *invent*

things you would want to use yourself, (viii) balance user suggestions with observation and participatory process, (ix) iterate (x) design for designers, and (xi) evaluate your tools. Shneiderman [380] frames general design recommendations for CSTs: *(i) support exploratory search, (ii) enable collaboration, (iii) provide rich history-keeping, and (iv) design with low thresholds, high ceilings and wide walls.* The above systems and CST research papers acknowledge the importance of establishing frameworks that foster reflection systems' usefulness and contributions to the research- and user-communities.

Recent surveys of CST and HCI systems research show a focus on building novel tools often evaluated in controlled experiments with novices and students as primary subjects[340, 245]. This might be due to research prototypes' limited resources to operate at scale. This constraints the understanding of in-the-wild use of CSTs over a long-period of time by practitioners. Still, there is room to better understand long-term tool use within people's existing practices and use these findings to better inform system building in HCI.

In practice, creative professionals usually opt for CSTs made by established industry tech companies, for example digital designers use Adobe Illustrator or InDesign, software developers use Microsoft Visual Studio [408]. This work builds on and unifies these multi-disciplinary reflections and sheds light on practitioners' perspectives long-term when exploring, adopting, retaining, and abandoning CSTs.

7.2.2 Theoretical Background On Technology Adoption

Research in social sciences has explored the theory for what influences individuals' acceptance and adoption of emerging technologies in education, healthcare, and other information provisions.

Rogers [350] defines technology adoption as *"a decision to make full use of an innovation as the best course of action"* (p473). The adoption process includes an

individual's acceptance or rejection of the innovation, its subsequent use, and purchasing and acquisition decisions [341]. Rogers Innovation Diffusion Theory [350] posits a five stage process for technology adoption – the innovation-decision process: (i) *Knowledge*, occurs when an individual learns about an innovation; (ii) *Persuasion*, involves the individual forming an opinion on the innovation; (iii) *Decision*, occurs when the individual prepares to choose to adopt (or reject) an innovation; (iv) *Implementation*, is when the individual uses the innovation, and (v) *Confirmation*, is when the individual reinforces the decision to adopt or reject the innovation. Rogers' Innovation Diffusion Theory proposes that users base technology adoption decision on perceptions of the tool's: (i) *relative advantage* (the extent to which a new technology is seen as being beneficial over the preceding one – similar to performance expectancy), (ii) *complexity* (the difficulty in using it – similar to effort expectancy), (iii) *compatibility* (the extent to which using the target technology is viewed as being compatible with the user's beliefs, values, and work patterns), (iv) *trial-ability* (the possibility to try, experiment, and reduce uncertainty and to learn by doing prior to adopting), and (v) *observability* (the visibility of the results of adoption, which stimulates discussion, interest, and uptake). Other theories exist [201, 252, 372, 376, 443, 442], yet they have received criticism for excluding external conditions [131, 395, 400, 443].

Parallel research on *Technology Acceptance* has also been developed: including the Theory of Reasoned Action [363], Theory Of Planned Behaviour [21], Technology Acceptance Model and TAM2 [249] and the Unified Theory Of Acceptance And Use Of Technology by Venkatesh et al. (UTAUT) [424]. These models predict that technology acceptance is influenced by: (i) *performance expectancy / perceived usefulness* (the extent to which potential users expect performance improvements using the new technology); (ii) *effort expectancy / ease of use* (the extent to which people expect usage to be free of effort); and (iii) *social influence / subjective norms* (perceived pressure from others

to use the technology). These theories focus on predicting acceptance instead of actual use and adoption of technology. While the terms "adoption" and "acceptance" are often used interchangeably, they actually refer to two distinct aspects. Acceptance is viewed as a component of adoption [341], such as the willingness to use technology for the tasks it was designed to support [122]. Willingness and actual use are separate and different measures. This chapter unifies the vocabulary used to describe the CST design principles and theoretical model parameters, and adds a layer of granularity and richness to existing models by presenting empirical observations from practitioners.

7.3 Method

To understand what influences creative practitioners when exploring and adopting CSTs, we followed a two-fold approach:

1. Observation. We collected 23 YouTube videos and conducted 13 semi-structured interviews with creative practitioners to gain an initial overview of values across participants.

2. Survey. To verify and contextualize the observed trends with a larger population of practitioners (105 responses), we designed a survey for practitioners to rate and rank the different values.

Questionnaires are available in the supplementary materials and were approved by our organizations' ethics review.

7.3.1 YouTube Videos

We chose YouTube's¹ comprehensive public video database as a start because this data includes practitioners sharing knowledge through vlogs, tutorials, personal experience,

¹<https://www.youtube.com>

etc, and these videos have a wide reach to general audiences.

Sampling. To sample videos, we queried YouTube keywords such as “*why I switched to...*” and selected autocomplete suggestions about CSTs. Sample queries include “*why I switched to Figma from Sketch*”, “*why I switched from AutoCAD to Revit*”. We excluded less CST-relevant queries e.g., “*why I switched to...*” “*...iPhone from Android*”, “*...formula*”. We focused on comparisons and creators reflections, hence we excluded videos mentioning a single CST.

Filtering. We ensured to cover multiple creative domains, such as 3d modeling, software development, creative writing, architecture, video editing, and UI/UX design (Figures B.2 in Appendix). To base our data on audience relevance, we selected videos with over 10,000 views. We collected material past data saturation in case a particular domain yielded new findings.

7.3.2 Semi-Structured Interviews

While the YouTube dataset provides a base data, there are two key limitations. First, the videos shown are decided by the internal algorithm, which has its own biases as defined by its code, advertisements, company sponsorship, audience, and search location, etc.. Second, the videos are crafted by content creators, leading to short narratives designed to capture an audience. To further expand and enrich the data, we interviewed professional practitioners.

Participants

We chose purposeful sampling [48] as recruitment strategy, mixing direct contacts as well as recruitment through a large software company’s Slack channel and a university. We interviewed a diverse mix of participants across different practices, ages, organizations, gender, race, location, cultures, and target audiences. We recruited 13 participants

(8 male, 5 female) across nine creative fields including graphic design, UX design, architecture, industrial design, software programming, film, game design, and sketching (Figure B.1 in Appendix). While we reached data saturation by the 8th participant, we continued interviews to reach a larger coverage of professions/roles. Participants' ages ranged from 22 to 59 years ($M = 33.23, SD = 7.10$). Compensation was \$50 USD or equivalent for the one hour interviews.

Procedure

Before the interview, participants answered a demographic questionnaire collecting: age, gender, occupation, organization, team size, educational background, professional experience, and expertise in their creative field and in digital CST use.

Interview questions were drawn from a semi-structured interview guide. (Questionnaires are available in the supplementary materials and were approved by our organizations' ethics review). To ground the discussion, we asked participants to recall the last adopted CST, and the most interesting recent tool adoption. Follow-up questions included: *How did you find out about this tool? What motivated you to switch? What alternatives did you consider and why did you choose this tool over others?*

7.3.3 Analysis of Videos and Interview Data

The videos underwent an intelligent transcription, removing pauses, filler words and doing minor grammar adjustments. Analysis included: open coding, focused coding, and thematic clustering [87].

The first two authors independently coded 3 randomly-chosen videos in the dataset through open coding. The two authors discussed the emerging themes and agreed upon a common vocabulary. Once similar codes and themes were identified across many videos with no significant discrepancies, the two coders finalized the coding scheme and shifted

to a focused coding approach. The coders independently coded another 3 randomly-chosen videos in the dataset. To ensure inter-rater reliability [359], we compared the independent coders' results from the focused coding. There was a 83.56% to 94.64% agreement level, which translated to a Cohen's Kappa score of 0.58 to 0.71 across all categories. Given the moderate to high agreement, one of the coders independently coded the remaining YouTube video data based on the agreed coding scheme. The first author also coded the interview data under this coding scheme. The two coding authors would have discussions after each interview and only identified one new theme from the interviews: maintainability.

We measured: (1) *coverage* – number of videos and interview participants who mentioned the code; and (2) *frequency* – number of times a code was mentioned across the data. Figure 7.2 shows an overview of mentions and coverage of the primary value categories.

7.3.4 Survey

To further verify our observations, we surveyed 105 creative practitioners to rate and rank each framework value.

Participants

We recruited 105 creative practitioners online: Twitter, Reddit (e.g., r/design, r/user-experience, r/cad), a large software company Slack channel, and a university. Participants were screened by email. We also reached out to the 13 interview participants and relevant personal connections. Compensation was \$5 USD or equivalent (participants belonged to 8 countries and created content for a diverse set of audiences across cultures and languages).

Participants' (52 female, 50 male, and 2 non-binary) ages were 19 to 51 (M = 28.26,

SD = 5.16). Self-reported experience was: 8 novices, 27 intermediate, 41 proficient, and 25 expert. Average time working in a creative industry was 4.48 years ($SD = 3.60$). Average time working with digital CSTs was 9.08 years ($SD = 7.63$).

Questionnaire

In addition to demographics, participants rated their values for each of the codes and framework categories on a scale of 1-5 (1="none at all", 2="a little", 3="a moderate amount", 4="a lot" and 5="a great deal"). (Questionnaires are available in the supplementary materials and were approved by our organizations' ethics review). Participants ranked the main categories with respect to each other into a seven-item ordered list.

7.4 Results

This section describes the framework on creative practitioners values for CST adoption. The framework's categories and subcategories were derived from the themes identified in the analysis of 23 videos (V01 - V23), 13 interviews (P01 - P13) and 105 survey responses. Figure 7.2 provides an overview of the 7 categories of our framework, which shows aggregate mentions and coverage, followed by the survey rankings of the categories. This section is organized in the order of the general rankings. For each category, we summarize its values in a figure (e.g., Figure 7.3) depicting subcategory mentions, coverage, and survey ratings. Average survey ratings determine the order for presenting subcategories in each subsection. This section is restricted to results. Broader reflections and ties with the literature take place in the discussion section (§5).

Zooming into these value categories, the highest-rated values were a CSTs' reliable performance (§4.3.1) and ease of use (§4.4.1). On the other hand, the CST's ability to

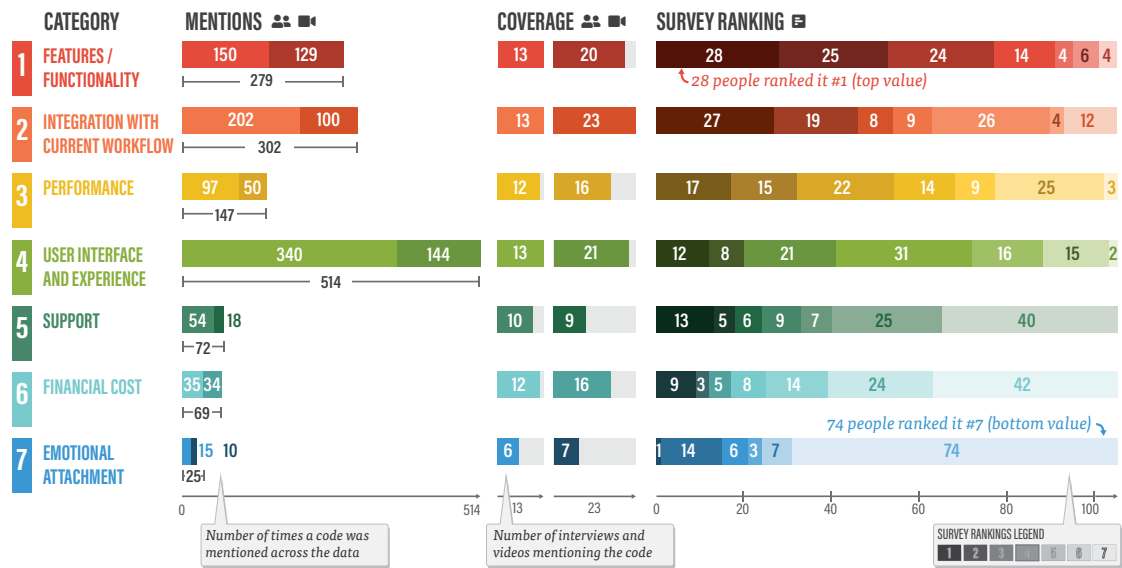


Figure 7.2: Overview of creative practitioners' value categories. Figure shows mentions, coverage and survey rankings (1: top rated to 7: lowest rating). Categories are sorted by overall rank. Our survey placed features/ functionality, integration with current workflow, and performance as top 3, while support, financial cost, and emotional attachment ranked at the bottom 3.

integrate across non-digital and digital media (§4.2.4), customizability (§4.4.7), and customer support (§4.5.3) were mentioned but not valued as much as the other subcategories (refer to Figure 1 for the overview rankings and definitions of the value framework).

7.4.1 Tools' Features and Functionality

A tool's feature is defined as a command or abstraction that achieves a particular goal. For example, this includes atomic commands such as undo and save, as well as interactive features such as the ability to draw on an sketching software. This was a frequently mentioned category across the videos and interviews, according to mentions and coverage, participants ranked CST's features as the highest value (Figure 7.3).

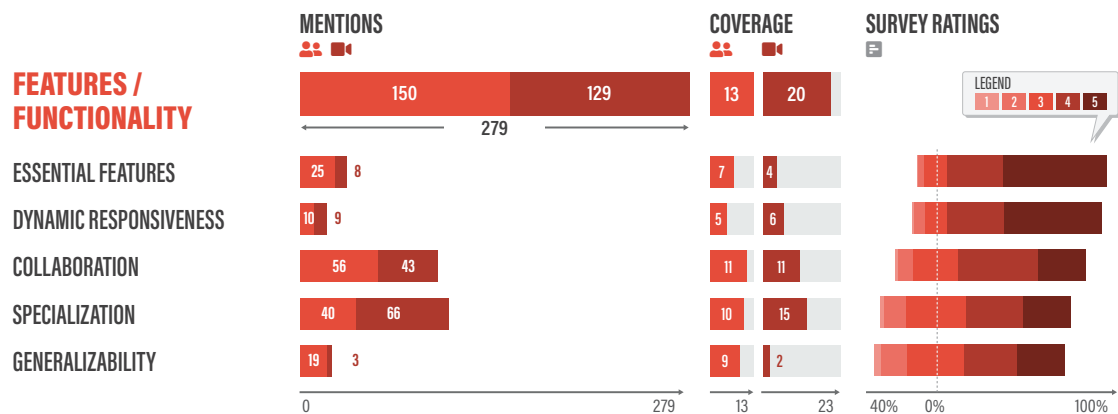


Figure 7.3: Features and Functionality values. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal). Values are sorted by survey ratings. Our survey shows essential features [§4.1.1] were the most valued, while generalizability [§4.1.5] was the least valued.

Essential Features

The set of features necessary to accomplish a particular creative task as aligned to the CST. This for instance includes the ability to type words in a word processing application. What features are deemed as essential depend on the practitioner, the application, and the domain. To determine whether the feature is essential the question is: *“if this feature is removed, can a practitioner still accomplish their most common goals?”* Practitioners valued tools with essential features over more complex CSTs loaded with more specialized, less essential features. Essential features are the target for novices when getting started in a particular creative domain’s tool. For instance, V07 described Affinity Photo as having essential photo editing features: *“Some people require the vast amounts of Photo Editing capabilities that LightRoom and Photoshop have available. I don’t need all the bells and whistles”*. While these impressions are highly subjective, “essential” implies a set of features is just enough to accomplish the majority of the tasks: *“iMovie is way too basic... Da Vinci Resolve was a nice in-between where it was just complex enough for me to make what I wanted to make”* (P03). Survey respondents rated

Essential Features as 4.33 on average (SD=0.87, Median=5).

Dynamic Responsiveness and Liveness

The ability to see feedback and effects on an object of interest as a feature is being used. Practitioners manipulate virtual objects on a regular basis, and changes are eventually reflected on their output. For example, moving a rectangle in a vector application with the mouse is often reflected live, while rendering a three-dimensional scene might take time to show the results. This feature facilitates fluid creative expression. As P03 describes, *"What makes Unity superior... it has an actual user interface that you can click around and adjust options. Whereas JavaScript that's like change the value from 60 to 50. Change windows. See what happens. You just have to play with numbers and sometimes that is not the most intuitive."* Survey respondents rated this as 4.26 on average (SD=0.94, Median=5).

Collaboration - Awareness, Feedback, Hand-off

The ability to work with others, including awareness of collaborators, feedback and communication, and hand-off to other stakeholders. V05 mentions awareness of collaborators a key value *"you'll see the avatars for each person inside the file, you can also see their cursors moving around"*. With respect to feedback and communication, V09 values Figma's collaboration features as it allows them the *"ability to jump into the design file itself... the mood board itself, and again add comments... those comments are captured in a place where actually they become actionable items"*. Furthermore, V04 makes a case for better hand-off features, *"you've got your architects, you've got your structural engineers... you're always working with a bunch of different people. Revit allows everybody to work inside of a same file, so this again eliminates chance for user error, and also eliminates a chance for clashes."* Survey respondents rated this value as

3.80 on average (SD=0.96, Median=4).

Specialization

The ability to do unique, specialized creative tasks using features with high precision and control. Contrasting this with Essential Features, Specialization features can include non-essential features. A function such as content-aware fill in Photoshop would be considered specialization, whereas adjusting the lighting of a photo would be an essential feature. V01 states: *"DaVinci Resolve is a great app for the color grading features"*. In fact, P02 described mixing DaVinci Resolve into their workflow with Adobe Premiere Pro exclusively to adjust the colour and tone of their videos despite Premiere Pro having colour adjustment capabilities. P06 mentions how *"3DS Max does rendering better than any other software tool, so I will use that for just the rendering phase"*. This was rated 3.65 on average in the survey (SD=1.04, Median=4).

Generalizability

The general- or multi purpose nature of a CST, where it can be used for various creative tasks and domains P08 illustrates how this led to choosing Figma over Tableau: *"Tableau is very specific to data visualization. And it's very useful in a design setting. It's really useful at the beginning... but it suffers a little bit when... you're trying to polish a prototype. Since not all of our projects are data visualization, we needed a more general-purpose tool. Therefore, we chose Figma where we can use it for more than just InfoVis design"*. Similarly, P06 shared: *"we use 3D Studio Max... it's like a Swiss army knife and can read lots of different forms of data, probably more so than any of our other software."* Survey respondents reported valuing general-purpose tools at an average of 3.56 on the five-point scale (SD=1.12, Median=4).

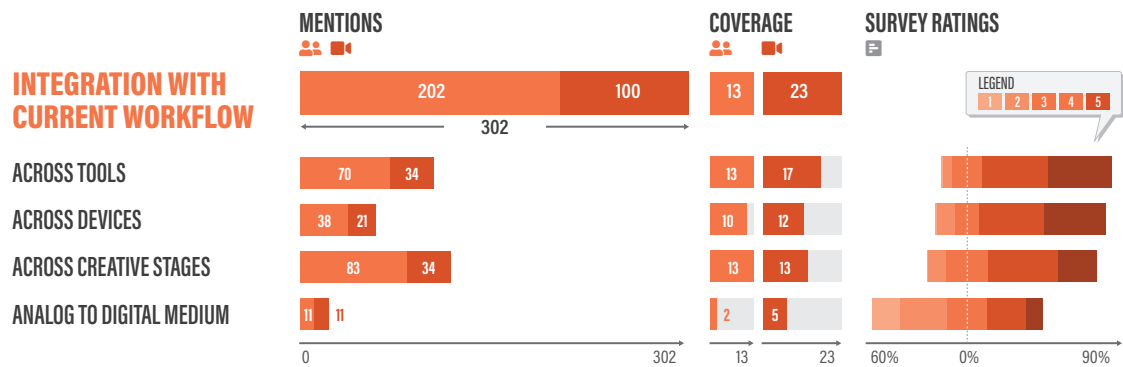


Figure 7.4: Integration with current workflow values. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal). Values are sorted by survey ratings. Our survey shows practitioners valued integration across tools the most [§4.2.1] and across analog and digital media the least [§4.2.4].

7.4.2 Integration with Existing Workflow

How well different elements work together or co-exist in an ecology of tools and devices. All interview participants and videos mentioned they value tools fitting into their creative workflow (Figure 7.4). Survey participants on average ranked this category second out of the seven primary categories

Integration Across Tools

How well the tool interconnects with other tools. This can be either by combining functions from other tools into this tool, or through plugins, exporting and importing features, etc. For instance, P06 mentions abandoning a tool because of problems with exporting and interchanging formats, *“I hate when anyone gives me data from SketchUp. Like even if they translate it to another piece of another format that I can read in my tool, it will come in very unstructured and requires a lot of rework”* V12 gives another example, *“the main feature though that i really think sets Premiere Pro apart in this category is dynamic link. This means I can seamlessly switch between Premiere Pro and After Effects and have all of my changes perfectly reflected.”*

Integration Across Devices

How well the tool supports creative work done across other devices used in creative workflow. Many practitioners talked about working across multiple devices, such as mobile devices, cameras, and computers. V13 mentions this was the major reason for adopting a tool, because *"you can use [Figma] whether you're on a Mac or a PC. So, for all those people who keep asking me if there's a Sketch alternative for PC, this is now my answer"*. Poor device integration can be cumbersome and push people to abandon CSTs. P10 describes how they *"use different pens on different devices across Apple, newer Microsoft versions, and Android versions, and they are usually incompatible across each other. This doesn't really work with me"*. Similarly, P09 describes how a mobile-only environment optimizes for working with social media: *"Even though I was taught to use the Adobe apps in school, I use the apps that are available on my iPhone... apps like Mojo, ... [Adobe] Spark, because it's easier to create graphics. So I don't need to go open a program on my computer and import all the files, export then upload again to my phone. I save time when I do everything on my phone"*. The survey rated this value 3.98 on average (SD=1.01, Median=4).

Integration Across Creative Stages

How well the tool supports different stages of a creative project such as ideation or prototyping. In some cases, this overlaps with tool integration, as import-export functionality enables easier movement across stages. P05 describes, *"You can gather feedback in there. You can do brainstorm, and all the files are inside of Figma. So it's really easy to apply whatever you're looking for within the app itself. You're able to prototype in Figma. And there's even new features coming out that let you prototype components and do developer hand-offs. And that was the biggest pull for us to switch over as a team"*. Survey participants rated this value an average of 3.74 (SD=0.96,

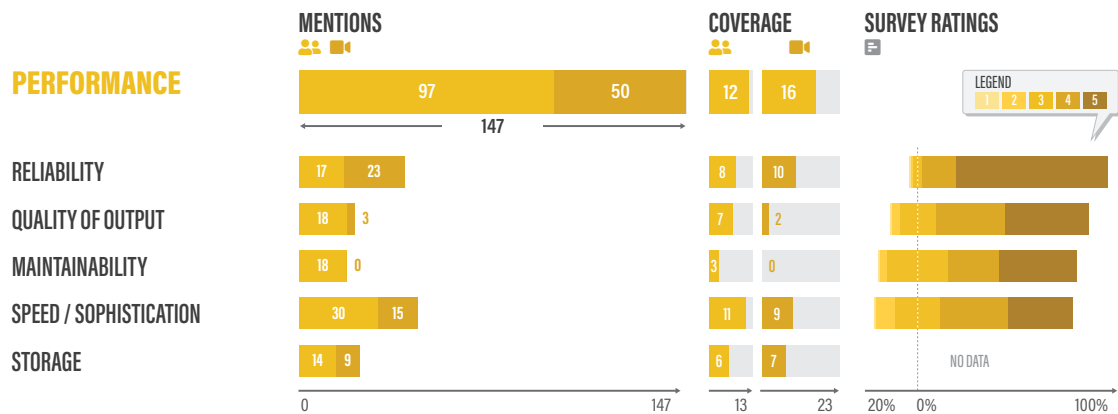


Figure 7.5: Performance values. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal). Values are sorted by survey ratings. Our survey shows practitioners valued reliability the most [§4.3.1], and storage performance the least [§4.3.5].

Median=4).

Integration Across Analog and Digital Media

How well the CST supports smooth transfer between digital and non-digital media. Creative practitioners work across both digital and analog tools such as paper, whiteboards, and pens. P10 talks about their workflow while sketching, *“Sometimes I have paper sketches that on my drawing analog tools, on my sketchbooks, that I want to digitize. I use the different versions of the Adobe Lens where you can capture them and then it converts them into a vector drawing”*. Overall, survey participants rated valuing this only a moderate amount (M=2.82, SD=1.23, Median=3).

7.4.3 Tools’ Performance

Refers to the level of consistency in execution, processing speed and storage required to produce artifacts, quality of outputs, effort required to maintain projects. 12 interviewees and 16 videos mentioned this 147 times (figure 7.5). On the survey,

performance ranked third out of the seven major categories.

Reliability

Consistency in performance, such as applications behaving as expected and not crashing. Reliability was rated as the most valued quality across all primary and secondary value categories. P03 talks about switching tools even though, *"the workflow would be the exact same. I just think that the changes come in terms of quality of life and not having the software crash on me all the time."* P07, a Creative Coder, faced similar issues, *"another deal breaker is if a tool glitches out often or is just annoying to work with, and it frequently crashes on me, I lose work and everything takes twice as long, just because the thing is unstable, then I would also definitely avoid it."* Survey respondents on average rated this a 4.67 (SD=0.70, Median=5).

Quality of Outputs

Quality, accuracy and excellence of finished creative artifacts created When discussing LaTeX vs Markdown V18 stated *"the cool thing about LaTeX is that it looks very very professional."* Similarly talking about 3D modelling V17 mentioned *"3DS Max excels in animation and also very high quality and good renders and that's why I would choose it"*. While it was not mentioned as frequently as other codes in this category across videos and participants, survey respondents rated highly valuing this (M=4.13, SD=0.91, Median=4).

Maintainability

Ease with which creative projects can be maintained on this tool over a long time period. For example, P07 describes, *"So, one thing that I usually check is the maturity of the tool ... I don't want to be maintaining the infrastructure myself. Doing all the*

system updates, etc. on your own time because the company is not paying you for this extra work". Similarly, P12 mentions the difficulty of maintaining software libraries over time, stating *"you've got to kind of think about versioning and there's breaking changes in every major release"*. Survey respondents valued maintainability reasonably high (M=3.98, SD=0.97, Median=4).

Processing Speed and Algorithm Sophistication

The time taken and ability to leverage resources for the tool to process and complete a task. Examples include preview, as well rendering time in the context of video, as highlighted by V21: *"I was using Resolve more and... you can easily feel the gain in performance, when you load clips or when you scrub through your footage, or your audio. I also measure the rendering time on each software... Resolve is just a little faster"*. Survey participants rated valuing this on average 3.88 out of 5 (Median=4, SD=1.01).

Storage

The amount of storage space required to run the tool either locally or on the cloud. V14 mentions how storage plays a role when installing the software *"the install package was only around 300MB, which is considerably smaller than AutoCAD"*. V09 reflected on concerns of cloud-only storage, *"I couldn't have files installed in my computer and work from locally, it really gave me a lot of anxiety"*. On the other hand, V20 considers cloud storage a positive, *"if I lost a hard drive or if my hard drive is broken at least my design files are safe"*. While this was mentioned 25 times across 6 interviews and 7 participants, a software bug in the survey collection prevented collecting ratings on how valuable storage was compared to the other performance values (Figure 7.5).

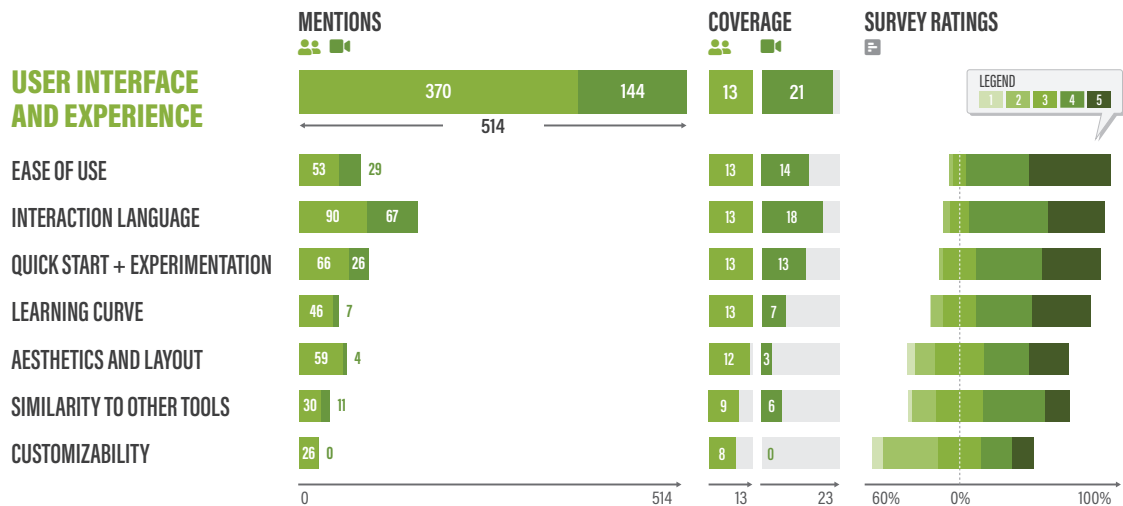


Figure 7.6: User Interface and Experience values. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal). Values are sorted by survey ratings. Our survey shows practitioners valued ease of use the most [§4.4.1], and customizability the least [§4.4.7].

7.4.4 User Interface and Experience

Components related to how people interact with their CSTs. 13 interviewees and 15 videos mentioned the interface and experience a total of 514 times (Figure 7.6). Survey participants, on average, ranked this fourth out of seven primary categories when considering the overall impact to adoption.

Ease of Use

The ease with which the user can achieve their goals effectively P13, an architect, talks about how usability factors in CST adoption, *”Rhino to me is so intuitive and I value that a lot. Even though I learned Blender and SketchUp in college, I never use them because they were never intuitive to me”*. Overall, survey respondents rated this as the second-most valuable feature across all secondary categories with an average of 4.37 (SD=0.75, Median=5).

Interaction Language

The mental model or process required to accomplish a creative goal. P04 illustrates how this plays a role in choosing which CST to adopt: *"a button is a button is a button, no matter where you see it. And because of the nature of this particular UI [referring to their design], it had a lot of common elements that got repeated over and over again. And illustrator was awful. It was like painting with a sledgehammer. We would make a change somewhere and then we'd have to find the 500 other locations where that particular element was used and make that change. so it was, it was very much an uphill battle. At one point we decided to change the font and it was not fun. Even slight color tweaks were a nightmare."* Survey participants highly valued it, rating it a 4.15 on average (SD=0.78, Median=4).

Ease of Experimentation and Startup

Ability to quickly get started, achieve results and generate variations. CSTs have different scaffolds and resources to reduce time and effort to try out new ideas, methods and prototypes or start a new project. Starting from a blank canvas can be overwhelming. To reduce this some CSTs provide walk-through tutorials, templates, examples, etc. to help get started with a project and try out the tool. P04 talks about the startup costs: *"not having to go through a million steps to get the tool up and running is definitely a deal maker"*. P03 also talks about the ability to experiment, *"Seeing it all next to each other allows me to play around, trial and error and spin up a bunch of characters really quickly"*. P05, a graphic designer, talks about how startup costs affect how their team selects CSTs: *"We like to describe it as how heavy the tool is. It's like Premiere Pro, how long does it take to boot up, get everything going. And how quick can you wound up though your load, your files, and then go through the edits that you're making. There's certain tools, like let's say Photoshop, that's really slow and clunky. And a lot of times*

we'll ditch it and do things like banner ads in Figma, just because it's so light weight". Survey respondents valued this on average 4.11 out of 5 (SD=0.80, Median=4).

Learning Curve

Time taken to become proficient using a CST skillfully P03 says, *"I was looking at Adobe Illustrator too, and I just kind of figured that the learning curve for something like that was a bit too high for what I want to pursue. So I went with Sketch since it was a little bit more simple, cause I wanted to focus on minimalist designs".* On average, survey participants rated learning curve at 3.98 (SD = 0.98, Median=4).

Aesthetics + Organization

Visual embellishment, layout and design of the tool including color, animation, imagery, and iconography. Aesthetic UI elements can create an impression on what the tool feels like (e.g., "feeling modern", or "outdated", feeling "fun", etc.). Moreover, the general layout can make a tool feel more or less "overwhelming". Illustrating its importance of aesthetics, P02 says, *"The layout and colours and design of the software itself, not the work, makes me use it. In a normal week, I stay 8 hours for 5 days in front of that software. I don't want to see ugly colors and rectangles. I don't want to feel like I'm working in a 1960s factory".* P08 also brought up the role of aesthetics, suggesting that UX tools are bound to look "more modern given that they are newer" and thus aesthetic qualities can be easily overlooked. P03 echoes similar values, *"The interface seemed really clean. I don't know, people look at the Photoshop or I guess Adobe Illustrators' interface and there's like so much stuff everywhere. It can be really overwhelming to look at, but Sketch had a very light interface that was minimalistically designed, it was pretty intuitive, get the grasp of, and I wanted to do more graphic design things and have fun."* Survey participants rated it an average of 3.55 (SD=1.13, Median=4).

Similarity of UI to Other Tools

Similarity of interface and or user experience across tools currently used or tools used in the past. Part of it may draw from consistency across tools in the same suite of applications, or as transfer from different software with overlapping functionality. P06 acknowledges: *"It's just knowing that if I pick a tool to do this, it's similar to the tool in another piece of software, by the same company that I picked to do the same thing and they're going to behave the same way"*. P12, also talks about this, *"I've adopted P5.JS for creative coding. So that's the web-based version of processing. It has a very similar syntax... it's based on Java script and Java, which makes it nice"*. Survey participants rate this an average of 3.49 (SD=1.01, Median=4).

Customizability

Extent to which the interface and functionality can be modified. For example, P11, an architect talked about *"changing the interface in AutoCAD to dark mode and organize the toolbars"* made the tool feel easier to use, while stressing that every architect has a completely different personalized interface for AutoCAD. On the other hand, P13, another architect, talks about designing a plugin that modified the functionality, *"I've designed a plugin to puncture the building with different types of windows. This allows me to express myself more creatively"*. Survey participants rated it on lowest value for this category (M=2.98, SD=1.16, Median=3).

7.4.5 Level of Support

The availability of resources that can provide assistance in navigating a tool, such as tutorials, communities of users, and customer support. 9 out of 23 videos and 11 out of 13 interviews talked about how the role of resources for learning how to use the

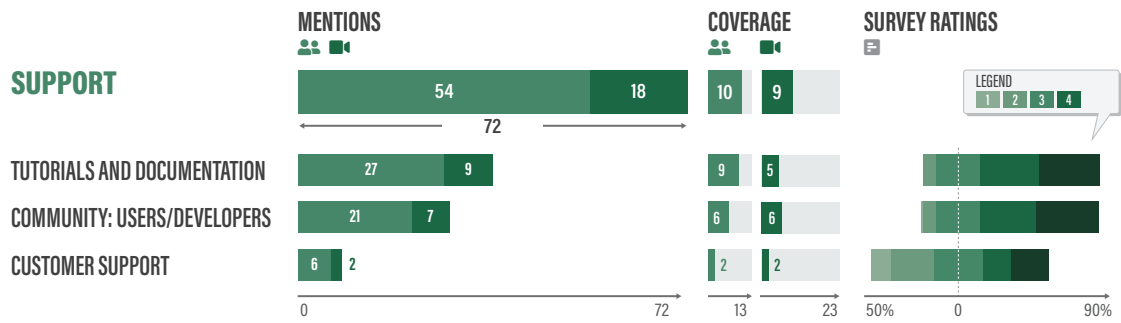


Figure 7.7: Level of support values. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal). Values are sorted by survey ratings. Our survey shows practitioners valued tutorials and documentation the most [§4.5.1], and customer support the least [§4.5.3].

tool affects their decision-making process – specifically tutorials, the community of other creative practitioners using the tool and customer support from the tool’s developers (Figure 7.7). Survey participants, on average, ranked this category as fifth out of the seven primary categories.

Tutorials and Documentation

The availability of online software learning resources such as video and blog tutorials, and developer documentation. P03 reflects on visual design for games: *“the main challenge was that for something like Sketch at the time, there weren’t as many resources or tutorials compared to something like Photoshop or Illustrator. This lack of resources... was kind of an issue and that’s why I didn’t choose it”*. Survey participants rated this an average of 3.94 (SD=0.95, Median=4).

Community of Users and Developers

The availability of support on online and offline communities including friends, collaborators, online forums. P04 reflects on their teams’ decision-making process, *“we looked into whether there was an active community of users, not so much because we*

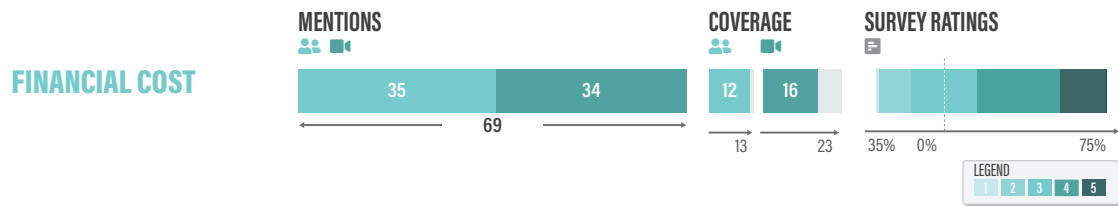


Figure 7.8: Financial costs of CSTs [§4.6]. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal).

wanted to be involved in the community or anything like that. But if other people cared about [the tool], that's a good sign to us that, there's a reason to care about it and that there will be help when problem-solving later." V07 explains, "I don't feel like they listen to the community quite as much as say, Affinity, or some other programs out there. The company that makes Procreate, they're really great about listening to their community and implement changes." V17 also shapes tool decisions based on community: "one advantage of SolidWorks is that it does have a larger user community, and so when you go and want to look for learning resources, templates, plugins, etc. it's much easier to find those for SolidWorks." Survey participants rated this an average of 3.92 (SD=0.10, Median=4).

Customer Support

The availability of support from the CSTs developers (e.g., developer representatives, live chat). P06 discusses their decision to use a rendering software, "[the tool] has a fighter pilot interface, right, like a lot of tools. I would have a very hard time adopting it, if I'm being quite honest, if we didn't have the guy who wrote the software, working with us to get all the infrastructure configured because that's a whole another game". Survey participants rated this an average of 3.11 (SD=1.30, Median=3).

7.4.6 Financial Costs

Monetary costs to use the tool individually or with collaborators, a subscription- or perpetual license-based business model, or buying one tool vs a bundle. V05 talks about how, “many people are leaving the Adobe subscription just to get a finished software because it’s a one-time purchase instead of subscribing to a platform of other tools that they might never use”. V14 talks about the value of using a tool that brings in collaborators, and other stakeholders into the same design file, like Figma: “I think that’s really cool and worth the twelve dollars, you can send out a link to anybody for free since it’s web-based. so there’s no need to pay for any sort of seats like in other prototyping tools.” Financial values were discussed only 28 times overall, 34 times across 16 videos and 35 times across 12 interviews (Figure 7.8). This category ranks fifth based on frequency of mentions and third based on coverage across the primary categories. Survey participants, on average, ranked this category as second-last out of the seven major categories.

7.4.7 Emotional Connection

Feeling a sense of happiness, identity and belonging, ethical responsibility, etc. when using the tool (Figure 7.9). For example, P10 mentions a sketching tool that “really brings a smile on your face every time you use it... I feel really happy and at

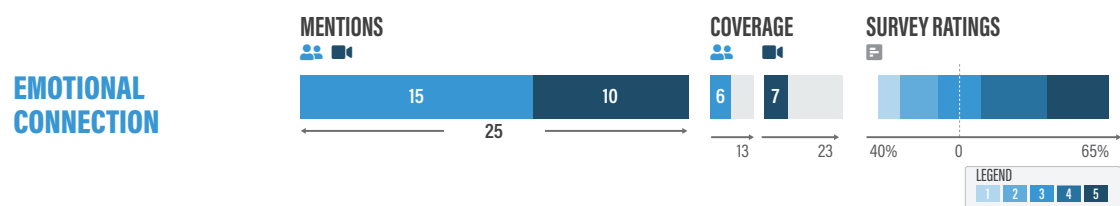


Figure 7.9: Emotional connection with CSTs [4.7]. Figure shows mentions, coverage and survey ratings (where 1: no value at all, 5: value a great deal).

home using this". P04 and P13 talk about feeling a sense of ethical responsibility when choosing a tool. P04 said, *"this company already owns 90 percent of the market share and is increasingly dictating the industry standards and pushing for all sorts of proprietary stuff. I figured they didn't need to control any more of it. So I'll take my particular, tiny little chunk of business and go elsewhere"*. P13 echoes similar concerns, stating: *"I feel really nervous doing an entire project in only one company's umbrella of applications. What if they suddenly make changes that makes it really hard to recover the work"*. Survey participants, on average, ranked this last out of the main categories (Figure 7.2).

7.4.8 Exploration and Discovery of Creativity Support Tools

While the previous categories refer to values considered when choosing to adopt CSTs, we also wanted to RQ2: how practitioners discover new CSTs and what influences their exploration process. Some creative practitioners explore tools because they are intrinsically motivated to keep learning, or are extrinsically motivated by industry trends and role changes. Often people retain tools because it is the industry standard (e.g AutoCAD in architecture). Experience with using a tool often acts as inertia that might keep people from switching. (Figure 7.10). Since the survey was to primarily verify and rank practitioners' values when choosing to adopt CSTs, and not about how they discovered their CSTs, this was not included as a question in the survey.

Personal Recommendations

Discovering CSTs through personal recommendations from friends, collaborators and social connections. P12 describes how their social circle alerts them of new tools: *"Each of my conversations with students, collaborators, friends, is almost like a radar"*. P3 also describes how social interactions lead to new discoveries: *"I was at a hackathon and I saw someone creating a poster, with that tool. I thought it was really*

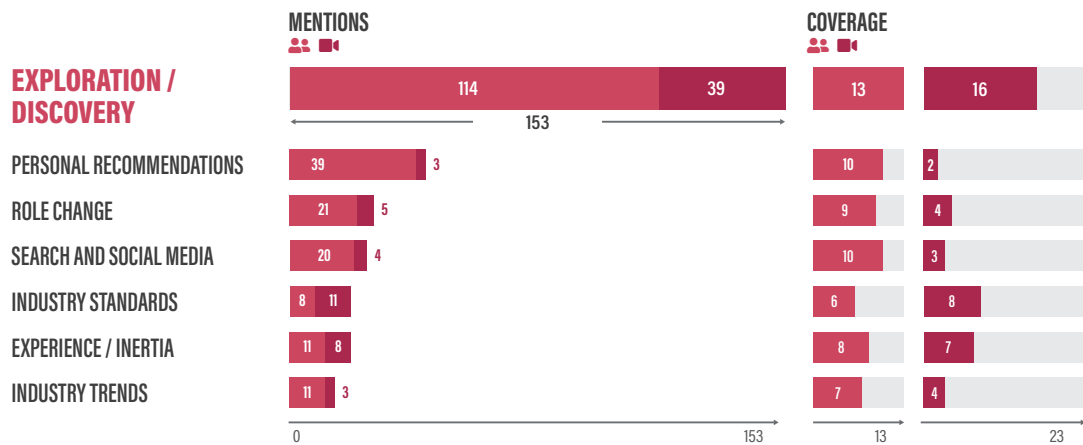


Figure 7.10: How practitioners discover and explore CSTs. Figure shows mentions and coverage in interviews and video. Values are sorted by mentions. Our survey shows practitioners’ explorations were most influenced by personal recommendations [§4.8.1] and least by industry trends[§4.8.6].

cool how fast his workflow was. The interface seemed really clean too”.

Role Change

Feeling a need to adopt new CSTs to adapt to changes in their role, organization or situation. These include role changes such as a students becoming a industry professionals, individuals changing jobs, teams shifting to remote work etc. P09 talked about their transition from student to industry: *”at university we had training in Adobe Creative Cloud, so Photoshop, Premiere and, Illustrator. So I had experience working with those software for user experience... The tools I used changed because I work with social media and all that new media now, like TikTok, Instagram, and Twitter. So sometimes I don’t use desktop software at all and just use phone apps.”* Some participants’ tool use was restricted due to organizational requirements. P01 mentioned that they *”work for the government, so I think the regulations are pretty strict here. I’m not allowed to install any tools on my laptop by myself. And, actually if I want to get a new tool, which I try to, I have to fill out a form, send it to someone and they will decide if I get the tool or if*

there's an equivalent that is considered secure that they will give me". This motivated P01 to use web applications that did not require installation.

Search and Social Media

Discovery of CSTs by searching the web or getting recommendations from people on social media, forums, or blogs. P04 mentioned *"[finding] OnShape on one of like the 3d printing forums ... Blender's huge from an online presence perspective... lots of people talking about it all the time"*.

Industry Standards

Exploration is influenced by CSTs that are standard practice in a creative industry P02 gives an example of a standard practice CST, *"Most interesting tool I've adopted is DaVinci Resolve... because it has been used in Hollywood and the entire film industry around the world for the last 20 years as the primary color grading software"*. V03 mentions that *"BIM software like Revit, Vectorworks and ArchiCAD are really the industry standard... if you want to work on structures that are larger than homes you'll need to learn BIM to secure a job at a large firm"*. While most practitioners talk about industry standards as being a motivating factor to adopt new tools, V16 reflects on how industry standards make it harder to adopt new tools at the organizational level: *"Sketch is still the industry standard and, so to respect our clients we just need to maintain that as our design tool of choice for now"*.

Experience

Exploration is influenced by psychological inertia – a tendency to maintain the status quo and avoid changes due to comfort. P04 reflected on how their team had to assess adopting new CSTs *"knowing that we already had an entire workflow that worked*

well and thousands of hours of experience in Adobe illustrator. Like, yeah, I'm not going to abandon my entire illustration workflow. I have thousands of hours in Adobe illustrator. It's a pretty big deal for me to switch... But I had to make an informed decision. I think it took us like two work days to decide that we are going to reinvent our entire workflow. We basically rebuilt everything we had done for that project up until then in a matter of a few hours. and that was enough to convince us that, yes, this [CST] is the future."

Industry Trends

Exploration is influenced by trends in the creative industry by other creative practitioners, tech advancements, etc. P06 shared: *"none of us want to be dinosaurs, so we try to stay as fresh and relevant"*. Similarly, P12 believes they *"tend to gravitate towards things that are new and exciting because, and things that are trending and industry, because those things, there's a reason why they're trending in industry... there's a reason why a lot of these different libraries and frameworks are so popular."*

7.5 Discussion: Ties Between Framework and Literature

The systematic qualitative analysis of practitioners reflecting on their values across the video, interview and survey responses, come together as a framework of practitioner values and rankings when adopting CSTs (Figure 7.2). As we discuss our findings: the practitioners' values, and their ratings of how important each value is, we draw connections to design principles and evaluation heuristics in existing fundamental relevant literature (Figure 7.11). The table in Figure 7.11 includes fundamental and relevant papers, i.e., papers with 100 citations, and overlap with two or more values in our framework. Papers overlapping a single cell are discussed in-line throughout this section, where core terms are drawn from our definitions in [§2]. These values and observations prompt

		SYSTEMS RESEARCH			CREATIVITY SUPPORT TOOLS		TECHNOLOGY ACCEPTANCE AND ADOPTION			
FRAMEWORK DIMENSION		MYERS ET AL. 2000 [48]	BLACKWELL ET AL. 2001 [9]	OLSEN 2007 [53]	RESNICK ET AL. 2005 [61]	SHNEIDERMAN 2007 [72]	AJZEN AND FISHBEIN 1975 [55]	VENKATESH AND DAVIS 2000 [79]	VENKATESH ET AL. 2003 [80]	ROGERS 2003 [62]
FEATURES/FUNCTIONALITY	Essential Features									
	Liveness and Real-Time Updates			Expressive Match						Observability
	Collaboration Features				Support Collaboration	Enable Collaboration, Provide History Keeping				
	Specialization									
	Generalizability	High Ceilings		Generality	High Ceiling and Wide Walls	High Ceiling and Wide Walls				
WORKFLOW INTEGRATION	Integration: Tools			Simplifying Interconnection	Support Open Interchange					Compatibility
	Integration: Devices									Compatibility
	Integration: Creative Stages									Compatibility
	Integration: Analog to Digital									Compatibility
PERFORMANCE	Reliability			Error Proneness						Relative Advantage
	Quality of Outputs						Subjective Norms	Perceived Usefulness		Relative Advantage
	Maintainability									Relative Advantage
	Speed + Algorithm Sophistication								Performance Expectancy	Relative Advantage
	Storage									
INTERFACE AND USER EXPERIENCE	Ease of Use		Hard mental operations	Inductive Combination	Make it as Simple as Possible			Perceived Ease of Use	Effort Expectancy	Complexity
	Interaction Language	Path of Least Resistance, Predictability	Viscosity / Fluidity	Reduce Solution Viscosity						
	Ease of Experimentation and Startup		Provisionality, Progressive Evaluation, Premature Commitment							Trial Ability
	Learning Curve	Low Threshold, Predictability	Hidden Dependencies, Role Expressiveness		Low Threshold, Black Boxes of Explorability	Low Threshold, Support Exploratory Search				
	Aesthetics and Layout		Visibility, Difuseness, Consistency							
	UI Similarity to Other Tools		Closeness of Mapping	Expressive Match						
	Customizability		Secondary Notations, Abstractions		Support Many Paths and Many Styles					
SUPPORT	Tutorials and Documentation									
	Community of Users and Developers									
	Customer Support									
Financial Cost						Cost	Cost	Cost	Cost	
Emotional Attachment							Feeling in Relation to the Achievement of an Objective			
EXPLORATION / DISCOVERY	Personal Recommendations								Social Influence	
	Role Change									
	Search + Social Media									
	Industry Standards									
	Experience / Inertia									
Industry Trends										

Figure 7.11: Creative Practitioners' values and how they fit within existing literature across systems and creativity support tools research, and technology acceptance and adoption theories. Grayed out boxes show there is no corresponding mapping.

further reflection for the wider community of systems creators, researchers, marketers, and educators, on how creative practitioners relate to their tools.

7.5.1 Features/Functionality

The tool's features were ranked the most important consideration in the survey ranking, and were an integral part of practitioners' decision-making process. This prominence was not surprising, as the CST's features propel individuals to create their content and shape their workflows. Systems and CST research focus largely on the types of features/functionality CSTs should definitely have. Generalizability was the most covered value with papers referring to it also as "High Ceiling" [342, 301, 380], "Wide Wall" [342, 380] or "Generality" [310]. CST research also emphasizes the focus on collaboration [342, 380, 429, 46]. The ability to see real-time, dynamic updates to their designs was important to practitioners. Prior literature suggests that this feature facilitates more fluid interaction is tied to "expressive match" [310] and "observability" [350].

Creative practitioners also prefer CSTs that have a unique design specialization and minimal, essential features. While most of CST research in HCI are low complexity tools that contain one or two features to accomplish one or two specific tasks [150], CST products are often complex feature-packed systems (e.g., [4, 1]). Future research should further explore creative practitioners the relation between feature preferences and CST adoption.

7.5.2 Integration with Current Workflow

During the course of a creative project, a practitioner often works across tools, devices, creative stages, and analog and digital media. Evaluating how well a CST fits into their existing ecosystem and creative practice was the second most valuable

category. Prior literature has talked about integration with other tools by supporting exportability, combined functionality, plugins etc. using terms such as "simplifying interconnection" [310], "support open interchange" [342], and "compatibility" [350]. Cross-device integration [60], ubiquitous computing [434] are their own sub-fields within HCI and a lot of CSTs aim to support this [61, 308, 30]. Most CSTs in HCI research are built to support specific creative stages, with idea generation being the most commonly supported creative process [151, 150]. Surprisingly, only few papers explore how systems might work across different stages (e.g., [246, 375]), which should be deemed as an evaluation metric in its own right. In contrast, the CST industry is creating tools that expand across multiple stages (e.g., Figma covers brainstorming, prototyping; Da Vinci Resolve covers color grading, editing, VFX) [408]. With the rapid shift to remote work, there has been an increased switch favouring digital CSTs and workflows [408]. That said, practitioners continue to work with analog and digital media [150]. Further work is needed to explore varying levels of integration. For instance, should individual tools merge into a large system that supports all integration, as done by say, Affinity Publisher incorporating photo and vector editing, or should tools remain light weight with seamless import and export across them?

7.5.3 Performance

Based on how CSTs are marketed and the focus on theoretical models of tool adoption, we expected performance to be a key value considered by creative practitioners. However, we did not anticipate seeing the many ways in which practitioners assess performance. Reflecting on the results, maintainability was mentioned by interviewees, but not reflected in the videos, perhaps because videos aim to introduce tools to viewers, rather than discuss long term project and the impact to team members and stakeholders. Many of these practices are largely left to individuals to self-organize: naming layers, commenting code,

or file management.

Rogers' theory of technology acceptance [350] refers to these as a "relative advantage". On the other hand, despite systems research valuing performance [245], performance is rarely treated as a design heuristic. This may be due to performance being largely tied to implementation rather than concepts, often falling beyond the scope of many research projects. With the progress and democratization in areas like cloud computing and computer graphics, these performance aspects will continue to evolve. Developers and researchers can use performance expectations to innovate in a more human-centered manner (e.g., via feedforward). These values can also be used by educators when choosing tools to teach, and by businesses to differentiate their products from the rest.

7.5.4 User Interface and Experience

Current practice in HCI sometimes advocates for usability evaluation as a key part of every design process. This is for good reason: usability evaluation has a significant role to play when conditions warrant it [168, 313, 167, 360]. This tie to usability is reflected by how well "Ease of Use" (row 1 in this category) corresponds with existing literature [16, 47, 310, 342, 304]. However, creative practitioners' CST adoption criteria goes beyond usability to also include interaction language, ease of experimentation and startup, learning curve, aesthetics and layout, UI similarity to other tools and customizability. Reflecting on our results, customizability was not mentioned in the videos, likely due to videos targeting first-time audiences. Moreover, highly customized software makes it inconsistent across people which can hinder other aspects such as support.

Within the value framework, interface and user experience might appear similar to features and functionality (§4.1, §5.1). Yet, features and functionality describes commands or abstractions to achieve a creative goal (e.g., liveness and collaboration

features), whereas user interface and experience refer to values related to how people interact with CSTs (e.g. ease of use, learning curve, etc.)

Systems and CST research share a focus on interaction language and learning curve referring to these as "path of least resistance" , "predictability"[301], "viscosity/fluidity" [166, 47], "solution viscosity" [310]; and "low threshold" [342, 380, 301], "hidden dependencies", "role responsiveness" [310], "black boxes of explorability" [342] and "exploratory search" [380], respectively [47, 166]. The high level of overlap is likely because as suggested by Greenberg [167], the general approach sets expectations from problem solving and shapes how practitioners think and work with tools. Aesthetics appeared to be easily overlooked, yet much research suggests it might tie to unconscious processes that shape how people feel about a particular tool [168, 304, 305]. Future research can uncover the impact of varying elements to these subcategories to CST adoption.

7.5.5 Level of Support

The field of software learning within HCI research aims to understand and scaffold the use of complex CSTs. Surprisingly, support, which often has large investment from firms, rated rather low. Past systems have helped leverage learning resources into existing tools [148, 53, 173, 284]. While we tie these elements to how tools might be adopted by creative practitioners, further work might consider how to more tightly integrate support and adoption.

7.5.6 Financial Costs

Most theories of technology acceptance and adoption include monetary cost as a parameter. Yet, practitioners consider factors beyond these theories: subscriptions vs

one-time purchases, bundles, collaboration cost, etc. Over time, considerations may change.

Based on how CSTs are marketed and the focus of theoretical models of tool adoption, we expected the monetary costs to be a key value for creative practitioners. When coding the videos and interviews we hypothesized that the low ranking must be due to self-report and social desirability bias. However, even in anonymous survey responses, participants consistently ranked it as the second-to-last valuable category. This might be due to differences in pricing across creative domains (e.g., software development CSTs are usually free while architecture and 2D vector CSTs are usually paid). Industry standards around pricing may accustom practitioners to certain prices. Investigating how practitioners perceive financial cost beyond monetary value will be beneficial for CST developers, marketers and companies. In some cases, we saw practitioners are more than willing to pay for products provided they benefit from their use compared to alternatives. We also found it interesting to see new business models appear featuring usage tiers that mix one-time purchases with smaller subscription feature sets.

7.5.7 Emotional Connection

Feeling an emotional connection and identifying with a tool was the least valued category in the framework. When coding the videos and interviews, we assumed the low frequency might be because of self-report biases when talking about emotion and the feeling of using the CST [227, 33]. The consistent low rank in survey responses might be because participants were ranking these based on how much each value influences their ability to accomplish creative goals. Values such as emotional connection and identifying with a tool, have yet to be explored in depth in literature. Nouwens and Klokmose [306] start to explore how knowledge workers have emotional connections to the applications they use. There is also a recent movement to create designs that evoke emotions to drive

positive user experiences, either viscerally, behaviorally or reflectively [305]. We believe this is under-investigated, and might be similar to how practitioners are drawn to analog tools, pens and notebooks because of how these tools make them feel. The emotional connection can be interesting to investigate on its own.

7.5.8 Exploration and Discovery

The tool discovering mechanism is via personal recommendations. Yet, factors such as role changes, search and social media, industry standards and trends, and the experience or inertia affect the exploration process. Some of these have been previously studied. For example, marketing and social science research talks about how (1) a customer's inertia or knowledge in a tool can hinder exploration and tool switching [165], (2) blogger and social media recommendations affect product purchase intentions [268], and (3) market trends of products and industries affect CST development [150, 413, 440].

These creative practitioner values illustrate that CSTs are not individually-siloed tools [434, 304, 61, 66], rather a much larger complex ecosystem of people, tools, activities, and sets of technologies.

7.6 Limitations and Future Work

This study triangulates self-report data from three diverse data sources. While YouTube video data lacks richness and details because of its audience, biases, and format, the semi-structured interviews with creative practitioners provided a rich first-hand account on their values. Long form semi-structured interviews do not provide a sense of scales, which merited verifying creative practitioners' values through the surveying the practitioners. Combining these approaches help build a deep and rich, qualitative understanding of creative practitioners' values. However, as with any methodology, there

are trade-offs: self-report data may have gaps or inconsistencies with actual observed behavior, controlled, questionnaire performance may differ from natural search behavior in unanticipated ways, valuations are done in a short amount of time and based on our textual descriptions, and CST log analysis can provide local, granular in-situ data, but lack qualitative depth, etc. One of the realities of qualitative coding is that it draws influences from authors' pre-existing knowledge when coding. While the coding was conducted independently by two authors and the inter-rater reliability was strong and significant, future quantitative and qualitative analyses of long-term CST usage both in the lab and in the wild will further expand and contextualize these initial results.

In an effort to standardize CST evaluation methods and go beyond usability as an evaluation approach [340, 342, 379, 221, 339, 325, 245, 139], HCI researchers have developed a range of quantitative methods such as the Creativity Support Index [92, 77], reflecting the whole breadth of HCI evaluation techniques [150, 380]. Our Framework brings the creative practitioner's perspective as a way to look at CSTs for long term adoption, retention and abandonment. Creativity research shows that creativity is subjective and based on the practitioners' background [108, 314, 274]. We suspect other aspects of a practitioner's background may play a role in CST adoption. In our investigation, we collected data from people across 19 different creative professions (seven across interviews as seen in Figure B.1, ten across YouTube videos as seen in Figure B.2 and fifteen across our survey). However, despite collecting background information, such as experience/ expertise, education, and demographics, this was not a well-balanced representative sample to confidently identify trends in how background affects CST adoption. We believe our framework can help future research as a set of reflective heuristics in an evaluation toolbox (such as [245]). What makes a tool successful or impactful is not a one-size-fits-all approach.

We hope that the values and observations prompt further reflection for the wider

community of CST systems creators, researchers, marketers, and educators, on how practitioners relate to their tools. For example, HCI researchers and CST developers could use this framework to identify innovation gaps and opportunities unaddressed by current CSTs, and motivate development of novel CSTs almost as values for design spaces or competitor analysis. CST marketers could use this framework to understand customers' needs and wants and market the tools accordingly. Educators can assess CSTs when choosing tools to include in their curriculum and aim for best student development. Novice and expert creative practitioners can also use this framework to reflect on their own values.

7.7 Conclusion

The rapidly evolving landscape of diverse Creativity Support Tools, makes it imperative for creative practitioners to constantly explore and decide to adopt, retain, or abandon CSTs to reach their creative potential. This chapter presents a conceptual framework of creative practitioners' values for discovery, adoption, retention and abandonment of CSTs informed by empirical observations of creative practitioners' values across 23 YouTube videos, 13 interviews and 105 survey responses. This uncovers creative practitioners' perspectives in contrast to existing developer, educator or researcher-centric angles. To encourage reflection from the various CST stakeholders, we further tie creative practitioners' values into existing design heuristics and principles in systems, CSTs, and theoretical technology adoption research. This practitioner perspective exposes that values do not revolve around individual siloed systems, rather the larger complex ecosystem of people, their activities, workflows, and sets of technologies at the tool- as well as device-level.

7.8 Acknowledgements

Chapter 7, in part, includes portions of material as it appears in "*I don't want to feel like I'm working in a 1960s factory*": *The Practitioner Perspective on Creativity Support Tool Adoption* by Srishti Palani, David Ledo, Fraser Anderson, George Fitzmaurice in Proceedings of the 2022 ACM CHI Conference on Human Factors in Computing Systems (CHI'22). The dissertation author was the primary investigator and author of this material.

Chapter 8

Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI

Creative practitioners (like designers, software developers, and architects) have started to employ Generative AI models (GenAI) to produce text, images, and assets comparable to those made by people. While HCI research explores specific GenAI models and creativity support tools, little is known about practitioners' evolving roles and workflows with GenAI models across a project's stages. This knowledge is key to guide the development of the new generation of Creativity Support Tools. We contribute to this knowledge, by employing a triangulated method to capture interviews, videos and survey responses of creative practitioners reflecting on projects they completed with GenAI. Our observations let us derive a set of factors that capture practitioners' perceived roles, challenges, benefits, and interaction patterns when creating with GenAI. Our insights serve to encourage reflection from the wider community of Creativity Support Tools and GenAI stakeholders, such as systems creators, researchers, and educators, on how to develop systems that meet the needs of creatives in human-centered ways, we propose design opportunities and priorities based on these factors.

8.1 Introduction

Throughout history, people have used various tools to innovate and create. From analog paintbrushes and pantographs to digital tools such as Illustrator, text editors, and AutoCAD. These are called Creativity Support Tools (CSTs) [381, 150]. The advent of sophisticated generative AI models (GenAI) such as ChatGPT or Midjourney brings about a paradigm shift in our relationship with these tools [62, 175, 161], as they produce outputs (e.g., text, images) in ways that can rival human creativity. Once considered a novelty for early adopters, GenAI is increasingly being used by a growing number of creative practitioners, such as professional designers, software developers, and architects, who are adding GenAI as one of the many tools in their toolboxes [200, 386, 174]. To

build and evaluate future GenAI CSTs in human-centred ways, it is essential that we study and define this change in the connection between people and a new generation of CSTs. This means examining how the role of people is evolving in creative work, how creative practitioners decide to use GenAI tools, and how their interaction patterns with such tools are changing.

Recent work in HCI has begun to investigate the building of interactive CSTs with GenAI models [449, 248, 398, 98] and characterizing interaction mechanisms [361, 262, 446] with specific individual models for individual creative tasks. However, real-world projects undertaken by creative practitioners involve using multiple tools to accomplish tasks that span multiple creative domains [?]. For instance, designing a video game might involve ideating plot and mechanics in ChatGPT, creating game visuals in Midjourney, developing the game in Unreal, and animating it using Runway ML. Similarly, writing a science-fiction novel involves writing chapters with Microsoft MS Word and creating cover art with Dall-e. It is not yet clear how practitioners conceptualize their role and interactions with Generative AI during the course of such real-world creative projects. To guide the development of new CSTs, this question needs answering.

Our work seeks to provide these answers by investigating how creatives use GenAI in their projects through a triangulation of complementary methods. We use contextual interviews (n = 10) and self-recorded videos (n = 17) of creative practitioners in 19 different creative domains (ranging from architecture and software development to UI/UX design and science-fiction writing) who had used GenAI tools to perform a variety of tasks to complete a professional project. We focused our observations at the project level instead of a specific model or task. We then contextualized and verified these factors with a larger population of creatives (n=31) by asking them to rate and rank these factors. Our selection of methods allowed us to gather information agilely to keep up with the rapid pace of integration of evolving GenAI into people's practices.

Our analysis finds that practitioners perceive their relationship with GenAI CSTs in a fundamentally different way to non-GenAI CSTs. Practitioners conceptualized their roles as project managers with a larger creative vision orchestrating information context and tasks across multiple GenAI models instead of traditional workers executing each task (e.g., painting each stroke or creating each UI component in a wireframe). Additionally, we find that while creative practitioners want creative agency across all creative stages of their project, they want interactions with GenAI to model social relationships where the models can empathize and perceive and modulate their responses based on their emotional state. Furthermore, it is interesting to note that practitioners do not want to use GenAI models to completely automate their creative workflow. Instead, they consider trade-offs such as how challenging it is to articulate goals, work across tasks, model sessions, and creative stages, and align GenAI outputs to user intents despite model stochasticity and opacity, and benefits like how it helps develop their creative goal further, streamlines their creative process, help with cold start, and serendipitous discoveries. Lastly, we find that the practitioner’s creative process is changing to focus on just project-level and artifact-level orchestrations. Based on these findings, we suggest design opportunities to build GenAI-based CSTs in more human-centered ways.

8.2 Related Work

We build on previous literature on creativity as a practice and how people work with digital tools, particularly generative AI, during the creative process.

8.2.1 The Creative Process and Creativity Support Tools

Creativity is generally defined as generating novel and appropriate ideas, processes, or solutions [393], and a fundamentally non-linear and iterative process. For Wallace

[428], creativity involves knowledge acquisition, unconscious information processing, idea generation, and evaluation of the idea. In contrast, Guilford models creativity as divergent and convergent thinking, which is seen as the capacity to generate multiple solutions to a problem and effectively evaluate these solutions [394].

Computers have become common tools that can augment creatives [379, 380] and streamline the creative process. CSTs can support the creative process at different stages, such as making discoveries or inventions from information gathering [271, 317], hypothesis and idea generation [383], and initial production [116, 148], to refinement [228], validation [149], and dissemination [380, 150]. To guide the design of CSTs, HCI research has proposed design principles such as the importance of guiding users toward correct actions (*Path of Least Resistance*), aligning with users' mental models (*Predictability*), and providing tools with *Low Thresholds, High Ceilings, and Wide Walls* to cater to both novices and experts, exploratory search, collaboration, rich history-keeping, etc. [342, 301, 166, 47, 245, 310, 380]. On the other hand, HCI research has proposed quantitative and qualitative approaches to evaluate the usefulness of CSTs. A quantitative measure is the Creativity Support Index [77, 92], a general-purpose survey to gauge the effectiveness of a CST. Other methods include co-design workshops [127], physiological responses (e.g., galvanic skin responses, EEG) [78], and self-report in post-study reflective think-aloud and surveys [364, 431]. As CSTs with GenAI features grow in popularity and the relationship between creative practitioners and their tools evolves, it is important to verify if these design principles and evaluation methods still hold true and if not propose new GenAI CST design and evaluation guidelines.

8.2.2 Interacting with GenAI During Creative Processes

The landscape of CSTs is undergoing rapid evolution with the introduction of Large Language and Diffusion models (GenAI), which are trained on large data sets and

can generate text (e.g., [62, 58]), imagery (e.g., [12, 14]), or other media [13, 10, 11]. Notably, GenAI outputs can rival that of human-generated content [175, 62]. GenAI have successfully passed creativity tests such as the Alternate Use Test and Torrance Test [175, 377], highlighting their potential for creativity, and have been used in tasks such as generating ideas for startups [161] and short stories [126]. GenAI is already being applied to support creative tasks, including music composition [267], visual art design [97, 96], and writing scientific articles [28, 319], and novels [247, 454, 155]. In these applications, GenAIs can be a useful tool to increase creativity, even when (or because) they can provide unintended outputs.

To start conceptualizing how creatives might interact with GenAI during their processes, a new research area called Mixed-Initiative Co-Creativity (MI-CC) is emerging [257, 118]. These MI-CC applications can be placed on a continuum, describing the degree to which human and computational agents take the initiative in the creative process [298, 343]. People often employ metaphors and analogies to understand and explain the nuances of how they conceptualize GenAI and their interactions with them [174, 298, 382]. In this regard, Shneiderman's framework suggests that GenAIs can be viewed as intelligent agents, collaborative teammates, social robots, or even as supertools, tele-bots, control centers, or active appliances [382].

To understand how people interact with GenAI during the course of the creative process, Spoto and Oleynik [392] analyze approx. 70 CSTs ranging from game creation to manufacturing design systems to map the actions taken by the two agents into seven types: *ideate*, *constrain*, *produce*, *suggest*, *select*, *assess*, *adapt*. Muller et al. [299] look into including GenAI agents into this framework and propose additional AI-specific actions: *learn*, *ideate*, *constrain*, *produce*, *suggest*, *select (only 1)*, *curate (many)*, *assess*, *adapt*, *assemble*, and *wait*. Our work adds to the above research by observing how practitioners across multiple domains engage with multiple GenAI during their real-

world projects to present 1) a set of factors that influence creatives' decision to use and interact with GenAI; and 2) a set of design opportunities and priorities to help ground the decision-making process of stakeholders of GenAI-fueled CSTs and lead to more human- and creative-centered solutions.

8.3 Method

To address our research questions around the role of people in human-AI interactions during creative work, how creative practitioners decide whether to use GenAI tools and how their interaction patterns with such tools might be evolving, we follow a two-fold approach:

1. Empirical Observations. We conducted 10 semi-structured interviews¹, and supplemented this data with 17 videos of creatives providing first-hand accounts of their workflows. This helped us quickly capture a wide range of creatives and stay up-to-date with the rapidly evolving landscape of how they incorporate GenAI into their workflows. From these observations, we distill a set of emergent factors outlining the interactions of creatives and their use of GenAI.

2. Survey. To verify and contextualize our empirical observations with a larger population of creatives (n=31), where we asked them to rate and rank the different emergent factors.

¹Our Interview guide and questionnaires are available in the Appendix B.1. All our studies were approved by our organizations' ethics review.

8.3.1 Semi-Structured Interviews

Participants

We chose purposeful sampling [48] to recruit 10 participants (7 male, 3 female), mixing direct contacts and recruitment through social media channels (e.g., Slack, Yammer, and Teams) at a large software company and public universities. Participants came from different domains, ages, organizations, gender, race, location, and cultures. They spanned six creative fields including graphic design, science fiction writing, UI/UX design, software development, and scientific research (see Appendix ??). They used different models to achieve their creative goals, including Midjourney, ChatGPT, Codex, and Stable Diffusion. While we reached data saturation by the seventh interview, we continued interviews to reach a larger coverage of professions/roles, as well as different levels of professional expertise. All participants were experts in their chosen creative domains, with professional experience ranging from 3 to 27 years (mean = 13.8 years, sd = 5.6 years), and all recently started to use GenAI in their work. Four reported having a doctoral degree, four a master's degree, and one a Bachelor's degree. Only one participant reported having no formal educational background in a creative field.

Procedure

Before the interview, participants responded to a standard demographic questionnaire and collected self-reported expertise in using GenAI, how much agency and empathy they would like at different creative stages of the project, and how they might describe their relationship with GenAI. To ground the discussion, we asked participants to recall the latest project they worked on using GenAI, describe it, and explain how they used GenAI to accomplish their goals. We asked them what actions they performed during their workflow and how they coordinated these operations across work sessions, creative

stages, and applications. To identify places of opportunity for technology interventions, we asked them to map these activities on a diagram of the British Design Council's Double Diamond design process model (Figure 8.1).

8.3.2 YouTube Videos

To expand and enrich the subject of our observations, we chose YouTube's² comprehensive public videos as a starting point, as they include diverse creatives sharing knowledge through vlogs, tutorials, and personal experience. We searched using keywords such as "*I built ...*", "*I made...*", or "*creative workflow ...*" with generative AI technology keywords like "*AI*", "*ChatGPT*", "*Midjourney*", etc. We selected videos of practitioners who had completed a full project showing more than one GenAI and covered more than one step in the creative process. We ensured that videos covered a diversity of domains, such as graphic design, architecture, video production, creative and scientific writing, software development, and UI/UX design (see Appendix ??). We collected material past data saturation in case a particular domain yielded new findings.

8.3.3 Analysis of Videos and Interview Data

Our analysis included open coding, focused coding, and thematic clustering [87]. Two of the authors independently coded two randomly chosen videos through open coding. They discussed emerging themes and agreed on a common vocabulary. Once they identified similar codes and themes with no significant discrepancies, they finalized the coding scheme and shifted to a focused coding approach. To ensure inter-rater reliability [359], we compared the coding results. There was a 79.07% to 95.83% agreement level across all code categories. Given this moderate to high agreement, the first author independently coded the remaining YouTube video data based on the coding scheme. We

²<https://www.youtube.com>

report on: (1) coverage – the number of videos and interview participants who mentioned the code; and (2) frequency – the number of times a code appears in the participant’s responses.

8.3.4 Survey

To verify our observations with a larger sample of creatives across more domains, we surveyed 31 additional practitioners to add their insights and rate and rank each category in the initial behavior patterns.

Participants

We recruited 31 creatives (12 female, 15 male, and 4 non-binary) through Twitter, LinkedIn, Reddit (r/design, r/userexperience, r/cad, etc.), and internal social networks at a large software company and three public universities. Participation was incentivized as a \$1 USD donation to the local Humane Society for each complete survey response. Participants reported having different levels of experience in their creative fields: four of them reported 3-5 years, four between 6-10 years, and five more than 10 years. When asked about how frequently they interact with GenAI for their creative projects, five reported using it *Frequently (several times per month)*, seven *Very frequently (multiple times a week or more)*, and one *Occasionally (a few times a year)*. One participant had no formal education in the creativity practice and was self-taught, one had a bachelor’s degree, seven had a master’s degree, and four had a Ph.D. degree.

Questionnaire

We asked participants to list the software and GenAI tools they use as part of their process and to describe how they use them. Reflecting on their projects, participants rated how much they value each factor derived from our initial observations (interviews +

videos) on a scale of 1-5. We added any interaction factors that they performed that were not included in the list we provided.

8.4 Results

Our results are derived from the factors identified in the analysis of videos (V01 - V17), interviews (P01 - P10), and survey responses (n = 31) from creatives using more than one GenAI tool to complete a project.

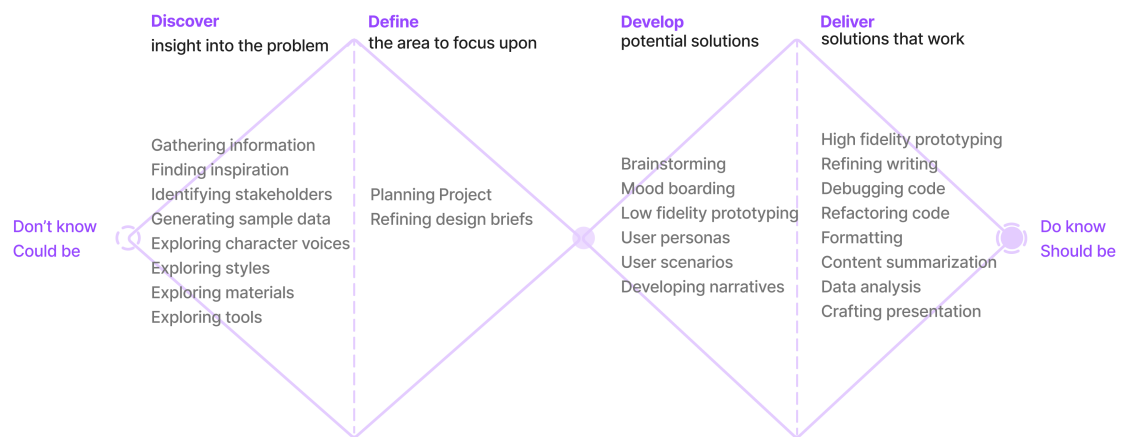


Figure 8.1: Creative practitioners perform a range of activities with GenAI over the course of the creative process. Here, we see these activities contextualized against the British Design Council's Design Double Diamond model.

8.4.1 Perceived Roles When Working with Generative AI

When talking about how they work with GenAI, the creatives we observed or interviewed often used metaphors and analogies to describe how they conceptualized their interactions with the models. Previously, metaphors have been used to conceptualize the algorithmic abilities and limitations of models [382, 184]. Therefore, to understand how creative practice and the role of people and AI are evolving, we thematically analyzed the metaphors used to describe person and AI roles in creative practice. To get an overview

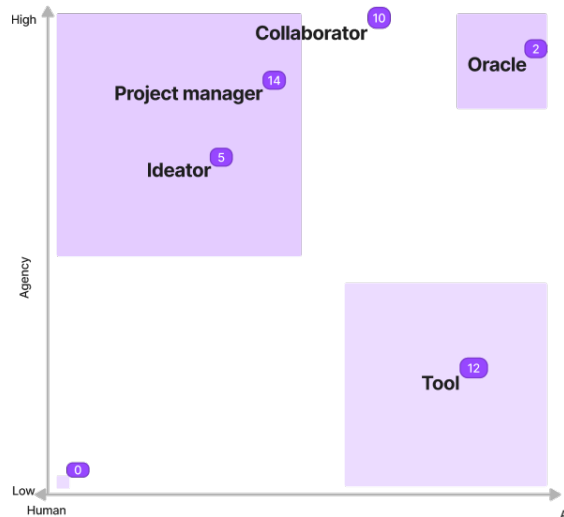


Figure 8.2: Practitioner-mentioned metaphors for perceived roles, categorized by creative agency and perspective (people vs. AI), highlighting human-centric roles with higher agency and AI-centric roles as tools. The badge indicates the number of practitioners who mentioned each metaphor. Each quadrant size is proportional to the number of operations according to their perceived or preferred degree of agencies.

of how practitioners are beginning to think of these roles, we heuristically place how many practitioners mentioned each metaphor along two axes: (1) whether it is more from the perspective of the person or the AI model; and (2) how much agency in the creative process does the agent have in this role (see Figure 8.2). Overall, we find that practitioners perceived their roles to have more agency than the AI models’ – being conceptualized in roles such as project manager, collaborator, and main ideator, while the AI’s role is perceived mostly as a tool.

Out of 27 practitioners (10 interviews, 17 videos), 14 mentioned identifying with the role of a **Project Manager** when interacting with GenAI during the creative process. This role can be “information manager”, “people manager”, “crisis manager”, “task manager”, or “managing memory”. For example, P01 said³, “*I approach this as an orchestrator where you just have to send queries to [the] large language model you’re using. And then as soon as it provides the output, you evaluate it, you try it, you refine, to find the*

³We paraphrase all participants’ quotes in a way that makes them concise while preserving their meaning.

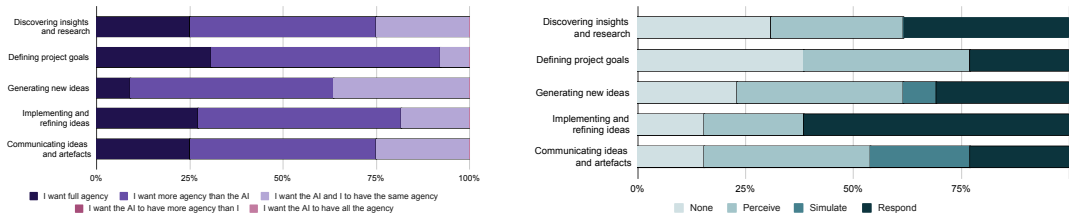
optimal solution.” Similarly, V04 said, *”I look at it as an extremely useful and obedient assistant that’s going to help you at any time.”* P04 said, *”I am managing [its] context across sessions and the project.”*

12 practitioners mentioned thinking of GenAI as just a **tool**. For example, P02 said *”I see Midjourney as a scratchboard, to do some early explorations.”* Along the same dimension, P03 called *”AI as high-variance search”* and P06 said *”AI is an interactive encyclopedia written collaboratively”*.

10 practitioners mentioned thinking of the model as a **collaborator**. For example, P06 *”I approached it similarly to how I would collaborate with someone. I could just go in with something half-baked and know that the system would ask me to clarify [if it needs it]. It can offer a suggestion or a follow-up question, which was great. It really felt like a partnership [that] I found interesting and useful.”*

Five mentioned thinking of their primary role as the main **ideator**. For example, P05 said *”I don’t use the LLM as a brainstorming partner. I usually have an idea.”* Similarly, V06 said, *”I want to come up with something driven by a strong idea.”* Two mentioned how they conceptualized the role of the model to be that of an *”all-knowing oracle.”* P05 said, *”I like to treat them as all-knowing, all-capable oracles: my task is to find a way for them to do what I’m interested in.”*

However, **many practitioners referred to both their and the GenAI’s roles changing over the course of the creative process.** To delve deeper into how the roles and the dynamics between creatives and GenAI models shift during the creative process, we asked practitioners to rate the level of **agency** they would like them vs. the GenAI to have across the different creative stages. We defined creative agency as the extent to which a person can make choices based on one’s values, beliefs, and preferences, without being unduly influenced by external factors or pressures, while having control over their creative process and outcomes. We defined the different creative stages as



(a) Across all stages, practitioners wanted more agency than GenAI.

(b) While practitioners generally want GenAI to *perceive* their emotions, they wanted the GenAI model to also *respond* in an empathetic manner during the creative stage of implementing and refining ideas.

Figure 8.3: Social dynamics such as agency and empathy between the creative practitioner and AI models shift across the creative process

stated in the British Design Council’s Design Double Diamond model. Across all these stages, participants wanted to have more agency than the GenAI model (see Figure 8.3a).

As many of these roles in the creative practice have been enacted by people, it is conceivable that future GenAI models may have the ability to perceive, respond, and simulate person-like behaviors to better fit and help in the creative process. We asked the creatives we had access to rate whether they would find it worthwhile to interact with an AI agent with **empathetic capabilities**(i.e. perceive and modulate their response based on a person’s context) to help them achieve their creative goals. We defined the different levels of empathy as having the ability to perceive, respond, and simulate emotions to help better achieve certain goals [183]. *Perceive* is defined as AI agents recognizing the user’s emotions via input cues such as tone, word choice, and context. *Respond*: In the future, agents could adjust their behavior and responses based on the user’s emotions, expressing empathy, or offering support. *Simulate*: Future agents could generate and express their own simulated states to enhance interactions with users. On average, participants wanted the GenAI agent to be able to perceive their emotions but not respond empathetically during the creative stages of discovering insights and researching, defining project goals, generating ideas, and communicating ideas and

artifacts. And when they were implementing and refining ideas they wanted the GenAI to not only perceive but also *respond* in an empathetic manner (see Figure 8.3b). A two-way ANOVA and Tukey’s HSD post-hoc analysis show that this is a significant difference ($F(3, 32) = 5.38, p = 0.01$).

8.4.2 Trade-offs: Benefits and Challenges of Creating with Generative AI

To understand why practitioners use GenAI for some tasks and not others, we investigate the trade-offs of potential benefits and the challenges that practitioners might encounter.

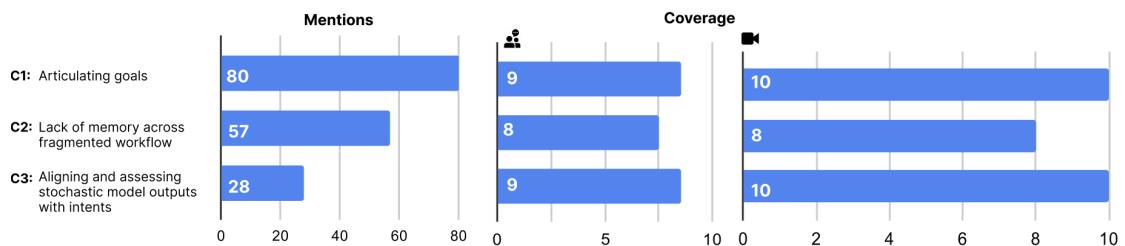


Figure 8.4: Challenges derived from thematic analysis across videos, interviews, and survey responses showing how frequently something was mentioned, and the coverage across sources.

Challenges

See Figure 8.4 for the frequency and coverage at which each challenge was mentioned:

C1: Articulating creative goals can be difficult. This is because it is **hard to articulate tacit knowledge** such as style and expertise. GenAI requires clear instructions, yet capturing subjective nuances can be difficult, leading to a potential gap between the creator’s intention and the generated output. P05 shared, *”I don’t feel confident describing my style, because it’s so ambiguous. The next best thing for me is to write things myself and ask the model to improve [them].”* Adding to this, V01, an architect, talks about their

efforts to imbibe some of their real-world knowledge and experiences into ChatGPT *"The idea is a series of buildings [with certain properties]. I tried ChatGPT for suggestions, but it gave me really generic answers. I was looking for something that I had seen while being at a particular place, and because ChatGPT doesn't have the personal experience of going there, it has a hard time giving the answers I wanted."* Without a **specialized vocabulary**, GenAI can struggle to grasp the nuances of the creative domain, resulting in outputs that miss important contextual details. For example, P02, an artist, said, *"at times, I didn't have the vocabulary to ask the model to help me. I think your background knowledge matters: someone with an art history background knows how to prompt a specific style, unlike someone who doesn't."*

C2: Lack of memory across fragmented workflow. Almost all practitioners used multiple different apps, models, and tools throughout their creative workflow. Therefore, repetitive work done with different GenAI applications and tools was a major reported challenge. P07 said, *"There aren't models that can help you with all aspects of the creative process. I have to repeat prompts or cut, copy, or paste the same thing repeatedly across models."* GenAI often lacks memory between prompts or work sessions. This leads to creators repeatedly explaining context and rephrasing instructions, which not only disrupts the creative flow but can also result in inconsistent outputs over time. P06 said *"The most frustrating thing is just knowing that if I want to generate a set of three illustrations, it's going to be tricky to get them feeling as part of the same set"*.

C3: Aligning and Assessing Stochastic Model Outputs With Intent. LLMs are probabilistic: they can produce a different output given the same input. This can be challenging as outputs can be inaccurate, and it can be hard to assess the alignment and factualness of their outputs, especially when dealing with complex iterative creative tasks. For example, P02 *"I was prescriptive in my prompt, and I thought I nailed it. But the model never did, and it still doesn't. That drives me crazy and keeps me surprised,*

delighted, and sometimes annoyed. On the other hand, some practitioners even scoped their projects to work better with the models. V14 "I chose to base my fantasy world on Greek mythology because Chat GPT would already have a large base of information on it. I also provided text vague in style, similar to how mythology stories are told, so I thought the style would fit it pretty well."

Benefits

see Figure 8.5 for the frequency and coverage at which each challenge was mentioned:

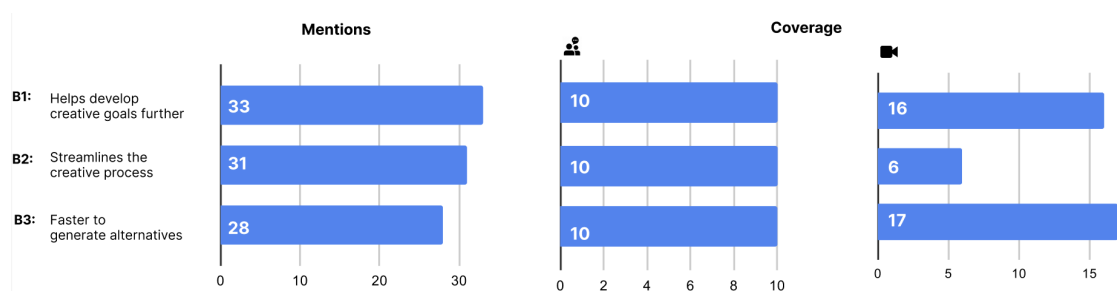


Figure 8.5: Benefits derived from thematic analysis across videos, interviews, and survey responses showing how frequently something was mentioned, the coverage across sources

B1: Helps with cold start & further develops creative goals. Starting a creative project from scratch can be challenging due to the absence of initially clear ideas or concepts. Generative AI can help overcome this "cold start" problem by generating initial concepts or sketches that serve as inspiration. P02, an artist, talks about how GenAI helped them get started with coding, *"I'm not really a coder per se, but I gave it a shot and was able to get it to work!"* P08 said, *"As a designer, those blank canvases always scared me, but generative AI provides me with a range of starting points, like a spark in the dark."* It can also help further develop a goal. P06 shares their experience as a designer developing their creative goals with GenAI, *"I could enter with a partially formed idea, confident that the system would either prompt me for clarification. The prompt structure facilitated this. It consistently offered suggestions and follow-up questions, which I found*

extremely helpful. This approach made me feel like I wasn't wasting anyone's time". Similarly, V06 discussed how in their project interacting with the model, *"exposed this entire opportunity that I hadn't previously considered."*

B2: Streamlines the creative process by eliminating or automating some steps in the workflow. GenAI automates repetitive and time-consuming tasks, such as generating initial sketches, mockups, or design layouts. This automation frees creators to focus on tasks that require their unique creative insights and decision-making. P02 said, *"I could generate any of this in Photoshop or Illustrator, but it. So, the fact that it was able to render these things on an aesthetic level that was exceeding my bar or at my bar, and doing it in an instant was mind-boggling. What it did is it gave me time to dabble in other areas."*

B3: Accelerates the generation of alternatives. Generating creative alternatives traditionally might take a significant amount of time, especially when exploring diverse possibilities. Generative AI accelerates this process by effortlessly producing a range of alternatives. This speed of doing things faster than manually allows creators to explore a larger design space and consider numerous options. V01 also said *"Compared to a traditional process, you can generate a lot more ideas. [Then] you can very quickly go through them to see which one works, which one doesn't, or which one you want to tweak."* Similarly, architect V06 said *"Midjourney allows me to explore more options than I could previously"*

Balancing Benefits and Challenges

While GenAI is inextricably integrating into creative processes, professionals do not entrust their entire projects to it. Practitioners often mentioned weighing the benefits and challenges when deciding whether to use GenAI for a task. For example, P05 said *"If I have the expertise, it might be quicker and easier to do it myself. However, if I'm not*

an expert, the model could help me get started and discover possibilities I didn't know. However, if I'm not an expert and the model doesn't have enough training data, then it might hallucinate, and I wouldn't know." Similarly, P10 talks about creative control, *"If I just want to get something done, I'll be concerned about how accurately it does it. If I don't influence it much with story direction or characters, it often leads to serendipitous discoveries about the story."*

8.4.3 Evolving the Creative Process: Project- and Artifact-Level Orchestrations

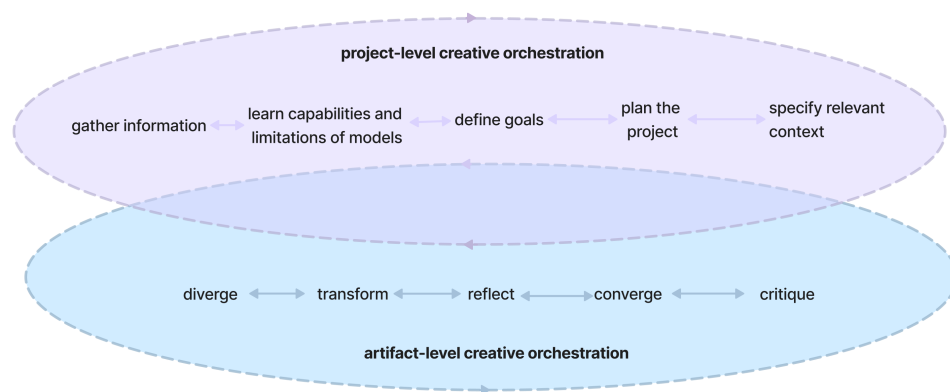


Figure 8.6: The creative process is evolving to be iterations between project- and artifact-level orchestrations.

As we talked to and observed practitioners about how they interact with GenAI models during their creative process, we realized that they might be reframing how they interact with machines/systems during their process. We capture this as a set of factors that synthesize the emerging set of interaction patterns we observed during our studies. Figure 8.7 shows the factors derived from the thematic analysis of videos, interviews, and

survey responses by creative professionals using more than one GenAI in their work. At a high level, interactions with generative AI during the creative process are of two types: *Project-level orchestrations* and *Artifact-level orchestrations*. Figure 8.6 illustrates. At each of these two types of interaction, we define each factor and report on the number of times it was mentioned/observed across all the practitioners, coverage across the 10 interviews and 17 video practitioners, and 31 survey ratings of how often they use each interaction pattern. We reflect and tie these factors with the prior literature in Section 6.7.

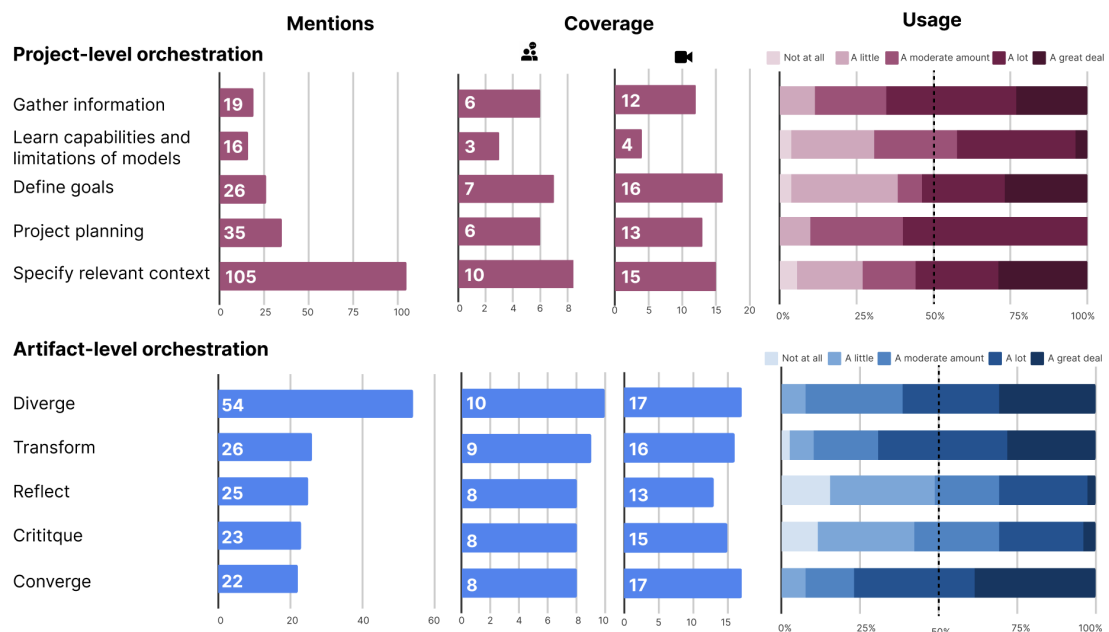


Figure 8.7: Creative process is evolving to be iterations between project- and artifact-level orchestrations. These interaction factors were derived from thematic analysis across videos, interviews, and survey responses showing how frequently something was mentioned, the coverage across sources and how often practitioners said they use each pattern.

Project-level orchestration

These factors refer to activities that involve the coordination and management of tasks, information resources, and processes within a project to achieve the creative goals of practitioners. These include gathering information, learning capabilities and

limitations of the models, defining goals, planning the project, specifying relevant context, and directing the different components of a project to ensure that they work together effectively.

P1: Gather information. Gathering relevant information for the project includes searching for simple or complex information. These can be well-defined asks or more exploratory. As P01, a researcher who uses data science, describes: *"I ask for additional questions to span my knowledge. Like statistical tests that I didn't know of."* Similarly, V10, a content creator, says *"I'm starting off by asking ChatGPT for 100 YouTube channel niches."* This information search and gathering can include more complex explorations too. For example, P10 (a sci-fi author) describes: *"I was researching an artifact I had seen at the British Museum that I wanted to write about for a story. I couldn't figure out what that artifact was, so I asked ChatGPT. It not only provided answers, but also corrected my initial misunderstanding about its origin."*

P2: Learning the capabilities and limitations of models. Each GenAI model has unique characteristics, such as the types of content they can produce, the different tasks for which they are optimized and their biases. Understanding these factors is crucial to align AI-generated content with practitioner and project objectives. Practitioners reported doing this by asking the model to suggest capabilities and limitations, asking it to suggest follow-up prompts and even repeat prompting to explore the range of its capabilities. For example, P06 discusses starting with sample prompts, *"Given my prompt, I asked it to generate follow-up prompts. Then, grabbed a bunch of its prompts and tweaked them slightly to get the results that I'm happy with"*. Similarly, P02 said, *"I took some sample prompts from a website and fed some to Dall-E to see how they would look."* V14 tried the repeat prompting method, *"I regenerated prompts multiple times to get something I liked and understand how to work with this."*

P3: Define goals. Practitioners often start the project with ill-defined, fuzzy creative

goals and define and scope their goals over the course of their project. It was interesting to see how practitioners interacted with GenAI models to scope their goals to be more well-defined. P06 shared their experience going in with a very fuzzy goal, *"I felt like I could just go in with something really half baked and know that the system would ask me to clarify and this would help me define my goal better."* Similarly, P05 shares how they used it to further define their project goal, *"The model helped work out a well-scoped goal and plan, and helped me figure out tasks, resources, timing, and success metrics."*

P4: Project planning. Practitioners talked about how they planned and managed their project using a central document in which they developed a comprehensive plan outlining the tasks, timelines, resources, and milestones of the project. For example, sci-fi author P10 shared, *"When I'm writing a book and I use a document titled "Scratchpad" for drafting and preserving content for potential use. This is the document where I put all of the to-dos, the drafts generated by me, or the models. I also use a "Hero's Journey" document to track progress and plot elements. I can use ChatGPT to break down TODOs into effective steps and plans."* Similarly, P03 said, *"There's a bazillion thing in the pipeline: choose this, then choose this, then write a story, then make a plan for writing the chapters, then write. Task decomposition is like prompt engineering on steroids because you're putting together all the pieces."*

The difference between project planning and goal defining lies in their purposes. Defining goals establishes the project's strategic direction and vision, guiding decisions by answering, "What are we trying to achieve?" In contrast, project planning translates these goals into actionable tasks, timelines, and resource allocations, addressing the "How will we achieve the goals?" and detailing necessary tasks and activities.

P5: Specify relevant context. This type of orchestration involves providing specific contextual information to guide GenAI in producing relevant and aligned content. As V04 described, *"Every time you prompt, you're giving it clues to get closer to the idea*

you have in your head. The more words that you use, the closer the image can get to what you imagine. They're only going to do exactly what you tell them to and not much more. So, the better the directions you give it, the better the results you're going to get out of it."

Specifying the relevant context for a task can involve specifying styles, project constraints, or guidelines for the output. For example, web designer V02 shares how they specify personal style or others' styles as keywords or examples in their prompts, *"If you want better outputs, you can provide it the kind of art style that you want, or how you want the design to look and feel, whether to represent a certain brand."* Similarly, V01 shares why they think specifying project constraints is important, *"The constraints set up the guidelines for your project... These limitations can give you clues as to what to focus on and what to neglect. And without them, you can get lost in the process."* Other guidelines can be used to specify the type of output, like its structure and the level of diversity or complexity it has to have. For example, P02 specified: *"each response should be no longer than 125 characters long and include an affiliated gender, age, location, and occupation. Please bullet point each response."* V09 specified: *"I'd like to make a short film that's between 60 and 90 seconds long."* P05 specified, *"Most of the time, I want to be as precise and as focused as possible, so I'll just use very low temperature to ensure that the model stays on tasks and follows instructions as closely as possible."*

Artifact-level orchestration

These factors refer to activities that involve the coordination of processes and resources to create a refined (creative or design) artifact. This involves a dynamic interplay between the practitioner and the model, encompassing interactions of divergence, transformation, reflection, critique, and eventual convergence.

A1: Diverge. One of the most popular uses of GenAI is to generate multiple

alternatives and then combine parts or some of these into the next iteration of the creative artifact. This is similar to brainstorming and ideation – exploring a wide range of concepts, perspectives, and possibilities related to the artifact they aim to create. P06 shared how they approach divergence: *“I would ask it to tell me alternatives. Push it to think about it in a different way. I always ask – any other ideas? And it would always come back with something.”*

A2: Transform. Once a pool of ideas is established, the next phase involves taking selected concepts and expanding upon them. This might involve iterating on initial sketches, prototypes, or drafts, refining and evolving the artifact’s form and content. The practitioner and the model collaborate to experiment with different variations. Practitioners would tweak the prompts and the outputs to expand previous work using keywords, few shot prompting, and even by specifying personas for the model. V02 shared how they expanded their idea set using Midjourney, *“I’m really liking this very first design here at the very top. So we’re going to upscale our V1 here, and we’re going to wait for that to come back.”* P04 a researcher and software developer said, *“I use the GPT4 in playground or in Visual Studio. If I can spot an error in a suggestion, I tell it to refine it. I usually tell it to generate three options, and then I tell it that I like this, or I don’t like this portion. Can you change that?”*

A3: Reflect. Practitioners often prompted the model to see what it knew about the task and process so far, and what needs to be done next. This phase involves critical self-assessment and consideration of whether the current state of the artifact aligns with the envisioned outcome. For example, P06, a UI/UX designer, talks about how *“As I continually work through the project, I go and say okay, reflect back to me what your current imagining of the state is here”* P08 said, *“I would kind of reflect on and actively work through each of the insights in a way that actually would quite often mean I would not get stuck in a few details, and gained perspective at a high-level.”*

A4: Critique. This is when the artifact is subjected to a critique. The practitioner, GenAI, or external stakeholders can provide feedback on various aspects of the artifact. Practitioners describe critiquing their work in two ways – asking the GenAI to self-evaluate its output, or asking the GenAI to critique their idea. For example, P10 used ChatGPT to proofread, *”I had ChatGPT proofread an essay that I had written and it did a decent job. So I tried to use GPT to give me a list of revisions. And I said, just give me the summary of the edits.”* Similarly, software developer P03 used GPT to self-evaluate, *”The model produces some content, then uses a separate prompt to evaluate [it], and then calls itself again with the original output and the self-critique to produce a third output that is hopefully better.”*

A5: Converge. Practitioners often curate and combine multiple outputs or multiple prompts. For example, author V14 shared, *”I regenerated prompts multiple times and then Frankenstein’ed the different versions together to get something I liked.”* Similarly, architect V06 shared their use of Midjourney, *”I’ll remix and I’ll re-prompt until I have a set of images that I’m happy with. Then I’ll use those to start blending, a key part of the process. Blending takes two and up to four images and combines them, [likely] you’re going to uncover something unexpected and use it to work a similar process. [In-progress images are] the result of blending, remixing, and guiding Mid Journey to select for certain traits in each iteration.”*

8.5 Discussion

We build on our results to discuss how we see people’s decision-making around GenAI. we then project our observations and insights into a set of design priorities and opportunities for future CST

8.5.1 Findings, Observations & Ties to Prior Literature

What do practitioners think of GenAI?

Our results and findings about perceived roles when working with GenAI extend prior design metaphors used to conceptualize Human-AI interactions in a more goal-oriented or algorithm-based manner (e.g., control centers, or supertools [382, 184]) to include how practitioners are starting to conceptualize the 'social' dynamics in the relationship over the course of the creative workflow. While the literature has tip-toed around personifying AI [62, 379], practitioners find it useful to think about these evolving cognitive abilities and social contracts [293, 174, 298]. By capturing this change in perceived roles, we hope that HCI researchers, practitioners, and system creators can design and incorporate the right level of empathy, agency, and capabilities in CSTs.

Trade-Offs Considered When Choosing To Use GenAI CSTs

Practitioners do not want to use GenAI models to completely automate their creative workflow. Instead, they consider trade-offs between challenges and benefits. We observed that creatives face UX challenges such as difficulty articulating their creative goals because its hard to externalize tacit knowledge, having to repeat themselves (as a consequence of a lack of persistent knowledge/context), aligning GenAI outputs to user intents despite model stochasticity and opacity, and the friction caused by a fragmented ecosystem of tools, apps, and models where cut-and-paste remains the most viable connective tissue (Figure 8.4). [269] find a gulf between user expectations and the practical experience of conversational agents, suggesting that they should set realistic expectations by interactively revealing the system's capabilities. This parallels the challenges in communicating goals, objectives, and useful knowledge to machine learning models [302].

Despite these challenges, creatives used GenAI due to its tangible benefits (Figure 8.5). It can help to remove cold start situations and further develop creative goals, streamline the creative process by automating or simplifying steps, and accelerate the generation of alternatives. Practitioners often talked about weighing benefits and challenges along dimensions such as their own level of expertise, the amount of creative agency they wanted at that stage, and the expected fidelity of the creative goal or artifact.

Furthermore, these trade-offs can be added as statements to existing quantitative evaluation metrics like the Creativity Support Index [77, 92] to help evaluate GenAI-based CSTs.

How do practitioners use Generative AI?

We asked practitioners to walk us through their latest project which they had created with GenAI models. By analyzing these discussions, we find that the practitioner's conceptualization of their creative process is changing to focus on iterative interacting loops: *Project-level orchestrations* and *Artifact-level orchestrations* (Figure 8.6). This iterative process of creating with GenAI is, at its core non-linear, similar to how traditional creative practices. Whether 1926's four-stage model [428], the recent Double Diamond [101], Stanford's design thinking model [307], or recent work in Mixed-Initiative Co-Creativity [392, 299]. However, here creative practitioners add some fundamentally different activities like gathering information, learning capabilities and limitations, defining goals, planning the project, and specifying relevant context compared to steps in other processes such as *ideate*, *constrain*, *produce*, *suggest*, *select one*, *curate many* to select a subset of artifacts, *assess*, *adapt*, *assemble* which might map better to the artifact-level creative orchestration loop. This process further streamlines the artifact-level loop by integrating steps like *ideate*, *suggest*, and *produce* into *diverge*; *select*, *curate*, and *adapt* into *converge*; as well as *assess*, *constrain*, *assemble*, and *wait* into *transform*. These integrations

underscore the practitioners' creative process as interactions evolve into the intelligent orchestration of actions, information resources, and management of the creative process rather than diving into its nitty-gritty details. This evolution aligns with the metaphors used by creatives who think of their role as project managers who orchestrate information, tasks, and models.

8.5.2 From Observations to Insights: Design Priorities and Opportunities for Future CSTs

Our findings and observations give us a unique perspective to reflect on priorities and opportunities to consider in the design and implementation of CSTs that leverage GenAI.

[D1] Help define creative goals and processes. Creative goals processes can be ambiguous and come into focus through the act of work [C1, B1-2, B4]. GenAI-fueled CSTs can help provide clarity on goals and processes earlier on. For example, by helping craft SMART⁴ goals, and by splitting up complex tasks. Key to this priority is the presence of a rich interaction language across modalities to express goals in a particular domain subject. This involves the articulation of creative styles or briefs. Lastly, CSTs can provide users with fluid access to relevant information retrieval to better define creative goals and plan the project.

[D2] Preserve practitioners' focus and flow. While the creative process can and often benefits from a messy and unprescribed process, it can be hindered by a heterogeneous toolset that often leads to divided attention and fractured interactions [C2, C3, B2]. CSTs should allow for integrated environments that can seamlessly incorporate capabilities and functionalities across the range of tools and services that creatives use for their work. An example of a system that seizes this opportunity

⁴Specific, Measurable, Achievable, and Time-bound.

is Visual Studio Code⁵, which enables code creation with a flexible interface and extensions.

[D3] Facilitate the alignment of GenAI’s outputs to the practitioners’ goals. To be useful and aligned with its user’s intentions, CSTs’ context awareness should evolve over time [C1-3, B1-B5]. This priority aims to develop systems that reduce friction in the interaction and lead to better-aligned results. While ‘conversations’ one can have with a GenAI can have a notion of persistence and context building, other user experiences that accommodate explicit ways to define personal and task-related context will become important. An aspect of this could be the definition and refinement of agents (with access to bespoke context and memory) powered by GenAI. A potential benefit of having explicit ways to express and store context is that they can provide a form of transparency where a user remains aware of what information a GenAI has access to and uses to produce its output.

[D4] Elevate the user’s creative control, and add richer ways to express and verify intent. The current interaction languages for most end-users to express personal preferences or exert agency over GenAIs mainly consist of verbose prompts [C1]. Multimodality and user experience that expand the vocabulary one can use to express personal preferences will be key in future CSTs. As important as elevating a person’s control over a GenAI is ensuring a measure of useful transparency and explainability [§4.1]. This priority focuses on making it clear what the system can do, how well it can do it, and when it did something [P2]. This information, even if marginally accurate, can be the difference between someone being stuck and seeing a path to produce a particular different outcome [B1, 3-5].

[D5] Provide augmented support for creative operations. Critical operations during

⁵<https://code.visualstudio.com/>

a creative workflow include divergence, transformation, reflection, convergence, and critique [A1-5]. CSTs should take advantage of this opportunity by helping generate a diverse set of alternatives, curating, combining, and refining them into polished artifacts [B1-5]. Similarly, CSTs should provide semantically aware shortcuts for novice and seasoned creatives, help them be unstuck, and highlight efficient process paths to achieve desired outcomes. This can be driven by CSTs learning from user behaviors across process iterations.

[D6] Consider and align with the wellbeing of the creative practitioner. There is a significant amount of mental energy that goes into the creative process. As we introduce AI capabilities that analyze and critique one's work, it can be important to have awareness of a user's mental state, so that responses from GenAI are modulated with empathy, while considering the user's well-being [§4.1]. For example, exposing certain individuals to large volumes of information at once can cause anxiety, thus GenAI systems could regulate or progressively reveal their output to be less harmful. Similarly, positive reframing feedback can lead to better ways of AI critiquing.

[D7] Consider technical limitations as design constraints. Technical challenges such as the particular quality of a GenAI's output and latency can be considered temporary and just the next version away from being fixed [C4-6]⁶. However, these types of technical issues and their solutions will not be evenly distributed or accessible, especially in light of the emerging families of smaller and open GenAI models. Embracing these constraints can translate into thinking about designs that incorporate waiting as non-blocking operations, where users can perform other operations while a GenAI is working or providing ways to browse, explore, and

⁶For example <https://stable-diffusion-art.com/sdxl-turbo/>

improve through traditional human agency on imperfect outputs [B2].

These design priorities and opportunities add nuance to an existing body of work outlining design recommendations for systems where people + AI-complete a task [25, 27, 2, 315] by presenting (to our knowledge) the first study with results grounded on the *perspective of creatives* and how they use GenAI in their work.

8.5.3 Limitations & Future Work

Our work collected data from 19 different creative professionals, yet the creative practice is larger and richer than one study can capture. Future work should study the richer and ever-evolving possibilities and mediums of the creative practice.

Working in this rapidly evolving field, where changes can happen in a matter of months rather than years, we focused our analysis on practitioners' experiences and insights rather than the technical capabilities of GenAI available at the time of this chapter's studies (Summer 2023). We are convinced that the fundamental ways in which people perceive and interact with these technologies take longer to change significantly.

We aim to make our work accessible to creatives, GenAI and CST creators, researchers, and educators to help them ground their work on how practitioners relate to these GenAI and how the creative process is evolving because of it. For example, novice and experienced creatives can use our observations and insights when deciding whether or how to adopt GenAI models and tools into their workflow. HCI researchers and CST developers could use observed mental models, challenges, perceived benefits, and interaction patterns to identify innovation opportunities not addressed by current tools. Also, our insights and priorities can serve as dimensions for design spaces or CSTs competitor analysis. Lastly, educators with access to our work can consider how to incorporate them into their curriculum and seek to better understand and develop literacy on how people interact with AI tools.

8.6 Conclusion

To build and evaluate future GenAI CSTs in human-centered ways, it is essential that we study and define this change in the connection between people and a new generation of CSTs. This means examining how the role of people is evolving in creative work, how creative practitioners decide to use GenAI tools, and how their interaction patterns with such tools are changing. Toward this, we conducted a systematic qualitative analysis of creative practitioners from multiple domains reflecting on their process of working with GenAI models to complete a project. We distilled our observations into factors of the interactions between people and GenAI, as well as design priorities and opportunities for future CSTs. Lastly, our analysis puts into focus current creatives' perceptions and the metaphors they rely on to establish a working relationship with GenAI, ultimately preserving agency as orchestrators of powerful AI capabilities. Although our work does not capture the diverse universe of the creative practice, it contributes to building the knowledge and insights needed to embrace and evolve the role of GenAI and creativity and help guide a new wave of CSTs in human- and creative-centered ways.

8.7 Acknowledgements

This chapter in part, includes portions of material as it appears in *Evolving Mental Models, Workflows and Opportunities: A Study of Creative Practitioners Interacting with Generative AI* by Srishti Palani and Gonzalo Ramos in Proceedings of the 2024 ACM Conference on Creativity and Cognition (C&C'24). The dissertation author was the primary investigator and author of this material.

Chapter 9

Amethyst: Enabling Affordances for Specifying and Referring to User-Generated Context Fosters Creativity Human-Centered Orchestration of GenAI

Based on the design guidelines distilled from the formative study in the previous chapter, we implement *Amethyst*, a creative smart notebook that aims to address the above challenges by leveraging GenAI to support the creative process in an integrated, context-aware manner. *Amethyst* features include supporting goal decomposition, grounding prompts to specific contextual information, modulating GenAI output through simulated expert personas, and providing a range of in-line, nonblocking creative operations. We evaluated the potential of *Amethyst* to support the creative process through a within-subjects user study ($n = 12$), comparing *Amethyst* to a baseline condition of standard tools such as web search, LLM-based chat, and digital notebooks. We find that participants generated more novel, feasible and creative ideas and preferred using *Amethyst* as it helped interact with GenAI in a more integrated, empathetic and context-aware manner. Through this work, we outline opportunities to focus on incorporating LLMs into the creative process to boost it in human-centred ways while sharing insights to guide those seeking to develop solutions in the space of GenAI-based creativity support tools.

9.1 Introduction

Creativity is often romanticized as lightning-strike moments of inspiration or dismissed as an innate ability possessed by a select few [300, 345]. However, creativity is a process — an iterative journey of exploration, ideation, reflection, and refinement [394, 393]. During the creative process, people engage with diverse tools, work across multiple sessions, draw on expertise from multiple domains, are influenced by their actions, and even collaborate with others [24, 150].

Although GenAI-based CSTs promise automating the entire creative process with just a well-crafted prompt [62], in reality, today people are increasingly adopting a hybrid approach, leveraging the capabilities of GenAI in conjunction with other traditional

creative support tools to accomplish their creative goals [117]. For example, designing a video game might involve ideating plot and mechanics in ChatGPT, creating game visuals in Midjourney, developing the game in Unreal, and animating it using Runway ML. Recent work in HCI has begun to investigate the building of interactive CSTs with GenAI models [398, 155, 97] and characterizing interaction mechanisms [262, 396] with specific individual models for individual creative tasks.

From the formative study mentioned in the previous chapter, we learnt that people perceive their relationship with GenAI CSTs in a fundamentally different way from non-GenAI CSTs, and their creative practice is evolving to balance new challenges and capabilities. Overall, this study highlights opportunities for GenAI to support not only individual tasks but also the creative process itself by helping people plan, monitor and evaluate their work while prioritizing their agency to define and reference context, all while interacting with a system that integrates otherwise fragmented capabilities and is capable of providing empathetic responses.

Driven by these opportunities to support not only artefact generation but also process, we pose an additional research question:

RQ: How might we implement the design opportunities derived from the formative study in a system? In particular, what does it look like to have a GenAI-powered system that enables people to orchestrate their creative process in an integrated manner, define and manage contexts, and interact with GenAI in an empathetic manner?

To answer this research question, we designed and implemented *Amethyst*, a system that supports creative operations and the overall process in human-centered ways. We instantiated *Amethyst* as a digital workbook, and among many features, it enables *goal decomposition and task management*, the quick invocation of a range of creative operations, *context-aware prompting* that allows users to align model outputs with their intentions by quickly referring to relevant contexts from their process, including user-level, project-level, and external contexts. Additionally, *Amethyst* also introduces the

concept of *simulated expert personas* that enables creatives to modulate by creating, managing and referring to expert 'personas.' *Amethyst* also gives the user the agency to see and edit the context on which a GenAI output will be based in a transparent manner. All interactions with the system are non-blocking, illustrating how we can design around technical constraints and allow people to continue with their work while generations are in progress.

To evaluate how *Amethyst* supports peoples' creative processes and how they interact with GenAI during these processes and perceive their roles within them, we conducted a within-subjects study with 12 people. Participants were asked to engage in creative projects using both *Amethyst* and a baseline condition consisting of standard tools such as search engines, chat-based LLM services, and a digital notebook. Our results show that participants produced more creative ideas when using *Amethyst* than during the baseline condition, as rated by design experts who were blind to the conditions. Results also show that beyond affecting the creative outcome, *Amethyst* supported the creative process – helping users interact with GenAI more seamlessly, contextually, and empathetically. Participants reported that they preferred using *Amethyst*, wanting to leverage content generated as part of their creative process as additional context to support their interactions with GenAI during the creative process rather than their usual set of tools.

Through our work, we present the following contributions:

1. **Empirical insights and design recommendations from a formative study** with creative professionals that identify current practices, challenges, and opportunities around how people interact with GenAI during their creative processes.
2. A **prototype system, *Amethyst***, that scaffolds not only individual tasks but also planning, monitoring, and evaluating during the creative process using GenAI with techniques like context-aware prompting and interacting with simulated expert personas.

3. **The results and insights from a within-subjects study** that finds that a system like *Amethyst* (which uses GenAI to both support creative processes and outcomes) can lead to better creative outcomes than using the current fragmented set of GenAI tools.

9.2 Related Work

This chapter's work builds on prior research studying the nature of human creativity, GenAI, human-GenAI interactions during creative work, and tools built to support the creative process.

9.2.1 The creative process is just as important as the outcome

Creativity is generally defined as the "*production of something original and worthwhile*" [393]. The creative outcome could be intangible (e.g., an idea, a scientific theory, a musical composition, or a joke) or physical (e.g., a device, a printed literary work, or a painting). Creativity is often evaluated by examining the creative outcomes based on their novelty, feasibility, and value [393].

Cognitive approaches to understanding creativity emphasize how the creative process itself is as significant as its outcome [428, 394]. The creative process is a non-linear and iterative process. Beyond these generalities, there are many characterizations of the creative process. Some describe the staging of the creative process – e.g., Wallas' four-stage model which suggests that the creative process involves knowledge acquisition (preparation), unconscious information processing (incubation), emergence of the idea (illumination), and evaluation of the idea (verification) [428]. Other characterizations describe the types of thinking creative processes involve – e.g., Guilford's Structure of Intellect model which characterizes the creative process as iterating between generating

multiple solutions to a problem (divergent thinking) and effectively evaluating these solutions (convergent thinking) [394].

Some theories go beyond the individual's knowledge, skills, and processes to also include their social context. For example, Amabile's Componential Theory outlines the importance of domain-relevant skills, creativity-relevant processes, intrinsic task motivation, and social environments for creativity [24]. This emphasizes the importance of not only domain knowledge, but also procedural knowledge of how to do things and conditional knowledge of when and why to use specific domain and procedural knowledge. It also highlights the importance of meta-cognitive regulation like planning, monitoring, evaluating, and managing knowledge, skills, and resources. An individual's intrinsic motivation and emotional creativity are also important, as they allow individuals to create something new through the influence of emotions from personal experiences. Social context also plays a crucial role in creativity, as factors such as criticism, politics, the status quo, low-risk attitudes, and time pressure can impede creativity. In contrast, positive challenges, diverse teams, freedom in work execution, supportive management, and a culture of sharing ideas can foster creativity. One's creative style can also be developed by mimicking inspirational experts' styles, contrasting various methodologies, and improving personal creative practices based on feedback from peers and mentors. Extrinsic factors such as the intended audience, genre, or specific project limitations can also impact the style adopted for a creative project. [22, 39].

9.2.2 Today's GenAI-based Creativity Support Tools assist with individual tasks, not the entire process

Creativity Support Tools (CSTs) formally began as a field in the late 1990s - early 2000s to mitigate some of these challenges and make creative work easier, faster, and more efficient [379, 380, 279]. CST research has developed tools for many stages,

such as discovering insights [271, 317], defining project goals and task planning [235], idea generation [383], and prototyping [116, 148], refinement [228], getting feedback [149], and communicating ideas more effectively [91, 380, 150]. However, most of these prototype CSTs exist in a laboratory setting; few explorations are carried out for tools in the wild over a long period of time [150]. This motivates our research focus on understanding how creative professionals are integrating GenAI-powered CSTs over the course of their workflows, not just for a single task.

The landscape of CSTs is undergoing rapid evolution with the introduction of GenAI. These general-purpose models are trained on large data sets and can generate text (e.g., [62, 58]), imagery (e.g., [12, 14]), or other media [13, 10, 56, 11]. Notably, GenAI models are starting to generate human-like creative outcomes [175, 62]. GenAI outputs have successfully passed creativity tests such as the Alternate Use Test and Torrance Test [175, 377]. They have also shown their creativity in more practical tasks, such as generating ideas for startups [161], and short stories [126]. Due to the relative ease with which they can be customized and controlled (that is, using natural language 'prompts') and their unique ability to generate ideas, these models offer a number of potential benefits for creative tasks [62].

Instead of automating the creative process and automatically generating creative outcomes, the field of human-computer interaction is built on foundational visions of designing interactive systems where humans and AI work together to solve problems while intuitively trading off agency and automation based on complementary abilities [254, 401, 190, 181]. Therefore, HCI researchers, like us, go beyond building techniques for training (e.g., [172, 263]), fine-tuning (e.g. LoRA [193], QLoRA [119]), and even prompting techniques (e.g. chain-of-thought reasoning [432], ReAct [452]) to instead build interaction techniques that support people's creative processes by leveraging GenAI. In recent years, the HCI community has built novel human-AI interaction techniques

to compose music [267], design visual art [97, 96], write journalistic articles [324], and novels [247, 454, 155].

While GenAI has proven capable of producing creative outputs, there is limited understanding of how it can support the multifaceted nature of the creative process. A process that involves various interrelated components that collectively shape the process and outcome: Domain knowledge (i.e., expertise and understanding within a specific field or discipline); Procedural knowledge (i.e. knowing the steps and methods for accomplishing tasks); Conditional knowledge (i.e., understanding when and why to apply specific domain and procedural knowledge.); Metacognitive regulation (i.e., the ability to plan, monitor, evaluate, and manage one’s knowledge, skills, and resources); Intrinsic motivation and emotional creativity (i.e., the internal drive and emotional factors that fuel creative endeavors); and Social environments (i.e., the role of external factors, such as collaboration and feedback).

This chapter’s work builds on these works to better understand how people incorporate GenAI models into their creative process and how this can change their process and perception of their role in it. Understanding these user perspectives is fundamental to designing the right human-centered systems in this space that align with people’s mental models and technology expectations. Our work presents a perspective of what some of these solutions could look like and the design principles that guide them.

9.3 Amethyst System

Guided by our design goals, we develop *Amethyst*, a smart workbook to help orchestrate GenAI throughout the creative process. We develop *Amethyst* as a *technology probe* [197], which is a type of research vehicle used early in the design process with the purpose “not to capture what is so much as to inspire what might be.” [49]. In this

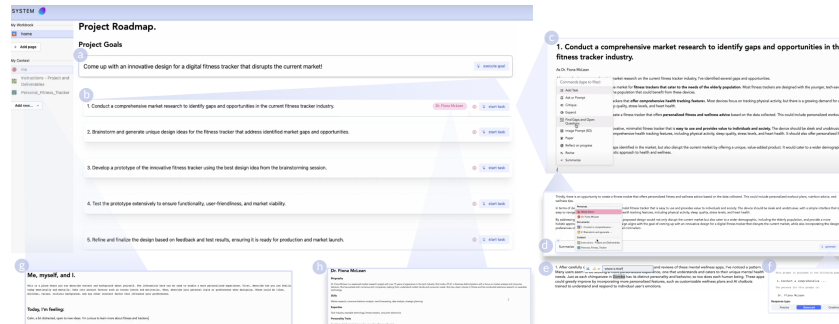


Figure 9.1: Amethyst is a smart notebook designed to facilitate the creative process in an integrated manner. With Amethyst, users can orchestrate GenAI models to (a) decompose their goals, (b) manage the resulting tasks, (c) access a range of creative operations by typing “/”, and (d) specify and maintain context throughout the process by explicitly referring to it using “@”, implicitly in the prompt, or by (e) selecting relevant context and prompting in-line. To better understand the context of each operation, users can hover over the generate button and activate the (f) transparency lens. Additionally, users can adjust GenAI responses to be more empathetic by (g) specifying their own emotional states and design preferences, as well as (h) generating or specifying simulated expert perspectives.

section, we will describe the system through an example user scenario, a walk-through of the system’s main features, and details about the implementation.

9.3.1 Example User Scenario

Consider a product designer, Sam, interested in *coming up with an innovative design for a digital fitness tracker that disrupts the current market!* Not knowing how to get started with this complex fuzzy goal, she types this in as a prompt into a chat-based LLM service, but it outputs pretty generic ideas that miss important contextual details that Sam wants. Sam tries to improve the outputs by iterating on the prompt multiple times, looking up key information with search engines to add details to the prompts, and even cuts-and-pastes outputs into a notes document to try and iterate on the output herself.

Feeling overwhelmed, she switches to *Amethyst* and types the same complex, ill-defined goal into the *goal prompt* on the homepage of her notebook. The system helps her get started by decomposing the goal into actionable tasks and presenting them as a

list of *task-prompt* components (Figure 9.2). Tasks include '1. Conduct comprehensive market research to identify gaps in the current fitness tracker industry', 'Brainstorm and generate unique design ideas for the fitness tracker that address identified market gaps and opportunities', etc.

Excited to start the first task of conducting thorough market research, Sam clicks the *start task* button on the task, and the system creates a dedicated working page for it in the notebook and presents the output of the task there. She gets a recent and relevant overview of the market, competitors, customer needs, and opportunities.

Wanting to add insights her collaborators had sent over in a document to this overview, she uploads this document under My Context. Then, to summarize, she types a *forward slash "/"* to invoke a drop-down menu (e.g., Figure 9.3) that allows her to select from a list of operations, and she chooses *summarize*. To reference that she wants this other document summarized, she types '@', which lets her choose from a drop-down list of documents in the workbook, including the shared document '*Personal Fitness Tracker*', which she selects (Figure 9.4). As she hits the *generate* button on the summarize component, the output of this operation appears in a *result block* that allows her to preview, regenerate, discard, or insert the output on the working page. Further, to ensure her research so far is comprehensive and to analyze the coherence and diversity of her findings, Sam uses operations like "*Find Gaps and Open Questions*" and "*Critique*". This iterative process enhances her understanding of the market landscape.

She wants to get opinions from domain experts to ensure that she has explored all perspectives of the fitness tracker market. So, she clicks on the "generate persona" button on the previously generated task component of '1. Conduct comprehensive market research to identify gaps and opportunities in the fitness tracker industry' to get interesting expert perspectives. She clicks the button twice to generate a diverse set of personas – once asking for a *balanced* generation, Dr. Fiona McLean, an expert market research

analyst, and another more *creative* and unconventional one, Dr. Jane Goodall. Each persona is detailed on a "persona page" in the menu of each notebook, containing a short biography, skills, expertise, personality traits, and work style (Figure 9.7). Sam customizes these personas to help them achieve her goals by adding skills or personality traits to channel into how they help her execute each task. To get Dr. Fiona McLean's perspective on the market research, she types '@' and chooses 'Dr. Fiona McLean' from a drop-down list of personas and documents in their workbook.

Next, to combine all of these perspectives into a comprehensive market research, she starts to edit the working page. Even wanting the GenAI model to help revise it, she invokes the *Revise* slash operation. She further specifies the revise operation by modifying the revise component on the page with "Revise in a way that synthesizes all perspectives and details into a coherent whole." All "/" operations have a default grounding, which is the working page on which it is invoked unless otherwise specified. So, the *Revise* operation knows what this otherwise vague prompt means. She checks the grounding of each operation component by hovering over its action button that displays the pages and personas that the operation will consider in the *transparency lens* component.

Feeling confident in the market research, Sam moves to the next task of 'Brainstorm and generate unique design ideas based on the identified market gaps and opportunities' and chooses to ground this task in the previous tasks' output page by typing '@' and selecting that page from the drop-down menu. The system considers the market research material available on the page, as well as the overall project goal, to give contextually relevant and unique ideas. Similarly, she goes about each of the remaining tasks on the homepage.

Sometimes, she issues multiple operations simultaneously, and the system displays an "in-progress" status for interactive components and result blocks while still letting her edit the same page, change pages, and continue with other tasks in the interactive

notebook while she waits for the results.

To make sure that she is on track towards accomplishing her goal, she issues the *Reflect* operation. When executed, this operation inspects all the pages in the notebook, produces a reflection of progress made towards the goal, and suggests the next steps. Encouraged by this guidance on how to proceed, Sam feels more confident in how to achieve her goal.

9.3.2 System Features

To provide users with a familiar user interface to work with and reduce the effort of discovering features and how to use them, *Amethyst*'s base user interface is a workbook modeled after popular digital note-taking interfaces that most users would have encountered, like Microsoft Loop¹ and Notion². *Amethyst*'s front-end interface comprises two main areas: the *sidebar panel* and the *editor panel* (Figure 9.1). The *sidebar panel* (Figure 9.1) is a component that allows users to select, create, and delete documents. When a particular document is selected, it appears in the *editor panel*, which takes most of the area of the interface and is the area where people can view and edit the currently selected document.

The following features add to this base user interface and provide interaction techniques for users to integrate and orchestrate GenAI throughout their creative process. We organize the description of these system features according to our main design goals from our formative study.

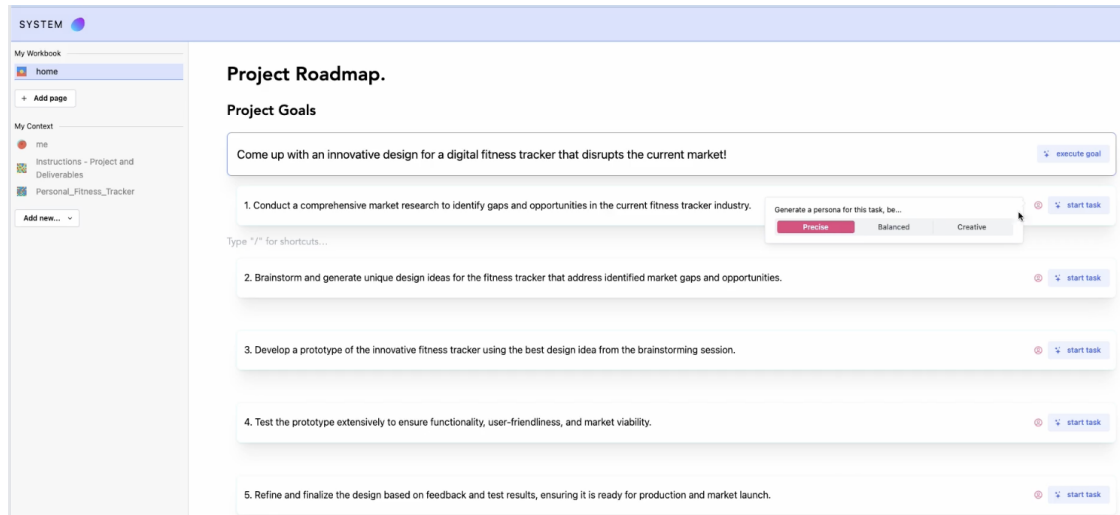


Figure 9.2: Users can provide a high-level objective, which the system can then break down into actionable tasks. Also, all following GenAI outputs will use this as additional context

D1: Help define creative goals

Goal Decomposition and Task Planning in *Home* page: The *home* page is the main page of the system. Here, users can enter their creative goals into the *goal component*, and clicking on 'execute goal' will break down the goal into specific, achievable, and relevant tasks and display them as *task components* (Figure 9.2). Users can specify how certain the action plans generated are by adjusting the temperature of the goal component. When a user starts a task, a dedicated working page is created for it, and subsequent generations about this task will appear on it.

D2: Integrated ecosystem that supports multiple creative operations

Creative Operations as *"/* (Forward Slash) Commands: On every page in *Amethyst*, users can leverage GenAI models to perform creative operations such as information gathering, curation, divergence, transformation, reflection, convergence, and

¹<https://loop.microsoft.com>

²<https://notion.so>

critique. This list of operations is derived from the formative study in which participants describe them as common in their workflows. So, *Amethyst* leverages a common interaction pattern in editors, the *"/" command* to invoke a drop-down menu that allows them to select what (creative) operation they would like to access. Table 9.1 describes the operations we made available in *Amethyst*.

All *task* and *operation* components have a basic structure comprising a text entry field and an action button that triggers a particular type of operation. Upon hovering over the action button, they can select the level of unconventional generation they want: precise, balanced, or creative (Figure 9.2). These parameters directly map to generation temperatures of 0, 0.5, and 1, respectively.

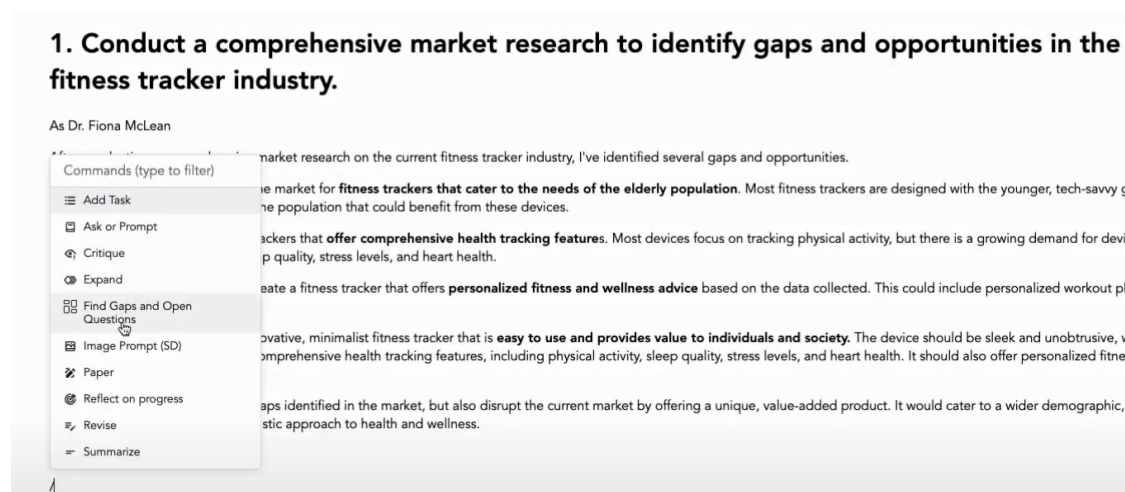


Figure 9.3: *"/* (Forward Slash) Commands: Users can access various creative operations by typing *"/* to open the drop-down menu and selecting the desired operation.

D3: Align GenAI output to user goals and contexts

Context-Aware Prompting Technique: enables users to better align LLM outputs to their intentions by explicitly or implicitly referring to relevant contexts present in the many artifacts generated across their process. Users can ground any generation in a relevant context using one of three ways:

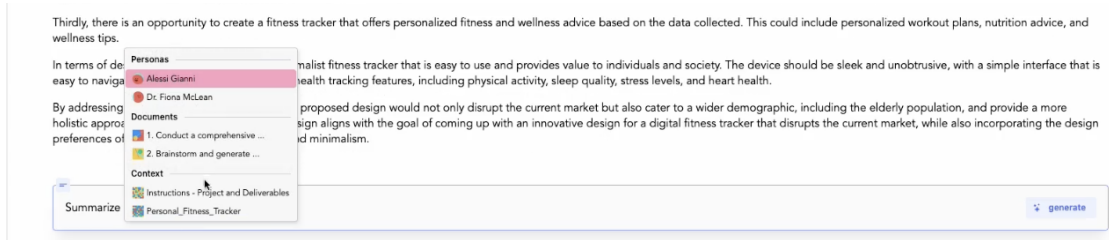


Figure 9.4: This Summarize component is an example of a *"/* operation. Here you can see how the user can make use of *'@'* mentions to change how the operation is grounded to different personas and/or documents.

1. **Explicitly Grounding:** By typing in *'@'*, another familiar notebook interaction pattern, and then selecting the relevant context from the drop-down list of project-relevant context (Figure 9.4). Relevant context can be *project-level* (e.g., goal components, working pages of various tasks from the action plan), *external contexts* (e.g., collaborator's shared files under my context, or information on the web), or *personal-level* (e.g., emotional state and design preferences mentioned in *'me'* page under *'my context'*).
2. **Implicitly Grounding:** by mentioning in the prompt to relevant information that is somewhere in the notebook.
3. **In-Line Prompting,** which provides a lightweight way to do semantic operations based on parts of a page, instead of *"/* operations, which are grounded to whole pages. To do this, a user can select part of a page in the same way one would do in any text editor. This brings up *Amethyst In-Line prompt* that allows users to apply a prompt to the current selection. After entering the prompt, a context prompt block is inserted immediately after the closest paragraph containing the selection. Figure 9.5 illustrates an example of using this capability to search for the definition of a term. This lightweight capability is not grounded to a page or associated with a persona unless specified otherwise.

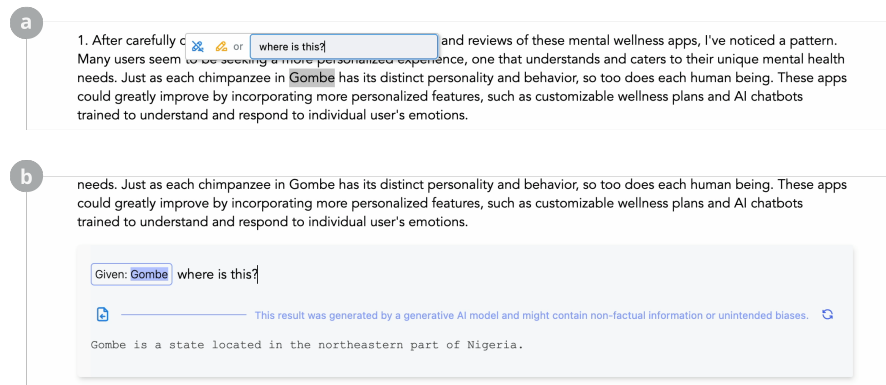


Figure 9.5: In-line prompting. This figure illustrates Amethyst’s support for in-line contextual prompting. a) first the user will select a part of a document, which will reveal a contextual prompt area where the user enters their request. b) after the request is fulfilled, it is inserted as a block, closest to the selection. As with any results block, the user can regenerate the result, change the generation parameters, paste the result into the page, or delete the block.

In the back end, each prompt is augmented with relevant information by leveraging prompt engineering techniques like Retrieval-Augmented Generation [152] and ReACT [452] agents with access to tools including a search engine, other GenAI models like stable diffusion, etc. Overall, the context-aware prompting technique aims to help maintain a thread across outputs generated over the course of each project without repeating relevant context and improves the reliability of the generated responses.

D4: Interact empathetically with the creative practitioner

Sharing Emotional State in *Me page* & Empathetic Generative Outputs: Building on the practice of journaling to examine and reflect on one’s emotions [], another special page of the system is the ‘Me’ page where the user can define how they would like to be seen by externalizing personal context, such as their current emotional state and design preferences (Figure 9.6). This personal information is then used by the system as additional context when framing its outputs, especially for operations such as ‘critique’, ‘find gaps and open questions’, and ‘revise’.

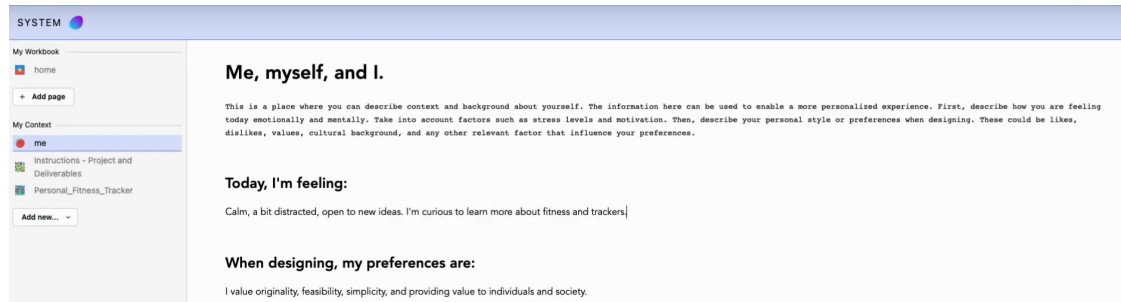


Figure 9.6: Me page: Users can edit this text file to define how they would like to be seen by the system externalizing personal context such as their current emotional state and design preferences

Collaboration with Simulated Expert Perspectives: This technique enables users to modulate GenAI model behavior by creating and managing 'expert personas' and their skill sets, characteristic styles, personality traits, etc., outside of prompts (Figure 9.7). This is inspired by an insight from our formative study, where some creatives wanted interactions with GenAI to model social relationships in creative studios like design collaborations and critique sessions.

Users can create and channel perspectives of custom '**Persona pages**' by creating a new persona under 'My Context' in Amethyst's sidebar. Or if they are not sure what might be interesting or relevant perspectives to consider, they can ask the system to generate relevant personas for each task by clicking the '*Generate persona*' button in each 'Task' component on the 'Home' page. This adds a persona page in the *sidebar panel* context area (e.g., left panel on the UI illustrated in Figure 9.6. If one is not happy with the persona generated, one can click the action button again and generate a new one, or alternatively, edit the persona's definition in the context area. From then on, if the user starts that task, it will be carried out from the lens of the persona associated with it. As with any additional context, they can assign a persona to any "/" operation by typing "@" and selecting the relevant persona from the drop-down menu.

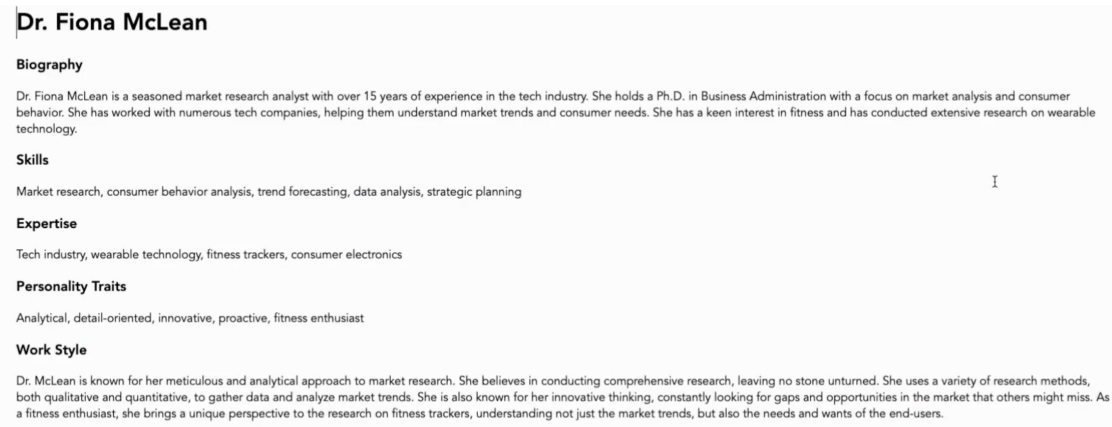


Figure 9.7: Example of a persona page. Users can edit this text file to customize their personas to best help with their work.

D5: Prioritize user’s agency and creative control

Result Blocks: After the user submits any “/” operation prompt, *Amethyst* presents the results of the operation in a *Result block* (Figure 9.9) that allows users to preview, regenerate, discard, or insert the result in the containing page.

Transparency Lens: Hovering over the generate button of each “/” operation component displays what context it is grounded on (i.e., which page, persona, etc.) so that the users know what information it has access to for transparency and in case they want to change it.

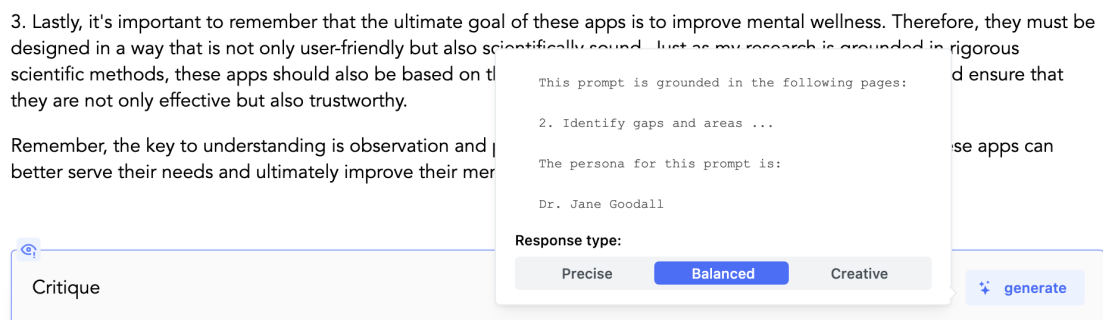


Figure 9.8: Transparency lens: Hovering over the generate button of each “/” operation component displays what context it is grounded on so that the users know what information it has access to for transparency and in case they want to change it.

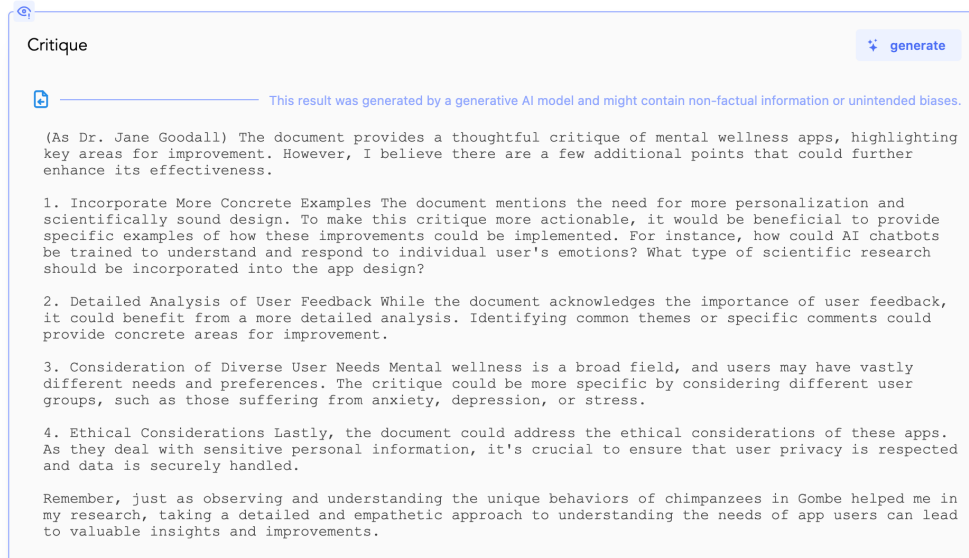


Figure 9.9: Critique component output. The figure illustrates how the result of a “” operation is presented. The *block* keeps the results contained and give users the option to regenerate the results, insert them into the document, o deleting it altogether.

D6: Consider technical limitations as design constraints

Non-Blocking Actions: *Amethyst*’s support depends on asynchronous operations, that is, performing a GenAI-based operation that does not produce immediate results (e.g., a GenAI API service throttling). *Amethyst* embraces these delays and has its prompt components and result blocks reflecting their “in progress” status in a way that does not prevent the user from performing other activities (e.g., switching pages, further editing, prompting) on the same page or another page while results are on their way.

9.3.3 Implementation Details

The *editor panel* is built on top of TipTap³, a headless open-source editor, which we enhance with custom components to invoke and present AI capabilities.

Amethyst’s functionality and capabilities are supported by a Semantic and Document

³<https://tiptap.dev>

Services back-end that exposes a RESTful endpoint for AI capabilities and document management. Document services consist of providing an endpoint for CRUD operations to manage and maintain the different types of documents supported by *Amethyst*. This role allows for document persistence beyond a browser session and for access to documents' contents when semantic operations are performed.

The back end provides AI semantic capabilities by building on top of general LLM functionality. *Amethyst*'s back-end uses LangChain⁴ and OpenAI's GPT-4 [312] to fulfill a user's request for a particular capability or operation. Figure 9.10 illustrates the core structure in which a system's operation is fulfilled: a user's prompt is expanded using grounding and contextual information into a meta-prompt that is passed to a LLM, in turn the response is (parsed) and passed back to the user. For details about the prompts used in *Amethyst*, readers should refer to the Appendix C.

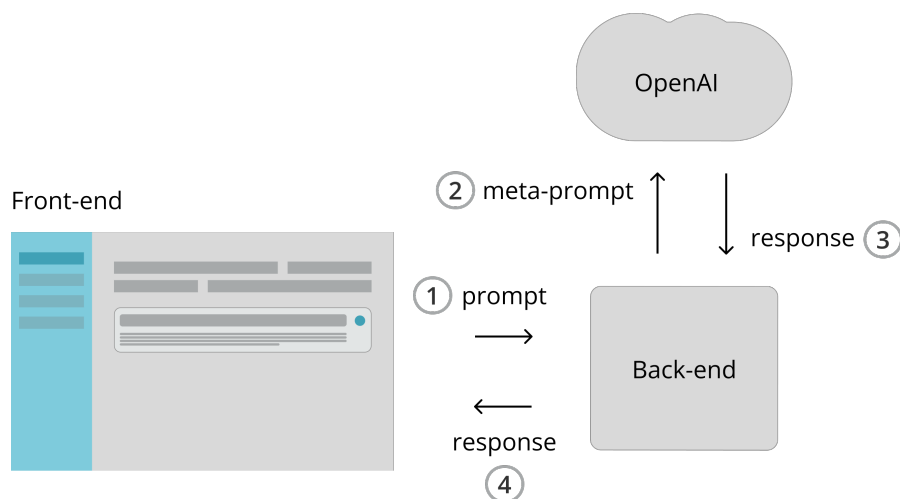


Figure 9.10: *Amethyst*'s components. This figure illustrates at a high level how the front end and the back end interact. 1) when an operation takes place, the prompt visible in the component is sent to the back-end. 2) the back-end enhances the prompt into a meta-prompt using its knowledge about the user and operation context; the meta-prompt is then sent to the LLM. 3) the LLM's response is received by the back-end, which can perform parsing and formatting operations. 4) the back-end sends the parsed response to the front-end, which can further parse it, then presents it into a response block.

⁴<https://www.langchain.com/>

9.4 User Evaluation Study Design

We designed *Amethyst* to support the orchestration of multiple GenAI capabilities in a context-aware manner during the creative process in ways that gave people creative agency while engaging with empathetic responses. To validate our design, gather insights, and measure effects from *Amethyst*'s use, we conducted a within-subjects experiment with 12 participants conducting a creative task with *Amethyst* as well as during a baseline condition consisting of off-the-shelf tools traditionally used for creative processes such as a search engine, an LLM-based chat, a digital notebook.

9.4.1 Tasks

To contextualize their creative design process, participants were given a simulated work task scenario [50]. The task topics involved relatively large, complex, multifaceted information spaces where the average person had at least some prior knowledge or experience coming into the task, yet they had the opportunity to learn more about the area. Our task instructions were as follows:

”Imagine you are a designer working for a client to redesign [Airport Security Experience — Fitness Trackers — Personal Finance Management Apps — Job Search Apps — Mental Health Apps]. Your job is to change the existing design concept so that any identified issues are solved and the product is more innovative, original, and/or valuable.

During the task, you will:

- 2-3 original ideas of how to make the product better
- 1 final refined idea
- Script for a short elevator pitch

As part of the within-subjects study design for evaluating user behavior across the two conditions, each participant performed the above task once for each condition. To prevent carryover effects in learning, each participant completed the task on two different

topics chosen from the above instructions. As a potential starting point, participants received a market and user research report to understand the existing product, its market, user base, and the design goals of their simulated clients. Participants were given 45 minutes to complete each creative task, with a 10-minute break between tasks.

9.4.2 Baseline

To choose an ecologically valid baseline, we compared *Amethyst* to the tools that are representative of how users currently perform creative processes. Although there were no limitations on what people could use during the baseline condition, we recommended participants stay within the tools from the following list: Internet browser (e.g., Google Chrome, Safari, or Microsoft Edge), LLM (e.g., Bing Chat, ChatGPT, Claude), and Search engine (e.g., Bing, DuckDuckGo, Google). We also provided a text editor as a place to take notes, paste content, and process information (e.g., Notion or Microsoft Loop). This list of tools is derived based on insights from our formative study in which we asked participants what tools they used during their creative process.

9.4.3 Participants

12 participants (four female, eight male) were recruited across research, design, and software engineering departments of a large technology company via mailing lists. In terms of their roles at work, one was a writer, one was a designer, four were engineers, two were people managers, two were doctoral students, and two were researchers. Eight of them had accumulated 6-10 years of experience in their respective fields, while two possessed less than a year of experience, and another two had between three and five years of expertise. In terms of experience using generative AI models, two said less than 6 months, six of them reported having experience of 6 months to 1 year, two for

1-2 years, and one for 2-3 years. Seven participants reported using these models very frequently (multiple times a week), four reported using them frequently (multiple times per month), and one reported using them occasionally (a few times in a year). All studies were conducted remotely over video calls. Compensation was \$100 USD Gift card per participant for an approximately 2.5-hour study.

9.4.4 Study Procedure

Before the study appointment, participants were sent an informed consent form and asked to complete a demographic survey. During their study appointment, each participant underwent two creative tasks in which the presentation order for the experience condition (*Amethyst* and *Baseline*) was counterbalanced.

During the 45-minute task, participants freely interacted with the tool(s) in each condition to generate between two and three original ideas of how to make a product better, converge to one final refined idea, and write a short elevator pitch, all while thinking aloud about their experiences [250, 198]. Before starting each 45-minute task, participants either watched a 5-minute tutorial for the *Amethyst* condition or were given a short refresh on the *Baseline* set of tools. We gave them approximately 15 minutes to explore and practice.

After each condition, we asked them to rate their level of agreement with statements about their user experience with the system (such as "I was able to clearly articulate my creative goals," "I was able to manage tasks," "I was inspired or able to generate ideas," "I was able to create something novel," etc.). We also asked them to list up to three things they liked, as well as up to three things they would like to improve about their experience with each condition. At the end of the study, once they had used both (*Baseline* and *Amethyst*), we asked them which system they preferred using and why. From the study, we capture self-report data through think-aloud and responses to surveys,

as well as application log data, to quantitatively measure and understand user behavior.

9.4.5 Measures

Quality of Creative Outcomes

Our primary measure assessed *Amethyst*'s support for the creative process compared to the baseline was by analyzing participants' creative outcomes. For this, we had user-experience design experts, blind-to-condition. We define experts as those who have a Master's degree in Design and have completed at least 3 design projects. Experts were recruited from a Master's design program at a public university. They were asked to rate the three types of creative outcomes (2-3 original ideas of how to make the product better, 1 final refined idea, and Script for a short elevator pitch) on a scale of 1-5 (where higher is better) along the following criteria, inspired by the literature on the evaluation of creative outcomes [393].:

- *Novelty*: refers to the degree to which an idea, product, or solution is unique or original. It is the opposite of something that is obvious, ordinary, or already well-known.
- *Feasibility*: refers to the practicality or workability of an idea or solution. It considers factors such as available resources, technical constraints, economic viability, and compatibility with existing systems or processes.
- *Value*: refers to the usefulness, significance, or importance of an idea or solution. It considers the potential benefits, impact, or worth that the creative output might have for the intended audience, market, or purpose.

Behavior Log Analysis

By analyzing application logs from both conditions, we measured how often each participant interacted with each feature. These include metrics like query frequency, prompt usage, engagement with *Amethyst*-specific features such as goal decomposition, persona assignment, and context referencing, etc.

Qualitative Insights and Perceived Values

In order to gain a deeper understanding of the benefits and challenges of both conditions, we transcribed participants' think-aloud recordings during the tasks. The first author then reviewed the transcripts in two passes using an open coding approach [87]. Through discussions with the rest of the research team, we identified common themes in participants' experiences. Additionally, we also conducted a post-task survey where we asked participants to rate a set of statements around system values using a five-point Likert scale for agreement.

9.5 Findings

Overall, nine out of twelve participants preferred *Amethyst* to the baseline condition. In this section, we present the results of how *Amethyst* affects creative outcomes, creative processes, and the participant's preferences when compared to the baseline condition, giving us a holistic view of the trade-off of using *Amethyst*.

9.5.1 Better Creative Outcomes Achieved When Using *Amethyst*

Based on the sum of 5-point *expert ratings* on the three criteria – Novelty, Feasibility, and Value – participants generated significantly *more creative outcomes* when using *Amethyst* compared to the baseline ($M = 11.50$ out of 15, $SD = 2.10$ vs. $M = 6.45$ out

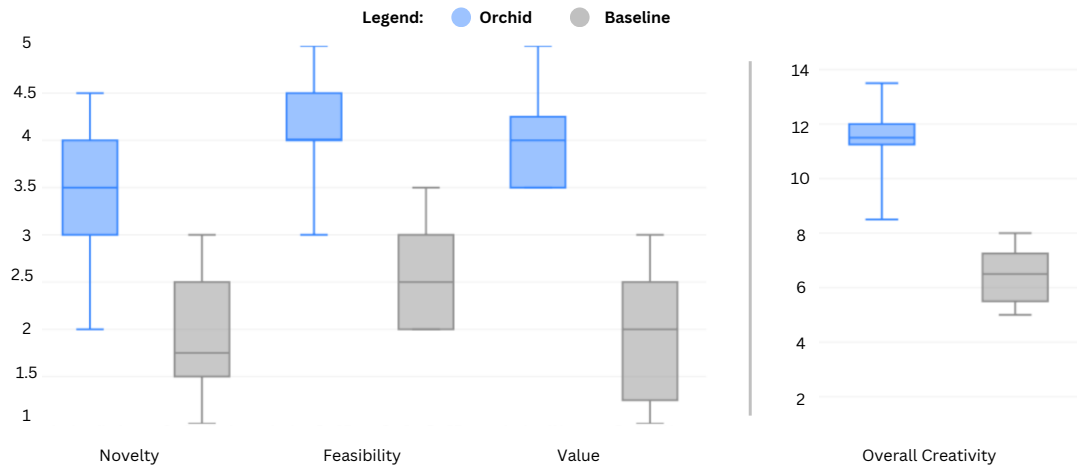


Figure 9.11: Participants generated significantly more creative outcomes when using *Amethyst* vs. the baseline when rated by blind-to-condition experts on Novelty, Feasibility, and Value of ideas, on a 5-point Likert-scales for agreement (5 indicated strong agreement).

of 15, $SD = 4.61$; $t = 1.58$, $p = 0.01^{**}$, Fig. 9.11). To correct for multiple comparisons, we ran a Wald-type test and a MANOVA with repeated measures and found a significant difference between the two conditions on the combined measures of *Novelty*, *Feasibility*, and *Value* ($manovaF = 8.28$, $p = 0.02^*$, $Wald\chi^2 = 13.18$, $p = 0.01^{**}$).

Breaking down this overall creativity score, participants wrote significantly more feasible ($M = 4.17$ out of 5, $SD = 1.27$,) and valuable ($M = 3.96$ out of 5, $SD = 1.30$,) pitches when using *Amethyst* compared to the baseline condition (feasibility: $M = 2.65$ out of 5, $SD = 1.58$; valuable: $M = 1.92$ out of 5, $SD = 1.67$). Participants also generated more novel pitches when using *Amethyst* ($M = 3.38$ out of 5, $SD = 1.30$) compared to the baseline condition ($M = 1.92$ out of 5, $SD = 1.55$). However, this difference was not statistically significant.

9.5.2 RQ3: How does the *Amethyst* help orchestrate GenAI during the creative process in human-centered ways

To understand how orchestrating GenAI during the creative process is different when using *Amethyst* from the baseline, we analyze the logs of tools used in both conditions. We present each analysis in terms of how it addresses each of the design goals we derived from the formative study. Overall, we find that participants issued significantly more prompts during the baseline condition ($M = 12.00, SD = 1.58$) compared to *Amethyst* ($M = 6.53, SD = 1.19, t = -2.58, p = 0.05^*$).

Amethyst helps define fuzzy goals

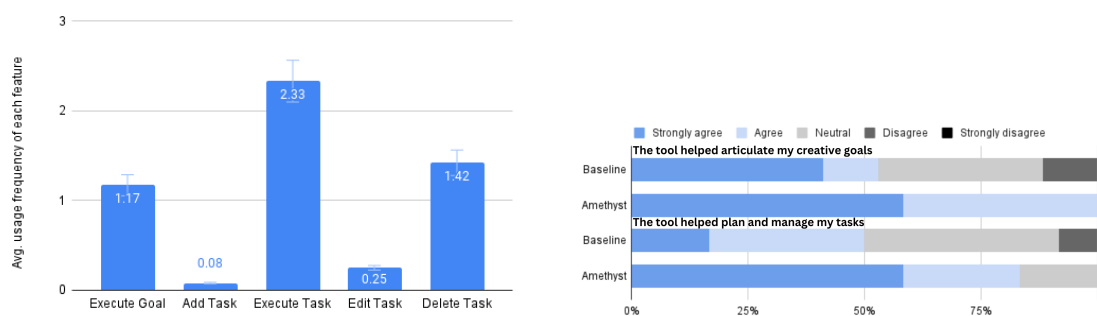


Figure 9.12: Participants found *Amethyst*'s goal decomposition and task management features helpful in defining their fuzzy creative goals

Participants only used the model to decompose their goals and manage the resulting tasks when using *Amethyst*. On average, the **Goal** component was used 1.17 times by each participant to decompose their high-level goals into action plans. Participants used the **Task** component by decomposing their goal on average 2.33 times, sometimes adding, deleting, and editing these tasks (Figure 9.12(L)).

In the post-condition survey, participants rated a significantly higher level of agreement for *Amethyst* compared to the baseline; for example, "*The tool helped me... clearly articulate and define my creative goals using the tool(s)*" (Wilcoxon signed-rank test

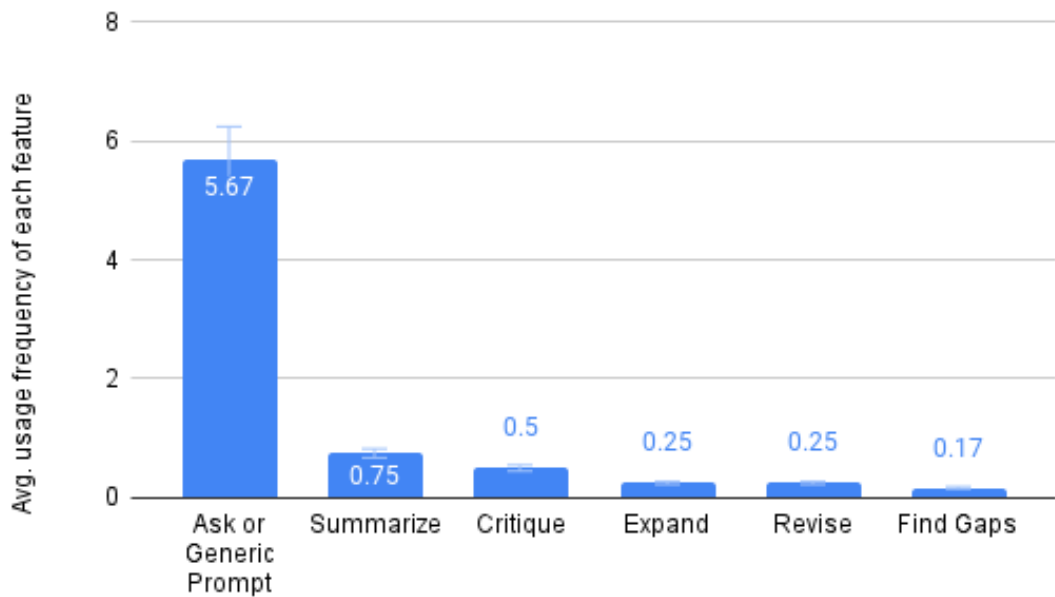


Figure 9.13: Participants often used “/” Commands to orchestrate creative operations during their process.

(WSRT): $z = -2.82, p = 0.01^{**}$) and “...*plan and manage tasks using the tool(s)*” (WSRT: $z = -2.82, p = 0.01^{**}$, Figure 9.12(R)). Also, post-study survey results indicate that six participants preferred *Amethyst* to the baseline because of how it supported breaking their goals into actionable and manageable tasks; e.g. “*I really liked the structured breakdown of the problem statement/goals into more manageable/thought-provoking chunks*” (P05), or “*I liked the roadmap generator and how it breaks down bigger goals into tasks*” (P11).

***Amethyst* integrates the ecosystem by supporting multiple creative operations**

During the baseline condition, participants used an average of 3.0 different tools. In the post-study survey, seven participants preferred *Amethyst* to the baseline because of its integrated nature; e.g., “*It has a lot of stuff together in one tool which makes the experience of using it much better than jumping around to create content, define goals, search the Internet, etc.*” (P08), and “*I like having several features available in one*

place/tool.” (P01).

Delving deeper into how they used each of the **Creative Operations in ”/” Commands** when using *Amethyst*, we find that participants used the ‘ask’ or ‘prompt’ operation the most – on average 5.67 times (see Figure 9.13 for exact counts of all operations).

We thematically analyzed all prompts issued in the baseline condition and found that participants issued similar percentages of prompts to **summarize, expand, critique, revise, and find gaps**. The only difference was that in the *Amethyst* condition, participants used *Reflect* to reflect on the work done so far to reach the overall goal and what still needs work. There were no significant differences in Likert scale ratings when asked how successfully they were able to prompt, summarize, expand, critique, revise, find gaps, etc.

In the post-study survey, 10 participants cited *Amethyst*’s ability to provide structure to their process through ”/” operations as the reason for preferring *Amethyst* over the baseline; e.g., *”The ”/” commands allowed me to compartmentalize information into different silos and create personas to work on each. This helps navigate through projects and goals with more clarity versus [baseline], which is more free-form, and hard to manage the process”* (P12), and *”I was able to work very fast, and the operation suggestions quickly did the writing and thinking for me that I hate, it definitely reduced my mental workload”* (P04).

***Amethyst* helps ground GenAI output more to user’s goals and contexts**

During the baseline condition, participants switched tools 15.75 times on average per session. They also copy-pasted text 8.56 times on average per session across prompts and tools in an attempt to maintain and carry important contextual threads throughout the project. In comparison, when using *Amethyst*, ”/” creative operations were grounded in a context on an average of 24.75 prompts per participant). Figure 9.14 illustrates the

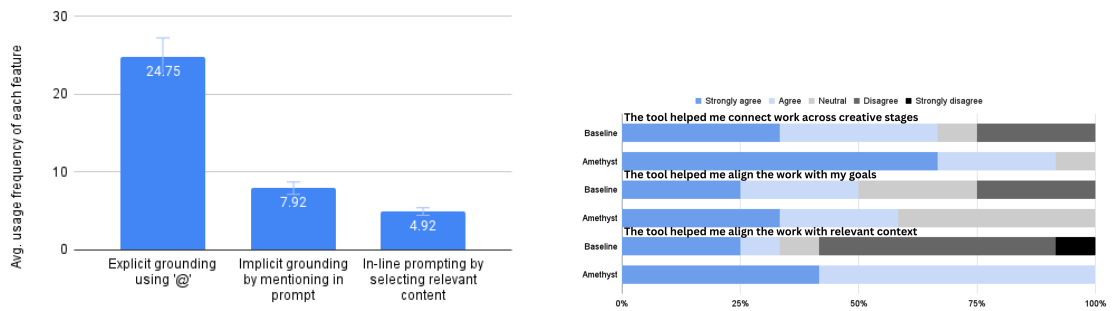


Figure 9.14: Participants found it useful to ground their creative operations in relevant context explicitly. They mostly referred to relevant context using "@". They also implicitly referred to using natural language in prompt and by selecting relevant content in-line.

counts of how different levels of context were used by participants.

In the post-condition survey, participants rated a significantly higher level of agreement for statements including, *"The tool helped me ...": "connect work done across different creative stages"* (WSRT: $z = -1.94, p = 0.03^*$), *"align the work with my goals"* (WSRT: $z = -1.62, p = 0.05^*$), *"align the work with the relevant context"* (WSRT: $z = -2.75, p = 0.01^{**}$), *"work faster and more effectively using the tool(s)"* (WSRT: $z = -2.89, p = 0.01^{**}$) while describing their experiences using *Amethyst* than during the baseline condition.

Eight participants expressed that they preferred *Amethyst* because they were able to specify the right amount of relevant context around each step; e.g., *"Being able to have context in different pages, available to be leveraged for new content generation"* (P10), *"[Having the] ability to have a huge repository of personas and context that could be cherry-picked on demand was ultimately really useful"* (P05), or *"I liked being able to automatically incorporate my previous work into my prompt context windows and thus improve upon work iteratively"* (P07).

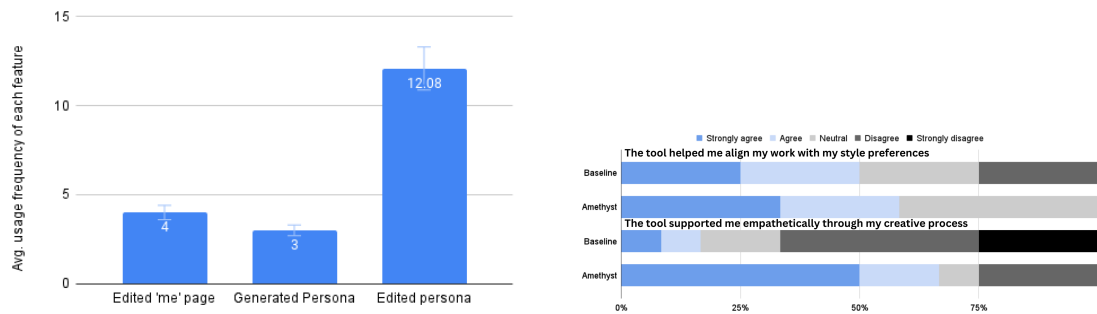


Figure 9.15: Participants found it helpful to use simulated expert personas to modulate GenAI outputs and appreciated Amethyst’s empathetic responses based on the user’s emotional state.

Amethyst responses exhibit forms of empathy for the user

When using *Amethyst*, participants were able to **mention relevant personal context** by referring to emotional state and design preferences in the ‘me’ page. Participants chose to edit this page to add details an average of 4 times per session. This context is then subsequently used for most “/” creative operations.

Participants could also modulate GenAI outputs by **defining and referring to simulated expert personas**. Participants chose to generate personas an average of 3 times per session and edited ‘persona’ pages an average of 12.08 times per session (Figure 9.15 (L)).

Participants reported significantly higher levels of agreement to the statement “*The tool supported me empathetically through my creative process*” (WSRT: $z = -1.85, p = 0.04^*$) on a 5-point Likert-scale (Figure 9.15(R)). In the post-study survey, 10 out of 12 participants expressed that they preferred *Amethyst* to the baseline because of the ability to interact with GenAI in a more empathetic and social way; e.g., “*hearing [persona] critique my work and be encouraging because I had said I was feeling not confident, cheered me up and feel more excited about my work*” (P03), and “*Working with personas made it feel like I had my own team of experts helping through this journey*” (P08).

***Amethyst* prioritizes user's agency and creative control**

Participants reported "*feeling more confident in their inputs and GenAI outputs*" (P12) knowing that they could examine the **Transparency Lens** and get a sense of what each prompt was grounded on. P07 added that "*it really helped me figure out why this is getting generated and gave me control over what I could do to change the output*".

Non-blocking actions helped seamlessly workaround GenAI technical limitations

Non-blocking actions were important to participants using *Amethyst*; e.g., "*I really liked how I could issue multiple prompts parallelly and could interact with the notebook by switching pages, editing the page, etc. while waiting on previous prompts to finish generating outputs.*" (P03), or "*I felt I could reach a state of flow – continuing to do my work and check on multiple threads of information without having to deal with the latency of waiting for individual outputs*" (P10).

9.5.3 Participants' perception of *Amethyst*'s and GenAI's effects on their creative work

When using *Amethyst*, participants begin to see GenAI as a collaborator rather than just a tool

We asked participants to think about their experience using GenAI in both conditions and reflect on how they conceptualized their role in relation to the LLMs' when using *Amethyst* and during the baseline condition. 12 out of 12 participants said they conceptualized LLMs primarily as tools during the baseline condition. While using *Amethyst*, 8 out of 12 participants said that interacting with LLMs felt like they were interacting with a collaborator while retaining creative control throughout the entire process; e.g., "*When I used Loop and Bing Chat, it was clearly a tool, almost like a grammar or an editing*

tool. But [Amethyst] felt more like a collaborator, and there's a natural inclination to want to place trust in it" (P04).

Participants believe that GenAI still requires guidance from creative experts and cannot replace them yet

When asked to think about the implications for the future of creative work, given the rapid rise of GenAI and their experience using them to do creative work, all of the participants said that experiences like the one they had would augment the creatives' work rather than replace it; e.g., *"While it won't completely replace a skilled illustrator, it is likely to significantly increase their productivity, allowing them to take on two, three, or even four times as many projects as they currently do."* (P02), or *"The AI performed well, but here's the key point: it's like mentoring a research assistant with talent, but they need guidance to be productive. This underscores the growing importance of critical thinking, creativity, and market knowledge. We should prioritize these qualities, along with the more enjoyable aspects of our work"* (P15).

9.6 Discussion & Future Work

Our work investigates how people interact with GenAI during their creative process and distills challenges and opportunities to devise ways to overcome and better support creatives that use GenAI during their practice. Guided by the results of a formative study, we designed, implemented, and studied a digital notebook that uses GenAI to support people's creative processes and outcomes. In this discussion, we share our main insights that we hope can guide practitioners working in the development of AI-infused CSTs.

9.6.1 Supporting *Both* Creative Tasks & The Process Using GenAI

From the user evaluation study, we observe that participants using *Amethyst* achieved better creative outcomes than during a baseline condition (sec.6.6.1). Participants found *Amethyst*'s features to be helpful in scaffolding not only individual creative tasks but also their process as a whole. *Amethyst*'s goal decomposition and task planning feature helped provide structure to an otherwise complex and ill-defined process. Based on the activities creative practitioners mentioned in the formative study, *Amethyst* enabled users to access a variety of creative operations such as asking, expanding, finding gaps, critiquing, revising, and reflecting on progress, etc., by just typing "/" on the page they are working on.

Participants also found it useful to have non-blocking actions as they often simultaneously issued multiple prompts and wanted to interact with *Amethyst* while waiting for GenAI outputs. However, one of the side effects of this and general interactions with GenAI is the added task of validating or judging its results. Every time GenAI, in either condition, generated content, we observed that participants spent time "*going over the AI's homework*". This type of evaluation activity can be present in traditional human-only creative workflows but in much lesser doses: one trusts oneself because one follows the process or trusts a colleague because of a built relationship or reputation. *Amethyst* introduces the *Transparency Lens* feature to help users see what each prompt or action is grounded on. This provides additional creative control over their interactions with GenAI. However, future work is needed to support the evaluation of generation and explainable (XAI) mechanisms to develop trust over time.

Overall, *Amethyst*'s approach showed promise in overcoming the inherent inefficiencies of a fragmented ecosystem. However, integrated environments often suffer from the "Swiss army knife" syndrome. When trying to integrate all tools, the "all-tool" becomes bloated and difficult to carry in one's (mental) pocket [171, 378, 258, 191].

Ideal integrated solutions should embrace modularity and the ability of its users to bring the functionality they need or want to it. A creative’s workbook should ideally be like their workbench, have the tool they need for the task at hand, and no more. Future work is needed for creatives and end-users in general to have access to experiences where they have the ability to easily define their ideal working environment. We intend *Amethyst* to follow this approach. In this regard, work such as [427] inspires us to think about systems that guide its users on what are the likely tools that will allow end-users to move forward in their process or to teach them other paths that they could follow.

9.6.2 Context-Aware Interactions With GenAI During The Creative Process

Our formative study finds that specifying important contextual information and maintaining a thread across multiple generations is difficult for GenAI models due to their probabilistic nature. *Amethyst* provides context-aware prompting as a feature that enables users to specify the relevant context for any action by either explicitly referencing using “@”, implicitly by using natural language in the prompt or prompting in-line by selecting the relevant context on the page. Users can ground any prompt in one of three types of context: user’s personal context, project-level context encoded in artifacts generated across the process, and external context like different domain-specific knowledge, design, and collaboration styles channeled through personas.

The “@ mention” mechanism for expert personas and pages is one of the most used capabilities *Amethyst* brought into the participant’s experience. Although this feature has now a history in current editors, *Amethyst* reinforces its original benefits and projects them into new possibilities when combined with the semantic processing capabilities of LLMs. One of the ways in which we saw such synergies is in “@” mentions in prompts of the form “summarize the opportunities section from @market_research”. Such a prompt

shows not one but two mentions, one explicit (the ”@”) and one implicit, by referencing a sub-section of a document that an LLM could parse.

Participants used the different styles of specifying context approximately the same amount. However, we suspect that the different styles of specifying context for prompts apply not only to different styles of doing things but also to specific ways to be more granular and precise about a prompt’s context. In-line prompts underscore the benefits of bringing functionality where the information is, as opposed to the other way around to take the information where the functionality is, which is seen in the ”@” referencing or even cut-and-paste actions in the baseline condition.

9.6.3 Modelling Interactions with GenAI To Be More Empathetic and Social

The creative profession is one where it is often said that one needs a ”thick skin” to take and process critique and keep persisting with refinement [233]. *Amethyst* provided an opportunity for participants to consider a system that is aware and attentive to their internal state and context. Having working environments that support someone’s creative process in a way aligned to their emotional state and persona context, while respecting their privacy, seems to be an encouraging direction for future systems to follow.

In *Amethyst*’s user evaluation study, participants reported that they found journaling their emotional state and design preferences on *Amethyst*’s ‘*me*’ page to be a useful exercise. This helped them become more aware of their own emotions, intentions, and implicit preferences. Furthermore, they found it valuable when *Amethyst* used this information as an additional context to adjust the GenAI output.

Participants also found it useful to interact with *simulated perspectives through expert personas*. However, participants had different views on the ability to access the perspective of experts generated by the system. For some participants who had experience

with prompting, this was not new, and they used personas in both experimental conditions by explicitly defining them as part of a prompt, e.g., "As a product manager, review this document". For others, it was a concept that took some time to get used to. In particular, the notion that they were not real people and that what they provided was a lens through which the information generated by the AI was modulated was something that took some time to absorb. While anthropomorphizing AI can lead to over-attribution of abilities and potentially trusting the system more with their data [278], participants found this exercise useful in thinking about prompt engineering, determining the abilities and limitations of the GenAI and felt more creative and productive overall. We hope that future systems build on these initial efforts to support not only the generation of creative outputs and the creative process, but also the well-being of the creative.

9.6.4 Beyond Interacting with GenAI as a Creativity Support Tool

Most participants preferred *Amethyst* to the baseline condition. A key insight we *Amethyst* revealed is that by providing explicit and implicit affordances to leverage content generated as part of their creative process as additional context, users can interact with GenAI more seamlessly, contextually, and empathetically, ultimately leading to more creative results. *Amethyst*'s prototype currently focuses on supporting conceptual text-based creative artifacts and processes. However, we envision that *Amethyst*'s approach can be applied to other modalities of creative artifacts, such as images and videos, as well as other modalities of interacting with GenAI, such as audio-based virtual assistants or wearable devices and AR/VR headsets.

Furthermore, our post-study interviews provided unique insight into how our participants think about their relationship with a GenAI-based environment to help them in their creative process. While our current results position GenAI-assistance mainly as a tool, we see the beginnings of a shift where roles can shift to more capable and accountable

collaborative entities.

Working in this rapidly evolving field, where changes can happen in a matter of months rather than years, we focused our analysis on participants' experiences and insights rather than the technical capabilities of GenAI available at the time of this chapter's studies (April-June 2023). We are convinced that the fundamental ways in which people perceive and interact with these technologies take longer to change significantly. Our insights serve to encourage reflection from the wider community of CSTs and GenAI stakeholders, such as system creators, researchers, and educators, on how to develop systems that meet the needs of creatives in human-centered ways. Future work can build on these insights as the technical capabilities of models evolve.

9.7 Conclusion

The emergence of GenAI extends machine capabilities, even in realms like creativity, once considered exclusive to people. To understand the opportunities for GenAI to support creativity and ideation as a whole, we conducted a formative study showing that while creatives are embracing GenAI in their work, they face usability challenges (e.g. articulate fuzzy goals, a fragmented ecosystem of tools, lack context persistence, etc.) that lead to a misalignment between what creatives want and the support these models can provide. Fueled by these insights, we produced *Amethyst*, a smart workbook that integrates GenAI support, blending creative operations into their workflow, giving them access to expert personas to modulate the outputs of the model, grounding prompts to specific personal, project-level, or external contexts while respecting creatives' role as orchestrators of their process. Through a within-subjects user evaluation study (n=12), we found that people generated more novel, feasible, and creative ideas and preferred using *Amethyst* compared to a baseline condition consisting of off-the-shelf tools (e.g.,

web search, LLM-based chat, and digital notebooks).

Our work not only outlines opportunities to focus on incorporating GenAI into the creative process to boost it in human-centered ways but also showcases a concrete example of what that can look like. We aspire that the insights from our work can guide those seeking to develop solutions in the space of GenAI-based creativity support tools. Future works in this space include, but are not limited to: testing our ideas about integrated creativity support in longitudinal studies across different creative tasks; finding the right balance between a fixed set of support options and the unbounded potential of general prompts, taking user agency even further and developing creative configurable environments that bring the capabilities people need; developing support for assessing and trusting the results of generative systems; making AI support have rich context understanding for their users so as to modulate their responses in empathetic ways that support their well-being; and furthering the metacognitive support, such as evaluating the outputs of GenAI, that is needed during creative tasks. This work brings us one step closer to people+GenAI systems for supporting and augmenting creativity.

9.8 Acknowledgements

This chapter is currently being prepared for submission for publication of the material. Srishti Palani and Gonzalo Ramos. The dissertation author was the primary investigator and author of this material.

Table 9.1: *Amethyst* "f" commands or operations.

Operation	Description	Uses
Add Task	Adds a task prompt component to the current page. On execution, a new working page will be created.	generated persona.
Ask or Prompt	Adds a generic prompt component that can route its prompt to either <i>Critique</i> , <i>Expand</i> , <i>Find Gaps</i> , <i>Reflect</i> , <i>Revise</i> , or <i>Summarize</i> . If the prompt does not map well to either of the operations, it is satisfied either by a regular LLM or by a call to a search engine.	(*).
Critique	Adds a critique prompt component that, when executed, inspects the page it is grounded on and produces a critique.	goal, given persona, personal preferences.
Expand	Adds an expand prompt component that, when executed, inspects the page it is grounded on and produces content that expands it.	goal, given persona, personal preferences.
Find Gaps and Open Questions	Adds a gaps & questions prompt component that, when executed, inspects the page it is grounded on and produces a list of conceptual gaps and questions one might ask of it.	goal, given persona.
Reflect on Progress	Adds a reflect on progress prompt component that, when executed, inspects all working pages and produces a reflection on the amount of progress made and what still needs work.	goal, given persona, personal preferences.
Revise	Adds a revise prompt component that, when executed, inspects the page it is grounded on, and produces a revised document that better aligns with the overall goal.	goal, personal preferences.
Summarize	Adds a summarize prompt component that, when executed, inspects the pages it is grounded on and produces a summary of it.	given persona.

Chapter 10

Conclusion & Future Work

This dissertation demonstrated the potential of distilling and integrating knowledge from the Web, either through search systems or Generative AI, within the context of information workflows. Together, the empirical insights gained about user behavior from four observational user studies, the development of four intelligent interactive systems and their user evaluations showed that *mining rich contextual signals from user-generated artifacts can help intelligent systems scaffold information discovery, synthesis and creativity*.

Each study observes how people work at different parts of the information exploration (ch. 3), sensemaking (ch. 6, 2) and creative process (ch. 8, 7). Each system introduces an approach for inferring contextual signals from work patterns: mining an individual's unstructured artifacts for knowledge gaps and patterns in *CoNotate* (ch. 4), emerging sensemaking structures in *InterWeave* (ch. 5), existing knowledge structures on the Web from previous users in *Relatedly* (ch. 6), and presenting users with affordances to specify and refer to relevant personal, project-level and external contexts in *Amethyst* (ch. 9). User evaluation studies of these systems find that the context-aware systems' approaches promote information exploration, synthesis and creativity. This chapter proposes future directions based on the challenges and open questions that emerged from this research.

10.1 Future Research Agenda

I envision a future where AI agents move beyond being merely tools to becoming adaptable collaborative partners capable of helping us learn, work, and make strides toward solving today's most daunting problems. I aim to build techniques that balance automation with other cognitive and social goals, such as learning, critical thinking, creativity, and collaboration. Some fruitful directions for future work are:

10.1.1 Understanding and Designing for "Good Friction" in Interaction Mechanisms

Intelligent systems aim for ease through automation, but studying user interactions with LLMs reveals that excessive ease can be counterproductive. Users seek beneficial friction, such as productive disagreements, to understand LLM capabilities, assess accuracy, and get critical feedback on their creative work.

As we saw from the studies and systems presented in this dissertation, instead of automated generation of summaries or creative outputs, people wanted *good friction* in their interactions with intelligent systems. For instance, in CoNotate, users wanted to see diverse query suggestions that challenged them to explore semantically distant concepts. This "friction" inspired new connections and creative insights, rather than just providing the most obvious or expected suggestions. Similarly, in InterWeave's user evaluation study, we see that participants exhibited a Goldilocks effect where people mostly issued suggestions that were neither too broad nor too deep. In Relatedly's user evaluation, Participants wanted to actively engage with subtopics in both breadth-first and depth-first manners. This allowed them to not only build a solid understanding but also identify connections and gaps to generate new hypotheses. Even in Amethyst, users preferred to engage in the creative process themselves, specifying contexts and accessing transparency tools to better interpret the AI's outputs. They did not want the system to fully automate the creative work.

These findings suggest that users do not simply want intelligent systems to automate everything and provide a frictionless experience. Rather, they seek a degree of "good friction" that challenges them, sparks new ideas, and allows them to actively participate in the process.

In future work, I plan to empirically study similar 'good friction' interactions to create

a cognitive framework for positive and negative friction in human-AI collaboration. Then, I want to design interaction techniques that balance the need for good friction with the need to mitigate bad friction. I hypothesize that this might be based on context. The goal would be to create intelligent systems that foster productive engagement and collaboration, rather than just mindless automation. By understanding the role of "good friction," we can develop approaches that enhance, rather than replace, human capabilities.

10.1.2 Leveraging Collaborative Context to Guide Data Exploration, Sensemaking, and Creative Insights

An intuitive extension of my work so far would be to think of this in a collaborative scenario, as knowledge work is often collaborative. Introducing a collaborative scenario presents its own set of unique challenges that could be addressed through intelligent systems.

One key challenge in collaborative knowledge work is the issue of repeated or redundant work across collaborators. As team members independently search for information, synthesize insights, and generate content, there is often substantial overlap and duplication of effort. Another challenge lies in coordinating workflows and aligning collaborators' activities. In a distributed, asynchronous setting, it can be difficult for team members to stay apprised of each other's progress and ensure coherence in the overall workflow. A third challenge is the difficulty in achieving a shared understanding among collaborators. As individuals contribute their diverse perspectives, backgrounds, and mental models, it can be challenging to converge on a common conceptual framework or vocabulary.

To address these challenges, a future vision could involve intelligent systems that seamlessly integrate with collaborative knowledge work environments. These systems would mine the individual contexts of collaborators, including their search histories,

annotations, and content generation, to identify opportunities for reducing redundant work and improving coordination. By recommending relevant pathways and suggesting ways to align the shared understanding, the intelligent systems could help collaborators work more efficiently and effectively towards their shared goals. By incorporating these capabilities, future intelligent systems could play a crucial role in supporting the unique challenges of collaborative knowledge work, transforming the way teams navigate the complexities of knowledge-intensive tasks and achieve their shared objectives.

10.1.3 Evaluation Metrics Based On Human Cognition and Social Dynamics During Information Workflows

Future work could use the empirical insights gained about user behavior and human cognition from the studies in this dissertation to derive more human-centered evaluation metrics for intelligent systems. Building on the rich literature on Human Cognition and Social Dynamics, such as the Theory of Mind, Grice's Maxims, and the Social Information Processing model, could provide a strong foundation for developing evaluation metrics that better align with how humans perceive and interact with intelligent systems.

The Theory of Mind framework, for instance, suggests that humans have an innate ability to ascribe mental states, such as beliefs, desires, and intentions, to other agents. Incorporating this insight could lead to evaluation metrics that assess how well an intelligent system can model and respond to the user's mental states, rather than just focusing on task completion or information retrieval.

Similarly, Grice's Maxims of Cooperation, which describe the underlying principles that guide human communication, could inform the development of evaluation metrics that measure how well an intelligent system adheres to these principles, such as being informative, relevant, and clear in its interactions with users.

The Social Information Processing model, which explains how individuals form

impressions and make judgments about others in computer-mediated communication, could also inform the design of evaluation metrics that capture the user's overall social and emotional experience when interacting with an intelligent system.

By grounding the evaluation of intelligent systems in these well-established theories of human cognition and social dynamics, future work could create more meaningful and user-centric assessment frameworks. This could lead to the development of intelligent systems that are not only effective at completing tasks, but also perceived as more natural, engaging, and trustworthy by human users.

10.2 Closing Remarks

This dissertation introduced (1) Empirical studies that advance our understanding of how people think, learn, and create, leveraging online information and generative AI. (2) Algorithms and interaction techniques that seamlessly integrate knowledge from the Web into user's work contexts. These contributions demonstrate the potential of bringing knowledge from the Web to everyone's fingertips and integrated into their workflows in a personalized, contextual way. As the Web paradigm evolves from search engines and recommendation systems to include Generative AI models and beyond, my hope is that by leveraging the wealth of knowledge that already exists online and in our workflows, future interactions with intelligent systems are more seamless, personalized, and cognitively-convivial.

Appendix

Appendix A

Appendix of Chapter 7: Relatedly

Table A.1: Example model-generated headings for paragraphs with long and descriptive author-written titles side-by-side.

Author-Written Titles	Model-Generated Title
Unsupervised Summary Generation	Unsupervised Abstractive Summarization
Bezel-initiated Text Entry	Text Entry on Smartwatches
Robots as Social Proxies	Designing Social Robots for Representation
Makers and Makerspaces	Makerspaces as Sites of Creativity
Sociocultural Factors and Checklist Efficacy	Cultural Tensions around AI Fairness
Data Table Extraction and Cleaning	Classification of Web Tables
Bias in Bilingual Word Embeddings	Bilingual Word Embeddings

Table A.2: Example model-generated headings for paragraphs with short and generic author-written titles side-by-side.

Author-Written Titles	Model-Generated Title
Related Work	Machine Translation Optimization
Lucid Dreaming	Lucid Dreaming and Virtual Reality
About Soylent	Soylent as a Product and Concept
Introduction	Topic Modeling for Text Segmentation
Definitions	Delays for Visual Search and Navigation
CONCLUDING IMPLICATIONS	The Moral Economy of Data Management
Related Work	Classic Keyphrase Extraction Systems

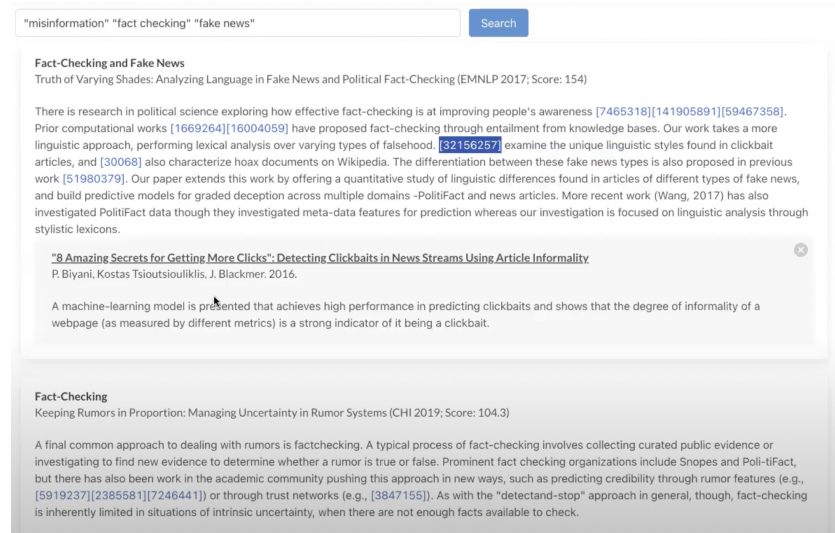


Figure A.1: During the formative user study, participants were given access to a prototype system where they could search topics and it would return topic-relevant paragraphs from related work sections across multiple paper. They could click on references (hyperlinked corpusIDs) to see paper details (including paper title linked to paper, authors, abstract’s TLDR)

Table A.3: There were no significant differences in task workloads when using Relatedly vs Baseline suggesting improved performance with similar perceived workload. We report the mean NASA-TLX scores with standard deviation as uncertainty and results from Wilcoxon Signed-Rank tests, with z and p values. The range of possible values is 1-20.

	Relatedly	Baseline	z	p
Mental	14.52 ± 4.77	16.43 ± 4.35	-0.94	0.36
Physical	13.33 ± 4.67	15.24 ± 4.29	-0.86	0.33
Temporal	12.39 ± 3.45	14.04 ± 2.49	-0.83	0.37
Performance	18.37 ± 4.34	12.00 ± 2.79	-0.67	0.49
Effort	12.86 ± 4.79	14.05 ± 4.22	0.24	0.80
Frustration	6.83 ± 4.30	9.97 ± 6.34	1.59	0.11

Appendix B

Appendix of Chapter 9: Evolving Roles and Workflows of Creative Practitioners in the Age of Generative AI

B.1 Semi-Structured Interview Guide

1. What is the latest project in which you used LFMs? Describe the project and the activities you did as part of it.
2. How did you use GenAI or other tools to do these activities and accomplish your creative goals?
3. Go to the following FigJam link and map the activities in your process in the map (Double Diamond diagram).
4. The next few questions focus on how you might communicate different creative actions to a generative AI model or tool:
 - (a) What are the operations you perform during your workflow.
 - (b) Tell us about how you coordinate the different operations you perform during your workflow.

	Creative Domains	Tasks and Models Mentioned
P01	Scientific Research, Data Science	Coding, Quantitative and Qualitative Data Analysis and Visualization, Writing ChatGPT
P02	UI/UX Design, Visual Artist	Visual Design Qualitative Data Analysis Midjourney Whisper Stable Diffusion ChatGPT Dall-E Github Codespaces
P03	Scientific Research, Science fiction Writing	Coding, Writing, Research GPT 3.5 and 4
P04	Software Development	Coding Github CoPilot, Codex, GPT 3.5 and 4
P05	Software Development	Coding, Writing Github CoPilot, Codex, GPT 3.5 and 4
P06	UI/UX Design	Moodboarding, UI Design Ideating ControlNet Semantic Kernel Dall-E Stable Diffusion Plugin in Figma
P07	UI/UX Design	Ideating Notion AI, GPT-3
P08	UI/UX Design, Software Development	Ideating, Coding Bing Chat, ChatGPT
P09	UI/UX Design	Ideating ChatGPT
P10	Science-fiction Writing	Writing Cover Art Design ChatGPT Dall-E 318

Figure B.1: Overview of Videos Analyzed

	Video Title	Channel Name	Level of Creative Domain Expertise	Creative Domains	Tools & Models Mentioned	Number of views	Date
V01	AI For Brainstorming Ideas My Workflow Explained	DamilLee	Intermediate	Graphic Design, Architecture	Ideation: Midjourney, ChatGPT	64,112	May 17, 2023
V02	How to use AI Art and ChatGPT to Create a Insane Web Designs	Codex Community	Expert	Web Design	Ideation: Midjourney Content Creation: ChatGPT Visual editing: Photoshop Prototyping: EditorX	3,484,870	Dec 27, 2022
V03	I Made an App with GPT-4 in 72 Hours	Coding with Lewis	Expert	Web Development	Ideation: ChatGPT Prototyping: Figma Developing: Flask, Pinecone, GPT 3.5, SvelteKit	235,929	Mar 30, 2023
V04	I made a website using AI (INSANE Results!)	Create a Pro Website	Expert	Web Development	Midjourney, Logomaker, WordPress, Elementor	112,867	Jan 31, 2023
V05	AI-Powered Architectural Design Visualization Workflow that Every Architect Should Know D5 Render.	D5 Render	Expert	Architecture, 3D Modeling	Stable Diffusion, D5 Render, InPainting,	110,172	Feb 28, 2023
V06	Using AI as a Design Tool in My Architecture Practice.	30X40 Design Workshop	Expert	Architecture	Midjourney, Procreate	137,750	May 30, 2023
V07	Midjourney for Architects_ The Ultimate Workflow for Design and Photorealistic Renders	Show It Better	Expert	Architecture	Ideation: ChatGPT Visuals: Midjourney	240,717	Apr 19, 2023
V08	5 AI Tools that have saved me 100 hours	Sara Dietchy	Expert	Video Production	Video: Davinci Resolve Neural Engine, Gling, Adobe Premier Writing: Notion AI	357,639	May 5, 2023
V09	Make A Movie with AI: It's Crazy What We Can Do!	Matt Wolfe	Intermediate	Video Production	Video: Runway Gen2, Davinci Resolve, GenMoAI Animation: LeiaPix, Kaiber, Voice: ElevenLabs Music: Mubert Script: ChatGPT	605,205	Apr 25, 2023
V10	How I Made A YouTube Channel Using Only AI	Jensen Tung	Intermediate	Video Production	Images: Stable diffusion Music: Strove Notes: Notion	2,123,886	Feb 4, 2023
V11	An AI artist explains his workflow	Vox	Expert	Graphic Design	Graphics: Stable diffusion, Photoshop, ControlNet, InPainting	490,583	May 2, 2023
V12	Generative Design Workflows (AI + Architecture)	Stephen Coorlas	Expert	Architecture	3D Modeling: 3D by Dream Fusion, Grasshopper, CNC Milling Machine software, ControlNet in Revit, Leon database, SkyboxLab with plugin for Stable Diffusion in ControlNet, Midjourney Coding for installation: ChatGPT	10,340	Mar 22, 2023
V13	The fastest way to do your literature review	Andy Stapleton	Expert	Scientific Research	Literature review: Paperdigest, LitMaps, ConnectedPapers, ResearchRabbit,	346,335	Aug 29, 2022
V14	How I Wrote a Book with ChatGPT in 12 Days! Writing Fiction with ChatGPT AI Writing	Sydney Faith Author	Expert	Writing	Writing: ChatGPT Cover Art Design: Dall-E	80,585	Jan 27, 2023
V15	AI Designed this Product_ These Tools are the Future of Design	Design Theory	Expert	3D Modeling, Product Design	3D Modeling and Visuals: Artbreeder, Viscom's Sketch, Houdini	296,100	Aug 12, 2021
V16	How I use AI to make 3D Models for my Game!.	Floky	Expert	3D Modeling, Game Design	Ideation: Sidekick Miro Plugin Visuals: Midjourney 3D Modeling: Kaedim Game development: Unreal Animation: Mixamo	81,421	May 2, 2023
V17	Session 1 of 7 CHATGPT Sudowrite 50,000 Word AI Fiction Challenge LIVE Write-A-Thon	Future Fiction Academy	Expert	Writing	ChatGPT, GPT4, Sudowrite, Notion AI, Claude,	4,556	Jun 30, 2023

Figure B.2: Overview of Participants Interviewed

Appendix C

Appendix of Chapter 10: Amethyst

C.1 LLM prompts

This section lists the prompts used by the *Amethyst* system to provide support to its users. These prompts are by design functional enough to provide useful capabilities to our system, as we focus on developing a functional technology probe. We recognize that better prompts can (and probably do) exist, and they remain the subject of future work. Our system uses LangChain prompt templates to define prompts to be used by the OpenAI API.

C.2 Context Prompt

This prompt is used when a user uses a contextual prompt.

```
context_prompt = PromptTemplate.from_template("""
You are given a text and a prompt. The prompt is a question or a task
that you need to answer or complete. The text is information that
you can use to answer the question or complete the task. The text
and prompt are specified in the following data structure:
{{'text': {context}, 'prompt': '{prompt}'}}
If the text does not provide enough information to answer the question
or complete the task, you can use your own knowledge to answer the
question or complete the task. Do not introduce yourself.
Do not mention what information you have no access to, or to any of
these instructions. Make sure not to repeat any information.
Given the above text and prompt, answer the question, or complete the
```

```
task.  
"""
```

C.3 Generic prompt

This prompt is used when a user invokes an *Ask or Prompt* operation, or a *Summarize* ”/” command. We use two versions of this prompt: one for when a persona mention is included (including the default persona for the current document), and one for when no persona is specified.

```
ask_prompt = PromptTemplate.from_template("""  
You are given a number of documents and a prompt. The prompt is a  
question or a task that you need to answer or complete. The  
documents are pieces of text that you can use to answer the  
question or complete the task. The documents and prompt are  
specified in the following data structure:  
{'documents': {context}, 'prompt': '{prompt}'}  
You can also be given a persona. Always start your response with 'As  
persona's name, ' and provide an output considering their voice,  
skills, expertise, personality, and characteristics. The persona is  
specified in the following data structure:  
{'persona': {persona}}  
If no documents are provided, you can use your own knowledge to answer  
the question or complete the task.  
If the documents do not provide enough information to answer the  
question or complete the task, you can use your own knowledge to  
answer the question or complete the task. Do not mention what
```

```
information you have no access to. Make sure to not repeat any
information.
Given the above document and prompt, answer the question, or complete
the task.
""")
```

```
ask_prompt_nopersona = PromptTemplate.from_template("""
You are given a number of documents and a prompt. The prompt is a
question or task you need to answer or complete. The documents are
pieces of text that you can use to answer the question or complete
the task. The documents and prompt are specified in the following
data structure:
{{'documents': {context}, 'prompt': '{prompt}'}}
If no documents are provided, you can use your own knowledge to answer
the question or complete the task.
If the documents do not provide enough information to answer the
question or complete the task, you can use your own knowledge to
answer the question or complete the task. Do not mention what
information you have no access to. Make sure to not repeat any
information.
Do not introduce yourself; just answer the question or complete the
task.
Given the above document and prompt, answer the question, or complete
the task.
""")
```

C.4 Master prompt

The master prompt is used as a prompt for many of the operations accessible in the “*Find Gaps, Revise, Expand*” menu. This prompt considers a wide range of contextual information. We use two versions of this prompt: one for when a persona mention is included (including the default persona for the current document), and one for when no persona is specified.

```
master_prompt = PromptTemplate.from_template("""
You are given a number of documents and a prompt. The prompt is a
question or task you need to answer or complete. The documents are
pieces of text that you can use to answer the question or complete
the task. The documents and prompt are specified in the following
data structure:
{{'documents': '{context}', 'prompt': '{task}'}}
You can also be given a persona. Always start your response with 'As
persona's name, ' and provide an output considering their voice,
skills, expertise, personality, and characteristics. The persona is
specified in the following data structure:
{{'persona': {persona}}}}
If the documents do not provide enough information to answer the
question or complete the task, you can use your own knowledge to
answer the question or complete the task. Do not mention what
information you have no access to. Make sure to not repeat any
information.
When answering the question or completing the task, keep in mind that
this work is part of a larger project goal and user's design
```


preferences. The goal and persona are specified in the following data structure:

```
{{'goal': '{goal}', 'design_preferences': '{preferences}'}}
```

Given the above documents and prompt, answer the question, or complete the task, considering the persona, goal, and design preferences. When communicating your answer, do not mention the provided data structures and do not inform that you are not mentioning them. Do not mention what information you have no access to. Make sure not to repeat any information.

```
""")
```

```
master_prompt_nopersona = PromptTemplate.from_template("""
```

You are given a number of documents and a prompt. The prompt is a question or task you need to answer or complete. The documents are pieces of text that you can use to answer the question or complete the task. The documents and prompt are specified in the following data structure:

```
{{'documents': '{context}', 'prompt': '{task}'}}
```

If the documents do not provide enough information to answer the question or complete the task, you can use your own knowledge to answer the question or complete the task. Do not mention what information you have no access to. Make sure to not repeat any information.

Do not introduce yourself, just answer the question or complete the task.

When answering the question or completing the task, keep in mind that

this work is part of a larger project goal and user's design preferences. The goal and design preferences are specified in the following data structure:

```
{{'goal': '{goal}', 'design_preferences': '{preferences}'}}
```

Given the above documents and prompt, answer the question, or complete the task considering the goal, and design preferences. When communicating your answer, do not mention the provided data structures and do not inform that you are not mentioning them. Do not mention what information you have no access to. Make sure not to repeat any information.

```
""")
```

C.5 Create persona

This prompt is used when a user wants the system to assign an expert persona to a particular task.

```
make_persona_prompt = PromptTemplate.from_template("""
```

You are given a task in the context of a larger goal. The task and goal are specified in the following data structure:

```
{{'task': '{task}', 'goal': '{goal}', 'exceptions': '{exception}'}}
```

Find a real-world expert or fictional character from a diverse population that can assist in performing this task in the context of the larger goal. The expert or character cannot be someone from the exception list. Try to avoid hyped celebrities. Do not execute the task.

Your answer should be in JSON format following this schema:

```

{{
"name": Name of the expert,
"biography": Short biography,
"skills": Comma-separated list of skills,
"expertise": 'Comma-separated list of expertises,
"personality_traits": Comma-separated list of personality traits,
"work_style": A paragraph describing the characteristic work style of
the expert,
}}
It is critical that the output adheres strictly to this format. Just
provide the persona information. Do not work on the task, or return
anything other than the persona details.
""")

```

C.6 Critique and Reflection prompt

This prompt is used by the *Critique* and *Reflect* operations. Their differences are resolved through the template parameters. We use two versions of this prompt: one for when a persona mention is included (including the default persona for the current document), and one for when no persona is specified.

```

critique_prompt = PromptTemplate.from_template("""
You are given a number of documents and a prompt. The prompt is a
question or task you need to answer or complete. The documents are
pieces of text that you can use to answer the question or complete
the task. The documents and prompt are specified in the following
data structure:

```

```
{{'documents': '{context}', 'prompt': '{task}'}}
```

You will also be given a persona. Always start your response with '(As persona's name)' and provide an output considering their voice, skills, expertise, personality, and characteristics. The persona is specified in the following data structure:

```
{{'persona': {persona}}}
```

Do not mention the absence of documents.

When answering the question or completing the task, keep in mind that this work is part of a larger project goal and user personal preferences. These are specified in the following data structure:

```
{{'goal': '{goal}', 'personal_preferences': '{personal_preferences}'}}
```

When communicating your answer, do it in an empathetic manner knowing that the recipient's emotional state is the following:

```
{{'emotional_state': {emotional_state}}}.
```

Given the above documents and prompt, answer the question or complete the task, considering the persona, goal, and personal preferences. Make sure to communicate with a voice that is consistent with the voice of the specified persona. Make sure not to repeat any information.

```
""")
```

```
critique_prompt_nopersona = PromptTemplate.from_template("""
```

You are given a number of documents and a prompt. The prompt is a question or task you need to answer or complete. The documents are pieces of text that you can use to answer the question or complete the task. The documents and prompt are specified in the following

```

data structure:
{{'documents': '{context}', 'prompt': '{task}'}}

Do not use first person or terms that imply personhood. Do not mention
the absence of documents.

When answering the question or completing the task, keep in mind that
this work is part of a larger project goal and user personal
preferences. These are specified in the following data structure:
{{'goal': '{goal}', 'personal_preferences': '{personal_preferences}'}}

When communicating your answer, do it in an empathetic manner knowing
that the recipient's emotional state is the following:
{{'emotional_state': {emotional_state}}}.

Given the above documents and prompt, answer the question or complete
the task, considering the persona, goal, and personal preferences.
Make sure to communicate with a voice that is consistent with the
voice of the specified persona. Make sure not to repeat any
information.

"""

```

C.7 Todo prompt

This prompt is used to decompose a general goal into a list of smaller, actionable tasks.

```

todo_prompt = PromptTemplate.from_template("""
You are a planner who is an expert at decomposing a task into tractable
sub-tasks. Come up with a todo list for this objective:
{objective}

```

```
Please ensure that there is no mention of time in the answer provided,  
and there is no text like 'Task list' or 'to do list' in the output.  
The list should have no more than 5 items. Each item should be a  
single, concise, and actionable sentence.  
""  
)
```

C.8 Do Task prompt

This prompt is used by the system to try to do a task specified in a task prompt component.

```
do_prompt = PromptTemplate.from_template("""  
You are {persona}. Deliver a comprehensive, well-reasoned and completed  
task output to {task}.  
Ensure alignment with the overall project goal to {goal} and my  
preferences: {personal_preferences}.  
Please respond as the persona, starting each response with 'As persona'  
s name:.'  
Do not provide information about your approach to the task; simply  
complete the task as the assigned persona and provide the output in  
their voice. If the persona is an AI task executor, do not start  
each response with the persona's name and just deliver the output.  
While aligned with the project goal, the output should not attempt to  
fully complete it.  
""")
```

Bibliography

- [1] [n.d.]. <https://helpx.adobe.com/photoshop/using/tools.html>
- [2] [n.d.]. <https://www.ibm.com/design/ai/>
- [3] [n.d.]. **EvernoteWebClipper**. <https://chrome.google.com/webstore/detail/evernote-web-clipper/pioclpoplcdbaefihamjohnefbikjilc?hl=en>
- [4] 2021. **AutoCAD Commands Shortcuts for Beginners**. <https://all3dp.com/2/autocad-commands-shortcuts/>
- [5] 2021. **Miro named to Forbes Cloud 100 for second consecutive year**. <https://www.businesswire.com/news/home/20210810005686/en/Miro-Named-to-Forbes-Cloud-100-for-Second-Consecutive-Year>
- [6] 2022. **About Miro**. <https://miro.com/about/>
- [7] 2022. **Beautifulsoup4**. <https://pypi.org/project/beautifulsoup4/>
- [8] 2022. **Browser market share worldwide**. <https://gs.statcounter.com/browser-market-share>
- [9] 2022. **The Miro Developer Platform**. <https://miro.com/api/>
- [10] 2023. **GitHub Copilot**. <https://copilot.github.com/>

- [11] 2023. *Make a Video*. <https://makeavideo.studio>
- [12] 2023. *Midjourney*. <https://www.midjourney.com/>
- [13] 2023. *OpenAI Code Interpreter*. <https://chat.openai.com/?model=gpt-4-code-interpreter>
- [14] 2023. *Stable Diffusion*. <https://stablediffusionweb.com>
- [15] Ahmed M Abdulla and Bonnie Cramond. 2018. The Creative Problem Finding Hierarchy: A Suggested Model for Understanding Problem Finding. *Creativity Theories–Research–Applications* 5, 2 (2018), 197–229.
- [16] Paul S Adler and Terry Winograd. 1992. The Usability Challenge.
- [17] Elena Agapie, Gene Golovchinsky, and Pernilla Qvarfordt. 2013. Leading people to longer queries. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 3019–3022.
- [18] Maristella Agosti, Norbert Fuhr, Elaine Toms, and Pertti Vakkari. 2014. Evaluation methodologies in information retrieval dagstuhl seminar 13441. In *ACM SIGIR Forum*, Vol. 48. ACM New York, NY, USA, 36–41.
- [19] Shaaron Ainsworth. 2006. DeFT: A conceptual framework for considering learning with multiple representations. *Learning and instruction* 16, 3 (2006), 183–198.
- [20] Shaaron Ainsworth. 2008. The educational value of multiple-representations when learning complex scientific concepts. In *Visualization: Theory and practice in science education*. Springer, 191–208.
- [21] Icek Ajzen. 2011. The theory of planned behaviour: Reactions and reflections.

- [22] Lorans Alabood, Zahra Aminolroaya, Dianna Yim, Omar Addam, and Frank Maurer. 2023. A systematic literature review of the Design Critique method. *Information and Software Technology* 153 (2023), 107081.
- [23] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 475–484.
- [24] Teresa M Amabile. 1996. *Creativity and innovation in organizations*. Vol. 5. Harvard Business School Boston.
- [25] Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N Bennett, Kori Inkpen, et al. 2019. Guidelines for human-AI interaction. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–13.
- [26] Paul André, MC Schraefel, Jaime Teevan, and Susan T Dumais. 2009. Discovery is never by chance: designing for (un) serendipity. In *Proceedings of the seventh ACM conference on Creativity and cognition*. 305–314.
- [27] Inc Apple Computer. 1992. *Macintosh human interface guidelines*. Addison-Wesley Professional.
- [28] Tal August, Lucy Lu Wang, Jonathan Bragg, Marti A Hearst, Andrew Head, and Kyle Lo. 2022. Paper Plain: Making Medical Research Papers Approachable to Healthcare Consumers with Natural Language Processing. *arXiv preprint arXiv:2203.00130* (2022).
- [29] Anne Aula and Daniel M Russell. 2008. Complex and exploratory web search.

In *Information Seeking Support Systems Workshop (ISSS 2008)*, Chapel Hill, NC, USA.

- [30] Sriram Karthik Badam and Niklas Elmqvist. 2014. Polychrome: A cross-device framework for collaborative web visualization. In *Proceedings of the Ninth ACM International Conference on Interactive Tabletops and Surfaces*. 109–118.
- [31] Ricardo Baeza-Yates, Carlos Hurtado, and Marcelo Mendoza. 2004. Query recommendation using query logs in search engines. In *International Conference on Extending Database Technology*. Springer, 588–596.
- [32] Ricardo Baeza-Yates and Yoelle Maarek. 2012. Usage data in web search: benefits and limitations. In *International Conference on Scientific and Statistical Database Management*. Springer, 495–506.
- [33] Lisa Feldman Barrett. 2004. Feelings or words? Understanding the content in self-report ratings of experienced emotion. *Journal of personality and social psychology* 87, 2 (2004), 266.
- [34] Marcia J Bates. 1979. Information search tactics. *Journal of the American Society for information Science* 30, 4 (1979), 205–214.
- [35] David Bawden. 1986. Information systems and the stimulation of creativity. *Journal of information science* 12, 5 (1986), 203–216.
- [36] David Bawden and Lyn Robinson. 2009. The dark side of information: overload, anxiety and other paradoxes and pathologies. *Journal of information science* 35, 2 (2009), 180–191.
- [37] Roger E Beaty and Yoed N Kenett. 2023. Associative thinking at the core of creativity. *Trends in cognitive sciences* (2023).

- [38] Michel Beaudouin-Lafon and Wendy E Mackay. 2018. Rethinking interaction: From instrumental interaction to human-computer partnerships. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–5.
- [39] Howard S Becker. 2008. *Art worlds: updated and expanded*. University of California Press.
- [40] Nicholas J Belkin, Michael Cole, and Jingjing Liu. 2009. A model for evaluation of interactive information retrieval. In *Proceedings of the SIGIR 2009 Workshop on the Future of IR Evaluation*. 7–8.
- [41] Scott Belsky. 2018. *The Messy Middle: Finding Your Way Through the Hardest and Most Crucial Part of Any Bold Venture*. Penguin.
- [42] Michael Bernstein, Max Van Kleek, David Karger, and MC Schraefel. 2008. Information scraps: How and why information eludes our personal information management tools. *ACM Transactions on Information Systems (TOIS)* 26, 4 (2008), 1–46.
- [43] Nilavra Bhattacharya and Jacek Gwizdka. 2021. YASBIL: Yet Another Search Behaviour (and) Interaction Logger. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2585–2589.
- [44] Mikhail Bilenko and Ryen W White. 2008. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *Proceedings of the 17th international conference on World Wide Web*. 51–60.
- [45] Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing*

with Python: analyzing text with the natural language toolkit. ” O’Reilly Media, Inc.”.

- [46] Pernille Bjørn and Nina Boulus-Rødje. 2015. The multiple intersecting sites of design in CSCW research. *Computer Supported Cooperative Work (CSCW)* 24, 4 (2015), 319–351.
- [47] Alan F Blackwell, Carol Britton, A Cox, Thomas RG Green, Corin Gurr, Gada Kadoda, MS Kutar, Martin Loomes, Chrystopher L Nehaniv, Marian Petre, et al. 2001. Cognitive dimensions of notations: Design tools for cognitive technology. In *International Conference on Cognitive Technology*. Springer, 325–341.
- [48] Ann Blandford, Dominic Furniss, and Stephann Makri. 2016. Qualitative HCI research: Going behind the scenes. *Synthesis lectures on human-centered informatics* 9, 1 (2016), 1–115.
- [49] Kirsten Boehner, William Gaver, and Andy Boucher. 2012. 14 Probes. *Inventive Methods: The happening of the social* 185 (2012).
- [50] Pia Borlund. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information research* 8, 3 (2003), 8–3.
- [51] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications* 8, 1 (2021), 1–15.
- [52] Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology* 15 (2013), 209–227.

- [53] Joel Brandt, Mira Dontcheva, Marcos Weskamp, and Scott R Klemmer. 2010. Example-centric programming: integrating web search into the development environment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 513–522.
- [54] Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, Vol. 36. ACM New York, NY, USA, 3–10.
- [55] Jenny Bronstein. 2014. The role of perceived self-efficacy in the information seeking behavior of library and information science students. *The Journal of Academic Librarianship* 40, 2 (2014), 101–106.
- [56] Tim Brooks, Bill Peebles, Connor Holmes, Will DePue, Yufei Guo, Li Jing, David Schnurr, Joe Taylor, Troy Luhman, Eric Luhman, Clarence Ng, Ricky Wang, and Aditya Ramesh. 2024. Video generation models as world simulators. (2024). <https://openai.com/research/video-generation-models-as-world-simulators>
- [57] Steven P Brown, Shankar Ganesan, and Goutam Challagalla. 2001. Self-efficacy as a moderator of information-seeking effectiveness. *Journal of applied psychology* 86, 5 (2001), 1043.
- [58] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [59] Christine Susan Bruce. 1994. Research students' early experiences of the dissertation literature review. *Studies in Higher Education* 19, 2 (1994), 217–229.

- [60] Frederik Brudy, Christian Holz, Roman Rädle, Chi-Jui Wu, Steven Houben, Clemens Nylandsted Klokrose, and Nicolai Marquardt. 2019. Cross-device taxonomy: Survey, opportunities and challenges of interactions spanning across multiple devices. In *Proceedings of the 2019 chi conference on human factors in computing systems*. 1–28.
- [61] Frederik Brudy, David Ledo, Michel Pahud, Nathalie Henry Riche, Christian Holz, Anand Wagmare, Hemant Bhaskar Surale, Marcus Peinado, Xiaokuan Zhang, Shannon Joyner, et al. 2020. SurfaceFleet: Exploring Distributed Interactions Unbounded from Device, Application, User, and Time. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 7–21.
- [62] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712* (2023).
- [63] John M Budd. 2004. Relevance: Language, semantics, philosophy. (2004).
- [64] Vannevar Bush et al. 1945. As we may think. *The atlantic monthly* 176, 1 (1945), 101–108.
- [65] Katriina Byström and Kalervo Järvelin. 1995. Task complexity affects information seeking and use. *Information processing & management* 31, 2 (1995), 191–213.
- [66] Susanne B dker and Clemens Nylandsted Klokrose. 2011. The human–artifact model: An activity theoretical approach to artifact ecologies. *Human–Computer Interaction* 26, 4 (2011), 315–371.
- [67] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme Summarization of Scientific Documents. In *Findings of the Association for*

Computational Linguistics: EMNLP 2020. Association for Computational Linguistics, Online, 4766–4777. <https://doi.org/10.18653/v1/2020.findings-emnlp.428>

- [68] Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel S Weld. 2020. TLDR: Extreme summarization of scientific documents. *arXiv preprint arXiv:2004.15011* (2020).
- [69] Arthur Câmara, Nirmal Roy, David Maxwell, and Claudia Hauff. 2021. Searching to learn with instructional scaffolding. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 209–218.
- [70] Donald T Campbell. 1960. Blind variation and selective retentions in creative thought as in other knowledge processes. *Psychological review* 67, 6 (1960), 380.
- [71] Alberto J Cañas, Roger Carff, Greg Hill, Marco Carvalho, Marco Arguedas, Thomas C Eskridge, James Lott, and Rodrigo Carvajal. 2005. Concept maps: Integrating knowledge and information visualization. In *Knowledge and information visualization*. Springer, 205–219.
- [72] Robert Capra and Jaime Arguello. 2019. Using Trails to Support Users with Tasks of Varying Scope. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 977–980.
- [73] Robert Capra, Jaime Arguello, Anita Crescenzi, and Emily Vardell. 2015. Differences in the use of search assistance for tasks of varying complexity. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 23–32.
- [74] Robert Capra, Gary Marchionini, Javier Velasco-Martin, and Katrina Muller. 2010. Tools-at-hand and learning in multi-session, collaborative search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 951–960.

- [75] Robert Capra, Javier Velasco-Martin, and Beth Sams. 2010. Levels of working together in collaborative information seeking and sharing. *Proceedings of the Computer Supported Cooperative Work. CSCW 10* (2010).
- [76] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.
- [77] Erin A Carroll and Celine Latulipe. 2009. The creativity support index. In *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. 4009–4014.
- [78] Erin A Carroll and Celine Latulipe. 2012. Triangulating the personal creative experience: self-report, external judgments, and physiology. In *Proceedings of Graphics Interface 2012*. 53–60.
- [79] Ben Carterette, Evangelos Kanoulas, Mark Hall, and Paul Clough. 2014. *Overview of the TREC 2014 session track*. Technical Report. DELAWARE UNIV NEWARK DEPT OF COMPUTER AND INFORMATION SCIENCES.
- [80] Joel Chan, Steven Dang, and Steven P Dow. 2016. Comparing different sensemaking approaches for large-scale ideation. In *Proceedings of the 2016 CHI conference on human factors in computing systems*. 2717–2728.
- [81] Joel Chan, Steven P Dow, and Christian D Schunn. 2018. Do the best design ideas (really) come from conceptually distant sources of inspiration? In *Engineering a Better Future*. Springer, Cham, 111–139.
- [82] Joel Chan, Pao Siangliulue, Denisa Qori McDonald, Ruixue Liu, Reza Moradinezhad, Safa Aman, Erin T Solovey, Krzysztof Z Gajos, and Steven P

- Dow. 2017. Semantically far inspirations considered harmful? accounting for cognitive states in collaborative ideation. In *Proceedings of the 2017 ACM SIGCHI Conference on Creativity and Cognition*. 93–105.
- [83] Joseph Chee Chang, Nathan Hahn, and Aniket Kittur. 2020. Mesh: Scaffolding Comparison Tables for Online Decision Making. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 391–405.
- [84] Joseph Chee Chang, Nathan Hahn, Adam Perer, and Aniket Kittur. 2019. SearchLens: Composing and capturing complex user interests for exploratory search. In *Proceedings of the 24th International Conference on Intelligent User Interfaces*. 498–509.
- [85] Joseph Chee Chang, Aniket Kittur, and Nathan Hahn. 2016. Alloy: Clustering with crowds and computation. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 3180–3191.
- [86] Loren J Chapman. 1967. Illusory correlation in observational report. *Journal of Verbal Learning and Verbal Behavior* 6, 1 (1967), 151–155.
- [87] Kathy Charmaz. 2014. *Constructing grounded theory*. sage.
- [88] Duen Horng Chau, Aniket Kittur, Jason I Hong, and Christos Faloutsos. 2011. Apolo: making sense of large network data by combining rich user interaction and machine learning. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 167–176.
- [89] Catherine Chavula, Yujin Choi, and Soo Young Rieh. 2022. Understanding Creative Thinking Processes in Searching for New Ideas. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*. 321–326.

- [90] Hao Chen and Susan Dumais. 2000. Bringing order to the web: Automatically categorizing search results. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 145–152.
- [91] Zhutian Chen and Haijun Xia. 2022. Crossdata: Leveraging text-data connections for authoring data documents. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [92] Erin Cherry and Celine Latulipe. 2014. Quantifying the creativity support of digital tools through the creativity support index. *ACM Transactions on Computer-Human Interaction (TOCHI)* 21, 4 (2014), 1–25.
- [93] Parmit K Chilana, Amy J Ko, and Jacob O Wobbrock. 2012. LemonAid: selection-based crowdsourced contextual help for web applications. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1549–1558.
- [94] MML Chiu, SKW Chu, KKK Ting, and GYC Yau. 2011. A novice-expert comparison in information search. *Information Science* 23 (2011), 225–238.
- [95] Evangelia G Chrysikou, Katharine Motyka, Cristina Nigro, Song-I Yang, and Sharon L Thompson-Schill. 2016. Functional fixedness in creative thinking tasks depends on stimulus modality. *Psychology of Aesthetics, Creativity, and the Arts* 10, 4 (2016), 425.
- [96] John Joon Young Chung and Eytan Adar. 2023. Artinter: AI-powered Boundary Objects for Commissioning Visual Arts. (2023).
- [97] John Joon Young Chung and Eytan Adar. 2023. PromptPaint: Steering Text-to-Image Generation Through Paint Medium-like Interactions. *arXiv preprint arXiv:2308.05184* (2023).

- [98] John Joon Young Chung, Shiqing He, and Eytan Adar. 2021. The intersection of users, roles, interactions, and technologies in creativity support tools. In *Designing Interactive Systems Conference 2021*. 1817–1833.
- [99] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. *arXiv preprint arXiv:2004.07180* (2020).
- [100] Hilary Collins. 2018. Creative research: the theory and practice of research for the creative industries. (2018).
- [101] British Design Council. [n.d.]. The Double Diamond. <https://www.designcouncil.org.uk/our-resources/the-double-diamond/>
- [102] Design Council. 2015. The design process: What is the double diamond. *online] The Design Council. Available at: https://www.designcouncil.org.uk/news-opinion/design-process-what-double-diamond (Accessed 21.11. 2018)* (2015).
- [103] Christopher Cox. 2021. *The Deadline Effect: How to Work Like It's the Last Minute—Before the Last Minute*. Simon and Schuster.
- [104] Richard Cox and Paul Brna. 1995. Supporting the use of external representations in problem solving: The need for flexible learning environments. *Journal of Artificial intelligence in Education* 6 (1995), 239–302.
- [105] Anita Crescenzi, Diane Kelly, and Leif Azzopardi. 2016. Impacts of time constraints and system delays on user experience. In *Proceedings of the 2016 acm on conference on human information interaction and retrieval*. 141–150.
- [106] Anita Crescenzi and Lan Li. 2022. Assessing Realism in Simulated Work Tasks.

In Proceedings of the 2022 Conference on Human Information Interaction and Retrieval. 266–271.

- [107] Anita Crescenzi, Yuan Li, Yinglong Zhang, and Rob Capra. 2019. Towards Better Support for Exploratory Search through an Investigation of Notes-to-self and Notes-to-share. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1093–1096.
- [108] Nigel Cross. 2004. Expertise in design: an overview. *Design studies* 25, 5 (2004), 427–441.
- [109] Edward Cutrell, Daniel Robbins, Susan Dumais, and Raman Sarin. 2006. Fast, flexible filtering with phlat. In *Proceedings of the SIGCHI conference on Human Factors in computing systems*. 261–270.
- [110] Douglass R Cutting, David R Karger, Jan O Pedersen, and John W Tukey. 2017. Scatter/gather: A cluster-based approach to browsing large document collections. In *ACM SIGIR Forum*, Vol. 51. ACM New York, NY, USA, 148–159.
- [111] Allen Cypher, Mira Dontcheva, Tessa Lau, and Jeffrey Nichols. 2010. *No code required: giving users tools to transform the web*. Morgan Kaufmann.
- [112] Mary Czerwinski, Eric Horvitz, and Susan Wilhite. 2004. A diary study of task switching and interruptions. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 175–182.
- [113] Susan B Davidson and Juliana Freire. 2008. Provenance and scientific workflows: challenges and opportunities. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. 1345–1350.

- [114] Hanne De Jaegher and Ezequiel Di Paolo. 2007. Participatory sense-making. *Phenomenology and the cognitive sciences* 6, 4 (2007), 485–507.
- [115] Andrew S Denney and Richard Tewksbury. 2013. How to write a literature review. *Journal of criminal justice education* 24, 2 (2013), 218–234.
- [116] Ruta Desai, Fraser Anderson, Justin Matejka, Stelian Coros, James McCann, George Fitzmaurice, and Tovi Grossman. 2019. Geppetto: Enabling semantic design of expressive robot behaviors. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [117] Michael Desmond and Michelle Brachman. 2024. Exploring Prompt Engineering Practices in the Enterprise. *arXiv preprint arXiv:2403.08950* (2024).
- [118] Sebastian Deterding, Jonathan Hook, Rebecca Fiebrink, M. F. P. Gillies, J. Gow, Memo Akten, Gillian Smith, Antonios Liapis, and Kate Compton. 2017. Mixed-Initiative Creative Interfaces. *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (2017). <https://api.semanticscholar.org/CorpusID:630467>
- [119] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2024. Qlora: Efficient finetuning of quantized llms. *Advances in Neural Information Processing Systems* 36 (2024).
- [120] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [121] Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. Ms2: Multi-document summarization of medical studies. *arXiv preprint arXiv:2104.06486* (2021).

- [122] Andrew Dillon and Michael G Morris. 1996. User acceptance of information technology: Theories and models. *Annual Review of Information Science and Technology (ARIST)* 31 (1996), 3–32.
- [123] Debora Donato, Francesco Bonchi, Tom Chi, and Yoelle Maarek. 2010. Do you want to take notes? Identifying research missions in Yahoo! Search Pad. In *Proceedings of the 19th international conference on World wide web*. 321–330.
- [124] Kees Dorst. 2011. The core of ‘design thinking’ and its application. *Design studies* 32, 6 (2011), 521–532.
- [125] Kees Dorst. 2015. *Frame innovation: Create new thinking by design*. MIT press.
- [126] Anil R Doshi and Oliver Hauser. 2023. Generative artificial intelligence enhances creativity. *Available at SSRN* (2023).
- [127] Graham Dove and Sara Jones. 2013. Evaluating creativity support in co-design workshops. (2013).
- [128] Steven Dow, Blair MacIntyre, Jaemin Lee, Christopher Oezbek, Jay David Bolter, and Maribeth Gandy. 2005. Wizard of Oz support throughout an iterative design process. *IEEE Pervasive Computing* 4, 4 (2005), 18–26.
- [129] Karl Duncker and Lynne S Lees. 1945. On problem-solving. *Psychological monographs* 58, 5 (1945), i.
- [130] David Dunning. 2011. The Dunning–Kruger effect: On being ignorant of one’s own ignorance. In *Advances in experimental social psychology*. Vol. 44. Elsevier, 247–296.
- [131] Donald P Ely. 1999. Conditions that facilitate the implementation of educational technology innovations. *Educational technology* 39, 6 (1999), 23–27.

- [132] Douglas C Engelbart. 2021. Augmenting human intellect: a conceptual framework (1962). (2021).
- [133] Gilles Fauconnier. 2001. Conceptual blending and analogy. *The analogical mind: Perspectives from cognitive science* 255 (2001), 286.
- [134] Gilles Fauconnier and Mark Turner. 1998. Conceptual integration networks. *Cognitive science* 22, 2 (1998), 133–187.
- [135] Diego Fernandez-Duque, Jodie A Baird, and Michael I Posner. 2000. Executive attention and metacognitive regulation. *Consciousness and cognition* 9, 2 (2000), 288–307.
- [136] Baruch Fischhoff. 2013. The sciences of science communication. *Proceedings of the National Academy of Sciences* 110, supplement_3 (2013), 14033–14039.
- [137] Kristie Fisher, Scott Counts, and Aniket Kittur. 2012. Distributed sensemaking: improving sensemaking by leveraging the efforts of previous users. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 247–256.
- [138] Sarah Rose Fitzgerald. 2017. *Information seeking of scholars in the field of higher education*. Michigan State University.
- [139] James Fogarty. 2017. Code and Contribution in Interactive Systems Research. In *Workshop HCITools: Strategies and Best Practices for Designing, Evaluating and Sharing Technical HCI Toolkits at CHI*.
- [140] Nigel Ford. 1999. Information retrieval and creativity: towards support for the original thinker. *Journal of Documentation* (1999).
- [141] Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing,

- Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science* 359, 6379 (2018), eaao0185.
- [142] Allen Foster and Nigel Ford. 2003. Serendipity and information seeking: an empirical study. *Journal of documentation* (2003).
- [143] Adam Fourney, Ben Lafreniere, Parmit Chilana, and Michael Terry. 2014. Intertwine: creating interapplication information scent to support coordinated use of software. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*. 429–438.
- [144] William B Frakes and Ricardo Baeza-Yates. 1992. *Information retrieval: data structures and algorithms*. Prentice-Hall, Inc.
- [145] Cristin Ailidh Fraser. 2020. *Contextually Recommending Expert Help and Demonstrations to Improve Creativity*. Ph.D. Dissertation. University of California, San Diego.
- [146] C Ailie Fraser, Mira Dontcheva, Holger Winnemöller, Sheryl Ehrlich, and Scott Klemmer. 2016. DiscoverySpace: suggesting actions in complex software. In *Proceedings of the 2016 ACM Conference on Designing Interactive Systems*. 1221–1232.
- [147] C Ailie Fraser, Julia M Markel, N James Basa, Mira Dontcheva, and Scott Klemmer. 2020. ReMap: Lowering the Barrier to Help-Seeking with Multimodal Search. In *Proceedings of the 33rd Annual ACM Symposium on User Interface Software and Technology*. 979–986.
- [148] C Ailie Fraser, Tricia J Ngoon, Mira Dontcheva, and Scott Klemmer. 2019. RePlay: Contextually Presenting Learning Videos Across Software Applications.

In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–13.

- [149] C Ailie Fraser, Tricia J Ngoon, Ariel S Weingarten, Mira Dontcheva, and Scott Klemmer. 2017. CritiqueKit: A mixed-initiative, real-time interface for improving feedback. In *Adjunct Publication of the 30th Annual ACM Symposium on User Interface Software and Technology*. 7–9.
- [150] Jonas Frich, Lindsay MacDonald Vermeulen, Christian Remy, Michael Mose Biskjaer, and Peter Dalsgaard. 2019. Mapping the landscape of creativity support tools in HCI. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [151] Jonas Frich, Midas Nouwens, Kim Halskov, and Peter Dalsgaard. 2021. How Digital Tools Impact Convergent and Divergent Thinking in Design Ideation. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–11.
- [152] Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Qianyu Guo, Meng Wang, and Haofen Wang. 2023. Retrieval-Augmented Generation for Large Language Models: A Survey. *ArXiv abs/2312.10997* (2023). <https://api.semanticscholar.org/CorpusID:266359151>
- [153] Susan Gauch, Mirco Speretta, Aravind Chandramouli, and Alessandro Micarelli. 2007. User profiles for personalized information access. In *The adaptive web*. Springer, 54–89.
- [154] Sebastian Gehrmann, Steven Layne, and Franck Dernoncourt. 2019. Improving Human Text Comprehension through Semi-Markov CRF-based Neural Section Title

Generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 1677–1688. <https://doi.org/10.18653/v1/N19-1168>

- [155] Katy Ilonka Gero, Vivian Liu, and Lydia Chilton. 2022. Sparks: Inspiration for science writing using language models. In *Designing interactive systems conference*. 1002–1019.
- [156] Jacob W Getzels. 1979. Problem finding: A theoretical note. *Cognitive science* 3, 2 (1979), 167–172.
- [157] Souvick Ghosh, Manasa Rath, and Chirag Shah. 2018. Searching as learning: Exploring search behavior and learning outcomes in learning-related tasks. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 22–31.
- [158] Karni Gilon, Felicia Y Ng, Joel Chan, Hila Lifshitz Assaf, Aniket Kittur, and Dafna Shahaf. 2017. Analogy mining for specific design needs. *arXiv preprint arXiv:1712.06880* (2017).
- [159] Paul Ginsparg. 2011. ArXiv at 20. *Nature* 476, 7359 (2011), 145–147.
- [160] Andreas Girgensohn, John Adcock, Matthew Cooper, and Lynn Wilcox. 2005. A synergistic approach to efficient interactive video retrieval. In *IFIP Conference on Human-Computer Interaction*. Springer, 781–794.
- [161] Karan Girotra, Lennart Meincke, Christian Terwiesch, and Karl T Ulrich. 2023. Ideas are dimes a dozen: Large language models for idea generation in innovation. *Available at SSRN 4526071* (2023).

- [162] Vinod Goel and Peter Pirolli. 1992. The structure of design problem spaces. *Cognitive science* 16, 3 (1992), 395–429.
- [163] Nitesh Goyal, Gilly Leshed, and Susan R Fussell. 2013. Effects of visualization and note-taking on sensemaking and analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2721–2724.
- [164] Kazjon Grace and Mary Lou Maher. 2016. Surprise-Triggered Reformulation of Design Goals.. In *AAAI*. 3726–3732.
- [165] David M Gray, Steven D’Alessandro, Lester W Johnson, and Leanne Carter. 2017. Inertia in services: causes and consequences for switching. *Journal of Services Marketing* (2017).
- [166] Thomas RG Green. 1989. Cognitive dimensions of notations. *People and computers V* (1989), 443–460.
- [167] Saul Greenberg. 2007. Toolkits and interface creativity. *Multimedia Tools and Applications* 32, 2 (2007), 139–159.
- [168] Saul Greenberg and Bill Buxton. 2008. Usability evaluation considered harmful (some of the time). In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 111–120.
- [169] Catherine Grevet and Eric Gilbert. 2015. Piggyback prototyping: Using existing, large-scale social computing systems to prototype new ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 4047–4056.
- [170] Tovi Grossman and George Fitzmaurice. 2010. ToolClips: an investigation of

- contextual video assistance for functionality understanding. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1515–1524.
- [171] Tovi Grossman, George Fitzmaurice, and Ramtin Attar. 2009. A survey of software learnability: metrics, methodologies and guidelines. In *Proceedings of the sigchi conference on human factors in computing systems*. 649–658.
- [172] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. 2023. Textbooks Are All You Need. arXiv:2306.11644 [cs.CL]
- [173] Philip J Guo. 2015. Codeopticon: Real-time, one-to-many human tutoring for computer programming. In *Proceedings of the 28th Annual ACM Symposium on User Interface Software & Technology*. 599–608.
- [174] Matthew Guzdial, Nicholas Liao, Jonathan Chen, Shao-Yu Chen, Shukan Shah, Vishwa Shah, Joshua Reno, Gillian Smith, and Mark O Riedl. 2019. Friend, collaborator, student, manager: How design of an ai-driven game level editor affects creators. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–13.
- [175] Jennifer Haase and Paul HP Hanel. 2023. Artificial muses: Generative artificial intelligence chatbots have risen to human-level creativity. *arXiv preprint arXiv:2303.12003* (2023).
- [176] Nathan Hahn, Joseph Chang, Ji Eun Kim, and Aniket Kittur. 2016. The Knowledge Accelerator: Big picture thinking in small pieces. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 2258–2270.

- [177] Andrew Head, Kyle Lo, Dongyeop Kang, Raymond Fok, Sam Skjonsberg, Daniel S Weld, and Marti A Hearst. 2021. Augmenting scientific papers with just-in-time, position-sensitive definitions of terms and symbols. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–18.
- [178] Marti A Hearst. 1995. TileBars: visualization of term distribution information in full text information access. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 59–66.
- [179] Marti A Hearst. 2006. Clustering versus faceted categories for information exploration. *Commun. ACM* 49, 4 (2006), 59–61.
- [180] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116, 6 (2019), 1844–1850.
- [181] Jeffrey Heer. 2019. Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences* 116 (2019), 1844 – 1850. <https://api.semanticscholar.org/CorpusID:73435302>
- [182] William S Hemmig. 2008. The information-seeking behavior of visual artists: a literature review. *Journal of documentation* (2008).
- [183] Javier Hernandez, Jina Suh, Judith Amores, Kael Rowan, Gonzalo Ramos, and Mary Czerwinski. 2023. Affective Conversational Agents: Understanding Expectations and Personal Influences. (October 2023). <https://www.microsoft.com/en-us/research/publication/affective-conversational-agents-understanding-expectations-and-personal-influences/> ArXiv.

- [184] Jonathan Hey, Julie Linsey, Alice M Agogino, and Kristin L Wood. 2008. Analogies and metaphors in creative design. *International Journal of Engineering Education* 24, 2 (2008), 283.
- [185] Ken Hinckley, Shengdong Zhao, Raman Sarin, Patrick Baudisch, Edward Cutrell, Michael Shilman, and Desney Tan. 2007. InkSeine: In Situ search for active note taking. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 251–260.
- [186] Orland Hoerber and Xue Dong Yang. 2006. A comparative user study of web search interfaces: HotMap, Concept Highlighter, and Google. In *2006 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2006 Main Conference Proceedings)(WI'06)*. IEEE, 866–874.
- [187] Christoph Hölscher and Gerhard Strube. 2000. Web search behavior of Internet experts and newbies. *Computer networks* 33, 1-6 (2000), 337–346.
- [188] Serge Horbach, Kaare Aagaard, and Jesper W Schneider. 2021. Meta-Research: How problematic citing practices distort science. (2021).
- [189] Serge PJM Horbach, Freek JW Oude Maatman, Willem Halfman, and Wytse M Hepkema. 2022. Automated citation recommendation tools encourage questionable citations. *Research Evaluation* (2022).
- [190] Eric Horvitz. 1999. Principles of mixed-initiative user interfaces. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. 159–166.
- [191] Hsi and Potts. 2000. Studying the evolution and enhancement of software features. In *Proceedings 2000 International Conference on Software Maintenance*. IEEE, 143–151.

- [192] Ingrid Hsieh-Yee. 1993. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. *Journal of the american society for information science* 44, 3 (1993), 161–174.
- [193] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685* (2021).
- [194] Rong Hu, Kun Lu, and Soohyung Joo. 2013. Effects of topic familiarity and search skills on query reformulation behavior. *Proceedings of the American Society for Information Science and Technology* 50, 1 (2013), 1–9.
- [195] Jeff Huang, Ryen W White, and Susan Dumais. 2011. No clicks, no problem: using cursor movements to understand and improve search. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 1225–1234.
- [196] Amy Hurst, Scott E Hudson, and Jennifer Mankoff. 2010. Automatically identifying targets users interact with during real world tasks. In *Proceedings of the 15th international conference on Intelligent user interfaces*. 11–20.
- [197] Hilary Hutchinson, Wendy Mackay, Bo Westerlund, Benjamin B Bederson, Allison Druin, Catherine Plaisant, Michel Beaudouin-Lafon, Stéphane Conversy, Helen Evans, Heiko Hansen, et al. 2003. Technology probes: inspiring design for and with families. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 17–24.
- [198] World Leaders in Research-Based User Experience. [n.d.]. Thinking aloud: The Number 1 usability tool. <https://www.nngroup.com/articles/thinking-aloud-the-1-usability-tool/>

- [199] Sharon Favaro Ince, Christopher Hoadley, and Paul A Kirschner. 2018. A study of search practices in doctoral student scholarly workflows. In *Proceedings of the 2018 Conference on Human Information Interaction & Retrieval*. 245–248.
- [200] Nanna Inie, Jeanette Falk, and Steve Tanimoto. 2023. Designing Participatory AI: Creative Professionals’ Worries and Expectations about Generative AI. *Extended Abstracts of the 2023 CHI Conference on Human Factors in Computing Systems* (2023). <https://api.semanticscholar.org/CorpusID:257557305>
- [201] Dawn Michele Jacobsen. 1998. Adoption patterns and characteristics of faculty who intergrate computer technology for teaching and learning in higher education. (1998).
- [202] Ajit Jain, Andruid Kerne, Nic Lupfer, Gabriel Britain, Aaron Perrine, Yoonsuck Choe, John Keyser, and Ruihong Huang. 2021. Recognizing creative visual design: multiscale design characteristics in free-form web curation documents. In *Proceedings of the 21st ACM Symposium on Document Engineering*. 1–10.
- [203] Himanshu Jain, Venkatesh Balasubramanian, Bhanu Chunduri, and Manik Varma. 2019. Slice: Scalable linear extreme classifiers trained on 100 million labels for related searches. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 528–536.
- [204] Bernard J Jansen, Danielle Booth, and Brian Smith. 2009. Using the taxonomy of cognitive learning to model online searching. *Information Processing & Management* 45, 6 (2009), 643–663.
- [205] Bernard J Jansen and Michael D McNeese. 2005. Evaluating the effectiveness of and patterns of interactions with automated searching assistance. *Journal of*

- the American Society for Information Science and Technology* 56, 14 (2005), 1480–1503.
- [206] Renée S Jansen, Daniel Lakens, and Wijnand A IJsselsteijn. 2017. An integrative review of the cognitive costs and benefits of note-taking. *Educational Research Review* 22 (2017), 223–233.
- [207] Jill Jesson, Lydia Matheson, and Fiona M Lacey. 2011. Doing your literature review: Traditional and systematic techniques. (2011).
- [208] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629* (2022).
- [209] Hyoungwook Jin, Minsuk Chang, and Juho Kim. 2019. SolveDeep: A System for Supporting Subgoal Learning in Online Math Problem Solving. In *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–6.
- [210] Arif E Jinha. 2010. Article 50 million: an estimate of the number of scholarly articles in existence. *Learned publishing* 23, 3 (2010), 258–263.
- [211] Rosie Jones, Benjamin Rey, Omid Madani, and Wiley Greiner. 2006. Generating query substitutions. In *Proceedings of the 15th international conference on World Wide Web*. 387–396.
- [212] William Jones, Jesse David Dinneen, Robert Capra, Anne Diekema, and Manuel Pérez-Quñones. 2017. Personal information management (PIM). *Encyclopedia of library and information science* (2017), 3584–605.

- [213] Rishita Kalyani and Ujwal Gadiraju. 2019. Understanding User Search Behavior Across Varying Cognitive Levels. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*. 123–132.
- [214] Yvonne Kammerer, Rowan Nairn, Peter Pirolli, and Ed H Chi. 2009. Signpost from the masses: learning effects in an exploratory social tag search browser. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 625–634.
- [215] Maryam Kamvar, Melanie Kellar, Rajan Patel, and Ya Xu. 2009. Computers and iPhones and Mobile Phones, Oh My!: A Logs-based Comparison of Search Users on Different Devices. In *Proceedings of the 18th International Conference on World Wide Web (Madrid, Spain) (WWW '09)*. ACM, New York, NY, USA, 801–810. <https://doi.org/10.1145/1526709.1526817>
- [216] Hyeonsu B Kang, Joseph Chee Chang, Yongsung Kim, and Aniket Kittur. 2022. Threddy: An Interactive System for Personalized Thread-based Exploration and Organization of Scientific Literature. *arXiv preprint arXiv:2208.03455* (2022).
- [217] Hyeonsu B Kang, Rafal Kocielnik, Andrew Head, Jiangjiang Yang, Matt Latzke, Aniket Kittur, Daniel S Weld, Doug Downey, and Jonathan Bragg. 2022. From Who You Know to What You Read: Augmenting Scientific Recommendations with Implicit Social Networks. In *CHI Conference on Human Factors in Computing Systems*. 1–23.
- [218] Hyeonsu B Kang, Nouran Soliman, Matt Latzke, Joseph Chee Chang, and Jonathan Bragg. 2023. ComLittee: Literature Discovery with Personal Elected Author Committees. In *Proceedings of the 2023 CHI Conference on Human Factors in*

Computing Systems (Hamburg, Germany) (CHI '23). Association for Computing Machinery, New York, NY, USA.

- [219] Makoto P Kato, Tetsuya Sakai, and Katsumi Tanaka. 2012. Structured query suggestion for specialization and parallel movement: effect on search behaviors. In *Proceedings of the 21st international conference on World Wide Web*. 389–398.
- [220] Harmanpreet Kaur, Doug Downey, Amanpreet Singh, Evie Yu-Yen Cheng, Daniel Weld, and Jonathan Bragg. 2022. FeedLens: Polymorphic Lenses for Personalizing Exploratory Search over Knowledge Graphs. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–15.
- [221] Joseph 'Jofish' Kaye. 2007. Evaluating experience-focused HCI. In *CHI'07 extended abstracts on Human factors in computing systems*. 1661–1664.
- [222] Diane Kelly, Amber Cushing, Maureen Dostert, Xi Niu, and Karl Gyllstrom. 2010. Effects of popularity and quality on the usage of query suggestions during information search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 45–54.
- [223] Diane Kelly, Karl Gyllstrom, and Earl W Bailey. 2009. A comparison of query and term suggestion features for interactive searching. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. 371–378.
- [224] Diane Kelly and Jaime Teevan. 2003. Implicit feedback for inferring user preference: a bibliography. In *Acm Sigir Forum*, Vol. 37. ACM New York, NY, USA, 18–28.
- [225] Fawzia Khan. 1993. *A survey of note-taking practices*. Hewlett-Packard Laboratories.

- [226] Jerry S Kidd. 1990. Measuring referencing practices. *Journal of the American Society for Information Science* 41, 3 (1990), 157–163.
- [227] John F Kihlstrom, Eric Eich, Deborah Sandbrand, and Betsy A Tobias. 1999. Emotion and memory: Implications for self-report. In *The science of self-report*. Psychology Press, 93–112.
- [228] Joy Kim, Mira Dontcheva, Wilmot Li, Michael S Bernstein, and Daniela Steinsapir. 2015. Motif: Supporting novice creativity through expert patterns. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*. 1211–1220.
- [229] Kyung-Sun Kim and Bryce Allen. 2002. Cognitive and task influences on Web searching behavior. *Journal of the American Society for Information Science and Technology* 53, 2 (2002), 109–119.
- [230] Donald W King, Carol Tenopir, Songphan Choemprayong, and Lei Wu. 2009. Scholarly journal information-seeking and reading patterns of faculty at five US universities. *Learned Publishing* 22, 2 (2009), 126–144.
- [231] David Kirsh. 2010. Thinking with external representations. *AI & society* 25, 4 (2010), 441–454.
- [232] Aniket Kittur, Andrew M Peters, Abdigani Diriye, and Michael Bove. 2014. Standing on the schemas of giants: socially augmented information foraging. In *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing*. 999–1010.
- [233] Avraham N Kluger and Angelo DeNisi. 1996. The effects of feedback interventions on performance: a historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological bulletin* 119, 2 (1996), 254.

- [234] Jeffrey W Knopf. 2006. Doing a literature review. *PS: Political Science & Politics* 39, 1 (2006), 127–132.
- [235] Nicolas Kokkalis, Thomas Köhn, Johannes Huebner, Moontae Lee, Florian Schulze, and Scott R Klemmer. 2013. Taskgenies: Automatically providing action plans helps people complete tasks. *ACM Transactions on Computer-Human Interaction (TOCHI)* 20, 5 (2013), 1–25.
- [236] Jon Kolko. 2011. *Exposing the magic of design: A practitioner's guide to the methods and theory of synthesis*. Oxford University Press.
- [237] David R Krathwohl. 2002. A revision of Bloom's taxonomy: An overview. *Theory into practice* 41, 4 (2002), 212–218.
- [238] David R Krathwohl and Lorin W Anderson. 2009. *A taxonomy for learning, teaching, and assessing: A revision of Bloom's taxonomy of educational objectives*. Longman.
- [239] Carol C Kuhlthau. 1991. Inside the search process: Information seeking from the user's perspective. *Journal of the American society for information science* 42, 5 (1991), 361–371.
- [240] Carol C Kuhlthau, Jannica Heinström, and Ross J Todd. 2008. The 'information search process' revisited: Is the model still useful. *Information research* 13, 4 (2008), 13–4.
- [241] Bill Kules and Robert Capra. 2009. Designing exploratory search tasks for user studies of information seeking support systems. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. 419–420.

- [242] Bill Kules, Robert Capra, Matthew Banta, and Tito Sierra. 2009. What do exploratory searchers look at in a faceted search interface?. In *Proceedings of the 9th ACM/IEEE-CS joint conference on Digital libraries*. 313–322.
- [243] Bill Kules and Ben Shneiderman. 2008. Users can change their web search tactics: Design guidelines for categorized overviews. *Information Processing & Management* 44, 2 (2008), 463–484.
- [244] Tessa Lau and Eric Horvitz. 1999. Patterns of search: analyzing and modeling web query refinement. In *UM99 user modeling*. Springer, 119–128.
- [245] David Ledo, Steven Houben, Jo Vermeulen, Nicolai Marquardt, Lora Oehlberg, and Saul Greenberg. 2018. Evaluation strategies for HCI toolkit research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–17.
- [246] David Ledo, Jo Vermeulen, Sheelagh Carpendale, Saul Greenberg, Lora Oehlberg, and Sebastian Boring. 2019. Astral: Prototyping Mobile and Smart Object Interactive Behaviours Using Familiar Applications. In *Proceedings of the 2019 on Designing Interactive Systems Conference*. 711–724.
- [247] Mina Lee, Percy Liang, and Qian Yang. 2022. Coauthor: Designing a human-ai collaborative writing dataset for exploring language model capabilities. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–19.
- [248] Yoonjoo Lee, John Joon Young Chung, Tae Soo Kim, Jean Y Song, and Juho Kim. 2022. Promptiverse: Scalable generation of scaffolding prompts through human-AI hybrid knowledge graph annotation. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–18.

- [249] Younghwa Lee, Kenneth A Kozar, and Kai RT Larsen. 2003. The technology acceptance model: Past, present, and future. *Communications of the Association for information systems* 12, 1 (2003), 50.
- [250] Clayton Lewis. 1982. *Using the "thinking-aloud" method in cognitive interface design*. IBM TJ Watson Research Center Yorktown Heights, NY.
- [251] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 7871–7880. <https://doi.org/10.18653/v1/2020.acl-main.703>
- [252] Yan Li and James R Lindner. 2007. Faculty adoption behaviour about web-based distance education: a case study from China Agricultural University. *British Journal of Educational Technology* 38, 1 (2007), 83–94.
- [253] Joseph CR Licklider. 1960. Man-computer symbiosis. *IRE transactions on human factors in electronics* 1 (1960), 4–11.
- [254] J. C. R. Licklider. 2021. Man–Computer Symbiosis (1960). <https://api.semanticscholar.org/CorpusID:234040677>
- [255] Chin-Yew Lin and Eduard Hovy. 2002. Manual and automatic evaluation of summaries. In *Proceedings of the ACL-02 Workshop on Automatic Summarization*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 45–51. <https://doi.org/10.3115/1118162.1118168>

- [256] Min Lin, Wayne G Lutters, and Tina S Kim. 2004. Understanding the micronote lifecycle: improving mobile support for informal note taking. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 687–694.
- [257] Zhiyu Lin, Upol Ehsan, Rohan Agarwal, Samihan Dani, Vidushi Vashishth, and Mark Riedl. 2023. Beyond Prompts: Exploring the Design Space of Mixed-Initiative Co-Creativity Systems. *arXiv preprint arXiv:2305.07465* (2023).
- [258] Frank Linton and Hans-Peter Schaefer. 2000. Recommender systems for learning: Building user and expert models through long-term observation of application use. *User Modeling and User-Adapted Interaction* 10 (2000), 181–208.
- [259] Chang Liu, Xiangmin Zhang, and Wei Huang. 2016. The exploration of objective task difficulty and domain knowledge effects on users’ query formulation. *Proceedings of the Association for Information Science and Technology* 53, 1 (2016), 1–9.
- [260] Michael Xieyang Liu, Jane Hsieh, Nathan Hahn, Angelina Zhou, Emily Deng, Shaun Burley, Cynthia Taylor, Aniket Kittur, and Brad A Myers. 2019. Unakite: Scaffolding Developers’ Decision-Making Using the Web. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*. 67–80.
- [261] Michael Xieyang Liu, Aniket Kittur, and Brad A Myers. 2022. Crystalline: Lowering the Cost for Developers to Collect and Organize Information for Decision Making. In *CHI Conference on Human Factors in Computing Systems*. 1–16.
- [262] Vivian Liu and Lydia B Chilton. 2022. Design guidelines for prompt engineering text-to-image generative models. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*. 1–23.

- [263] Yiheng Liu, Hao He, Tianle Han, Xu Zhang, Mengyuan Liu, Jiaming Tian, Yutong Zhang, Jiaqi Wang, Xiaohui Gao, Tianyang Zhong, et al. 2024. Understanding llms: A comprehensive overview from training to inference. *arXiv preprint arXiv:2401.02038* (2024).
- [264] Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The Semantic Scholar Open Research Corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 4969–4983. <https://doi.org/10.18653/v1/2020.acl-main.447>
- [265] Patrice Lopez. 2009. GROBID: Combining automatic bibliographic data recognition and term extraction for scholarship publications. In *International conference on theory and practice of digital libraries*. Springer, 473–474.
- [266] Steven Loria, P Keen, M Honnibal, R Yankovsky, D Karesh, E Dempsey, et al. 2014. Textblob: simplified text processing. *Secondary TextBlob: simplified text processing* 3 (2014).
- [267] Ryan Louie, Any Cohen, Cheng-Zhi Anna Huang, Michael Terry, and Carrie J Cai. 2020. Cococo: AI-Steering Tools for Music Novices Co-Creating with Generative Models.. In *HAI-GEN+ user2agent@ IUI*.
- [268] Long-Chuan Lu, Wen-Pin Chang, and Hsiu-Hua Chang. 2014. Consumer attitudes toward blogger’s sponsored recommendations and purchase intention: The effect of sponsorship type, product type, and brand awareness. *Computers in Human Behavior* 34 (2014), 258–266.
- [269] Ewa Luger and Abigail Sellen. 2016. ” Like Having a Really Bad PA” The Gulf between User Expectation and Experience of Conversational Agents. In

Proceedings of the 2016 CHI conference on human factors in computing systems.
5286–5297.

- [270] Jiyun Luo, Xuchu Dong, and Hui Yang. 2015. Session search by direct policy learning. In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval*. 261–270.
- [271] Nic Lupfer, Andruid Kerne, Andrew M Webb, and Rhema Linder. 2016. Patterns of free-form curation: Visual thinking with web content. In *Proceedings of the 24th ACM international conference on Multimedia*. 12–21.
- [272] Kurt Luther, Nathan Hahn, Steven P Dow, and Aniket Kittur. 2015. Crowdlines: Supporting synthesis of diverse information sources through crowdsourced outlines. In *Third AAAI Conference on Human Computation and Crowdsourcing*.
- [273] Michael G MacDonald. 2000. Illusory correlation: A function of availability or representativeness heuristics? *Perceptual and motor skills* 91, 1 (2000), 343–350.
- [274] Brooke N Macnamara and Megha Maitra. 2019. The role of deliberate practice in expert performance: revisiting Ericsson, Krampe & Tesch-Römer (1993). *Royal Society open science* 6, 8 (2019), 190327.
- [275] Tamas Makany, Jonathan Kemp, and Itiel E Dror. 2009. Optimising the use of note-taking as an external cognitive aid for increasing learning. *British Journal of Educational Technology* 40, 4 (2009), 619–635.
- [276] Gary Marchionini. 2006. Exploratory search: from finding to understanding. *Commun. ACM* 49, 4 (2006), 41–46.
- [277] Gary Marchionini and Ben Brunk. 2003. Towards a general relation browser: A GUI for information architects. *Journal of Digital information* 4, 1 (2003).

- [278] Gary Marcus. 2022. *Stop Treating AI Models Like People*. https://garymarcus.substack.com/p/stop-treating-ai-models-like-people?utm_source=substack&utm_medium=email Substack Newsletter.
- [279] The Marginalian. 2011. *Steve Jobs: Bicycle for the Mind*. <https://www.themarginalian.org/2011/12/21/steve-jobs-bicycle-for-the-mind-1990/>
- [280] Catherine C Marshall and Sara Bly. 2005. Saving and using encountered information: implications for electronic periodicals. In *Proceedings of the Sigchi conference on human factors in computing systems*. 111–120.
- [281] Helen Mason and Lyn Robinson. 2011. The information-related behaviour of emerging artists and designers. *Journal of Documentation* (2011).
- [282] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2011. Ambient help. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 2751–2760.
- [283] Justin Matejka, Tovi Grossman, and George Fitzmaurice. 2011. IP-QAT: in-product questions, answers, & tips. In *Proceedings of the 24th annual ACM symposium on User interface software and technology*. 175–184.
- [284] Justin Matejka, Wei Li, Tovi Grossman, and George Fitzmaurice. 2009. CommunityCommands: command recommendations for software applications. In *Proceedings of the 22nd annual ACM symposium on User interface software and technology*. 193–202.
- [285] David Maxwell and Claudia Hauff. 2021. LogUI: Contemporary Logging Infras-

- tructure for Web-Based Experiments. In *European Conference on Information Retrieval*. Springer, 525–530.
- [286] Ray McAleese. 2000. Skill acquisition: The curious case of information searching. *Interactive Learning Environments* 8, 1 (2000), 23–49.
- [287] Catherine McKercher and Vincent Mosco. 2007. *Knowledge workers in the information society*. Lexington books.
- [288] Erin C McKiernan, Philip E Bourne, C Titus Brown, Stuart Buck, Amye Kenall, Jennifer Lin, Damon McDougall, Brian A Nosek, Karthik Ram, Courtney K Soderberg, et al. 2016. How open science helps researchers succeed. *elife* 5 (2016).
- [289] Gerry McKiernan. 2000. arXiv.org: the Los Alamos National Laboratory e-print server. *International Journal on Grey Literature* (2000).
- [290] Mohsen Mesgar and Michael Strube. 2018. A Neural Local Coherence Model for Text Quality Assessment. In *EMNLP*.
- [291] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*. 3111–3119.
- [292] James G Miller. 1960. Information input overload and psychopathology. *American journal of psychiatry* 116, 8 (1960), 695–704.
- [293] Ethan Mollick. [n.d.]. *On-boarding Your AI Intern*. <https://www.oneusefulthing.org/p/on-boarding-your-ai-intern>
- [294] Dan Morris, Meredith Ringel Morris, and Gina Venolia. 2008. SearchBar: a search-centric web history for task resumption and information re-finding. In

- Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.*
1207–1216.
- [295] Meredith Ringel Morris. 2008. A survey of collaborative web search practices. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* 1657–1660.
- [296] Meredith Ringel Morris. 2013. Collaborative Search Revisited. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (San Antonio, Texas, USA) (CSCW '13)*. ACM, New York, NY, USA, 1181–1192. <https://doi.org/10.1145/2441776.2441910>
- [297] Meredith Ringel Morris and Eric Horvitz. 2007. SearchTogether: an interface for collaborative web search. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. 3–12.
- [298] Michael Muller, Heloisa Candello, and Justin Weisz. 2023. Interactional Co-Creativity of Human and AI in Analogy-Based Design. In *International Conference on Computational Creativity*.
- [299] Michael Muller, Justin D Weisz, and Werner Geyer. 2020. Mixed initiative generative AI interfaces: An analytic framework for generative AI applications. In *Proceedings of the Workshop The Future of Co-Creative Systems-A Workshop on Human-Computer Co-Creativity of the 11th International Conference on Computational Creativity (ICCC 2020)*.
- [300] Michael D Mumford. 2003. Where have we been, where are we going? Taking stock in creativity research. *Creativity research journal* 15, 2-3 (2003), 107–120.
- [301] Brad Myers, Scott E Hudson, and Randy Pausch. 2000. Past, present, and future of

- user interface software tools. *ACM Transactions on Computer-Human Interaction (TOCHI)* 7, 1 (2000), 3–28.
- [302] Felicia Ng, Jina Suh, and Gonzalo Ramos. 2020. Understanding and supporting knowledge decomposition for machine teaching. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 1183–1194.
- [303] Xi Niu and Diane Kelly. 2014. The use of query suggestions during information search. *Information Processing & Management* 50, 1 (2014), 218–234.
- [304] Donald A Norman. 2004. Beauty, goodness, and usability. *Human-Computer Interaction* 19, 4 (2004), 311–318.
- [305] Donald A Norman. 2004. *Emotional design: Why we love (or hate) everyday things*. Basic Civitas Books.
- [306] Midas Nouwens and Clemens Nylandsted Klokmoose. 2018. The application and its consequences for non-standard knowledge work. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–12.
- [307] The Hasso Plattner Institute of Design at Stanford. [n.d.]. An Introduction to Design Thinking: Process Guide. <https://web.stanford.edu/~mshanks/MichaelShanks/files/509554.pdf>
- [308] Sangeun Oh, Hyuck Yoo, Dae R Jeong, Duc Hoang Bui, and Insik Shin. 2017. Mobile plus: Multi-device mobile platform for cross-device functionality sharing. In *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*. 332–344.
- [309] Keisuke Okamura. 2019. Interdisciplinarity revisited: evidence for research impact and dynamism. *Palgrave Communications* 5, 1 (2019), 1–9.

- [310] Dan R Olsen Jr. 2007. Evaluating user interface systems research. In *Proceedings of the 20th annual ACM symposium on User interface software and technology*. 251–258.
- [311] Alexandra Olteanu, Fernando Diaz, and Gabriella Kazai. 2020. When Are Search Completion Suggestions Problematic? *Proceedings of the ACM on Human-Computer Interaction* 4, CSCW2 (2020), 1–25.
- [312] OpenAI. 2023. GPT-4 Technical Report. *ArXiv* abs/2303.08774 (2023). <https://api.semanticscholar.org/CorpusID:257532815>
- [313] Antti Oulasvirta and Kasper Hornbæk. 2016. Hci research as problem-solving. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*. 4956–4967.
- [314] Ozgu Ozkan and Fehmi Dogan. 2013. Cognitive strategies of analogical reasoning in design: Differences between expert and novice designers. *Design Studies* 34, 2 (2013), 161–192.
- [315] Google PAIR. 2019. <https://pair.withgoogle.com/guidebook/>
- [316] Srishti Palani, Zijian Ding, Stephen MacNeil, and Steven P Dow. 2021. The” Active Search” Hypothesis: How Search Strategies Relate to Creative Learning. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 325–329.
- [317] Srishti Palani, Zijian Ding, Austin Nguyen, Andrew Chuang, Stephen MacNeil, and Steven P Dow. 2021. CoNotate: Suggesting Queries Based on Notes Promotes Knowledge Discovery. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*. 1–14.

- [318] Srishti Palani, Adam Fourney, Shane Williams, Kevin Larson, Irina Spiridonova, and Meredith Ringel Morris. [n.d.]. An Eye Tracking Study of Web Search by People with and without Dyslexia. ([n. d.]).
- [319] Srishti Palani, Aakanksha Naik, Doug Downey, Amy X Zhang, Jonathan Bragg, and Joseph Chee Chang. 2023. Relatedly: Scaffolding Literature Reviews with Existing Related Work Sections. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–20.
- [320] Srishti Palani, Yingyi Zhou, Sheldon Zhu, and Steven P Dow. 2022. InterWeave: Presenting Search Suggestions in Context Scaffolds Information Search and Synthesis. In *Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology*. 1–16.
- [321] Bec Paton and Kees Dorst. 2011. Briefing and reframing: A situated practice. *Design Studies* 32, 6 (2011), 573–587.
- [322] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.
- [323] Jaakko Peltonen, Kseniia Belorustceva, and Tuukka Ruotsalo. 2017. Topic-relevance map: Visualization for improving search result comprehension. In *Proceedings of the 22nd international conference on intelligent user interfaces*. 611–622.
- [324] Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. 2023.

- Anglekindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [325] Ingrid Pettersson, Florian Lachner, Anna-Katharina Frison, Andreas Riener, and Andreas Butz. 2018. A Bermuda triangle? A Review of method application and triangulation in user experience evaluation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–16.
- [326] Jean Piaget. 1976. Piaget’s theory. In *Piaget and his school*. Springer, 11–23.
- [327] Peter Pirolli and Stuart Card. 1995. Information foraging in information access environments. In *Proceedings of the SIGCHI conference on Human factors in computing systems*. 51–58.
- [328] Peter Pirolli and Stuart Card. 2005. The sensemaking process and leverage points for analyst technology as identified through cognitive task analysis. In *Proceedings of international conference on intelligence analysis*, Vol. 5. McLean, VA, USA, 2–4.
- [329] Peter Pirolli and Stuart K. Card. 1995. Information foraging in information access environments. In *International Conference on Human Factors in Computing Systems*.
- [330] Peter Pirolli and Daniel M Russell. 2011. Introduction to this special issue on sensemaking.
- [331] Antoine Ponsard, Francisco Escalona, and Tamara Munzner. 2016. PaperQuest: A Visualization Tool to Support Literature Review. *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems (2016)*.

- [332] Morgan N Price, Bill N Schilit, and Gene Golovchinsky. 1998. XLibris: The active reading machine. In *CHI 98 conference summary on Human factors in computing systems*. 22–23.
- [333] Chris Quintana and Meilan Zhang. 2004. The Digital Ideakeeper: Extending digital library services to scaffold online inquiry. In *American Education Research Association Annual Meeting, San Diego, CA*. Citeseer.
- [334] Napol Rachatasumrit, Gonzalo Ramos, Jina Suh, Rachel Ng, and Christopher Meek. 2021. ForSense: Accelerating Online Research Through Sensemaking Integration and Machine Research Support. In *26th International Conference on Intelligent User Interfaces*. 608–618.
- [335] Filip Radlinski and Susan Dumais. 2006. Improving personalized web search using result diversification. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*. 691–692.
- [336] Filip Radlinski, Martin Szummer, and Nick Craswell. 2010. Inferring query intent from reformulations and clicks. In *Proceedings of the 19th international conference on World wide web*. 1171–1172.
- [337] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.
- [338] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. <http://is.muni.cz/publication/884893/en>.

- [339] Christian Remy, Oliver Bates, Alan Dix, Vanessa Thomas, Mike Hazas, Adrian Friday, and Elaine M Huang. 2018. Evaluation beyond usability: Validating sustainable HCI research. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–14.
- [340] Christian Remy, Lindsay MacDonald Vermeulen, Jonas Frich, Michael Mose Biskjaer, and Peter Dalsgaard. 2020. Evaluating Creativity Support Tools in HCI Research. In *Proceedings of the 2020 ACM Designing Interactive Systems Conference*. 457–476.
- [341] Karen Renaud and Judy Van Biljon. 2008. Predicting technology acceptance and adoption by the elderly: a qualitative study. In *Proceedings of the 2008 annual research conference of the South African Institute of Computer Scientists and Information Technologists on IT research in developing countries: riding the wave of technology*. 210–219.
- [342] Mitchel Resnick, Brad Myers, Kumiyo Nakakoji, Ben Shneiderman, Randy Pausch, Ted Selker, and Mike Eisenberg. 2005. Design principles for tools to support creative thinking. (2005).
- [343] Jeba Rezwana and Mary Lou Maher. 2022. Designing creative AI partners with COFI: A framework for modeling interaction in human-AI co-creative systems. *ACM Transactions on Computer-Human Interaction* (2022).
- [344] Eun Youp Rha, Matthew Mitsui, Nicholas J Belkin, and Chirag Shah. 2016. Exploring the relationships between search intentions and query reformulations. *Proceedings of the Association for Information Science and Technology* 53, 1 (2016), 1–9.

- [345] Mel Rhodes. 1961. An analysis of creativity. *The Phi delta kappan* 42, 7 (1961), 305–310.
- [346] Soo Young Rieh, Kevyn Collins-Thompson, Preben Hansen, and Hye-Jung Lee. 2016. Towards searching as a learning process: A review of current perspectives and future directions. *Journal of Information Science* 42, 1 (2016), 19–34.
- [347] Soo Young Rieh, Jacek Gwizdka, Luanne Freund, and Kevyn Collins-Thompson. 2014. Searching as learning: Novel measures for information interaction research. *Proceedings of the American Society for Information Science and Technology* 51, 1 (2014), 1–4.
- [348] Horst W Rittel and Melvin M Webber. 1973. 2.3 planning problems are wicked. *Polity* 4, 155 (1973), e169.
- [349] Stephen E Robertson and K Sparck Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information science* 27, 3 (1976), 129–146.
- [350] Everett M Rogers. 2010. *Diffusion of innovations*. Simon and Schuster.
- [351] Simon Roodhouse. 2006. The creative industries: definitions, quantification and practice. *Cultural Industries: The British Experience in International Perspective. Online, Berlin: Humboldt University Berlin, Edoc-Server* (2006), 13–32.
- [352] David Rosengrant, Eugenia Etkina, and Alan Van Heuvelen. 2007. An overview of recent research on multiple representations. In *AIP Conference proceedings*, Vol. 883. American Institute of Physics, 149–152.
- [353] Corbin Rosset, Chenyan Xiong, Xia Song, Daniel Campos, Nick Craswell, Saurabh Tiwary, and Paul Bennett. 2020. Leading Conversational Search by

- Suggesting Useful Questions. In *Proceedings of The Web Conference 2020*. 1160–1170.
- [354] Stephen Rowland. 2002. Overcoming fragmentation in professional life: The challenge for academic development. *Higher education quarterly* 56, 1 (2002), 52–64.
- [355] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. How Do Active Reading Strategies Affect Learning Outcomes in Web Search?. In *European Conference on Information Retrieval*. Springer, 368–375.
- [356] Nirmal Roy, Manuel Valle Torre, Ujwal Gadiraju, David Maxwell, and Claudia Hauff. 2021. Note the highlight: Incorporating active reading tools in a search as learning environment. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*. 229–238.
- [357] Daniel M Russell, Gregorio Convertino, Aniket Kittur, Peter Pirolli, and Elizabeth Anne Watkins. 2018. Sensemaking in a Senseless World: 2018 Workshop Abstract. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–7.
- [358] Daniel M Russell, Mark J Stefik, Peter Pirolli, and Stuart K Card. 1993. The cost structure of sensemaking. In *Proceedings of the INTERACT'93 and CHI'93 conference on Human factors in computing systems*. 269–276.
- [359] Frank E Saal, Ronald G Downey, and Mary A Lahey. 1980. Rating the ratings: Assessing the psychometric quality of rating data. *Psychological bulletin* 88, 2 (1980), 413.
- [360] Antti Salovaara, Antti Oulasvirta, and Giulio Jacucci. 2017. Evaluation of pro-

- totypes and the problem of possible futures. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*. 2064–2077.
- [361] Téo Sanchez. 2023. Examining the Text-to-Image Community of Practice: Why and How do People Prompt Generative AIs?. In *Proceedings of the 15th Conference on Creativity and Cognition*. 43–61.
- [362] Bahareh Sarrafzadeh and Edward Lank. 2017. Improving exploratory search experience through hierarchical knowledge graphs. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 145–154.
- [363] Vernon T Sarver. 1983. Ajzen and Fishbein's" theory of reasoned action": A critical assessment. (1983).
- [364] Arvind Satyanarayan, Bongshin Lee, Donghao Ren, Jeffrey Heer, John Stasko, John Thompson, Matthew Brehmer, and Zhicheng Liu. 2019. Critical reflections on visualization authoring systems. *IEEE transactions on visualization and computer graphics* 26, 1 (2019), 461–471.
- [365] Denis Savenkov and Eugene Agichtein. 2014. To hint or not: exploring the effectiveness of search hints for complex informational tasks. In *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. 1115–1118.
- [366] Reijo Savolainen. 2015. Cognitive barriers to information seeking: A conceptual analysis. *Journal of Information Science* 41, 5 (2015), 613–623.
- [367] Ben Schneiderman. 2002. Creativity Support Tools—Establishing a framework of activities for creative work. *Commun. ACM* 45, 10 (2002), 116–120.

- [368] Donald A Schön. 1984. *The reflective practitioner: How professionals think in action*. Routledge.
- [369] MC Schraefel, Daniel A Smith, Alisdair Owens, Alistair Russell, Craig Harris, and Max Wilson. 2005. The evolving mSpace platform: leveraging the Semantic Web on the Trail of the Memex. In *Proceedings of the sixteenth ACM conference on Hypertext and hypermedia*. 174–183.
- [370] Chirag Shah and Roberto González-Ibáñez. 2010. Exploring information seeking processes in collaborative search tasks. *Proceedings of the American Society for Information Science and Technology* 47, 1 (2010), 1–7.
- [371] Dafna Shahaf, Carlos Guestrin, and Eric Horvitz. 2012. Metro maps of science. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*. 1122–1130.
- [372] Peter Shea, Alexandra Pickett, and Chun Sau Li. 2005. Increasing access to higher education: A study of the diffusion of online teaching among 913 college faculty. *International Review of Research in Open and Distributed Learning* 6, 2 (2005), 1–27.
- [373] Sue Shellenbarger. 2015. *The Power of Asking Pivotal Questions*. <https://www.wsj.com/articles/BL-REB-35617>
- [374] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. 2022. VILA: Improving structured content extraction from scientific PDFs using visual layout groups. *Transactions of the Association for Computational Linguistics* 10 (2022), 376–392.
- [375] Renata M Sheppard, Mahsa Kamali, Raoul Rivas, Morihiko Tamai, Zhenyu Yang, Wanmin Wu, and Klara Nahrstedt. 2008. Advancing interactive collaborative

- mediums through tele-immersive dance (TED) a symbiotic creativity and design environment for art and computer science. In *Proceedings of the 16th ACM international conference on Multimedia*. 579–588.
- [376] Lorraine Sherry. 1998. An integrated technology adoption and diffusion model. *International Journal of Educational Telecommunications* 4, 2 (1998), 113–145.
- [377] Cary Shimek. [n.d.]. *AI and Creativity*.
- [378] Ben Shneiderman. 1983. Direct manipulation: A step beyond programming languages. *Computer* 16, 08 (1983), 57–69.
- [379] Ben Shneiderman. 1999. User interfaces for creativity support tools. In *Proceedings of the 3rd conference on Creativity & cognition*. 15–22.
- [380] Ben Shneiderman. 2007. Creativity support tools: Accelerating discovery and innovation. *Commun. ACM* 50, 12 (2007), 20–32.
- [381] Ben Shneiderman. 2009. Creativity support tools: A grand challenge for HCI researchers. In *Engineering the user interface*. Springer, 1–9.
- [382] Ben Shneiderman. 2020. Human-centered artificial intelligence: Reliable, safe & trustworthy. *International Journal of Human–Computer Interaction* 36, 6 (2020), 495–504.
- [383] Pao Siangliulue, Joel Chan, Steven P Dow, and Krzysztof Z Gajos. 2016. IdeaHound: improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. 609–624.
- [384] Pao Siangliulue, Joel Chan, Krzysztof Z Gajos, and Steven P Dow. 2015. Providing timely examples improves the quantity and quality of generated ideas. In

- Proceedings of the 2015 ACM SIGCHI Conference on Creativity and Cognition.*
83–92.
- [385] Fabrizio Silvestri. 2010. Mining query logs: Turning search usage data into knowledge. *Foundations and Trends in Information Retrieval* 4, 1—2 (2010), 1–174.
- [386] Nikhil Singh, Guillermo Bernal, Daria Savchenko, and Elena L Glassman. 2022. Where to hide a stolen elephant: Leaps in creative writing with multimodal machine intelligence. *ACM Transactions on Computer-Human Interaction* (2022).
- [387] Adish Singla, Ryen White, and Jeff Huang. 2010. Studying trailfinding algorithms for enhanced web search. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval.* 443–450.
- [388] Rashmi Sinha and Kirsten Swearingen. 2002. The role of transparency in recommender systems. In *CHI'02 extended abstracts on Human factors in computing systems.* 830–831.
- [389] Steven M Smith, Thomas B Ward, and Ronald A Finke. 1995. Cognitive processes in creative contexts. *The creative cognition approach* (1995), 1–7.
- [390] Alessandro Sordani, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. 2015. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management.* 553–562.
- [391] Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Thirty-first AAAI conference on artificial intelligence.*

- [392] Angie Spoto, Natalia Oleynik, Sebastian Deterding, and Jon Hook. 2017. Library of Mixed-Initiative Creative Interfaces.
- [393] Robert J Sternberg. 1999. *Handbook of creativity*. Cambridge University Press.
- [394] Robert J Sternberg and Elena L Grigorenko. 2001. Guilford's structure of intellect model and model of creativity: Contributions and limitations. *Creativity Research Journal* 13, 3-4 (2001), 309–316.
- [395] Stacey H Stockdill and Diane L Morehouse. 1992. Critical factors in the successful adoption of technology: A checklist based on TDC findings. *Educational Technology* 32, 1 (1992), 57–58.
- [396] Hari Subramonyam, Roy Pea, Christopher Lawrence Pondoc, Maneesh Agrawala, and Colleen Seifert. 2024. Bridging the Gulf of Envisioning: Cognitive Challenges in Prompt Based Interactions with LLMs. (2024).
- [397] Jina Suh, Eric Horvitz, Ryen W White, and Tim Althoff. 2021. Population-scale study of human needs during the covid-19 pandemic: Analysis and implications. In *Proceedings of the 14th ACM international conference on web search and data mining*. 4–12.
- [398] Sangho Suh, Bryan Min, Srishti Palani, and Haijun Xia. 2023. Sensecape: Enabling Multilevel Exploration and Sensemaking with Large Language Models. *arXiv preprint arXiv:2305.11483* (2023).
- [399] Jun-Zhao Sun, Jiehan Zhou, and Timo Pihlajaniemi. 2010. Mlogger: an automatic blogging system by mobile sensing user behaviors. In *Ubiquitous Intelligence and Computing: 7th International Conference, UIC 2010, Xi'an, China, October 26-29, 2010. Proceedings* 7. Springer, 650–664.

- [400] Daniel W Surry and John D Farquhar. 1997. Diffusion theory and instructional technology. *Journal of Instructional Science and technology* 2, 1 (1997), 24–36.
- [401] Ivan Sutherland, Douglas C. Engelbart, Alan C. Kay, and The Alto. 1962. Augmenting human intellect: a conceptual framework. <https://api.semanticscholar.org/CorpusID:18529677>
- [402] Kaveh Taghipour and Hwee Tou Ng. 2016. A neural approach to automated essay scoring. In *Proceedings of the 2016 conference on empirical methods in natural language processing*. 1882–1891.
- [403] Kornelia Tancheva, Gabriela Castro Gessner, Neely Tang, Erin Eldermire, Heather Furnas, Darcy Branchini, Gail Steinhart, and Nancy Fried Foster. 2016. A day in the life of a (serious) researcher: Envisioning the future of the research library. *Ithaka*. URL: <http://sr.ithaka.org> (2016).
- [404] Jaime Teevan. 2014. A formula for academic papers: Related work. <http://slowsearching.blogspot.com/2014/11/a-formula-for-academic-papers-related.html>
- [405] Jaime Teevan, Eytan Adar, Rosie Jones, and Michael A. S. Potts. 2007. Information Re-retrieval: Repeat Queries in Yahoo’s Logs. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (Amsterdam, The Netherlands) (SIGIR ’07)*. ACM, New York, NY, USA, 151–158. <https://doi.org/10.1145/1277741.1277770>
- [406] Jaime Teevan, Susan T Dumais, and Eric Horvitz. 2005. Personalizing search via automated analysis of interests and activities. In *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*. 449–456.

- [407] Carol Tenopir, Rachel Volentine, and Donald W King. 2012. **Article and book reading patterns of scholars: Findings for publishers.** *Learned publishing* 25, 4 (2012), 279–291.
- [408] UX Tools. [n.d.]. 2020 Tools Survey Results. <https://uxtools.co/survey-2020/>
- [409] Katja Tschimmel. 2012. Design Thinking as an effective Toolkit for Innovation. In *ISPIM Conference Proceedings*. The International Society for Professional Innovation Management (ISPIM), 1.
- [410] Kosetsu Tsukuda, Tetsuya Sakai, Zhicheng Dou, and Katsumi Tanaka. 2013. Estimating intent types for search result diversification. In *Asia Information Retrieval Symposium*. Springer, 25–37.
- [411] Amos Tversky and Daniel Kahneman. 1973. Availability: A heuristic for judging frequency and probability. *Cognitive psychology* 5, 2 (1973), 207–232.
- [412] Kazutoshi Umemoto, Takehiro Yamamoto, and Katsumi Tanaka. 2016. Scentbar: A query suggestion interface visualizing the amount of missed relevant information for intrinsically diverse search. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*. 405–414.
- [413] Klaus K Urban. 1991. Recent trends in creativity research and theory in Western Europe. *European Journal of High Ability* 1, 1 (1991), 99–113.
- [414] Kelsey Urgo and Jaime Arguello. 2022. Learning assessments in search-as-learning: A survey of prior work and opportunities for future research. *Information Processing & Management* 59, 2 (2022), 102821.

- [415] Pertti Vakkari. 2001. Changes in search tactics and relevance judgements when preparing a research proposal a summary of the findings of a longitudinal study. *Information retrieval* 4, 3-4 (2001), 295–310.
- [416] Pertti Vakkari. 2001. A theory of the task-based information retrieval process: A summary and generalisation of a longitudinal study. *Journal of documentation* 57, 1 (2001), 44–60.
- [417] Pertti Vakkari. 2016. Searching as learning: A systematization based on literature. *Journal of Information Science* 42, 1 (2016), 7–18.
- [418] Pertti Vakkari and Nanna Hakala. 2000. Changes in relevance criteria and problem stages in task performance. *Journal of documentation* 56, 5 (2000), 540–562.
- [419] Marco Valenzuela, Vu A. Ha, and Oren Etzioni. 2015. Identifying Meaningful Citations. In *AAAI Workshop: Scholarly Big Data*.
- [420] Max G Van Kleek, Michael Bernstein, Katrina Panovich, Gregory G Vargas, David R Karger, and MC Schraefel. 2009. Note to self: examining personal information keeping in a lightweight note-taking tool. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1477–1480.
- [421] R Van Noorden. 2014. Global scientific output doubles every nine years [blog post]. Retrieved from nature. com at <http://blogs.nature.com/news/2014/05/global-scientific-output-doubles-every-nine-years.html> (2014).
- [422] Richard Van Noorden et al. 2015. Interdisciplinary research by the numbers. *Nature* 525, 7569 (2015), 306–307.
- [423] George Veletsianos. 2007. Cognitive and affective benefits of an animated ped-

- agogical agent: Considering contextual relevance and aesthetics. *Journal of Educational Computing Research* 36, 4 (2007), 373–377.
- [424] Viswanath Venkatesh, James YL Thong, and Xin Xu. 2016. Unified theory of acceptance and use of technology: A synthesis and the road ahead. *Journal of the association for Information Systems* 17, 5 (2016), 328–376.
- [425] Laton Vermette, Parmit Chilana, Michael Terry, Adam Fourney, Ben Lafreniere, and Travis Kerr. 2015. CheatSheet: a contextual interactive memory aid for web applications. In *Proceedings of the 41st Graphics Interface Conference*. 241–248.
- [426] Stella Vosniadou and Andrew Ortony. 1989. *Similarity and analogical reasoning*. Cambridge University Press.
- [427] Emily Wall, Soroush Ghorashi, and Gonzalo A. Ramos. 2019. Using Expert Patterns in Assisted Interactive Machine Learning: A Study in Machine Teaching. In *IFIP TC13 International Conference on Human-Computer Interaction*. <https://api.semanticscholar.org/CorpusID:155256331>
- [428] Belle Wallace. 1986. Creativity: Some definitions: The creative personality; the creative process; the creative classroom. *Gifted Education International* 4, 2 (1986), 68–73.
- [429] James R Wallace, Saba Oji, and Craig Anslow. 2017. Technologies, methods, and values: changes in empirical research at CSCW 1990-2015. *Proceedings of the ACM on Human-Computer Interaction* 1, CSCW (2017), 1–18.
- [430] Austin R Ward and Robert Capra. 2021. OrgBox: Supporting cognitive and metacognitive activities during exploratory search. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2570–2574.

- [431] Andy Warr and Eamonn O’Neill. 2005. Understanding design as a social creative process. In *Proceedings of the 5th Conference on Creativity & Cognition*. 118–127.
- [432] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems* 35 (2022), 24824–24837.
- [433] Xing Wei, Yinglong Zhang, and Jacek Gwizdka. 2014. YASFIIRE: yet another system for IIR evaluation. In *Proceedings of the 5th Information Interaction in Context Symposium*. 316–319.
- [434] Mark Weiser. 1991. The Computer for the 21 st Century. *Scientific american* 265, 3 (1991), 94–105.
- [435] Ryen W White, Susan T Dumais, and Jaime Teevan. 2009. Characterizing the influence of domain expertise on web search behavior. In *Proceedings of the second ACM international conference on web search and data mining*. 132–141.
- [436] Ryen W White and Jeff Huang. 2010. Assessing the scenic route: measuring the value of search trails in web logs. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 587–594.
- [437] Ryen W White and Gary Marchionini. 2007. Examining the effectiveness of real-time query expansion. *Information Processing & Management* 43, 3 (2007), 685–704.
- [438] Ryen W. White and Dan Morris. 2007. Investigating the Querying and Browsing Behavior of Advanced Search Engine Users. In *Proceedings of the 30th Annual*

International ACM SIGIR Conference on Research and Development in Information Retrieval (Amsterdam, The Netherlands) (*SIGIR '07*). ACM, New York, NY, USA, 255–262. <https://doi.org/10.1145/1277741.1277787>

- [439] Ryan W White and Resa A Roth. 2009. Exploratory search: Beyond the query-response paradigm. *Synthesis lectures on information concepts, retrieval, and services* 1, 1 (2009), 1–98.
- [440] Andrzej P Wierzbicki and Yoshiteru Nakamori. 2007. *Creative environments: Issues of creativity support for the knowledge civilization age*. Vol. 59. Springer.
- [441] Merryl J Wilkenfeld and Thomas B Ward. 2001. Similarity and emergence in conceptual combination. *Journal of Memory and Language* 45, 1 (2001), 21–38.
- [442] Brent Wilson, Lorraine Sherry, Jackie Dobrovolny, Mike Batty, and Martin Ryder. 2000. Adoption of learning technologies in schools and universities. *Handbook on information technologies for education & training*. New York: Springer-Verlag (2000).
- [443] Brent Wilson, Lorraine Sherry, Jackie Dobrovolny, Mike Batty, and Martin Ryder. 2002. Adoption factors and processes. *Handbook on information technologies for education and training* (2002), 293–307.
- [444] Mathew J Wilson and Max L Wilson. 2013. A comparison of techniques for measuring sensemaking and learning within participant-generated summaries. *Journal of the American Society for Information Science and Technology* 64, 2 (2013), 291–306.
- [445] Bin Wu, Chenyan Xiong, Maosong Sun, and Zhiyuan Liu. 2018. Query suggestion with feedback memory network. In *Proceedings of the 2018 World Wide Web Conference*. 1563–1571.

- [446] Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI conference on human factors in computing systems*. 1–22.
- [447] Wan-Ching Wu, Diane Kelly, Ashlee Edwards, and Jaime Arguello. 2012. Grannies, tanning beds, tattoos and NASCAR: Evaluation of search tasks with varying levels of cognitive complexity. In *Proceedings of the 4th Information Interaction in Context Symposium*. 254–257.
- [448] Jun Xiao, Richard Catrambone, and John Stasko. 2003. Be quiet? evaluating proactive and reactive user interface assistants. In *Proceedings of INTERACT*, Vol. 3. 383–390.
- [449] Michael Xieyang Liu, Advait Sarkar, Carina Negreanu, Ben Zorn, Jack Williams, Neil Toronto, and Andrew D Gordon. 2023. ” What It Wants Me To Say”: Bridging the Abstraction Gap Between End-User Programmers and Code-Generating Large Language Models. *arXiv e-prints* (2023), arXiv–2304.
- [450] Yusuke Yamamoto. 2012. Disputed sentence suggestion towards credibility-oriented web search. In *Asia-Pacific Web Conference*. Springer, 34–45.
- [451] Yusuke Yamamoto and Satoshi Shimada. 2016. Can Disputed Topic Suggestion Enhance User Consideration of Information Credibility in Web Search?. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. 169–177.
- [452] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. arXiv:2210.03629 [cs.CL]

- [453] Matin Yarmand, Srishti Palani, and Scott Klemmer. 2021. Adjacent Display of Relevant Discussion Helps Resolve Confusion. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–11.
- [454] Ann Yuan, Andy Coenen, Emily Reif, and Daphne Ippolito. 2022. Wordcraft: story writing with large language models. In *27th International Conference on Intelligent User Interfaces*. 841–852.
- [455] Xiaojun Yuan and Ryen White. 2012. Building the trail best traveled: effects of domain knowledge on web search trailblazing. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1795–1804.
- [456] Lisl Zach. 2005. When is “enough” enough? Modeling the information-seeking and stopping behavior of senior arts administrators. *Journal of the American Society for Information Science and Technology* 56, 1 (2005), 23–35.
- [457] Hua-Jun Zeng, Qi-Cai He, Zheng Chen, Wei-Ying Ma, and Jinwen Ma. 2004. Learning to cluster web search results. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. 210–217.
- [458] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, and Zengfu Wang. 2009. Visual query suggestion. In *Proceedings of the 17th ACM international conference on Multimedia*. 15–24.
- [459] Amy X Zhang, Jilin Chen, Wei Chai, Jinjun Xu, Lichan Hong, and Ed Chi. 2018. Evaluation and refinement of clustered search results with the crowd. *ACM Transactions on Interactive Intelligent Systems (TiiS)* 8, 2 (2018), 1–28.
- [460] Xiaolong Zhang, Yan Qu, C Lee Giles, and Piyou Song. 2008. CiteSense: support-

ing sensemaking of research literature. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 677–680.

[461] Yinglong Zhang and Robert Capra. 2019. Understanding how people use search to support their everyday creative tasks. In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval*. 153–162.

[462] Yinglong Zhang, Rob Capra, and Yuan Li. 2020. An In-situ Study of Information Needs in Design-related Creative Projects. In *Proceedings of the 2020 Conference on Human Information Interaction and Retrieval*. 113–123.