

UCLA

UCLA Electronic Theses and Dissertations

Title

Meaning, Computation, and Reasoning: Four Studies

Permalink

<https://escholarship.org/uc/item/5471f6f0>

ISBN

9798190944815

Author

Yildiz, Zeynep Senahan

Publication Date

2026-09-11

Peer reviewed|Thesis/dissertation

UNIVERSITY OF CALIFORNIA

Los Angeles

Meaning, Computation, and Reasoning: Four Studies

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

by

Zeynep Senahan Yildiz

2026

© Copyright by

Zeynep Senahan Yildiz

2026

ABSTRACT OF THE DISSERTATION

Meaning, Computation, and Reasoning

Four Studies

by

Zeynep Senahan Yildiz

Doctor of Philosophy in Philosophy

University of California, Los Angeles, 2026

Sean Walsh, Co-Chair

Michael Rescorla, Co-Chair

This dissertation brings together four studies. The first two address practical representation, drawing on programming-language semantics and motor representation. The latter two are collaborative studies that develop and test methods for evaluating psychological depth in large language models.

The first study examines Carlotta Pavese's account of practical modes of presentation, in which operational semantic values serve as practical Fregean senses and algorithms as their referents. I show that the account does not establish how distinct operational

semantic values can be more fine-grained than an algorithm while sharing it as their referent. I argue that algorithms cannot occupy the intermediate level of granularity between operational semantic values and the functions they compute.

The second study develops a positive account of practical representation as a distinctive representational kind. I frame the central dispute as one between a mode of presentation view, in which a task is represented in a distinctively practical way, and a mode of use view, in which ordinary task representations are practically deployed in action. I examine how motor representational format bears on this dispute. I then take Pavese's capacity-relative view of practical modes of presentation as the starting point for a positive account, and I introduce practical constancies to specify the capacities that play the relevant constitutive role. I propose that motor equivalences instantiate a distinctively practical form of representational constancy by preserving the same operationally articulated task across variations in execution.

The third study introduces the Psychological Depth Scale, which is a reader-centered framework for assessing five dimensions of human- and LLM-generated fiction. The study first tests whether human readers can apply the PDS consistently, then examines whether those judgments can be approximated automatically. It finally uses the scale to compare human- and LLM-authored fiction, finding that GPT-4-generated stories score higher than highly rated human-written stories on empathy and narrative complexity and comparably on the remaining dimensions.

The fourth study asks whether an LLM-based evaluator validated against human judgments on this original dataset continues to track human judgments when deployed on closely matched frontier-model outputs. Human readers show no uniform preference

for reasoning-enhanced models, whereas the automated judge shows an almost uniform preference for reasoning-model outputs. Agreement with human ratings on one dataset therefore does not guarantee that an automated evaluator will track human judgments on a different distribution of outputs, especially when those judgments are subjective.

Through philosophical and empirical inquiry, the dissertation seeks to address questions concerning meaning, computation, and reasoning.

The dissertation of Zeynep Senahan Yildiz is approved.

Gabriel Greenberg

Amit Sahai

Sean Walsh, Committee Co-Chair

Michael Rescorla, Committee Co-Chair

University of California, Los Angeles

2026

CONTENTS

ABSTRACT OF THE DISSERTATION	ii	
CONTENTS	vi	
ACKNOWLEDGEMENTS	ix	
VITA	xi	
PUBLICATIONS	xi	
Chapter 1 CAN ALGORITHMS BE FREGEAN REFERENTS OF PROGRAMS?		
RESPONSE TO PAVESE’S “PRACTICAL SENSES”	12	
Chapter 1 1 Introduction	12	
Chapter 1 2 Overview of Pavese’s account	13	
Chapter 1 3 Why algorithms cannot serve as referents of practical senses	15	
Chapter 1 3.1 The program–algorithm distinction and Pavese’s paradigm case	16	
Chapter 1 3.2 Algorithms as equivalence classes of OSV	19	
Chapter 1 3.3 Unresolved granularity between algorithms and OSVs: implications on Hesperus–Phosphorus cases	21	
Chapter 1 4 Rethinking practical reference as functions	26	
Chapter 1 5 Conclusion	29	
Chapter 1 6 References	29	
Chapter 2 MODE OF PRESENTATION VS. MODE OF USE IN PRACTICAL REPRESENTATION.....		34
Chapter 2 1 Introduction	34	
Chapter 2 2 Mode of presentation vs. mode of use	39	
Chapter 2 3 Examining representational format in practical representation	45	

Chapter 2 4 Pavese’s capacity-relative account of practical modes of presentation..	59
Chapter 2 5 Practical constancies	64
Chapter 2 6 Conclusion	70
Chapter 2 7 References	72
Chapter 3 PSYCHOLOGICAL DEPTH IN LARGE LANGUAGE MODELS	75
Chapter 3 1 Introduction	75
Chapter 3 2 Study 1: Measuring Psychological Depth in Language Models	79
Chapter 3 2 1 Motivation and General Approach	79
Chapter 3 2 2 The Five Dimensions of Psychological Depth	81
Chapter 3 2 3 Construction of the PSYCHDEPTH Dataset	84
Chapter 3 2 4 Human Evaluation	86
Chapter 3 2 5 Reliability of Human PDS Judgments	86
Chapter 3 2 6 Automating Psychological-Depth Evaluation	87
Chapter 3 2 7 Psychological Depth in Human and LLM Stories	89
Chapter 3 2 8 Discussion and Limitations of the Initial PDS Study	91
Chapter 3 3 Study 2: Reasoning-Enhanced Models and the Validity of Automated Psychological-Depth Evaluation	93
Chapter 3 3 1 Motivation	93
Chapter 3 3 2 Construction of the Frontier Story-Pair Dataset	94
Chapter 3 3 3 Construction of the Automated PDS Evaluator	96
Chapter 3 3 4 Human Evaluation	97
Chapter 3 3 5 Statistical Analysis	99
Chapter 3 3 6 Automated Evaluation Strongly Favors Reasoning Outputs	100

Chapter 3 3 7 Human Readers Do Not Show a Universal Reasoning Advantage	101
Chapter 3 3 8 Human Agreement Falls Sharply on Closely Matched Stories	102
Chapter 3 3 9 Reader Disagreement Shows Structure	103
Chapter 3 3 10 Surface Properties and Automated Judge Scores	104
Chapter 3 3 11 Development-Set Agreement Does Not Guarantee Deployment Validity	106
Chapter 3 3 12 Limitations	108
Chapter 3 4 Conclusion	108
Chapter 3 5 References	109

ACKNOWLEDGEMENTS

I would like to thank my committee co-chairs, Sean Walsh and Michael Rescorla, for their guidance and support throughout the development of this dissertation. I am also grateful to Gabriel Greenberg and Amit Sahai for serving on my committee and for their comments, advice, and support. I would also like to thank my collaborators and colleagues at UCLA whose contributions made the collaborative research included in this dissertation possible. I am grateful as well to my family and friends for their encouragement and support throughout graduate school.

Chapter 3 contains material from two coauthored studies. The first study is adapted from Harel-Canada, Fabrice Y.; Zhou, Hanyu; Muppalla, Sreya; Yildiz, Zeynep Senahan; Kim, Miryung; Sahai, Amit; and Peng, Nanyun. “Measuring Psychological Depth in Language Models.” Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 17162–17196. DOI: 10.18653/v1/2024.emnlp-main.953. Amit Sahai and Nanyun Peng jointly directed and supervised the research as equal co-advisors. Fabrice Harel-Canada, Hanyu Zhou, Sreya Muppalla, and Miryung Kim contributed as coauthors to the development, execution, analysis, and preparation of the study. The dissertation author conducted a revised literature review underlying the Psychological Depth Scale, rewrote the section establishing the theoretical and empirical foundations of its five components, clarified the conceptual distinctions among those components, and contributed to revisions of the interpretation and presentation of the empirical results.

Chapter 3 also contains material from Zheng, Ruichen; Wang, Yihe; Harel-Canada, Fabrice Y.; Khosravi, Sara; Yildiz, Zeynep Senahan; Sahai, Amit; and Peng, Nanyun. “Does Reasoning Improve Psychological Depth in Large Language Models? Depends on Who’s Judging.” Amit Sahai and Nanyun Peng jointly directed and supervised the research. Ruichen Zheng, Yihe Wang, Fabrice Y Harel-Canada, and Sara Khosravi contributed as coauthors to the development, execution, analysis, and preparation of the study. The dissertation author contributed to the conceptual development of the project, including the formulation of hypotheses concerning the relationship between different forms of reasoning and psychological depth; the classification and selection of candidate reasoning tasks and datasets; the design of prompts for generating narrative samples; the annotation of model-generated stories using the Psychological Depth Scale; and the interpretation of results concerning differences among models across the five PDS dimensions.

VITA

2019 Bachelor of Arts in Philosophy, Boğaziçi University

2021 Master of Arts in Philosophy, University of California, Los Angeles

PUBLICATIONS

Harel-Canada, Fabrice Y.; Zhou, Hanyu; Muppalla, Sreya; Yildiz, Zeynep Senahan;

Kim, Miryung; Sahai, Amit; and Peng, Nanyun. “Measuring Psychological Depth in Language Models.” Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2024, pp. 17162–17196

Chapter 1 CAN ALGORITHMS BE FREGEAN REFERENTS OF PROGRAMS?

RESPONSE TO PAVESE'S "PRACTICAL SENSES"

Chapter 1 1 Introduction

How can agents act skillfully without consciously formulating what they do?

Philosophers have variously characterized practical forms of thought which directly guide action—as “action-based modes of presentation” (Peacocke 1986, p. 49; 2007, p. 360), “executable ideas” (Israel et al., 1993, p. 537), “action concepts” (Pacherie 2000, p. 408; 2011, p. 69), and “practical modes of presentation” (Stanley and Williamson, 2001, p. 429). The challenge, however, is to specify what these representations are and how they guide action.

Pavese (2015) offers a concrete proposal: she argues that *computer programs*, understood as *operational semantic values of program texts*, constitute blueprints for practical ways of thinking. This is an appealing framework, since operational semantics is used for modeling the *behavior* of a machine “in interpreting and compiling program texts” (Pavese, 2015, p. 5). In turn, operational semantic values are formal specifications of how computer programs are executed, so they capture the operational meaning of program texts. Compiling a program text is akin to understanding how to carry out an instruction: computers acquire an *ability* to follow the text’s operational semantic value. Compilers thus ‘grasp’ an operational semantic value as the distinctively practical sense of a program text. Rather than understanding this sense as a human would, their grasp consists in rendering the system constitutively able to act in accordance with it.

However, Pavese makes the further claim that *algorithms* are the Fregean referents of programs. This claim creates a difficulty: if programs are ways of thinking of

algorithms, then programs must map many-to-one onto algorithms, which themselves map many-to-one onto functions. Algorithms must therefore occupy an intermediate granularity between the programs (operational semantic values) that determine them and the tasks they execute. In this paper, I argue that no account secures this position without either mishandling coreference or violating Fregean assumptions. That is why one cannot sustain an account of practical senses if one takes their referents to be algorithms.

Chapter 1 2 Overview of Pavese's account

Operational semantics defines the meaning of program expressions in terms of “the behaviour of programs during execution” (Pitts, 2002, p. 379). It is the most commonly used formal semantics for programming languages (Barman, 2004, p. 43; Steingartner et al., 2020, p. 1182). Pavese's approach is to explicate programs as the *operational semantic values* (OSVs) of program texts (Pavese, 2015, p. 5). She thereby draws a distinction between *programs* and *program texts*, where programs, as OSVs, are the Fregean senses of program texts. Pavese further considers algorithms to be the referents of OSVs, where algorithms constitute “ways to execute tasks” (2015, p. 2). In this picture, an *algorithm* for calculating a function f is a way of performing the task of calculating f . Program texts are signs which designate algorithms as their referents. If the task of calculating f is abbreviated as ϕ -ing, the OSV of a program text to ϕ is the Fregean practical sense containing a practical mode of presentation of the algorithm for ϕ -ing.

Pavese argues that OSVs, as Fregean senses, have a distinctively practical nature because grasping them (1) requires “rule-following capacities” and (2) entails “rule-following abilities” (p. 10). When one has a capacity to follow a rule to ϕ , one is in a

position to grasp the practical sense of that rule. That is, if properly instructed, one would become “able to ϕ by following that instruction” (p. 10). Successfully grasping a practical sense thus entails that one becomes able to ϕ “whenever all the conditions for the exercise of that ability [are] present”.

OSVs are inferential rules for interpreters and compilers built for processing program texts. For an interpreter to “grasp the OSV of a program text to ϕ ”, the interpreter must “understand the primitive rules corresponding to each part” of the OSV (Pavese 2015, p. 11). Moreover, the interpreter must also understand relevant primitive rules about how parts make wholes. Grasping those primitive rules in turn “endows the interpreter with the ability to follow each part” in the appropriate order, allowing the system to acquire “the ability to follow a rule to ϕ ”. The fact that a computer can successfully ϕ upon processing an OSV to ϕ is what explains its distinctively practical nature. Suppose one computer runs a compiler on a program text, while another runs a word processor to display the same text. Both process the same text, yet only the compiler produces executable instructions, thereby realizing the OSV of the text. In Pavese’s framework, this is because OSVs are distinctively practical senses of program texts. The compiler’s software design enables it to ‘grasp’ the OSV of a program text, through which the computer acquires the ability to follow it as an inferential rule. The word processor, on the other hand, may be said to ‘grasp’ only a perceptual or demonstrative sense of the text, understood as tokens to be displayed and arranged in fonts, sizes, and layouts. Similarly, two people may receive the same instruction on how to ϕ , with one grasping its practical sense and the other only a non-practical sense. This difference can explain why one might fail to acquire a skill despite unlimited instruction. On Pavese’s account, generalizing from OSVs to human cases of skill acquisition thus

serves to establish practical senses as a genuine and distinct category of mode of presentation.

Chapter 1 3 Why algorithms cannot serve as referents of practical senses

Algorithms stand in a many-to-one relation to functions. This relation is predicated on the fact that algorithms are demonstrably more finely-grained than functions. If an algorithm is understood as a *way* of ϕ -ing, then the task of ϕ -ing can be modeled as the function computed by that algorithm (Pavese, 2015, p. 2). A function which models the task of ϕ -ing maps any starting point of undertaking ϕ to the appropriate end point of achieving ϕ . Algorithms, on the other hand, specify particular paths from starting points to corresponding outcomes. Such paths could thus capture different ways of breaking down the same task. That is why algorithms are more finely grained than functions: they capture distinct ways of detailing their functions.

If algorithms are supposed to be the referents of OSVs, then OSVs must stand in a many-to-one relation to algorithms, in the same way that algorithms stand to functions. Yet in the case of algorithms and functions, the many-to-one relation is well understood: it is grounded in a distinct difference of granularity. By contrast, OSVs and algorithms exhibit no comparably clear granularity difference that could ground such a relation. To justify algorithms¹ as the referents of OSVs, one must therefore give an account of granularity under which OSVs are strictly finer than algorithms. This is crucial for a satisfactory model of practical senses: it must not only allow for coreference but also provide a principled explanation of it.

In the rest of this section, I evaluate how far Pavese's account bears on this challenge. Her treatment of programs qua *modes of presentation of algorithms* unfolds

in three steps. First, she appeals to the conventional program–algorithm distinction in computer science, claiming that it could be “modeled after the sense–reference relation” (Pavese, 2015, p. 4). Second, she presents what she takes to be a paradigm case: two program texts that are meant to show that distinct programs can determine the same algorithm. Third, on the basis of her paradigm case and by appealing to equivalence class views of algorithms, she suggests that algorithms can be conceived as classes of OSVs. I show that these steps exhibit a reliance on interpretive errors and that the account is fundamentally compromised by a loose notion of algorithm that yields problematic results about coreference. These issues ultimately show that algorithms cannot plausibly serve as Fregean referents of practical senses.

Chapter 1 3.1 The program–algorithm distinction and Pavese’s paradigm case

Pavese claims that the program–algorithm distinction, which she notes “computer scientists routinely draw” (2015, p. 3; citing Brookshear 2007 and Knuth 1966, 1974), can be used to model a Fregean sense–reference relation. This proposal quickly runs into trouble. Fregean senses, by Pavese’s own description, must be “non-concrete” and “[n]on-linguistic” (Pavese, 2015, p. 4). Yet on the conventional distinction, as the very works Pavese cites exemplify, treats programs as concrete and linguistic. Brookshear, for example, likens the algorithm–program relation to “that between a story and a book” (2017, p. 214): programs are code-based instances of abstract algorithms, much like a book is a “physical representation of a story”. Knuth (1966, 1974) likewise describes programs as their linguistic embodiments of algorithms in particular programming languages. In these works, and in computer science literature more broadly, programs typically mean concrete bits of code written in a programming language.

By contrast, Pavese claims that the conventional treatment of programs has two components, one for *program texts* and one for *semantic values*, and she proposes to reserve the term *program* for the semantic-value component, which she takes to be “more clearly non-linguistic” (2015, p. 5). The distinction is clear, but she does not show that the word program in fact tracks semantic values. The conventional use of program fits what she calls program texts rather than semantic values. Hence the conventional program–algorithm distinction cannot underwrite a model of the sense–reference relation. After this non-standard redefinition, Pavese seeks motivation. She holds that it should be possible to think of “the same algorithm ... under different (practical) modes of presentation” (2015, p. 4, n. 8). However, she offers no independent argument that algorithms must admit different modes of presentation. Instead, she presents a paradigm case: two program texts purported to show distinct practical modes of presenting the same algorithm.

The example involves two program texts, A and B, written in different languages $\mathcal{L}_1$ and $\mathcal{L}_2$, respectively. Both A and B aim to multiply n and m by adding n to itself $m - 1$ times. $\mathcal{L}_1$ has the binary addition rule $ADD(n, m)$ as one of its primitive OSVs, which takes n and m and returns $n + m$. To multiply n and m , A calls itself on $(n, m - 1)$ to use ADD in sequence:

PROGRAM TEXT A

MULT(n, m): If $m = 0$, then 0,

else SET v to MULT($n, m-1$); SET res to ADD(n, v); res . (Pavese, 2015, p. 6)

Program text A calls itself on $(n, m - 1)$ and adds n to the value returned. The call on $(n, 0)$ is preset to return 0. On any (n, m) , A thus builds up from the base call,

adding n at each step. $\mathcal{L}_2$, by contrast, has another primitive rule $ADD_n(m)$, which takes n and m and then adds n to itself $m - 1$ times. This rule can carry on multiplication of n and m , since it adds m many n 's:

$$ADD_n(m) = \underbrace{(n + n) + \dots}_{m-1} = \underbrace{n + \dots}_m$$

So, the Program Text B, written in $\mathcal{L}_2$ consists only of the rule $ADD_n(m)$:

Program Text B

MULTIPLY(n, m): $ADD_n(m)$ (Pavese, 2015, p. 6)

The OSV of Program Text A is derived from the primitive OSVs in $\mathcal{L}_1$ for binary addition, value assignment, conditional execution, and sequencing (Pavese, 2015, pp. 5–6). In contrast, the OSV of Program Text B is “defined in terms of the primitive basic command $ADD_n(m)$ ” (p. 6). Since both A and B perform multiplication by adding n to itself $m - 1$ times but are assigned different OSVs, Pavese concludes that their OSVs constitute different modes of presentation of the same algorithm.

This example fails to illustrate distinct modes of presentation of the same algorithm for two key reasons. First, while Pavese claims that A and B are “linguistically encoding the same algorithm” (2015, p. 7), she does not clarify what this means. She characterizes the algorithm only as “the algorithm that consists in multiplying n by m by adding n to itself $m - 1$ times in a row” (p. 6), without clarifying what notion of algorithm underlies this description. In effect, she appears to rely on the conventional program–algorithm distinction. This reliance is problematic because, as established previously, the conventional view treats programs as concrete expressions of abstract algorithms. Second, the example suggests *only* that the corresponding semantic values differ under

Pavese's formalism. It does not establish such differences as true differences in Fregean sense. Formal operational semantics is not only a system that assigns meaning; it itself is yet another system of concrete expression. As such, it may generate notational distinctions between OSVs that solely underpin differences of syntactic configuration rather than real differences in operational meaning.

The OSVs of A and B are thus represented in this system of concrete expression. Their seeming differences may thus arise entirely from their respective syntactic configurations. Pavese herself concedes that the difference between the OSVs of A and B “depends in part on the syntax of the relevant programming language” (2015, p. 7). Crucially, however, she does not specify what else, beyond syntax, justifies a distinction between the OSVs of A and B qua non-concrete Fregean senses. Her paradigm example therefore leaves open the possibility that these differences are entirely syntactic—and thus irrelevant to the meaning they express.

Chapter 1 3.2 Algorithms as equivalence classes of OSV

Pavese concludes her paradigm case by proposing that algorithms can be conceived as “classes of OSVs” (2015, p. 7). She cites computer-science discussions of *the equivalence-class view of algorithms* (Blass et al. 2009; Moschovakis 1998, 2001), claiming that it is “how many computer scientists think of algorithms” (Pavese, 2015, p. 7). On that view, an algorithm is an equivalence class of programs; a formal equivalence relation fixes algorithmic sameness. The trouble is that the view faces serious difficulties, and even the very sources Pavese invokes undercut the view. For example, Blass et al. (2009) claim that algorithm sameness is “a subjective notion” even amongst experts (p. 156), and that “different purposes will yield different notions of sameness” (p. 159). Likewise, Moschovakis contends that “a separate notion of

algorithm” is required to establish program equivalence (Moschovakis, 1998, p. 81). In short, the equivalence class view of algorithms face the difficulty of specifying an equivalence relation that captures “the intuitive notion of ‘the same algorithm’” (Blass et al., 2009, p. 145). Crucially, both discussions of program equivalence still target program texts. Moschovakis describes programs as “syntactic objects” (1998, p. 79) and treats them as Fregean signs, with algorithms as their meanings. Blass et al. also describe programs as written in a language (2009, p. 145). Their worries thus track a familiar philosophical problem, akin to Quine (1951) on synonymy: when do distinct linguistic expressions convey the same meaning?

The issue is that Pavese does not treat programs as linguistic expressions. She identifies programs (OSVs) with *senses* of such expressions and distinguishes OSVs from intuitive algorithms: an OSV is a mode of presentation of an intuitive algorithm. Consequently, on Pavese’s account, program equivalence is no longer about two texts expressing the same algorithm. Instead, it concerns the claim that two distinct Fregean senses, grasped from different texts, determine the same algorithm. That is a harder task. For distinct coreferential senses are supposed to be cognitively distinct ways of presenting the same referent: ways of thinking that are, by definition, different in cognitive significance.

If one instead defined algorithms as classes of concrete, linguistic program representations, the relevant relation would be synonymy of expressions. The difficulties noted by Moschovakis and Blass et al. would persist; at the limit, Quine (1951) famously rejects a robust notion of synonymy. Fregean senses might then face a *synonymy crisis*: one might argue that whether different signs, such as ‘horse’ and ‘steed’, are synonymous is up to subjective evaluation.

However, Pavese's framework introduces a further challenge, which we may call a *coreference crisis*. Briefly, the equivalence-class view must now ground coreference of nonsynonymous program texts. Yet what exactly does it mean for distinct OSVs to determine the same algorithm as their referent? As Fregean senses, distinct coreferential OSVs must carry distinct cognitive significance: they must capture different ways of thinking of the same algorithm. So, differences between coreferential OSVs to ϕ must *cause* one to understand them as different ways of ϕ -ing. Yet coreferential OSVs must also fix the distinctive way of ϕ -ing that sets *their* algorithm apart from rival algorithms for ϕ -ing. Hence a granularity problem: at what level do we individuate algorithms, and how fine-grained must OSVs be to count as genuinely distinct while still coreferring?

Chapter 1 3.3 Unresolved granularity between algorithms and OSVs: implications on Hesperus–Phosphorus cases

The coreference crisis returns us to the central difficulty motivating this critique: a satisfactory model of practical senses as OSVs must establish a granularity difference between OSVs and their referents. Pavese neither states nor justifies such a difference in grain. The question is whether OSVs can be more fine-grained than algorithms in a way that the many-to-one relation without collapse into trivial differences. With concrete distinctions off the table, how should we interpret OSVs as more finely individuated than algorithms? Relatedly, how should we place algorithms between coarse-grained functions and very fine-grained OSVs?

The closest Pavese comes to a statement of the granularity relation is at p. 12: algorithms are “effective ways of breaking down a task into parts,” while OSVs “present an algorithm by presenting every part of a task (as broken down by that algorithm) as a

particular inferential rule” (Pavese, 2015, p. 12). Yet this comparison seems to place them at the same level of detail rather than marking a finer level for OSVs. Worse, she attributes algorithms a cognitive or intensional profile. To effectively break down a task is to think through the task. Algorithms are then ways of thinking (about tasks). Since OSVs, as Fregean senses, are ways of thinking about their referents, it follows that OSVs are ways of thinking about ways of thinking. We then face the odd result that coreferential OSVs are different ways of thinking about the same way of thinking. But how could coreferential OSVs be differentiated when they are ways of thinking about something which is itself a way of thinking? What relation between OSVs could secure coreference without erasing their distinct cognitive contributions?

The challenge is clearer in *Hesperus–Phosphorus* cases, where differences between coreferential senses support differences of epistemic attitude. Coreferential Fregean senses must accommodate ignorance, exploration, and discovery regarding their coreferentiality, and their differences must explain why such epistemic possibilities arise. This follows from Frege’s *criterion of difference* (Perry, 1977; Evans, 1982, pp. 18–19): two senses differ if one can grasp both and yet adopt divergent epistemic attitudes about them without committing oneself to irrationality (pp. 18-19). Evans adds that ways of thinking should not be so fine-grained to block differences in epistemic attitudes (1982, p. 20). Thus, on the one hand, OSVs must be more finely individuated than algorithms so that distinct OSVs can determine the same algorithm. On the other hand, OSVs must not be so fine-grained that their differences fail to register genuine cognitive difference.

A natural reply would be that distinct coreferential OSVs capture the same algorithm at different explanatory levels, understood as *levels of abstraction*. Recall

Pavese's paradigm example of program texts A and B, written in $\mathcal{L}_1$ and $\mathcal{L}_2$, for the task of multiplication. $\mathcal{L}_2$ includes a primitive OSV for sequential addition that effectively implements multiplication, so the OSV of B consisted only of that primitive. $\mathcal{L}_1$ lacks that primitive, so the OSV of A uses multiple primitives for binary addition, control, and sequence. Yet both OSVs are specified under the same operational semantic framework, where $+$ specifies both A's binary addition primitive $ADD(n, m)$ and B's n-ary addition primitive $ADD_n(m)$. In fact, the same operator $+$ must count as a derived OSV in a radically low-level language such as Turing's system. So one could argue that B's primitive OSV is derived just in the way A's is; B merely packages that derivation as a primitive of $\mathcal{L}_2$. This packaging is *abstraction*. Namely, the OSVs of A and B differ because their corresponding parts involve different levels of abstraction, which would be the sort of difference Pavese's account needs.

In general, Fregean senses that represent the same referent at different levels of abstraction may ground differences of epistemic attitudes, where abstraction more broadly applies to detail or description. We may say, for example, the mathematical expressions $(1 + 1) + 1$ and $2 + 1$ both denote the number three, but their senses present it at different levels of detail. These senses would satisfy Frege's criterion of difference, since one can grasp the sense of a mathematical expression without having to calculate it. But this route cannot secure the required difference in Pavese's framework. Pavese applies abstraction to algorithms as instructions, not to representations in a broad descriptive domain. Subpersonal descriptions (e.g., motor-cortex activity during bicycle riding) do not supply an instructional level at which an agent can act. The right contrast, if any, must *involve levels of instructional precision*.

However, if practical senses satisfy Frege's criterion only in virtue of differences in levels of instructional precision, a further difficulty arises. This difficulty concerns the distinctively practical nature of these senses and what it means to grasp them. Because OSVs are inferential rules, grasp of a high-level OSV requires derivation of its components, which bottoms out in grasp of a lower-level OSV. Grasping a *practical* sense entails an ability, so grasping them across different abstraction levels translate to *the same practical skill*. In programming practice, higher- and lower-level programs all compile to machine code, which is the lowest level of program execution. During compilation, primitives of higher-level languages correspond to derived OSVs of lower-level ones, and differences disappear at execution. *Intermediate languages* (ILs) reinforce this point: ILs are intermediary representations generated *during the compilation* of high-level code. Some ILs are specifically designed to serve as a common lower-level step for multiple high-level languages. Program texts are translated into ILs *as part of the compilation process*. Microsoft's Common Intermediate Language within .NET, for example, serves multiple source languages such as C#, F#, and VB.NET. It is possible to give operational semantics for such languages and their common IL in such a way that accounts for compilation of the higher-level languages into their common IL. We might even interpret Pavese's program text A as the intermediate result of compiling program text B, where $\mathcal{L}_1$ is the lower-level IL that the higher-level is $\mathcal{L}_2$ compiled into. This interpretation is consistent with how high-level instructions are compiled into more explicit instruction sequences. The supposed difference in instructional precision thus disappears at the level of transforming into practical abilities.

In fact, human action shows the same pattern. Consider:

ihigh: *Hold on the handlebars and push down on the pedal with one foot and get on the bicycle. Start pedaling forward.*

ilow: *With your left foot on the ground, hold the handlebars with both hands and take your right foot over the bicycle onto the right pedal. Make sure that the pedal is at its highest point before placing your foot on it. Then push down the pedal as you raise up and place your left foot on the left pedal. Sit on the seat as you start pedaling forward.*

A subject who grasps either instruction acquires the same skill. The details that further specify the instruction must have been cognitively accessible to them as prior practical skills. These could be acquired via prior experience in the case of humans, or a built-in capacity in the case of language processors. It is difficult to imagine someone can contemplate practical senses that translate to the same practical skill, while at the same time holding different epistemic attitudes toward them. Differences in instructional precision do not obviously ground stable differences in epistemic attitude of the Fregean sort. Thus, levels of instructional precision seem to constitute a questionable basis for differences of epistemic attitude. A theoretically useful Fregean model should have a reasonable amount of intuitive and obvious Hesperus-Phosphorus cases. Pavese's framework has not shown promise in accounting for any natural and obvious instances of them. Without a principled and explanatory way of distinguishing coreferential OSVs, one cannot establish a many-to-one determination relation between OSVs and algorithms. The result is that algorithms cannot plausibly serve as the referents of practical modes of presentation, which undermines the explanatory promise of Pavese's account.

Chapter 1 4 Rethinking practical reference as functions

If algorithms cannot plausibly be treated as referents of practical senses, then we should explore how Pavese's account fares without this assumption. Functions are conventionally treated as Fregean referents, and Pavese herself introduces them as a second layer of the sense–reference relation (Pavese, 2015, p. 4, footnote 8). In this section, I briefly explore how Pavese's framework can be developed by treating functions as the referents of OSVs.

Fregean approaches to computational expressions typically consider referents of such expressions to be *extensionally given functions* computed by them (Church, 1941, 1993; Moschovakis, 1994; Klement 2001; van Heijenoort, 1977; Dummett, 1978; Horty, 2007; Girard, 1989; Martin-Löf, 2021; Bentzen, 2020; Muskens, 2005). Given that there already exists a prominent model of Fregean reference for computational expressions, why does Pavese not adopt it? Pavese rules out this possibility in a footnote, claiming that the Fregean many-to-one determination is not, or “at least not obviously”, established from probabilistic algorithms to their functions (Pavese, 2015, p. 4, footnote 8). This claim does not succeed since nondeterministic algorithms do have extensionally defined functions, which can be given as an array of possible outputs on each input, which may incorporate probability distributions and external oracles. Nevertheless, Pavese suggests that the algorithm-function relationship may be incorporated into her account as another layer of sense-reference (Pavese, 2015, p. 4, footnote 8). The two-layered Fregean framework appears in a table (p. 13, Table 2) further down in her paper, but was left unexplained.

Let us define *the function model of reference* as the minimally modified version of Pavese's account, in which functions, rather than algorithms, are taken as the referents

of practical senses. The referents would be, following Church's terminology, "functions in extension", which capture *output values*, *range of output values*, or *sets of ordered pairs of input-output values* (Church, 1941, p. 3). The referent of a computational expression could be the output value that is determined by computing the expression on a given input value. And the computational expressions that are not fed input values would be incomplete expressions, and denote functions. We may adopt a set-theoretic definition of functions as a set of ordered pairs of input-output values—an exhaustive set of every input-output pair generated by computing the function on every input on the function's domain. The senses of computational expressions would then be the *rules* or *methods* by which referents are computed, which is consistent with modeling them as OSVs.

If OSVs are taken to determine functions instead of algorithms, how are algorithms to be understood within such a framework? First, we may reinterpret the link between OSVs and algorithms and ask: why not say OSVs are algorithms? On this approach, OSVs capture what we have in mind when we speak of the intuitive notion of an algorithm. Pavese already introduces an unconventional conception of programs—as OSVs of program texts. But in the computer science literature, programs are usually taken to be concrete representations of algorithms. We have seen that all of the works Pavese cites to draw on the program–algorithm distinction and the equivalence class view treat programs as what Pavese considers program texts. If we carry this conventional assumption into Pavese's framework, programs would no longer be explicated as OSVs but rather they would be identified with program texts. And the role of OSVs would be more naturally assigned to algorithms rather than programs. This would mean that algorithms are understood through grasping the practical sense of the

relevant program text, which amounts to thinking about the relevant task under a practical mode of presentation. On this recasting, the many-to-one relation between practical senses and their referents would be securely established by the familiar many-to-one relation between algorithms and functions. This option does not resolve every difficulty, but it faces significantly fewer problems than the original view. In particular, it avoids the awkward characterization of OSVs as ways of thinking about ways of thinking. Instead, OSVs would simply be the algorithms that program texts express.

The second option is to drop the informal notion of algorithm altogether from Pavese's framework. The term algorithm is notoriously vague and contested. It invites unresolved questions about algorithm sameness and equivalence, and complicates any rigorous reasoning that hinges on it. This is not a new problem. Historically, the lack of precision around intuitive algorithms created the demand for a formalization of the intuitive notion, which culminated in the Church–Turing thesis. Any account of algorithms outside concrete formalizations will almost inevitably have an unclear or contested ontology, with unsatisfactory answers to questions about algorithm sameness or equivalence (Blass et al., 2009). One might then commit to speaking of algorithms only as defined within formalizations such as Church's lambda calculus, Turing machines, or any programming language. Algorithms would be nothing over and above their concrete formalizations: the program texts themselves. Any talk of “informal” or “intuitive” algorithms would drop out of the ontology. This would also bring Pavese's framework into closer alignment with the methodological rigor of theoretical computer science, where informal algorithmic descriptions are grounded in formal systems in virtue of the Church-Turing thesis. But the benefit is equally clear: it avoids ontological

ambiguities and secures a precise framework in which OSVs are practical modes of presentation of functions.

Chapter 1 5 Conclusion

Pavese makes an important contribution by showing how operational semantics can model practical modes of presentation. She provides a formal framework for understanding how grasping a practical sense can entail the acquisition of a practical ability. However, modeling the referents of OSVs as algorithms leaves the account unable to satisfy crucial Fregean requirements. In particular, this model fails to show how OSVs are more fine-grained than algorithms in a way that can ground coreference. That is why Pavese's framework cannot secure the explanatory power it seeks. To preserve the distinctive character of practical senses while allowing for principled and cognitively significant coreference, a different type of object must model their referents.

Chapter 1 6 References

- Barman, K. D. (2004). *Type theoretic semantics for programming languages* [Doctoral dissertation, Université Nice Sophia Antipolis]. Theses.fr.
- Bentzen, B. (2020). Sense, reference, and computation. *Perspectiva Filosófica*, 47(2), 179–203.
- Brookshear, G. (2007). *Computer science: An overview* (Pearson International ed.). Pearson.
- Burge, T. (2005). *Truth, thought, reason: Essays on Frege*. Oxford University Press.
- Buss, Samuel R., Alexander S. Kechris, Anand Pillay, and Richard A. Shore. 2001. "The Prospects for Mathematical Logic in the Twenty-first Century". *Bulletin of Symbolic Logic* 7 (2): 169–96.
- Carroll, L. (1895). What the tortoise said to Achilles. *Mind*, 4(14), 278–280.

- Chirimar, J., Gunter, C. A., & Riecke, J. G. (1992). Proving memory management invariants for a language based on linear logic. In *Proceedings of the 1992 ACM Conference on LISP and Functional Programming* (pp. 139–150). ACM Press.
- Church, A. (1941). *The calculi of lambda-conversion*. Princeton University Press.
- Church, A. (1993). A revised formulation of the logic of sense and denotation. *Noûs*, 27(2), 141–157.
- Demers, A., Weiser, M., Hayes, B., Boehm, H., Bobrow, D., & Shenker, S. (1990). Combining generational and conservative garbage collection: Framework and implementations. In *Proceedings of the 17th ACM SIGPLAN-SIGACT Symposium on Principles of Programming Languages* (pp. 261–269). ACM Press.
- Dummett, M. (1978). Frege’s distinction between sense and reference. In *Truth and other enigmas* (pp. 116–144). Harvard University Press
- Elsman, M. (2003). Garbage collection safety for region-based memory management. In *Proceedings of the 2003 ACM SIGPLAN International Workshop on Types in Languages Design and Implementation* (pp. 123–134). ACM Press.
- Evans, G. (1982). *The varieties of reference*. Clarendon Press.
- Evans, G. (1985). Understanding Demonstratives. In *G. Evans, Collected Papers* (pp. 291–321). Oxford University Press.
- Frege, G. (1997). On *Sinn* and *Bedeutung*. In M. Beaney (Ed.), *The Frege reader* (pp. 151–171). Blackwell Publishing. (Original work published 1892)
- Frege, G. (1997). Thought. In M. Beaney (Ed.), *The Frege reader* (pp. 325–345). Blackwell Publishing. (Original work published 1918)
- Girard, J. (1989). *Proofs and types*. Cambridge University Press.

- Hawley, K. (2003). Success and knowledge-how. *American Philosophical Quarterly*, 40(1), 19–31.
- Heidegger, M. (1996). *Being and time* (J. Stambaugh, Trans.). State University of New York Press. (Original work published 1927; based on the 1976 German edition)
- Horty, J. F. (2007). *Frege on definitions: A case study of semantic content*. Oxford University Press.
- Hudak, P. (1986). A semantic model of reference counting and its abstraction. In *Proceedings of the 1986 ACM Conference on LISP and Functional Programming* (pp. 351–363). ACM Press.
- Hunter, R., & Krishnamurthi, S. (2003). A model of garbage collection for OO languages. In *10th International Workshop on Foundations of Object-Oriented Languages (FOOL10)*.
- Israel, D., Perry, J., & Tutiya, S. (1993). Executions, motivations, and accomplishments. *Philosophical Review*, 102(4), 515–540. <https://doi.org/10.2307/2185682>
- Klement, K. C. (2001). *Frege and the logic of sense and reference*. Routledge.
- Knuth, D. (1966). Algorithm and program; information and data. *Communications of the ACM*, 9(9), 654.
- Knuth, D. (1974). Computer science and its relation to mathematics. *The American Mathematical Monthly*, 81(4), 323–343.
- Leal, M., & Ierusalimsky, R. (2005). A formal semantics for finalizers. *Journal of Universal Computer Science*, 11(7), 1198–1214. <https://doi.org/10.3217/jucs-011-07-1198>
- Martin-Löf, P. (2021). The sense/reference distinction in constructive semantics. *Bulletin of Symbolic Logic*, 27(4), 501–513.

- Merleau-Ponty, M. (2002). *Phenomenology of perception* [*Phénoménologie de la perception*] (C. Smith, Trans.). Routledge.
<https://doi.org/10.4324/9780203994610> (Original work published 1945; English translation originally published 1962)
- Miller, D. Reciprocal Influences Between Proof Theory and Logic Programming. *Philos. Technol.* 34, 75–104 (2021).
- Morrisett, G., Felleisen, M., & Harper, R. (1995). Abstract models of memory management. In *Proceedings of the 1995 ACM Conference on Functional Programming Languages and Computer Architecture* (pp. 66–77). ACM Press.
- Morrisett, G., & Harper, R. (1998). Semantics of memory management for polymorphic languages. In *Higher order operational techniques in semantics* (pp. 175–226). Cambridge University Press.
- Moschovakis, Y. N. (1994). Sense and denotation as algorithm and value. In J. Väänänen & J. Oikkonen (Eds.), *Logic colloquium '90* (Vol. 2, pp. 210–249). Association for Symbolic Logic.
- Muskens, R. (2005). Sense and the computation of reference. *Linguistics and Philosophy*, 28, 473–504. <https://doi.org/10.1007/s10988-004-7684-1>
- Pacherie, E. (2000). The content of intentions. *Mind & Language*, 15(4), 400–432.
- Pavese, C. (2015). Practical senses. *Philosophers' Imprint*, 15, 1–25.
- Pavese, C. (2020). Practical representation. In E. Fridland & C. Pavese (Eds.), *The Routledge handbook of philosophy of skill and expertise* (pp. 233–245). Routledge.
- Peacocke, C. (1986). *Thoughts: An essay on content*. Blackwell.

- Peacocke, C. (2007). Mental action and self-awareness (I). In B. P. McLaughlin & J. Cohen (Eds.), *Contemporary debates in philosophy of mind* (pp. 358–376). Blackwell.
- Perry, J. (1977). Frege on Demonstratives. *The Philosophical Review*, 86(4), 474–497.
- Pitts, A.M. (2002). Operational Semantics and Program Equivalence. In Barthe, G., Dybjer, P., Pinto, L., Saraiva, J. (Eds.), *Applied Semantics*. APPSEM 2000. Lecture Notes in Computer Science, vol 2395. Springer, Berlin, Heidelberg.
- Pierce, B. C. (2002). *Types and programming languages*. MIT Press.
- Quine, W. V. O. (1951). Two dogmas of empiricism. *The Philosophical Review*, 60(1), 20–43.
- Rumfitt, I. (2011). Inference, deduction, and logic. In J. Bengson & M. Moffett (Eds.), *Knowing how: Essays on knowledge, mind, and action* (pp. 334–350). Oxford University Press.
- Ryle, G. (1946). Knowing how and knowing that: The presidential address. *Proceedings of the Aristotelian Society*, 46, 1–16. <http://www.jstor.org/stable/4544405>
- Ryle, G. (1965). *The concept of mind* (6th printing). Barnes & Noble. (Original work published 1949)
- Stanley, J., & Williamson, T. (2001). Knowing how. *Journal of Philosophy*, 98(8), 411–444.
- Stanley, J. (2011). *Know how*. Oxford University Press.
- Steingartner, W., Novitzka, V., & Schreiner, W. (2020). Coalgebraic Operational Semantics for an Imperative Language. *Computing and Informatics*, 38(5), 1181–1209.

Chapter 2 MODE OF PRESENTATION VS. MODE OF USE IN PRACTICAL REPRESENTATION

Chapter 2 1 Introduction

Consider an experienced ballet dancer executing an arabesque, or a surgeon tying a suture. Performances like these exhibit a structure that unfolds as a sequence of coordinated movements. For the dancer, the sequence may involve lengthening the spine, rotating the hip, and extending one leg while keeping the other grounded. For the surgeon, it may involve positioning the needle, driving it through tissue, drawing the thread to the appropriate tension, and so forth. All of these are intentional actions themselves. Yet the dancer and the surgeon typically do not attend to all these details in the course of their performance, just as we carry out various actions every day without thinking about each movement involved. At the same time, agents can, to varying degrees and at different levels of skill, prepare for, attend to, and intervene in such fine-grained aspects of performance that would otherwise go unnoticed. How should we understand the representational states that make such guidance possible? Call such action-guiding representations *practical representation*. The question at hand is: what is the nature of practical representations?

Contemporary discussion of practical representation has developed along two lines. One line emerges within the context of intellectualism, the view that knowledge of how to do things is propositional knowledge. In this setting, practical representation appears in the form of *practical modes of presentation* (“PMOP” or “practical MOP”), originally introduced by Stanley and Williamson (2001) in defense of intellectualism. Stanley and William’s proposal reignited debate over intellectualism about knowledge-

how and prompted further elaboration and criticism of the notion of *practical* MOPs (Fantl, 2008; Glick, 2015; Koethe, 2002; Mosdell, 2018; Noë, 2005; Pavese, 2015; Rosefeldt, 2004; Schiffer, 2002; Stanley, 2011, 2012). Because these views are articulated through debates over intellectualism, however, they are often entangled with commitments that obscure their significant explanatory and critical contributions to the notion of distinctively practical representation.

Outside the knowledge-how debate, cognitive scientists and philosophers of action have treated practical representation as a distinctive explanatory family of states, usually through the framework of motor control and sensorimotor representation (Buehler, 2025; Fridland, 2017; Mylopoulos & Pacherie, 2017; Mylopoulos et al., 2026; Rescorla, 2016). In this literature, the main questions concern the representational role, content, and format of the states that mediate between intention and execution, or that guide performance within motor planning and control. This line of inquiry has yielded increasingly detailed proposals about the format and structure of such states at subintentional and subpersonal levels (Buehler, 2025; Mylopoulos et al., 2026). Nevertheless, the question of whether practical representation constitutes a distinct representational kind in its own right has not been squarely addressed at the appropriate level of generality, much less what would have to be shown to establish it as such.

In this paper, I undertake to answer that question. I start by isolating the central explanatory challenge that drives much of the critical and clarificatory work on action-guiding representational states, where the notion of a distinctively practical type of representation comes into play. The challenge is to settle the question whether the

notion is grounded in a distinctively practical kind of *mode of presentation*, or that “practical” simply marks the action-guiding *mode of use* or deployment of ordinary representations. To establish PMOPs as a distinctive representational kind, one must settle this question in favor of the former over the latter, thereby defending the *mode of presentation* view of practical representation over the *mode of use* view. That is why an account of PMOPs has to explain why the practical element belongs to the way an action outcome is represented and not to the use of an ordinary representation in executing that action.

Representational kinds are not individuated in only one way. A putative representational kind may be marked by its content, its format or structure, or by the ways it is acquired and processed (Beck, 2015, p. 2; Martínez-Manrique, 2010, p. 60; Weiskopf, 2009, p. 151). Indeed, outside the PMOP literature, practical representation is often explored through questions about *motor representational format* (Buehler, 2025; Mylopoulos et al., 2026). It should be noted that accounts of format in practical representation more generally do not plausibly fall within either the mode of presentation view of practical representation or the mode of use view simply in virtue of being accounts of representational format. Such accounts aim is to characterize the structure of motor representations, which by itself does not suffice to decide whether the practical element belongs to representational presentation or to action-guiding use. For representational format could either mark a difference in representational presentation of an objective entity, or a difference in how the same representation can be encoded and used.

Nevertheless, accounts of motor representational format are relevant here because they indicate where a representational difference might be located, if there is one. The mode of presentation view of practical representation must be able to identify some feature of the representation itself in virtue of which it presents an object, task, or outcome practically. A motor representational format that is *distinctive to motor representation* is a candidate for such a feature. While a reasonable proposal of a distinctively motoric representational format would not by itself establish the mode of presentation view or rule out the mode of use view, it could give the mode of presentation view a supported feature to develop. By contrast, an otherwise reasonable proposal of motor representational format in which the format is also available in, say, perceptual representation, makes the mode of use view harder to rule out. A successful defense of practical modes of presentation must show either that such proposal does not in fact favor the mode of use view or that any support it provides does not undermine the case for practical modes of presentation. For this reason, I will be examining Buehler's (2025) account as it is such a proposal. Buehler argues that certain motor representations governing eye movements exhibit a *topographic* structure, which may also organize visual perceptual states. His account therefore presents a case where the proposed format of a subset of practical representations is not distinctively practical but also found in perceptual representation.

At first glance, Buehler's account might seem to favor a mode of use view of practical representation. The same general representational format appears in both perceptual and motor contexts; in turn, it may seem that the supposed practical element lies in how the representation is *used* by a motor system, rather than in the way the

represented outcome is *presented*. As I will argue, however, this case involves a structural element that is not captured by the topographic format-type while being central to the representation's distinctive, command-like mode. More specifically, I show that the relevant representational format consists in not only a map-like topographic field but also the structure of the represented eye movement as a *directed transition* from one position to another. This transition-structure cannot be captured by a solely perceptual representational state; it belongs to the organization of the action-guiding representation itself. The objection I develop is therefore that the topographic-format proposal underdescribes the motor representational structure of the relevant eye movement, which further motivates the search for a distinctively practical mode of presentation.

My positive account to establish practical representation as a distinctive representational kind is as follows. I adopt Pavese's (2015) view, where practical MOPs of tasks or action outcomes are modeled after operational semantic values of computer programs. In her recent work that builds on her operational-semantic approach, Pavese (2026) proposes that systems possess directive capacities to follow through on commands: representing an action outcome as a function of these capacities constitutes representing it under a practical MOP (p. 209). I agree with Pavese that practical MOPs are capacity-relative, but this account is incomplete without an explanation of which capacities play the relevant constitutive role. I supply that account by drawing a parallel with *perceptual* modes of presentation: just as perceptual MOPs are tied to *perceptual constancies*, practical MOPs are tied to *practical constancies*. Motor equivalences are empirically established capacities to represent one and the

same task or action outcome across significant variations in bodily and environmental implementation. I argue that they showcase practical constancies which constitute capacities to represent a task as a constant operational semantic value across varied practical representations.

Chapter 2 2 Mode of presentation vs. mode of use

The notion of *practical modes of presentation* (“PMOP” or “practical MOP”) purports to pick out modes of presentation that are distinctively practical, thereby marking practical representation as a distinctive representational kind. In this section, I argue that the central explanatory burden in theorizing about practical representation as a distinctive representational kind is to show that “practical” lies in the representational presentation of tasks (or action outcomes) rather than the mere use of otherwise ordinary representations in task execution.

Stanley and Williamson originally introduce PMOP to defend *intellectualism about knowledge-how*, which is the view that knowledge-how is a species of propositional knowledge. A standard objection intellectualism faces is that one can learn any number of propositions about how to do something without actually learning how to do it themselves. A subject watching someone riding a bicycle may acquire the *knowledge that* the demonstrated way is a way for her to ride a bicycle without acquiring the *knowledge how* to do it (Stanley & Williamson, 2001, p. 428). Stanley and Williamson propose that the relevant knowledge-how ascription requires the subject to stand in the knowledge-that relation to a Russellian proposition: the proposition contains *a way of riding a bicycle*, and the same proposition may be entertained under different modes of presentation (p. 427). More specifically, they argue that the same way of

doing something may be represented under different modes. In the example above, the learner represents how to ride a bicycle under a demonstrative mode of presentation, thus failing to learn how to do it herself. To acquire knowledge how to do something, one must represent the relevant way of acting under a practical mode of presentation, which is bound up with “certain complex dispositions” connected to action (p. 429).

Criticisms of Stanley and Williamson’s proposal center on the issue of PMOPs as a distinctive representational kind, more so than issues proprietary to the intellectualism/anti-intellectualism debate. Koethe (2002) claims that Stanley and Williamson do not explain what it is to represent a way practically, as opposed to merely knowing that it is a way (p. 325). Likewise, Rosefeldt (2004) argues that PMOPs must admit characterization independent of the abilities they aim to explain (p. 378). If a PMOP is characterized by reference to the very dispositions or abilities it is meant to explain, then the account becomes circular. Schiffer (2002) claims that PMOPs cannot be understood just by appealing to a practical relation between a subject and a process in their own body (p. 200). If PMOPs simply require having a practical relation to a process that occurs in one’s body, then every person would know, under a practical mode of presentation, how stem cells divide, which is implausible. Schiffer therefore concludes that the theory fails to sufficiently specify what PMOPs are supposed to be (Schiffer, 2002, p. 201). Noë (2005) raises a circularity objection, arguing that Stanley and Williamson take for granted that the relevant knowledge-how attributions concern same proposition then use PMOPs to explain how they differ in mode of presentation (p. 288). Since the notion of a practical mode is introduced only to make that assumption plausible, he deems the proposal circular. Glick (2015) later presents a more robust

formulation of this problem. He argues that PMOPs are introduced as the extra condition needed to distinguish merely knowing that a method is a way from actually knowing how, but the stock claims about them still do not tell us what they are (Glick, 2015, p. 547). His “schmode” analogy serves to show that relabeling the difference does not explain it (p. 549).

The unifying concern behind these criticisms is a problem about the practical element itself. Namely, if PMOPs are to explain anything, they must be characterized independently of the action-guiding competence they are invoked to account for. To address this problem, Pavese (2015) offers a more detailed account of PMOPs. She proposes a model drawn from operational semantics in computer science. In operational semantics, a program receives a semantic value that functions as a rule-like description of the system’s behavior in executing that program. Operational semantic values are structured like inferential rules that describe how one configuration transforms into another (Pavese, 2015, p. 7). Configurations are ordered pairs consisting of a command and an assignment of values to variables. Differences in operational semantic values track distinct transitions between configurations, and Pavese argues that these rule-like structures can model PMOPs. Operational semantic inference works through command execution steps, and Pavese claims that PMOPs resemble rules that guide execution (Pavese, 2015, p. 8). She argues that the ability to perform an action therefore involves grasping such rule-like representations.

Read independently of the intellectualism dispute, these proposals and criticisms point to a more general explanatory issue concerning practical representation as a distinctive kind and the idea that there are practical ways of thinking. The proposal side

introduces the idea that tasks or methods may be represented under a distinctively practical mode. The critical side demands an account of what makes that mode practical. Indeed, Pavese (2021) notes that even though the idea of practical concepts has become closely associated with intellectualism, the main reaction to it has been skepticism about whether such a posit is intelligible and non ad hoc at all (p. 7662). More than an anti-intellectualist claim, such skepticism fundamentally involves a demand for a theory of practical representation. The relevant question is whether we can think of a task in a distinctively practical way, not merely semantically or observationally (p. 7660). In an earlier paper, Pavese (2019) answers the same question in representational terms: namely, she argues that procedural representations are distinct from both purely conceptual and purely perceptual representations, and their modes of presentation are distinctively practical (p. 785).

A representational kind counts as distinctive qua representational only if the difference belongs to the representation itself. If the difference lies merely in what a subject does with an otherwise ordinary representation, then one has not identified a new representational kind at all. At most, one would have identified a mode of use of an ordinary representational kind, which is not the point of contention in the discourse around PMOPs. A mode of presentation is supposed to constitute how something is represented or entertained, such that its distinctive features are distinctive *representational* features. The practical element is introduced as marking a distinct representational kind, and problematized by critics as such. That is why the critics press the question of what practical modes of presentation themselves are. The explanatory issue concerns whether that added practical element can be shown to describe the

representational *presentation* over representational *use*: we want to know if the action-guiding state involves a practical presentation of a task or some action-guiding deployment of an otherwise ordinary representational content.

The same mode of presentation can be available for different uses. It can be used in theoretical or epistemic activity, like judgment and inquiry. It can also be used in practical activity, like deliberation and action guidance. The same way of representing the glass as containing water seems equally able to figure in purely epistemic activity, such as believing or checking, and in actions such as reaching or drinking. Since it is the same way of representing, it will have the same correctness conditions across these cases. And since representations are characterized by modes of presentation and their correctness conditions, what makes a representational state “theoretical” or “epistemic” seems not so much about how an objective entity is *presented* to the mind, but how that mode of presentation is *used*. If the same goes for action-guiding representational states, then representations involved in such states are not themselves “practical”. Rather, they are simply used for a practical purpose, just like a perceptual representation (e.g., of a glass of water) can be deployed for a practical use (e.g., to guide the action of reaching). This would mean that “practical” only marks a difference in *mode of use* rather than a difference in *mode of presentation*. “Practical representation” would be a misnomer as a representational kind, since the representations would only comprise ordinary modes of presentation put to practical use. Any account that posits practical representations thus incurs a basic explanatory burden: it has to say what, exactly, is supposed to be distinctively practical about the

mode of presentation. This is the central explanatory challenge of practical representation:

The Practical Presentation Challenge: the challenge of accounting for what makes a representation distinctively practical *as a way of presenting a task*.

This challenge turns on the following question:

Practical Mode vs. Practical Use (“MODE vs. USE”): the question of whether practical representation involves a distinctively practical mode of presentation, such that the practical element belongs to the way a task is presented (“the MODE view”), or whether it instead involves the use of a non-practical mode of presentation in action, such that the practical element belongs to the action-guiding deployment of an otherwise ordinary representation (“the USE view”).

One cannot rely solely on the fact that some representational states guide action. The central explanatory challenge of practical representation as a distinctive representational kind involves demonstrating that there is a representational presentation of tasks that is practical, rather than a capacity to use ordinary representational contents in task execution.

Indeed, Pavese’s (2015) account of practical senses approaches the MODE vs. USE question from the MODE view. She does this by modeling practical MOPs after operational semantic values. Different program texts may execute the same task while differing in their operational semantic values because they determine different rule-governed transitions between configurations (Pavese, 2015, p. 7). Practical MOPs can thus be explicated as a kind of execution-guiding transition structure. In contrast, Waights Hickman (2019) defends an account that appears to side with the USE view on

the MODE vs. USE question. She argues that what is practical in knowledge-how need not belong to a special mode of presentation at all. It may belong instead to a practical way of knowing rules, methods, or procedures, namely knowing them “in the executive way” of being able to apply them (p. 1). She claims that the work attributed to practical modes of presentation might instead be done by a practical knowledge relation (p. 22). This is the explicit rival position generated from within the PMOP debate. The critical sentiment underlying this account, as well as much of the criticism of PMOP, is that the supposed practical element could well belong to the application or exercise of a representation rather than to the representation itself.

The PMOP discourse exhibits these two views if we consider the proposals and critiques separately from the intellectualism debate. The motivating thought is that a task may be represented in a distinctively practical way. The central criticism is that the proposals do not sufficiently characterize the practical element as a representational feature in its own right and that they merely invoke the notion expecting it to do explanatory work.

Chapter 2 3 Examining representational format in practical representation

Recent discussions on motor representation have focused on the format or structure of action-guiding representations (Buehler, 2025; Mylopoulos et al., 2026). Yet the taxonomic significance of such claims depends on what the relevant difference in format amounts to. Not every appeal to “practical,” “actional,” “action-oriented,” or “action-guiding” representation commits one to the MODE view. If such terms are used to mean a mental state that guides action by drawing on perceptual or conceptual MOPs, then one may be implicitly operating within the USE view. What is practical

(actional, action-guiding, etc.) would then be what the state does, rather than the representational kind involved. The MODE view is at issue only when one takes practical representation to mark a distinct representational kind.

The question for our purposes is whether accounts of motoric format can also illuminate what would make practical representation a distinct representational kind. Such accounts do not themselves constitute positions in the MODE vs. USE dispute. Their aim is to characterize the format of motor representations, which does not by itself determine whether the relevant practical element belongs to the representational presentation or to action-guiding use. A distinctively motoric representational format may provide support for the MODE view if it captures a difference in how something is represented, and thus in the presentation-structure of the representation itself. However, the defender of the USE view could still argue that such format is merely the motor system's proprietary way of encoding, processing, or deploying representational content in the service of action. Thus, while an account of distinctively motoric format would show that motor representation differs *structurally* from perceptual or cognitive representation, but that would not suffice to show that this difference is a difference in practical presentation rather than a difference in motor-system implementation.

By contrast, if the purported format found in motor representation is also available to different types of representation, then the format marks the encoding of what could otherwise be a perceptual or conceptual presentation for action-guiding processing and control. This is, in fact, how both Buehler (2025) and Mylopoulos et al. (2026) account for motor format: their developed proposals identify motor representations with configurations of structural features that are also available outside

the motor system (Buehler, 2025, p. 24; Mylopoulos et al., 2026, p. 472). Mylopoulos et al.'s (2026) proposal is that motor representation has a *hybrid* format, in which *analog* parameters are embedded within a *symbolic* scheme (p. 470). They discuss empirical evidence from motor imagery, grasping, and temporal synchronization that supports *analogicity* (pp. 462-463), together with evidence from motor planning, motor control, and motor learning for *symbolicity* (pp. 464-469). The evidence for both formats suggests that while motor plans exhibit core markers of symbolic format (which include compositional structure, discrete constituents, abstract contents, and possibly role-filler independence) (p. 469), certain motor parameters are encoded at least partially analogically by exploiting structural properties of their representational vehicles (p. 470). Mylopoulos et al. propose that such analog parameters are “embedded within a broader symbolic scheme” of the motor plan. Motor representation is thus described as “a symbolic scheme that utilizes analog representations as cognitive tools” for motor planning (p. 472).

Mylopoulos et al. suggest that sensory and motor representations may employ the same analog format (p. 462) and that the embedding of analog constituents within symbolic structures can be found in numerical cognition and visual contour representation (p. 473). Neither the symbolic scheme nor the analog resources it incorporates are characterized in this account as distinctive to motor representation. While their account identifies a distinctive way in which representational resources are combined and deployed in motor planning, Mylopoulos et al. leave open the stronger possibility that motor representations constitute “a novel type of format” (Mylopoulos et al., 2026, p. 469).

Buehler (2025) takes a more case-specific approach by examining one particular type of motor representation: the commands that control rapid eye movements called saccades. Saccade-commands are issued by the oculomotor system and specify shifts of foveal fixation from one location in the visual field to another. Buehler's goal is to show that these representations have a topographic, map-like format. Like Mylopoulos et al., Buehler does not consider the structural resources identified by his account to be *distinctively motoric*. However, Buehler's proposal has a stronger implication. Mylopoulos et al. characterize motor representation as hybrid, where analog parameters are embedded within a symbolic representational scheme. Because their proposal integrates features associated with two recognized format types, it remains possible that this hybrid organization constitutes a novel format category. As mentioned above, Mylopoulos et al. leave open the possibility that such distinctive integration may constitute a novel motoric format with some features previously associated with symbolic formats and others with analog formats. Buehler, on the other hand, characterizes the representational structure of the saccade-command by reference to a single format-type: topographic structure. His developed proposal thus does not contain a hybrid level like what Mylopoulos et al. propose, where the combination of different format-types might constitute a distinctively motoric format. Because the topographic structure may also organize visual perceptual representations, Buehler assigns the relevant representational distinctiveness to what he specifies as the saccade-command's *mode* and its *non-structural* constituent-representations.

Saccade-commands participate in motor planning, initiate eye movements, and guide those movements toward completion. Buehler analyzes a saccade-command as a

mental state with a mode and a content. Its mode is “that of being a saccade-command”: namely, the state directs the oculomotor system to produce a saccade and sets the movement in motion (2025, p. 3). Its content specifies a motor plan in the form of a *displacement-vector*, that is, a *shift of fixation* through a determinate direction and distance. In virtue of the state’s command-mode, the gaze shift is represented as “to-be-performed”. Buehler characterizes this mode as “intention-like” (p. 21). Accordingly, Buehler argues that saccade-commands have world-to-mind satisfaction conditions: a saccade-command’s content is satisfied when it brings about the saccade as specified (p. 22).

My concern here is whether the topographic format exhausts the structure of a saccade-command’s representation of a gaze shift across the visual field. I argue that it does not: the complete command crucially involves a further form of organization. The following examination reveals the structural contribution involved in presenting the displacement vector, where map-positions are represented as a transition from the start point of the gaze shift to the end point. Establishing this additional transition-structure will provide the starting point for my subsequent account defending the MODE view and distinctively practical modes of presentation.

Buehler takes format as *representational structure* which is the way constituent representations combine to form complex representational contents. The content of a saccade-command has a “topographic representational structure” and the marker “shift”, which is a “primitive representational constituent” of this structure (p. 4). The marker shift contributes the representation of a shift in foveal fixation. The topographic structure supplies a field of positions whose geometric relations correspond to

geometric relations in visual space. The marker is depicted as an arrow extending from map-position A to map-position D (Buehler, 2025, p. 3, fig. 1). Position A corresponds to the current fixation-point, position D corresponds to the target fixation-point, and the geometric relation between them determines the direction and distance of the represented shift. The marker's insertion into those positions therefore yields a complex content representing a displacement of fixation from A to D (p. 4).

The command-content thus includes a representation of shifting fixation together with the spatial path of that shift. The marker supplies the shifting component, while the topographic structure encodes the shift's spatial parameters. Buehler associates the marker with the organism's basic practical ability to shift fixation from one location to another (Buehler, 2025, p. 23). As the marker enters the topographic structure, the resulting content specifies a fully articulated fixation-shift. Buehler maintains that computations that generate saccade-commands and their neural implementation lend support to the topographic structure view. Operations such as merging and averaging depend on spatial relations within the neural map: aligned neural positions preserve information about corresponding locations, while patterns of activity determine centers of gravity among clusters of priority. The neural organization exhibits the functional profile of topographic representational structure, which Buehler takes to support the attribution of topographic structure to the representational states implemented there (p. 21).

Crucially, Buehler describes the marker shift as part of the command-content while distinguishing it from the topographic representational structure. He calls the marker a "non-structural constituent-representation" (Buehler, 2025, p. 24): the marker

is non-structural relative to the map-like scheme of combination. At the same time, Buehler maintains that the marker has its own combinatorial structure consisting in “how it combines to form well-formed contents” with that scheme (p. 23). Buehler’s classification therefore places the marker’s command-specific contribution within content while excluding it from the representational structure proper. He maintains that saccade-commands and perceptual states may share the same topographic structure while differing in their non-structural constituent-representations. This seems to imply an additional source of difference within the content itself, such that the distinctiveness of saccade-commands consists not only in their mode but also the non-structural constituent-representations. Yet Buehler maintains that the distinctive “to-be-performed” status comes from the mode and applies to the *complete* displacement-content, and that the mode likewise does not determine how the marker combines with the map. The marker contributes the representation of shifting fixation, which receives its determinate spatial articulation through the topographic structure; and the mode presents the resulting displacement-content as to be performed. At the same time, Buehler claims that the marker’s meaning is connected to its “functional role in commanding saccades” and to the organism’s practical ability to shift fixation (p. 23, n. 40). Nevertheless, he does not explain how the intention-like mode of the state and the marker shift relate in terms of such motor-functional considerations, identifying the individuation of markers and modes as questions for future research (p. 5, n. 14; p. 25).

The state-mode and the marker shift seem to belong to different levels of Buehler’s analysis. The mode is the mode of the mental state as a whole, which is the saccade-command. Buehler individuates this mode through the state’s intention-like

functional role in motor planning, initiation, and guidance. In virtue of that mode, the displacement specified by the content is represented as to-be-performed, and the state has world-to-mind satisfaction conditions. The marker shift, by contrast, is a constituent of the command's content. It contributes a representation of shifting foveal fixation. When the marker combines with positions in the topographic map, the result is a well-formed content specifying a particular displacement vector with a determinate direction and distance. The marker's own structure consists in this combinatorial profile: namely, the way it can combine with the topographic scheme to form such contents.

Yet this division raises an issue concerning the representational structure of the complete saccade-command. A fixation-shift has an ordered organization that proceeds from a current fixation-position toward a target fixation-position. The geometrical relations within a topographic map determine the direction and distance between two positions, while they do not by themselves assign those positions the roles of source and target within a single transition. An ordering of this sort must be contributed either by the marker shift or by some combination rule through which the marker combines with map-positions.

While the topographic structure supplies the geometric field within which a displacement receives its spatial parameters, the marker shift supplies the representation of a directed transition from one fixation-position to another. Since the marker belongs constitutively to the command-content, its contribution forms part of the *complete* representational organization of the saccade-command. I argue that Buehler's topographic characterization underdescribes the format of the complete saccade-command. A representation of a shift must distinguish its source from its target. Bare

topographic relations among A and D do not determine whether the represented shift proceeds from A to D or from D to A. The source-target ordering must therefore be supplied either by the marker itself or by the rule through which the marker combines with the map. If it is supplied by the marker, then the marker is internally or relationally structured in a way that the description “non-structural constituent-representation” obscures. If it is supplied by a further combination rule, then the complete representational format includes more than topographic organization. Either way, topographic structure does not capture the structure of the complete saccade-command.

To clarify the additional organization contributed by the marker shift, consider three maps. The first is MAP: a bare map that represents a spatial layout of a topographic structure: its map-positions stand in relations that mirror relations of distance, direction, and adjacency within the represented space. MAP makes the relative locations of places available within a single spatial configuration. A user who uses MAP to navigate through that space must figure out where a chosen destination stands relative to them, and determine which locations should be traversed, and in what order. While MAP’s topographic format provides the spatial relations from which a route may be calculated, it does not itself provide orientation, and does not specify a route. Next, suppose we add a *north arrow* to MAP. Call this MAP+COMPASS. The north arrow establishes an orientation for interpreting the map and thereby helps the user relate directions on the map to directions in the represented environment. Its contribution does not select or specify a path through the represented space. The user must still determine a destination, identify a route, and order the relevant locations for

traversal. MAP+COMPASS therefore contains an additional representational element while retaining the same basic topographic organization of locations.

The final map involves a further modification: it highlights a path and marks its direction with directive arrows. Call this MAP+PATH. Here MAP+PATH would be akin to a route-guidance system: the highlighted path presents a route for the user to follow. The complete path may comprise a sequence of steps, but consider a single highlighted step that extends from location A to location B. MAP represents A and B, together with their distance and an unoriented direction relative to one another. MAP+COMPASS additionally situates that relation within a global orientation. Neither representation assigns A the role of the point from which the user should proceed and B the role of the next point to be reached. In MAP+PATH, on the other hand, the tail of the arrow is located at A and its head at B. Under the relevant interpretive rule, this arrow represents an ordered transition from A to B. Reversing the arrow would represent a different transition even though the underlying locations and their topographic relations remained unchanged. MAP+PATH therefore retains the spatial organization represented by MAP while adding a *rule* that projects directional organization in the sign onto transition-order in the content. The topographic relation between A and B still determines the spatial character of the step, including its direction and distance. The arrow assigns the two locations distinct roles within that step: A is represented as the position from which the transition begins, while B is represented as the position toward which it proceeds. Across the complete highlighted path, the same rule orders successive locations as stages of a route. While the action-guiding arrow in MAP+PATH is etched into the

representational organization, its representational format cannot be reduced to the topographic representation.

In MAP, relations among map-positions represent spatial relations among locations. MAP+PATH changes how the represented layout is structured for a subject by making an ordered route available without requiring the subject first to derive that route from the topographic layout alone. A user can now move from represented layout to route-following guidance without first supplying the relevant direction from outside the representation. The represented space is now given under an ordered transition-structure that already privileges one way through the space. MAP+PATH thus adds a rule under which the asymmetry between an arrow's tail and head represents an ordered transition from a source-position to a target-position. An arrow from A to B and a reversed arrow from B to A therefore represent different transitions while preserving the same underlying topographic layout. MAP+PATH consequently contains a systematic transition-structure in addition to its topographic organization.

As Camp (2007) points out, cartographic systems can be cognitively efficient by making spatial relations "automatically available" (p. 161). However, when the map is extended with further elements that employ abstract interpretive principles, the map becomes a more articulated representational system (p. 165). MAP+PATH changes how the represented layout is structured for the subject because its sign structure is governed by an additional semantic rule: the ordering and directionality of the highlighted segments represent the ordering and directionality of a route through the space. The representation therefore makes available a transition from spatial layout to route-following guidance without requiring the subject first to derive that route from the

layout alone. The represented space is presented together with an ordered transition-structure that privileges one way of traversing it. Camp's account thus explains the cognitive consequence of encoding the transitional relation within the representation: the ordered route becomes available through competent interpretation of the map rather than through a separate calculation from the spatial layout alone.

Buehler's marker shift is better captured within MAP+PATH than both MAP and MAP+COMPASS. The marker shift is a constituent that, when entered into the map, yields a well-formed, actional representation of a fixation-shift from one location to another (Buehler, 2025, p. 24). Insofar as the marker belongs constitutively to the representation, the format can no longer be adequately described as merely topographic. This is because the representation contains a transition-structure poised toward execution that is not specified by a spatial layout alone. If the marker shift does the work it must do, the complete saccade-command representation cannot be captured by the topographic structure alone.

The marker shift should not be considered a non-structural constituent. A gaze shift is intrinsically *directed*: it proceeds from one location to another. Merely associating the marker with two positions such as A and D in Buehler's Fig. 1 would not distinguish a shift from A to D from one from D to A. Consequently, the displacement would not be adequately distinguished from a static connection or a relation of proximity between those positions. Some element of the representation must therefore assign different roles to the two map positions, where one is treated as the origin and the other as the destination. In Buehler's diagram, the arrowhead indeed yields this treatment. Yet that

arrow seems merely illustrative; that is why an equivalent source-target asymmetry must be encoded somewhere in the representation.

If the marker shift itself encodes the distinction between origin and destination, then it possesses an internal, role-structured format. Its contribution must then specify an ordered relation whose argument places are filled by positions in the map. In that case, describing it simply as a “non-structural constituent-representation” obscures a structurally significant feature of the representation. If, on the other hand, the origin-destination pair is supplied by the larger representational scheme, then the scheme structure involves more than topographic relations. More specifically, the representational structure must further contain a rule for assigning ordered roles to map-positions and combining them into a representation of directed movement. The format of the complete content would consequently involve an additional operational structure. The complete representation is not merely a topographic layout of positions, nor merely a representation of their spatial relations. Rather, it represents one position as the source of a fixation-shift and another as its target. This sort of functional/role-differentiation cannot be supplied by topographic structure alone. It belongs to the organization through which the represented outcome is given as a transition from one position to another. The marker shift, or the rule by which it combines with the topographic map, therefore contributes a structural feature of the complete command representation.

A defender of the USE view might resist this conclusion by claiming that the source-target organization need not be treated as a feature of representational structure. Namely, the representation is topographically structured, which makes

available positions, distances, and directions within a map-like field. The oculomotor system then *uses* that topographically structured representation in a command-like way. This objection treats the source-target organization as simply the positions from which the system begins and toward which the system moves the eyes: the ordering from source to target would therefore belong to the motor system's deployment of the representation, rather than the content of the representation itself.

However, such a treatment would only work if the relevant content could be fully specified before the source-target ordering enters. Buehler's own model of the content of saccade-commands does not meet that condition since the relevant content is not merely a topographic registration of two positions but a displacement-vector, representing a shift of fixation from one location to another. The marker shift is introduced as a representational constituent which, when entered into map-positions, yields a well-formed content representing a displacement. A representation of this kind would not be captured by a representation of two positions and their topographic relation. The distinction between A-to-D and D-to-A cannot be reduced to how the same content is used since it is a distinction of which displacement-content has been formed.

If source-target ordering is excluded from the representational content of the saccade-command, then the state would not have the content of the displacement-vector Buehler attributes to it. The source-target ordering is the command-specific transition-structure that contributes to the formation of the saccade-command's representational content, and it is not captured by topographic structure. The result is that Buehler's proposed format type may fail to track the representational structure that is central to the representational state's command-like mode. The command-specific

element helps determine how the outcome is represented in the first place, namely as a transition to be carried out.

Chapter 2 4 Pavese's capacity-relative account of practical modes of presentation

Buehler's account classifies saccadic commands as topographically structured, which is a structure also available in visual perception. Yet the command-like character of these states turns on a source-target organization through which the represented outcome is given as a transition. This result motivates a clear account of the MODE view since the practical structure of such states therefore remains to be accounted for. The next task is to characterize the transition-structure of practical modes of presentation, through which an outcome is represented as executable, not merely as a spatial item, relation, or content available for action-guiding deployment.

Pavese's account is the strongest existing candidate to meet this demand as she provides the most well-developed account of practical representation. Her operational-semantic model treats PMOPs as execution-guiding structures, and her later work explains practical representation through system-relative capacities to follow through on commands. She accounts for the structural feature that I argued in the previous section cannot be captured by the topographic format: namely, the representation of an outcome both organized and supplied through system-relative executable parts and transitions. In this section, I examine Pavese's account and argue that her appeal to practical capacities does not suffice to identify the representational capacities that constitute a practical mode of presentation.

Pavese (2015) argues that computer programs, understood as the operational semantic values of program texts, provide a model for practical ways of thinking (pp. 5-

8). The appeal of this model is that operational semantic values specify the behavior of computers “in interpreting and compiling program texts” (Pavese, 2015, p. 5). Compiling a program text is thus analogous to understanding how to carry out an instruction: computers acquire an ability to follow the text’s operational semantic value. Pavese’s (2026) account of practical modes of presentation builds on the operational-semantic approach while focusing on the role of a system’s operational capacities in producing practical modes of presentation. Her starting point is that “[m]ental representation is perspectival”: representing anything is representing it “in some way”, and this way constitutes a “perspective” or mode of presentation (p. 205). Conceptual and perceptual modes of presentation differ according to the perspectival ways constituted by the relevant representational functions and capacities (p. 206). In the conceptual case, the relevant function is to represent things in a way “conducive to thinking and reasoning about them” (p. 207) and the relevant capacities are basic conceptual capacities for genuine thought (p. 208). In the perceptual case, the relevant function is “tracking features of the environment”, while the relevant capacities are perceptual capacities to register changes in the environment in a systematic fashion (p. 207).

This functional and capacity-based approach to modes of presentation then serves “to introduce the theoretical possibility of practical representation” (Pavese, 2026, p. 208). Just as conceptual and perceptual representations present things in ways conducive, respectively, to thinking and reasoning about them and to tracking them, practically representing an action outcome is to represent it in a way “conducive to performing it”. Pavese proposes that this function is served by a *practical* perspective constituted by capacities that are *directive* capacities to follow through on commands

(pp. 208-209). Directive capacities have what Pavese calls a “directive” direction of fit: unlike perceptual capacities, which track how the world is, directive capacities bring about an action outcome in the world that conforms to the represented command (p. 209). Qua system-relative operational capacities, they are more basic than agential abilities, habits, or skills. They are at the level of subpersonal motor systems and even systems such as thermostats may be said to possess directive capacities. Practical representation can therefore occur subpersonally and nonconceptually as well as in conscious practical thought (p. 209; p. 233). For example, the motor system’s representation of how to adjust one’s hand while catching a falling glass would be practical without any personal-level thought involved.

Pavese gives practical representation a two-part structure. A representation is practical only if it is prescriptive in that it is explicitly imperative or used as an instruction, and it decomposes the task into units that are executable by that system (pp. 211-212). Her formulation is that a practical representation “breaks down the task in terms of parts that the system can execute and in terms of transitions that the system can execute”. Prescription alone is insufficient since an instruction such as “Tie the surgical knot” presents something to be done, but is not yet practical in the appropriate sense for a surgeon in training who cannot execute the presupposed steps. A practical representation would be produced by relevant directive capacities corresponding to the more basic tasks of how to position the needle holder, loop the thread, reverse the grip, and tighten the knot.

The task decomposition should be *specific*, *compact*, and *efficient*, and the practicality of a representation could be defective when these are compromised

(Pavese, 2026, p. 211). Specificity requires enough detail to guide performance; compactness requires termination at operations already basic for the system; efficiency concerns avoiding unnecessary steps. Pavese compares practical representations to programs whose terminal nodes prescribe primitive operations and whose higher structure organizes them into executable sequences or trees (p. 213). For example, an experienced driver may represent the task of parallel parking as one unified routine, whereas a beginner may need mirror checks, steering angles, and corrective movements. How each driver represents the parking task turns on what relevant system-relative operations each can execute and how those operations end up at the stopping points of task decomposition. Because every component and transition in a genuinely practical representation must be executable, practical representation is marked by the corresponding capacity. Pavese claims that if a system represents a task practically, “it cannot fail to have the capacity to perform it” (p. 212). Executability is built into what qualifies a representation as practical.

Practical modes are system-relative and time-relative because basic operations vary among systems and can change through learning (Pavese, 2026, p. 209; p. 212). The task of performing an arabesque may have become basic for an expert dancer while highly complex to a beginner, who must represent the movement through instructions concerning weight placement, spinal extension, turnout, and balance. The two dancers’ respective practical representations of the same action outcome therefore differ because their directive capacities differ, which in turn differentiates the executable manner in which the task is presented to each dancer. Training changes the practical representation of a task through what Pavese calls “chunking,” where separate

sequence elements become grouped into one elementary operation, reducing memory demands and increasing speed (p. 214). Pavese maintains that chunking creates new basic directive capacities and thereby changes the representation itself, since the task is now represented through different elementary operations.

Pavese provides the most developed version of the MODE view. She explains practical modes of presentation by appeal to system-relative capacities for execution: a task is represented practically when it is represented through parts and transitions that the relevant system can carry out. Her account avoids treating the practical element as a mere use of an ordinary representation and explains why different agents, or the same agent at different stages of training, may represent the same task under different practical modes.

However, Pavese's account leaves the MOP-constituting role of practical capacities underspecified in the sense relevant to the Practical Presentation Challenge. If practical modes are constituted by capacities, what kind of capacities play the relevant MOP-constituting role? An appeal to capacities must explain why the practical element belongs to representational presentation rather than to successful action-guiding use. It cannot be enough, for that purpose, to say that the relevant capacities are directive capacities. Such characterization captures merely that practical representations are connected to systems that can follow through on commands. It does not illuminate the capacities marked by a practical way of presenting tasks. The class of directive capacities is too broad: it includes subpersonal motor systems, skilled agents, and simple regulatory systems such as thermostats. What is needed is an account of the capacities through which a system reliably represents a task as executable across

decomposition, recombination, adjustment, and variation in the conditions of performance.

Chapter 2 5 Practical constancies

Pavese is right that practical representation is relative to a system's capacities: how a task is practically represented depends on what a system can execute as basic, what must be decomposed into further steps, and how those steps can be organized into transitions. Yet this account does not explain practical presentation. More specifically, the relevant capacities must account for a form of sameness across variation. A task can be carried out through different bodily movements, effectors, force profiles, temporal patterns, and environmental adjustments while still being represented as the same task. Conversely, small changes in movement may amount to a different task when they alter the operational organization of the outcome. An account of PMOPs therefore needs a way to distinguish variation in execution from variation in the practical object represented.

I propose to supply this account through the notion of *practical constancy*. Burge's (2022) account of *perceptual constancies* clarifies the relevant form of representational explanation. In perceptual representation, constancy capacities allow a system to represent a *stable* environmental particular or attribute across variation in proximal stimulation. In practical representation, the constancy capacities allow a system to represent a stable task or action outcome across variation in the conditions of execution. Studies on motor equivalences (also called motor constancies) provide clear empirical cases of this phenomenon. They show how an action outcome can remain fixed for a system even when the detailed bodily implementation changes. Such

capacities therefore explain how practical modes of presentation can be both capacity-relative and genuinely representational: they organize heterogeneous executions as presentations of one and the same operationally articulated task.

Burge characterizes “perceptual constancies” as capacities to represent a particular or attribute as the same under significant variation in proximal stimulation (Burge, 2022, p. 53). He treats perceptual constancies as constitutive of perceptual representation itself. A perceptual state counts as genuinely representational only insofar as it belongs to a capacity to represent a given environmental particular or attribute as *that* very particular or attribute across significant differences in proximal stimulation. These constancies form part of a broader process of *objectification* in which a perceptual system separates features of stimulation tied to specific occasions from cues that support representation of a stable environmental attribute (p. 55).

The relevance of this account is that, if practical representation is to mark a distinct representational kind, then it too must exhibit a form of constancy that is constitutive. The practical analogue of perceptual constancy, I suggest, is exhibited in motor control. Namely, a single movement class can be executed through many distinct parameter specifications while maintaining the same functional organization of the action. Schmidt’s (1975) influential schema theory proposes that “generalized motor programs” store an invariant structure for a movement while allowing variation in force, timing, or muscle recruitment across performances (p. 232). Schmidt later maintains that one “memorial structure” can generate a large set of executions for what counts as the same action, provided that the invariant relations among components of the movement remain fixed (2003, p. 367).

Furthermore, studies of motor constancy or motor equivalence show how the nervous system keeps an action outcome constant even when the detailed motor commands change. Mattos et al. (2015) examine this using a task in which participants must produce a fixed total force with several fingers. Because the same total force can be produced through many different combinations of individual finger forces, the task allows researchers to separate two kinds of change: variations that keep the total force constant and variations that change it (Mattos et al., 2015, p. 492). Mattos et al. call the first type motor-equivalent change, because the task outcome remains the same even though the contributions of the individual fingers differ. When the system is perturbed, they report that most of the resulting adjustments occur in this motor-equivalent direction, which means the system reorganizes finger forces in ways that preserve the required total force (p. 498). Similarly, Wing notes that writing on a vertical board versus on paper involves different muscles, torques, and gravitational forces, yet the written word remains recognizably the same, which shows constancy of the produced form despite large differences in the underlying motor implementation (2000, p. R245).

Burge explains perceptual constancy in terms of how different perceptual states can represent one and the same attribute across variation in proximal stimulation. In this respect, motor constancies/equivalences should be understood as instantiating a distinctively practical form of representational constancy: they are capacities to systematically represent a task or action outcome as *that* task breakdown or *that* action outcome across significant variations in bodily movement, effector, trajectory, posture, speed, force-profile, and environmental circumstance. Through these capacities, agents can represent heterogeneous executions as instances of one operationally articulated

task whose structure remains stable across variation in the sensorimotor conditions of performance. This way, practical representations can remain answerable to one and the same operational content.

One might object by saying that motor constancies/equivalences are merely practical capacities that exploit perceptual and proprioceptive content. As such, their existence does not warrant postulating a distinct form of practical content. Such an objection captures part of the phenomenon, since perceptual and proprioceptive contents figure in action-guiding representations, and Burge himself includes proprioception within the domain in which objectification can separate constant bodily subject matter from variation in stimulation (Burge, 2022, p. 55). However, such an objection treats variation-tolerant performance strictly as a feature of control mechanisms that operate over perceptual information about bodily states.

While it is true that perceptual and proprioceptive contents figure in action guidance, there is reason to think of practical constancies as extending beyond motor constancies. Studies suggest that abstract rule structures often develop in interaction with motor organization. Schack (2004) examines the volleyball attack hit to see how the internal structure of skilled movement is represented in memory. He analyzes the hit into basic movement units and asks whether expert players mentally organize those units in a structured way (p. 415). The results show this is the case: experts represent the skill in a hierarchical structure that mirrors the functional phases of the movement itself, whereas less skilled players do not (p. 417). Schack also reports related studies in gymnastics and volleyball that found significant correlations between kinematic parameters and corresponding parts of mental representation structure (Schack, 2004,

p. 428). In a rule-learning experiment done by Collins and Frank (2016), subjects perform a keyboard task in which they press keys according to learned rules (pp. 4-5). In performing the task, the subjects could organize keyboard response configurations. Even though the valid sequence of motor actions stayed the same whichever configuration was used, the subjects chose the configuration that afforded more motor clustering (p. 6). In the subsequent simulations, the authors found stronger results. A pure motor bias in choice did not explain the empirical pattern. But when representational overlap between adjacent finger actions also affected *outcome prediction*, the simulations reproduced the observed preference for the structure that afforded greater motor clustering. The study thus suggests that motor organization influences abstract rule representations (p. 7).

Crucially, certain types of reasoning and critical thinking skills may be considered to be constancies of the cognitive sort. Muller and Haberman (2008) describe abstraction in algorithmic tasks through pattern recognition and structure identification, and they treat success here as the capacity to extract a problem's main task and organize relations among sub-tasks across analogous cases (pp. 189-190; p. 200). Chen offers the same general lesson at a more formal level: analogical solution can proceed through recognition of a conceptual pattern and transfer of its structure into a new case by grammar mapping (1993, p. 1709). These studies therefore lend support to a notion of practical constancy that extends beyond motor constancies: one and the same rule, pattern, or operational structure can guide cognitive skills across different abstract contexts, just as one and the same operational structure can guide bodily action across different physical contexts.

In fact, these capacities (motor/concrete and cognitive/abstract constancies) can be seen as corresponding to two types of operational primitives: *sensorimotor primitive operational representations* and *abstract-conceptual primitive operational representations*. Sensorimotor primitive operational representations guide bodily actions across variable executions. Abstract-conceptual primitive operational representations organize rule application and pattern uptake across variable token cases. In both subclasses, the explanatory identity of the state turns on a stable operation or pattern that recurs through heterogeneous realizations. That is why inferential constancies and pattern-recognition invariances can count as marks of operational objectification when many different token transitions count as instances of a single rule-governed operation or one tracked structural pattern.

In Burge's formulation, perceptual constancies do not involve an internal identity judgment to the effect that one is representing the same thing. Perceptual content is not propositional in that way. Instead, the sameness is coded via a competence that links different perceptual states with different representational contents to a single constant representatum. The relevant states represent the same attribute in different perceptual ways.

Motor constancy describes a capacity to treat heterogeneous movement-patterns as variant practical presentations of one and the same task. The relevant sameness is not the sameness of kinematic profile, nor is it the sameness of external outcome considered apart from the system's capacities: it is sameness of an operationally articulated task as preserved across variation in bodily position, effector, trajectory, timing, force, and environmental circumstance. Just as perceptual constancy

contributes to individuating perceptual states by their relation to constant distal attributes rather than by proximal stimulation alone, practical constancy contributes to individuating practical states by their relation to constant operational structure rather than by kinematic profile alone. What Burge shows for perception is that constancies are what make representational content possible in the first place. His example is that a single edge or length can be represented under different slants and stimulus conditions through different attributives, while the constancy competence links those different attributives to one and the same attribute. Likewise, in the case of practical representation, dancer may execute the same arabesque under changes in tempo, balance conditions, stage orientation, or bodily starting position. A surgeon may execute the same suture under changes in angle of approach, wrist orientation, tissue resistance, or handedness. These are cases in which the system represents one and the same operationalization of a task through different practical presentations. What remains constant is an operationally articulated task-outcome: namely, what is to be achieved, by what sequence of available abilities, under the present conditions of performance. Practical constancy therefore contributes to the individuation of practical representational states. Just as perceptual constancy contributes to individuating perceptual states by their relation to stable environmental particulars or attributes, practical constancy contributes to individuating practical states by their relation to stable task-structures.

Chapter 2 6 Conclusion

I have argued that practical representation should be understood as a distinctive representational kind. The proposal that tasks can be represented under practical

modes of presentation introduces a genuine representational claim, while the major criticisms of PMOPs press for an account of what makes such modes practical in the first place. The central challenge is to explain why the practical element belongs to the way a task or action outcome is represented (the MODE view), over the alternative position that it belongs to the use of an otherwise ordinary representation in action (the USE view).

Buehler's account appears compatible with the USE view: the topographic representational format is not practical in kind, while the practical element seems to enter through the state's command-mode or its role in motor control. Yet the complete saccade command cannot be understood as a topographically structured representation because it contains a source-target organization through which represented outcomes are given as transition schemes. Pavese's operational-semantic model accounts for such structural aspect, and her account constitutes the strongest existing version of the MODE view. She treats practical modes of presentation as execution-guiding structures whose parts and transitions must be executable by the relevant system. Her account explains why different agents, or the same agent at different stages of training, may represent the same task under different practical modes. However, Pavese's appeal to directive capacities does not exactly showcase them as the capacities that constitute practical presentation. A capacity to follow through on commands is too broad unless it is connected to a representational capacity to preserve a task across variation in the conditions of execution.

I have argued that motor constancies/equivalences show how a system can preserve one and the same action outcome across differences in bodily position,

effector, trajectory, timing, force, and environmental circumstance. They instantiate a form of representational constancy: while perceptual constancies allow a system to represent stable environmental subject matter across variation in proximal stimulation, practical constancies allow a system to represent stable task-structures across variation in execution. PMOPs are therefore practical modes of presentation in virtue of the practical capacities that preserve task-structures across variable conditions of execution. The model of practical constancy thus strengthens the case for the MODE view.

Chapter 2 7 References

- Beck, J. (2015). Analogue magnitude representations: A philosophical introduction. *The British Journal for the Philosophy of Science*, 66(4), 829–855.
- Buehler, D. (2025). Topographic motor representations. *Mind*. Advance article, fzaf023.
- Burge, T. (2022). *Perception: First form of mind*. Oxford University Press.
- Camp, E. (2007). Thinking with maps. *Philosophical Perspectives*, 21, 145–182.
- Chen, Z. (1993). Grammar mapping for conceptual patterns. *Pattern Recognition*, 26(11), 1709–1716.
- Collins, A. G. E., & Frank, M. J. (2016). Motor demands constrain cognitive rule structures. *PLoS Computational Biology*, 12(3), e1004785.
- Fantl, J. (2008). Knowing-how and knowing-that. *Philosophy Compass*, 3(3), 451–470.
- Fridland, E. (2017). Skill and motor control: Intelligence all the way down. *Philosophical Studies*, 174(6), 1539–1560.
- Glick, E. (2015). Practical modes of presentation. *Noûs*, 49(3), 538–559.

- Koethe, J. (2002). Stanley and Williamson on knowing how. *The Journal of Philosophy*, 99(6), 325–328.
- Martínez-Manrique, F. (2010). On the distinction between semantic and conceptual representation. *Dialectica*, 64(1), 57–78.
- Mattos, D., Schöner, G., Zatsiorsky, V. M., & Latash, M. L. (2015). Motor equivalence during multi-finger accurate force production. *Experimental Brain Research*, 233, 487–502.
- Mosdell, M. (2018). Modeling practical thinking. *Mind & Language*.
- Muller, O., & Haberman, B. (2008). Supporting abstraction processes in problem solving through pattern-oriented instruction. *Computer Science Education*, 18(3), 187–212.
- Mylopoulos, M., & Pacherie, E. (2017). Intentions and motor representations: The interface challenge. *Review of Philosophy and Psychology*, 8(2), 317–336.
- Mylopoulos, M., Pacherie, E., & Shepherd, J. (2026). Motoric representational format. *Philosophical Studies*, 183(2), 455–479.
- Noë, A. (2005). Against intellectualism. *Analysis*, 65(4), 278–290.
- Pavese, C. (2015). Practical senses. *Philosophers' Imprint*, 15(29), 1–25.
- Pavese, C. (2019). The psychological reality of practical representation. *Philosophical Psychology*, 32(5), 785–822.
- Pavese, C. (2021). Practical concepts and productive reasoning. *Synthese*, 199, 7659–7688.
- Pavese, C. (2026). *The practical mind: Skill, knowledge, and intelligence*. Cambridge University Press.

- Rescorla, M. (2016). Bayesian sensorimotor psychology. *Mind & Language*, 31(1), 3–36.
- Rosefeldt, T. (2004). Is knowing-how simply a case of knowing-that? *Philosophical Investigations*, 27(4), 370–380.
- Schack, T. (2004). The cognitive architecture of complex movement. *International Journal of Sport and Exercise Psychology*, 2(4), 403–438.
- Schiffer, S. (2002). Amazing knowledge. *The Journal of Philosophy*, 99(4), 200–202.
- Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review*, 82(4), 225–260.
- Schmidt, R. A. (2003). Motor schema theory after 27 years: Reflections and implications for a new theory. *Research Quarterly for Exercise and Sport*, 74(4), 366–375.
- Stanley, J. (2011). *Know how*. Oxford University Press.
- Stanley, J. (2012). Replies to Dickie, Schroeder, and Stalnaker. *Philosophy and Phenomenological Research*, 85(3), 778–787.
- Stanley, J., & Williamson, T. (2001). Knowing how. *The Journal of Philosophy*, 98(8), 411–444.
- Wrights Hickman, N. (2019). Knowing in the “executive way”: Knowing how, rules, methods, principles and criteria. *Philosophy and Phenomenological Research*, 99(2), 311–335.
- Weiskopf, D. A. (2009). The plurality of concepts. *Synthese*, 169, 145–173.
- Wing, A. M. (2000). Motor control: Mechanisms of motor equivalence in handwriting. *Current Biology*, 10, R245–R248.

Chapter 3 PSYCHOLOGICAL DEPTH IN LARGE LANGUAGE MODELS

Chapter 3 1 Introduction

This chapter brings together two stages of a collaborative research project on the Psychological Depth Scale (PDS), which is a framework for evaluating the psychological impact of human- and machine-generated narratives. The first study develops and empirically validates the PDS as a reader-centered measure of five dimensions of psychological depth: authenticity, empathy, engagement, emotion provocation, and narrative complexity. The second study extends this work to more recent reasoning-enhanced language models and examines whether automated PDS evaluation continues to provide a valid approximation of human judgment when it is applied to closely matched outputs from frontier models.

My contribution to the first study began with the theoretical foundations of the PDS. I examined the literature review and dataset underlying the initial draft. The initial version contained approximately 140 evaluative components drawn from 95 publications, from which the five main components of the PDS had been extracted. Although the empirical framework had already been developed, the relationships among these five components and the rationale for selecting them required further articulation. This concern had also been raised in reviewer feedback on an earlier submission. I therefore reconducted the literature review with the aim of establishing a clearer theoretical foundation for the scale and explaining how its five dimensions emerged from the broader literature.

I rewrote the section of the paper devoted to the conceptual foundations of the PDS, developing both a theoretical rationale and empirical support for each component.

I also worked to distinguish dimensions that could otherwise appear to overlap. Engagement and emotion provocation, for example, both concern aspects of a reader's response to a story, but the relevant literatures characterize them as separable experiences. Engagement concerns sustained attentional involvement and transportation into a narrative, whereas emotion provocation concerns the strength and character of the affective responses elicited by it. Similar distinctions were developed across the remaining dimensions. My contribution situated the PDS more explicitly as a reader-centered measure and addressed concerns that its dimensions were insufficiently motivated or differentiated.

Because I had not participated in the original empirical design or initial data collection, I worked closely with the research team to understand the generation, annotation, and evaluation procedures. This gave me practical familiarity with prompt engineering, LLM-based narrative generation, and human annotation pipelines. I also contributed to revisions intended to ensure that the interpretation of the empirical results remained appropriately constrained by what the data supported.

The follow-up project approached psychological depth through a different question: whether improvements in reasoning ability might transfer to improvements in psychological depth. One working proposal was that training on extended multistep reasoning could strengthen planning capacities that also contribute to narrative construction. A competing possibility was that highly constrained reasoning tasks, especially those organized around exact answers and narrow correctness conditions, could encourage forms of generation less conducive to psychologically rich narrative writing.

My initial role was to help identify and classify reasoning datasets that could support controlled comparisons between these possibilities. Early automated PDS results appeared to suggest that DeepSeek-R1 outperformed DeepSeek-V3 on PDS, raising the question of what aspect of reasoning training might be responsible. Separate analysis within the project suggested that the reasoning-oriented model's outputs tended to score higher in dimensions such as prompt adherence, logical strictness, and novelty. This led me to consider whether different forms of reasoning might have different relationships to creative generation. The emergence of distinct LLM modes such as search and extended reasoning heightened interest in the possibility that apparently similar improvements in problem-solving ability might turn on importantly different processes. In exploratory comparisons, retrieval-oriented modes often appeared to produce flatter narratives than extended reasoning modes.

I therefore proposed reframing the question from whether reasoning generally improves psychological depth to which forms of reasoning support psychological depth, which do not, and why. I developed a preliminary framework organized along two dimensions: type of inference and degree of required precision. The first dimension distinguished deductive, inductive, and abductive reasoning. The second concerned the extent to which a task constrains the form and specificity of an acceptable answer. Formal proofs and code, for example, impose highly constrained correctness conditions, whereas other natural-language reasoning tasks permit a wider range of acceptable inferential paths. I used this framework to examine candidate datasets. GSM8K and parts of the LSAT were considered as possible cases of comparatively less formalized deductive reasoning, while UNCommonsense provided a candidate for abductive

reasoning. I also examined LogicBench, including its classical and nonmonotonic reasoning branches, and identified tasks involving default reasoning under open domains and default reasoning with conflicting information as especially relevant to reasoning under uncertainty. The underlying hypothesis was that forms of reasoning permitting plausible inferential movement under incomplete information might be more closely related to some of the constructive capacities involved in psychologically rich narrative generation than highly precision-oriented reasoning.

As the project developed, however, the original training-based question became difficult to isolate empirically. Apparent gains in models such as DeepSeek-R1 could not straightforwardly be attributed to reasoning training because the models also differed in other aspects of post-training and fine-tuning. The project therefore shifted toward controlled comparisons among off-the-shelf frontier models. My earlier taxonomy of reasoning tasks remains a possible direction for a future experiment designed specifically to isolate different kinds of reasoning training.

I also contributed to the empirical development of the follow-up study. I worked on prompt design for narrative generation, aiming to keep the outputs sufficiently controlled for meaningful comparison while preserving enough freedom for differences in psychological depth to emerge. I also participated in the evaluation of model-generated stories according to the PDS dimensions, providing initial evidence about how different models varied across the scale.

The eventual direction of the follow-up study emerged from an unexpected empirical problem. Automated PDS evaluation strongly appeared to favor reasoning-enhanced models. When human readers were subsequently asked to compare the

same outputs, however, their judgments were far less uniform and differed sharply across model families and individual readers. The resulting question was therefore broader than whether reasoning improves psychological depth. It concerned the validity of automated evaluation itself: when an LLM-based evaluator has correlated well with human PDS ratings on one dataset, under what conditions can it be trusted to approximate human judgment on a new and more difficult distribution?

The two studies that follow trace this progression. The first develops the PDS and establishes its initial empirical reliability. The second places the scale under a substantially more demanding form of evaluation and shows that agreement between an automated evaluator and human judgment can deteriorate even when the automated evaluator previously performed well against human annotations.

Chapter 3 2 Study 1: Measuring Psychological Depth in Language Models

Chapter 3 2 1 Motivation and General Approach

Evaluations of LLM-generated stories commonly concentrate on properties of the text itself, such as fluency, coherence, style, diversity, discourse structure, creativity, toxicity, and bias. These properties are useful for determining whether a generated text is well formed or satisfies particular structural criteria, but they do not directly characterize what the narrative does to a reader. A story can be fluent and stylistically polished while remaining emotionally flat, psychologically unconvincing, or difficult to become invested in. Conversely, a story containing stylistic imperfections can still produce a strong emotional or imaginative response.

The first study therefore shifts the unit of evaluation from the text considered in isolation to the interaction between the text and its reader. Psychological depth is

treated as a property of the reading experience that can be elicited by features of a narrative. The Psychological Depth Scale was developed to capture five aspects of this experience: authenticity, empathy, engagement, emotion provocation, and narrative complexity.

This approach draws particularly from reader-response criticism and text world theory. Reader-response approaches emphasize that literary meaning is partly constituted through the reader's activity rather than residing entirely in textual properties. The relevant question is therefore not only what a sentence or narrative means, but what it does in the experience of a reader (Fish, 1970; Mailloux, 1976; Babae and Montashery, 2012). Text world theory provides a complementary cognitive perspective. On this view, readers construct and continuously revise mental representations of the narrative world as they encounter new information (Gavins, 2007; Canning, 2017). A psychologically effective narrative can therefore be understood partly in terms of the cognitive and affective processes it elicits while that mental world is constructed.

The final literature review surveyed 95 peer-reviewed articles and books across areas including cognitive psychology, media studies, literary and narrative theory. From this literature, 143 candidate evaluative components were extracted and organized thematically. These components were consolidated into the five dimensions of the PDS. The objective was not to claim that these five dimensions exhaust every psychologically relevant feature of narrative experience. Instead, the scale was designed to provide a tractable set of relatively broad dimensions with substantial coverage of the phenomena repeatedly discussed in the literature.

Chapter 3 2 2 The Five Dimensions of Psychological Depth

Emotion provocation concerns a narrative's capacity to elicit strong affective responses in a reader. These responses can be positive or negative and need not correspond straightforwardly to the amount of emotionally charged language contained in the text. Research on emotion indicates that affective responses depend on interactions among valence, arousal, context, and interpretation. Increasing the explicit intensity of emotional language can therefore fail to increase emotional effect and can sometimes reduce it through phenomena such as compassion fatigue or psychic numbing (Kinnick et al., 1996; Slovic, 2007; Citron et al., 2014; Maier, 2015; Maier et al., 2016; Megalakaki et al., 2019). Emotion provocation consequently measures the achieved response of the narrative rather than the superficial presence of emotional vocabulary.

The relevance of emotion extends beyond immediate affect. Emotionally evocative narratives can influence attention, memory, and subsequent reflection. Evoked emotions can persist or recur after the reading experience itself has ended (Mar et al., 2011). A psychologically deep story can therefore have an emotional consequence that is temporally broader than the moment at which a particular sentence is encountered.

Empathy concerns the capacity of a narrative to elicit identification with characters and cognitive participation in their experiences. Empathetic engagement can involve imaginatively occupying another perspective and temporarily privileging the fictional perspective over one's ordinary point of view (Miall and Kuiken, 2001; Oatley, 2002; Kaufman and Libby, 2012; Zaki and Ochsner, 2012). Both behavioral and

neurocognitive work supports the importance of such perspective shifts in narrative experience.

Empathetic narrative response also bears on introspection and social cognition. Perspective-taking can facilitate reflection on experiences removed from the reader's own circumstances, while literary devices such as metaphor and metonymy can connect individual fictional situations with more general aspects of human experience (Oatley, 2002; Johnson-Laird and Oatley, 2022). Empathy is therefore treated as a distinct component of psychological depth rather than as a synonym for emotional intensity. A reader can experience strong emotion without identifying with a character, and can understand or inhabit a character's perspective without experiencing an equally intense emotional response.

Engagement measures a story's ability to capture and maintain the reader's attention. It is closely related to the literature on narrative transportation, in which readers become sufficiently absorbed in a story that attention to the external environment decreases (Gerrig, 1993; Green and Brock, 2000; Busselle and Bilandzic, 2008, 2009). Highly engaged readers may lose track of time or fail to notice changes in their surroundings, whereas disengagement is associated with distraction and frustration.

Engagement interacts with emotion but remains conceptually distinguishable from it. Transportation has both attentional and affective components, and a narrative can retain attention without producing an especially strong emotional reaction. Similarly, a brief passage may provoke a powerful emotion without sustaining engagement across an entire story. Treating engagement and emotion provocation separately therefore

permits the PDS to distinguish narratives that are compelling to continue reading from narratives that are affectively intense.

Authenticity concerns whether a narrative presents experiences, emotions, and relationships in a way that feels psychologically genuine. Authenticity does not require literal similarity between the fictional world and ordinary life. A speculative narrative involving radically unfamiliar settings or entities can still feel authentic if the motivations, reactions, tensions, or emotional relationships represented in it are experienced as recognizably human.

The concept draws from philosophical accounts of authenticity as well as empirical research on realism, belonging, and narrative persuasion. Authentic expression has been associated with the manifestation or construction of a self and with the articulation of features of human existence (Kierkegaard, 1980; Heidegger, 1962; Taylor, 1991; Berlin, 2000; Rousseau, 2016). Empirical work also indicates that judgments of authenticity and realism affect interest, identification, and persuasive impact (Hall, 2003; Green, 2004; Newman and Smith, 2016; Petraglia, 2009; Zwarun and Hall, 2012). The PDS accordingly treats authenticity as the extent to which a story conveys the complexity of experience in a way that a reader can recognize or inhabit.

Narrative complexity concerns richness in plot and character development, especially when a narrative invites the reader to revise or elaborate an interpretation as the story unfolds. It can be produced through nonlinear structures, perspective shifts, plot reversals, concealed information, or other forms of narrative organization that require active cognitive restructuring (Kiss and Willemsen, 2017; Cutting, 2019).

Narrative complexity must also be distinguished from linguistic or narrational difficulty. A simple story can be told obscurely, while a structurally or psychologically complex story can be expressed in straightforward prose. Complexity therefore concerns the structure of the represented narrative and the interpretive activity that structure solicits, rather than the surface difficulty of the language itself. This distinction is especially important when evaluating LLM-generated stories, since surface sophistication can otherwise be mistaken for underlying narrative richness.

The five PDS dimensions are designed to characterize different aspects of the reader's psychological encounter with a narrative. Their boundaries are not intended to imply that the underlying processes are independent. Engagement may facilitate emotion, empathy may contribute to perceived authenticity, and complexity may sustain engagement. The scale nevertheless separates these dimensions because they track different aspects of reader experience and can vary independently enough to provide diagnostically useful information.

Chapter 3 2 3 Construction of the PSYCHDEPTH Dataset

The empirical study used a dataset of English-language premise-response pairs with authorship metadata. The full dataset consisted of 495 stories, including 45 human-authored stories and 450 LLM-generated stories. Stories averaged approximately 450 words. Because manual annotation of the full collection would have been costly and time intensive, a stratified subset of 97 stories was constructed for the human study. Sampling was designed to balance story premises, authorship, and generation strategies.

Human-authored stories were collected from Reddit's r/WritingPrompts community. The structure of that community was useful because writers respond to shared premises and readers can upvote their responses, providing a rough public signal of reception. Human stories were divided into three quality strata according to their ranking. The Human-Advanced category contained highly ranked stories and averaged 1,434 upvotes. Human-Intermediate stories averaged 263 upvotes, while Human-Novice stories averaged 9. Stories used in a comparison were selected from posts appearing within a sufficiently narrow time window to reduce differences in exposure as an explanation of voting differences.

The machine-generated portion of the original study included Llama-2 models with 7, 13, and 70 billion parameters, Vicuna-33B, and GPT-4. This selection allowed the study to examine both variation across model families and possible relationships between model scale and psychological depth.

Two generation strategies were used. The Writer Profile strategy relied on in-context impersonation. The model was given a description of an accomplished writer whose work was characterized by psychological richness, complex emotional states, and engaging narrative development. The profile was intended to prime the model toward qualities associated with psychological depth. The Plan+Write strategy separated generation into a preliminary character-development stage and subsequent story composition. Before composing the story, the model developed portraits of the major characters, including information about their emotional states and internal perspectives. The story was then written with these portraits available as context. Early experimentation considered additional explicit planning stages for setting, plot, and

outline, but adding more stages tended to reduce coherence in these relatively short stories. The final procedure therefore concentrated on character portraits.

Chapter 3 2 4 Human Evaluation

The human study recruited undergraduate participants from UCLA's English and Psychology departments. Forty-seven students applied, and five were selected based on their interests and prior experience with literary or psychological analysis. The aim was to recruit informed nonexpert readers rather than either professional literary critics or anonymous crowd workers. Participants received an introduction to the five PDS dimensions and completed tutorial annotations intended to calibrate their understanding of the rubric. They then evaluated the stories independently and remotely over a seven-day period. Stories were divided into batches of twenty to reduce fatigue.

For every story, participants rated authenticity, empathy, engagement, emotion provocation, and narrative complexity on five-point Likert scales. They also rated how likely they believed the story was to have been written by a human, on a scale ranging from 1, indicating an LLM, to 5, indicating a human author. Participants could additionally provide written explanations for their judgments. Each story in the human-study sample was evaluated by all five annotators. Participants spent approximately 7.8 hours on the annotation task on average and were paid \$100. The resulting dataset contained 2,425 individual PDS ratings, 485 authorship-likelihood judgments, and 1,128 free-text justifications.

Chapter 3 2 5 Reliability of Human PDS Judgments

The first research question concerned whether different readers could use the PDS with sufficient consistency for the framework to function as a meaningful evaluative

instrument. Agreement was measured with Krippendorff's alpha using an ordinal distance metric appropriate to the Likert ratings. Agreement was similar across the five dimensions. Krippendorff's alpha was 0.71 for authenticity, 0.74 for empathy, 0.70 for engagement, 0.71 for emotion provocation, and 0.74 for narrative complexity. Averaged across dimensions, alpha was 0.72. These values suggested substantial agreement among the five human readers in the original study. Importantly, the agreement concerned judgments about reader-centered properties that might initially appear highly subjective. The result suggested that, at least for the distribution represented by the original dataset, informed readers could apply the PDS dimensions with considerable consistency.

Chapter 3 2 6 Automating Psychological-Depth Evaluation

Human evaluation is costly and difficult to scale. The second research question therefore examined whether contemporary LLMs could approximate human PDS judgments. Four models were evaluated as automated judges: Llama-3-8B, Llama-3-70B, GPT-3.5, and GPT-4o. These models received instructions analogous to those given to human annotators and were asked both to assign numerical scores and to provide explanations. Constrained decoding was used for locally hosted Llama models to guarantee parseable outputs without requiring the model to devote generation capacity to formatting.

Two evaluation conditions were compared. The baseline used ordinary zero-shot prompting. The second condition used a Mixture-of-Personas strategy. GPT-4o was first used to generate five personas relevant to literary and psychological analysis. The evaluation was then repeated from each of these perspectives. The motivation was to

introduce structured diversity into automated judgment, approximating some of the variation produced by multiple human readers. The persona-based judges remained considerably more internally consistent than the human annotators. Average Krippendorff's alpha across the five dimensions was approximately 0.94 for both Llama-3-8B and Llama-3-70B, 0.90 for GPT-3.5, and 0.92 for GPT-4o, compared with 0.72 among human raters. The personas therefore induced some variation, although less than the variation observed across people.

The more consequential question was whether this diversity improved agreement with human consensus. It did. Across models and dimensions, Mixture-of-Personas prompting increased the average Spearman correlation between LLM and human ratings by approximately 20.43 percent relative to the corresponding zero-shot evaluations. For Llama-3-8B, the mean correlation across PDS dimensions increased from 0.299 to 0.375. For Llama-3-70B, it increased from 0.431 to 0.479. GPT-3.5 increased from 0.343 to 0.434, and GPT-4o increased from 0.417 to 0.507. Some of the largest improvements occurred for authenticity and empathy. Llama-3-8B's correlation with human authenticity judgments rose from 0.079 to 0.318. Across all evaluated models, authenticity correlations improved by approximately 33.81 percent with Mixture-of-Personas prompting. For empathy, Llama-3-70B improved from 0.579 to 0.679, the highest individual human-correlation result in the study.

No single model was best across all five PDS dimensions. GPT-4o produced the highest average correlation with humans, approximately 0.51, and performed especially well on authenticity and engagement. Llama-3-70B was strongest on empathy, with a correlation of approximately 0.68, narrative complexity at approximately 0.62, and

emotion provocation at approximately 0.57. This variation suggested that automated PDS evaluation might benefit from using different evaluators for different dimensions rather than assuming that a single LLM constitutes the best general judge.

Chapter 3 2 7 Psychological Depth in Human and LLM Stories

The third research question compared the psychological depth of human- and LLM-authored narratives. GPT-4 received the strongest overall PDS scores among the machine-generated stories. Its mean authenticity score was 3.89, empathy 3.68, engagement 3.94, emotion provocation 3.53, and narrative complexity 3.80. By comparison, the Human-Advanced stories received mean scores of 3.60 for authenticity, 2.95 for empathy, 3.90 for engagement, 3.65 for emotion provocation, and 2.95 for narrative complexity. Human-Intermediate stories received means of 3.53, 2.93, 3.80, 3.27, and 3.00 respectively, while Human-Novice stories received 2.73, 2.07, 3.27, 2.67, and 2.20. The remaining LLM families generally received lower scores than GPT-4. Llama-2-7B received mean scores of 2.92 for authenticity, 2.62 for empathy, 2.77 for engagement, 2.64 for emotion provocation, and 2.48 for narrative complexity. Llama-2-13B received 2.96, 2.73, 2.51, 2.53, and 2.43. Vicuna-33B received 2.76, 2.59, 2.61, 2.59, and 2.55. Llama-2-70B performed somewhat better, with corresponding means of 3.09, 2.99, 3.01, 2.94, and 2.73.

Pairwise statistical comparisons indicated that GPT-4 exceeded the highly rated Reddit stories with statistical significance on empathy and narrative complexity. On authenticity, engagement, and emotion provocation, its scores were statistically indistinguishable from the advanced human stories. GPT-4 also showed relatively small

variance compared with many of the other author categories, suggesting comparatively consistent performance.

The relationship between perceived human authorship and psychological depth was also informative. Human stories received an average human-authorship rating of 4.18, while GPT-4 received 3.91. Thus, although readers still regarded human-written stories as somewhat more likely to have been produced by humans, the GPT-4 stories often received equal or higher psychological-depth ratings. Across the study, participants correctly distinguished human from LLM authorship only 56 percent of the time. Classification accuracy for GPT-4 stories was substantially lower, at approximately 27 percent. In other words, readers classified nearly three quarters of GPT-4 stories as human. GPT-4 itself also performed poorly when asked to distinguish human from machine-written stories, reaching approximately 39 percent accuracy.

Free-text justifications provided additional information about the cues readers used. A thematic analysis identified recurring explanations for authorship judgments. GPT-4 stories were frequently characterized as creative and nuanced: 89 percent were associated with creativity-related observations and 94 percent with nuance, compared with approximately 53 percent for most human-generated stories. GPT-4 stories also contained grammatical irregularities more frequently than the other evaluated LLMs, with such issues observed in approximately 17 percent of its stories. Interestingly, readers sometimes interpreted these irregularities as evidence of human authorship. GPT-4 also tended to avoid cues commonly associated with machine-generated fiction, such as simplistic character names or formulaic endings that explicitly state a moral.

Model size by itself was only weakly related to psychological depth. Across the evaluated LLMs, the Pearson correlation between parameter count and depth ratings was approximately 0.31. Smaller models occasionally performed competitively with larger ones, indicating that increases in parameter count alone did not explain narrative quality.

Chapter 3 2 8 Discussion and Limitations of the Initial PDS Study

The first study provided evidence for three claims. First, readers could apply the five PDS dimensions with substantial agreement on the original dataset. Second, automated evaluators could recover a meaningful portion of human judgment, particularly when multiple personas were used to introduce diversity into evaluation. Third, the strongest model in the generation study, GPT-4, produced short narratives whose measured psychological depth was comparable to, and on some dimensions higher than, highly rated human stories from Reddit.

These results differ from work finding much larger human advantages in creative writing. Chakrabarty et al. (2024), for example, reported substantial human advantages using the Torrance Test of Creative Writing. Several methodological differences may contribute to the contrast. Their narratives were considerably longer, while the PDS study used stories of roughly 450 words. Long-form generation places different demands on coherence and planning. The PDS prompts were also relatively elaborate and often provided complex narrative premises. Finally, the two evaluation frameworks operationalize creative quality differently. PDS specifically targets effects on the reader rather than creativity in a broader sense.

The study has several limitations and potential risks. First, using human-written content from Reddit's r/WritingPrompts does not guarantee that all selected stories were fully written by humans, despite the community's rules prohibiting AI-generated content. In addition, although Reddit's voting system and large user base provide a reasonable signal of writing quality, the selected stories may not represent the highest caliber of human writing. The study addressed this in part by introducing multiple levels of human writing quality, while future work could identify more reliable sources of high-quality stories. Second, although the five components of psychological depth are grounded in an extensive literature review, they are not intended to comprehensively cover every psychological aspect of reading. The aim was to keep annotation tractable while maximizing semantic coverage. The components provide a structured framework for analyzing narrative quality and reader response. Third, the study relies on a relatively small dataset of short fictional stories and therefore does not capture the full diversity of storytelling styles and narrative structures. The PDS components were also designed primarily for this form of creative writing. Additional evaluation would be required to determine whether the framework generalizes to other forms of writing, including screenplays, scripts, and speeches. Fourth, the prompts used for both story generation and evaluation were carefully reviewed and iteratively improved but are still likely to be suboptimal. Future work could refine these prompts using improved prompting methods and current best practices.

Finally, the methods developed for enhancing and assessing psychological depth carry potential risks if misapplied. Techniques for increasing or automatically measuring psychological engagement could be used to disseminate misinformation more

effectively, since emotionally charged or psychologically resonant messages may be more memorable and influential. As LLM-generated content increasingly resembles human creative writing, distinguishing between human and machine-generated content may also become more difficult, potentially undermining trust in digital communications, especially in domains such as journalism and political discourse.

Chapter 3 3 Study 2: Reasoning-Enhanced Models and the Validity of Automated Psychological-Depth Evaluation

Chapter 3 3 1 Motivation

LLM-as-a-Judge methods are increasingly used to compare model outputs, evaluate checkpoints, and provide scalable feedback for training. Their usefulness depends on whether the evaluator measures the intended property in the context in which it is deployed. A judge can correlate strongly with human ratings on one dataset while relying on cues that cease to be valid, or acquire different significance, on another. Prior work has documented several forms of evaluator bias, including position effects, verbosity preferences, self-preference, and inconsistency (Zheng et al., 2023; Panickssery et al., 2024; Stureborg et al., 2024; Wang et al., 2024; Li et al., 2025). Pairwise LLM judging can improve agreement with human preferences in some settings while also amplifying systematic biases (Liu et al., 2025; Jeong et al., 2025). This problem is particularly difficult for subjective creative evaluation. Fluency-based metrics may reward polished writing without capturing the qualities that make fiction psychologically effective (Gómez-Rodríguez and Williams, 2023; Chakrabarty et al., 2024; Lu et al., 2026). Human disagreement in subjective tasks can also reflect genuine differences in preference or interpretation rather than annotation error alone (Clark et

al., 2021; Sandri et al., 2023; Kirk et al., 2024). If people disagree for structured reasons, a single point estimate produced by an automated judge may obscure important features of the target being measured.

This study investigates this problem through the PDS. Reasoning-enhanced models provided a useful test case because models such as GPT-5 and DeepSeek-R1 are routinely evaluated on mathematical, coding, and factual benchmarks. Whether reasoning-oriented post-training also affects open-ended creative generation is less clear. The empirical question therefore had two layers. The first concerned the outputs themselves: do reasoning-enhanced models produce stories that humans regard as psychologically deeper than those produced by comparison models? The second concerned measurement: does an automated PDS evaluator selected for its agreement with human ratings on the original PSYCHDEPTH dataset reproduce human preferences when applied to these newer and more closely matched outputs?

Chapter 3 3 2 Construction of the Frontier Story-Pair Dataset

The study constructed 60 blinded, prompt-matched pairs across three model comparisons. Twenty pairs compared DeepSeek-R1 with DeepSeek-V3. Twenty compared GPT-5 operating with its automatic reasoning-effort setting with GPT-4o. A further twenty compared GPT-5 using high reasoning effort with GPT-4o. Twenty narrative premises were used. Fifteen were selected from r/WritingPrompts and five from Reedsy. The Reddit prompts were restricted to material posted after the relevant models' training-data cutoffs, ranked using community reception, and manually screened for suitability. They largely provided high-concept speculative scenarios. The

Reedsy prompts supplied broader literary instructions involving character relationships, concealed information, warnings, fear, and unlikely friendships.

For every premise, the relevant model pair received closely matched instructions. The generation prompt described an accomplished writer whose stories were emotionally rich, psychologically attentive, and concerned with complex relationships and conflicts. Models were asked to produce approximately 500 words. DeepSeek-V3 and DeepSeek-R1 received the same additional instruction requiring at least 400 words because both models otherwise tended to produce responses below the target range. Since this modification was applied identically within the DeepSeek comparison, it did not distinguish R1 from V3.

All generations used a temperature of 1.0 and otherwise default sampling settings. GPT-5's high-effort condition used the model's high reasoning-effort setting. DeepSeek-R1's internal thinking output was removed before the resulting story was presented to readers. Stories were required to fall between 400 and 600 words. When an output fell outside this range, generation was retried up to ten times and the sample closest to 500 words was retained. Retry counts by model were not recorded, which means that unequal sampling opportunities cannot be ruled out as a source of variation. GPT-5 in automatic mode used an average of approximately 7,846 reasoning tokens on retained generations, with a standard deviation of 3,117. The high-effort condition used approximately 11,606 reasoning tokens on average, with a standard deviation of 3,557.

One retained sample was used for each model and premise condition. Consequently, differences between models include ordinary generation-level sampling variation in addition to systematic model differences. The model comparisons should

also be interpreted carefully. DeepSeek-R1 is closely related to V3 and provides the study's strongest comparison involving reasoning-oriented post-training. GPT-5 and GPT-4o, by contrast, differ in more ways than reasoning effort alone. The term “reasoning output” is therefore used as a descriptive label for outputs from the reasoning-oriented model rather than as a claim that any observed difference was causally produced by reasoning.

Chapter 3 3 3 Construction of the Automated PDS Evaluator

The automated evaluator was built using the original PSYCHDEPTH dataset and human PDS annotations from Study 1. The objective was to construct an evaluator that performed substantially better than the earlier zero-shot judge before deploying it on the new story pairs. The 97-story annotated dataset was divided into a 20 percent demonstration pool and an 80 percent development and selection split. Representative few-shot examples were selected from the demonstration pool by clustering the PDS score vectors with k-means and choosing examples that represented different regions of the rating space. Candidate evaluators included Llama 3.1 70B, Llama 3.3 70B, and Qwen 3.5 397B under different numbers of demonstrations and, in one condition, Chain-of-Thought prompting. DSPy was used to enforce structured outputs, while the Mixture-of-Personas strategy from the original PDS work was retained.

Performance varied by PDS dimension. For authenticity, the strongest development-set configuration was Llama 3.1 70B with five demonstrations, with a Spearman correlation of 0.588 against human ratings. For emotion provocation, Llama 3.1 70B with ten demonstrations reached 0.702. For empathy, Llama 3.3 70B with ten demonstrations also reached 0.702. Qwen 3.5 397B with seven demonstrations

performed best on engagement at 0.564. Narrative complexity was assigned to a ten-demonstration Chain-of-Thought configuration of Llama 3.1 70B, which reached 0.664. These dimension-specific configurations were combined into a routed evaluator. Across routed predictions, the resulting evaluator achieved a Spearman correlation of 0.646 with human PDS ratings, exceeding every individual candidate configuration and the approximately 0.51 correlation achieved by the earlier GPT-4o zero-shot PDS evaluator.

The value of 0.646 requires an important qualification. The same 80 percent subset was used both to compare candidate configurations and to determine which configuration would be used for each dimension. The reported correlation is therefore a development or selection-set estimate. It is not an estimate from an untouched test set. Such design reflects a common practical pattern in which an evaluator is selected on an available labeled dataset and subsequently reused, but it means that the correlation should not be interpreted as independent evidence of generalization.

During deployment, every story received a score from 1 to 5 on each PDS dimension. The higher score within a pair determined the evaluator's preference for that dimension. Exact ties contributed one-half of a preference to each model when pooled rates were calculated. An aggregate PDS score was also computed by averaging the five dimension scores and comparing the two stories. This aggregate score differs conceptually from the human readers' separate question asking which story they preferred overall.

Chapter 3 3 4 Human Evaluation

Seven undergraduate readers from English and Psychology departments participated in the follow-up study. They were selected from 70 applicants based on

English proficiency, interest in the project, and experience evaluating short stories. The final number of participants was constrained by the study budget.

Participants completed online onboarding covering the five PDS dimensions and the annotation procedure. The study was reviewed by the relevant Institutional Review Board and determined to be exempt because participants served as annotators rather than as the subjects of the research. For each of the 60 story pairs, the two stories were displayed side by side in randomized order. Readers saw the definitions of the five PDS dimensions while performing the task. They made six forced-choice judgments for every pair: one judgment for each of the five PDS dimensions and a sixth judgment indicating which story they preferred overall. No tie or confidence option was available. Readers could also provide free-text explanations. The study therefore produced 2,520 pairwise judgments, calculated as seven readers multiplied by 60 pairs multiplied by six questions. Readers also supplied 112 written justifications.

Several checks were used to assess annotation quality. Median evaluation time per pair varied considerably across readers, from 108 to 418 seconds, with an overall median of approximately 5.2 minutes. One participant completed the task substantially faster than the others, with a median of 108 seconds per pair, but had professional reading experience and maintained a stable pace throughout the annotation sequence. A leave-one-reader-out analysis showed that removing this participant changed pooled reasoning-model preference by approximately two to four percentage points depending on the criterion. Removing other readers produced changes of comparable magnitude. The unusually fast reader therefore did not appear to be responsible for the aggregate pattern.

Post-study responses were similarly reassuring. All seven participants reported relatively high confidence in their ability to use the PDS dimensions, generally rating their confidence at 4 or 5 out of 5. Narrative complexity produced the lowest confidence. Four readers reported no fatigue. Of the three who reported some fatigue, only one believed that it would have changed their annotations, while two were uncertain. All seven rated the instructions and their overall experience highly. These checks cannot eliminate annotation noise, but they provide little evidence that the main results were produced by a single inattentive reader or a general collapse in task engagement.

Chapter 3 3 5 Statistical Analysis

Human preferences were analyzed separately for each combination of model comparison and evaluative criterion. Intercept-only logistic mixed-effects models included crossed random intercepts for the story premise and the reader. This structure accounts for the repeated use of the same readers and the fact that individual premise pairs can differ systematically in difficulty. The resulting confidence intervals were pointwise 95 percent intervals. Eighteen model-by-criterion comparisons were examined, and the intervals were not adjusted for multiplicity. They are therefore most appropriately interpreted descriptively as indicators of the pattern and uncertainty of preferences.

Inter-reader agreement was characterized using Fleiss' kappa, Krippendorff's alpha, and raw pairwise agreement. Reader-level logistic regressions examined how the five PDS dimensions related to each reader's overall preference. Cohen's kappa was also computed within readers to measure correspondence between each dimension-specific judgment and the reader's overall preference. The study additionally examined

whether surface properties of the stories were associated with evaluator scores. These properties included word count, mean sentence length and other sentence-length statistics, moving-average type-token ratio (MATTR), Measure of Textual Lexical Diversity (MTLD), and perplexity under Llama 3.1 70B. Because these analyses were exploratory and the observations shared prompts and model families, they are treated as descriptive associations rather than causal effects.

Chapter 3 3 6 Automated Evaluation Strongly Favors Reasoning Outputs

The automated evaluator produced an exceptionally consistent preference for outputs from the reasoning-oriented models. Across all 300 pair-by-dimension comparisons, consisting of 60 story pairs evaluated on five PDS dimensions, the routed evaluator favored the reasoning-model output at an even-split rate of 89.0 percent. Its preference was strongest for engagement, at 95.8 percent, followed by emotion provocation at 94.2 percent, empathy at 87.5 percent, authenticity at 85.0 percent, and narrative complexity at 82.5 percent. The aggregate PDS score was still more extreme. The reasoning-model story received the higher aggregate score in 59 of the 60 story pairs, or 98.3 percent. GPT-5 received the higher aggregate PDS score in all 40 comparisons with GPT-4o. DeepSeek-R1 received the higher aggregate score in 19 of its 20 comparisons with DeepSeek-V3.

The magnitude of the judge-score difference was substantial. For GPT-5 in automatic reasoning mode relative to GPT-4o, the mean paired judge-score difference across the routed dimensions was approximately 0.518 points, corresponding to a standardized descriptive difference of 1.14. For GPT-5 high reasoning effort relative to GPT-4o, the difference was 0.500 points, with a standardized value of 1.12. DeepSeek-

R1's advantage over V3 was smaller but remained positive, approximately 0.214 points with a standardized value of 0.68. Pooled across all three comparisons, the mean difference was approximately 0.411 points with a standardized value of 0.95. This pattern did not arise solely from the routing procedure. Every individual evaluator configuration also favored reasoning outputs strongly. Depending on the configuration, the even-split preference rate ranged from 77.5 percent to 96.3 percent. Thus, substituting one of the individual judges for the dimension-routed ensemble would have weakened or strengthened the effect but would not have eliminated its basic direction.

Chapter 3 3 7 Human Readers Do Not Show a Universal Reasoning Advantage

Human evaluations produced a very different pattern than the automated evaluation results. Across all 2,100 dimension-level human votes, reasoning-model stories received 58.0 percent of preferences. Across the 420 overall-preference votes, they received 55.2 percent. More importantly, these pooled percentages conceal substantial differences among the three model comparisons.

For DeepSeek-R1 versus DeepSeek-V3, readers preferred R1 on authenticity in 47.9 percent of judgments, with a 95 percent interval from 39.0 to 56.4 percent. R1 received 52.9 percent of empathy preferences, with an interval from 44.4 to 62.2; only 39.3 percent of engagement preferences, with an interval from 29.6 to 46.6; 61.4 percent of emotion-provocation preferences, with an interval from 53.6 to 69.9; and 58.6 percent of narrative-complexity preferences, with an interval from 50.9 to 67.7. On the separate overall-preference question, readers chose R1 only 42.9 percent of the time, with an interval from 33.7 to 50.9 percent. Thus, although R1 performed relatively well

on emotion provocation and narrative complexity, readers as a group leaned toward V3 overall.

By contrast, in the automatic-effort comparison with GPT-4o, GPT-5 received 55.7 percent of authenticity preferences, 68.6 percent of empathy preferences, 58.6 percent of engagement preferences, 70.7 percent of emotion-provocation preferences, and 55.0 percent of narrative-complexity preferences. Readers selected GPT-5 as their overall preferred story in 60.0 percent of judgments, with a 95 percent interval from 52.6 to 69.5 percent. When GPT-5 used high reasoning effort, its authenticity preference rate was 52.1 percent, empathy 69.3 percent, engagement 65.0 percent, emotion provocation 62.9 percent, and narrative complexity 52.9 percent. Its overall-preference rate was 62.9 percent, with a 95 percent interval from 55.9 to 72.7 percent. The human data therefore provide some evidence favoring GPT-5 over GPT-4o, particularly on empathy, engagement, and emotion provocation. They do not support a model-family-independent reasoning advantage. DeepSeek-R1 was disfavored overall relative to V3 despite the automated evaluator's preference for R1 in 19 of 20 aggregate comparisons.

That reasoning models received 55.2 percent of overall human preferences thus conceals a qualitative split. GPT-5 was modestly preferred to GPT-4o, whereas DeepSeek-R1 was modestly disfavored relative to V3.

Chapter 3 3 8 Human Agreement Falls Sharply on Closely Matched Stories

Average Krippendorff's alpha across the six judgments was only 0.070. Raw pairwise agreement averaged 53.8 percent, only modestly above the 50 percent expected from two-option forced-choice judgments. Agreement varied somewhat across dimensions but was low throughout. Krippendorff's alpha was 0.037 for authenticity,

0.082 for emotion provocation, 0.185 for empathy, 0.061 for engagement, 0.018 for narrative complexity, and 0.038 for overall preference. Corresponding raw agreement rates were 52.5 percent for authenticity, 54.1 percent for emotion provocation, 59.1 percent for empathy, 53.3 percent for engagement, 52.1 percent for narrative complexity, and 51.9 percent for overall preference. Fleiss' kappa produced a very similar average of 0.068.

The low agreement was also present within each model comparison. Average Krippendorff's alpha across dimensions was approximately 0.063 for R1 versus V3, 0.062 for GPT-5 automatic versus GPT-4o, and 0.080 for GPT-5 high effort versus GPT-4o. Thus, the overall disagreement cannot be attributed to a single unusually difficult model pairing. This represents a dramatic drop from the approximately 0.72 average alpha reported in the original PDS study using the same rubric on larger-gap items. The frontier pairs are far closer in quality, and the low agreement may reflect limited rubric discrimination on close pairs, reader subjectivity, or noise.

Chapter 3 3 9 Reader Disagreement Shows Structure

Low aggregate agreement does not by itself establish that the human judgments are random. Several analyses suggest that at least some of the disagreement was structured. Reader-level logistic regressions examined how each participant's dimension-specific preferences related to that participant's overall story preference. A second analysis computed Cohen's kappa between each dimension and overall preference within each reader. These two methods produced similar patterns.

Readers 1 and 3 were especially similar to one another and tended to place comparatively greater weight on engagement and authenticity. Reader 2 was related to this cluster but less closely. Readers 4 and 7 formed another relatively similar pair and tended to assign positive weight across a broader collection of dimensions. Across readers, empathy and emotion provocation were the two dimensions with the most similar weighting patterns. Written justifications were also consistent with different readers prioritizing different properties. Some readers emphasized narrative clarity and coherence, while others placed greater weight on immediate emotional force, characterization, or other aspects of the PDS. This looks different from a situation in which all readers attend to the same features but respond inconsistently.

The sample contains only seven readers, so these patterns do not justify claims about stable reader archetypes in the population. They do, however, provide evidence that a portion of the disagreement reflects recurring individual differences in how narrative qualities are weighted. A sensitivity analysis reinforced the need for caution about the small reader sample. Under the crossed-random-intercept mixed model, eleven model-by-dimension comparisons met conventional significance thresholds. A cluster bootstrap resampling story pairs yielded ten, whereas resampling readers yielded five. The reduction under reader-level resampling reflects the uncertainty associated with having only seven participants. The substantive pattern is therefore better characterized through effect directions and intervals than through counts of statistically significant cells.

Chapter 3 3 10 Surface Properties and Automated Judge Scores

The automated evaluator's nearly uniform reasoning preference raised the question of what textual properties might be driving its scores. Several surface-level features were associated with the judge's ratings. Mean sentence length showed the strongest relationship. At the story level, mean sentence length correlated at approximately -0.32 with authenticity score, -0.43 with emotion provocation, -0.17 with empathy, -0.70 with engagement, and -0.09 with narrative complexity. The particularly large engagement association suggests that stories written in shorter sentences tended to receive considerably higher engagement ratings from the automated evaluator.

Lexical diversity showed another systematic relationship. MATTR correlated negatively with all five judge dimensions, with coefficients of approximately -0.37 for authenticity, -0.33 for emotion provocation, -0.35 for empathy, -0.29 for engagement, and -0.14 for narrative complexity. MTLD produced weaker negative associations for four dimensions and essentially no relationship with narrative complexity. Within-pair analyses designed to remove premise-level variation produced similar patterns. Mean sentence length had standardized coefficients of approximately -0.56 for authenticity, -0.47 for emotion provocation, -0.27 for empathy, -0.59 for engagement, and 0.03 for narrative complexity. Word count was also negatively associated with several judge dimensions, most strongly authenticity and engagement. MATTR remained negatively associated with authenticity, emotion provocation, empathy, and engagement.

Human overall preference also showed a relationship with sentence length. Across the 60 pairs, the difference in mean sentence length correlated at approximately -0.59 with preference for the reasoning-model story. Within the 40 GPT-family pairs, the corresponding correlation was approximately -0.55. The DeepSeek stories showed

relatively little pairwise sentence-length difference and no comparable association. These results should not be interpreted as evidence that shorter sentences themselves cause either psychological depth or human preference. The relevant features covary with many other stylistic properties, and the observations are dependent through shared premises and model families. They nevertheless indicate that the automated evaluator is sensitive to relatively superficial textual statistics.

Lastly, reasoning-model stories were also slightly more locally lexically uniform, with an average MATTR of 0.896 compared with 0.907 for comparison-model stories. At the same time, they were less predictable to Llama 3.1 70B. GPT-5 stories had an average perplexity of 16.3 compared with 9.5 for GPT-4o, while DeepSeek-R1 averaged 8.9 compared with 5.1 for V3. The evaluator's reasoning preference therefore cannot be reduced to a simple tendency to reward texts that are more familiar to the evaluator model.

Chapter 3 3 11 Development-Set Agreement Does Not Guarantee Deployment Validity

The main methodological result of the follow-up study is the divergence between development-set performance and deployment behavior. The routed evaluator reached a Spearman correlation of 0.646 with human PDS annotations on its selection distribution. On the frontier story pairs, however, it generated an almost uniformly pro-reasoning signal. Human readers did not. The judge preferred reasoning outputs on 89.0 percent of dimension-level comparisons and on 59 of 60 aggregate-PDS pairings, while human overall preferences differed by model family and averaged only 55.2 percent in favor of reasoning outputs. The mismatch, however, should not be

interpreted as a clean demonstration that the evaluator "failed" in an absolute sense, because the target itself changed.

The original PDS dataset used scalar ratings and contained larger quality differences. The follow-up uses forced-choice comparisons among much more closely matched stories. The selection-set correlation was also computed on data used to choose the evaluator configurations. The study instead illustrates why a development-set correlation should be understood as contextual evidence. An evaluator that predicts human ratings relatively well under one data-generating process can behave differently after the distribution of outputs changes. Furthermore, agreement among automated judges does not solve the problem. All five candidate configurations favored reasoning outputs, often more consistently than human readers agreed with one another. High inter-judge consistency could therefore have been interpreted as confidence even though the judge ensemble did not preserve the distribution of human judgments.

The surface-feature analysis provides a possible explanation. Reasoning-oriented models produced systematic stylistic differences, including shorter sentences and slightly different lexical-diversity profiles. Because judge scores were associated with these properties, a model can acquire a strong apparent advantage if its generation style happens to align with features rewarded by the evaluator. Such a relationship does not require the evaluator to have been explicitly trained to prefer "reasoning models." The broader implication is that LLM-as-a-Judge validation should be target matched. Evaluation reports should describe how a judge was selected, the distribution on which its human agreement was measured, the distribution on which it will subsequently be

used, and any known sensitivity to incidental features. When the deployment distribution changes substantially, a human check on that new distribution is warranted.

Chapter 3 3 12 Limitations

This study is exploratory. It uses twenty premises, one retained generation per condition, seven readers from one university, and one domain. Forced choice without ties or confidence may turn indifference into preference; limited PDS discrimination on close pairs, fatigue, and noise may all contribute to low agreement. The human target changed and the surface results are correlational. Broader cohorts, repeated generations, target-matched judges, and out-of-fold validation are needed.

Chapter 3 4 Conclusion

The two studies in this chapter examine psychological depth from complementary directions.

The first study develops and empirically validates the Psychological Depth Scale as a reader-centered framework for evaluating authenticity, empathy, engagement, emotion provocation, and narrative complexity in human- and LLM-generated stories. Human annotators applied the five dimensions with substantial agreement, with an average Krippendorff's alpha of 0.72. The study also examined whether PDS evaluation could be automated, finding that Mixture-of-Personas prompting improved agreement between LLM evaluators and human judgments, with GPT-4o achieving the highest average correlation of approximately 0.51. In the story-generation comparison, GPT-4 produced narratives that were statistically stronger than highly rated Reddit stories on empathy and narrative complexity and statistically indistinguishable from them on the remaining PDS dimensions.

The second study evaluates psychological depth in prompt-matched stories generated by GPT-5, GPT-4o, DeepSeek-R1, and DeepSeek-V3, while also examining the validity of automated PDS evaluation on these frontier-model outputs. The automated evaluator strongly favored reasoning-model outputs, selecting them in 89.0 percent of dimension-level comparisons and in 59 of 60 pairs on aggregate PDS. Human judgments were substantially less uniform: GPT-5 was modestly preferred to GPT-4o, whereas DeepSeek-R1 trailed DeepSeek-V3 overall, and inter-reader agreement was low. The study also found structured differences in how readers weighted PDS dimensions and associations between automated judge scores and surface properties such as sentence length and lexical diversity. These findings motivate target-matched human validation and independent evaluation before an automated judge is reused on a new distribution.

Chapter 3 5 References

- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5), 997–1013.
- Anthropic. (2023). *Introducing Claude*.
- Babae, R., & Montashery, I. (2012). Stanley Fish with respect to the reader. *International Journal of English and Literature*, 3, 18–24.
- Baer, J. (2017). Why you are probably more creative (and less creative) than you think. In *The creative self* (pp. 259–273). Elsevier.
- Baer, J., & Kaufman, J. C. (2019). Assessing creativity with the consensual assessment technique. In *The Palgrave handbook of social creativity research* (pp. 27–37). Palgrave Macmillan.

- Baer, J., & McKool, S. S. (2014). The gold standard for assessing creativity. *International Journal of Quality Assurance in Engineering and Technology Education (IJQAETE)*, 3(1), 81–93.
- Bal, P. M., & Veltkamp, M. (2013). How does fiction reading influence empathy? An experimental investigation on the role of emotional transportation. *PLOS ONE*, 8(1), e55341.
- Barrett, L. F., Lewis, M., & Haviland-Jones, J. M. (Eds.). (2016). *Handbook of emotions* (4th ed.). Guilford Press.
- Berlin, I. (2000). *Three critics of the Enlightenment: Vico, Hamann, Herder*. Princeton University Press.
- Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D. M., Wu, J., Winter, C., ... Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, 33 (pp. 1877–1901).
- Bsharat, S. M., Myrzakhan, A., & Shen, Z. (2023). *Principled instructions are all you need for questioning LLaMA-1/2, GPT-3.5/4* [Preprint]. arXiv.
- Busselle, R., & Bilandzic, H. (2008). Fictionality and perceived realism in experiencing stories: A model of narrative comprehension and engagement. *Communication Theory*, 18(2), 255–280.
- Busselle, R., & Bilandzic, H. (2009). Measuring narrative engagement. *Media Psychology*, 12(4), 321–347.

- Canning, P. (2017). Text world theory and real world readers: From literature to life in a Belfast prison. *Language and Literature*, 26(2), 172–187.
- Carson, S. H. (2019). Creativity and mental illness. In J. C. Kaufman & R. J. Sternberg (Eds.), *The Cambridge handbook of creativity* (2nd ed., pp. 296–318). Cambridge University Press.
- Chakrabarty, T., Laban, P., Agarwal, D., Muresan, S., & Wu, C.-S. (2024). Art or artifice? Large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*. Association for Computing Machinery.
- Chan, L. T.-H. (2016). Reader response and reception theory. In C. V. Angelelli & B. J. Baer (Eds.), *Researching translation and interpreting* (pp. 146–154). Taylor & Francis.
- Chase, H. (2022). *LangChain* [Computer software].
- Chen, C., & Shu, K. (2024). Can LLM-generated misinformation be detected? In *The Twelfth International Conference on Learning Representations*.
- Chiang, W.-L., Li, Z., Lin, Z., Sheng, Y., Wu, Z., Zhang, H., Zheng, L., Zhuang, S., Zhuang, Y., Gonzalez, J. E., Stoica, I., & Xing, E. P. (2023). *Vicuna: An open-source chatbot impressing GPT-4 with 90% ChatGPT quality*. LMSYS Org.
- Citron, F., Gray, M., Critchley, H., Weekes, B., & Ferstl, E. (2014). Emotional valence and arousal affect reading in an interactive way: Neuroimaging evidence for an approach-withdrawal framework. *Neuropsychologia*, 56, 79–89.
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All that's "human" is not gold: Evaluating human evaluation of generated text. In

- Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 7282–7296). Association for Computational Linguistics.
- Clark-Polner, E., Wager, T. D., Satpute, A. B., & Barrett, L. F. (2016). Neural fingerprinting: Meta-analysis, variation, and the search for brain-based essences in the science of emotion. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 146–165). Guilford Press.
- Cutting, J. E. (2019). Simplicity, complexity, and narration in popular movies. In M. Grishakova & M. Poulaki (Eds.), *Narrative complexity: Cognition, embodiment, evolution* (pp. 200–222). University of Nebraska Press.
- Davis, M. H. (1980). *A multidimensional approach to individual differences in empathy. JSAS Catalog of Selected Documents in Psychology*, 10, 85.
- Davis, M. H. (1994). *Empathy: A social psychological approach*. Brown & Benchmark.
- DeepSeek-AI. (2025). *DeepSeek-V3 technical report* [Technical report]. arXiv.
- Empson, W. (1947). *The structure of complex words*. Chatto & Windus.
- Fish, S. (1970). Literature in the reader: Affective stylistics. *New Literary History*, 2(1), 123–162.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Forster, M. N. (2022). Johann Gottfried von Herder. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*.
- Gavins, J. (2007). *Text world theory: An introduction*. Edinburgh University Press.

- Gehrmann, S., Adewumi, T., Aggarwal, K., Ammanamanchi, P. S., Anuoluwapo, A., Bosselut, A., Chandu, K. R., Clinciu, M.-A., Das, D., Dhole, K. D., Du, W., Durmus, E., Dušek, O., Emezue, C. C., Gangal, V., Garbacea, C., Hashimoto, T., Hou, Y., Jernite, Y., ... Zhou, J. (2021). The GEM benchmark: Natural language generation, its evaluation and metrics. In *Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics (GEM 2021)* (pp. 96–120). Association for Computational Linguistics.
- Gerdes, K. E., Segal, E. A., & Lietz, C. A. (2010). Conceptualising and measuring empathy. *British Journal of Social Work*, 40(7), 2326–2343.
- Gerrig, R. (1993). *Experiencing narrative worlds*. Routledge.
- Ghazarian, S., Weischedel, R., Galstyan, A., & Peng, N. (2020). Predictive engagement: An efficient metric for automatic evaluation of open-domain dialogue systems. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05), 7789–7796.
- Goffman, E. (1961). Fun in games. In *Encounters: Two studies in the sociology of interaction* (pp. 15–81). Bobbs-Merrill.
- Goldfarb-Tarrant, S., Feng, H., & Peng, N. (2019). Plan, write, and revise: An interactive system for open-domain story generation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)* (pp. 89–97). Association for Computational Linguistics.

- Grant, A. M. (2008). Does intrinsic motivation fuel the prosocial fire? Motivational synergy in predicting persistence, performance, and productivity. *Journal of Applied Psychology*, 93(1), 48–58.
- Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., Yang, A., Fan, A., Goyal, A., Hartshorn, A., Yang, A., Mitra, A., Sravankumar, A., Korenev, A., Hinsvark, A., ... Ma, Z. (2024). *The Llama 3 herd of models* [Preprint]. arXiv.
- Green, M. C. (2004). Transportation into narrative worlds: The role of prior knowledge and perceived realism. *Discourse Processes*, 38(2), 247–266.
- Green, M. C., Kass, S., Carrey, J., Feeney, R., Herzig, B., & Sabini, J. (2008). Transportation across media: Print versus film comparisons. *Media Psychology*, 11(4), 512–539.
- Green, M., & Brock, T. (2000). The role of transportation in the persuasiveness of public narrative. *Journal of Personality and Social Psychology*, 79, 701–721.
- Guidance AI. (2022). *Guidance* [Computer software]. GitHub.
- Guo, D., Yang, D., Zhang, H., Song, J., Wang, P., Zhu, Q., Xu, R., Zhang, R., Ma, S., Bi, X., Zhang, X., Yu, X., Wu, Y., Wu, Z. F., Gou, Z., Shao, Z., Li, Z., Gao, Z., Liu, A., ... Zhang, Z. (2025). DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081), 633–638.
- Gómez-Rodríguez, C., & Williams, P. (2023). A confederacy of models: A comprehensive evaluation of LLMs on creative writing. In *Findings of the Association for Computational Linguistics: EMNLP 2023* (pp. 14504–14528). Association for Computational Linguistics.

- Hall, A. (2003). Reading realism: Audiences' evaluations of the reality of media texts. *Journal of Communication*, 53, 624–641.
- Harel-Canada, F. Y., Zhou, H., Muppalla, S., Yildiz, Z. S., Kim, M., Sahai, A., & Peng, N. (2024). Measuring psychological depth in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17162–17196). Association for Computational Linguistics.
- Heidegger, M. (1962). *Being and time* (J. Macquarrie & E. Robinson, Trans.). Harper & Row. (Original work published 1927)
- Hoffman, M. L. (1991). Empathy, social cognition, and moral action. In W. M. Kurtines & J. L. Gewirtz (Eds.), *Handbook of moral behavior and development* (Vol. 1, pp. 275–301). Lawrence Erlbaum Associates.
- Holland, N. N. (1989). *The dynamics of literary response*. Columbia University Press.
- Jeong, H., Park, C., Hong, J., Lee, H., & Choo, J. (2025). The comparative trap: Pairwise comparisons amplifies biased preferences of LLM evaluators. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP* (pp. 79–108). Association for Computational Linguistics.
- Jiang, A. Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D. S., de las Casas, D., Hanna, E. B., Bressand, F., Lengyel, G., Bour, G., Lample, G., Lavaud, L. R., Saulnier, L., Lachaux, M.-A., Stock, P., Subramanian, S., Yang, S., ... El Sayed, W. (2024). *Mixtral of experts* [Preprint]. arXiv.

- Johnson, D. R. (2012). Transportation into a story increases empathy, prosocial behavior, and perceptual bias toward fearful expressions. *Personality and Individual Differences*, 52(2), 150–155.
- Johnson-Laird, P. N., & Oatley, K. (2022). How poetry evokes emotions. *Acta Psychologica*, 224, 103506.
- Kaniss, P. (1991). *Making local news*. University of Chicago Press.
- Kaufman, G., & Libby, L. (2012). Changing beliefs and behavior through experience-taking. *Journal of Personality and Social Psychology*, 103, 1–19.
- Ke, P., Wen, B., Feng, A., Liu, X., Lei, X., Cheng, J., Wang, S., Zeng, A., Dong, Y., Wang, H., Tang, J., & Huang, M. (2024). CritiqueLLM: Towards an informative critique generation model for evaluation of large language model generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 13034–13054). Association for Computational Linguistics.
- Khattab, O., Singhvi, A., Maheshwari, P., Zhang, Z., Santhanam, K., Vardhamanan, S., Haq, S., Sharma, A., Joshi, T. T., Moazam, H., Miller, H., Zaharia, M., & Potts, C. (2023). *DSPy: Compiling declarative language model calls into self-improving pipelines* [Preprint]. arXiv.
- Kierkegaard, S. (1980). *The sickness unto death: A Christian psychological exposition for upbuilding and awakening* (H. V. Hong & E. H. Hong, Trans.). Princeton University Press. (Original work published 1849)

- Kinnick, K. N., Krugman, D. M., & Cameron, G. T. (1996). Compassion fatigue: Communication and burnout toward social problems. *Journalism & Mass Communication Quarterly*, 73(3), 687–707.
- Kirk, H. R., Whitefield, A., Röttger, P., Bean, A., Margatina, K., Ciro, J., Mosquera, R., Bartolo, M., Williams, A., He, H., Vidgen, B., & Hale, S. A. (2024). The PRISM alignment dataset: What participatory, representative and individualised human feedback reveals about the subjective and multicultural alignment of large language models. In *Advances in Neural Information Processing Systems*, 37 (pp. 105236–105344).
- Kiss, M., & Willemsen, S. (2017). *Impossible puzzle films: A cognitive approach to contemporary complex cinema*. Edinburgh University Press.
- Krippendorff, K. (2011). *Computing Krippendorff's alpha-reliability*. University of Pennsylvania.
- Kunjanman, S., & Abdul Aziz, A. (2021). Reader-response theory: A systematic literature review. *Malaysian Journal of Social Sciences and Humanities (MJSSH)*, 6(4), 252–260.
- Kushneruk, S., Afanasyeva, O., Kurochkina, M., & Mineeva, M. (2017). Metaphors in the text-worlds of commercial advertising. In *Proceedings of the 4th International Multidisciplinary Scientific Conference on Social Sciences and Arts SGEM 2017* (Book 3, Vol. 1, pp. 133–140).
- Langer, S. K. (1942). *Philosophy in a new key: A study in the symbolism of reason, rite, and art*. Harvard University Press.

- Lazarus, R. S. (1991). Cognition and motivation in emotion. *American Psychologist*, 46(4), 352–367.
- Lee, M., Liang, P., & Yang, Q. (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*. Association for Computing Machinery.
- Li, C., Wang, J., Zhang, Y., Zhu, K., Hou, W., Lian, J., Luo, F., Yang, Q., & Xie, X. (2023). *Large language models understand and can be enhanced by emotional stimuli* [Preprint]. arXiv.
- Li, D., Jiang, B., Huang, L., Beigi, A., Zhao, C., Tan, Z., Bhattacharjee, A., Jiang, Y., Chen, C., Wu, T., Shu, K., Cheng, L., & Liu, H. (2025). From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 2757–2791). Association for Computational Linguistics.
- Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 2511–2522). Association for Computational Linguistics.
- Liu, Y., Su, Y., Shareghi, E., & Collier, N. (2024). Unlocking structure measuring: Introducing PDD, an automatic metric for positional discourse coherence. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*

- (*Volume 2: Short Papers*) (pp. 92–100). Association for Computational Linguistics.
- Liu, Y., Zhou, H., Guo, Z., Shareghi, E., Vulić, I., Korhonen, A., & Collier, N. (2025). *Aligning with human judgement: The role of pairwise preference in large language model evaluators* [Preprint]. arXiv.
- Lu, L.-C., Liu, M., Lu, P. C., Tian, Y., Sun, S.-H., & Peng, N. (2026). Rethinking creativity evaluation: A critical analysis of existing creativity evaluations. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 6329–6352). Association for Computational Linguistics.
- Ma, Y., Nguyen, K. L., Xing, F. Z., & Cambria, E. (2020). A survey on empathetic dialogue systems. *Information Fusion*, 64, 50–70.
- Maier, S. R. (2015). Compassion fatigue and the elusive quest for journalistic impact: A content and reader-metrics analysis assessing audience response. *Journalism & Mass Communication Quarterly*, 92(3), 700–722.
- Maier, S. R., Slovic, P., & Mayorga, M. (2017). Reader reaction to news of mass suffering: Assessing the influence of story form and emotional response. *Journalism*, 18(8), 1011–1029.
- Mailloux, S. (1976). Stanley Fish's "Interpreting the Variorum": Advance or retreat? *Critical Inquiry*, 3(1), 183–190.
- Mar, R. A., Oatley, K., Djikic, M., & Mullin, J. (2011). Emotion and narrative fiction: Interactive influences before, during, and after reading. *Cognition & Emotion*, 25(5), 818–833.

- Mar, R. A., Oatley, K., Hirsh, J., Dela Paz, J., & Peterson, J. B. (2006). Bookworms versus nerds: Exposure to fiction versus non-fiction, divergent associations with social ability, and the simulation of fictional social worlds. *Journal of Research in Personality, 40*(5), 694–712.
- Megalakaki, O., Ballenghein, U., & Baccino, T. (2019). Effects of valence and emotional intensity on the comprehension and memorization of texts. *Frontiers in Psychology, 10*, 179.
- Mendes, W. B. (2016). Emotion and the autonomic nervous system. In L. F. Barrett, M. Lewis, & J. M. Haviland-Jones (Eds.), *Handbook of emotions* (4th ed., pp. 166–181). Guilford Press.
- Miall, D. S., & Kuiken, D. (2001). Shifting perspectives: Readers' feelings and literary response. In W. van Peer & S. Chatman (Eds.), *New perspectives on narrative perspective* (pp. 289–301). SUNY Press.
- Neal, T., Sundararajan, K., Fatima, A., Yan, Y., Xiang, Y., & Woodard, D. (2017). Surveying stylometry techniques and applications. *ACM Computing Surveys, 50*(6), Article 86.
- Newman, G. E., & Smith, R. K. (2016). Kinds of authenticity. *Philosophy Compass, 11*(10), 609–618.
- Nussbaum, M. C. (2004). Emotions as judgments of value and importance. In R. C. Solomon (Ed.), *Thinking about feeling: Contemporary philosophers on emotions* (pp. 183–199). Oxford University Press.

- Oatley, K. (2002). Emotions and the story worlds of fiction. In M. C. Green, J. J. Strange, & T. C. Brock (Eds.), *Narrative impact: Social and cognitive foundations* (pp. 39–69). Lawrence Erlbaum Associates.
- OpenAI. (2023). *GPT-4 technical report* [Technical report]. arXiv.
- OpenAI. (2024a). *GPT-4o system card* [Technical report]. arXiv.
- OpenAI. (2024b). *OpenAI o1 system card* [Technical report]. arXiv.
- Ortony, A., Clore, G. L., & Collins, A. (2022). *The cognitive structure of emotions*. Cambridge University Press.
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, 37 (pp. 68772–68802).
- Petraglia, J. (2009). The importance of being authentic: Persuasion, narration, and dialogue in health communication and education. *Health Communication*, 24(2), 176–185.
- Pryzant, R., Iter, D., Li, J., Lee, Y. T., Zhu, C., & Zeng, M. (2023). Automatic prompt optimization with “gradient descent” and beam search. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing* (pp. 7957–7968). Association for Computational Linguistics.
- Qwen Team. (2026, February). *Qwen3.5: Accelerating productivity with native multimodal agents*.
- Rajani, N., Lambert, N., Han, S., Wang, J., Nitski, O., Beeching, E., Turnstall, L., Launay, J., & Wolf, T. (2023). *Can foundation models label data like humans?* Hugging Face Blog.

- Rousseau, J.-J. (2016). A discourse on the origin of inequality. In *Democracy: A reader* (pp. 127–129). Columbia University Press. (Original work published 1755)
- Sabour, S., Zheng, C., & Huang, M. (2022). CEM: Commonsense-aware empathetic response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 36 (pp. 11229–11237).
- Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., & Akata, Z. (2023). In-context impersonation reveals large language models’ strengths and biases. In *Advances in Neural Information Processing Systems*, 36 (pp. 72044–72057).
- Sandri, M., Leonardelli, E., Tonelli, S., & Jezek, E. (2023). Why don’t you do it right? Analysing annotators’ disagreement in subjective tasks. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2428–2441). Association for Computational Linguistics.
- Sartre, J.-P. (1971). *Sketch for a theory of the emotions*. Routledge. (Original work published 1939)
- Scott, L. M. (1994). The bridge from text to mind: Adapting reader-response theory to consumer research. *Journal of Consumer Research*, 21(3), 461–480.
- Singer, T., Seymour, B., O’Doherty, J. P., Stephan, K. E., Dolan, R. J., & Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439, 466–469.
- Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A. J., Ananthram, A., Nathan, A., Luo, A., Helyar, A., Madry, A., Efremov, A., Spyra, A., Baker-Whitcomb, A., Beutel, A., Karpenko, A., ... Wang, Z. (2025). *OpenAI GPT-5 system card* [Technical report]. arXiv.

- Slovic, P. (2007). “If I look at the mass I will never act”: Psychic numbing and genocide. *Judgment and Decision Making*, 2(2), 79–95.
- Snell, C., Lee, J., Xu, K., & Kumar, A. (2024). *Scaling LLM test-time compute optimally can be more effective than scaling model parameters* [Preprint]. arXiv.
- Somasundaran, S., Flor, M., Chodorow, M., Molloy, H., Gyawali, B., & McCulla, L. (2018). Towards evaluating narrative quality in student writing. *Transactions of the Association for Computational Linguistics*, 6, 91–106.
- Spangher, A., Peng, N., Gehrmann, S., & Dredze, M. (2024). Do LLMs plan like human writers? Comparing journalist coverage of press releases with LLMs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 21814–21828). Association for Computational Linguistics.
- Speer, N. K., Reynolds, J. R., Swallow, K. M., & Zacks, J. M. (2009). Reading stories activates neural representations of visual and motor experiences. *Psychological Science*, 20(8), 989–999.
- Steiner, G. (1978). On difficulty. *The Journal of Aesthetics and Art Criticism*, 36(3), 263–276.
- Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). *Large language models are inconsistent and biased evaluators* [Preprint]. arXiv.
- Su, H., Kasai, J., Wu, C. H., Shi, W., Wang, T., Xin, J., Zhang, R., Ostendorf, M., Zettlemoyer, L., Smith, N. A., & Yu, T. (2022). *Selective annotation makes language models better few-shot learners* [Preprint]. arXiv.
- Taylor, C. (1991). *The ethics of authenticity*. Harvard University Press.

- Tian, L., & Wang, H. (2019). Towards a text-world approach to translation and its pedagogical implications. *International Journal of Translation, Interpretation, and Applied Linguistics*, 1, 14–26.
- Tian, Y., Huang, T., Liu, M., Jiang, D., Spangher, A., Chen, M., May, J., & Peng, N. (2024). Are large language models capable of generating human-level narratives? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 17659–17681). Association for Computational Linguistics.
- Tikka, P., & Kaipainen, M. (2017). Intersubjectivity, idiosyncrasy and narrative deixis: A neurocinematic approach. In M. Poulaki & M. Grishakova (Eds.), *Narrative complexity and media: Experiential and cognitive interfaces*. University of Nebraska Press.
- Tkachenko, M., Malyuk, M., Holmanyuk, A., & Liubimov, N. (2020–2025). *Label Studio: Data labeling software* [Computer software]. HumanSignal.
- Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., Bikel, D., Blecher, L., Canton Ferrer, C., Chen, M., Cucurull, G., Esiobu, D., Fernandes, J., Fu, J., Fu, W., ... Scialom, T. (2023). *Llama 2: Open foundation and fine-tuned chat models* [Preprint]. arXiv.
- Uddin, L. Q., Iacoboni, M., Lange, C., & Keenan, J. P. (2007). The self and social cognition: The role of cortical midline structures and mirror neurons. *Trends in Cognitive Sciences*, 11(4), 153–157.
- Vaezi, M., & Rezaei, S. (2019). Development of a rubric for evaluating creative writing: A multiphase research. *New Writing*, 16(3), 303–317.

- Verma, V., Fleisig, E., Tomlin, N., & Klein, D. (2024). Ghostbuster: Detecting text ghostwritten by large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 1702–1717). Association for Computational Linguistics.
- Wang, B., Chen, W., Pei, H., Xie, C., Kang, M., Zhang, C., Xu, C., Xiong, Z., Dutta, R., Schaeffer, R., Truong, S. T., Arora, S., Mazeika, M., Hendrycks, D., Lin, Z., Ho, E., Chen, X., Song, Z., & Koyejo, S. (2023). DecodingTrust: A comprehensive assessment of trustworthiness in GPT models. In *Advances in Neural Information Processing Systems*, 36.
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Kong, L., Liu, Q., Liu, T., & Sui, Z. (2024). Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 9440–9450). Association for Computational Linguistics.
- Wang, T., Chen, J., Jia, Q., Wang, S., Fang, R., Wang, H., Gao, Z., Xie, C., Xu, C., Dai, J., Liu, Y., Wu, J., Ding, S., Li, L., Huang, Z., Deng, X., Yu, T., Ma, G., Xiao, H., ... Zhou, W. (2024). *Weaver: Foundation models for creative writing* [Preprint]. arXiv.
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q. V., & Zhou, D. (2022). Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*, 35 (pp. 24824–24837).

- Woodruff, A. H., & Griffin, R. A. (2017). Reader response in secondary settings: Increasing comprehension through meaningful interactions with literary texts. *Texas Journal of Literacy Education*, 5(2), 108–116.
- Xu, G., Liu, R., Harel-Canada, F., Chandra, N. R., & Peng, N. (2022). EnDex: Evaluation of dialogue engagingness at scale. In *Findings of the Association for Computational Linguistics: EMNLP 2022* (pp. 4884–4893). Association for Computational Linguistics.
- Yang, K., Tian, Y., Peng, N., & Klein, D. (2022). Re3: Generating longer stories with recursive reprompting and revision. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing* (pp. 4393–4479). Association for Computational Linguistics.
- Yao, L., Peng, N., Weischedel, R., Knight, K., Zhao, D., & Yan, R. (2019). Plan-and-write: Towards better automatic storytelling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 33 (pp. 7378–7385).
- Yuan, A., Coenen, A., Reif, E., & Ippolito, D. (2022). Wordcraft: Story writing with large language models. In *Proceedings of the 27th International Conference on Intelligent User Interfaces* (pp. 841–852). Association for Computing Machinery.
- Zaki, J., & Ochsner, K. N. (2012). The neuroscience of empathy: Progress, pitfalls and promise. *Nature Neuroscience*, 15(5), 675–680.
- Zhao, Z., Song, S., Duah, B., Macbeth, J., Carter, S., Van, M. P., Bravo, N. S., Klenk, M., Sick, K., & Filipowicz, A. L. S. (2023). More human than human: LLM-generated narratives outperform human-LLM interleaved narratives. In

- Proceedings of the 15th Conference on Creativity and Cognition* (pp. 368–370). Association for Computing Machinery.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, 36 (pp. 46595–46623).
- Zhou, J., Zhang, Y., Luo, Q., Parker, A. G., & De Choudhury, M. (2023). Synthetic lies: Understanding AI-generated misinformation and evaluating algorithmic and human solutions. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Article 436, pp. 1–20). Association for Computing Machinery.
- Zhou, Y., Muresanu, A. I., Han, Z., Paster, K., Pitis, S., Chan, H., & Ba, J. (2023). *Large language models are human-level prompt engineers* [Conference paper]. International Conference on Learning Representations.
- Zunshine, L. (2006). *Why we read fiction: Theory of mind and the novel*. University of Chicago Press.
- Zwarun, L., & Hall, A. (2012). Narrative persuasion, transportation, and the role of need for cognition in online viewing of fantastical films. *Media Psychology*, 15(3), 327–355.