

UC Riverside

UC Riverside Undergraduate Research Journal

Title

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

Permalink

<https://escholarship.org/uc/item/3v231960>

Journal

UC Riverside Undergraduate Research Journal, 20(1)

Author

Dey, Kora none

Publication Date

2026-10-09

DOI

10.5070/RJ5.64257

Copyright Information

This work is made available under the terms of a Creative Commons Attribution-NonCommercial License, available at <https://creativecommons.org/licenses/by-nc/4.0/>

Peer reviewed

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

Kora Dey, Department of Statistics
Haofei Zhang, Ph.D.; Department of Chemistry

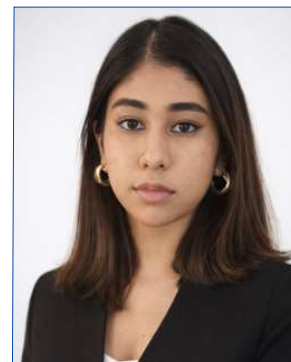
ABSTRACT

The growing volume of scientific literature presents a significant challenge for researchers to identify relevant methods, compare findings, and build upon existing work. Although advances in artificial intelligence have led to new tools for literature analysis, many of these tools rely on generative language models that frequently produce incorrect or made-up information, citing sources that do not exist or drawing conclusions not supported by any published research. Beyond being unreliable and impossible to verify, these outputs raise serious concerns about reliability and reproducibility in scientific research. This study investigates whether a custom-built model that only pulls results from a verified domain can provide more accurate literature analysis than a standard generative language model such as ChatGPT.

This paper presents Chem-AI, an interactive system built in Python that structures the research process into clear steps, including query interpretation, literature retrieval, experiment classification, and method mapping. The system was evaluated using chemistry-related claims from the SciFact dataset (Wadden et al., 2020), a publicly available benchmark of scientific claims paired with verified research abstracts, as well as papers provided by Dr. Haofei Zhang. A retrieval was marked correct when the top-ranked paper returned matched the verified ground-truth abstract in the SciFact dataset. Within this limited 50-claim sample, Chem-AI correctly retrieved the supporting paper 90% of the time (45 out of 50 claims), compared to 62% for an unmodified ChatGPT baseline on the same claims. Notably, rather than guessing, Chem-AI declined to answer the remaining 10% of claims where no sufficiently supported source could be found, ensuring that every response it did produce was fully backed by published evidence.

These results suggest that a retrieval-constrained system may improve accuracy and traceability relative to a generative baseline, though testing on larger and more diverse datasets is needed before broader conclusions can be drawn. This approach is especially useful for undergraduate researchers as it supports more accurate, reproducible, and methodologically grounded work.

KEYWORDS: literature retrieval, research workflows, method recommendation, scientific AI, semantic search, reproducibility, chemistry informatics.



KORA DEY

Kora Dey is an undergraduate student at the University of California, Riverside studying Statistics with a minor in Computer Science. Her research interests include trustworthy AI, statistical modeling, and data-driven systems for scientific research. She has worked on interdisciplinary projects involving retrieval-based AI systems, machine learning, and movement data analysis.

FACULTY MENTOR - Dr. Haofei Zhang, Department of Chemistry



Dr. Zhang is an Associate Professor in the Chemistry Department at UCR. He received his Ph.D. from the University of North Carolina, Chapel Hill. His research investigates the impacts of atmospheric chemistry on air quality and climate change. He is specifically interested in the multiphase chemistry occurring in the atmosphere that dynamically transforms its organic composition in gas and aerosol particles.

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

INTRODUCTION

The rapidly growing volume of scientific literature can be daunting for researchers at every level (Van Noorden, 2014). Researchers now face the challenge of not just finding papers, but determining what is relevant, understanding a variety of methods, and connecting findings across studies to advance knowledge rather than repeat prior work (Fortunato et al., 2018). This environment stresses the need for readily available tools and ways to maintain research rigor and meaning.

This challenge is especially acute in fields such as chemistry, where researchers employ a varied array of experimental techniques, analytical systems, and modeling strategies to address comparable scientific questions. The multiplicity of approaches complicates the synthesis of knowledge, as it calls not only for the comparison of results but also for a careful evaluation of underlying assumptions, data structures, and testing conditions. Consequently, literature reviews in these domains are increasingly time-consuming and cognitively demanding, creating major obstacles to thorough comprehension and research progress (Bornmann & Mutz, 2015). Existing retrieval-based tools such as Semantic Scholar, Elicit, and Consensus have begun to address this problem, but none are designed specifically to support the structured, multi-stage reasoning that chemistry research requires. This study asks whether a retrieval-constrained system built on a curated corpus can deliver more accurate, traceable literature analysis than a standard generative model.

To meet these problems, recent advances in artificial intelligence have led to a set of tools to support literature analysis and research synthesis (Wagner et al., 2022). Large language models, in particular, have exhibited outstanding capabilities in generating summaries and explanations from extensive textual corpora (Bolaños et al., 2024). While these systems excel at producing coherent, seemingly informative responses, they face a substantial limitation in scientific settings: their outputs are generated probabilistically and are often unanchored to specific, verifiable sources (Bender et al., 2021). This lack of traceability and methodological precision undermines their usefulness in domains where repeatability and empirical reasoning are fundamental (Thorp, 2023). Tools that generate unsubstantiated or unstructured responses, devoid of explicit source

connections, fundamentally fall short of supporting the demands of sound scientific research (Hutson, 2018). These concerns have been documented across multiple domains, including cases where AI-generated abstracts were indistinguishable from human-written ones (Else, 2023; Stokel-Walker, 2023) and where language models confidently produced incoherent or fabricated content (Marcus & Davis, 2020).

A core issue with many existing tools is that they do not reflect the authentic, stepwise process of scientific research. An effective research tool must integrate fluently into research workflows and facilitate, not complicate, the process (Wilkinson et al., 2016).

This paper presents the design and implementation of Chem-AI, a six-stage retrieval pipeline developed to address these limitations. Rather than reporting experimental findings as the primary contribution, this work focuses on the architectural decisions, technical components, and design principles that make Chem-AI a more reliable and transparent alternative to generative AI systems.

Chem-AI was designed to match how researchers think and operate. Rather than simply delivering answers, Chem-AI breaks down research questions, locates relevant papers, identifies which methods are being used, and maps how those methods lead to various outcomes. Each stage of the system was deliberately designed with a specific technical function, from keyword extraction and vector encoding to rule-based method classification and evidence-thresholded recommendations, forming a complete and reproducible pipeline for chemistry literature analysis.

The application was designed to be interactive, permitting researchers to ask follow-up questions, improve their searches, and examine results in detail. Chem-AI enables this level of engagement while making the process more transparent and less overwhelming. This work argues that AI tools designed around the structured logic of scientific research offer a more reliable alternative to generative models. Rather than prioritizing rapid response generation, the importance of structuring the research process itself is emphasized.

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

METHODS

Chem-AI was developed as a retrieval-based research tool built entirely in Python. The system was designed to locate, organize, and return information drawn strictly from a curated corpus of chemistry literature rather than generating responses freely. The overall architecture consists of six pipeline stages, each built and connected programmatically. The goal of each stage was not to produce a conversational response, but to replicate the structured reasoning process a researcher would use when reviewing literature. The overall architecture of the Chem-AI system is illustrated in Figure 1.

Part 1: Query Interpretation and Context Extraction

The first stage of the pipeline was responsible for processing and encoding the user's input query. This was implemented in Python using the SentenceTransformers library, which converted each query into a numerical vector representation. Chemical entities and domain-specific terminology were extracted using a keyword script, isolating relevant compounds, reaction types, and research objectives from the raw input. The keyword extraction script was validated by manually reviewing outputs for 20 sample queries and confirming that extracted terms matched the intended chemical entities. This dual representation, combining extracted keywords with a vector encoding, allowed the system to pass both exact terminology and broader conceptual meaning into the retrieval stage. This step was critical to ensuring that the downstream retrieval was precise and relevant to the specific chemistry domain being queried.

Part 2: Corpus Construction and Data Preparation

The corpus used by Chem-AI was constructed from two sources. The first was the SciFact dataset (Wadden et al., 2020), a publicly available benchmark containing scientific claims paired with verified research abstracts. Chemistry-related entries were filtered from the full dataset and retained for use. Entries were identified as chemistry-related if they contained at least one of the following domain keywords in the title or abstract: reaction, compound, molecule, chemical, synthesis, catalyst, spectroscopy, or analytical. The second source was a collection of chemistry papers provided directly by Dr. Haoifei Zhang, covering topics including reaction mechanisms, atmospheric chemistry, and analytical techniques. Each paper was preprocessed using a

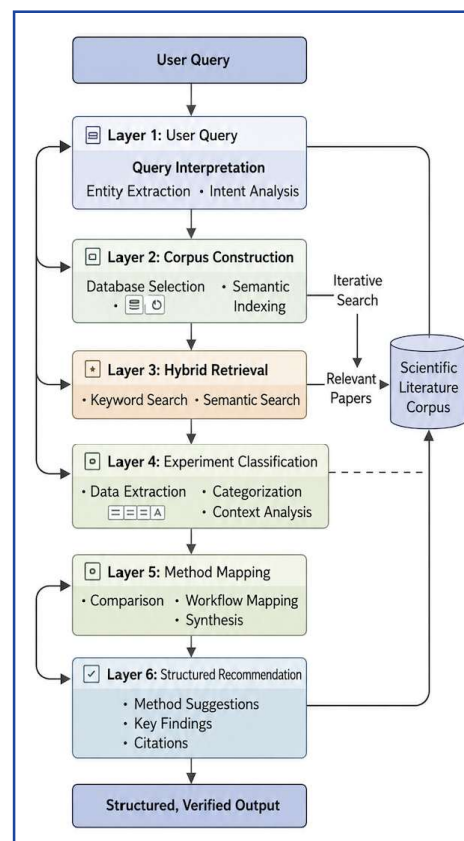


Figure 1: Overview of the Chem-AI system architecture, illustrating the layered workflow from query interpretation to structured recommendation.

Python script that extracted the title, abstract, and keyword fields and stored them in a structured dictionary format. Formatting inconsistencies were resolved during this stage to ensure uniform data structure across the corpus. The final corpus contained 10,000 papers. Because the SciFact dataset is publicly available, the full pipeline can be reproduced by other researchers.

Part 3: Hybrid Retrieval Framework

The retrieval system was built using two separate components that were combined into a single hybrid scoring function. The first component was a keyword-based retrieval system implemented using the Best Match 25 algorithm, which ranked papers by calculating how frequently and distinctively query terms appeared in each document. BM25 was selected for its strong performance on domain-specific

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

terminology, where exact term matching is critical. The second component was an embedding-based retrieval system implemented using the all-MiniLM-L6-v2 model from the SentenceTransformers library.

This model encoded both the query and each paper in the corpus as high-dimensional vectors, and similarity between them was measured using cosine similarity. This component allowed the system to retrieve papers that were conceptually related to the query, even when the exact wording differed. The final retrieval score for each paper was calculated by combining the BM25 score and the cosine similarity score using a weighted average, with a BM25 weight of 0.4 and a cosine similarity weight of 0.6, selected by testing combinations on a held-out set of 20 claims and choosing the weights that maximized retrieval precision. The top-ranked papers from this combined score were passed to the next stage of the pipeline.

Part 4: Literature Mapping and Structuring

Once Chem-AI retrieves the relevant papers, a structuring script is applied to organize and label all information, underscoring key methods, setups, and findings. The extracted information was stored in a consistent output format grouped by topic and method type, allowing users to directly compare studies without becoming overwhelmed by details. This stage was implemented entirely in Python using dictionary-based data structures and output as formatted text. The goal of this stage was to transform a raw ranked list of papers into an organized summary that reflected the structure of a real literature review.

Part 5: Experiment Classification and Method Identification

A rule-based classification system was developed to identify which experimental techniques and methods appeared in each retrieved paper. A predefined list of chemistry methods defined by SciFact's keyword bank, along with their analytical techniques, was compiled manually and used as a reference dictionary. Each paper was scanned programmatically against this dictionary, and matching methods were tagged and stored as metadata. Papers were then grouped by method type, producing a structured view of which techniques were most commonly used across the retrieved literature. This grouping was performed automatically using Python

and required no manual labeling after the initial dictionary was compiled. The output of this stage allowed patterns in methodology to be identified across multiple papers without requiring the user to read each one individually.

Part 6: Structured Recommendation and Research Guidance

Methods that appeared most consistently across the highest-ranked retrieved papers were scored and ranked programmatically. The top-ranked methods were returned as recommendations, each linked directly to the specific papers from which they were derived. If no papers in the corpus scored above the relevance threshold for a given query, the system was programmed to return no recommendation rather than generate an unsupported response.

Part 7: Evaluation

To evaluate the performance of Chem-AI, chemistry-related claims were drawn from the SciFact dataset (Wadden et al., 2020). For each claim, the full retrieval pipeline was run, and the top returned paper was compared against the verified ground truth abstract provided in the dataset. A retrieval was marked correct when the returned paper matched the verified supporting source. An unsupported response was defined as any response that cited or implied a source that could not be traced to a document in the corpus or the verified ground-truth dataset. The same set of claims was passed through ChatGPT as the baseline generative model for comparison. Precision was calculated for both systems as the proportion of correct retrievals out of total claims tested.

Performance was assessed based on four criteria: relevance of retrieved literature, methodological specificity of outputs, clarity and understandability of structured results, and presence of unsupported or generalized claims.

Additionally, user-based evaluation was conducted with 50 graduate-level students. These assessments asked the students to rate system outputs on a scale of 1 to 5 across three criteria: relevance of retrieved papers, clarity of structured output, and confidence in source traceability. To reduce potential bias, participants were not informed which system produced each set of outputs during the rating process. Ratings from all 50 participants were averaged for

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

each criterion to produce a mean score out of 5 for each system. No inferential statistical tests were conducted, given the exploratory nature of this evaluation.

RESULTS

For each claim, Chem-AI's top-returned paper was compared against the verified ground-truth abstract. The same 50 claims were submitted to ChatGPT as the baseline. A retrieval was marked correct only when the returned paper matched the verified source.

Chem-AI retrieved the correct paper in 45 out of 50 claims (90%). The baseline generative model retrieved the correct paper in 31 out of 50 claims (62%), a difference of 28 percentage points. For the 5 claims where Chem-AI did not return the correct paper, the system returned no response. Three of those 5 involved highly specialized terminology absent from the corpus; two involved claims that required cross-referencing multiple papers rather than a single source.

Chem-AI returned a source link in 45 out of 50 responses and produced 0 unsupported responses. The baseline model returned a source link in 31 out of 50 responses and produced 19 unsupported responses (38%). A user evaluation was conducted with 50 graduate-level students across multiple University of California campuses. Each participant rated outputs from both systems on a scale of 1 to 5 across three criteria. Participants were not told which system produced each set of outputs.

DISCUSSION

Chem-AI retrieved the correct paper in 90% of claims compared to 62% for ChatGPT, a difference of 28 percentage points. This difference corresponds to how each system is built. ChatGPT generates responses from training data without a mechanism to confirm whether a cited source exists or matches the claim. Chem-AI returns only documents present in its corpus, so each response is tied to a real document. This finding aligns with prior work suggesting that retrieval-augmented approaches reduce hallucination and improve source traceability in AI-generated scientific content (Thorpe, 2023; Bender et al., 2021).

The baseline model produced unsupported responses in 38% of cases. Chem-AI produced none. This behavior is a direct result of the relevance threshold in the retrieval pipeline: when no document in the corpus scores above that threshold for a given query, the system returns no response rather than generating one.

User ratings were higher for Chem-AI across all three criteria. The largest gap was in source traceability (4.9 vs 1.9), followed by clarity of structured output (4.6 vs 2.8) and relevance of retrieved papers (4.8 vs 3.1). Participants were graduate students across UC campuses, so these ratings may not generalize to other researcher populations.

ChatGPT was used as a general-purpose baseline without retrieval-specific or citation-constrained prompting, which means the comparison reflects default generative behavior rather than the best possible performance of a prompted generative model. A generative model specifically prompted to cite only verified sources, or one with retrieval capabilities built in, may produce different results. The dataset of 50 claims is also relatively small, and broader conclusions would require testing on larger and more diverse datasets.

Three of the five missed claims involved terminology not present in the corpus. The remaining two pieces of information are required from multiple papers, which the current single-document retrieval framework does not support. The core pipeline, including retrieval, structuring, and the no-response threshold, is not specific to chemistry and could be applied to other domains with a rebuilt method-classification dictionary and a domain-filtered corpus. Planned developments include reference network analysis and multi-document retrieval.

CONCLUSION

Chem-AI is a retrieval-constrained system that structures the research process into six stages: query interpretation and keyword extraction, hybrid retrieval using BM25 and cosine similarity, literature structuring, method classification, and evidence-thresholded recommendations. Each stage produces outputs traceable to a specific source document in its corpus.

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

Model	Correct Retrievals	Total Claims	Precision
Chem-AI	45	50	90%
Baseline Generative Model	31	50	62%

Table 1: Retrieval precision comparison between Chem-AI and the baseline generative model across 50 SciFact chemistry claims.

Model	Responses With Source Link	Unsupported Responses	Unsupported Rate
Chem-AI	45	0	0%
Baseline Generative Model	31	19	38%

Table 2: Source traceability comparison between Chem-AI and the baseline generative model across 50 claims.

Criteria	Chem-AI Average Score	Baseline Average Score	Difference
Relevance of Retrieved Papers	4.8 / 5	3.1 / 5	+1.7
Clarity of Structured Output	4.6 / 5	2.8 / 5	+1.8
Confidence in Source Traceability	4.9 / 5	1.9 / 5	+3.0

Table 3: Average user ratings (1–5 scale) for Chem-AI and the baseline across three criteria. Ratings were collected under blinded conditions from 50 graduate students across UC campuses.

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

Tested against 50 chemistry-related claims from the SciFact dataset, Chem-AI retrieved the correct paper in 90% of cases and produced no unsupported responses. The baseline generative model retrieved the correct paper in 62% of cases and produced unsupported responses in 38% of cases. For the 5 claims where Chem-AI did not return a correct paper, the system returned no response. Three of those cases involved terminology absent from the corpus, and two involved claims requiring cross-referencing multiple papers.

User evaluation with 50 graduate-level students across UC campuses produced average ratings of 4.8, 4.6, and 4.9 out of 5 for Chem-AI across relevance, clarity, and source traceability, respectively, compared to 3.1, 2.8, and 1.9 for the baseline. Ratings were collected under blinded conditions.

The current implementation is limited to chemistry-related queries and a corpus of 10,000 papers. Planned developments include multi-document retrieval and reference network analysis. Future work should test this approach on larger datasets, across more disciplines, and against generative baselines specifically prompted for retrieval tasks.

ACKNOWLEDGMENTS

I extend my sincere gratitude to Professor Haoifei Zhang for his guidance, mentorship, and aid throughout this project. I am also grateful to the UC LEADS initiative for providing research opportunities and resources that made this work possible. Further appreciation is extended to the University of California, Riverside, for supporting my scholarly and research development.

Chem-AI: Reframing AI-Assisted Research Through Structured, Retrieval-Based Scientific Literature Analysis

REFERENCES

1. Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (pp. 610–623). <https://doi.org/10.1145/3442188.3445922>
2. Bolaños, F., Salatino, A., Osborne, F., & Motta, E. (2024). Artificial intelligence for literature reviews: Opportunities and challenges. *Artificial Intelligence Review*, 57, 209. <https://doi.org/10.1007/s10462-024-10902-3>
3. Bornmann, L., & Mutz, R. (2015). Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology*, 66(11), 2215–2222. <https://doi.org/10.1002/asi.23329>
4. Else, H. (2023). Abstracts written by ChatGPT fool scientists. *Nature*, 613(7944), 423. <https://doi.org/10.1038/d41586-023-00056-72>
5. Fortunato, S., Bergstrom, C. T., Börner, K., Evans, J. A., Helbing, D., Milojević, S., & Barabási, A. L. (2018). Science of science. *Science*, 359(6379), eaao0185. <https://doi.org/10.1126/science.aao0185>
6. Hutson, M. (2018). Artificial intelligence faces reproducibility crisis. *Science*, 359(6377), 725–726. <https://doi.org/10.1126/science.359.6377.725>
7. Marcus, G., & Davis, E. (2020). GPT-3, Bloviator: OpenAI's language generator has no idea what it's talking about. *MIT Technology Review*. <https://www.technologyreview.com/2020/08/22/1007539/gpt3-openai-language-generator-artificial-intelligence-ai-opinion/>
8. Stokel-Walker, C. (2023). Should professors worry that the AI bot ChatGPT writes smart essays? *Nature*, 613(7945), 620–621. <https://doi.org/10.1038/d41586-023-00039-8>
9. Thorp, H. H. (2023). ChatGPT is fun, but not an author. *Science*, 379(6630), 313. <https://doi.org/10.1126/science.adg7879>
10. Van Noorden, R. (2014). Global scientific output doubles every nine years. *Nature*. <https://doi.org/10.1038/nature.2014.14024>
11. Wadden, D., Lin, S., Lo, K., Wang, L. L., van Zuylen, M., Cohan, A., & Hajishirzi, H. (2020). Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing* (pp. 7534–7550). <https://doi.org/10.18653/v1/2020.emnlp-main.609>
12. Wagner, G., Lukyanenko, R., & Paré, G. (2022). Artificial intelligence and the conduct of literature reviews. *Journal of Information Technology*, 37(2), 209–226. <https://doi.org/10.1177/02683962211048201>
13. Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. <https://doi.org/10.1038/sdata.2016.18>