Evaluating Large Language Model Literature Reviews In Interdisciplinary Science: A Systems Biology Perspective

Charvi Jain[1], Sahar Vahdati[1,2,7], Nandu Gopan[1,3,5], Ivo F. Sbalzarini[1,2,3,5,6] and Jens Lehmann[1,2,4,*]

[1]Dresden University of Technology, Faculty of Computer Science, Nöthnitzer Str. 46, 01187 Dresden, Germany
[2]Center for Scalable Data Analytics and Artificial Intelligence ScaDS.AI, Dresden/Leipzig, Germany
[3]Center for Systems Biology Dresden, Pfotenhauerstr. 108, 01307 Dresden, Germany
[4]Amazon, Germany
[5]Max Planck Institute of Molecular Cell Biology and Genetics, Pfotenhauerstr. 108, 01307 Dresden, Germany
[6]Cluster of Excellence Physics of Life, TU Dresden, Dresden, Germany
[7]Leibniz University Hanover, Welfengarten 1, 30167 Hannover Germany

Abstract

We evaluate the effectiveness of current large language model (LLM) literature review systems in interdisciplinary domains. While LLMs can support and accelerate reviewing the scientific literature, it is unclear how they cope with interdisciplinary science, where sources from multiple fields must be integrated according to relevance defined by context. We study this from the perspective of systems biology, a field that combines biology, mathematics, physics, and computer science. Using a set of expert-defined research questions, we assess the ability of LLMs to meaningfully integrate cross-domain knowledge and correctly reflect relevance. Specifically, we evaluate the quality of generated reports and the relevance of retrieved references from five different review models. We find that LLMs are a valuable augmentative tool for literature reviews, but trade off report quality for completeness in interdisciplinary domains. We address these limitations by proposing a novel method, termed AURORA, which is particularly designed for interdisciplinary applications. On the interdisciplinary systems biology benchmark, AURORA offers good coverage with high-quality reports.

Keywords

Literature Review, Large Language Models, Scientific Literature, Interdisciplinary Science, Systems Biology

1. Introduction

Scientific research usually starts by reviewing prior art from the literature. This involves extensive search across multiple sources, followed by critical evaluation of the relevance of each result and the semantic relations between them. This time-consuming and somewhat repetitive task can be accelerated and augmented with the help of LLMs, including generating the final written report. However, LLMs cannot directly be used for literature surveys, due to the risk of hallucinating nonexistent references or generating unverifiable reports. Instead, LLMs must be carefully integrated with literature databases into robust automated workflows for scientific literature review (SLR). This has been successfully practiced in several approaches, including Elicit [1], Scite [2, 3], and Undermind [4]. Such systems have demonstrated their potential for reducing workload, but their effectiveness in interdisciplinary research domains remains unclear.

Literature reviews in interdisciplinary fields require integrating sources from multiple disciplines in a contextually meaningful way. This hinges on matching potentially different domain-specific vocabularies and assessing results based on their semantic relevance to the question. Here, we empirically benchmark current state-of-the-art LLM SLR methodologies in an interdisciplinary setting. We specifically consider the example of systems biology, which integrates knowledge from biology, physics, mathematics, and computer science in order to provide predictive mechanistic understanding of living

EKAW 2024: EKAW 2024 Workshops, Tutorials, Posters and Demos, 24th International Conference on Knowledge Engineering and Knowledge Management (EKAW 2024), November 26-28, 2024, Amsterdam, The Netherlands.
*The work is done outside Amazon
charvi.jain@tu-dresden.de (C. Jain); sahar.vahdati@tu-dresden.de (S. Vahdati); nandu.gopan@tu-dresden.de (N. Gopan); ivo.sbalzarini@tu-dresden.de (I. F. Sbalzarini); jens.lehmann@tu-dresden.de (J. Lehmann)


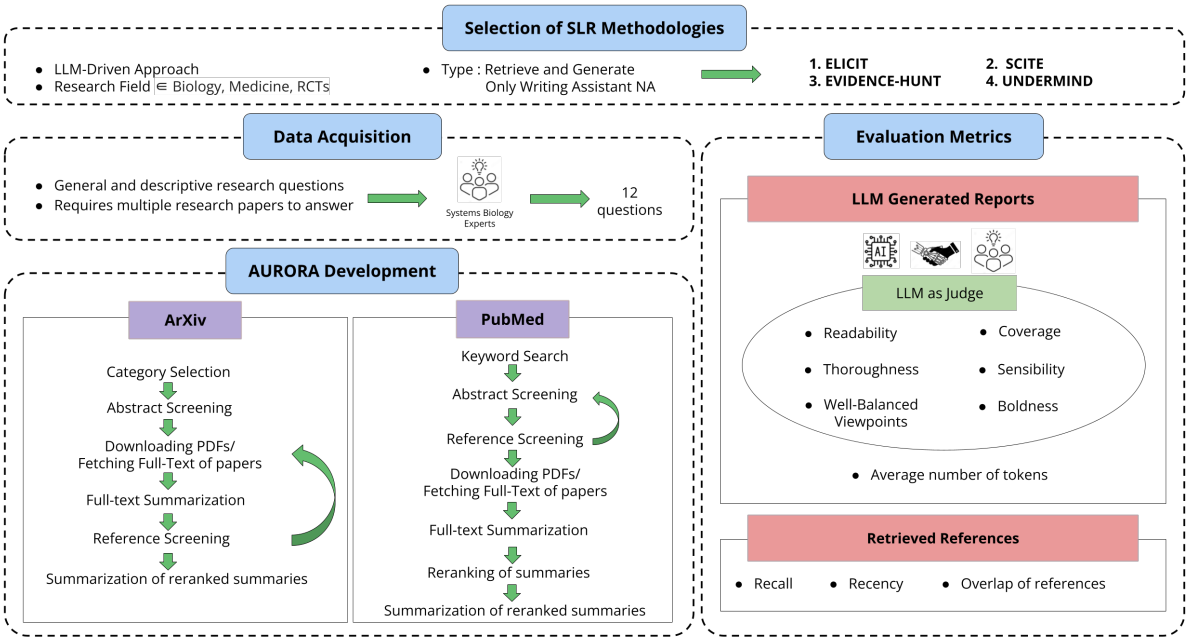

Figure 1: Overview of the present comparative evaluation methodology.

systems [5, 6]. The breadth of the research field of systems biology creates several data integration and integrative analysis challenges [7, 8] that are difficulty even for human scientists. In order to evaluate how well LLM-based SLR systems perform in systems biology, we introduce an end-to-end systematic comparison framework based on research questions formulated by domain experts and quantitative evaluation of results. We find that previous LLM SLR systems focus on certain performance dimensions while neglecting others. Based on this observation, we propose a new approach: AURORA — Automated Understanding and Review Of Research Articles. The proposed AURORA method balances well on report quality and reference retrieval. This allows us to identify opportunities for further development of LLM SLR systems in interdisciplinary science.

2. Methodology

We design an end-to-end evaluation framework as illustrated in Fig. 1. The key ingredients are: selection of SLR methods, dataset acquisition, AURORA development, and evaluation metrics.

Selection of SLR Methods: From amongst the previous LLM SLR methods Elicit[1], Scite[2], Perplexity[3], EvidenceHunt[4], MirrorThink[5], Scispace[6], and Undermind[7], we selected four state-of-the-art approaches by the following criteria: the approach should be LLM-driven; it should either focus on biology, medicine, or randomized controlled trials as their domain; it should be a retrieve-and-generate assistant rather than a simple search wizard or writing assistant.

Dataset acquisition: We organized a workshop with a total of 15 practicing domain experts from systems biology. This included doctoral students, doctoral researchers, and principal investigators from the Center for Systems Biology Dresden. We asked them to formulate broad and open-ended research questions that were not previously answered in any single publication. We collected 12 such research questions of diverse type and nature (see appendix A).

[1]https://elicit.com/
[2]https://scite.ai/assistant
[3]https://www.perplexity.ai/
[4]https://evidencehunt.com/
[5]https://mirrorthink.ai/
[6]https://typeset.io/
[7]https:/undermind.ai/

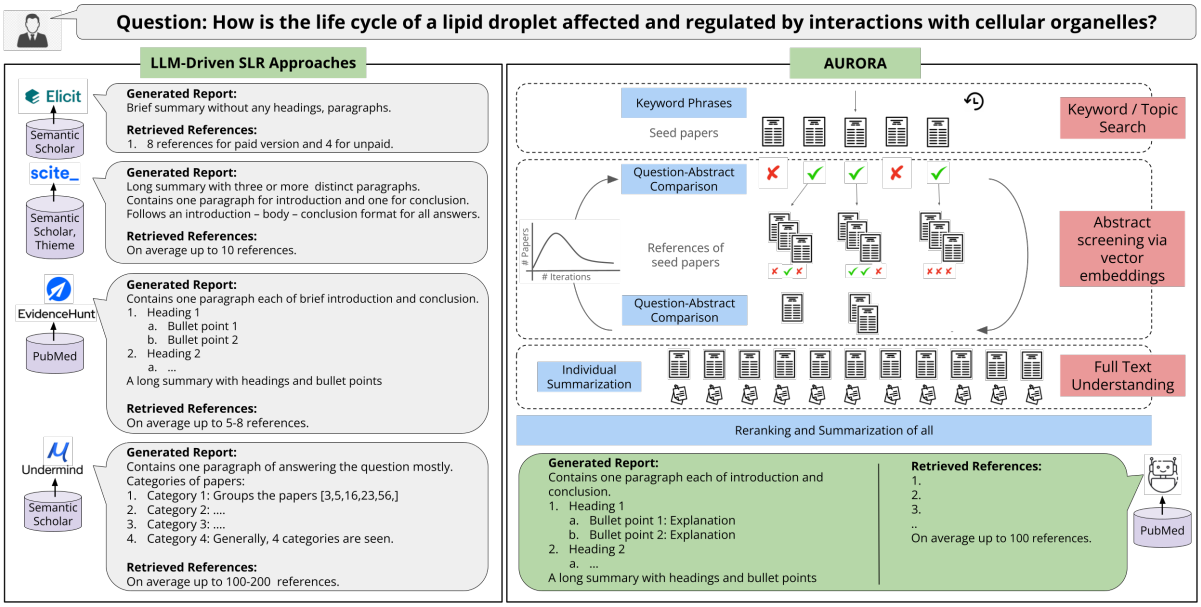

Figure 2: Overview of the proposed approach AURORA using the PubMed and PubMed Central literature databases and report formats for all the approaches.

AURORA development: We design AURORA to perform end-to-end systematic review of all literature relevant to a given research question. It combines keyword search, abstract screening, full-text understanding of individual papers, and combining information from multiple papers. As data sources, AURORA uses the public APIs of ArXiv and PubMed, providing almost complete coverage of the literature relevant to systems biology. PubMed's metadata supports convergent search over multiple iterations, which is not possible for ArXiv. Hence, the evaluation below uses the PubMed database.

The overall design of AURORA is summarized in Fig. 2. We automate keyword search by prompting a LLM to generate five search phrases for a given research question, selecting the one yielding the most results. We screen the abstracts of the resulting papers in a vector-space embedding. We repeat the process for successively longer search time horizons until we find 5–10 "seed papers". Next, we scan the references of these seed papers using PubMed's metadata, collect their unique identifiers (PubMedID-PMID, PubMedCentralID-PMCID, DOI), and screen the abstracts of all cited papers. This is repeated until the result set converges (i.e., no new papers are found). Upon convergence, the full texts of all found papers are fetched and collectively fed into a long-context LLM. If full text is not available (or pay-walled) only the abstract is used. The LLM summarizes each paper, and a re-ranking model is applied to score the summaries. From this, the final written report is generated using the LLM.

Implementation details: Overall, AURORA uses one embedding model, one re-ranking model, and three LLMs (for keyword–phrase generation, full-text summarizing, and report compilation). For the evaluation below, all three LLM instances were GPT-4o-mini. For abstract screening, embeddings were generated using the nomic-embeddings model[8] [9], which supports a sequence length of 8192 tokens. The search time horizon was iteratively extended in 12-month steps until seed papers were found. The maximum number of scans allowed is 20 per research question. The summaries were re-ranked using the Jina re-ranker[9] based on their relevance to the initially given research question.

3. Evaluation and Results

We use evaluation metrics from Fig. 1 to perform qualitative analysis of generated reports and quantitative analysis of references retrieved for Elicit, Scite, Evidence Hunt, Undermind, and AURORA.

[8]https://huggingface.co/nomic-ai/nomic-embed-text-v1
[9]https://huggingface.co/jinaai/jina-reranker-v1-turbo-en

3.1. LLM-Generated Reports

LLM-as-judge evaluation: We use the state-of-the-art models GPT-4o and Claude 3.5-sonnet as judges to rate the reports generated by the different LLM SLR approaches across all 12 research questions on a Likert scale of 1–5 based on the criteria in Table 1. Higher ratings indicate better reports. The spider charts of the results from both judges are shown in Fig. 3. From those, the performance of Evidence Hunt and AURORA were similar for almost all criteria, whereas Scite showed moderate performance, and Elicit and Undermind under-performed.

Metric	Prompt description
Coverage	Assess how well the report addresses the research question by considering all relevant aspects, perspectives, and sub-questions.
Thoroughness	Examine how thoroughly and in-depth each aspect is explained in the report, along with how well it is supported by evidence and clear explanations.
Boldness	Examine whether the report presents well-supported arguments with confidence, handles ambiguous points with clarity and free of vague language.
Readability	Assess how easily the report can be understood and followed by evaluating its structure, including the use of headings, subheadings, and sufficient context.
Balance of viewpoints	Evaluate if the report presents balanced perspectives on conflicting literature, avoids extreme views unless well supported, and acknowledges different sides of argument.
Sensibility	Assess the report's alignment with common sense and logical coherence. Evaluate the rationality and chronology of references in the report.

Table 1: Metrics for LLM-as-judge evaluation, scored on a 1–5 Likert scale (higher is better).

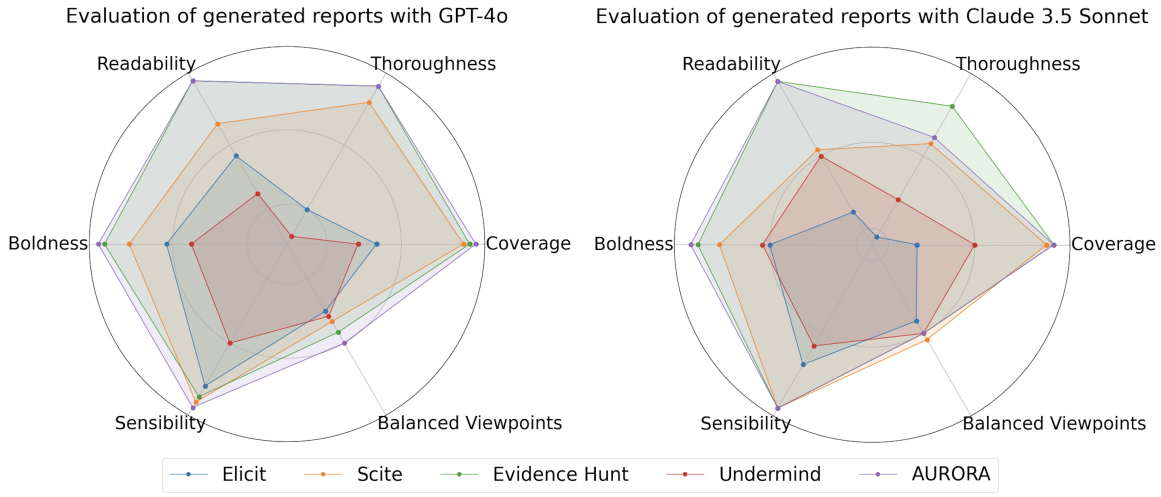

Figure 3: Spider charts of the results using GPT-4o (left) and Claude 3.5-sonnet (right) as judges to evaluate the generated reports of all five LLM SLR systems (colors in legend) based on Table 1.

SLR System:	Elicit	Scite	Evidence Hunt	Undermind	AURORA
Avg. report length:	193.25	761.42	782.83	218.25	1235.73

Table 2: Average (across research questions) lengths of generated literature reports in numers of tokens.

Average length of generated reports: We compare the average number of tokens generated for each research question and approach. As shown in Table 2, Elicit and Undermind produced shorter reports than Scite and Evidence Hunt, suggesting they may not comprehensively address open-ended literature research questions. This aligns with Fig. 3, where Elicit and Undermind received lower ratings for coverage and thoroughness. AURORA used the most tokens, indicating it may provide more comprehensive contextual information, whereas Evince Hunt and Scite were almost equal in the middle.

3.2. Retrieved References

Elicit, Scite and Undermind use Semantic Scholar having access to 200 million articles [10] whereas Evidence Hunt and AURORA use PubMed with approx. 37 million articles [11] (cf. Table 3). We used the initially retrieved references without further asking for extension to ensure fairness across approaches.

Overlap of references: We count the number of common references reported by any two SLR approaches. Since approaches are uncorrelated, this allows estimating the fraction of all references any method *A* finds as: (Number of common references between *A* and *B*)/(Number of references found by *B*) [4]. This metric is called "overlap", since a value of 1 indicates complete overlap of results. A color map of the overlap between any two approaches is shown in Fig. 4 where rows are *A* and columns *B*. The diagonal is set to 0 to better utilize the color range. Undermind shows the highest overlap with other methods, likely because it finds the most references. The overlap of AURORA is almost as high, while Elicit, Scite, and Evidence Hunt show significantly lower overlaps (below 0.075) with other methods, indicating limited efficiency to find the best, most relevant, or foundational papers for the question.

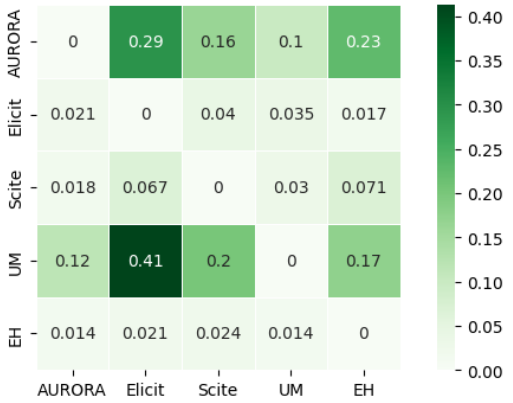

Figure 4: Reference overlap between approaches.

SLR Approach	Articles (million)	Recall (%)	Recency (avg. #)
Elicit	200	4.46	0.75
Scite	200	6.64	3.0
Evidence-Hunt	37	3.88	2.17
Undermind	200	47.13	17.41
AURORA	37	47.98	7.25

Table 3: Recall and recency scores.

Recall of references: Recall is defined as the fraction of references retrieved by any SLR method compared to the total set of references found by all methods combined. This generalizes the pair-wise overlap metric to considering all other methods as a reference set. Table 3 shows that Evidence Hunt has the lowest recall score, suggesting that it fails to find a sufficient number of relevant references despite high scores in other criteria (see Fig. 3). AURORA mitigates this limitation and stands close to Undermind in its ability to retrieve highly relevant references.

Recency of references: We count the average number of references found from recent years (2022, 2023, 2024), across all research questions. Table 3 shows that Elicit has only a 75% chance to find one recent paper on average, whereas all other approaches find several recent papers. AURORA's iterative temporal approach helps it capture more recent papers than Evidence Hunt from the same database. Undermind's very high recency score is likely due to its use of Semantic Scholar, which is a much larger literature database than PubMed.

4. Conclusion

We presented an initial study on using LLM SLR methods in an interdisciplinary field of science, here systems biology. We found that the performance of state-of-the-art methods is complementary in terms of report quality and reference recall/recency, reflecting a trade-off in interdisciplinary searches. The AURORA approach proposed here could provide a way of addressing this limitation, as we found it to maintain the best balance. This hints at its ability of retrieving contextually relevant references while still achieving good literature coverage. This hinges on iterating temporal search with LLM-based full-text understanding until convergence. In the future, we will extend AURORA to additional database sources, including databases of gene and protein sequences.

Acknowledgments

This work was supported by the German Federal Ministry of Education and Research BMBF through DAAD project 57616814 (SECAI, School of Embedded Composite Artificial Intelligence). We thank all participants of the workshop at the Center for Systems Biology Dresden for contributing research questions for the evaluation.

References

[1] J. Kung, Elicit (product review), Journal of the Canadian Health Libraries Association / Journal de l'Association des bibliothèques de la santé du Canada 44 (2023). URL: https://journals.library.ualberta.ca/jchla/index.php/jchla/article/view/29657. doi:10.29173/jchla29657.

[2] J. Nicholson, M. Mordaunt, P. Lopez, A. Uppala, D. Rosati, N. Rodrigues, P. Grabitz, S. Rife, Scite: A smart citation index that displays the context of citations and classifies their intent using deep learning, Quantitative Science Studies 2 (2021) 1–38. doi:10.1162/qss_a_00146.

[3] J. M. N. Sean C. Rife, Domenic Rosati, scite: The next generation of citations, Learned Publishing 34 (2021).

[4] T. Hartke, J. Ramette, Benchmarking the undermind search assistant, 2024.

[5] Z. Szallasi, J. Stelling, V. Periwal, System modeling in cellular biology: From concepts to nuts and bolts, 2006. doi:10.7551/MITPRESS/9780262195485.001.0001.

[6] L. Hood, L. Rowen, D. Galas, J. Aitchison, Systems biology at the institute for systems biology., 2008. doi:10.1093/bfgp/eln027.

[7] E. D. . th. Manchester, Data integration in the life sciences : 6th international workshop, dils 2009, manchester, uk, july 20-22, 2009 : Proceedings, 2009.

[8] L. Zhu, A. Bhattacharyya, E. Kurali, A. Anderson, A. Menius, K. R. Lee, Review of integrative analysis challenges in systems biology, 2011. doi:10.1198/sbr.2010.09027.

[9] Z. Nussbaum, J. X. Morris, B. Duderstadt, A. Mulyar, Nomic embed: Training a reproducible long context text embedder, 2024. URL: https://arxiv.org/abs/2402.01613. arXiv:2402.01613.

[10] U. of Calgary, Semantic scholar resources statistics, 2024. URL: https://libguides.ucalgary.ca/c.php?g=732144&p=5260798, last accessed 15 September 2024.

[11] N. L. of Medicine, Pubmed resources statistics, 2024. URL: https://pubmed.ncbi.nlm.nih.gov/, last accessed 15 September 2024.

A. Online Resources

- The research questions used in the evaluation are available as online spreadsheet A.
- The reports from all approaches with criteria scores are available as online spreadsheet B.