Beyond Search: Measuring LLM Performance for Scientific Literature Discovery

Ching Nam Hang*[§], Pei-Duo Yu[†], Chee Wei Tan[‡], Dah Ming Chiu[§]
[*] Yam Pak Charitable Foundation School of Computing and Information Sciences, Saint Francis University, Hong Kong
[†] Department of Applied Mathematics, Chung Yuan Christian University, Taiwan
[‡] College of Computing and Data Science, Nanyang Technological University, Singapore
[§] Data Science Research Centre for Social Policies and Services, Saint Francis University, Hong Kong
Email: cnhang@sfu.edu.hk, peiduoyu@cycu.edu.tw, cheewei.tan@ntu.edu.sg, dmchiu@sfu.edu.hk

Abstract—**The exponential growth of scientific publications has made traditional literature reviews increasingly time-consuming and error-prone, particularly in fast-moving fields such as artificial intelligence (AI) and machine learning (ML), thereby highlighting the need for intelligent discovery tools. With the recent development of large language models (LLMs), this paper evaluates their capability to perform automated literature search using only paper abstracts. We propose a three-metric framework, reference recall, supplementary suggestion rate, and hallucination rate, to measure how effectively an LLM retrieves known citations, recommends additional relevant works, and avoids generating fabricated entries. Using three influential papers in AI and ML as test cases, we compare GPT-4, Gemini 2.5, and DeepSeek-V3 in both a normal configuration and an augmented, retrieval-enabled deep research mode. GPT-4 with deep research augmentation achieves the highest citation accuracy, while Gemini 2.5 and DeepSeek-V3 exhibit more variable gains depending on their respective retrieval strategies. Supplementary suggestions consistently enhance the bibliography, particularly in areas with broader foundational literature. Only a few hallucinations are observed, indicating that the well-known problem of fabricated references has been significantly mitigated as LLM technology has evolved. Our study suggests that modern LLMs, especially when using deep research mode, can serve as effective research assistants for literature discovery in AI and ML.**

Index Terms—**Literature search, large language models, deep research, research methodologies, information retrieval.**

I. Introduction

The rapid growth of scientific publications has made it increasingly challenging for researchers to keep up with related work in their fields [1]. In the domains of artificial intelligence (AI) and machine learning (ML), tens of thousands of new papers are published every year [2], [3]. For instance, the 40th Annual AAAI Conference on Artificial Intelligence in 2025 received nearly 29,000 submissions to its main track, an unprecedented volume [3]. This explosion of literature, with the number of AI and ML papers roughly doubling every two years, places immense pressure on traditional literature review practices. Academic search engines and recommendation systems provide some assistance, but researchers, especially students and early-career scholars, still spend considerable time identifying relevant prior work. Consequently, there is growing interest in using AI assistants to support literature reviews [4], [5]. For example, large language models (LLMs) could suggest relevant papers based on a given research topic or abstract [6]. Such tools have the potential to enhance learning and research in educational settings by helping users navigate the overwhelming "deluge" of publications. However, the capabilities and reliability of LLMs in this role remain under-examined.

Recent advances in LLMs, such as the Generative Pretrained Transformer (GPT) series, have achieved human-level performance on numerous benchmarks [7]–[10], raising the question of whether they can also address more advanced research tasks. An LLM capable of passing professional examinations may also be able to retrieve and synthesize academic literature at a level comparable to trained researchers. The emergence of "deep research" features, where the model autonomously conducts in-depth, multi-step research across the public web, further enables fine-tuning on diverse online sources and the generation of fully documented, properly cited reports. Such capabilities can lead to more accurate and comprehensive literature surveys, including the discovery of niche or non-intuitive information that typically requires navigating multiple sources and websites on complex topics.

On the other hand, LLMs are known to generate plausible-sounding but incorrect information, a phenomenon known as "hallucination" [11]–[14]. Even the most advanced models can produce fabricated references or citations in academic content. For example, the Galactica model [15], trained specifically on scientific texts, was criticized for producing fictitious research citations that appeared convincingly authoritative. This issue was publicly highlighted by Michael Black, a renowned computer vision researcher and Director at the Max Planck Institute for Intelligent Systems, who warned that such confident fabrications could mislead researchers. These concerns underscore the importance of rigorously evaluating LLMs before relying on them as tools for literature reviews or research assistance.

In this paper, we present an evaluation of state-of-the-art LLMs on the task of identifying related work for a given paper, effectively assessing their literature discovery capabilities. Our focus is on the AI and ML domains, chosen both for their exceptionally high publication volume and for their technological maturity, which makes it likely that LLMs have been trained on many influential papers in these fields. In educational contexts, such as graduate programmes in engineering or computer science, tools that assist with literature reviews in AI and ML could be particularly valuable, helping students efficiently identify relevant prior work for theses or research projects. Our study specifically examines GPT-4 [16], Gemini 2.5 [17], and DeepSeek-V3 [18], comparing their performance with and without an augmented "deep research" feature that enables the model to perform real-time online searches. GPT-4 and Gemini 2.5 are leading proprietary models, widely regarded as industry benchmarks for reasoning, general knowledge, and language capabilities. DeepSeek-V3, on the other hand, is an open-source model reported to achieve strong performance on reasoning benchmarks, offering transparency and reproducibility advantages not found in closed-source systems. We deliberately focus on these three models because they represent a balance of proprietary and open-source options, are accessible for testing, and are among the most widely cited or discussed in the AI research community. By evaluating these models in both standard configurations and enhanced modes with deep research capabilities, we aim to assess the current limitations of LLMs in performing literature discovery tasks.

II. Related Work

Literature search and the identification of relevant academic publications have long been fundamental components of research methodology. Early empirical studies in information systems highlighted disciplinary diversity and examined journal characteristics to better understand how researchers select and retrieve literature [19]. With the rapid growth of academic output, research paper recommendation systems have increasingly emerged as valuable tools to help scholars navigate the expanding literature [20]. These systems employ a range of approaches, including content-based filtering, collaborative filtering, and hybrid techniques, to suggest relevant publications [21]. However, these earlier systems often relied heavily on structured data, including citation networks and metadata, which limited their adaptability in complex research contexts or emerging interdisciplinary fields.

Recent advancements in LLMs open up new possibilities for automating systematic literature reviews and literature searches. For example, the authors in [22] introduce srBERT, a domain-adapted Bidirectional Encoder Representations from Transformers (BERT) model designed to automate key steps of systematic reviews by classifying included articles, streamlining literature screening and article classification for systematic review tasks. The study in [23] examines the ability of GPT-4 to conduct systematic reviews, evaluating its performance in screening and data extraction across multiple languages. In [24], the authors compare AI-generated and human-conducted literature reviews on physician–patient relational dynamics and show that GPT-4, when guided by structured prompt engineering, can rapidly produce broad overviews but still requires expert evaluation to achieve the depth, accuracy, and contextual understanding of traditional human reviews.

A major challenge in using LLMs for academic content generation is the well-documented problem of hallucinations, where models produce references that appear plausible, but are incorrect or nonexistent. For instance, the study in [25] examines the authenticity and accuracy of medical references produced by ChatGPT-3.5, showing that while the model can quickly generate referenced medical articles, most of its citations are fabricated or inaccurate. The authors in [26] assess the ability of ChatGPT to answer medical questions and provide supporting references, finding that while the model can generate plausible-looking responses, its answers have limited scientific quality and the majority of its citations are fabricated. Recent studies show that LLM-based literature reviews focus on narrative quality while neglecting reference checks [24].

Our study addresses an underexplored research gap by evaluating how leading LLMs perform in generating relevant and authentic references when provided only with a paper abstract. We compare their capabilities in both a normal and an augmented configuration. This comparative approach allows us to assess both the baseline performance of LLMs in literature discovery as well as the practical benefits of incorporating external retrieval to improve accuracy, reduce hallucinations, and enhance overall reliability. This work therefore provides insights that are valuable to researchers, students, and engineers who aim to use LLMs as dependable tools for literature search across diverse topics in AI and ML.

III. Methodology

We formulate the literature review task as follows: given the abstract of a published research paper, an LLM is prompted to generate a list of references that would appropriately situate the paper within its research context. The expected output is a set of reference entries, each comprising a title, author list, publication year, and source. We do not require the references to follow a specific citation style. We use abstracts as input because they succinctly summarize the content and contributions of a paper. Abstracts are commonly used by human researchers to understand the scope of a paper, and thus should provide sufficient context for an LLM to infer relevant prior work. Importantly, using abstracts prevents models from simply reproducing the actual reference list of a paper, allowing us to evaluate their ability to identify relevant citations independently. Applying abstracts uniformly also simulates a realistic scenario in which a researcher, equipped only with a summary of a new idea, seeks to discover related work through the assistance of an LLM.

To assess performance, we compare the LLM-generated reference list with the actual references cited in the target paper. We measure how many of the original references the model successfully retrieves and examine additional suggestions, including any hallucinated entries. Our evaluation is based on the assumption that the references selected by the original authors are a reasonable proxy for the key prior work in the field, an assumption that generally holds for well-written, peer-reviewed papers.

A. Dataset

To make the evaluation concrete, we select three seminal AI and ML papers from the past decade as test cases:

- [27]: The authors introduce Generative Adversarial Networks (GANs), which represent a novel framework for generative modeling.
- [28]: The authors introduce ResNet, which enables the training of very deep convolutional neural networks.
- [29]: The authors introduce the Transformer architecture, which serves as the foundation for modern LLMs.

We select these three landmark papers because each introduces a novel neural network architecture based on a distinct fundamental principle, illustrating the diversity of modern deep learning. The authors in [27] pioneer adversarial training for generative modeling, a framework that can be interpreted as minimizing a divergence between the model distribution and the data distribution. ResNet in [28] enables the training of ultra-deep models in computer vision by introducing skip connections that allow signals to bypass layers. In [29], the authors transform sequence modeling in natural language processing (NLP) with an architecture centered on attention, building on the attention mechanism first proposed by [30]. These three works are highly influential and exemplify a broad architectural diversity in AI and ML.

We choose these works not only for their impact but also because their reference lists serve as well-curated representations of related literature in their respective subfields of generative modeling, computer vision, and NLP. By focusing on such high-quality papers, we ensure that the “ground truth” references are both reliable and meaningful. Additionally, selecting papers from diverse areas within AI and ML allows us to assess the ability of LLMs to generalize across topics. This study is limited to the AI and ML domains, while performance evaluations in other areas, such as biomedical research or the social sciences, are left for future work.

B. Models and Settings

We evaluate three LLMs: GPT-4 [16], Gemini 2.5 [17], and DeepSeek-V3 [18]. GPT-4 is a flagship model renowned for its extensive knowledge and strong performance across a wide range of tasks. Gemini 2.5 is a multimodal model that has been positioned as a direct competitor to GPT-4. DeepSeek-V3 is an open-source model developed through a collaborative research effort. According to its authors, DeepSeek surpasses LLaMA-2 70B on multiple benchmarks and demonstrates particularly strong reasoning and coding capabilities, making it a leading candidate among open models. All three models are based on the Transformer architecture and are accessed through text-based interfaces.

For each model, we evaluate two modes of operation:

- Normal Mode: The model is provided only with the abstract text and an instruction to generate related work, without access to any external information. This mode relies solely on the internal knowledge of the model from its training data. It simulates the behavior of a base LLM, such as ChatGPT without plugins, essentially addressing the question of what the model can infer on its own.
- Augmented Mode: The model is allowed to perform external searches and read retrieved content before generating its response. In this setting, we enable the deep research or deep thinking mode, allowing the LLM to issue search queries (typically using key terms or concepts from the abstract) and read snippets of papers or articles it retrieves. It then incorporates this external information into its output. This mode is analogous to an LLM equipped with a research assistant plugin, capable of fetching real papers in real time. It is intended to reduce errors such as hallucinated citations, as the model can reference verified sources.

Each model is tested in two modes across three paper abstracts, resulting in six evaluations per model. We apply the same prompt format to all models, making minor wording adjustments to ensure the output is a list of papers rather than a prose response. To keep the output manageable, we limit it to the 10 most relevant references. Although the actual reference lists typically contain around 40 citations, we do not expect the models to reproduce all of them. To fairly evaluate each model in identifying prior work, we also include a publication year constraint in the prompt, instructing the LLM to list only papers published on or before the year of the target paper. Without this constraint, a model may reasonably suggest many papers published after 2017, especially since the target papers date from 2014 to 2017 and recent literature from 2020 onward often dominates when the input is limited to the abstract. While researchers often aim to find cutting-edge publications in practice, our objective here is to measure the capacity of each LLM to retrieve existing citations that could have informed the original work. By restricting suggestions to the appropriate timeframe, we avoid penalizing models for returning otherwise relevant but later publications, and we keep the evaluation focused solely on literature published prior to the target paper. This approach aligns the task with typical researcher expectations, where the initial focus is on uncovering prior art before exploring more recent developments. The prompt used for evaluation is provided below.

Evaluation Prompt

> You are given the abstract of a research paper.
>
> Abstract: [Insert Abstract]
>
> Please conduct a literature review based on the key terms and concepts in the abstract, and list the 10 most relevant related works published before [year].
>
> For each entry, include the title, author list, publication year, and source (journal or proceedings).
>
> Do not include any additional commentary or text. Only provide the list of 10 related works.

C. Evaluation Criteria

Our primary evaluation metric is the number of relevant references generated by the LLM. We define "relevant" using two criteria:

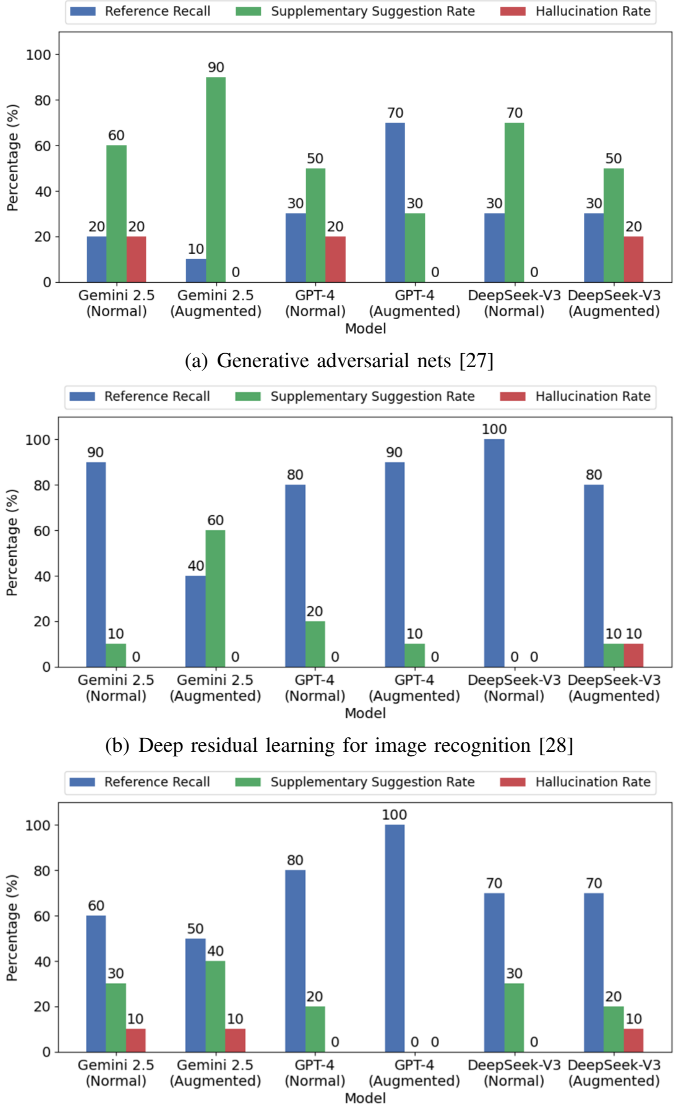
1) Overlap with the actual reference list of the target paper: If the authors of the target paper cited a work, that work is considered definitively relevant. We count how many such overlaps the model produces.
2) Expert human judgment: For any references suggested by the model that do not appear in the original list, we manually assess whether each one is topically appropriate and suitable for citation based on the abstract.

In applying the second criterion, we verify the authenticity of each suggestion by cross-checking every listed title and author using Google Scholar. If a paper cannot be found or the citation appears distorted or incomplete, we flag it as fabricated and count it as an error. Any real and thematically appropriate paper that was not cited by the original authors is treated as a valuable supplement rather than an error. Therefore, only fabricated or irrelevant references are penalized, while all genuine prior works are credited as useful additions.

To quantify these outcomes, we define three complementary metrics for each model output:

- Reference Recall: It measures the proportion of true positives, referring to the percentage of generated references that overlap with the actual reference list of the target paper.
- Supplementary Suggestion Rate: It captures the percentage of real, thematically appropriate works published before the target paper that were proposed by the model but not included in the original citations. These are considered valuable additions based on human judgment.
- Hallucination Rate: It denotes the percentage of fabricated or post-dated references, meaning those that cannot be verified or were published after the target paper, and are therefore counted as errors.

These metrics are computed as the exact count divided by the total number of ten suggested references, expressed as percentages.



(a) Generative adversarial nets [27]

(b) Deep residual learning for image recognition [28]

(c) Attention is all you need [29]

Fig. 1. Performance of different LLMs in normal and augmented modes on automated literature search across three seminal AI and ML papers, highlighting their ability to recover true citations, identify valuable supplementary works, and avoid hallucinated references.

IV. Results

In Fig. 1, we present the performance of different LLMs in normal and augmented modes on automated literature search across three seminal AI and ML papers, evaluated using our three proposed metrics. Overall, all models retrieve a high proportion of relevant works with minimal hallucinations, particularly for the papers [28] and [29], where average reference recall exceeds 80 percent and hallucination rates are close to zero. In contrast, the case of paper [27] demonstrates lower reference recall for all models (see Fig. 1(a)). The literature of GANs covers a broader and more diverse range of probabilistic and deep generative frameworks, including VAEs, Boltzmann machines, and stochastic networks. This diversity makes it more difficult for an LLM to identify the exact citations used by the original authors. However, the high supplementary suggestion rates for the case of paper [27] help compensate for this limitation: each model proposes many additional, thematically relevant papers (published before [27]) that enrich the reference list despite lower overlap. This compensatory pattern is also observed in the cases of [28] and [29], where supplementary suggestions enhance overall coverage beyond the core references. Importantly, no model suggests irrelevant works. The three metrics always sum to 100%, confirming that aside from hallucinations, all outputs are either accurate matches or valid additions.

Focusing on the reference recall metric, GPT-4 augmented with deep research performs relatively well in conducting literature search, achieving the highest recall without any hallucinated entries, whereas Gemini 2.5 augmented with deep research demonstrates comparatively weaker performance. Upon examining the activities carried out by these two LLMs and analyzing their internal reasoning processes, several differences become evident. GPT-4 augmented with deep research adopts a focused and strategic approach. It begins by extracting key concepts and terms from the abstract, then formulates targeted queries and compares the abstracts of retrieved papers. On average, it performs more than 50 searches per paper before selecting the ten most relevant works. This iterative, comparison-driven method allows the model to filter out loosely related references and prioritize those with strong conceptual alignment to the target abstract. This precision contributes to its higher Reference Recall and minimal hallucination rate.

In contrast, Gemini 2.5 augmented with deep research follows a more generalized workflow. Instead of directly targeting key terms from the abstract, it attempts to construct a comprehensive academic report based on the abstract content. This includes generating multiple paragraphs on the foundational background of the concepts of an abstract. On average, it performs 32 searches per paper, which are primarily used to support the writing of the report. As a result, the final list of ten references is often broader and less specific to the target paper, leading to a lower overlap with the original reference list. This difference in search methodology explains the performance gap. The targeted and analytical strategy of GPT-4 enhances its ability to retrieve citations that are both topically accurate and directly relevant, while the narrative-based approach of Gemini 2.5 dilutes its focus, resulting in reduced precision in identifying core related works.

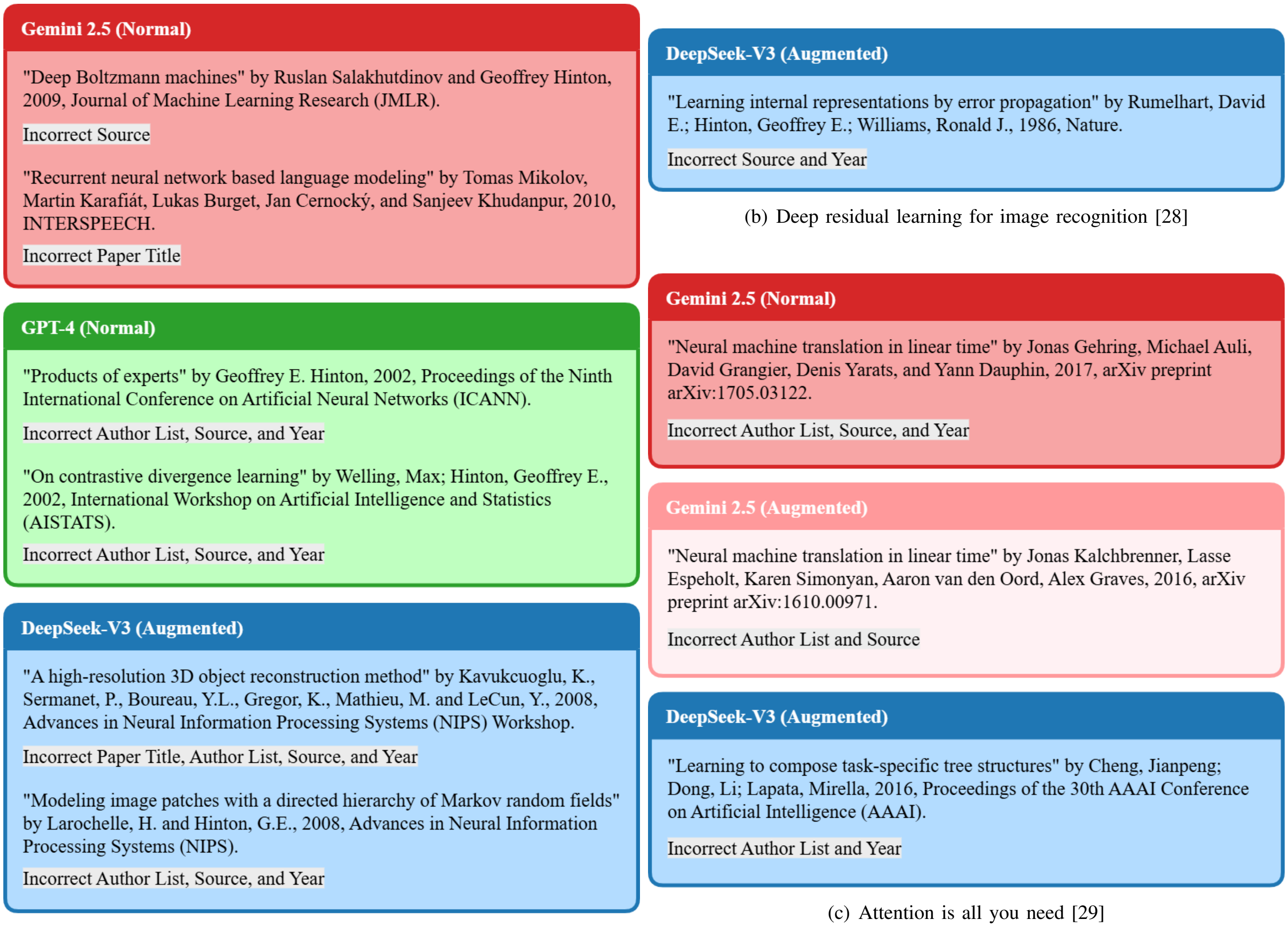
For Gemini 2.5 and DeepSeek-V3, it is noteworthy that enabling the augmented mode does not lead to better performance than the normal mode in conducting literature search. In the case of Gemini 2.5 augmented with deep research, as previously discussed, its tendency to generate a comprehensive academic report contributes to the inclusion of many general foundational works, rather than focusing on the most directly relevant citations derived from the abstract. For DeepSeek-V3 augmented with deep thinking and web search, analysis of its activities and reasoning reveals that although it performs an average of 49 searches per paper, many of these searches target broadly trustworthy sources such as Wikipedia. Within those sources, the model extracts cited works without conducting comparisons across multiple references. This lack of cross-source comparison may result in overlooking more directly relevant papers, as the model relies on citations within a single source rather than synthesizing across different materials. For both Gemini 2.5 and DeepSeek-V3, it can be inferred that the training data available to them in normal mode already provides strong baseline performance. The introduction of web search in augmented mode may introduce excessive information, which could overwhelm the models and reduce their focus. This is partly influenced by the quality and relevance of the search results retrieved through the external search engines, which do not always align precisely with the abstract content or task intent.

To further evaluate hallucination cases, we present all identified instances in Fig. 2. It is observed that only one case, produced by DeepSeek-V3 in augmented mode with deep thinking and web search, contains a completely fabricated reference that cannot be found through verification. All other hallucinations involve minor issues such as incorrect author lists (mostly accurate but including a few extraneous names), incorrect sources or publication years, or slight typographical errors in the paper titles. In the case of the fully fabricated citation from DeepSeek-V3 augmented, further analysis shows that the reasoning and search activities consistently reference real papers. However, the final list of ten suggested works includes a citation that does not appear at any point in the analysis process of the model. This anomaly suggests that even with transparent retrieval steps, there may still be unpredictable moments where unrelated content is introduced, potentially due to internal inconsistencies during the synthesis phase of generation. Overall, these findings indicate that the well-known issue of hallucination in literature search has been substantially mitigated through advancements in LLM development. Most hallucinations are now reduced to minor inaccuracies rather than fully invented references, reflecting improved reliability in citation-related tasks.

Finally, the overall time required for augmenting the LLMs to perform literature search varies across the models, with DeepSeek-V3 requiring the shortest time and Gemini 2.5 taking the longest. This difference is mainly due to their respective goals and strategies. Gemini 2.5 focuses on composing a complete academic report, which involves generating extensive background content before selecting references. In contrast, DeepSeek-V3 aims to locate related works by targeting trustworthy sources. For example, after identifying a relevant Wikipedia page based on the abstract, it scans the content to extract cited works without conducting deeper analysis or cross-comparison. Considering the overall performance of all LLMs in the augmented mode, the time spent on augmentation is deemed worthwhile for conducting literature search.

V. Conclusion

In this paper, we present a systematic evaluation of LLMs, GPT-4, Gemini 2.5, and DeepSeek-V3, in conducting automated literature search based on paper abstracts. By introducing three metrics, reference recall, supplementary suggestion rate, and hallucination rate, and applying them to three influential papers in AI and ML, we demonstrate that incorporating deep research with web search augmentation significantly improves citation accuracy. GPT-4, in particular, achieves the highest recall and avoids fabrication errors through a focused, keyword-driven retrieval strategy. Gemini 2.5 and DeepSeek-V3 also show benefits, though to varying extents depending on their retrieval workflows. Across all experiments, no genuinely irrelevant works are suggested, and the supplementary suggestions help compensate for lower overlap in more challenging domains. These findings confirm that modern LLMs, when paired with intelligent retrieval, can act as reliable assistants for compiling high-quality bibliographies.



Gemini 2.5 (Normal)

"Deep Boltzmann machines" by Ruslan Salakhutdinov and Geoffrey Hinton, 2009, Journal of Machine Learning Research (JMLR).

Incorrect Source

"Recurrent neural network based language modeling" by Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernocký, and Sanjeev Khudanpur, 2010, INTERSPEECH.

Incorrect Paper Title

GPT-4 (Normal)

"Products of experts" by Geoffrey E. Hinton, 2002, Proceedings of the Ninth International Conference on Artificial Neural Networks (ICANN).

Incorrect Author List, Source, and Year

"On contrastive divergence learning" by Welling, Max; Hinton, Geoffrey E., 2002, International Workshop on Artificial Intelligence and Statistics (AISTATS).

Incorrect Author List, Source, and Year

DeepSeek-V3 (Augmented)

"A high-resolution 3D object reconstruction method" by Kavukcuoglu, K., Sermanet, P., Boureau, Y.L., Gregor, K., Mathieu, M. and LeCun, Y., 2008, Advances in Neural Information Processing Systems (NIPS) Workshop.

Incorrect Paper Title, Author List, Source, and Year

"Modeling image patches with a directed hierarchy of Markov random fields" by Larochelle, H. and Hinton, G.E., 2008, Advances in Neural Information Processing Systems (NIPS).

Incorrect Author List, Source, and Year

(a) Generative adversarial nets [27]

DeepSeek-V3 (Augmented)

"Learning internal representations by error propagation" by Rumelhart, David E.; Hinton, Geoffrey E.; Williams, Ronald J., 1986, Nature.

Incorrect Source and Year

(b) Deep residual learning for image recognition [28]

Gemini 2.5 (Normal)

"Neural machine translation in linear time" by Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann Dauphin, 2017, arXiv preprint arXiv:1705.03122.

Incorrect Author List, Source, and Year

Gemini 2.5 (Augmented)

"Neural machine translation in linear time" by Jonas Kalchbrenner, Lasse Espeholt, Karen Simonyan, Aaron van den Oord, Alex Graves, 2016, arXiv preprint arXiv:1610.00971.

Incorrect Author List and Source

DeepSeek-V3 (Augmented)

"Learning to compose task-specific tree structures" by Cheng, Jianpeng; Dong, Li; Lapata, Mirella, 2016, Proceedings of the 30th AAAI Conference on Artificial Intelligence (AAAI).

Incorrect Author List and Year

(c) Attention is all you need [29]

Fig. 2. Fabricated references produced by LLMs during the literature search process.

Our findings indicate that LLMs have significant potential as teaching aids in research methodology for both educators and student supervisors. Students can be trained to use LLM-based assistants to speed up their literature searches by drawing on the ability of the model to rapidly identify key and foundational papers on a topic. Instructors should place strong emphasis on critical thinking and verification so that students learn to cross-check LLM-suggested references for accuracy and relevance and remain fully aware of the limitations of these tools such as occasional citation errors. By systematically integrating these tools into coursework, research training, and thesis supervision with proper guidance, educators can help students develop efficient and well structured literature review skills while upholding high standards of scholarly accuracy and reliability. This method also familiarizes students with advanced digital research methods, preparing them to use similar tools responsibly in their future academic or professional work.

Looking ahead, an important direction is to extend this framework to other research domains such as biomedical sciences and social sciences, using larger and more diverse sets of papers to evaluate its generalizability. We anticipate that certain fields may present new challenges. For example, social science research often focuses on theoretical concepts and qualitative findings that are not easily represented by specific keywords, which may require LLMs to interpret context more deeply. Another important consideration is language diversity, as supporting literature searches in languages beyond English will be essential for fair access and to allow researchers around the world to benefit from these tools. A key next step is to benchmark language model-based literature search against established academic search engines such as Google Scholar and assess not only recall and precision but also retrieval efficiency and overall user experience. Future research may also explore hybrid systems that combine language model reasoning with structured database queries to create more reliable and interactive tools for literature discovery. With the recent release of GPT-5, it is expected that LLM-based literature search tools will reach even higher levels of accuracy, reliability, and inclusiveness, further strengthening their role as effective and valuable research assistants. As these models continue to evolve, the demand for more rigorous and meaningful evaluation frameworks will grow correspondingly, highlighting the timeliness and relevance of this study as a foundation for assessing next-generation LLM capabilities in literature discovery.

Acknowledgment

This work was supported in part by the Saint Francis University Institute Development Grant (Conference) (No. CL0141/IDGC250105), in part by the Singapore Ministry of Education Academic Research Fund RG91/22 and MOE-T2EP20224-0009, in part by the National Science and Technology Council, R.O.C, under Grant 114-2115-M-033-003-MY2, and in part by the Research Grants Council of Hong Kong under its Institutional Development Scheme Research Infrastructure Grant (No. UGC/IDS(R)11/21) and Collaborative Research Grant (No. UGC/IDS(C)11/E01/24).

References

[1] L. Bornmann, R. Haunschild, and R. Mutz, "Growth rates of modern science: A latent piecewise growth curve approach to model publication numbers from established and new literature databases," *Humanities and Social Sciences Communications*, vol. 8, no. 1, pp. 1–15, 2021.

[2] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, and H. Larochelle, "Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program)," *Journal of Machine Learning Research*, vol. 22, no. 164, pp. 1–20, 2021.

[3] C. N. Hang, P.-D. Yu, C. W. Tan, and D. M. Chiu, "When ideas go viral: Measuring scholarly novelty and viral influence via citation network analysis," in *GLOBECOM 2025-2025 IEEE Global Communications Conference*, 2025, pp. 1–6.

[4] G. Wagner, R. Lukyanenko, and G. Paré, "Artificial intelligence and the conduct of literature reviews," *Journal of Information Technology*, vol. 37, no. 2, pp. 209–226, 2022.

[5] Q. Jin, R. Leaman, and Z. Lu, "PubMed and beyond: Biomedical literature search in the age of artificial intelligence," *eBioMedicine*, vol. 100, 2024.

[6] S. A. Antu, H. Chen, and C. K. Richards, "Using LLM (large language model) to improve efficiency in literature review for undergraduate research," *Llm@ Aied*, pp. 8–16, 2023.

[7] C. N. Hang and S. M. Ho, "Personalized vocabulary learning through images: Harnessing multimodal large language models for early childhood education," in *2025 IEEE Integrated STEM Education Conference (ISEC)*, 2025, pp. 1–7.

[8] K. Huang, J. Guo, Z. Li, X. Ji, J. Ge, W. Li, Y. Guo, T. Cai, H. Yuan, R. Wang *et al.*, "MATH-Perturb: Benchmarking LLMs' math reasoning abilities against hard perturbations," *arXiv:2502.06453*, 2025.

[9] C. N. Hang, C. W. Tan, and P.-D. Yu, "MCQGen: A large language model-driven MCQ generator for personalized learning," *IEEE Access*, vol. 12, pp. 102 261–102 273, 2024.

[10] M.-F. Wong, S. Guo, C.-N. Hang, S.-W. Ho, and C.-W. Tan, "Natural language generation and understanding of big code for AI-assisted programming: A review," *Entropy*, vol. 25, no. 6, p. 888, 2023.

[11] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, and P. Fung, "Survey of hallucination in natural language generation," *ACM Computing Surveys*, vol. 55, no. 12, pp. 1–38, 2023.

[12] F. Liu, Y. Liu, L. Shi, H. Huang, R. Wang, Z. Yang, L. Zhang, Z. Li, and Y. Ma, "Exploring and evaluating hallucinations in LLM-powered code generation," *arXiv:2404.00971*, 2024.

[13] G. Sriramanan, S. Bharti, V. S. Sadasivan, S. Saha, P. Kattakinda, and S. Feizi, "LLM-check: Investigating detection of hallucinations in large language models," *Advances in Neural Information Processing Systems*, vol. 37, pp. 34 188–34 216, 2024.

[14] C. N. Hang, P.-D. Yu, and C. W. Tan, "TrumorGPT: Graph-based retrieval-augmented large language model for fact-checking," *IEEE Transactions on Artificial Intelligence*, 2025.

[15] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," *arXiv:2211.09085*, 2022.

[16] OpenAI, "GPT-4 technical report," *arXiv:2303.08774*, 2023.

[17] G. Comanici, E. Bieber, M. Schaekermann, I. Pasupat, N. Sachdeva, I. Dhillon, M. Blistein, O. Ram, D. Zhang, E. Rosen *et al.*, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," *arXiv:2507.06261*, 2025.

[18] A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang, C. Ruan *et al.*, "DeepSeek-V3 technical report," *arXiv:2412.19437*, 2024.

[19] I. Vessey, V. Ramesh, and R. L. Glass, "Research in information systems: An empirical study of diversity in the discipline and its journals," *Journal of Management Information Systems*, vol. 19, no. 2, pp. 129–174, 2002.

[20] X. Bai, M. Wang, I. Lee, Z. Yang, X. Kong, and F. Xia, "Scientific paper recommendation: A survey," *IEEE Access*, vol. 7, pp. 9324–9339, 2019.

[21] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: A literature survey," *International Journal on Digital Libraries*, vol. 17, no. 4, pp. 305–338, 2016.

[22] S. Aum and S. Choe, "srBERT: Automatic article classification model for systematic review using BERT," *Systematic Reviews*, vol. 10, no. 1, p. 285, 2021.

[23] Q. Khraisha, S. Put, J. Kappenberg, A. Warraitch, and K. Hadfield, "Can large language models replace humans in systematic reviews? Evaluating GPT-4's efficacy in screening and extracting data from peer-reviewed and grey literature in multiple languages," *Research Synthesis Methods*, vol. 15, no. 4, pp. 616–626, 2024.

[24] M. Mostafapour, J. H. Fortier, K. Pacheco, H. Murray, and G. Garber, "Evaluating literature reviews conducted by humans versus ChatGPT: Comparative study," *JMIR AI*, vol. 3, p. e56537, 2024.

[25] M. Bhattacharyya, V. M. Miller, D. Bhattacharyya, L. E. Miller, and V. Miller, "High rates of fabricated and inaccurate references in ChatGPT-generated medical content," *Cureus*, vol. 15, no. 5, 2023.

[26] J. Gravel, M. D'Amours-Gravel, and E. Osmanlliu, "Learning to fake it: Limited responses and fabricated references provided by ChatGPT for medical questions," *Mayo Clinic Proceedings: Digital Health*, vol. 1, no. 3, pp. 226–234, 2023.

[27] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," *Advances in Neural Information Processing Systems*, vol. 27, 2014.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[29] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[30] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *arXiv:1409.0473*, 2014.