Large Language Models for Automated Literature Review: An Evaluation of Reference Generation, Abstract Writing, and Review Composition

Xuemei Tang[1] **Xufeng Duan**[3]* **Zhenguang G. Cai**[2,3]*
[1]The Department of Language Science and Technology, The Hong Kong Polytechnic University
[2]Department of Linguistics and Modern Languages, The Chinese University of Hong Kong
[3]Brain and Mind Institute, The Chinese University of Hong Kong
xuemeitang00@gmail.com
{xufengduan, zhenguangcai}@cuhk.edu.hk

Abstract

Large language models (LLMs) have emerged as a potential solution to automate the complex processes involved in writing literature reviews, such as literature collection, organization, and summarization. However, it is yet unclear how good LLMs are at automating comprehensive and reliable literature reviews. This study introduces a framework to automatically evaluate the performance of LLMs in three key tasks of literature review writing: reference generation, abstract writing, and literature review composition. We introduce multidimensional evaluation metrics that assess the hallucination rates in generated references and measure the semantic coverage and factual consistency of the literature summaries and compositions against human-written counterparts. The experimental results reveal that even the most advanced models still generate hallucinated references, despite recent progress. Moreover, we observe that the performance of different models varies across disciplines when it comes to writing literature reviews. These findings highlight the need for further research and development to improve the reliability of LLMs in automating academic literature reviews. The dataset and code used in this study are publicly available in our GitHub repository [1].

1 Introduction

The literature review is a critical component of academic writing that aims to synthesize, critique, and assess the current state of knowledge in a particular field. It involves a comprehensive examination of published research articles, theoretical frameworks, and research methodologies related to a specific topic. Conducting a thorough literature review often necessitates extensive reading and summarizing of pertinent literature, which can be a complex and time-consuming process, especially in well-established fields where the number of relevant references can range from dozens to hundreds. To alleviate this burden, researchers have recently turned to advanced deep learning models as a potential tool to aid in the automated generation of literature reviews (Aliyu et al., 2018; Kontonatsios et al., 2020).

The emergence of LLMs has introduced a promising avenue for automating key aspects of literature review writing, including identifying relevant sources, summarizing findings, and generating coherent syntheses (Wang et al., 2024b; Agarwal et al., 2024; Hsu et al., 2024).

While techniques such as Retrieval-Augmented Generation (RAG) can enhance the domain-specific knowledge of LLMs—by providing access to real literature databases and helping generate more accurate content—in practice, most researchers still rely on vanilla LLMs, such as ChatGPT, for literature review writing without the use of RAG (Wang et al., 2024a). Consequently, it is crucial to evaluate the performance of these naive LLMs in the context of literature review writing to determine their effectiveness and limitations.

Therefore, in this paper, we propose a framework for automatically assessing the literature review writing ability of LLMs, using human-written literature reviews as the gold standard and designing metrics for a comprehensive evaluation. We first collect a dataset of human-written literature reviews to serve as a benchmark for evaluating the performance of LLMs. We then ask LLMs to complete three tasks based on the collected dataset: generating references, writing an abstract, and writing a complete literature review based on a given topic. Finally, we evaluate the generated results from several dimensions, including the presence of hallucinations in the references, as well as the semantic coverage and factual consistency of the generated abstract and literature review compared

*Corresponding Author
[1]https://github.com/tangxuemei1995/Eval_LLM_Literature_Review

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 1603–1618
November 4-9, 2025

to the human-written context. By assessing the performance of LLMs across these tasks and evaluating their output using our proposed metrics, we aim to provide a comprehensive understanding of their capabilities and limitations in writing literature reviews.

Our contribution can be summarized as follows.

- First, we propose a framework for automatically evaluating the literature review writing ability of LLMs, without requiring any human involvement. This framework encompasses multiple stages, including the compilation of a literature review dataset construction, the collection of LLM-generated output, and the evaluation of LLM performance.
- Second, we collect 1,105 literature reviews from 51 journals across six disciplines as the ground truth. We then design three tasks for accessing LLMs in literature writing: reference generation, abstract writing, and literature composition on a given topic.
- Then, we evaluate the generated results of LLMs from multiple perspectives, including the hallucination rate in generated references, factual consistency, and semantic coverage compared to human-written content.
- Finally, we assess five LLMs using the proposed framework. By analyzing the experimental results, we find that hallucinated references remain a prevalent issue for current LLMs. Furthermore, the performance of LLMs in writing literature reviews varies across different disciplines.

2 Related Work

Recent studies have explored LLMs for literature review generation. For example, Wang et al. (2024b) proposed AutoSurvey, which incorporates up-to-date papers via retrieval-augmented generation. Agarwal et al. (2024) examined zero-shot LLM review generation using a two-step retrieval and outlining process. More recently, Liang et al. (2025) presented SurveyX, an efficient system that optimizes retrieval, extraction, and outline generation, supporting multimodal outputs such as figures and tables.

Additionally, recent efforts to evaluate literature review generation by LLMs have increasingly focused on assessing hallucinations in reference citations. For instance, Chelli et al. (2024) analyzed hallucination rates in 11 systematic reviews on shoulder rotator cuff pathology generated by ChatGPT, GPT-4, and Bard, finding Bard exhibited significantly higher hallucination rates. Similarly, Agrawal et al. (2024) evaluated hallucinations across 200 computer science topics by generating reference titles with LLMs and verifying their existence via the BING Search API, further probing whether LLMs could detect hallucinated references through direct and indirect queries. Athaluri et al. (2023) examined hallucinations in 50 ChatGPT-generated research proposals, manually validating references and DOIs using Scopus, Google, and PubMed, reporting 109 valid DOIs among 178 references. Additionally, Aljamaan et al. (2024) introduced the Reference Hallucination Score (RHS) by generating references for five medical topics across multiple LLMs, assigning weighted hallucination scores to citation components such as title and publication date. While these studies provide valuable insights, they are generally limited to specific domains and rely heavily on manual evaluation, lacking a comprehensive, scalable assessment framework for LLM reference hallucinations.

3 Methodology

In this section, we propose a framework for evaluating LLMs' literature review writing ability. The framework, as shown in Figure 1, consists of three main stages: dataset construction and task design for evaluation, collection LLM-generated output, and assessment of the generated output.

3.1 Dataset Construction

Assessing the ability to write literature reviews is a challenging task, as evaluating the quality of content is inherently complex. In this paper, we use human-written reviews as the gold standard, which simplifies the evaluation process to some extent. As illustrated in Figure 1, we first collect publicly available information of literature reviews (i.e., the title, authors, abstract, keywords, and content) from the Annual Reviews website [2]. Annual Reviews, an independent nonprofit publisher, produces 51 review journals spanning various scientific disciplines. Invited experts write comprehensive, authoritative reviews that synthesize and summarize the most significant primary research literature in their field, providing a valuable resource for re-

[2] https://www.annualreviews.org/

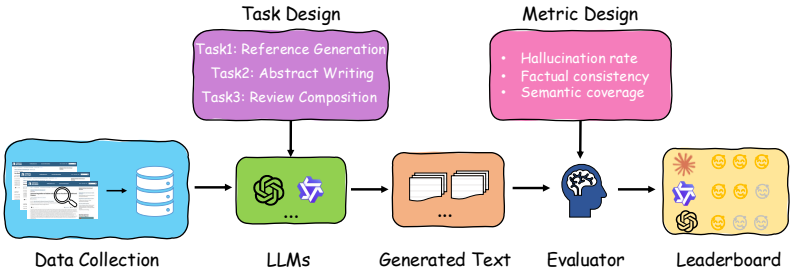


Figure 1: Illustration of the evaluation framework.

searchers to stay current with the latest developments. We crawl all articles published in 2023, including their title, keywords, abstracts, contents, and references, and then clean them to create the experimental dataset.

Then, the dataset D is the article set from 51 journals, $D = \{p_0, ..., p_i, ..., p_M\}$, where M represents the number of articles. Each article $p_i = \{t_i, w_i, a_i, c_i, R_i\}$, where t_i, w_i, a_i, c_i, R_i represent the title, keywords, abstract, context, and reference set $R_i = \{r_1, ..., r_k, ..., r_K\}$, and K represents the length of the reference set.

3.2 Task Design

Since literature review writing primarily involves the collection and synthesis of relevant research, we design three independent tasks as follows to evaluate LLMs' capabilities in different aspects of literature review writing.

- **Reference Generation:** Given the article title t_i and keywords w_i, ask LLMs to find the N most relevant studies $R_i^g = \{r_1^g, ..., r_n^g, ..., r_N^g\}$ to the research topic. Each citation study must include 7 metadata elements: title, authors, journal, year, volumes, first page, and last page, $r_n^g = \{T, A, J, Y, V, FP, LP\}$. In this task, we evaluate whether LLMs can recommend reliable references based on the given topic. *Note that these references are not reused in later tasks.*

- **Abstract Writing:** Given an article title t_i and its associated keywords w_i, the LLMs are prompted to generate an abstract a_i^g that aligns with the research topic. The length of the generated abstract is constrained to match that of the original. This task serves as a proxy for literature review planning, as abstracts often outline the key components of a study—such as its objectives, methods, and covered subtopics—which are also critical in structuring comprehensive literature reviews. By evaluating the model's ability to generate coherent and topic-relevant abstracts, we assess its potential to assist researchers in the early planning stages of literature review writing.

- **Review Composition:** Given the article t_i and keywords w_i, and abstract a_i, ask LLMs to write a short literature review c_i^g according to the research topic provided in the title, keywords, and abstract. To facilitate evaluation and accommodate computational budget constraints, the length of each literature review is limited to approximately 1000 words. LLMs also need to back up claims by citing relevent studies $R_i^g = \{r_1^g, ..., r_n^g, ..., r_N^g\}$ (with a total of N citations in the literature review). These citations are newly generated to support the content of the review. In this task, we evaluate whether LLMs can write a high-quality literature review and cite truth studies.

We designed the three tasks as independent tasks for two main reasons: ***a.*** Each task has a different goal. Task 1 (Reference Generation) aims to evaluate the LLMs' ability to recommend relevant papers, which is a common practical use case — retrieving relevant literature based on a specific topic. Task 2 (Abstract Writing) evaluates the LLMs' ability to outline and plan a literature review through abstract writing. Task 3 (Review Composition) assesses whether LLMs can organize and synthesize multiple sources into a coherent review while also providing verifiable references. ***b.*** We intentionally separated the tasks to avoid cross-task interference, which would make it difficult to isolate and evaluate specific capabilities of LLMs.

Three task prompts are shown in Appendix Table 5.

3.3 Evaluation Metrics

Based on the type of generated text, we divide the evaluation of the model's results into two parts: first, the hallucination rate of the references generated by LLMs, and second, a comparison of the generated context with human-written results, including two dimensions: factual consistency and semantic coverage.

Reference hallucination evaluation metrics. Given that LLMs are trained on vast corpora, including academic sources, we aim to evaluate whether they can generate true references. In this section, we introduce the calculation process of the reference precision ***Precision***, reference overlap rate with human-cited references ***Overlap rate***, and title search rate $\boldsymbol{S_t}$ for each LLM. **A higher precision metric indicates a lower hallucination rate**. A higher overlap indicates that the LLM-generated references cover more of the ground-truth citations used by human authors, reflecting a better ability to identify key prior work relevant to the topic.

For each article $p_i \in D$, each LLM generates N references $R_i^g = \{r_1^g, ..., r_n^g, ..., r_N^g\}$ in both Reference Generation and Review Composition tasks, each r_n^g and includes 7 elements. Each element corresponding to a state label represents whether it is accurate or not $\{e_d\}_{d=0}^6$, e_d = 1 or 0.

Next, we describe how to obtain $\{e_d\}_{d=0}^6$. First, we use the generated titles T and the first author in A as the queries and search them separately from external academic search engines. This results in two sets of candidate articles, Z_t and Z_a respectively, We then merge the two sets and remove duplicates to obtain the final candidate set $Z = \{z_1, ..., z_j, ..., z_J\}$. Subsequently, we compare the generated r_n^g with the article z_j from candidate sets Z. For example, if the title of a candidate article z_j matches the title of r_n^g, then $e_0 = 1$. Finally, we find the best candidate article based on the sum of $\{e_d\}_{d=0}^6$, and the one with the largest sum is the best candidate article z_j of r_n^g.

Then, we compare the alignment degree between the generated reference r_n^g and the best-matching candidate article z_j to determine whether r_n^g corresponds to a real article (as shown in Eq. 1). We consider r_n^g to be reliable under either of the following two conditions:

Title-based matching: If the title T is correct (i.e., $e_0 = 1$), and at least one other metadata element (e.g., author, journal, year, etc.) also matches, the reference is deemed reliable. To allow for minor variations, we consider the title to be correct if it achieves a match rate of at least 80% with the ground-truth title—a threshold determined through human evaluation.

Metadata-based matching: If the title T is incorrect (i.e., $e_0 = 0$), we still consider the reference reliable if at least three of the remaining metadata elements (author, journal, year, volume, first page, last page) match those of a real article. This allows us to identify true references even when the title is noisy or incomplete.

$$\text{True}(r_n^g) = \begin{cases} 1 & \text{if } \left(e_0 = 1 \text{ and } \sum_{i=1}^6 e_i \geq 1\right) \\ & \text{or } \left(e_0 = 0 \text{ and } \sum_{i=1}^6 e_i \geq 3\right) \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

For each paper p_i in the dataset, we compute the reference precision of the LLM-generated references, denoted as $\textit{Precision}(p_i)$, as defined in Eq. 2. We then obtain the overall *Precision* score for each LLM by averaging $\textit{Precision}(p_i)$ across all papers in the dataset, as shown in Eq. 3.

$$Precision_{(p_i)} = \frac{1}{N} \sum_{n=0}^{N} \text{True}(r_n^g) \quad (2)$$

$$Preicison = \frac{1}{M} \sum_{i=0}^{M} Precison(p_i) \quad (3)$$

Precision is measured by comparing the LLM-generated references with external academic databases. We also evaluate *Overlap rate* by comparing the references generated by the LLM with those cited in the human-written original articles. The key difference between precision and overlap rate in our setting lies in the candidate set Z: for precision, Z is constructed from external academic search results, whereas for overlap rate, Z consists of the references actually cited in the human-written articles.

Additionally, the title is intuitively the most critical element in determining the faithfulness of a generated reference. In the work of Agrawal et al. (2024), ground-truth labels were assigned based on results returned by the Bing Search API. Inspired by their approach, we also calculate the title search score for each LLM to estimate how many generated titles correspond to real publications.

$$S_t = \frac{1}{MN} \sum_{M} \sum_{N} s_{p_i}^{(n)} \quad (4)$$

$$s_{p_i}^{(n)} = \sum_{r_n^g \in R_i^g} \begin{cases} 1 & \text{if the } T \in r_n^g \text{ has return value} \\ & \text{from external Scholar API,} \\ 0 & \text{otherwise} \end{cases} \quad (5)$$

Here, $s_{p_i}^{(n)}$ indicates whether the generated title in reference r_n^g for paper p_i can be found using an external academic search engine. This metric helps estimate the proportion of references with verifiable titles among the total generated references.

Context evaluation metrics. In our study, we use the human-written article as the gold truth and then evaluate LLM-generated context from **factual consistency** and **semantic coverage** aspects. The resemblance of natural language inference (NLI) to factual consistency evaluation has led to utilizing NLI models for measuring factual consistency (Gao et al., 2023). Encouraged by previous works, we also use the NLI method to evaluate the factual consistency between LLMs generated and human-written text. For example, we calculate the NLI score $Entail_{p_i}$ between the original article abstract a_i and the LLM-generated abstract a_i^g as follows.

$$Entail_{p_i} = \theta_{\text{NLI}}(a_i^g, a_i) = \begin{cases} 1 & \text{if } a_i^g \text{ entails } a_i, \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

where θ_{NLI} denotes the NLI model. Finally, we obtain the NLI score $Entail$ for each model according to Eq 7.

$$Entail = \frac{1}{M} \sum_{i=0}^{M} Entail_{p_i} \quad (7)$$

Additionally, we use commonly employed **semantic similarity metrics** and **Key Point Recall (KPR)** to calculate the semantic coverage between the context generated by LLMs and human-written context. Specifically, for the Abstract Writing task, we apply cosine similarity and the ROUGE metric for semantic coverage evaluation. For the Review Composition task, we use the ROUGE metric and KPR to measure the semantic coverage of the literature review generated by the LLMs relative to human-written content.

KPR, first proposed by Qi et al. (2024), is a metric designed to evaluate the effectiveness of LLMs in utilizing RAG for long documents. Since human-written literature reviews are lengthy and difficult to compare directly, we adopt the KPR method to measure the extent to which LLM-generated content covers the key points in human-written literature reviews. Specifically, we first use GPT-4 to extract q key points $X_i = [x_{i_1}, x_{i_2}, ..., x_{i_q}]$ from the human-written literature review c_i, and then calculate the coverage of these key points by the model-generated literature review as Eq. 8.

$$KPR = \frac{1}{M} \sum_{i=0}^{M} \frac{\sum_{x \in X_i} \theta_{\text{NLI}}(c_i^g, x)}{|X_i|} \quad (8)$$

where c_i^g denotes the literature review generated by LLMs.

Finally, we also concatenate key points and compute the ROUGE metric between the key points and c_i^g.

4 Experiments

4.1 Experimental Settings

Dataset. We collect 1,105 literature review articles published in 2023 from the Annual Reviews website. The distribution of articles across journals is shown in Appendix B, Figure 5.

LLMs Selection. We evaluate five LLMs: **Claude-3.5-Sonnet-20240620**, **GPT-4o-2024-08-16**, **Qwen-2.5-72B-Instruct**, **DeepSeek-V3**, and **Llama-3.2-3B-Instruct**. All model outputs were generated via their official APIs with temperature set to 0 for consistency.

In Reference Generation and Review Composition tasks, we set N as 10, each model generates 10 references. For the generated reference evaluation, we use **Semantic Scholar** as the external database. Recently, LLMs-as-judges has become more common (Chen et al., 2024; Zheng et al., 2023; Shangyu et al., 2024). So, for the Abstract Writing task, we employ **TRUE** (Honovich et al., 2022), along with **GPT-4o** as NLI models for evaluating factual consistency in context; to compute semantic similarity, we use **text-embedding-3-large** to convert texts into embeddings. For the Review Composition task, we employ **GPT-4o** as the NLI model in Eq 8, and set q as 10.

4.2 Main Results

We present the results for the three tasks in Table 1, 2, and 3. The performance of different models on each task is analyzed as follows.

Results for Reference Generation. As shown in Table 1, Claude-3.5-Sonnet achieves the highest precision, overlap rate, and S_t, while Llama-3.2-3B performs the worst on three metrics. When evaluating the author dimension of LLM-generated references, we consider the reference to match in this dimension if the first author is correctly matched. Applying this criterion results in a 1–3% increase in precision scores across all models. This suggests that generating complete and accurate author lists remains a major challenge for LLMs.

We further conduct a year-wise analysis of the correctly generated references, as illustrated in Figure 2. The results reveal that the majority of accu-

rate citations produced by the models are concentrated in the period between 2010 and 2020, a trend consistent across nearly all LLMs evaluated in this task.

Results for Abstract Writing. As shown in Table 2, Claude-3.5-Sonnet achieves the best overall performance across most evaluation metrics. It generates abstracts with the highest average semantic similarity to human-written ones (81.17%) and shows strong factual consistency, achieving a TRUE score of 78.10%. DeepSeek-V3 also performs well in factual consistency, with the highest GPT-4o-based assessment score (96.84%). In contrast, Llama-3.2-3B obtains the highest ROUGE-L score but does not show clear advantages on other metrics. These results highlight the importance of using multiple evaluation metrics to comprehensively assess the diverse outputs of LLMs.

Results for Review Composition. As shown in Table 3, compared to the Reference Generation task, all LLMs demonstrate a significant increase in precision when generating references within the Review Composition task. Prior research indicates that grounding generated text with real external citations can effectively reduce hallucination rates (Gao et al., 2023). Consistently, our experiments reveal that when LLMs generate references alongside the literature review, the accuracy of these references improves markedly. This suggests a mutual constraint between the generated references and the review text, leading to enhanced overall reliability.

On the other hand, Claude-3.5-Sonnet achieves the highest performance on the *KPR* metric, indicating that its generated literature reviews recall the greatest number of claims from the human-written versions. Meanwhile, the literature reviews produced by DeepSeek-V3 excel on the ROUGE metrics, demonstrating stronger overlap with reference texts in terms of lexical similarity.

4.3 Analyze LLM-Generated References from Different Dimensions

In both Reference Generation and Review Composition tasks, we ask LLMs to generate references. The overall performance was discussed in the previous section. In this section, we provide a detailed comparison of the accuracy of LLM-generated references across various dimensions, as shown in Figure 3. As seen in Figure 3(a), Claude-3.5-Sonnet demonstrates a clear advantage over other models across all dimensions in the Reference Generation

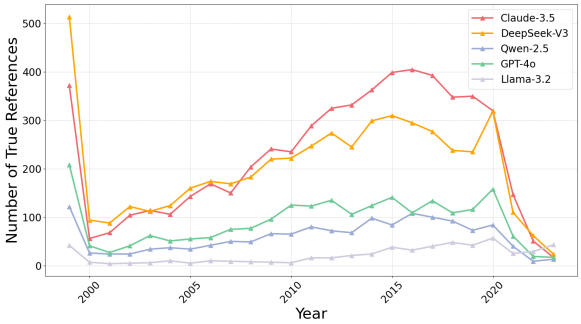
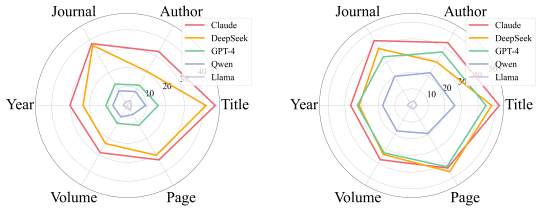


Figure 2: Distribution of LLM-generated true references over years.



(a) Reference Generation (b) Review Composition

Figure 3: Radar chart of the accuracy of LLM-generated references across various dimensions.

task. Additionally, the accuracy of reference generation in the Reference Generation task for Claude-3.5-Sonnet, GPT-4o, and Qwen-2.5-72B follows a consistent trend across all dimensions, with the highest accuracy observed in the title dimension. Accuracy for journal name, page, and author is also relatively high. However, DeepSeek-V3 performs worse in the author dimension compared to the other dimensions. In contrast, Llama-3.2 demonstrates higher accuracy in the page and author dimensions than in other dimensions. However, overall, Llama-3.2-3B does not exhibit a competitive advantage in reference generation accuracy.

Next, we examine the accuracy of LLM-generated references across various dimensions in Review Composition, as shown in Figure 3(b), and comparing it with Figure 3(a), we observe improvements across all dimensions for Claude-3.5-Sonnet, DeepSeek-V3, GPT-4o, and Qwen-2.5-72B, with particularly obvious gains in the author dimension. The possible reason is that, in the generated text, the LLMs tend to cite the first author's name, which may lead the models to place more emphasis on this dimension. Notably, the accuracy of DeepSeek-V3 and GPT-4o in certain dimensions approaches or even exceeds that of Claude-3.5-Sonnet. However, the performance of LLaMA-3.2-3B remains

Models	Reference Generation				
	S_t↑	*P*	*Overlap*	*P (first author)*	*Overlap (first author)*
Qwen-2.5-72B	21.80	12.25	12.60	17.58	13.27
Llama-3.2-3B	16.62	3.45	8.48	6.95	8.67
DeepSeek-V3	56.04	46.33	19.72	50.66	20.50
GPT-4o	32.07	21.65	18.76	24.65	19.50
Claude-3.5-Sonnet	**64.82**	**51.59**	**24.34**	**55.77**	**25.21**

Table 1: The experimental results of the five LLMs in Reference Generation. "S_t" refers to the title search rate as defined in Eq 4, while "*P*" represents the *Precision*, "*Overlap*" denotes the *Overlap rate*. "first author" refers to when evaluating the accuracy of references, the author dimension only comparing the first author.

Models	Abstract Writing					
	Similarity↑	Entail(TRUE)↑	Entail(GPT-4o)↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
Qwen-2.5-72B	80.22	69.52	95.02	40.61	8.78	20.12
Llama-3.2-3B	79.28	62.39	92.14	40.35	8.96	**20.52**
DeepSeek-V3	80.96	78.55	**96.84**	**41.13**	8.98	20.33
GPT-4o	80.96	77.91	96.50	40.70	8.56	19.86
Claude-3.5-Sonnet	**81.17**	**78.90**	96.77	**41.13**	**8.99**	20.00

Table 2: Compare the performance of four LLMs on Abstract Writing.

Models	Review Composition								
	References					Literature Review			
	S_t↑	*P*	*Overlap*	*P(first author)*	*Overlap(first author)*	*KPR*↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
Qwen-2.5-72B	40.02	28.91	17.36	33.64	18.31	38.82	29.95	9.01	15.14
Llama-3.2-3B	21.78	4.86	8.28	7.28	8.44	29.07	28.07	7.77	15.46
DeepSeek-V3	62.29	52.81	26.79	55.38	27.30	56.02	**35.65**	**10.40**	**17.46**
GPT-4o	60.05	50.62	27.88	54.16	28.86	59.18	30.78	9.72	15.54
Claude-3.5-Sonnet	**66.43**	**59.06**	**31.90**	**63.06**	**33.25**	**62.32**	28.59	8.90	14.41

Table 3: The experimental results of the four LLMs in Review Composition. "S_t" refers to the title retrieval rate as defined in Eq 4. While "*P*" represents the *Precision*, "*Overlap*" denotes the *Overlap rate*. "*KPR*" means the *Key Point Recall* rate."first author" refers to when evaluating the accuracy of references, the author dimension only comparing the first author.

Discipline	Citation Count		Precision	
	DeepSeek	Claude	DeepSeek	Claude
Biology	763	678	55.55	58.00
Mathematics	**2288**	**1984**	**60.00**	**62.22**
Physics	894	652	47.62	56.19
Chemistry	1334	1079	43.14	43.80
Social Science	1321	1151	46.80	56.70
Technology	904	748	44.88	49.01

Table 4: Average citation counts and reference precision across disciplines.

suboptimal.

4.4 Cross-Disciplinary Analysis

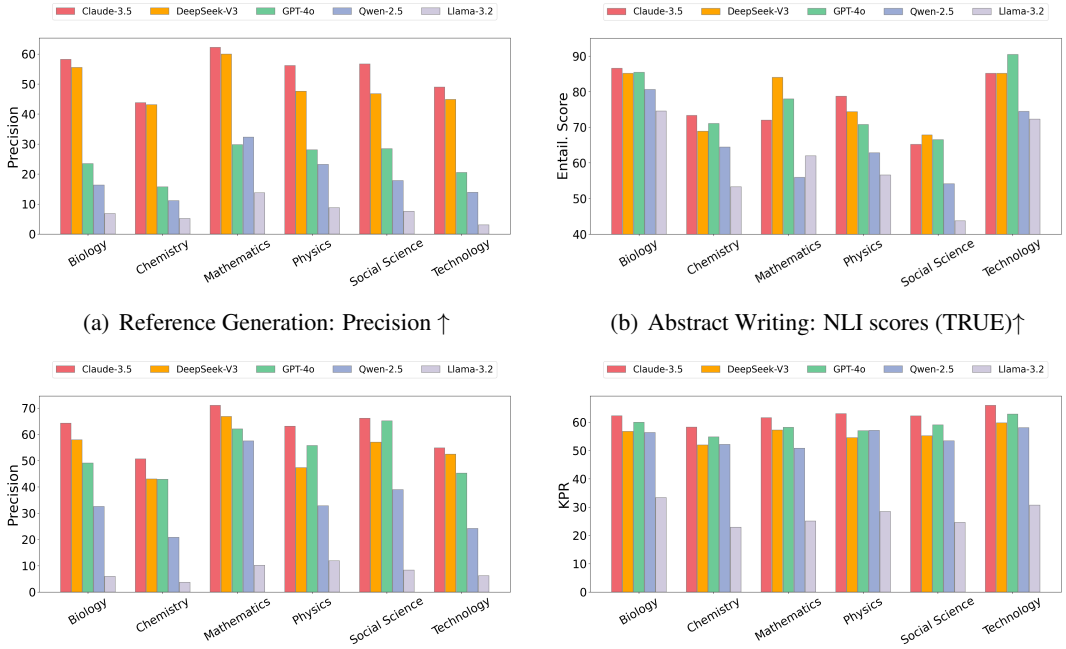
In this section, we compare the performance of LLMs across different disciplines. First, based on Dewey's Decimal Classification, we categorize 51 journals into six disciplines: Biology, Chemistry, Mathematics, Physics, Social Science, and Technology. After categorization, there are 460 articles in the Biology category, 90 in Chemistry, 50 in Mathematics, 113 in Physics, 299 in Social Science, and 94 in Technology.

We then present bar charts in Figures 4, which illustrate the performance of different models across various tasks and disciplines.

First, we observe that in the Reference Generation task, as shown in Figure 4(a), almost all models exhibit the highest precision in the Mathematics discipline and the lowest precision in the Chemistry discipline. To validate these differences, we conduct one-way ANOVA tests for each LLM across five disciplines. Significant differences are found for all models except Llama-3.2-3B. Detailed ANOVA results are reported in Appendix I.

Secondly, as shown in Figure 4(b), the NLI scores evaluated by TRUE in the Abstract Writing task indicate that all models perform the worst in Social Science. GPT-4o performs best in Technology, while Claude 3.5-Sonnet achieves the highest performance in Biology. One-way ANOVA tests reveal significant differences across disciplines for all models. See Appendix I for detailed results.

Thirdly, we examine the references precision of each model across five disciplines in the Review Composition task, as illustrated in Figure 4(c). It is evident that the precision of Claude 3.5 Sonnet, DeepSeek-V3, and GPT-4o is significantly higher



(a) Reference Generation: Precision ↑

(b) Abstract Writing: NLI scores (TRUE)↑

(c) Review Composition: Precision ↑

(d) Review Composition: KPR scores ↑

Figure 4: Three tasks evaluation scores across different disciplines.

than that of Qwen-2.5-72B and LLaMA-3.2-3B across all disciplines. Furthermore, Claude 3.5 Sonnet, DeepSeek-V3, and Qwen-2.5-72B exhibit the highest precision in Mathematics, while GPT-4o performs best in Social Science. ANOVA tests confirm significant differences across disciplines for all models (see Appendix I).

Finally, we observe the KPR metric across different disciplines in Review Composition, as shown in Figure 4(d). The results from the figure indicate that the differences between models—Claude-3.5-Sonnet, DeepSeek-V3, GPT-4o, and Qwen-2.5-72B—are not significant across various disciplines, a finding that is also supported by statistical tests (see Appendix I).

Citation Frequency and Precision Across Disciplines. Additionally, we report statistics on the citation frequency of correctly generated references by LLMs in the Reference Generation task, as shown in Table 4, using Claude-3.5 and DeepSeek-V3 as examples. The data indicates that the references generated by the LLMs are highly cited, which might be due to their frequent presence in online sources, making them more likely to appear in the LLMs training datasets. As a result, LLMs tend to generate more accurate metadata (e.g., author, year) for these well-known references.

Furthermore, when analyzing different disciplines, we observe that the Mathematics discipline has the highest precision, and the relevant references generated for Mathematics also have the highest citation count. We compute the correlation between citation precision and average citation counts, finding that the correlation coefficient for Claude-3.5 is 0.4, and for DeepSeek-V3 it is 0.51, indicating a positive relationship between the two.

4.5 Human Evaluation

To evaluate the reliability of our automatic assessment method for identifying hallucinated references, we conduct a comparative analysis involving 100 LLM-generated references. These references were assessed by three annotators and the final manual results were obtained by majority vote. The results demonstrated a kappa agreement of 0.71 between the automatic and human assessments, signifying a relatively high level of consistency and supporting the reliability of our method. Furthermore, when using human assessment results as the gold standard, the automatic assessment method achieved an accuracy of 86%, further validating its effectiveness.

5 Conclusion

In this paper, we present a framework to assess the literature review writing abilities of LLMs. This framework includes three tasks designed to evaluate LLMs' literature review writing capabil-

ities. The generated outputs are then evaluated from multiple dimensions using various tools, such as Semantic Scholar and NLI models, focusing on aspects like hallucination rate, semantic coverage, and factual consistency compared to human-written texts. Finally, we analyze the performance of LLMs in writing literature reviews from the perspective of different academic disciplines.

Limitations

In this paper, we evaluate the ability of LLMs to write literature reviews. However, several limitations remain:

First, instead of evaluating the generated reviews from conventional perspectives such as fluency or topic coverage, we primarily compare LLM-generated results with human-written ones. As such, our current evaluation metrics may not be comprehensive. In the future, we plan to incorporate additional aspects of review quality to improve the completeness of our evaluation. These may include the coverage of cited works (i.e., whether the review offers a comprehensive overview of the relevant field) and the coherence of the overall structure (i.e., whether the review is organized in a way that facilitates information-seeking).

Second, there is a possibility that our test data overlaps with the training data of the LLMs. When we initiated this study in August 2024, the dataset from the Annual Reviews website had not yet been updated to include 2024 articles, so we relied on the complete 2023 dataset. To further address potential data contamination, we also conducted an additional evaluation using 2025 data to test GPT-5 (as shown in Appendix E), whose knowledge cutoff is September 2024. To mitigate potential data leakage more broadly, we plan to deploy a leaderboard on Hugging Face to continuously evaluate the performance of various LLMs in literature review writing, with real-time updates to the test dataset. However, due to the rapid iteration of LLMs, data leakage cannot be completely ruled out. That said, our experimental results—particularly those related to reference generation—show that all models still perform poorly. If data contamination were present, the actual scores would likely be lower than those reported. This reinforces, rather than undermines, our conclusion that significant challenges remain in using LLMs for literature review generation.

Additionally, when processing LLM-generated outputs, we often encountered abbreviated author names and journal titles. Although these issues have been carefully addressed (see Appendix F), minor discrepancies may remain.

Finally, to verify the precision of LLM-generated references, we primarily used Semantic Scholar as our auxiliary tool. Although we also experimented with Google Scholar, its lack of an accessible API led us to rely on the freely available Semantic Scholar API for consistency and ease of access. However, this may have resulted in incomplete reference retrieval.

Ethics Statement

The human evaluations conducted in this study were carried out by members of the research team. No personal or sensitive information was collected, and all participants were fully informed of the purpose of the evaluation. Therefore, the study does not raise any ethical concerns.

Acknowledgements

The research was supported by a direct grant from the Faculty of Arts, the Chinese University of Hong Kong. We thank Yicheng Li for his valuable assistance with data collection.

References

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2024. Llms for literature review: Are we there yet? (arXiv:2412.15249). ArXiv:2412.15249 [cs].

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Tauman Kalai. 2024. Do language models know when they're hallucinating references?

Muhammad Bello Aliyu, Rahat Iqbal, and Anne James. 2018. The canonical model of structure for data extraction in systematic reviews of scientific research articles. In *2018 Fifth International Conference on Social Networks Analysis, Management and Security (SNAMS)*, page 264–271.

Fadi Aljamaan, Mohamad-Hani Temsah, Ibraheem Altamimi, Ayman Al-Eyadhy, Amr Jamal, Khalid Alhasan, Tamer A. Mesallam, Mohamed Farahat, and Khalid H. Malki. 2024. Reference hallucination score for medical artificial intelligence chatbots: Development and usability study. *JMIR Medical Informatics*, 12(1):e54345. Company: JMIR Medical Informatics Distributor: JMIR Medical Informatics Institution: JMIR Medical Informatics Label: JMIR Medical Informatics publisher: JMIR Publications Inc., Toronto, Canada.

Sai Anirudh Athaluri, Sandeep Varma Manthena, V S R Krishna Manoj Kesapragada, Vineel Yarlagadda, Tirth Dave, and Rama Tulasi Siri Duddumpudi. 2023. Exploring the boundaries of reality: Investigating the phenomenon of artificial intelligence hallucination in scientific writing through chatgpt references. *Cureus*, 15(4):e37432.

Mikaël Chelli, Jules Descamps, Vincent Lavoué, Christophe Trojani, Michel Azar, Marcel Deckert, Jean-Luc Raynier, Gilles Clowez, Pascal Boileau, and Caroline Ruetsch-Chelli. 2024. Hallucination rates and reference accuracy of chatgpt and bard for systematic reviews: Comparative analysis. *Journal of Medical Internet Research*, 26:e53164.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. (arXiv:2402.10669). ArXiv:2402.10669 [cs].

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. (arXiv:2305.14627). ArXiv:2305.14627 [cs].

Or Honovich, Roee Aharoni, Jonathan Herzig, Hagai Taitelbaum, Doron Kukliansy, Vered Cohen, Thomas Scialom, Idan Szpektor, Avinatan Hassidim, and Yossi Matias. 2022. TRUE: Re-evaluating factual consistency evaluation. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3905–3920, Seattle, United States. Association for Computational Linguistics.

Chao-Chun Hsu, Erin Bransom, Jenna Sparks, Bailey Kuehl, Chenhao Tan, David Wadden, Lucy Wang, and Aakanksha Naik. 2024. CHIME: LLM-assisted hierarchical organization of scientific studies for literature review support. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 118–132, Bangkok, Thailand. Association for Computational Linguistics.

Georgios Kontonatsios, Sally Spencer, Peter Matthew, and Ioannis Korkontzelos. 2020. Using a neural network-based feature extraction method to facilitate citation screening for systematic reviews. *Expert Systems with Applications: X*, 6:100030.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu li. 2025. Surveyx: Academic survey automation via large language models. (arXiv:2502.14776). ArXiv:2502.14776 [cs].

Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. 2024. Long^2rag : $Evaluatinglong-contextlong-formretrieval-augmentedgenerationwithkeypointrecall$. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, page 4852–4872, Miami, Florida, USA. Association for Computational Linguistics.

Xing Shangyu, Zhao Fei, Wu Zhen, An Tuo, Chen Weihao, Li Chunhui, Zhang Jianbing, and Dai Xinyu. 2024. Efuf: Efficient fine-grained unlearning framework for mitigating hallucinations in multimodal large language models. page 1167–1181.

Jiyao Wang, Haolong Hu, Zuyuan Wang, Song Yan, Youyu Sheng, and Dengbo He. 2024a. Evaluating large language models on academic literature understanding and review: An empirical study among early-stage scholars. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, page 1–18, Honolulu HI USA. ACM.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. (arXiv:2406.10252). ArXiv:2406.10252.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. (arXiv:2306.05685). ArXiv:2306.05685 [cs].

A Prompts for Tasks

Prompt	Content
Prompt 1	Imagine you are an experienced academic researcher with access to a vast library of scientific literature. I would like you to find the 10 studies that are most relevant to the research topic provided in the "Title" and the "Keywords" below. Please cite the studies according to the following JSON format. There is no need to provide any explanation before or after the JSON output. Ensure that the "authors" field lists the names of all authors and not exceeding 10 authors, and that there are no duplicate author names nor abbreviations such as "et al.". { "References": [{ "title": "", "authors": "", "journal": "", "year": "", "volumes": "", "first page": "", "last page": "", }] } Title: title Keywords: keywords
Prompt 2	Imagine you are an experienced academic researcher with access to a vast library of scientific literature. I would like you to write an abstract according to the research topic provided in the "Title" and the "Keywords" below. Please write the abstract for about xx words, according to the JSON format as follows. There is no need to provide any explanation before or after the JSON output. {"Abstract": ""} Title: title Keywords: keywords
Prompt 3	Imagine you are an experienced academic researcher with access to a vast library of scientific literature. I would like you to write a literature review according to the research topic provided in the "Title", “Abstract” and "Keywords" below. The literature review should be about 1000 words long. I would like you to back up claims by citing previous studies (with a total of 10 citations in the literature review). The output should be in JSON format as follows: { "Literature Review": "xxx", "References": [{ "title": "", "authors": "", "journal": "", "year": "", "volumes": "", "first page": "", "last page": "", }] } The "Literature Review" field should be about 1000 words. The "References" field is a list of 10 references, and ensures that the "authors" field lists the names of all authors and not exceeding 10 authors, and that there are no duplicate author names nor abbreviations such as "et al.". Title: title Keywords: keywords Abstract: abstract

Table 5: Prompts for tasks.

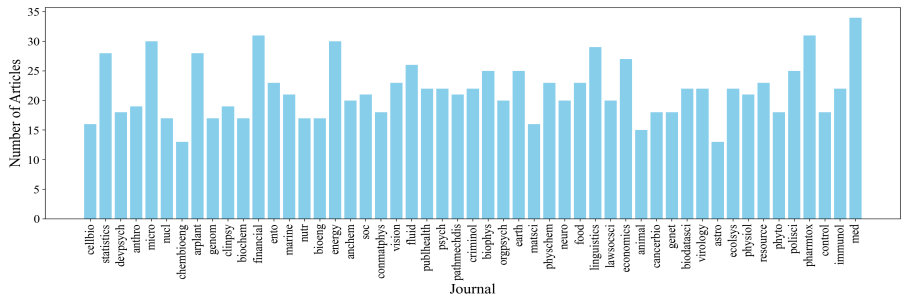


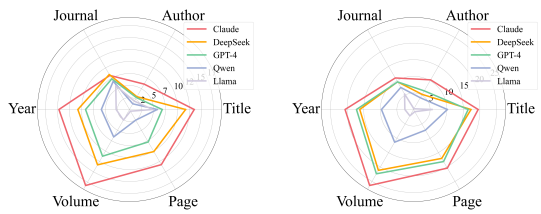
Figure 5: Statistics of dataset.

B Data Distribution

Statistics of the dataset are shown in Figure 5.

C Comparison of LLM-Cited and Human-cited References from Different Dimensions

We provide a more detailed comparison of the LLM-cited and human-cited references across various dimensions. As shown in Figure 6, for Reference Generation, we observe that the overlap rate is higher in the "Title" and other numerical dimensions, while the overlap rates for the "Journal" and "Author" dimensions are relatively lower. For Review Composition, Claude-3.5-Sonnet and GPT-4o exhibit a higher overlap rate on the "Author" dimension compared to Reference Generation. This trend is consistent with the findings in Figure 6, as the citation of author names in the literature for Review Composition leads to the generation of more accurate author information.



(a) Task1 (b) Task3

Figure 6: Radar chart of the accuracy of LLM-generated references with human-written references in the original article.

D Discussion

We select five LLMs for task evaluation and find that Claude-3.5-Sonnet outperforms DeepSeek-V3, GPT-4o, Qwen-2.5-72B, and Llama-3.2-3B across all three tasks, particularly excelling in the task of generating accurate references. This advantage is likely influenced by the training data of each model. Additionally, we observed that each model has different strengths across disciplines. Overall, for the reference generation task, nearly all models perform better in Mathematics, while their performance is weaker in Chemistry and Technology. However, when writing abstracts, all models exhibit the lowest factual consistency in Social Science, as indicated by the entailment scores, compared to human-written texts.

When comparing the references generated by the models in Reference Generation and Review Composition, we find that in Review Composition, nearly all models generate more accurate references. This suggests that LLMs cite references during the writing process, which improves the authenticity of the references. Moreover, the inclusion of the first author's name in the generated context also enhances the accuracy of the author dimension.

E Mitigating Potential Data Contamination

To address potential data contamination, we updated the dataset by crawling articles from the Annual Reviews website between January 1, 2025, and August 13, 2025, resulting in a total of 651 articles. Based on these newly collected articles, we evaluate **gpt-5-2025-08-07** (knowledge cutoff: September 2024) on three tasks. The experimental results are presented in Table 6, 7, and 8. The results indicate that even GPT-5 exhibits a substantial hallucination rate when generating references, further confirming that large language models continue to face significant challenges in literature generation tasks.

F Data Processing Strategy

Author name and journal title variations often pose challenges when aligning LLM-generated references with articles from Semantic Scholar. To address this, we adopt the following normalization strategies:

Author names. When comparing author names between LLM-generated references and candidate articles, if an exact match is not found (e.g., "John Smith"), we consider common variants such as "Smith, John", "Smith, J.", or "J. Smith" to account for different citation formats.

Journal titles. For journal names like Journal of Chemical Physics, we incorporate standard abbreviation forms (e.g., J. Chem. Phys.) based on widely used abbreviation conventions. Nonetheless, certain non-standard or ambiguous cases may still be unmatched.

G Significance Testing for Abstract Writing Metrics

We conduct statistical significance tests to rigorously examine whether ROUGE and similarity met-

Models	Reference Generation				
	S_t↑	*P*	*Overlap*	*P(first author)*	*Overlap(first author)*
GPT-5	26.91	20.62	19.88	21.79	19.75

Table 6: The experimental results of the GPT-5 in Reference Generation.

Models	Abstract Writing					
	Similarity↑	Entail(TRUE)↑	Entail(GPT-4o)↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
GPT-5	74.13	79.82	81.65	30.55	4.77	14.92

Table 7: The performance of GPT-5 on Abstract Writing.

Models	Review Composition								
	References					Literature Review			
	S_t↑	*P*	*Overlap*	*P(first author)*	*Overlap(first author)*	*KPR*↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑
GPT-5	40.92	26.38	20.36	29.72	20.93	80.80	19.42	4.88	9.19

Table 8: The experimental results of the GPT-5 in Review Composition.

rics can meaningfully differentiate model performance in Abstract Writing.

We apply the Friedman test followed by Nemenyi post-hoc comparisons on four automatic metrics (ROUGE-1, ROUGE-2, ROUGE-L, and similarity) across five LLMs (GPT-4o, Claude, Qwen, LLaMA, and Deepseek).

The results demonstrate that all four metrics reveal statistically significant differences across models ($p < 0.001$ for all Friedman tests).

- GPT-4o consistently outperforms other models on all metrics, with significant pairwise differences (e.g., $p < 0.01$ vs. Claude and Qwen).
- Claude, despite relatively high average scores, exhibits larger variance and does not differ significantly from Qwen or LLaMA in some comparisons.
- The similarity metric effectively differentiates GPT-4o and Claude from the remaining models ($p <$ 1e-12 vs. Qwen, LLaMA, and Deepseek), confirming its discriminative power.

These results confirm that ROUGE and similarity scores can capture meaningful and statistically significant differences in model performance, beyond what is apparent from mean values alone.

Tables 9–12 present the pairwise significance test results (p-values) for each metric using the Nemenyi test.

These results confirm that the differences in ROUGE and similarity metrics are statistically meaningful, validating the discriminative power of these metrics for evaluating abstract writing performance across LLMs.

H LLM Output Example

Here is one example of the output generated by Claude-3.5-Sonnet for the literature review composition task.

{"Literature Review": "CRISPR-Cas systems have emerged as a powerful tool in the field of biosensing, offering unprecedented capabilities for detecting nucleic acids, proteins, and small molecules. This review explores the technical developments and application prospects of CRISPR-based biosensing strategies, highlighting their potential to revolutionize diagnostics and molecular detection.The core biochemical properties of CRISPR-Cas systems provide the foundation for their application in biosensing. These properties include diverse reaction temperatures, programmability in design, high reaction efficiency, and recognition specificity. Chen et al. (2018) demonstrated the versatility of CRISPR-Cas9 in detecting DNA targets across a wide temperature range, from 20°C to 60°C, enabling its use in various environmental conditions. The programmability of CRISPR systems allows for the easy design of guide RNAs (gRNAs) to target specific sequences, as shown by Gootenberg et al. (2017) in their development of the SHERLOCK (Specific High-sensitivity Enzymatic Reporter unLOCKing) platform for nucleic acid detection.The high reaction efficiency of CRISPR-Cas systems contributes to their sensitivity in biosensing applications. Li et al. (2019) reported a CRISPR-Cas12a-based assay capable of detecting attomolar concentrations of DNA targets, demonstrating the potential for ultra-

	GPT-4o	Claude	Qwen	LLaMA	Deepseek
GPT-4o	1.0000	0.0200	0.9888	0.5536	0.0033
Claude	0.0200	1.0000	0.0039	0.0001	0.9849
Qwen	0.9888	0.0039	1.0000	0.8419	0.0005
LLaMA	0.5536	0.0001	0.8419	1.0000	0.0000
Deepseek	0.0033	0.9849	0.0005	0.0000	1.0000

Table 9: Pairwise significance test (p-values) for ROUGE-1 scores across five models using Nemenyi test.

	GPT-4o	Claude	Qwen	LLaMA	Deepseek
GPT-4o	1.0000	0.0000	0.0246	0.0000	0.0000
Claude	0.0000	1.0000	0.2415	0.9998	0.9963
Qwen	0.0246	0.2415	1.0000	0.3225	0.4411
LLaMA	0.0000	0.9998	0.3225	1.0000	0.9996
Deepseek	0.0000	0.9963	0.4411	0.9996	1.0000

Table 10: Pairwise significance test (p-values) for ROUGE-2 scores across five models using Nemenyi test.

	GPT-4o	Claude	Qwen	LLaMA	Deepseek
GPT-4o	1.0000	0.1650	0.0089	4.22e-15	2.29e-10
Claude	0.1650	1.0000	0.8322	3.80e-08	8.54e-05
Qwen	0.0089	0.8322	1.0000	1.26e-05	0.0061
LLaMA	4.22e-15	3.80e-08	1.26e-05	1.0000	0.5980
Deepseek	2.29e-10	8.54e-05	0.0061	0.5980	1.0000

Table 11: Pairwise significance test (p-values) for ROUGE-L scores across five models using Nemenyi test.

	GPT-4o	Claude	Qwen	LLaMA	Deepseek
GPT-4o	1.0	0.9156	4.88e-14	0.0	0.9783
Claude	0.9156	1.0	1.11e-16	0.0	0.6107
Qwen	4.88e-14	1.11e-16	1.0	3.87e-11	4.12e-12
LLaMA	0.0	0.0	3.87e-11	1.0	0.0
Deepseek	0.9783	0.6107	4.12e-12	0.0	1.0

Table 12: Pairwise significance test (p-values) for Similarity scores across five models using Nemenyi test.

sensitive detection. The recognition specificity of CRISPR systems allows for precise discrimination between closely related sequences, as evidenced by the work of Myhrvold et al. (2018), who developed a CRISPR-based diagnostic platform capable of distinguishing between Zika and Dengue virus strains with single-base resolution.Recent efforts have focused on improving these core properties to enhance the performance of CRISPR-based biosensors. For example, Harrington et al. (2018) engineered Cas12a variants with improved activity and specificity, expanding the range of targetable sequences and reducing off-target effects. Similarly, Gootenberg et al. (2018) developed an enhanced version of the SHERLOCK system, incorporating lateral flow readouts and multiplexing capabilities, further improving its utility in point-of-care diagnostics.One of the key areas of technical development in CRISPR-based biosensing is improving sensitivity and quantification capabilities. Wang et al. (2020) introduced a CRISPR-Cas12a-based digital PCR method that achieved attomolar sensitivity and precise quantification of target nucleic acids. This approach combines the specificity of CRISPR recognition with the quantitative power of digital PCR, opening new avenues for highly sensitive and accurate molecular diagnostics.Multiplex detection is another important aspect of CRISPR biosensing development. Ackerman et al. (2020) demonstrated a CRISPR-Cas13-based multiplexed

assay capable of simultaneously detecting and distinguishing multiple SARS-CoV-2 variants, showcasing the potential of CRISPR technology in rapid and accurate pathogen identification. The ability to detect multiple targets in a single reaction not only saves time and resources but also provides more comprehensive diagnostic information.Efforts to develop convenient one-pot assays have led to simplified and more user-friendly CRISPR-based detection methods. Joung et al. (2020) reported a one-pot, CRISPR-Cas12-based assay for SARS-CoV-2 detection that could be performed at a single temperature, eliminating the need for complex thermal cycling equipment. This advancement brings CRISPR-based diagnostics closer to point-of-care applications, where simplicity and ease of use are crucial.The creation of advanced sensors has further expanded the capabilities of CRISPR-based biosensing. Bruch et al. (2019) developed a CRISPR-Cas13a-based electrochemical sensor for nucleic acid detection, demonstrating the potential for integrating CRISPR technology with electronic readout systems. This approach offers the possibility of developing portable, electronic CRISPR-based diagnostic devices for rapid and sensitive detection in resource-limited settings.While CRISPR-based biosensing has shown great promise, there are still obstacles to its widespread commercial application. These include challenges in sample preparation, the need for standardization across different CRISPR systems and assay formats, and regulatory considerations for diagnostic use. Addressing these issues will be crucial for the successful translation of CRISPR biosensing technology from the laboratory to real-world applications.Looking ahead, the future of CRISPR-based biosensing appears bright, with numerous opportunities for further development and application. Potential directions include the integration of CRISPR biosensors with microfluidic devices for automated sample processing and analysis, the development of CRISPR-based wearable sensors for continuous health monitoring, and the application of machine learning algorithms to improve the design and performance of CRISPR biosensing systems.In conclusion, CRISPR-based biosensing strategies represent a rapidly evolving field with immense potential to transform molecular detection and diagnostics. As researchers continue to refine and expand upon the core capabilities of CRISPR systems, we can expect to see increasingly sensitive, specific, and versatile biosensing applications emerge, with far-reaching implications for healthcare, environmental monitoring, and beyond.", "References": [{"title": "CRISPR–Cas12a target binding unleashes indiscriminate single-stranded DNase activity", "authors": "Chen, Janice S., Ma, Enbo, Harrington, Lucas B., Da Costa, Maria, Tian, Xiang, Palefsky, Joel M., Doudna, Jennifer A.", "journal": "Science", "year": "2018", "volumes": "360", "first page": "436", "last page": "439"},{"title": "Nucleic acid detection with CRISPR-Cas13a/C2c2", "authors": "Gootenberg, Jonathan S., Abudayyeh, Omar O., Lee, Jeong Wook, Essletzbichler, Patrick, Dy, Aaron J., Joung, Julia, Verdine, Vanessa, Donghia, Nina, Daringer, Nichole M., Freije, Catherine A.", "journal": "Science", "year": "2017", "volumes": "356", "first page": "438", "last page": "442", "DOI": "10.1126/science.aam9321"}, {"title": "CRISPR-Cas12a-assisted nucleic acid detection", "authors": "Li, Suwei, Cheng, Qingmei, Wang, Jianming, Li, Xiaoyu, Zhang, Zhiwei, Gao, Shan, Cao, Rong, Zhao, Guoping, Wang, Jin", "journal": "Cell Discovery", "year": "2019", "volumes": "5", "first page": "1", "last page": "4"}, {"title": "Field-deployable viral diagnostics using CRISPR-Cas13", "authors": "Myhrvold, Cameron, Freije, Catherine A., Gootenberg, Jonathan S., Abudayyeh, Omar O., Metsky, Hayden C., Durbin, Ann F., Kellner, Max J., Tan, Amanda L., Paul, Lauren M., Parham, Leda A.", "journal": "Science", "year": "2018", "volumes": "360", "first page": "444", "last page": "448"}, {"title": "Enhanced proofreading governs CRISPR–Cas9 targeting accuracy", "authors": "Harrington, Lucas B., Paez-Espino, David, Staahl, Brett T., Chen, Janice S., Ma, Enbo, Kyrpides, Nikos C., Doudna, Jennifer A.", "journal": "Nature", "year": "2018", "volumes": "563", "first page": "621", "last page": "625"}, {"title": "Multiplexed and portable nucleic acid detection platform with Cas13, Cas12a, and Csm6", "authors": "Gootenberg, Jonathan S., Abudayyeh, Omar O., Kellner, Max J., Joung, Julia, Collins, James J., Zhang, Feng", "journal": "Science", "year": "2018", "volumes": "360", "first page": "439", "last page": "444"}, {"title": "Ultrasensitive and visual detection of SARS-CoV-2 using all-in-one dual CRISPR-Cas12a assay", "authors": "Wang, Xiaoxia, Zhong, Minjie, Liu, Yue, Ma, Pengfei, Dang, Lei, Meng, Qing, Wan, Wanying, Ma, Xiaowei, Liu, Jing, Yang, Guohua", "journal": "Nature Communications", "year": "2020", "volumes": "11", "first page":

"4711", "last page": "4711"}, {"title": "Detection of SARS-CoV-2 with SHERLOCK One-Pot Testing", "authors": "Joung, Julia, Ladha, Alim, Saito, Makoto, Kim, Nam-Gyun, Woolley, Ann E., Segel, Michael, Barretto, Robert P. J., Ranu, Antonija, Macrae, Rhiannon K., Faure, Guilhem", "journal": "New England Journal of Medicine", "year": "2020", "volumes": "383", "first page": "1492", "last page": "1494"}, {"title": "CRISPR-Cas13-based electrochemical biosensing of viral RNA: Application to detection of SARS-CoV-2", "authors": "Bruch, Richard, Baaske, Johannes, Chatelle, Claire, Meirich, Maren, Madlener, Sibylle, Weber, Wilfried, Dincer, Can, Urban, Gerald A.", "journal": "Angewandte Chemie International Edition", "year": "2019", "volumes": "58", "first page": "17571", "last page": "17575"}, {"title": "Scalable and robust SARS-CoV-2 testing in an academic center", "authors": "Ackerman, Cheri M., Myhrvold, Cameron, Thakku, Shiv G., Freije, Catherine A., Metsky, Hayden C., Yang, David K., Ye, Simon H., Boehm, Chloe K., Kosoko-Thoroddsen, Tinna-Solveig F., Kehe, Jared", "journal": "Nature Biotechnology", "year": "2020", "volumes": "38", "first page": "927", "last page": "931"}]}

I ANOVA Test Results

We report the p-values from one-way ANOVA tests across disciplines for each model and task:

- Reference Generation task: Claude-3.5-Sonnet (p<.0001), DeepSeek-V3(p<.0001), GPT-4o(p<.0001), Qwen-2.5-72B (p<.0001), Llama-3.2-3B (p=0.065).
- Abstract Writing task (NLI scores): Claude-3.5-Sonnet(p<.0001), DeepSeek-V3(p<0.05), GPT-4o(p<.01), Qwen2.5-72B (p<.001), Llama-3.2-3B (p<.0001).
- Review Composition (Reference accuracy): Claude-3.5-Sonnet(p<.0001), DeepSeek-V3(p<.0001), GPT-4o(p<.0001), Qwen-2.5-72B(p<.0001), Llama-3.2-3B(p<.001).
- Review Composition (KPR metric): Claude-3.5-Sonnet (p=0.46), DeepSeek-V3 (p=0.23), GPT-4o (p=0.18), Qwen-2.5-72B (p=0.10), and Llama-3.2-3B (p<0.001).