SurveyGen: Quality-Aware Scientific Survey Generation with Large Language Models

Tong Bao[1,2*], **Mir Tafseer Nayeem**[2], **Davood Rafiei**[2 †], **Chengzhi Zhang**[1 †]
[1]Nanjing University of Science and Technology
[2]University of Alberta
{tbao,zhangcz}@njust.edu.cn, {mnayeem,drafiei}@ualberta.ca

Abstract

Automatic survey generation has emerged as a key task in scientific document processing. While large language models (LLMs) have shown promise in generating survey texts, the lack of standardized evaluation datasets critically hampers rigorous assessment of their performance against human-written surveys. In this work, we present **SurveyGen**, a large-scale dataset comprising over 4,200 human-written surveys across diverse scientific domains, along with 242,143 cited references and extensive quality-related metadata for both the surveys and the cited papers. Leveraging this resource, we build QUAL-SG, a novel quality-aware framework for survey generation that enhances the standard Retrieval-Augmented Generation (RAG) pipeline by incorporating quality-aware indicators into literature retrieval to assess and select higher-quality source papers. Using this dataset and framework, we systematically evaluate state-of-the-art LLMs under varying levels of human involvement—from fully automatic generation to human-guided writing. Experimental results and human evaluations show that while semi-automatic pipelines can achieve partially competitive outcomes, fully automatic survey generation still suffers from low citation quality and limited critical analysis [1].

1 Introduction

Survey articles play a crucial role in summarizing previous research on a specific topic, providing a comprehensive understanding of the field, and supporting further advancements (Torraco, 2005). However, writing a survey is a complex task as it typically requires summarizing hundreds of relevant studies. The rapid expansion of academic publications further adds to the difficulty, making it increasingly challenging for researchers to keep up with the latest findings. Given these challenges, the development of automatic survey generation systems has become a key focus in the field of scientific document processing (Wang et al., 2024).

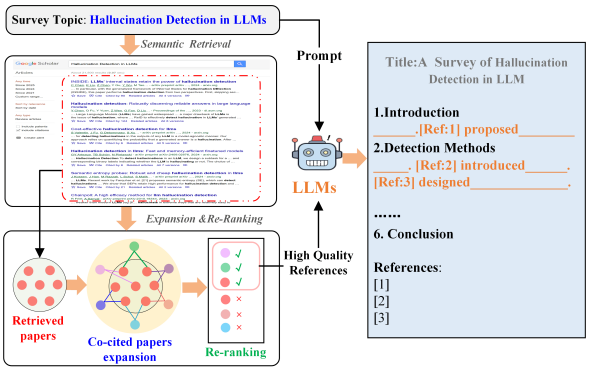


Figure 1: Overview of the proposed QUAL-SG: a quality-aware framework that leverages semantic retrieval and citation expansion to select high-quality literature and support more reliable survey generation.

Leveraging the strong text generation capabilities of large language models (LLMs) (Brown et al., 2020), recent studies on automatic survey generation have adopted Retrieval-Augmented Generation (RAG) techniques (Izacard et al., 2022; Borgeaud et al., 2022; Gao et al., 2024; Agarwal et al., 2025a) to augment them with external knowledge sources, yielding promising results (Wang et al., 2024; Tang et al., 2025; Wu et al., 2025; Liang et al., 2025). However, these approaches fall short in two key aspects: **(1)** the retrieval of high-quality literature, and **(2)** the rigorous evaluation against the human-authored gold standard.

In the retrieval stage, most prior works rely on semantic and syntactic similarity between user-provided survey topics and publication abstracts to identify relevant studies. These methods typically do not assess the quality, impact, or influence of the retrieved literature. Yet, a well-crafted survey is expected to not only summarize existing research but also highlight seminal works and major advance-

*This work was partly done at the University of Alberta.
†Corresponding authors
[1]Code and data are available at https://github.com/tongbao96/SurveyGen

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 2713–2737
November 4-9, 2025

ments in the field (Snyder, 2019; Paul and Criado, 2020; Kanellos et al., 2021). As a result, retrieving articles based purely on textual relevance risks including low-impact or marginal studies, which in turn diminishes the quality and credibility of the generated survey.

In the evaluation stage, although recent works have employed both automatic and human evaluations to assess LLM-generated surveys (Wu et al., 2025; Lai et al., 2025; Agarwal et al., 2025b; Wang et al., 2024; Liang et al., 2025), the lack of large-scale benchmarks has hindered systematic comparisons with human-written surveys, which remain the gold standard. In particular, critical evaluation dimensions such as citation quality, structural consistency, and domain-specific variation remain underexplored. Without comprehensive benchmarks, it is difficult to rigorously evaluate whether LLM-generated surveys meet the quality, reliability, and scholarly standards expected in academic writing.

To address the above limitations, we first introduce **SurveyGen**, a large-scale dataset comprising over 4,200 human-written surveys from the Semantic Scholar Open Research Corpus (S2ORC; Lo et al., 2020), along with 242,143 cited references within these surveys and extensive metadata for all referenced papers for evaluation purposes. Building on this resource, we propose QUAL-SG, a novel quality-aware literature retrieval framework designed to enhance the reliability and relevance of retrieved articles for survey generation. As shown in Figure 1, QUAL-SG first expands the candidate reference pool via citation graph analysis, then reranks articles based on quality indicators, ensuring both citation reliability and broad literature coverage for survey generation. We design three targeted tasks, each equipped with domain-appropriate evaluation metrics, to provide a comprehensive analysis of LLMs' effectiveness across different stages of the survey generation pipeline.

Our contributions can be summarized as follows:

- We introduce **SurveyGen**, a large-scale dataset comprising over 4,200 human-written surveys with section-level structures, including cited references and rich metadata capturing citation performance, author influence, and venue reputation. **SurveyGen** supports comprehensive evaluation across content quality, citation quality, and structural consistency in scientific survey generation tasks.
- We propose **QUAL-SG**, a novel quality-aware framework that extends Naive-RAG by incorporating literature quality assessment into the survey generation pipeline. Our results show that QUAL-SG significantly improves citation reliability and enhances the overall content quality and structure consistency of the generated surveys.
- We benchmark several state-of-the-art LLMs under varying levels of involvement in the survey generation process and conduct extensive evaluations—both automatic and human—to analyze model performance, identify key limitations, and offer actionable insights for future research on LLM-assisted academic writing.

2 Approach

In this section, we first introduce the design of the survey generation tasks (§2.1), then present our **SurveyGen** dataset (§2.2, §2.3) and the proposed QUAL-SG framework (§2.4).

2.1 Task Design

Given that humans may engage LLMs at different stages during survey generation depending on their specific goals (e.g., literature retrieval, outline generation, or content drafting), the level of involvement can vary considerably. We define *three* representative tasks to systematically evaluate LLMs' generation capabilities across these different levels: **(1) Fully LLM-based**, **(2) RAG-based**, and **(3) Human-guided Survey Generation**.

The *distinct focus* of the three tasks is as follows: Task 1 evaluates LLM's capability to generate a complete survey without access to external sources; Task 2 evaluates its performance under the standard RAG setting, where relevant literature is first retrieved from an external database and then used to support survey generation; and Task 3 evaluates the generated survey when LLMs are provided with human-selected references and a human-written predefined outline, simulating a fully guided writing setting. The definitions of these three tasks are detailed below:

Task 1: Fully LLM-based Survey Generation: Given only a survey topic t_i, the LLMs are prompted to generate the entire survey, including a structured outline, corresponding content, and a relevant list of references. No external documents or human-crafted materials are provided.

Task 2: RAG-based Survey Generation: This task follows the standard RAG pipeline, where a retriever identifies relevant literature from an external database, and a generator writes the survey's outline and content. Given a survey topic t_i, we

Dataset	Domains	#Docs	#Input Len	#Target Len	#Input Docs	Structural Outline	Quality Indicators	Multi-level Citation	For Survey Generation
PubMed (2018)	Bio	133K	3016	203	1	✓	✗	✗	✗
ArXiv (2018)	Mixed	215K	4938	220	1	✓	✗	✗	✗
SciSummNet (2019)	CL	1K	4417	151	61.00	✗	✗	✗	✗
Multi-XScience (2020)	CS	40.5K	778	116	4.42	✗	✗	✗	✗
BigSurvey (2022)	Mixed	4.4K	11893	1051	76.30	✗	✗	✗	✗
SciReviewGen (2023)	CS	10.2K	12503	8082	68.00	✓	✗	✗	✓
SurveyGen(ours)	Mixed	4.2K	11423	5115	57.58	✓	✓	✓	✓

Table 1: Comparison with other scientific document summarization datasets. **SurveyGen(ours)** and SciReviewGen (Kasanishi et al., 2023) are the only two suitable for survey generation. Compared to SciReviewGen, our dataset provides additional quality indicators and second-level references, supporting more accurate document selection and citation network analysis. In addition, **SurveyGen** includes surveys from multiple domains, such as Computer Science, Medicine, Biology, and Psychology, whereas SciReviewGen is limited to only Computer Science.

retrieve the top-n most relevant papers to form the initial candidate set $D = \{a_1, a_2, \ldots, a_n\}$. Then, based on D, LLMs are prompted to first create a survey outline to avoid brief outputs from one-shot generation, and then expand each section in parallel to construct the final survey.

Task 3: Human-guided Survey Generation: In this task, we remove the retrieval stage of RAG; instead, the survey is generated based on a gold-standard survey outline and selected references, both extracted from human-written surveys. This setup simulates a realistic scenario in which authors, having already selected relevant literature and a predefined outline, can then focus on guiding LLMs to write the survey.

To provide publicly accessible input, the abstracts of the cited papers are used as the primary information in our study. The input and output for the three tasks are detailed in the Appendix B.

2.2 SurveyGen: Dataset Construction

We developed **SurveyGen** based on S2ORC (Lo et al., 2020), a large dataset containing 81.1 million English academic papers. In the preliminary search, we extracted articles by filtering titles that either contain *"a survey"*, *"survey of"*, *"a review"*, *"literature review"*, *"overview"* with full-text data available and publication years after 2010[2]. This resulted in a total of 8,676 candidate papers.

Since title-based filtering may still include non-survey articles, we applied an additional filtering step using abstracts to further refine the candidate set. Specifically, inspired by previous work that LLMs are effective as NLI models for evaluating factual consistency (Gubelmann et al., 2023; Chiang et al., 2024), we prompted three LLMs to classify whether a candidate paper is a survey-type article based on its title and abstracts, following three criteria: **(1)** Explicit declaration of survey intent (e.g., *"conducts a survey"* or *"provides a survey"*). **(2)** Focus on survey papers, rather than proposing novel methodologies or experimental results. **(3)** Discussion of field-specific trends, challenges, or future directions. Papers without abstracts were excluded at this step. Based on these criteria, 6,851 out of 8,676 papers were identified as survey articles by a majority vote of the LLMs.

We then retrieved the full-text data of these surveys using their paper IDs from the S2ORC bulks[3]. Here, the full-text data includes the **full body** of the survey with **section divisions**, as well as the **citation locations** of each reference within the survey. This allows us to obtain the structural outline of each survey paper and map references to their corresponding sections, which serve as the key input for Task 3. At this point, we removed papers that had fewer than 30 references or fewer than three top-level sections, as they are too short to serve as meaningful surveys. Finally, we obtained 4,205 papers suitable for survey generation and constructed the **SurveyGen** dataset, which includes 115,376 sections, 242,143 references directly cited within the surveys, and 5,062,596 references cited by these cited papers.

The data format is outlined in Appendix H. Table 1 compares **SurveyGen** with other datasets for scientific document summarization.

2.3 Quality-Related Indicators Supplement

To facilitate citation-based evaluation, we first supplemented all survey papers and their directly cited references with basic metadata (e.g., abstract, DOI,

[2]Data collected from S2ORC up to December 2024.

[3]https://api.semanticscholar.org/api-docs/

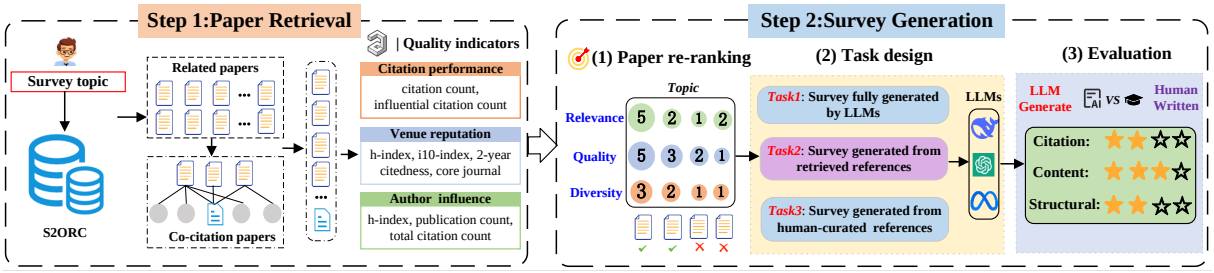


Figure 2: Overview of the QUAL-SG framework, which comprises two main stages: **paper retrieval** and **survey generation**. The retrieval stage includes three steps: **(1)** retrieving topic-relevant papers, **(2)** expanding with frequently co-cited papers, and **(3)** enriching them with quality-related metadata. Based on the retrieved set, the generation stage first re-ranks the papers from three key aspects, then prompts LLMs to perform tasks under different input conditions. Finally, we evaluate the generated surveys against human-written ones across multiple dimensions.

publication venue, date, and research fields) from S2ORC, linked via corpus IDs. However, S2ORC does not provide sufficient metadata to measure the impact of academic papers.

To address this limitation, we used DOIs of the involved papers to retrieve their corresponding metadata from the OpenAlex[4] (Priem et al., 2022) database and enriched them with additional quality-related signals. Specifically, we incorporated three well-known bibliometric indicators to measure the quality of scientific publications: **(1) citation performance**: citation count and influential citation count[5]; **(2) author influence**: h-index, publication count, and total citation count; and **(3) venue reputation**: h-index, i10-index, and CORE status of the publication venue (journal or conference) (Hicks et al., 2015; Donthu et al., 2021). As a result, each survey is now paired with its full text, complete with section divisions, and linked to its directly cited references. These references are further enriched with comprehensive multi-level quality indicators retrieved from OpenAlex, providing a robust foundation for evaluating generated surveys across multiple dimensions, such as citation accuracy, content quality, and structural consistency.

Second-Level References Supplement: In some cases, influential works relevant to a survey may not be semantically aligned with its primary topic. For example, when retrieving literature on *"Deep Learning"*, seminal works such as the *"Backpropagation algorithm"* (Rumelhart et al., 1986) may obtain low semantic similarity scores as their titles and abstracts do not explicitly mention the topic. However, such papers are frequently cited by other retrieved references and are widely recognized as foundational to the field. To address this issue and support advanced citation network analysis, we further enriched the metadata for 5.06 million references cited by the papers referenced in all surveys. For each of these references, we extracted essential bibliographic details, including the title, abstract, DOI, and citation count.

2.4 QUAL-SG: QUALity-aware Literature Retrieval for Survey Generation

We propose QUAL-SG, a quality-aware extension of the naive RAG framework designed to improve the quality of retrieved literature for survey generation (Task 2). The overall framework is illustrated in Figure 2. As described in Section 2.1, in the Naive-RAG framework, the survey topic is used as a query to retrieve relevant papers from external databases. Formally, let q denote the topic derived from the human-written survey. Each candidate paper d_i in the external database is represented by its abstract embedding. We define the semantic similarity score as:

$$\text{Sim}(q, d_i) = \cos(\mathbf{v}_q, \mathbf{v}_{d_i}) = \frac{\mathbf{v}_q \cdot \mathbf{v}_{d_i}}{\|\mathbf{v}_q\| \, \|\mathbf{v}_{d_i}\|} \quad (1)$$

where $\mathbf{v}_q$ and $\mathbf{v}_{d_i}$ are the embedding vectors of the query and the abstract of document d_i, respectively.

The top-n with the highest embedding similarity scores are selected to form the initial candidate set $D = \{d_1, d_2, \ldots, d_n\}$, where n is set to exceed the number of references in the corresponding human-written survey to ensure sufficient candidate coverage.

Although the documents in D are topically relevant, certain papers may not exhibit strong semantic similarity to the query but still have a substantial impact within the research area (e.g., as seen in cases like *"Backpropagation algorithm"* to

[4] https://openalex.org/

[5] Citations identified by Semantic Scholar as impactful in context, rather than mere mentions in the bibliography.

"Deep Learning"). Therefore, we expand D via a co-citation expansion: any paper cited by at least two papers in D is added to the set. Let D_{ex} denote this expanded set.

Beyond topical relevance, crafting a high-quality survey also requires careful selection of cited papers (Paul and Criado, 2020). High-impact publications such as those published in reputable venues or frequently cited by other works generally contribute more significantly to the field (Kanellos et al., 2021). Therefore, for each document in D_{ex}, we further collect a set of quality-related indicators, including citation performance, author influence, and venue reputation, as described in Section 2.2.

Then, we evaluate the quality of each candidate paper from three perspectives: **topical relevance**, **academic impact**, and **content diversity**. Specifically, for topical relevance, we employ LLMs-as-judge to assess the alignment between each candidate paper $a_i \in D_{ex}$ and the survey topic t. The relevance score is denoted as:

$$S_t = \text{LLM}_{\text{judge}}(\boldsymbol{a}_i, t) \tag{2}$$

For academic impact, we compute a weighted score that integrates three components: citation performance $C(a_i)$, author influence $A(a_i)$, and venue reputation $V(a_i)$, since these factors are commonly associated with paper quality (Hicks et al., 2015; Donthu et al., 2021). Each component is computed using a group-based scoring strategy, where raw indicator values are categorized into four ordinal levels based on percentile ranks. The overall academic impact score is defined as:

$$S_a = \alpha \cdot C(a_i) + \beta \cdot A(a_i) + \gamma \cdot V(a_i) \tag{3}$$

where α, β, and γ are control variables that can be adjusted based on specific application needs.

For content diversity, we select papers that are topically relevant yet semantically distinct from others in the candidate pool to broaden the survey's perspectives. To achieve this, we use the abstract of each paper as input and define the diversity of a candidate paper a_i to a set of papers $S \subseteq D_{ex}$ as the average semantic distance:

$$S_d(a_i, S) = \frac{1}{|S|} \sum_{a_j \in S} \text{Dist}(a_i, a_j) \tag{4}$$

Finally, all candidate papers in D_{ex} are re-ranked based on their average ranks across S_t, S_a, and S_d. The top-$\mathcal{K}$ papers are selected to form the final set for survey generation, where $\mathcal{K}$ matches the number of references in the corresponding human-written survey to ensure a fair comparison.

3 Experiments

3.1 Baselines

We selected three baselines for comparison.

- **Fully-LLMGen** (Tang et al., 2025): Surveys are generated by LLMs based only on the given topic, without external inputs.
- **Naive-RAG** (Wu et al., 2025): Candidate papers are retrieved from an external literature database based on semantic similarity between the abstract and the survey topic. We use the same input fields as QUAL-SG to prompt LLMs for survey generation.
- **Human-written**: The human-written surveys are selected from our **SurveyGen** dataset.

For generation stages, we employed six LLMs as agents, including three **Open-source LLMs:** *GLM-4-Flash* (GLM et al., 2024), *LLaMA-3.1-70B* (Meta, 2024), and *DeepSeek-V3* (DeepSeek-AI et al., 2025), and three **Closed-source LLMs:** *GPT-4.1-2025-04-14* (OpenAI, 2025), *Gemini-2.0-Flash* (Team et al., 2024), and *Claude-3.7-Sonnet-20250219* (Anthropic, 2025). Implementation details are provided in Appendix C.

To be cost-effective, our experiments are conducted on 120 highly cited surveys from **SurveyGen**, with 30 selected from each of four domains: Biology, Medicine, Psychology, and Computer Science. For Task 1 and Task 3, we directly report the performance of different LLMs. For Task 2, we provide a comparative analysis between our QUAL-SG and the baseline methods. A subset of survey examples is provided in Appendix G.

3.2 Evaluation Metrics

We consider human-written surveys as the ground truth for both automatic and human evaluations.

Automatic evaluation: The automatic evaluation includes three parts: ***citation quality***, ***content quality***, and ***structural consistency***. The formulas for the metrics and calculation details in this section are provided in the Appendix D.

(1) Citation quality evaluation. First, we assess how closely the references retrieved by RAG or generated by LLMs match those selected by humans. To address variations in title phrasing and formatting of the same article, we compute the textual similarity between each generated or retrieved reference and the human-selected ones. A reference is considered matched if the similarity exceeds a predefined threshold. We use precision,

Model	Citation Quality				Content Quality			Structural Consistency	
	Acc. ↑	P ↑	R ↑	F1 ↑	Sim. ↑	R-L ↑	KPR ↑	Rel_{LLM}	Overlap (%)
Open-source LLMs									
GLM-4-Flash	9.27	9.03	3.26	4.79	81.27	15.04	41.71	2.44	10.62
LLaMA-3.1-70B	15.43	11.48	2.74	4.42	**82.43**	**15.36**	44.36	2.62	13.48
DeepSeek-V3	33.63	10.85	4.09	5.94	82.05	14.18	43.53	2.57	11.03
Closed-source LLMs									
GPT-4.1	21.07	**12.31**	3.72	5.71	79.51	13.48	39.21	2.39	10.95
Gemini-2.0-Flash	22.20	8.97	3.59	5.13	80.20	14.65	42.67	2.50	12.39
Claude-3.7-Sonnet	**35.84**	11.79	**5.78**	**7.76**	81.32	13.77	**46.59**	**2.65**	**14.89**

Table 2: Performance comparison of different LLMs on Task 1. "Acc" indicates whether the generated references are factually accurate and correspond to real papers. "Sim", "R-L", and "KPR" represent "Semantic similarity", "ROUGE-L", and "Key Point Recall", respectively. "Rel_{LLM}" represents structural consistency in LLM evaluations. The best results are marked **bold** and the second-best are underlined.

recall, and F1 score to evaluate citation overlap. Additionally, for Task 1, we compute citation accuracy to check whether the generated references are fabricated or hallucinated.

(2) Content quality evaluation. We first compute the semantic similarity between the LLM-generated and human-written surveys, and then report ROUGE[6] score to quantify their textual overlap. Apart from semantic similarity evaluation, we employ Key Point Recall (KPR) (Qi et al., 2024; Tang et al., 2025) to evaluate how effectively LLM-generated surveys capture the key points conveyed in human-written ones.

(3) Structural consistency evaluation. In scientific writing, a well-structured survey typically features clear section divisions and coherent thematic development (Wee and Banister, 2016; Paul and Criado, 2020). To evaluate structural consistency, we adopt two metrics: Overlap score and $\text{Relevance}_{\text{LLM}}$. Specifically, the Overlap score is defined as the number of sections between the generated and human-written surveys with semantic similarity exceeding a predefined threshold. Then, we prompt the LLM-as-judge to evaluate the structural alignment between LLM-generated and human-written surveys using a 5-point scale.

Human evaluation: Inspired by previous works (Kasanishi et al., 2023; Liang et al., 2025), we also conduct human evaluation to compare the LLM-generated and human-written surveys from the following four aspects: ***topic relevance***, ***information coverage***, ***critical analysis***, and ***overall rating***. The evaluation criteria and the detailed annotation process are provided in the Appendix E.

[6] https://pypi.org/project/pyrouge/. All reported Rouge scores have a 95% confidence interval in this paper.

4 Results and Analysis

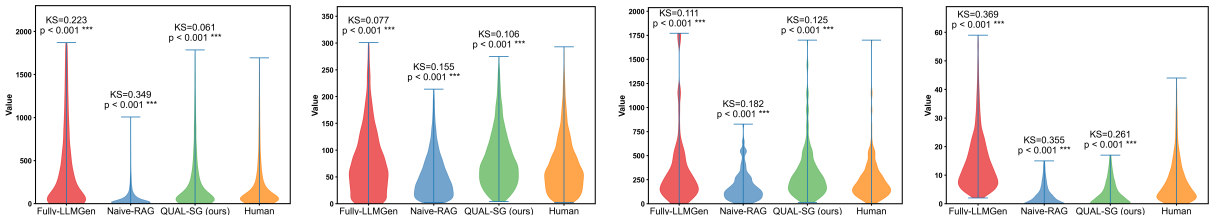
4.1 Main Results

Results for Task 1: We report the evaluation results of different LLMs on Task 1. As shown in Table 2, Claude 3.7-Sonnet achieves the best overall performance across citation quality, KPR, and structural consistency. In content evaluation, LLaMA-3.1-70B achieves the highest similarity to human-written surveys (82.43%) and the highest ROUGE-L (15.36%). However, citation accuracy remains a major limitation: the best-performing model achieves only 35.84%, indicating that **relying solely on LLMs for survey generation is insufficient for ensuring reliable reference generation**. Furthermore, compared to human-written surveys, although the LLM-generated content is semantically similar, it still shows significant gaps in key point coverage (46.59%) and structural overlap (14.89%). Lastly, closed-source and open-source LLMs exhibit distinct strengths: closed-source models consistently surpass open-source models in citation quality and structural consistency, while open-source models deliver comparable results in content generation.

Results for Task 2: Table 3 summarizes the results of different models on Task 2. Compared with the Fully-LLMGen approach, the Naive-RAG method, despite retrieving authentic literature from external databases, yields the lowest citation quality. In contrast, **our proposed QUAL-SG achieves the highest citation quality** (F1 score of 16.73%), outperforming Naive-RAG and Fully-LLMGen by 10.80% and 8.97%, respectively. QUAL-SG also surpasses both baselines in content quality (Similarity +0.73%, ROUGE-L +1.40%, KPR +3.66%)

Model	Citation Quality			Content Quality			Structural Consistency	
	P ↑	R ↑	F1 ↑	Sim. ↑	R-L ↑	KPR ↑	Rel_{LLM}	Overlap (%)
Fully-LLMGen	11.79	5.78	7.76	81.32	13.77	46.59	2.65	14.89
Naive-RAG	5.18	6.94	5.93	82.37	12.90	42.17	2.43	12.22
QUAL-SG (Ours)	**15.87**†	**17.71**†	**16.73**†	**83.10**†	**15.17**†	**50.25**†	**2.81**†	**24.76**†

Table 3: Performance of different models on Task 2. For Fully-LLMGen (Tang et al., 2025), we directly report the results from Task 1. In the Naive-RAG setting (Wu et al., 2025), retrieval is based on the semantic similarity between the survey topic and candidate abstracts. Claude-3.7-Sonnet is used as the backbone for all methods. The best results are marked **bold**. † denotes significant differences to baselines (p-value < 0.001).



(a) Citation Count (b) Author H-Index (c) Venue H-Index (d) Years Gap

Figure 3: Comparison of reference selection distributions across models. "KS" denotes the Kolmogorov–Smirnov statistic against the human baseline (lower values indicate closer alignment), "p" is the associated p-value, and "Years Gap" denotes the difference in publication years between the reference and the survey. For Fully-LLMGen, the survey year is set to 2025. Claude-3.7-Sonnet is used as the backbone LLM for all methods.

and structural consistency (LLM evaluation +0.16 on a 5-point scale, semantic overlap +12.54%).

The results suggest that while the Naive-RAG framework can improve the factual accuracy of generated references, it remains limited in identifying truly human-preferred or high-quality references from the large-scale academic database. In contrast, QUAL-SG mitigates this limitation via a re-ranking module that integrates topical relevance, academic impact, and content diversity, yielding reference selections better aligned with human preferences. This improvement in citation quality, in turn, enhances the overall content quality and structural consistency of the generated surveys.

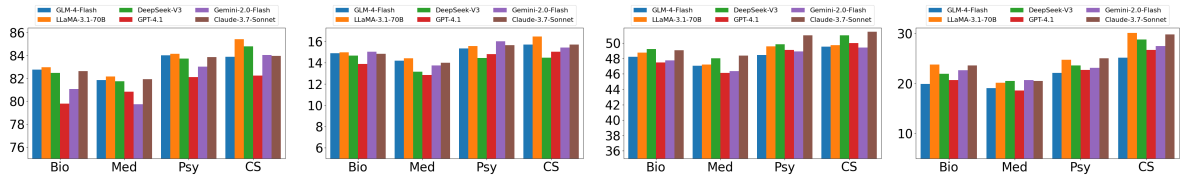
Results for Task 3: For Task 3, since both the candidate references and the outline are directly extracted from human-written surveys, we only report the content evaluation results of different LLMs, as shown in Table 4. We observe that **when LLMs are provided with more accurate references and outlines, their generated content quality improves accordingly** compared to Task 2, which involves no human intervention. Among the models, the open-source LLaMA-3.1-70B still achieves the highest content similarity (84.39%) and ROUGE-L (17.16%), while Claude-3.7-Sonnet obtains the highest KPR (54.67%). Overall, with human intervention, open-source models exhibit a strong capability to compete with advanced closed-source models in the survey generation task.

Model	Sim. ↑	R-L ↑	KPR ↑
Open-source LLMs			
GLM-4-Flash	82.04	<u>16.29</u>	46.88
LLaMA-3.1-70B	**84.39**	**17.16**	<u>52.13</u>
DeepSeek-V3	<u>83.97</u>	15.25	49.50
Closed-source LLMs			
GPT-4.1	82.59	13.82	50.02
Gemini-2.0-Flash	83.74	15.62	51.76
Claude-3.7-Sonnet	84.22	15.43	**54.67**

Table 4: Content quality evaluation results of different LLMs on Task 3. The best results are marked **bold** and the second-best are <u>underlined</u>.

4.2 Further Analysis of Reference Selection

We further analyze the distribution of references yielded by different models in Task 2, as shown in Figure 3. The results indicate that QUAL-SG exhibits the closest alignment to human-written surveys in citation count and temporal distribution of selected references, and achieves competitive performance in author H-index and venue H-index (Figure 3a ~ 3d). Specifically, Fully-LLMGen exhibits a pronounced long-tail distribution in reference selection, with most selected papers concentrated in the less-cited studies. The poor performance of Naive-RAG highlights the limitation of



(a) Similarity (b) Rouge-L (c) KPR (d) Structural Overlap

Figure 4: Performance comparison of different models across disciplines on Task 2. "Bio", "Med", "Psy", and "CS" denote Biology, Medicine, Psychology, and Computer Science, respectively. "KPR" refers to Key Point Recall.

purely semantic retrieval, as many retrieved articles, although semantically relevant, do not meet the quality standards expected for survey writing. Regarding the temporal distribution, human-written surveys tend to favor papers published within the preceding decade, while Fully-LLMGen often overlooks recent studies due to outdated training data.

4.3 Cross-Disciplinary Comparison of LLMs

We extend the analysis from Task 2 to compare the performance of each LLM powering QUAL-SG across academic disciplines. As shown in Figure 4a, the models yield relatively stable content similarity across domains. This observation is further confirmed by one-way ANOVA tests conducted for each model, which reveal no statistically significant differences across disciplines: GLM-4-Flash (p=0.17), DeepSeek-V3 (p=0.42), LLaMA-3.1-70B (p=0.31), GPT-4.1 (p=0.21), Gemini-2.0-Flash (p=0.39), and Claude-3.7-Sonnet (p=0.32).

We then report the ROUGE-L scores for different LLMs across disciplines. As shown in Figure 4b, scores in Computer Science and Psychology are generally higher than those in Medicine and Biology, with LLaMA-3.1-70B consistently outperforming other models. Moreover, all models exhibit statistically significant performance differences across disciplines (p<.001).

Similarly, KPR scores (Figure 4c) follow the same trend, with higher scores in Computer Science and Psychology across all models. Claude-3.7-Sonnet consistently achieves the best KPR score. However, the differences across disciplines are not statistically significant for individual models: GLM-4-Flash (p=0.12), DeepSeek-V3 (p=0.36), LLaMA-3.1-70B (p=0.25), GPT-4.1 (p=0.40), Gemini-2.0-Flash (p=0.27), and Claude-3.7-Sonnet (p=0.33).

For structural consistency (Figure 4d), LLaMA-3.1-70B achieves the best performance in Computer Science, Biology, and Psychology, while Gemini-2.0-Flash leads in Medicine. All models show statistically significant differences in structural consistency across disciplines (p<.001).

4.4 Comparison with Other Ranking Models

We compare our QUAL-SG with UPR (Sachan et al., 2022) and RankGPT (Sun et al., 2023), both designed for ranking candidates in the RAG pipeline. Since the generation stage mainly depends on the selected references as sources, we report their performance only in the retrieval stage, as this more directly reflects the impact of candidate ranking. As shown in Table 5, our method outperforms UPR, which relies on probability-based token-level ranking. While RankGPT incorporates this criterion through its instructions, QUAL-SG employs a more direct strategy through weighted aggregation, demonstrating greater robustness when handling multiple ranking criteria.

Model	P%↑	R%↑	F1%↑
UPR (Sachan et al., 2022)	10.28	10.63	10.45
RankGPT (Sun et al., 2023)	<u>14.55</u>	<u>15.09</u>	<u>14.81</u>
QUAL-SG (ours)	**15.87**	**17.71**	**16.73**

Table 5: Citation quality comparison of different ranking models. For RankGPT, we prompt it to rank papers according to the same three criteria (§2.4) used in our QUAL-SG. The best results are marked **bold** and the second-best are <u>underlined</u>.

4.5 Human Evaluation Results

The human evaluation results are presented in Table 6 in the Appendix E. We can observe that Task 3 is generally rated as more acceptable by human evaluators. This highlights the importance of key preprocessing steps, such as high-quality reference selection and effective outline construction, in guiding LLMs to generate more reliable scientific surveys. However, despite the comparable performance in terms of topic relevance, the generated surveys currently ***fail to provide sufficient information coverage and critical analysis***.

5 Discussion and Future Directions

LLM for Automatic Survey Generation: Are We There Yet? The results in Section 4.1 indicate that neither Fully LLM-based nor RAG-based approaches have achieved human-level performance. As highlighted in (Liang et al., 2025; Tang et al., 2025), hallucinated information, such as fabricated references and factual inaccuracies, remains a critical challenge in LLM-generated surveys. Although RAG-based methods reduce hallucinations by retrieving external sources, the retrieved papers are often only topically relevant and misaligned with human preferences. While LLMs have demonstrated efficiency and the ability to generate content considered useful by human evaluators (Wang et al., 2024), our human evaluation results (§4.5) reveal that, despite strong topical relevance, LLM-generated surveys exhibit limited information coverage and in-depth analysis, both essential for high-quality scientific surveys. Therefore, while LLMs can assist in survey generation, they are still unable to independently craft surveys that meet academic standards at the current stage.

Future Directions for Enhancing Survey Generation As shown in Section 4.1, quality-based ranking of candidate references effectively improves the citation performance of generated surveys. This can be further enhanced through several strategies. For example, citation network analysis can be used to capture global relationships among papers and identify influential studies. Furthermore, analyzing human citation behavior, such as intent, frequency, and location of citation in the textual context, can inform better reference selection mechanisms. Training reference selection models on human-annotated datasets is also a potential option to collect relevant literature for survey generation. In the generation stage, relying solely on abstracts as input significantly limits the information coverage, as it fails to fully capture the paper's broader details. Future work could leverage full-text information to enable more comprehensive contextual understanding, as well as explore human-in-the-loop discourse control, factual consistency verification, and advanced long-document modeling to improve survey generation.

Real-World Applicability and Deployment Our framework is designed with modular components, including embedding-based retrieval, co-citation expansion, and re-ranking, which can be parallelized or extended. For the retrieval stage, we choose S2ORC (Lo et al., 2020) as the external database because its papers are peer-reviewed and have full dataset downloads, which can be stored locally and used for a one-time embedding computation. In practice, it can be replaced with other sources such as arXiv[7] or PubMed[8], depending on user needs. Additionally, numerous well-established embedding models are available on the MTEB leaderboard[9], offering a range of trade-offs between accuracy, model size, and computational efficiency. For the co-citation expansion module, we rely on the OpenAlex (Priem et al., 2022) database for citation analysis. OpenAlex also provides free APIs and allows bulk download of citation data. Similarly, users can replace OpenAlex with other citation databases, such as Scopus (Elsevier, 2025) and SciSciNet (Lin et al., 2023). As for the re-ranking, we assume it is highly adaptable to different downstream needs. Since we will release the **SurveyGen**, users can customize re-ranking strategies according to their specific preferences. In the generation stage, the results (§4.1) show that open-source models (e.g., LLaMA-3.1-70B) can achieve competitive performance compared to closed-source commercial LLMs such as GPT-4.1. This offers users greater flexibility based on their budget, deployment needs, and infrastructure.

6 Conclusion

We introduce **SurveyGen**, a new dataset designed to support scientific survey generation. Building on this resource, we propose QUAL-SG, an enhanced RAG framework that improves upon Naive-RAG by identifying higher-quality references during literature retrieval. Experimental results show that QUAL-SG outperforms semantic similarity-based RAG methods across key aspects, including citation quality, content quality, and structural consistency of the generated surveys. Finally, we conduct a human evaluation to assess the impact of human intervention at different stages of the survey generation process. Our findings show that providing more accurate references and a well-structured outline enables LLM to generate surveys more aligned with human-written ones; however, there remains considerable room for improvement in both citation and content quality to meet human expectations.

[7] https://arxiv.org/

[8] https://pubmed.ncbi.nlm.nih.gov/

[9] https://huggingface.co/spaces/mteb/leaderboard

Limitations

We acknowledge several limitations in our work.

Input Limitation. For copyright reasons, our approach is restricted to using only abstracts and bibliographic metadata of the retrieved papers, without access to full-text content. This limitation may hinder the LLM's ability to capture finer-grained details and structural elements that are often present in full-length papers. Hence, the generated surveys may lack depth and completeness compared to human-written surveys that draw on the entire papers.

Post-generation Refinement. To reduce API costs, we did not perform post-generation refinement, such as language polishing, citation formatting, structural adjustments, and PDF/Latex export. These post-processing steps could further improve the personalization and overall quality of the generated surveys. Also, our work focuses on generating textual survey content and does not include visual elements such as figures, tables, or diagrams, which are often present in published scientific surveys. Lastly, for longer surveys, models like *Claude-3-Haiku*[10] may offer superior performance due to their extended context handling capabilities.

Data Contamination. We acknowledge the possibility of data contamination, as some surveys or key references (§3.1) used are open access and may have been included in the training data of the LLMs, potentially leading to slightly different performance estimates. Although we do not explicitly control for this factor in our evaluation process, such contamination is a general challenge in benchmarking LLMs on open-domain generation tasks (Xu et al., 2024). Moreover, since all baselines in our study are based on mainstream LLMs, any potential contamination would be shared and thus unlikely to impact the relative comparison.

Evaluation Sample Scope. While our empirical evaluation focuses on a subset of 120 relatively short surveys spanning multiple disciplines—selected to balance cost and feasibility in *academic settings*—we expect similar performance trends to hold across the full dataset. We encourage the broader research community to further benchmark models using our dataset and framework to extend our findings across broader contexts.

[10] https://www.anthropic.com/news/claude-3-haiku

Ethics Statement

Data Collection, Ethics, and Licensing. Our `SurveyGen` dataset is constructed based on S2ORC (Lo et al., 2020), a large corpus of scientific papers released under the CC BY-NC 4.0[11]. The dataset includes metadata extracted from the papers, such as author names, venue names, citation counts, and h-index values. No sensitive personal data (e.g., contact details or affiliations) is included. All metadata was collected in compliance with the terms of their sources and is used strictly for non-commercial academic research. The dataset is not intended for ranking or evaluating individuals or venues. We are committed to handling the data responsibly and ethically and will release our dataset under the same non-commercial license to ensure transparency and responsible data usage.

Caution about Use of LLMs. While our QUAL-SG framework leverages LLMs to generate scientific surveys and strives to maintain the factual accuracy of the literature, there remains a concern of factual inconsistencies during the generation process. We advise users to critically evaluate the generated content, especially when using it for subsequent scientific research or practical applications. The LLM-generated survey is for reference only and should not be regarded as a substitute for peer-reviewed articles or expert judgment.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (No.72074113) and the Natural Sciences and Engineering Research Council of Canada (NSERC). We gratefully acknowledge the Digital Research Alliance of Canada (CCDB) for providing GPU resources. Mir Tafseer Nayeem is supported by a Huawei PhD Fellowship. We thank Yi Zhao, Heng Zhang, Wenqing Wu, and the anonymous reviewers for their valuable feedback. Tong Bao also thanks his parents and his girlfriend (S. Song) for supporting him during his visit to the University of Alberta.

References

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025a. Litllm: A toolkit for scientific literature review. *Preprint*, arXiv:2402.01788.

[11] https://creativecommons.org/licenses/by-nc/4.0/deed.en

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy DJ Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. 2025b. Litllms, llms for literature review: Are we there yet? *Preprint*, arXiv:2412.15249.

Anthropic. 2025. Claude 3.7 sonnet and claude code. https://www.anthropic.com/news/claude-3-7-sonnet. Accessed: 2025-02-24.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego De Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, and 9 others. 2022. Improving language models by retrieving from trillions of tokens. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2206–2240.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901.

Isabel Cachola, Kyle Lo, Arman Cohan, and Daniel Weld. 2020. TLDR: Extreme summarization of scientific documents. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4766–4777, Online. Association for Computational Linguistics.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 8359–8388.

Arman Cohan, Franck Dernoncourt, Doo Soon Kim, Trung Bui, Seokhwan Kim, Walter Chang, and Nazli Goharian. 2018. A discourse-aware attention model for abstractive summarization of long documents. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 615–621, New Orleans, Louisiana. Association for Computational Linguistics.

Edilson A. Corrêa Jr., Filipi N. Silva, Luciano da F. Costa, and Diego R. Amancio. 2017. Patterns of authors contribution in scientific manuscripts. *Journal of Informetrics*, 11(2):498–510.

DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025. Deepseek-v3 technical report. *Preprint*, arXiv:2412.19437.

Jay DeYoung, Iz Beltagy, Madeleine van Zuylen, Bailey Kuehl, and Lucy Lu Wang. 2021. MS^2: Multi-document summarization of medical studies. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7494–7513, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Naveen Donthu, Satish Kumar, Debmalya Mukherjee, Nitesh Pandey, and Weng Marc Lim. 2021. How to conduct a bibliometric analysis: An overview and guidelines. *Journal of Business Research*, 133:285–296.

Moe Elbadawi, Hanxiang Li, Abdul W. Basit, and Simon Gaisford. 2024. The role of artificial intelligence in generating original scientific research. *International Journal of Pharmaceutics*, 652:123741.

Elsevier. 2025. Scopus search api. https://dev.elsevier.com/documentation/ScopusSearchAPI.wadl.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Team GLM, :, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, and 40 others. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *Preprint*, arXiv:2406.12793.

Reto Gubelmann, Aikaterini-lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023. When truth matters - addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 24–39, Toronto, Canada. Association for Computational Linguistics.

Diana Hicks, Paul Wouters, Ludo Waltman, Sarah De Rijcke, and Ismael Rafols. 2015. Bibliometrics: the leiden manifesto for research metrics. *Nature*, 520(7548):429–431.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2022. Atlas: Few-shot learning with retrieval augmented language models. *Preprint*, arXiv:2208.03299.

Ilias Kanellos, Thanasis Vergoulis, Dimitris Sacharidis, Theodore Dalamagas, and Yannis Vassiliou. 2021. Impact-based ranking of scientific publications: A survey and experimental evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 33(4):1567–1584.

Chandrakant S. Karigar and Shwetha S. Rao. 2011. Role of microbial enzymes in the bioremediation of pollutants: A review. *Enzyme Research*, 2011(1):805187.

Tetsu Kasanishi, Masaru Isonuma, Junichiro Mori, and Ichiro Sakata. 2023. SciReviewGen: A large-scale dataset for automatic literature review generation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 6695–6715, Toronto, Canada. Association for Computational Linguistics.

Yuxuan Lai, Yupeng Wu, Yidan Wang, Wenpeng Hu, and Chen Zheng. 2025. Instruct large language models to generate scientific literature survey step by step. In *Natural Language Processing and Chinese Computing*, pages 484–496.

Vincent Larivière, Nadine Desrochers, Benoît Macaluso, Philippe Mongeon, Adèle Paul-Hus, and Cassidy R Sugimoto. 2016. Contributorship and division of labor in knowledge production. *Social Studies of Science*, 46(3):417–435. PMID: 28948891.

Shira A. Lehr, Aylin Caliskan, Sanjaya Liyanage, and Mahzarin R. Banaji. 2024. Chatgpt as research scientist: Probing gpt's capabilities as a research librarian, research ethicist, data generator, and data predictor. *Proceedings of the National Academy of Sciences of the United States of America*, 121(35):e2404328121.

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. *Preprint*, arXiv:2502.14776.

Zihang Lin, Yian Yin, Lu Liu, and Dashun Wang. 2023. Sciscinet: A large-scale open data lake for the science of science research. *Scientific Data*, 10(1):315.

Shuaiqi LIU, Jiannong Cao, Ruosong Yang, and Zhiyuan Wen. 2022. Generating a structured summary of numerous academic papers: Dataset and method. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 4259–4265. Main Track.

Kyle Lo, Lucy Lu Wang, Mark Neumann, Rodney Kinney, and Daniel Weld. 2020. S2ORC: The semantic scholar open research corpus. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4969–4983, Online. Association for Computational Linguistics.

Yao Lu, Yue Dong, and Laurent Charlin. 2020. Multi-XScience: A large-scale dataset for extreme multi-document summarization of scientific articles. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8068–8074, Online. Association for Computational Linguistics.

Meta. 2024. Introducing llama 3.1: Our most capable models to date. `https://ai.meta.com/blog/meta-llama-3-1/`. Accessed: 2024-07-23.

OpenAI. 2024. Gpt-4o. `https://openai.com/index/hello-gpt-4o/`. Accessed: 2025-05-11.

OpenAI. 2025. Introducing gpt-4.1 in the api. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-05-11.

Justin Paul and Alex Rialp Criado. 2020. The art of writing literature review: What do we know and what do we need to know? *International Business Review*, 29(4):101717.

Jason Priem, Heather Piwowar, and Richard Orr. 2022. Openalex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. In *26th International Conference on Science, Technology and Innovation Indicators (STI 2022)*.

Zehan Qi, Rongwu Xu, Zhijiang Guo, Cunxiang Wang, Hao Zhang, and Wei Xu. 2024. $long^2rag$: Evaluating long-context & long-form retrieval-augmented generation with key point recall. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 4852–4872, Miami, Florida, USA. Association for Computational Linguistics.

David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. 1986. Learning representations by back-propagating errors. *Nature*, 323:533–536.

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Hannah Snyder. 2019. Literature review as a research methodology: An overview and guidelines. *Journal of Business Research*, 104:333–339.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. 2023. Is ChatGPT good at search? investigating large language models as re-ranking agents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14918–14937, Singapore. Association for Computational Linguistics.

Liyan Tang, Zhaoyi Sun, Betina Idnay, Jordan G Nestor, Ali Soroush, Pierre A Elias, Ziyang Xu, Ying Ding, Greg Durrett, Justin F Rousseau, and 1 others. 2023. Evaluating large language models on medical evidence summarization. *npj Digital Medicine*, 6:158.

Xuemei Tang, Xufeng Duan, and Zhenguang G. Cai. 2025. Large language models for automated literature review: An evaluation of reference generation, abstract writing, and review composition. *Preprint*, arXiv:2412.13612.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, and 1331 others. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Raymond J. Torraco. 2005. Writing integrative literature reviews: Guidelines and examples. *Human Resource Development Review*, 4(3):356–367.

Athanasios Voulodimos, Nikolaos Doulamis, Anastasios Doulamis, and Eftychios Protopapadakis. 2018. Deep learning for computer vision: A brief review. *Computational intelligence and neuroscience*, 2018(1):7068349.

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. 2024. Autosurvey: Large language models can automatically write surveys. In *Advances in Neural Information Processing Systems*, volume 37, pages 115119–115145.

Bert Van Wee and David Banister. 2016. How to write a literature review paper? *Transport Reviews*, 36(2):278–288.

Shican Wu, Xiao Ma, Dehui Luo, Lulu Li, Xiangcheng Shi, Xin Chang, Xiaoyun Lin, Ran Luo, Chunlei Pei, Changying Du, Zhi-Jian Zhao, and Jinlong Gong. 2025. Automated literature research and review-generation method based on large language models. *National Science Review*, 12(6):nwaf169.

Ruijie Xu, Zengzhi Wang, Run-Ze Fan, and Pengfei Liu. 2024. Benchmarking benchmark leakage in large language models. *Preprint*, arXiv:2404.18824.

Michihiro Yasunaga, Jungo Kasai, Rui Zhang, Alexander R Fabbri, Irene Li, Dan Friedman, and Dragomir R Radev. 2019. Scisummnet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 7386–7393.

Supplementary Material: Appendices

A Related Work

Dataset for Scientific Literature Summarization: While scientific literature summarization has been extensively studied, most available datasets are limited to single-document scenarios. For instance, SciTLDR (Cachola et al., 2020) contains both author-written and expert-derived summaries for scientific paper summarization tasks. Cohan et al. (2018) introduced a dataset from PubMed and arXiv for long document summarization. However, real-world scientific writing often integrates insights from multiple studies, which requires multi-document summarization datasets. To address this, Lu et al. (2020) proposed Multi-XScience, which extends single-document summarization by incorporating multiple source papers to generate a cohesive summary. DeYoung et al. (2021) proposed MS² for summarizing numerous medical studies to generate comprehensive surveys.

The work most similar to ours is SciReviewGen (Kasanishi et al., 2023), which created a dataset of over 10,000 surveys in the computer science domain with cited references within the surveys. Our dataset differs in that **SurveyGen** additionally provides extensive metadata for all the referenced papers for evaluation purposes, including bibliographic information for papers (e.g., title, abstract, topics), citation performance (e.g., citation count, influential citation count), author-level influence indicators (e.g., publication count, h-index, and total citations), and venue-level reputation metrics (e.g., h-index, mean-citedness, i10-index). Unlike SciReviewGen, which focuses primarily on survey generation, our **SurveyGen** offers a more comprehensive benchmark for assessing citation reliability, content quality, and structural alignment in LLM-generated surveys. Finally, **SurveyGen** also supports the evaluation of surveys across multiple disciplines, while SciReviewGen is limited to computer science.

Automatic Literature Survey Generation with LLMs: While LLMs have demonstrated impressive performance in text generation tasks, generating content that meets the accuracy, structure, and logical coherence required for scientific surveys remains a challenge (Tang et al., 2023; Lehr et al., 2024; Elbadawi et al., 2024).To address this issue, some studies integrate RAG techniques with LLMs and define output templates to control the structure of the generated content (Lai et al., 2025; Agarwal et al., 2025b; Tang et al., 2025). For instance, Wang et al. (2024) proposed AutoSurvey, which employs a two-stage generation strategy: first, retrieving relevant literature to generate a detailed outline, and then drafting individual sections and integrating them into a cohesive review. Similarly, Liang et al. (2025) introduced SurveyX, which employs online reference retrieval to gather relevant literature and utilizes a pre-processing method called AttributeTree to extract and organize key information from these sources. Wu et al. (2025) implemented a multi-layered quality control strategy to mitigate hallucination issues during the literature review generation process. While the above studies provide valuable insights into this task, our work offers more reliable sources, improved retrieval strategies, and a more rigorous evaluation against human-written surveys to explore the upper limits of LLMs.

B Input and Output Settings

The input and output texts for the three tasks are as follows:

Task 1: The LLMs are provided only with the survey topic. They are first prompted to generate a structured outline along with brief descriptions for each section, and then to produce the full survey content based on that outline.

Task 2: During the retrieval stage, the survey topic is used as a query to retrieve relevant literature from external databases. In the generation stage, the input includes the survey topic, along with the titles, abstracts, and quality-related metadata of the retrieved papers. The generation process follows the same steps as in Task 1, where the LLMs are instructed first to generate an outline with section descriptions and then write the corresponding content for each section to form the final survey.

Task 3: In this task, all references are sourced from human-written surveys, and the bibliographic information provided for each reference is consistent with that used in Task 2. In addition, we provide the outline of each human-written survey, with all cited references grouped under their corresponding sections. The LLMs are then instructed to generate each section using the selected references and the corresponding outline information.

C Implementation Details

During the literature retrieval stage, we utilize **Semantic Scholar**[12] as the external literature database, and use the **bge-large-en-v1.5**[13] as the embedding model to compute semantic similarity throughout the pipeline. To mitigate potential self-evaluation bias, **GPT-4o** (OpenAI, 2024) is selected as the LLM agent for evaluation purposes.

To implement the QUAL-SG framework, we first use the S2ORC API[14] to retrieve the 300 papers most relevant to the survey topic, as a trade-off between literature coverage and computational efficiency. To ensure temporal consistency, we restrict this set to works published before the survey's publication date. The retrieved sets are then ranked by abstract–topic semantic similarity, and we further identify the 50 most frequently co-cited papers and add them to the original candidate pool.

For literature re-ranking, we use the paper's citation count[15], the sum of the first and last author's h-index—the last author often being the corresponding or supervising author associated with publication quality (Larivière et al., 2016; Corrêa Jr. et al., 2017), and the venue's h-index to represent citation performance, author influence, and venue reputation, respectively. We assign $\gamma = 0.5$ to citation performance, $\beta = 0.3$ to venue reputation, and $\alpha = 0.2$ to author influence. Based on the final weighted scores, we rank the articles and select a final subset of references that matches the reference count of the corresponding human-written survey for evaluation. The weights were chosen based on the author's intuition and preliminary analysis to reflect the relative importance of citation performance, venue reputation, and author influence. While not exhaustive, these values offer a practical starting point for evaluation. Importantly, our framework is modular and supports alternative weight configurations based on downstream needs.

For structural consistency evaluation, we removed non-content sections such as *"funding"*, *"acknowledgements"*, *"author contributions"*, *"competing interests"*, and *"supplementary material"*.

[12] https://api.semanticscholar.org/api-docs/datasets

[13] https://huggingface.co/BAAI/bge-large-en-v1.5

[14] https://api.semanticscholar.org/graph/v1/paper/search

[15] Citation counts are normalized by the number of years since publication to control for citation accumulation bias.

D Evaluation Metrics

D.1 Metric Formulations

Citation Quality: We compute the precision, recall, and F1 score of LLM-generated or RAG-retrieved candidate references with human-selected references as follows:

$$\text{Precision}_{\text{cite}} = \frac{R_L \cap R_H}{R_L}$$

$$\text{Recall}_{\text{cite}} = \frac{R_L \cap R_H}{R_H}$$

$$\text{F1}_{\text{cite}} = 2 \times \frac{\text{Precision}_{\text{cite}} \times \text{Recall}_{\text{cite}}}{\text{Precision}_{\text{cite}} + \text{Recall}_{\text{cite}}}$$

Here, R_L and R_H denote the sets of references generated or retrieved by the LLM and those selected by humans, respectively, and $\cap$ denotes set intersection. A reference is considered a match if its textual similarity exceeds 0.95, as determined from our preliminary experiments.

To evaluate the accuracy of LLM-generated references, we perform title searches to check whether the generated reference yields an exact match with an existing publication in S2ORC databases.

Content Quality: To measure Key Point Recall (KPR) (Qi et al., 2024) for generated surveys, we first instruct the LLMs to extract key points from the human-written survey. We then verify whether each extracted key point is captured in the corresponding LLM-generated survey using a question-answering (QA) approach. The KPR is defined as follows:

$$KPR(H_i, G) = \begin{cases} 1 & \text{if } H_i \text{ is present in } G, \\ 0 & \text{otherwise.} \end{cases}$$

where H_i is the i-th key point extracted in the human-written survey. G is the LLM-generated survey. A higher KPR score indicates that the LLM-generated survey covers more key points from the human-written ones.

Structural Consistency: In structural consistency evaluation, for the structural overlap, we set the semantic similarity threshold to 0.8, as our preliminary experiments showed it to be optimal for identifying valid matches. The calculation formula is defined as follows:

$$M(S_H^i, S_G^j) = \begin{cases} 1 & \text{if } S_H^i \text{ and } S_G^j \text{ are matching,} \\ 0 & \text{otherwise.} \end{cases}$$

where S_H^irepresents the i-th section from the human-written survey. S_G^j represents the j-th section from the LLM-generated survey. We define the

Task	Comparison	Topic Relevance	Information Coverage	Critical Analysis	Overall Rating
Task 1	Comparable	33.3%	33.3%	26.7%	20.0%
	LLM-Generated > Human-written	20.0%	26.7%	26.7%	13.3%
Task 2	Comparable	33.3%	46.7%	40.0%	26.7%
	LLM-Generated > Human-written	33.3%	20.0%	20.0%	13.3%
Task 3	Comparable	40.0%	53.3%	46.7%	26.7%
	LLM-Generated > Human-written	26.7%	20.0%	20.0%	20.0%

Table 6: Human evaluation results across tasks. Each task includes five surveys from the Computer Science domain, all generated using Claude-3.7-Sonnet. For Task 2, the surveys were generated from the QUAL-SG pipeline.

structural consistency between the two as follows:

$$S_{\text{struct}} = \frac{2 \times (S_H \cap S_G)}{S_H + S_G}$$

where S_H, S_G represent the number of sections in the human-written and LLM-generated surveys, respectively, $S_H \cap S_G$ is the number of matching sections between the two.

We then use LLM-as-judge to score structural relevance between the LLM-generated survey and the corresponding human-written ones as follows:

$$\text{Relevance}_{\text{LLM}} = \frac{1}{|S|} \sum_{s \in S} \mathbb{I}_{\text{relevant}}(s, H)$$

D.2 Comparison with Existing Metrics

The *two main differences* in our evaluation design compared to prior work (Wang et al., 2024; Liang et al., 2025; Tang et al., 2025) are as follows:

First, in citation quality evaluation, previous studies primarily assess recall—i.e., how many human-selected citations are recovered by the LLM-generated or RAG-based retrieval. However, we argue that measuring recall alone may overestimate model performance. For example, an LLM might generate 8 out of 10 citations from a human-written survey (80% recall), but also includes over 20 additional irrelevant references; the overall citation reliability is significantly compromised. Therefore, we further introduce citation precision to provide a more balanced assessment of citation quality.

Second, in structural evaluation, prior work mainly examines how well the content aligns with the research topic. In contrast, we directly compare the LLM-generated outline with the survey structure of human-written surveys. This is motivated by the fact that human-written outlines, which are carefully designed and peer-reviewed, better reflect the scope and logic of the survey. Taking human-written surveys as a gold standard, this finer-grained structural evaluation enables a more precise assessment to identify which types of sections are well-covered, missing, or over-generated by the LLMs, helping reveal the specific aspects where LLMs still fall short.

Human evaluation criteria

Topic Relevance: *whether the survey maintains a clear focus on the assigned topic, with each section contributing to the central topic without digressions?*

Information Courage: *whether the survey includes key papers, major developments, and diverse research approaches relevant to the topic?*

Critical Analysis: *whether the survey compares methods or findings, identifies limitations or open challenges, and offers insight rather than descriptive summaries?*

Overall Rating: *whether the survey is well-written, logically structured, and academically appropriate, and would be considered the better survey in comparison?*

Figure 5: Human evaluation criteria

E Human Evaluation Protocol

Given that scientific survey evaluation requires specific domain expertise and is time-consuming, we randomly select 5 surveys from the computer science domain for each task, resulting in a total of 15 surveys for human evaluation. Each LLM-generated survey is paired with its corresponding human-written survey to form an evaluation pair. We then invite three second-year PhD students with a background in computer science as annotators, each of whom has published at least one peer-reviewed paper, to compare the LLM-generated and human-written surveys from the following four aspects: ***topic relevance***, ***information coverage***, ***critical analysis***, and ***overall rating***. For each pair, annotators are asked to compare the two surveys and judge which one is better, comparable, or worse. To mitigate potential bias, all identifying information was removed, and annotators **were not informed** whether the surveys were LLM-generated or human-written. The human evaluation criteria are illustrated in Figure 5,

Ablation Setting	P ↑	R ↑	F1 ↑
QUAL-SG	**15.87**	**17.71**	**16.73**
w/o co-cited expansion	10.07 (↓5.80)	11.52 (↓6.19)	10.75 (↓5.98)
w/o topical relevance	11.54 (↓4.33)	13.15 (↓4.56)	12.29 (↓4.44)
w/o academic impact	8.76 (↓7.11)	9.28 (↓8.43)	9.01 (↓7.72)
w/o content diversity	<u>13.16</u> (↓2.71)	<u>14.34</u> (↓3.37)	<u>13.72</u> (↓3.01)

Table 7: Ablation study of QUAL-SG in the literature retrieval stage. The best results are marked **bold** and the second-best are <u>underlined</u>.

Key References		Models				
S2ORC ID	Citation Count	Human Selected	Fully-LLMGen	Naive-RAG	QUAL-SG (Ours)	Generated Context
Title: Deep Learning for Computer Vision: A Brief Review (Voulodimos et al., 2018)						
57246310	61709	✓	✓	×	✓*	*. . . , the availability of large annotated datasets, exemplified by* **[ref]** *, provided. . .*
10328909	60469	✓	×	×	✓*	*. . . and its variants are commonly employed in object detection frameworks Faster R-CNN* **[ref].**
2930547	38960	✓	✓	×	✓*	*. . . achieving unprecedented accuracy on the ImageNet Large Scale Visual Recognition* **[ref].**
2309950	16043	✓	×	×	✓*	*. . . and seminal work like Spatial Pyramid Pooling in Deep Convolutional Networks for Visual Recognition* **[ref]**
Title: Role of Microbial Enzymes in the Bioremediation of Pollutants: A Review (Karigar and Rao, 2011)						
83928450	812	✓	✓	×	✓*	*. . . among enzymatic methods, laccases stand out for phenol degradation and lignin transformation* **[ref]** *, making them valuable in...*
1624267	519	✓	×	×	✓*	*. . . with recent advances in enzyme engineering and DNA shuffling* **[ref]** *, enhancing...*
84754528	119	✓	×	✓	✓	*. . . White rot fungi can degrade chlorinated phenolics via enzyme systems for paper industry cleanup* **[ref].**

Table 8: Case study with two surveys from our *SurveyGen*. The S2ORC ID refers to the article's ID in the S2ORC corpus. An * indicates that the paper was not retrieved via semantic similarity but was identified as a highly co-cited reference and therefore included in the candidate pool.

and the corresponding evaluation results are summarized in Table 6 of Section 4.5.

F Additional Results

F.1 Ablation Study

We present an ablation study of the QUAL-SG framework by individually removing each of the four key components to assess their contributions: **(1)** co-cited paper expansion; **(2)** relevance-based ranking; **(3)** academic impact-based ranking; and **(4)** content diversity-based ranking.

As shown in Table 7, QUAL-SG shows performance drops across all ablation settings. Removing academic impact-based ranking (-7.72%) and co-cited paper expansion (-5.98%) caused the most significant drops. This highlights the importance of expanding candidate pools via citation analysis and identifying high-impact research for reference selection. Furthermore, topical relevance and content diversity were also shown to contribute positively.

F.2 Case Analysis

We conduct a case analysis using two surveys from the Computer Science and Biology domains. As shown in Table 8, both Fully-LLMGen and Naive-RAG failed to identify several crucial, human-selected references. Notably, Naive-RAG retrieves only one valid reference (out of seven) due to weak semantic similarity between reference abstracts and the topic; however, these papers are frequently cited by other works, indicating their academic influence despite low semantic similarity. QUAL-SG succeeds in recovering all key papers through two core strategies: first, expanding the candidate pool via co-citation analysis, which allows the inclusion of semantically distant yet influential works; and second, ranking candidates by quality to identify the most impactful studies and better highlight their contributions in the generated survey.

G Topic Examples for Survey Generation

S2ORC ID	Topic	Citation
	Biology	
13599358	Microbial Enzymes in Pollutant Bioremediation	628
11116464	Lactic Acid Bacteria and Bacteriocins	457
6209474	Mathematical Models of Malaria	361
5068313	Effects of Deoxynivalenol and Type B Trichothecenes on the Intestine	284
19692413	Perio-Pathogenic Bacteria in Oral Carcinogenesis	187
15915856	Monosodium Glutamate Toxic Effects and Implications for Human Intake	140
17610865	Marine N-3 Fatty Acids and Type 2 Diabetes Risk	129
39756789	Neonicotinoid Insecticides and Developmental Neurotoxicity	109
220843996	Detection of Human Intestinal Protozoan Parasites in Vegetables and Fruits	89
1100406	Cyanobacterial Natural Products: Structure, Properties, and Applications	87
	Medicine	
17136958	Dietary Sugars and Body Weight in Randomised Controlled Trials	1583
52095775	Flavonoids and Phenolic Compounds from Medicinal Plants	1408
263077223	Patient Engagement in Research	1135
17464731	Delirium Outcomes in Critically Ill Patients	793
212709676	Traditional Chinese Medicine in Treating SARS-CoV-2 Infections	769
4893818	Education and Dementia in the Context of the Cognitive Reserve Hypothesis	740
51958985	ADHD Medications: Efficacy and Tolerability	705
13897386	Short Term Air Pollution Exposure and Stroke	705
219607020	Amyotrophic Lateral Sclerosis: Clinical Perspectives	657
6017773	Maternal Smoking During Pregnancy and Associated Birth Defects	626
	Psychology	
1845793	Neuroimaging Studies of Internet and Gaming Addiction	377
52293261	Technology-Delivered Interventions for Youth Depression and Anxiety	212
18781074	Music Therapy and Cognitive Function in Alzheimer's Disease	174
20918937	Gender Dysphoria and Autism Spectrum Disorder	173
4033267	Figurative Language Comprehension in Autism Spectrum Disorder	172
2802244	Propranolol in Anxiety Disorder Treatment	170
4152509	ADHD Prevalence in Chinese Children and Adolescents	161
54447675	Influence of Role Models on Gender and Careers	156
3916988	Motivation in Health Education and Self-Determination Theory	152
12043081	Fundamental Criteria for Eating Disorder Recovery	141
	Computer Science	
10137425	Multimodal Machine Learning Taxonomy	2737
3557281	Deep Learning Applications in Computer Vision	2709
218474694	IoT Sensing with RFID and Wireless Sensor Networks	269
235410640	Deep Multimodal Learning for Computer Vision	266
258180322	Fairness and Bias in Artificial Intelligence	216
232300174	Facial Micro-Expression Analysis	180
231802191	Synthetic-CT Generation in Radiotherapy and PET	131
208035941	Detecting Sleep Apnea Using Deep Learning	115
186207561	Head-Mounted Eye Gaze Tracking Devices	99
150036628	Electrical Impedance Tomography and AI Applications	90

H SurveyGen Data Format

```
{
  "corpusId": integer,
  "metadata": {
     "title": string,
     "abstract": string,
     "authors": [{"authorId": string, "h-Index": string, "...": "..." }],
     "journal": {"name": string, "h-Index": integer, "...": "..."},
     "...": "..."},
  "sections": [
    {
     "index": integer,
     "title": string,
     "paragraphs": [
       {
         "text": string,
         "citations": [{"ref_id": string,  "title": string, "...": "..." }]
       } ],
     "citations_in_section": [ref_id]
    }
  ],
  "references": [
    {
     "ref_id": string,
     "matched_paper_id": integer,
     "metadata": {
       "title": string,
       "abstract": string,
       "authors": [{"authorId": string, "h-Index": string, "...": "..."}],
       "journal": { "name": string, "h-Index": integer, "...": "..." },
       "...": "..."
     } } ]
}
```

I Prompt used in this study

Prompt for Survey-type Article Classification

You are an academic expert in scientific literature classification. Your task is to determine whether the following paper is a survey-type article, based on its title and abstract below:

Title: {title}
Abstract: {abstract}

Please make your judgment based on the following three criteria:
1. The abstract explicitly declares a survey intent (e.g., phrases like *"conducts a survey", "provides an overview", "this survey..."*, etc.).
2. The focus of the paper is on reviewing or summarizing existing work, rather than proposing new methodologies or reporting novel experimental results.
3. The abstract provides a forward-looking perspective on the field by synthesizing the reviewed literature to identify key trends, highlight significant open challenges or gaps, and suggest promising directions for future investigation.

Your output should be only one word: **"TRUE"** if the paper qualifies as a survey-type article, or **"FALSE"** if it does not. Do not include any additional commentary, explanation, or formatting instructions.

Prompt for Topic Relevance Evaluation

You are an academic expert helping to write a survey on the topic: "**{TOPIC}**".
You will be provided with the title and abstract of a research paper. Your task is to rating the paper's relevance to the survey topic based on the following criteria:

Score-1 (poor): The paper is unrelated to the topic or only mentions it in passing, with no meaningful contribution.
Score-2 (low): The paper is loosely connected or provides general background, but not focused on the topic.
Score-3 (moderate): The paper discusses a specific sub-aspect of the topic; somewhat useful, but not central.
Score-4 (high): The paper substantially addresses key elements of the topic; would likely be cited in the survey.
Score-5 (very high): The paper is entirely focused on the topic, offering essential insights; likely foundational to the survey.

Your output should only include a single score from 1 to 5. Do not provide any explanation or additional text.

Title:title of the paper Abstract:abstract of the paper

Topical relevance score:

Prompt for Outline Generation (Task1)

You are an academic expert in the field of "**{TOPIC}**" with deep expertise in survey writing. Your task is to generate a well-structured outline for a survey on this topic.

Please follow the instructions below:

Step 1: Based on the given topic, identify 3 to 7 major thematic sections that define the overall scope and objectives of the survey. For each section, provide an academically styled title, along with a brief description summarizing its focus and relevance.

Step 2: For the major thematic sections, list several subsections representing more specific research areas, concepts, or points to be covered. Subsections should be conceptually related to their parent section and serve to further structure the survey.

Step 3: Your output should be in JSON format and must include the survey title, a structured outline with section titles, descriptions, and subsections. Ensure the structure is logically coherent, well-aligned with the topic, and suitable for developing a full-length academic survey.

The output example should follow the format below:

```
{
  "title": "TITLE OF THE SURVEY",
  "outline": [
    {
      "section_title": "SECTION TITLE 1",
      "description": "A brief description of this section.",
      "subsections": {
        "subsection title 1": "content",
        "...": "...",
        "subsection title n": "content"
      }
    },
    {
      "section_title": "SECTION TITLE 2",
      "description": "A brief description of this section.",
      "subsections": {
        "subsection title 1": "content",
        "...": "...",
        "subsection title n": "content"
      }
    }
  ]
}
```

Now, based on the given topic {Topic}, please generate the outline by following the steps above. Do not include any additional commentary, explanation, or formatting instructions; only return the structured JSON output as specified.

Prompt for Section Content Generation (Task 1)

You are an academic expert in the field of "**{TOPIC}**" with deep expertise in survey writing. Your task is to write a subsection of a survey, based on the following information:

The overall structure of the survey is as follows:

- {{outline}}

You are now asked to write the following subsection:

- Subsection: {{subsection_title}}

Instructions for generating subsection content:
1. Ensure the content directly addresses the specific topic of the subsection, and the generated content should be a minimum of 300 words.

2. Ensure that all claims are fully supported by relevant academic literature. Cite each reference using in-text citations in the format `ref [1]`, `ref [2]`, etc. If a source is cited multiple times, use only the reference number assigned to its first occurrence.

3. Maintain alignment with the parent section and overall survey topic, ensuring thematic and conceptual consistency.

4. Use a formal academic tone, with logically structured arguments and scholarly language.

5. Your output should strictly be a JSON object with the following two fields.

```
{
    "content": "the content of the subsection",
    "references": [
    {
        "refNo": "reference number",
        "authors": "full author list",
        "year": "year of publication",
        "title": "title of the paper",
        "venue": "publication source",
        "doi": "DOI"
    }
    ]
}
```

Now, please generate the content for the given subsection: {subsection_title}. Please ensure the output is formatted according to the requirements mentioned above. Do not include any explanation, commentary, or preamble in your response.

Prompt for Outline Generation (Task 2)

You are an academic expert in the field of "**{TOPIC}**" with deep expertise in survey writing. Your task is to generate a well-structured outline for a survey on this topic.

The following highly relevant papers, including their titles and abstracts, are provided for reference:

- {{references list}}

Please follow the instructions below:

Step 1: Based on the given topic, identify 3 to 7 major thematic sections that define the overall scope and objectives of the survey. For each section, provide an academically styled title, along with a brief description summarizing its focus and relevance.

Step 2: For the major thematic sections, list several subsections representing more specific research areas, concepts, or points to be covered. Subsections should be conceptually related to their parent section and serve to further structure the survey.

Step 3: Your output should be in JSON format and must include the survey title, a structured outline with section titles, descriptions, and subsections. Ensure the structure is logically coherent, well-aligned with the topic, and suitable for developing a full-length academic survey.

The output example should follow the format below:

```
{
  "title": "TITLE OF THE SURVEY",
  "outline": [
    {
      "section_title": "SECTION TITLE 1",
      "description": "A brief description of this section.",
      "subsections": {
        "subsection title 1": "content",
        "...": "...",
        "subsection title n": "content"
      }
    },
    {
      "section_title": "SECTION TITLE 2",
      "description": "A brief description of this section.",
      "subsections": {
        "subsection title 1": "content",
        "...": "...",
        "subsection title n": "content"
      }
    }
  ]
}
```

Now, based on the given topic {Topic}, please generate the outline by following the steps above. Do not include any additional commentary, explanation, or formatting instructions; only return the structured JSON output as specified.

Prompt for Section Content Generation (Task 2)

You are an academic expert in the field of "**{TOPIC}**" with deep expertise in survey writing. Your task is to write a subsection of a survey, based on the following information:

The overall structure of the survey is as follows:

- {{outline}}

You are now asked to write the following subsection:

- Subsection: {{subsection_title}}

The following highly relevant papers, including their titles and abstracts, are provided for reference:

- Subsection: {{references_list}}

Instructions for generating subsection content:

1. Carefully analyze the provided references and identify those highly relevant to the subsection topic as the basis for your generation.

2. Ensure the content directly addresses the specific topic of the subsection, and the generated content should be a minimum of 300 words.

3. Ensure that all claims are fully supported by relevant academic literature. Cite each reference using in-text citations in the format `ref [1]`, `ref [2]`, etc. If a source is cited multiple times, use only the reference number assigned to its first occurrence.

4. Maintain alignment with the parent section and overall survey topic, ensuring thematic and conceptual consistency.

5. Use a formal academic tone, with logically structured arguments and scholarly language.

6. Your output should strictly be a JSON object with the following two fields.

```
{
    "content": "the content of the subsection",
    "references": [
    {
        "refNo": "reference number",
        "title": "title of the paper",
    }
    ]
}
```

Now, please generate the content for the given subsection: {subsection_title}. Please ensure the output is formatted according to the requirements mentioned above. Do not include any explanation, commentary, or preamble in your response.

Prompt for Section Content Generation (Task 3)

You are an academic expert in the field of "**{TOPIC}**" with deep expertise in survey writing. Your task is to write a subsection of a survey, based on the following information:

The overall structure of the survey is as follows:

- {{outline}}

You are now asked to write the following subsection:

- Subsection: {{subsection_title}}

The following papers are provided for this subsection and should be summarized accordingly:

- {{references_list}}

Instructions for generating subsection content:

1. You must generate the content based on the provided references. Do not incorporate or cite any external references.

2. Ensure the content directly addresses the specific topic of the subsection, and the generated content should be a minimum of 300 words

3. Ensure that all claims are fully supported by relevant academic literature. Cite each reference using in-text citations in the format `ref [1]`, `ref [2]`, etc. If a source is cited multiple times, use only the reference number assigned to its first occurrence.

4. Maintain alignment with the parent section and overall survey topic, ensuring thematic and conceptual consistency.

5. Use a formal academic tone, with logically structured arguments and scholarly language.

Now, please generate the content for the given subsection: {subsection_title}. Please ensure the output is formatted according to the requirements mentioned above. Do not include any explanation, commentary, or preamble in your response.

Prompt for Structural Consistency Evaluation

You are an academic expert tasked with evaluating a draft outline for a literature survey on the topic of "**{TOPIC}**".

You will be provided with two outlines:
1. A draft outline as a preliminary version.
– **LLM-generated Outline**
2. A gold-standard outline created by domain experts.
– **Human-written Outline**

Your task is to rate the structural consistency between the two outlines following the criteria below:

Score 1 (Very Poor): The draft outline is largely inconsistent with the gold-standard; major sections are missing or irrelevant.
Score 2 (Poor): The draft outline only partially overlaps with the gold-standard; some key themes are omitted or misplaced.
Score 3 (Moderate): The draft includes some of the core topics but lacks structural alignment or completeness.
Score 4 (Good): The draft mostly aligns with the gold-standard; minor deviations in structure or scope.
Score 5 (Excellent): The draft closely mirrors the gold standard in both structure and content coverage.

Your output should only include a single score from 1 to 5. Do not provide any explanation or additional text.