DeepSurvey-Bench: Evaluating Academic Value of Automatically Generated Scientific Survey

Anonymous ACL submission

Abstract

The rapid development of automated scientific survey generation technology has made it increasingly important to establish a comprehensive benchmark to evaluate the quality of generated surveys. Nearly all existing evaluation benchmarks rely on flawed selection criteria such as citation counts and structural coherence to select human-written surveys as the ground truth survey datasets, and then use surface-level metrics such as structural quality and reference relevance to evaluate generated surveys. However, these benchmarks have two key issues: (1) the ground truth survey datasets are unreliable because of a lack academic dimension annotations; (2) the evaluation metrics only focus on the surface quality of the survey such as logical coherence. Both issues lead to existing benchmarks cannot assess to evaluate their deep "academic value", such as the core research objectives and the critical analysis of different studies. To address the above problems, we propose **DeepSurvey-Bench**, a novel benchmark designed to comprehensively evaluate the academic value of generated surveys. Specifically, our benchmark propose a comprehensive academic value evaluation criteria covering three dimensions: informational value, scholarly communication value, and research guidance value. Based on this criteria, we construct a reliable dataset with academic value annotations, and evaluate the deep academic value of the generated surveys. Extensive experimental results demonstrate that our benchmark is highly consistent with human performance in assessing the academic value of generated surveys[1].

1 Introduction

In recent years, the rapid development of artificial intelligence (Huynh-The et al., 2023; Wang et al., 2023) and the exponential growth of research papers have made it increasingly difficult for researchers to efficiently organize vast amounts of information, resulting in greater challenges in writing surveys manually. The rise of large language models (LLMs) (Achiam et al., 2023; Team et al., 2024; Liu et al., 2024) provides a new promising avenue to addressing this challenge. By combining their powerful text generation capabilities with retrieval-augmented generation (RAG) technology (Fan et al., 2024; Chang et al., 2025), researchers can automate both literature retrieval and scientific survey generation (Wang et al., 2024; Liang et al., 2025; Bao et al., 2025). With the rapid development of automated scientific survey generation technology, establishing a comprehensive benchmark to evaluate the quality of generated surveys has become an urgent need.

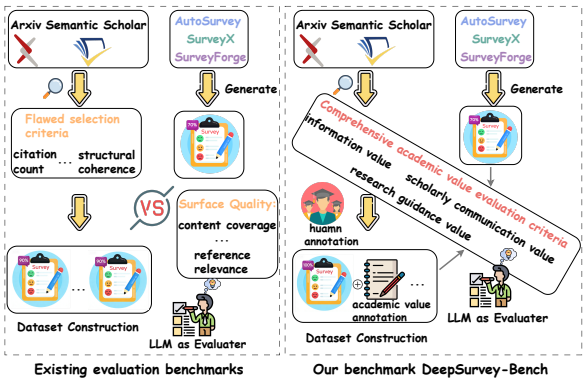

Figure 1: Comparison between existing evaluation frameworks and our proposed DeepSurvey-Bench.

Existing evaluation benchmarks (Yan et al., 2025; Su et al., 2025; Shi et al., 2025) have adopted selection criteria such as citation counts and structural coherence to select high-quality human-written surveys from the two open-access repositories arXiv[2] and semantic scholar[3] as ground truth

[1]The code and datasets of this paper can be obtained from https://anonymous.4open.science/r/DeepSurvey-Bench-AB42/

[2]https://arxiv.org/

[3]https://www.semanticscholar.org/

survey datasets. Then, the automatically generated survey is then compared with the ground truth surveys, and further evaluated under an LLM-as-a-judge paradigm using surface quality metrics such as content coverage and reference relevance, as illustrated in the left half of Figure 1.

However, these evaluation benchmarks face two problems: (1) their selection criteria are relatively one-sided and lack academic dimension annotation, leading to unreliable ground truth survey datasets that fail to accurately reflect the quality of generated surveys; (2) the evaluation metrics solely evaluate the surface quality of surveys. Both problems prevent these benchmarks from evaluating the deep academic value of generated surveys, such as core research objectives and critical analysis of the relationships between different studies. As a result, some generated surveys are often perform well on surface quality metrics, but are actually superficial in content and offer limited substantive academic contribution.

From the perspective of expert evaluation of the quality of a survey, a high-quality survey must possess academic value: it should be grounded in a comprehensive and carefully selected body of relevant literature, clearly state the research objectives or core questions (Torraco, 2005; Denney and Tewksbury, 2013), and conduct in-depth comparison, evaluation, and interpretation of key studies in the field (Snyder, 2019). It should further offer original critical insights, identify gaps and unresolved core issues, and propose new research directions or topics based on practical considerations.(Xiao and Watson, 2019; Kraus et al., 2022).

Inspired by the above insights, we propose **DeepSurvey-Bench**, a novel benchmark designed to comprehensively evaluate academic value of automatically generated scientific surveys, to address the two key issues of existing benchmarks. As illustrated in the right half of Figure 1. Specifically, based on the aforementioned expert insights and with the assistance of researchers experienced in survey writing, the proposed benchmark establishes a comprehensive academic value evaluation criteria across three dimensions: information value that refers to whether existing research is transformed into a well-structured, coherent, and well-argued knowledge system; scholarly communication value that pertains to whether key research can be compared, evaluated, and critically analyzed in depth; and research guidance value that concerns whether it identifies gaps in existing research and proposes feasible new directions. Then, we manually annotate and select surveys based on the criteria, thereby constructing a reliable and high-quality ground truth survey dataset with academic value annotations. Furthermore, to make the evaluation process more specific and feasible, we decompose the criteria into seven quantifiable academic value evaluation metrics to evaluate the academic value of the generated survey.

We conducted a comprehensive evaluation of existing automated survey generation methods on our proposed benchmark. Extensive experimental results demonstrate that our benchmark is aligned with human judgment in assessing the academic value of surveys.

In conclusion, our main contributions are as follows:

- We propose DeepSurvey-Bench, a novel benchmark for comprehensively evaluating academic value of automatically generated scientific survey.
- We propose a comprehensive academic value evaluation criteria. Based on this criteria, we construct a high-quality dataset with academic value annotations and further evaluate the academic value of the generated surveys.
- The experimental results demonstrate that our benchmark is reliable and highly consistent with human performance in assessing the academic value of generated surveys.

2 Related Works

2.1 Survey Generation

Research on the automatic generation of surveys has been ongoing for more than a decade. Early studies primarily relied on multi-document summarization techniques to automatically generate relevant sections within surveys (Chen et al., 2021; Jiang et al., 2019; Erera et al., 2019). In recent years, the emergence of LLMs has created new opportunities for this field. Many recent studies have proposed end-to-end pipelines that integrate RAG and multi-agent strategies to enable the automated generation of long-form surveys (Wang et al., 2024; Liang et al., 2025; Wen et al., 2025). AutoSurvey (Wang et al., 2024) adopts a two-stage generation strategy: it first retrieves relevant literature to construct detailed outlines, and then employs multiple LLMs in parallel to generate individual chapters,

which are subsequently integrated into a coherent survey. Similarly, SurveyX (Liang et al., 2025) decomposes the survey generation process into preparation and generation stages, and incorporates online reference retrieval along with AttributeTree-based preprocessing and a re-polishing process to improve generation efficiency. SurveyForge (Yan et al., 2025) further enhances outline quality and citation accuracy by analyzing the logic of manually written outlines and integrating them with retrieved literature, while leveraging academic navigation agents to identify high-quality papers. In addition, LiRA (Go et al., 2025) employs multiple specialized agents for outline construction, subsection writing, editing, and reviewing, thereby improving readability and factual accuracy.

2.2 Evaluation for Survey Generation

Existing evaluation benchmarks typically use human-written surveys as the gold standard and then evaluate the quality of generated surveys based on the LLM-as-a-judgment paradigm. For example, SurveyBench (Yan et al., 2025) directly adopts human-written surveys selected by researchers based on their experience knowledge as the gold standard, primarily evaluating the performance of the generated surveys in terms of outline structure, topic coverage, and content relevance. In contrast, SurveyScope (Shi et al., 2025) constructs its gold standard by combining objective metrics such as publication time and citation count with researchers' judgment, and evaluates the content quality, organizational logic, and citation accuracy of generated surveys. Furthermore, SurGE (Su et al., 2025) selects its gold-standard surveys through human annotation by researchers based on criteria such as citation count and content coverage, and evaluates the generated survey from multiple dimensions including citation accuracy, structural organization, and content quality. However, these evaluation benchmarks lack human annotation for the academic dimensions of the surveys, resulting in unreliable gold standard datasets. Furthermore, the evaluation metrics only assess the surface quality and fail to evaluate the deep academic value of generated surveys.

3 Survey Generation Task Definition

The objective of automated scientific survey generation is to produce a survey $\mathcal{S}$, given a topic description t and a large academic paper corpus $R = \{r_1, r_2, ..., r_n\}$. The generated survey is expected to comprehensively introduce, analyze, and synthesize research findings within a given field over a specific period. The generation process generally consists of the following three stages: (i) literature retrieval stage, which first retrieves an initial set of papers related to the topic t and then filters them according to predefined criteria to obtain the most relevant set of reference papers $R_s \subseteq R$; (ii) Outline generation stage, which produces a well-structured and logically coherent outline $O = \{o_1, o_2, ..., o_m\}$ based on a given t and a set of relevant references R_s; (iii) survey generation stage, which generates a comprehensive and reliable survey $\mathcal{S}$ based on the topic t, outline O, and relevant references R_s, including appropriate in-text citations and a list of references.

4 DeepSurvey-Bench

To address the issues of unreliable datasets in existing evaluation benchmarks and the inability of current metrics to evaluate the academic value of generated surveys, we propose a reliable and comprehensive benchmark, named DeepSurvey-Bench. Figure 2 illustrates the datasets construction process, which consists of the following three steps: Initial Survey Collection (§4.1), Parsing and Filtering (§4.2) and Human Annotation (§4.3).

4.1 Initial Survey Collection

To ensure the timeliness and relevance of the selected papers, we first identify surveys published between January 1, 2022, and November 30, 2025, whose titles explicitly contain the terms "survey", "literature review", or "review", thereby focusing on recent research trends and the latest developments in the field. In this process, we adopt differentiated citation thresholds based on publication year, which balances the use of citation counts as a reliable metric of academic influence and recognition with the consideration that recently published high-quality papers may not yet have accumulated a large number of citations. Finally, we obtain 1,101 papers preliminarily identified as surveys after the initial collection.

4.2 Parsing and Filtering

Since title-based filtering may still include non-survey papers, inspired by prior work (Bao et al., 2025), we prompt an LLM to evaluate the abstracts of the initially collected papers according to predefined criteria, further filtering out those that con-

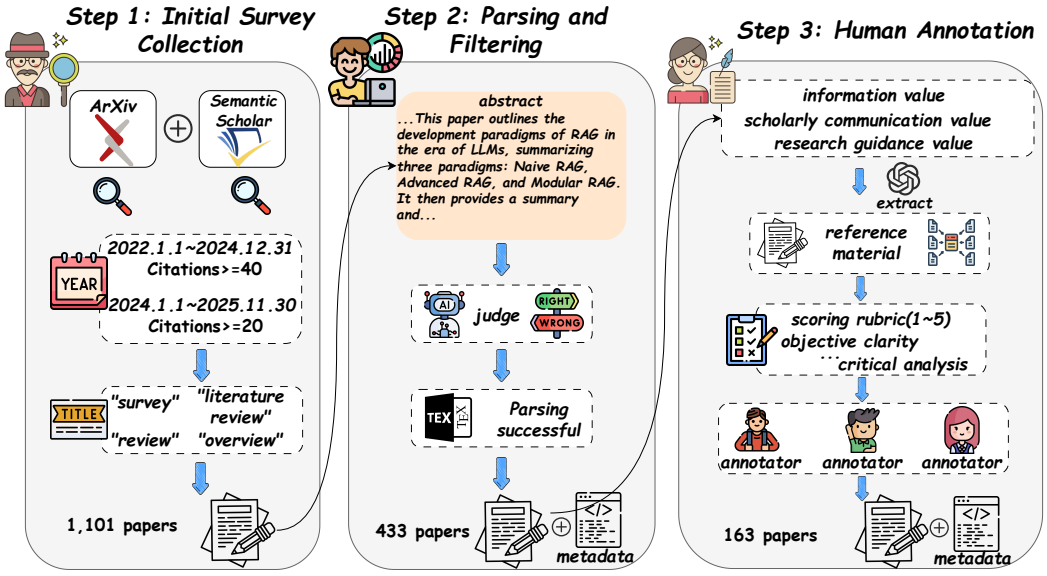

Figure 2: The dataset construction pipeline for our proposed DeepSurvey-Bench.

tained terms such as "survey" or "review" in the title but were not actually surveys. Then we parse the retained surveys to extract their full texts and complete metadata, including authors, citation counts and publication year, and we further enrich the metadata with details about the first author information and the publishing venue. Based on the above parsing and filtering process, we retain 433 surveys that meet the required the criteria. Appendix A.1 details the parsing and filtering process.

4.3 Human Annotation

After completing the two stages described above, we further refine the dataset through human annotation to ensure its reliability and academic value. First, we prompt an LLM to preliminarily extract and summarize the academic value of 457 surveys from three dimensions: information value, scholarly communication value, and research guidance value, providing reference material for subsequent human annotation. Prior to the annotation, we invite two researchers with backgrounds in text generation and survey writing to establish a comprehensive academic value evaluation criteria based on the evaluation guidelines (Torraco, 2005; Denney and Tewksbury, 2013; Snyder, 2019; Xiao and Watson, 2019; Kraus et al., 2022) for scientific surveys in top-tier computer science venues. This criteria considers three core dimensions: information value, academic communication value, and research guidance value. Furthermore, to make the evaluation process more specific and feasible, the researchers further decompose the criteria into seven quantifiable academic value evaluation metrics (details in §5.2). We additionally invited three computer science graduate students in the filed of long text generation to evaluate the rationality of the academic value evaluation criteria. The results showed that they reached a high degree of consensus on the rationality of the dimension design and scoring criteria, which can be used for subsequent human and automated evaluation. Details of the rationality assessment are provided in Appendix A.2. Subsequently, these three graduate students form the annotation team and perform further annotation. During the annotation process, the annotators independently evaluate each survey using a unified scoring rubric (1-5), and label each survey as either "select" or "discard" based on its average score[4]. A survey is included in the dataset only if all three annotators independently mark it as "select". To assess annotation consistency, we compute Cohen's Kappa coefficient and obtain a value of 0.76, indicating substantial inter-annotator agreement. Ultimately, we construct a dataset comprising 163 rigorously validated and academically annotated high-quality scientific surveys. Detailed prompts and scoring rubric are provided in Appendix D.1.

4.4 Statistics

Our benchmark dataset contains 163 high-quality ground truth surveys, covering 8,715 outline sections and including 5,692,2 directly cited references, along with corresponding academic value

[4]Label "select" if average score ≥ 4, otherwise "discard".

Dataset	Domains	#Survey Nums	#Average Citations	#Metadata Type Nums	Academic value Annotation	Surface Quality Evaluation	Academic Value Evaluation
SurveyScope	CS	46	195.87	10	✗	✓	✗
SurveyBench	CS	100	220.17	7	✗	✓	✗
SurGE	CS	205	275.80	9	✗	✓	✗
DeepSurvey-Bench	Mixed	163	235.41	14	✓	✓	✓

Table 1: Comparison with other scientific survey datasets. CS denotes Computer Science, while Mixed indicates coverage spanning both Computer Science and other disciplines.

annotations for each survey. The dataset format is detailed in Appendix A.3. Table 1 compares DeepSurvey-Bench with existing survey generation datasets. Specifically, our benchmark covers four primary disciplines: Computer Science, Physics, Astronomy, and Electronics Systems, and includes 16 different topics, such as LLMs, Graph Representation Learning, Multimodal, and Agents, supporting systematic and interdisciplinary evaluation of automated survey generation methods. The distribution of different topics is shown in Figure 3.

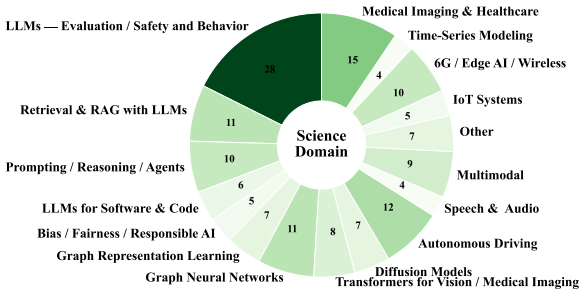

Figure 3: Distribution our benchmarks across different topics. Other disciplines have fewer surveys, so we categorized them as "Other".

5 Evaluation Metrics

To comprehensively evaluate the quality of automatically generated scientific surveys, we not only evaluate their surface quality but also their academic value. Specifically, Section 5.1 will elaborate on the evaluation metrics for surface quality, while Section 5.2 will focus on the evaluation metrics for academic value.

5.1 Surface Quality

Building on the prevailing three-stage process for generating surveys (Wang et al., 2024; Liang et al., 2025; Yan et al., 2025), we conducted a detailed evaluation of the surface quality of the generated survey from three perspectives:

Outline Quality: A detailed and well-structured survey outline lays a solid foundation for generating a high-quality survey. Drawing on previous work (Shao et al., 2024), we introduce a quantitative metric, **Heading Soft Recall (HSR)**, to assess the semantic similarity between automatically generated outline titles and ground truth outline titles. In addition, we introduce the LLM-as-a-judge paradigm to assess the outline's **Hierarchical Clarity**, **Logical Coherence**, and **Guidance for Content Generation** from a more macroscopic perspective that mimics human experts. Details of the metrics are provided in Appendix B.1. The prompt and scoring rubric (1-5) is shown in Appendix D.2.

Content Quality: We first employ the widely recognized automatic evaluation metrics **ROUGE** (Lin, 2004) and **BLEU** (Papineni et al., 2002) from the field of long-text generation to quantify the n-gram overlap between the generated survey and the ground truth survey. In addition, inspired by Autosurvey (Wang et al., 2024), we adopt the LLM-as-a-judge paradigm to comprehensively evaluate the true content quality of the survey across three dimensions: **Coverage**, **Structure**, **Relevance**, and **Fluency**. The prompt and scoring rubric (1-5) is shown in Appendix D.2.

Reference Quality: References are the foundation of a scientific survey, and their selection directly determines the content composition of the survey. For the evaluation of reference quality, we adopted the **Citation Recall** and **Citation Precision** metrics proposed by ALCE (Gao et al., 2023) and calculate the **F1** score. The detailed descriptions and calculation methods for metrics are provided in Appendix B.2.

5.2 Academic Value

A high-quality scientific survey is not only perfect in its surface quality, but more importantly, delivers substantive academic value to researchers. To make the evaluation of the survey's academic value

Backbones	Methods	Content Quality					
		ROUGE-L	BLEU	Coverage	Structure	Relevance	Language Fluency
GPT-4o	AutoSurvey	14.38	10.07	4.55	3.70	4.10	4.10
	SurveyX	13.74	12.37	4.35	3.05	3.95	3.60
	SurveyForge	14.75	10.43	4.85	3.95	4.55	4.00
Claude-3-5-haiku	AutoSurvey	12.98	8.36	3.95	3.60	3.90	3.95
	SurveyX	14.18	13.09	4.25	3.10	3.95	3.65
	SurveyForge	13.25	8.44	4.70	3.95	4.40	4.00
DeepSeek-v3	AutoSurvey	12.19	8.22	5.00	3.80	4.15	4.05
	SurveyX	13.62	12.13	4.30	3.20	3.90	3.65
	SurveyForge	13.10	8.76	4.85	3.70	4.45	4.00

Table 2: The surface quality evaluation results of the generated survey's content.

Backbones	Methods	Outline Quality				Reference Quality		
		HSR	Guidance	Hierarchical	Logical	CR	CP	F1
GPT-4o	AutoSurvey	63.11	4.00	3.90	3.95	46.71	40.78	43.54
	SurveyX	65.44	3.65	4.05	4.20	53.44	50.97	52.18
	SurveyForge	64.35	4.00	4.00	4.15	43.90	36.97	40.14
Claude-3-5-haiku	AutoSurvey	49.17	3.95	4.00	4.25	55.80	53.24	54.50
	SurveyX	66.26	3.70	4.00	4.05	54.36	52.10	53.21
	SurveyForge	59.63	4.00	4.00	4.00	66.01	62.83	64.38
DeepSeek-v3	AutoSurvey	75.27	4.05	3.35	3.35	30.95	24.23	27.18
	SurveyX	65.95	3.75	4.00	4.10	45.34	43.60	44.45
	SurveyForge	63.63	3.95	4.00	4.05	28.34	19.03	22.77

Table 3: The surface quality evaluation results of the generated survey's outline and reference. CR and CP denotes Citation Recall and Citation Precision, respectively.

more specific and feasible, we further decompose the three dimensions (§4.3) into seven quantifiable evaluation metrics. The prompts and scoring rubric (1-5) used for these seven academic metrics are all in the appendix D.2.

Information Value: We further decompose informational value into three key metrics: **Objective Clarity**, **Classification-Evolution Coherence**, and **Dataset & Metric Coverage**. The objective clarity evaluates the clarity of research objectives in scientific survey. The classification–evolution coherence assesses the clarity of method classification and the evolution of technological development in the survey. The dataset & metric coverage evaluates the coverage of datasets and the rationality of the evaluation metrics in the paper.

Scholarly Communication Value: We further decompose scholarly communication value into two key metrics: **In-depth Comparison** and **Critical Analysis**. The in-depth comparison assess whether the survey systematically compares multiple methods, clearly describing the advantages and disadvantages of each method. The critical analysis evaluates the critical analysis of different methods in the survey, focusing on explaining the underlying reasons for the differences between methods.

Research Guidance Value: We further decompose research guidance value into two key metrics: **Research Gaps** and **Future Work**. The research gaps evaluates whether a survey can systematically organize existing research findings and deeply identify and analyze key unresolved issues and shortcomings in the current research field. The future work evaluates whether a survey can propose forward-looking and innovative future research directions based on identified research gaps and practical needs.

6 Experiments and Analysis

6.1 Baselines

We used three different automated survey generation methods to generate surveys in order to compare the performance differences of different baselines for our proposed evaluation metrics. Detailed descriptions of these baselines are provided in Appendix C.1.

6.2 Implementation Details

We selected the relatively weaker model Claude-3-5-haiku , together with two stronger and comparable models, GPT-4o and DeepSeek-v3, as the backbone models for survey generation, in order

Backbones	Methods	Information				Scholarly com.			Research gui.		
		OC	CEC	DMC	Avg	IC	CA	Avg	RG	FW	Avg
GPT-4o	AutoSurvey	3.45	3.95	2.85	3.42	3.25	3.55	3.40	4.15	4.00	4.08
	SurveyX	3.85	3.25	2.80	3.30	2.70	3.00	2.85	3.70	3.85	3.78
	SurveyForge	3.40	3.85	3.05	3.43	3.45	3.90	3.68	3.80	4.00	3.90
Claude-3-5-haiku	AutoSurvey	2.80	3.80	2.50	3.03	2.60	3.15	2.88	3.50	3.75	3.63
	SurveyX	3.85	3.00	3.00	3.28	2.65	2.90	2.78	3.70	3.95	3.83
	SurveyForge	3.40	3.85	2.80	3.35	2.85	3.50	3.18	3.70	3.95	3.83
DeepSeek-v3	AutoSurvey	3.50	3.80	2.90	3.40	3.10	3.65	3.38	4.10	3.95	4.05
	SurveyX	3.80	3.20	2.85	3.28	2.60	3.05	2.83	3.60	3.80	3.70
	SurveyForge	3.50	3.69	3.10	3.43	3.35	3.75	3.55	3.85	4.05	3.95

Table 4: The academic value evaluation results of the generated survey (To keep the table concise and readable, we have standardized the names of the seven metrics to their initial letter abbreviations). Avg denotes the average value.

to effectively test the discriminative power of the evaluation metrics across models with different capability levels. With respect to survey generation configuration, all experiments adopted the default settings reported in the original papers for each model. For automated evaluation, we selected GPT-5.1 as the evaluator[5]. To improve evaluation stability, we perform three evaluation runs with different random seeds for each survey and use the mean score as the final result. Furthermore, to mitigate the impact of internal randomness in LLMs on the results (Bouras, 2024), we uniformly set the temperature parameter to 0 during the evaluation process to achieve deterministic decoding.

6.3 Experimental Results

This section presents the evaluation results of the three survey-generation baselines under different backbone models. Table 2 and Table 3 report the surface quality evaluation results of the generated surveys, and Table 4 presents the academic value evaluation results. Based on these results, we derive several notable observations:

Across backbone models with different capability levels, we observe that the three baselines generally perform well on surface-quality metrics, with only limited variation. For example, all baselines achieve ROUGE and BLEU scores within a similar range; the highest structure (3.95), logical (4.25), CR (66.01), and CP (62.83) scores appear under Claude-3.5-Haiku, the highest coverage (5.00) and guidance (4.05) score appear under DeepSeek-v3, while the highest relevance (4.55), language fluency (4.10) and hierarchical (4.05) scores are obtained under GPT-4o. Despite the substantial differences in their overall modeling capabilities, these models remain comparable in surface metrics, and in some cases even perform almost identically. In contrast, our academic value metrics reveal more discriminative differences. GPT-4o attains the highest average scores in all three dimensions: informational value (3.43), scholarly communication value (3.68), and research guidance value (4.08). Although Claude-3.5-Haiku achieves similar performance on surface quality metrics, its scores on these academic dimensions are substantially lower, with an average of only 2.78 in scholarly communication value. Compared with the relatively minor differences observed in surface quality, our academic value system can more clearly and more robustly distinguish the differences in the ability of different backbone models to generate surveys with academic quality.

Based on the data in Table 4 using GPT-4o as the backbone model, we compare the academic value of the surveys generated by the three baselines. The results show that SurveyX had the lowest average scores in information value, scholarly communication value, and research guidance value. Additionally, AutoSurvey, SurveyX, and SurveyForge score almost all below 4 points in seven academic metrics, while their surface quality scores in terms of outline and content—evaluated using LLMs as the standard—are almost all above 4 points. This indicates that high surface quality in surveys does not necessarily equate to high academic value. Existing automatically generated survey methods often produce surveys with perfect surface content but lack substantial academic contributions, highlighting the complexity of automatically generating scientific surveys. In addition, the generation time for each baseline and detailed API costs are shown

[5]The correlation between GPT-5.1 and human evaluation is verified in §6.4

in the appendix C.2.

6.4 Correlation Analysis between LLM-as-a-judge and Human Evaluation

Since the LLM-as-a-judge paradigm may introduce a potential preference bias toward AI-generated surveys, we randomly sample 20 surveys covering diverse research topics, generated by three LLM backbones and three different baselines. For each survey, three graduate students who had participated in the dataset construction process (§4.3) were asked to conduct a human evaluation using the same scoring rubric as the LLMs evaluation, in order to perform a correlation analysis. We have demonstrated the internal consistency among human annotators in Section 4.3.

Accordingly, we first computed the mean human rating for each metric, and then compute three widely-used inter-annotator agreement (IAA) metrics among humans and LLMs: percent agreement (Zheng et al., 2023), Spearman correlation (Lu et al., 2025) and Cohen's Kappa (Thakur et al., 2025). The results in Table 5 show that the LLM-based evaluator exhibits a strong and consistent alignment with human expert judgment across all three academic value dimensions. Furthermore, we constructed a pairwise comparison set in which each pair of reviews was generated by different baseline systems for the same topic. For each pair, human annotators selected the review they considered to possess higher academic value. We then selected the superior review according to the academic-value score from DeepSurvey-Bench and compared this selection with those obtained using surface-level metrics and randomly selected baseline systems. Detailed results are in Appendix C.3, demonstrating that our academic value metrics are highly consistent with human preferences.

Dimension	PA	Spearman ρ	Cohen κ
Informational	0.88	0.83	0.79
Scholarly Com.	0.85	0.81	0.76
Research Gui.	0.91	0.86	0.82

Table 5: Correlation results between LLM-as-a-judge and human evaluation. PA denotes Percent Agreement.

6.5 Weak Academic Value Capabilities Analysis

To further analyze the major weaknesses of automatically generated surveys in terms of academic value, we randomly sample 20 surveys from the bottom 20% of academic value scores and categorize their deficiency types along the three dimensions of informational value, scholarly communication value, and research guidance value. Meanwhile, we manually verify and conduct attribution analysis on the statistics by referring to the reasoning and evidence provided by the LLM during evaluation. As shown in Figure 4, half of the samples exhibit issues concentrated in the scholarly communication dimension, while another portion is focused on the objective metrics in the informational value dimension. This indicates that although most generated surveys perform well in terms of content organization and information presentation, they still fall short in areas such as in-depth comparison, critical analysis, and systematic discourse, highlighting significant room for improvement in the deeper analytical and integrative capabilities of current automated survey generation methods.

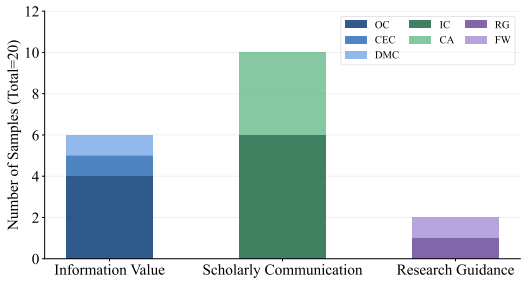

Figure 4: Distribution of deficiencies in academic value among 20 low-quality surveys.

7 Conclusion and Future Work

In this paper, We propose DeepSurvey-Bench, a novel benchmark for comprehensively evaluating "academic value" of automatically generated scientific survey. DeepSurvey-Bench constructs a more reliable survey dataset with academic value annotations based on a comprehensive quality criteria, and it not only evaluates the "surface quality" of generated surveys, but also focuses on assessing their deeper academic value across three dimensions: information value, scholarly communication value, and research guidance value. Extensive experimental results demonstrate the comprehensiveness and reliability of our proposed benchmark, while also highlighting the complexity of the task of automatically generating scientific surveys. In future work, we will explore methods for enhancing the academic value of generated surveys in automated survey generation and conduct more fine-grained evaluations of the academic value in survey.

Limitations

Closed-loop Verification of Our Benchmark A key limitation of our benchmark lies in the absence of a closed-loop validation process that links benchmark-based evaluation to subsequent model improvement. Although we have verified the reliability of the benchmark, including its inter-rater agreement and internal consistency, and demonstrated its effectiveness in assessing academic value through comparison with human-annotated reference scores, we have not yet established a complete "evaluation–optimization–reassessment" pipeline. The future work will address this by designing academic-value-aware generation strategies aligned with the benchmark's diagnostic insights, then re-evaluating optimized models to validate the benchmark's end-to-end guidance capability. For example, the proposed academic dimensions and metrics have not yet been applied as a reward mechanism in model training, and their ability to guide models toward generating higher-quality surveys remains to be explored. In future work, we plan to fine-tune a lightweight model on this benchmark to further examine its effectiveness in supporting model optimization. Therefore, in future work, we plan to fine-tune a lightweight model based on this benchmark in order to further verify its effectiveness in guiding model optimization.

Cost of LLM-based Evaluation Another major limitation of our benchmark is the high API cost of LLM-based review generation. The high cost of generating reviews objectively limits the scale of the evaluation, thus limiting the number of review generation methods, evaluation models, and research topics considered. Therefore, our experiments focused only on three representative review generation pipelines and three LLM evaluators, covering 20 topics, without expanding to a wider range of model architectures or performance levels.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Tong Bao, Mir Tafseer Nayeem, Davood Rafiei, and Chengzhi Zhang. 2025. Surveygen: Quality-aware scientific survey generation with large language models. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2712–2736.

Andrew Bouras. 2024. Integrating randomness in large language models: A linear congruential generator approach for generating clinically relevant content. *arXiv preprint arXiv:2504.05732*.

Chia-Yuan Chang, Zhimeng Jiang, Vineeth Rakesh, Menghai Pan, Chin-Chia Michael Yeh, Guanchu Wang, Mingzhi Hu, Zhichao Xu, Yan Zheng, Mahashweta Das, and 1 others. 2025. Main-rag: Multi-agent filtering retrieval-augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2607–2622.

Xiuying Chen, Hind Alamro, Mingzhe Li, Shen Gao, Xiangliang Zhang, Dongyan Zhao, and Rui Yan. 2021. Capturing relations between scientific papers: An abstractive model for related work section generation. In *Association for Computational Linguistics (ACL)*.

Andrew S Denney and Richard Tewksbury. 2013. How to write a literature review. *Journal of criminal justice education*, 24(2):218–234.

Shai Erera, Michal Shmueli-Scheuer, Guy Feigenblat, Ora Peled Nakash, Odellia Boni, Haggai Roitman, Doron Cohen, Bar Weiner, Yosi Mass, Or Rivlin, and 1 others. 2019. A summarization system for scientific documents. *EMNLP-IJCNLP 2019*, page 211.

Wenqi Fan, Yujuan Ding, Liangbo Ning, Shijie Wang, Hengyun Li, Dawei Yin, Tat-Seng Chua, and Qing Li. 2024. A survey on rag meeting llms: Towards retrieval-augmented large language models. In *Proceedings of the 30th ACM SIGKDD conference on knowledge discovery and data mining*, pages 6491–6501.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 6465–6488. Association for Computational Linguistics.

Gregory Hok Tjoan Go, Khang Ly, Anders Søgaard, Amin Tabatabaei, Maarten de Rijke, and Xinyi Chen. 2025. Lira: A multi-agent framework for reliable and readable literature review generation. *arXiv preprint arXiv:2510.05138*.

Thien Huynh-The, Quoc-Viet Pham, Xuan-Qui Pham, Thanh Thi Nguyen, Zhu Han, and Dong-Seong Kim. 2023. Artificial intelligence for the metaverse: A survey. *Engineering Applications of Artificial Intelligence*, 117:105581.

Xiao-Jian Jiang, Xian-Ling Mao, Bo-Si Feng, Xiaochi Wei, Bin-Bin Bian, and Heyan Huang. 2019. Hsds: An abstractive model for automatic survey generation. In *International conference on database systems for advanced applications*, pages 70–86. Springer.

Sascha Kraus, Matthias Breier, Weng Marc Lim, Marina Dabić, Satish Kumar, Dominik Kanbach, Debmalya Mukherjee, Vincenzo Corvello, Juan Piñeiro-Chousa, Eric Liguori, and 1 others. 2022. Literature reviews as independent studies: guidelines for academic practice. *Review of managerial science*, 16(8):2577–2595.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, and 1 others. 2025. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Yi-Fan Lu, Xian-Ling Mao, Tian Lan, Tong Zhang, Yu-Shi Zhu, and He-Yan Huang. 2025. Seoe: A scalable and reliable semantic evaluation framework for open domain event detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7201–7218.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Yijia Shao, Yucheng Jiang, Theodore Kanell, Peter Xu, Omar Khattab, and Monica Lam. 2024. Assisting in writing wikipedia-like articles from scratch with large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6252–6278.

Xiaofeng Shi, Qian Kou, Yuduo Li, Ning Tang, Jinxin Xie, Longbin Yu, Songjing Wang, and Hua Zhou. 2025. Scisage: A multi-agent framework for high-quality scientific survey generation. *arXiv preprint arXiv:2506.12689*.

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

Weihang Su, Anzhe Xie, Qingyao Ai, Jianming Long, Jiaxin Mao, Ziyi Ye, and Yiqun Liu. 2025. Surge: A benchmark and evaluation framework for scientific survey generation. *arXiv preprint arXiv:2508.15658*.

Qwen Team and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3).

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2025. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. In *Proceedings of the Fourth Workshop on Generation, Evaluation and Metrics (GEM^2)*, pages 404–430.

Richard J Torraco. 2005. Writing integrative literature reviews: Guidelines and examples. *Human resource development review*, 4(3):356–367.

Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, and 1 others. 2023. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

Zhiyuan Wen, Jiannong Cao, Zian Wang, Beichen Guo, Ruosong Yang, and Shuaiqi Liu. 2025. Interactivesurvey: An llm-based personalized and interactive survey paper generation system. *arXiv preprint arXiv:2504.08762*.

Yu Xiao and Maria Watson. 2019. Guidance on conducting a systematic literature review. *Journal of planning education and research*, 39(1):93–112.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Lei Bai, and Bo Zhang. 2025. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12444–12465.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Dataset Construction

A.1 Parsing and Filtering

In the parsing and filtering stage, we first remove papers that cannot be fully parsed, retaining 468 survey papers that are successfully parsed. Then, we use an LLM to evaluate the abstracts of each paper based on the following three criteria: (1) whether the purpose of the review is clearly stated (e.g., using phrases like "survey research" or "provide a review"); (2) whether the focus is on the nature of the survey rather than proposing new

methods or reporting experimental results; and (3) whether it discusses trends, challenges, or future directions in a specific field. Based on the LLM judgment, 433 out of the 468 papers are identified as genuine surveys. In parallel, we also extract the references cited in these surveys along with their associated metadata, which serve as the basis for subsequent evaluation of citation quality in generated surveys.

A.2 Human Annotation

We provide the three graduate students with a complete academic value evaluation criteria, which includes the descriptions of the three academic value dimensions (Information Value, Scholarly Communication Value, and Research Guidance Value), as well as a scoring rubric (1-5) for seven specific metrics (Objective Clarity, Classification-Evolution Coherence, Dataset & Metric Coverage, In-depth Comparison, Critical Analysis, Research Gaps, and Future Work). Subsequently, they evaluate the academic value evaluation criteria at both the dimension level and the metric level.

For each of the three dimensions, the three graduate students provide 1-5 ratings from three perspectives: *Coverage*, whether the dimension captures the core meaning of the corresponding scholarly value, *Necessity*, whether it is necessary to define it as an independent dimension, and *Clarity*, whether it is clearly distinguishable from the other dimensions. The results are shown in Table 6. All three dimensions score significantly above 4.0 in terms of coverage, necessity, and clarity, indicating that our proposed dimensional structure is conceptually reasonable and well-defined.

Dimension	*Coverage*	*Necessity*	*Clarity*
Information	4.6 ± 0.3	4.1 ± 0.2	4.5 ± 0.3
Scholarly	4.4 ± 0.4	4.6 ± 0.2	4.3 ± 0.4
Research	4.7 ± 0.2	4.3 ± 0.2	4.3 ± 0.3

Table 6: Results of the rationality evaluation of the three academic value dimensions (1–5 Likert scale: 1 = Very Poor, 5 = Very Good).

For the seven specific scoring criteria, the three graduate students scored them from 1 to 5 points based on four aspects: *Definition Clarity*, *Operability*, *Non-redundancy*, and *Importance*. The results are shown in Table 7. All metrics achieve scores above 4.0 across aspects, suggesting that the metrics are clearly defined, practically applicable, and very reasonable.

Indicator	Definition Clarity	Operability	Non-redundancy	Importance
OC	4.4 ± 0.3	4.6 ± 0.3	4.7 ± 0.2	4.1 ± 0.2
CEC	4.4 ± 0.4	4.2 ± 0.5	4.5 ± 0.4	4.5 ± 0.3
DMC	4.5 ± 0.3	4.4 ± 0.4	4.6 ± 0.3	4.6 ± 0.3
IC	4.6 ± 0.3	4.6 ± 0.3	4.7 ± 0.2	4.6 ± 0.2
CA	4.5 ± 0.4	4.4 ± 0.4	4.6 ± 0.3	4.6 ± 0.3
RG	4.7 ± 0.2	4.6 ± 0.3	4.8 ± 0.2	4.2 ± 0.2
FW	4.6 ± 0.3	4.5 ± 0.3	4.7 ± 0.2	4.7 ± 0.2

Table 7: Results of the rationality evaluation of the seven scoring rubric (1–5 Likert scale: 1 = Very Poor, 5 = Very Good).

A.3 Statistics

The dataset format is shown in Figure 5, and each field is briefly described in 8.

DeepSurvey-Bench Data Format

```
{
  "authors": [
    "string"
  ],
  "literature_review_title": "string",
  "year": "string",
  "date": "string",
  "category": "string",
  "abstract": "string",
  "structure": [
    {
      "section_title": "string",
      "level": "string",
      "content": "string",
      "origin_cites_number": "number"
    }
  ],
  "literature_review_id": "number",
  "meta_info": {
    "cite_counts": "number",
    "venue": "string",
    "influential_citation_count": "number",
    "Author_info": {
      "Publicationsh": "number",
      "h_index": "number",
      "Citations": "number",
      "Highly Influential Citations":
      "number"
    },
    "all_cites_title": [
      "string"
    ]
  }
}
```

Figure 5: The dataset format of DeepSurvey-Bench.

B Evaluate Metrics

B.1 Outline Quality

HSR evaluates how well the generated outline covers the specific headings present in the ground truth outline. Formally, HSR is defined as the soft cardinality overlap between the predicted heading set (H_P) and the ground truth heading set (H_{GT}):

$$\text{HSR} = \frac{\mathcal{S}(H_P \cap H_{GT})}{\mathcal{S}(H_{GT})}$$

where $\mathcal{S}(A)$ denotes the "soft cardinality" of a heading set A. Intuitively, this metric counts the number

Metadata	Description
authors	List of contributing researchers.
literature_review_title	The title of the survey
year	The publication year of the survey
date	The timestamp of publication
category	The classification or type of the survey
abstract	A brief summary of the survey's content, aims, and findings
literature_review_id	A unique identifier for the survey
cite_counts	The total number of citations for the survey
Conference_journal_name	The conference or journal where the survey was published
influential_citation_count	The number of citations that are considered highly influential
publication	The total number of publications by the first author
h_index	The h-index of the first author
citations	The total number of citations for the first author's work
all_cites_title	The titles of all the references cited in the survey

Table 8: Results of outline quality evaluation on the FreshWiki dataset.

of semantically unique headings in a set. Specifically, the contribution of each heading is inversely proportional to its aggregated similarity with all other headings in the set:

$$\mathcal{S}(A) = \sum_{i=1}^{K} \frac{1}{\sum_{j=1}^{K} \text{sim}(A_i, A_j)}$$

Here, $\text{sim}(A_i, A_j)$ is the cosine similarity between the embeddings of headings A_i and A_j. A standard set intersection would be too strict for comparing paraphrased headings. Therefore, we define the soft intersection cardinality using the inclusion-exclusion principle:

$$\mathcal{S}(H_P \cap H_{GT}) = \mathcal{S}(H_P) + \mathcal{S}(H_{GT}) - \mathcal{S}(H_P \cup H_{GT})$$

The complete prompt and scoring rubric (1-5) of Hierarchical Clarity, Logical Coherence, and Guidance for Content Generation is shown in Appendix D.2

B.2 Reference Quality

We define a set of claims extracted from the review as $C = \{c_1, c_2, ...\}$, and use a natural language inference (NLI) model h to determine whether a claim c_i is supported by its references $\text{Ref}_i = \{r_{i1}, r_{i2}, ...\}$ (where r_{ik} denotes a cited paper). $h(c_i, \text{Ref}_i) = 1$ means that the references can support the claim, and $h(c_i, \text{Ref}_i) = 0$ otherwise.

Citation Recall: Measures whether all statements in the generated text are fully supported by the cited passages, which is calculated as:

$$\text{Recall} = \frac{\sum_{i=1}^{|C|} h(c_i, \text{Ref}_i)}{|C|}$$

Citation Precision: Identifies irrelevant citations, ensuring that the provided citations are pertinent and directly support the statements. Before listing the formula for precision, a function g is defined as:

$$g(c_i, r_{ik}) = (h(c_i, \{r_{ik}\}) = 1) \cup (h(c_i, \text{Ref}_i \setminus \{r_{ik}\}) = 0) \quad (1)$$

$$\text{Precision} = \frac{\sum_{i=1}^{|C|} \sum_{k=1}^{|\text{Ref}_i|} h(c_i, \text{Ref}_i) \cap g(c_i, r_{ik})}{\sum_{i=1}^{|C|} |\text{Ref}_i|} \quad (2)$$

C Experiment and Analysis

C.1 Baselines

- AutoSurvey adopts a two-stage generation strategy: it first retrieves relevant literature to construct a detailed outline, then employs multiple LLMs in parallel to generate individual chapters, and finally integrates these chapters into a coherent survey article.

- SurveyForge first generates an outline by analyzing the logical structure of human-written outlines and referring to the retrieved domain-related articles. It then automatically generates and refines the article content using an academic navigation agent that retrieves high-quality papers from memory.

- SurveyX decomposes the survey writing process into two stages: a preparation stage and a generation stage. By incorporating online reference retrieval, a preprocessing method called AttributeTree, and a refinement procedure, SurveyX significantly improves the efficiency of survey writing.

C.2 Experimental Results

As shown in Figure 6, SurveyX incurs substantially higher latency than AutoSurvey and SurveyForge, with total runtime being approximately 9–10× longer, primarily due to the dominance of the text generation stage. In contrast, AutoSurvey and SurveyForge exhibit consistently low latency, as database retrieval introduces negligible overhead and text generation is only slightly slower than outline formulation. In terms of token cost, text generation constitutes the main expense across all pipelines, while overall cost differences across LLMs align with their pricing tiers (GPT-4o > Claude-3.5-haiku > DeepSeek-v3). Notably, SurveyForge achieves the best cost–efficiency.

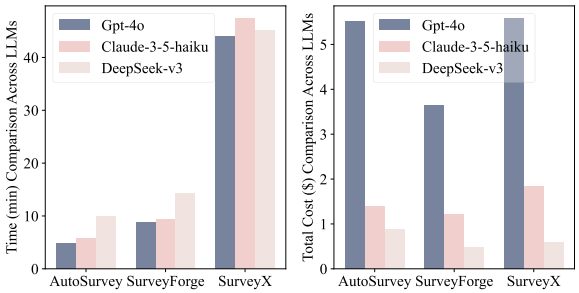

Figure 6: Statistics of defective chapters and sentences in four randomly selected articles

C.3 Correlation Analysis between LLM-as-a-judge and Human Evaluation

D Prompts and Scoring Rubric

D.1 Human Annotation Prompt

The prompt provided to annotators and the scoring criteria are shown in Figure 10. Since the dimensions and standards used are consistent with Appendix D.2, we only provide one example for reference.

D.2 Quality Evaluation Prompt

Human Annotation: The Objective Clarity of Information Value

Role Description:
We provide you with (1) an excerpt containing the academic-value related content extracted from a real scientific survey, and (2) the full survey paper in PDF form for reference. Based on these two sources, and strictly following our provided scoring rubric (1–5), please evaluate the academic value of this survey and assign a final score.

Evaluation Dimensions:
- **Research Objective Clarity**: Evaluate whether the research objective is specific, clear, and closely aligned with the core issues in the field.
- **Background and Motivation**: Evaluate whether the background and motivation are sufficiently explained, especially how they support the research objective.
- **Practical Significance and Guidance Value**: Evaluate whether the research objective demonstrates clear academic value and practical guidance for the field.

Scoring Criteria:

```
{
    - **5 points**: The research objective is clear, specific, and the background and motivation
    are clearly explained. The objective has significant academic and practical value. For
    example, the objective is closely tied to the core issues of the field, and the review
    provides a thorough analysis of the current state and challenges.

    - **4 points**: The research objective is clear, but the background or motivation may be
    somewhat brief. The objective has noticeable academic or practical value, and it mostly guides
    the research direction clearly.

    - **3 points**: The research objective is present, but the background and motivation lack
    depth. The academic and practical value of the objective is not fully explained, and the
    objective is somewhat vague or lacks clear direction.

    - **2 points**: The objective is unclear, and the background and motivation are inadequately
    explained. The academic value of the objective is not clear, and some parts of the objective
    are vague or repetitive.

    - **1 point**: The review does not present a clear research objective, lacks background and
    motivation, and the research direction is not specified. The objective is almost nonexistent.
}
```

Output Requirements:
- A final academic-value score (1–5) and a brief justification explaining the key reasons for your score.

Figure 7: A prompt and score rubric of Human Annotation

Guidance for Content Generation

Input Standard:
Here is an academic survey outline about the topic [topic]:
–
[content]
–

Evaluation Description:
Criteria Description Guidance for Content Generation: Does the outline effectively guide content generation, ensuring comprehensive coverage of the topic?

Scoring Criteria:

```
{
    Score 1 Description The outline fails to guide content generation, omitting significant
    aspects of the topic or providing insufficient direction.

    Score 2 Description The outline provides limited guidance, covering some key areas but lacking
    depth or completeness in addressing the topic.

    Score 3 Description The outline provides moderate guidance for content generation, addressing
    most key areas but leaving some gaps or ambiguities
    .
    Score 4 Description The outline effectively guides content generation, covering all
    significant aspects with clear direction, though minor refinements could enhance
    comprehensiveness.

    Score 5 Description The outline is exemplary in guiding content generation, thoroughly
    addressing all aspects of the topic with clear, detailed direction and no significant gaps.
}
```

Output Requirements:
Return the score without any other information.

Figure 8: The prompt and score rubric of Guidance metric

Hierarchical Clarity

Input Standard:
Here is an academic survey outline about the topic [topic]:
–
[content]
–

Evaluation Description:
Criteria Description Hierarchical Clarity: Does the outline clearly define a hierarchy of topics and subtopics, with a logical, diverse structure that is easy to understand?

Scoring Criteria:

```
{
    Score 1 Description The outline exhibits no discernible hierarchical structure. Topics and
    subtopics are jumbled together without logical separation or clear levels, making it nearly
    impossible to follow or identify any organization.

    Score 2 Description The outline attempts to establish a hierarchy but fails to maintain
    logical consistency. Main topics and subtopics are frequently misclassified, and the structure
    is overly rigid or disjointed. Subtopics may be missing, misplaced, or redundant, making it
    hard to grasp the intent of the structure.

    Score 3 Description The outline has a recognizable hierarchical structure but lacks diversity
    in organization style. While main topics are somewhat clear, subtopics occasionally overlap,
    are misaligned, or follow a repetitive format. This restricts flexibility and introduces mild
    confusion in certain areas.

    Score 4 Description The outline displays a clear, logical, and diverse hierarchical structure.
    Main topics are distinct, and subtopics are properly nested. While most elements are
    well-placed, there may be minor redundancies or opportunities to introduce more diverse
    formats for subtopics. Slight adjustments could achieve better precision and variety in style.

    Score 5 Description The outline showcases an exceptional, flawless hierarchical structure.
    Each main topic is distinct, and subtopics are logically nested with absolute clarity and
    stylistic diversity. The outline demonstrates flexibility in structure and organization,
    adapting its style where appropriate for the content and logic. No further refinement is
    necessary.

}
```

Output Requirements:
Return the score without any other information.

Figure 9: The prompt and score rubric of Hierarchical Clarity metric

Logical Coherence

Input Standard:

Here is an academic survey outline about the topic [topic]:

–

[content]

–

Evaluation Description:

Criteria Description Logical Coherence: Does the outline logically organize topics and subtopics, ensuring a smooth and natural flow of ideas with clear logical transitions?

Scoring Criteria:

```
{
    Score 1 Description The outline is highly disjointed and incoherent. Topics and subtopics
    appear in a random, unordered manner, with no logical flow or sense of progression. Major
    conceptual gaps and illogical jumps are present throughout the structure.

    Score 2 Description The outline shows some attempt at logical organization, but it contains
    frequent inconsistencies, abrupt shifts, or logical missteps. Topics and subtopics are
    misaligned or lack proper transitions, making the reader work hard to follow the structure.

    Score 3 Description The outline demonstrates a basic level of logical coherence. Most topics
    follow a general sequence, but some sections feel forced, with weak or unclear transitions.
    There are small jumps in logic, causing slight confusion or loss of flow at certain points.

    Score 4 Description The outline exhibits a strong sense of logical flow, with ideas presented
    in a mostly smooth and connected manner. Transitions between topics and subtopics are clear,
    but a few minor adjustments could make the flow more seamless or natural. The logic is sound,
    but room for refinement exists.

    Score 5 Description The outline achieves exceptional logical coherence. Each topic and
    subtopic follows a deliberate, thoughtful progression, with clear, natural, and intuitive
    transitions. The reader experiences a seamless flow of ideas, and no adjustments are required
    to improve logical consistency or flow.
}
```

Output Requirements:

Return the score without any other information.

Figure 10: The prompt and score rubric of Logical Coherence metric

Coverage

Task Description:
"coverage": Here is an academic survey about the topic [topic]:
—
[content]
—
<instruction>
Please evaluate this survey about the topic [topic] based on the criteria above provided below, and give a score from 1 to 5 according to the score description:
—

Evaluation Description:
Coverage: Coverage assesses the extent to which the survey encapsulates all relevant aspects of the topic, ensuring comprehensive discussion on both central and peripheral topics.

Scoring Criteria:

```
{
    Score 1 Description: The survey has very limited coverage, only touching on a small portion of
    the topic and lacking discussion on key areas.

    Score 2 Description: The survey covers some parts of the topic but has noticeable omissions,
    with significant areas either underrepresented or missing.

    Score 3 Description: The survey is generally comprehensive in coverage but still misses a few
    key points that are not fully discussed.

    Score 4 Description: The survey covers most key areas of the topic comprehensively, with only
    very minor topics left out.

    Score 5 Description: The survey comprehensively covers all key and peripheral topics,
    providing detailed discussions and extensive information.
}
```

Output Requirements:
Return the score without any other information.

Figure 11: The prompt and score rubric of Coverage metric

Structure

Task Description:
"structure": Here is an academic survey about the topic [topic]:
—
[content]
—
<instruction>
Please evaluate this survey about the topic [topic] based on the criteria above provided below, and give a score from 1 to 5 according to the score description:
—

Evaluation Description:
Structure: Structure evaluates the logical organization and coherence of sections and subsections, ensuring that they are logically connected.

Scoring Criteria:

```
{
    Score 1 Description: The survey lacks logic, with no clear connections between sections,
    making it difficult to understand the overall framework.

    Score 2 Description: The survey has weak logical flow with some content arranged in a
    disordered or unreasonable manner.

    Score 3 Description: The survey has a generally reasonable logical structure, with most
    content arranged orderly, though some links and transitions could be improved such as repeated
    subsections.

    Score 4 Description: The survey has good logical consistency, with content well arranged and
    natural transitions, only slightly rigid in a few parts.

    Score 5 Description: The survey is tightly structured and logically clear, with all sections
    and content arranged most reasonably, and transitions between adajecent sections smooth
    without redundancy.
}
```

Output Requirements:
Return the score without any other information.

Figure 12: The prompt and score rubric of Structure metric

Relevance

Task Description:
"relevance": Here is an academic survey about the topic [topic]:
—
[content]
—
<instruction>
Please evaluate this survey about the topic [topic] based on the criteria above provided below, and give a score from 1 to 5 according to the score description:
—

Evaluation Description:
Relevance: Relevance measures how well the content of the survey aligns with the research topic and maintain a clear focus.

Scoring Criteria:

```
{
Score 1 Description: The  content is outdated or unrelated to the field it purports to review,
offering no alignment with the topic.

Score 2 Description: The survey is somewhat on topic but with several digressions; the core
subject is evident but not consistently adhered to.

Score 3 Description: The survey is generally on topic, despite a few unrelated details.

Score 4 Description: The survey is mostly on topic and focused; the narrative has a consistent
relevance to the core subject with infrequent digressions.

Score 5 Description: The survey is exceptionally focused and entirely on topic; the article is
tightly centered on the subject, with every piece of information contributing to a comprehensive
understanding of the topic.
}
```

Output Requirements:
Return the score without any other information.

Figure 13: The prompt and score rubric of Relevance metric

Language

Task Description:
"language": Here is an academic survey about the topic [topic]:
—
[content]
—
<instruction>
Please evaluate this survey about the topic [topic] based on the criteria above provided below, and give a score from 1 to 5 according to the score description:
—

Evaluation Description:
Language: Language assesses the academic formality, clarity, and correctness of the writing, including grammar, terminology, and tone.

Scoring Criteria:

```
{
        Score 1 Description: The language is highly informal, contains frequent grammatical
        errors, imprecise terminology, and numerous colloquial expressions. The writing lacks
        academic tone and professionalism.

        Score 2 Description: The writing style is somewhat informal, with several grammatical
        errors or ambiguous expressions. Academic terminology is inconsistently used.

        Score 3 Description: The language is mostly formal and generally clear, with only
        occasional minor grammatical issues or slightly informal phrasing.

        Score 4 Description: The language is clear, formal, and mostly error-free, with only rare
        lapses in academic tone or minor imprecisions.

        Score 5 Description: The writing is exemplary in academic formality and clarity, using
        precise terminology throughout, flawless grammar, and a consistently scholarly tone.
}
```

Output Requirements:
Return the score without any other information.

Figure 14: The prompt and score rubric of Language metric

Information Value: Objective Clarity

Role Description:

You are now acting as an **experienced literature review evaluator** with years of academic review experience. You are proficient in evaluating the clarity of research objectives, the articulation of background and motivation, and the clarity of research direction in academic papers. You will evaluate the **Abstract** and **Introduction** sections of the paper in detail.

Evaluation Dimensions:

- **Research Objective Clarity**: Evaluate whether the research objective is specific, clear, and closely aligned with the core issues in the field. - **Background and Motivation**: Evaluate whether the background and motivation are sufficiently explained, especially how they support the research objective.
- **Practical Significance and Guidance Value**: Evaluate whether the research objective demonstrates clear academic value and practical guidance for the field.

Scoring Criteria:

```
{
    - **5 points**: The research objective is clear, specific, and the background and motivation
    are clearly explained. The objective has significant academic and practical value. For
    example, the objective is closely tied to the core issues of the field, and the review
    provides a thorough analysis of the current state and challenges.

    - **4 points**: The research objective is clear, but the background or motivation may be
    somewhat brief. The objective has noticeable academic or practical value, and it mostly guides
    the research direction clearly.

    - **3 points**: The research objective is present, but the background and motivation lack
    depth. The academic and practical value of the objective is not fully explained, and the
    objective is somewhat vague or lacks clear direction.

    - **2 points**: The objective is unclear, and the background and motivation are inadequately
    explained. The academic value of the objective is not clear, and some parts of the objective
    are vague or repetitive.

    - **1 point**: The review does not present a clear research objective, lacks background and
    motivation, and the research direction is not specified. The objective is almost nonexistent.
}
```

Output Requirements:

- Please **first provide the score** for this section (1-5 points).
- **Then provide a detailed explanation** of why you assigned this score, and specifically mention which parts of the paper (including chapters and sentences) support your scoring.
- Please ensure that the score is entirely consistent with the content, and make a reasonable judgment based on the actual content of the paper.

Figure 15: The prompt and score rubric of Objective Clarity metric

Information Value: Classification-Evolution Coherence

Role Description:

You are now acting as a **senior literature review evaluator** with many years of academic review experience. You are proficient in evaluating the clarity and reasonableness of the method classification system and the evolution of the technological progression in literature reviews. You will evaluate the **Method** and/or **Related Work** sections of the paper. If the paper does not explicitly use headings such as "Method" or "Related Work", focus on the content after the **Introduction** and before the **Experiments/Evaluation** sections for a detailed evaluation.

Evaluation Dimensions:

- **Method Classification Clarity**: Evaluate whether the method classification is clear and reasonable and whether it reflects the technological development path in the field.
- **Evolution of Methodology**: Evaluate whether the evolution process of methods is systematically presented and whether the technological or methodological trends are shown.

Scoring Criteria:

```
{
    - **5 points**: The method classification is completely clear, and the evolution process is
    systematically presented, well revealing the technological advancements and field development
    trends. Each category is clearly defined with inherent connections, and the evolutionary
    direction of methods is clear and innovative.

    - **4 points**: The method classification is relatively clear, and the evolution process is
    somewhat presented, but the connections between some methods are unclear, and some
    evolutionary stages are not fully explained. Overall, it reflects the technological
    development of the field.

    - **3 points**: The method classification is somewhat vague, and the evolution process is
    partially clear, but lacks a detailed analysis of the inheritance between methods. Some
    evolutionary directions are unclear.

    - **2 points**: The method classification is unclear, the evolution process is not
    well-defined, and there is no analysis of the relationships between methods. It is difficult
    to clearly present the technological progress of the field.

    - **1 point**: The method classification is chaotic, and the evolution process is almost
    unrecognizable. There is no mention of technological progress or the relationships between
    methods.
}
```

Output Requirements:

- Please **first provide the score** for this section (1-5 points).
- **Then provide a detailed explanation** of why you assigned this score, and specifically mention which parts of the paper (including chapters and sentences) support your scoring.
- Please ensure that the score is entirely consistent with the content, and make a reasonable judgment based on the actual content of the paper.

Figure 16: The prompt and score rubric of Classification-Evolution Coherence metric

Information Value: Dataset & Metric Coverage

Role Description:

You are now acting as a **senior literature review evaluator** with many years of academic review experience. You are proficient in evaluating the coverage of datasets and the applicability of evaluation metrics in the literature review. You will carefully evaluate the **Data**, **Evaluation**, and/or **Experiments** sections of the paper.

Evaluation Dimensions:

- **Diversity of Datasets and Metrics**: Evaluate whether the review covers a variety of datasets and evaluation metrics, and whether it includes important datasets and metrics in the field.
- **Rationality of Datasets and Metrics**: Evaluate whether the choice of datasets is reasonable and sufficiently supports the research objective, and whether the evaluation metrics are academically sound and practically meaningful.

Scoring Criteria:

```
{
    - **5 points**: The review comprehensively covers multiple datasets and evaluation metrics,
    providing detailed descriptions of each dataset's scale, application scenario, and labeling
    method. The choice and use of evaluation metrics are highly targeted and reasonable, covering
    the key dimensions of the field.

    - **4 points**: The review includes multiple datasets and evaluation metrics, and the
    description of each dataset is fairly detailed. The choice of evaluation metrics is generally
    reasonable, but some aspects of the dataset's application scenarios or the use of metrics may
    not be fully explained.

    - **3 points**: The review covers a limited set of datasets and evaluation metrics, and the
    descriptions lack detail. The choice of metrics does not fully reflect key dimensions of the
    field, and the characteristics of datasets are not sufficiently explained.

    - **2 points**: The review includes few datasets or evaluation metrics, and the descriptions
    are not clear or detailed. There is a lack of analysis of the rationale behind the choices,
    and some datasets or metrics are not described in detail.

    - **1 point**: No datasets or evaluation metrics are mentioned, or the descriptions are
    extremely simple and not practical.
}
```

Output Requirements:

- Please **first provide the score** (1-5 points) for this section.
- **Then provide a detailed explanation**, explaining why you assigned this score, and specifically mention which parts of the paper (including chapters and sentences) support your scoring.
- Please ensure that the score is entirely consistent with the content, and make a reasonable judgment based on the actual content of the paper.

Figure 17: The prompt and score rubric of Dataset & Metric Coverage metric

Scholarly communication value: In-depth Comparison

Role Description:

You are now acting as a **senior literature review evaluator** with many years of academic review experience. You are proficient in evaluating the comparison of different research methods and the analysis of their advantages, disadvantages, similarities, and distinctions in literature review papers. You will carefully evaluate the **Method** and/or **Related Work** sections of the paper. If the paper does not explicitly use "Method" or "Related Work" as section titles, focus on the content after the **Introduction** and before the **Experiments/Evaluation** sections for a detailed evaluation.

Evaluation Dimensions:

Evaluate the **clarity, rigor, and depth** of the review's **comparison of different research methods**. This evaluation focuses on whether the paper:

- systematically compares methods across multiple dimensions - clearly describes **advantages and disadvantages** - identifies **commonalities and distinctions** - explains differences in terms of **architecture, objectives, or assumptions** - avoids superficial or fragmented listing of methods

This dimension emphasizes **objective and structured comparison**, rather than subjective commentary.

Scoring Criteria:

```
{
    - **5 points**:
    The review presents a **systematic, well-structured, and detailed comparison** of multiple
    methods, clearly summarizing their advantages, disadvantages, commonalities, and distinctions
    across multiple meaningful dimensions (e.g., modeling perspective, data dependency, learning
    strategy, application scenario). The comparison is technically grounded and reflects a
    comprehensive understanding of the research landscape.

    - **4 points**:
    The review provides a **clear comparison** of major advantages and  disadvantages of the
    methods and identifies their similarities and differences, but some comparison dimensions are
    not fully elaborated, or certain aspects of the comparison remain at a relatively high level.

    - **3 points**:
    The review mentions the **pros and cons** or **differences** between methods, but the
    comparison is **partially fragmented or superficial**, lacking systematic structure or
    sufficient technical depth in contrasting the methods.

    - **2 points**:
    The review mainly **lists the characteristics or outcomes** of different methods, with limited
    explicit comparison. Advantages and disadvantages are mentioned in isolation, and **the
    relationships among methods are not clearly contrasted**.

    - **1 point**:
    The review does not provide meaningful comparison. Methods are described **independently**,
    with no discussion of similarities, differences, or advantages and disadvantages.
}
```

Output Requirements:

- Please **first provide the score** (1-5 points) for this section.
- **Then provide a detailed explanation**, explaining why you assigned this score, and specifically mention **which sections and sentences** in the paper support your scoring.
- Please ensure that the score is **entirely consistent with the content**, and make a reasonable judgment based on the actual content of the paper.

Figure 18: The prompt and score rubric of In-depth Comparison metric

Scholarly communication value: Critical Analysis

Role Description:

You are now acting as a **senior literature review evaluator** with many years of academic review experience. You are proficient in evaluating the **critical analysis, interpretation, and reflective commentary** provided in literature review papers. You will carefully evaluate the **Method** and/or **Related Work** sections of the paper. If the paper does not explicitly use "Method" or "Related Work" as section titles, focus on the content after the **Introduction** and before the **Experiments/Evaluation** sections for a detailed evaluation.

Evaluation Dimensions:

Evaluate the **depth, reasoning, and insightfulness** of the review's **critical analysis of different methods**. This evaluation focuses on whether the paper:

- explains the **fundamental causes** of differences between methods - analyzes **design trade-offs, assumptions, and limitations** - synthesizes relationships across research lines - provides **technically grounded explanatory commentary** - extends beyond descriptive summary to offer **interpretive insights**

This dimension emphasizes **analytical reasoning and reflective interpretation**, not merely reporting or summarization.

Scoring Criteria:

```
{
    - **5 points**:
    The review provides **deep, well-reasoned, and technically grounded critical analysis**,
    clearly explaining the **underlying mechanisms, design trade-offs, and fundamental causes** of
    methodological differences. It synthesizes connections across research directions and offers
    **insightful, evidence-based personal commentary** that meaningfully interprets the
    development trends and limitations of existing work.

    - **4 points**:
    The review offers **meaningful analytical interpretation** of method differences and provides
    reasonable explanations for some underlying causes, but the depth of analysis is **uneven
    across methods**, or some arguments remain partially underdeveloped.

    - **3 points**:
    The review includes **basic analytical comments** or evaluative statements, but the analysis
    remains **relatively shallow**, focusing more on descriptive remarks than on rigorous
    technical reasoning. Explanations of fundamental causes are **limited or implicit**.

    - **2 points**:
    The review provides only **brief or generic evaluative comments** without explaining why such
    differences or limitations arise. Arguments lack analytical depth and do not meaningfully
    interpret relationships between methods.

    - **1 point**:
    The review lacks **critical analysis**. The paper only presents methods descriptively, with no
    interpretive commentary, reasoning, or reflective insight.
}
```

Output Requirements:

- Please **first provide the score** (1-5 points) for this section.
- **Then provide a detailed explanation**, explaining why you assigned this score, and specifically mention **which sections and sentences** in the paper support your scoring.
- Please ensure that the score is **entirely consistent with the content**, and make a reasonable judgment based on the actual content of the paper.

Figure 19: The prompt and score rubric of Critical Analysis metric

Research guidance value: Research Gaps

Role Description:
You are now acting as a **senior literature review evaluator** with many years of academic review experience. You are proficient in evaluating the identification and analysis of research gaps in literature review papers. You will carefully evaluate the **Gap/Future Work** section of the paper.

Evaluation Dimensions:
Evaluate whether the review systematically identifies, analyzes, and explains the key issues and shortcomings that need to be addressed in the future of the research field, based on the current achievements. This evaluation focuses not only on whether the "unknowns" are pointed out but also on the **depth of analysis** regarding why these issues are important and what impact they have.

Scoring Criteria:

```
{
    - **5 points**:
    Based on the review content, the **major research gaps** are comprehensively identified and
    deeply analyzed, covering **data**, **methods**, and other dimensions. The analysis is
    detailed and discusses the **potential impact** of each gap on the development of the field.

    - **4 points**:
    The review points out several research gaps, but the analysis is somewhat **brief** and does
    not delve deeply into the **impact** or **background** of each gap. The gaps are identified in
    a comprehensive way, but the discussion is not fully developed.

    - **3 points**:
    The review lists some research gaps but lacks in-depth analysis or discussion. Although some
    gaps are identified, their **impact** or **reasons** are not fully explored.

    - **2 points**:
    The research gap analysis is **limited**, and the review does not fully discuss the identified
    gaps. The gaps are mentioned in passing but not explored in detail.

    - **1 point**:
    The review **does not identify or discuss any research gaps**, and there is a lack of analysis
    of the **major issues** in the field.
}
```

Output Requirements:
- Please **first provide the score** (1-5 points) for this section.
- **Then provide a detailed explanation**, explaining why you assigned this score, and specifically mention **which sections and sentences** in the paper support your scoring.
- Please ensure that the score is **entirely consistent with the content**, and make a reasonable judgment based on the actual content of the paper.

Figure 20: The prompt and score rubric of Research Gaps metric

Research guidance value: Future Work

Role Description:
You are now acting as a **senior literature review evaluator** with many years of academic review experience. You are proficient in evaluating the identification of future research directions and their innovative analysis in literature review papers. You will carefully evaluate the **Gap/Future Work** section of the paper.

Evaluation Dimensions:
Evaluate whether the paper proposes **forward-looking research directions** based on the **existing research gaps** or **real-world issues**, and whether it offers **new research topics** or **suggestions** in response to these gaps, aligning with real-world needs.

Scoring Criteria:

```
{
    - **5 points**:
    The review tightly integrates the **key issues** and **research gaps** in the field, proposing
    **highly innovative** research directions that effectively address **real-world needs**. The
    review presents **specific and innovative research topics or suggestions** and provides a
    thorough analysis of their **academic and practical impact**, offering a **clear and
    actionable path** for future research.

    - **4 points**:
    The review identifies several **forward-looking research directions** based on key issues and
    research gaps, addressing **real-world needs**, but the analysis of the **potential impact**
    and **innovation** is somewhat shallow. The directions are innovative, but the discussion is
    brief and does not fully explore the **causes** or **impacts** of the research gaps.

    - **3 points**:
    The proposed research directions are **broad** and lack an in-depth discussion of their
    **forward-looking nature**. The review does not clearly explain how these directions address
    the **existing research gaps** or meet **real-world needs**.

    - **2 points**:
    The future research directions are unclear and lack an in-depth analysis of their
    **innovation**. The proposed directions are **traditional**, and the review does not clearly
    explain their **academic significance** or **practical value**.

    - **1 point**:
    The review does not propose any **future research directions**, nor does it discuss the
    **forward-looking nature** of the research.
}
```

Output Requirements:
- Please **first provide the score** (1-5 points) for this section.
- **Then provide a detailed explanation**, explaining why you assigned this score, and specifically mention **which sections and sentences** in the paper support your scoring.
- Please ensure that the score is **entirely consistent with the content**, and make a reasonable judgment based on the actual content of the paper.

Figure 21: The prompt and score rubric of Future Work metric