SurveyForge: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing

Xiangchao Yan[♣,*] **Shiyang Feng**[♣,♠,*] **Jiakang Yuan**[♣,♠] **Renqiu Xia**[♣,♢]
Bin Wang[♠] **Lei Bai**[♣,†] **Bo Zhang**[♣,†]
[♣]Shanghai Artificial Intelligence Laboratory [♠]Fudan University
[♢] Shanghai Jiao Tong University

https://github.com/Alpha-Innovator/SurveyForge
https://huggingface.co/datasets/U4R/SurveyBench

Abstract

Survey paper plays a crucial role in scientific research, especially given the rapid growth of research publications. Recently, researchers have begun using LLMs to automate survey generation for better efficiency. However, the quality gap between LLM-generated surveys and those written by human remains significant, particularly in terms of outline quality and citation accuracy. To close these gaps, we introduce SurveyForge, which first generates the outline by analyzing the logical structure of human-written outlines and referring to the retrieved domain-related articles. Subsequently, leveraging high-quality papers retrieved from memory by our scholar navigation agent, SurveyForge can automatically generate and refine the content of the generated article. Moreover, to achieve a comprehensive evaluation, we construct SurveyBench, which includes 100 human-written survey papers for win-rate comparison and assesses AI-generated survey papers across three dimensions: reference, outline, and content quality. Experiments demonstrate that SurveyForge can outperform previous works such as AutoSurvey.

1 Introduction

With the rapid development of science and technology, the number of published research articles has been growing exponentially, particularly in fast-evolving fields like Artificial Intelligence (AI). The rapid growth of the literature makes it increasingly difficult for researchers to gain in-depth knowledge of a specific scientific field. Survey papers, which systematically integrate existing studies and provide comprehensive developments and trends in the specific domain, have become a vital starting point of the scientific research cycle. However, traditional human-driven survey writing requires researchers to review a vast number of articles which is time-consuming and makes it challenging to keep up-to-date with the latest advancements in the field.

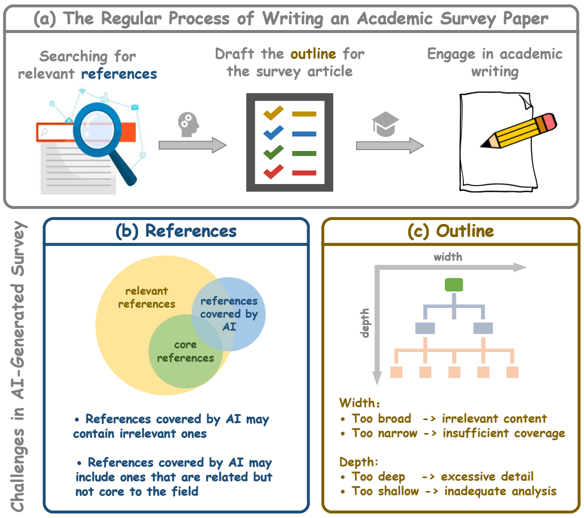

Figure 1: Compared to human-written surveys, AI-generated surveys face two primary challenges. First, regarding the outline, these papers may often lack coherent logic and well-structured organization. Second, with respect to references, they frequently fail to include truly relevant and influential literature.

Inspired by the remarkable advancement and capabilities of Large Language Models (LLMs) (Achiam et al., 2023; Anthropic, 2024; Touvron et al., 2023; Cai et al., 2024), researchers have begun utilizing them to automatically review the literature and generate survey papers. As a pioneer, GPT-Researcher (Assafelovic, 2023) generates survey papers based on the abstract of topic-relevant articles retrieved from multiple online academic databases. To identify more relevant literature to the survey topic, AutoSurvey (Wang et al., 2024c) constructs a local literature database based on arXiv, establishes vector indices for each literature, and concurrently generates content for each subsection. To further align the writing style of LLM-generated content with that of humans, OpenScholar (Asai et al., 2024) proposes a large-scale scientific literature dataset, and fine-tunes

[*]Core Contributor
[†]Corresponding Authors

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 12444–12465
July 27 - August 1, 2025

the LLMs based on this dataset to obtain a model specifically designed for answering scientific questions.

Most of these automated survey generation methods follow the traditional academic survey writing workflow: from literature search, to outline drafting, and finally academic writing, as illustrated in Fig. 1. However, despite the promising achievements of the aforementioned methods, several significant challenges still remain. **Firstly**, the structure of AI-generated surveys often lacks coherent logic and is often poorly-organized. For example, as shown in Fig. 1, existing works may suffer from structural imbalance in both width and depth, such as overly detailed sectioning or inadequate coverage of key topics. **Secondly**, AI-generated surveys often fail to reference key influential literature, reducing the overall depth and value of surveys. As shown in Fig. 1, they may cite irrelevant works while overlooking important contributions in the field. **Lastly**, the evaluation of AI-generated surveys mainly relies on LLMs, focusing on the overall quality of the long-form content. This approach lacks fine-grained analysis of critical aspects such as outline quality, reference relevance, and structural coherence. Moreover, the absence of objective evaluation criteria makes it difficult to establish consistent quality benchmarks or compare different methods effectively.

To address the aforementioned challenges, we propose an automated framework for generating survey papers, namely SURVEYFORGE which contains two stages: Outline Generation and Content Generation. In the first stage, SURVEYFORGE employs a heuristic learning approach to leverage topic-relevant literature and structural patterns from human-written surveys, generating semantically comprehensive and well-organized outlines. In the second stage, a memory-driven scholar navigation agent, with a temporal-aware reranking engine, retrieves high-quality literature for each subsection. Then, the content for each section is combined and refined into a coherent and comprehensive survey. Furthermore, we construct **SurveyBench**, a multi-dimensional benchmark to facilitate systematic assessment of automated survey generation systems.

Extensive results highlight the unique strengths of SURVEYFORGE across multiple dimensions, including its ability to generate well-structured outlines, retrieve high-quality and highly relevant references, and produce coherent, comprehensive content. SURVEYFORGE not only delivers measurable improvements in these areas but also demonstrates a remarkable ability to bridge the gap between AI-generated and human-written surveys. These findings underscore its potential as a robust framework for automated survey generation, setting a new standard for quality and reliability in this domain.

Our contribution can be summarized as follows.

- We propose SURVEYFORGE, a novel automated framework for generating high-quality academic survey papers.
- We propose a heuristic outline generation method and a memory-driven scholar navigation agent, which together ensure a well-structured survey framework and high-quality content generation.
- To facilitate objective evaluation, we establish SurveyBench, a comprehensive benchmark featuring quantifiable metrics for assessing outline quality, reference quality, and content quality.

2 Related Work

Autonomous Scientific Discovery. With the advancement of LLMs (Achiam et al., 2023; Anthropic, 2024; Chen et al., 2024a), an increasing number of researchers have begun exploring their potential for autonomous scientific discovery (Xia et al., 2024b; Li et al., 2024; Xia et al., 2024a; Huang et al., 2024; Ghafarollahi and Buehler, 2024; Chen et al., 2024b). Several studies (Li et al., 2024; Hu et al., 2024a; Kumar et al., 2024; Wang et al., 2024b; Su et al., 2024) have focused on leveraging LLMs for novel scientific idea generation. For instance, COI-Agent (Li et al., 2024) introduces an innovative chain-structured literature organization framework. SCIPIP (Wang et al., 2024b) proposes a hybrid approach combining literature-based and brainstorming-based generation to improve both the novelty and feasibility of the generated ideas. Beyond these specific applications, researchers have also developed comprehensive systems for scientific discovery. AI-Scientist (Lu et al., 2024) designs a comprehensive pipeline that covers idea generation, experimental design, and manuscript writing. More recently, Dolphin (Yuan et al., 2025) develops a closed-loop LLM-driven framework to boost the automation level of scientific research.

Automated Survey Generation. With the rapid proliferation of scientific papers, it has become increasingly challenging for researchers to track developments in specific fields. Early methods

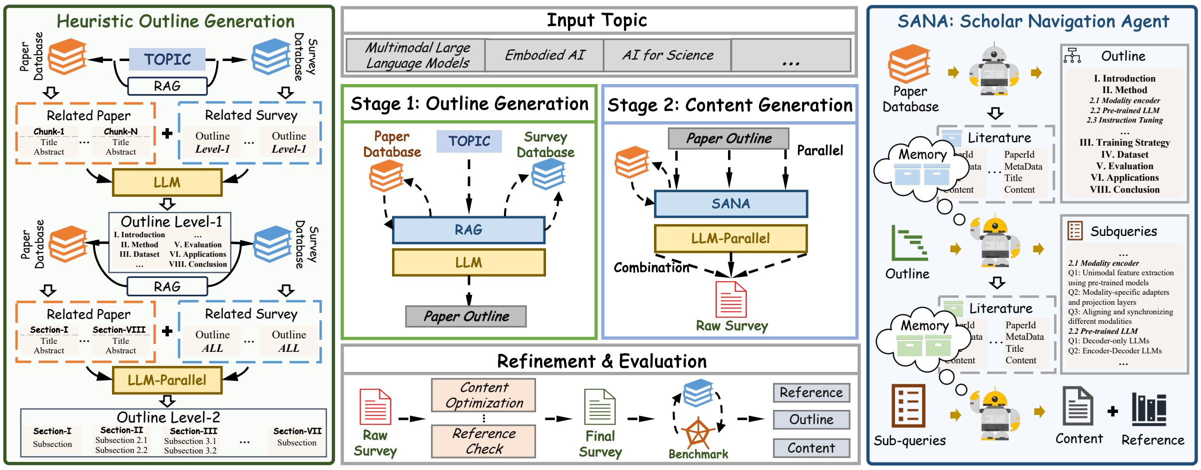

Figure 2: The overview of SURVEYFORGE. The framework consists of two main stages: Outline Generation and Content Writing. In the Outline Generation stage, SURVEYFORGE utilizes heuristic learning to generate well-structured outlines by leveraging topic-relevant literature and structural patterns from existing surveys. In the Content Writing stage, a memory-driven Scholar Navigation Agent (SANA) retrieves high-quality literature for each subsection and LLM generates the content of each subsection. Finally, the content is synthesized and refined into a coherent and comprehensive survey.

(Hoang and Kan, 2010; Hu and Wan, 2014; Jha et al., 2015; Chen and Zhuge, 2019) primarily rely on content models to select and organize sentences from papers, often resulting in outputs lacking coherence and readability. Sun et al. (Sun and Zhuge, 2019) introduce a template tree that generates content recursively based on nodes, which improves coherence but remains inflexible. Recognizing the need for more flexible and coherent solutions, the emergence of LLMs has introduced new opportunities for enhancing the automated survey generation. Researchers have begun to leverage LLMs to facilitate efficient literature comprehension and review (Wang et al., 2024c; Hu et al., 2024b). Zhu et al. (Zhu et al., 2023) introduce a novel task of hierarchical catalogue generation for surveys, along with corresponding semantic and structural metrics for evaluation, but it is limited to outline generation with fixed reference papers. AutoSurvey (Wang et al., 2024c) proposes a two-stage LLM-based method for survey generation but fails to focus on the analysis of human academic writing styles and key references, which are crucial for producing high-quality surveys. Subsequently, HiReview (Hu et al., 2024b) introduces a taxonomy-driven framework to explore paper relationships hierarchically, enhancing LLMs' understanding of inter-paper connections. However, relying on 2-hop citation networks from existing surveys instead of commonly-cited papers limits its broader applicability.

3 Method

In this section, we propose SURVEYFORGE, a novel framework based on LLMs for automatically retrieving relevant literature and generating comprehensive survey papers. As shown in Fig. 2, our framework consists of two main stages: outline generation stage and content writing stage. The outline generation stage leverages both research papers and existing survey structures through a heuristic learning mechanism, producing academically structured outlines. The content generation stage employs a memory-driven scholar navigation agent with key paper retrieval strategy to synthesize the content of the survey. Finally, we propose a benchmark **SurveyBench** for automated survey generation tasks. The details are elaborated in Sec. 3.1, Sec. 3.2 and Sec. 3.3, respectively.

3.1 Heuristic Outline Generation

The outline of a survey paper is crucial as it defines the logical organization and knowledge structure of the entire work. While LLMs excel at generating textual content, they often fall short in crafting well-structured survey outlines. Common issues include a lack of hierarchical depth, insufficient theoretical grounding, and a tendency toward report-like structures rather than scholarly frameworks. These limitations can be attributed to the limited understanding of academic writing conventions and the organizational principles underlying survey design. To address these challenges, we propose a top-down heuristic learning approach, enabling LLMs to understand the established theoretical frameworks and organizational paradigms from human-written survey outlines. Our approach is underpinned by two domain-specific knowledge bases: a Research Paper Database, which encodes

Algorithm 1: SURVEYFORGE

Input: Survey Topic T; Research Paper Database $\mathcal{D}_R$; Survey Outline Database $\mathcal{D}_S$
Output: Final Survey Document F

/ Outline Generation */*
Retrieve relevant papers and outlines for T: $\mathcal{P}_R$, $\mathcal{P}_S$;
Generate first-level outline $\mathcal{O}_i$ and queries $\{Q_i\}$;
foreach *first-level* $\mathcal{O}_i$ **do**
    Retrieve relevant papers and outlines for Q_i: $\mathcal{P}_{R_i}, \mathcal{P}_{S_i}$;
    Generate second-level outline $\mathcal{O}_{ij}$ and queries $\{q_{ij}\}$;
    Store $\mathcal{P}_{R_i}$ as memory M_i;
Store $\mathcal{P}_R$ as overall memory M;

/ Content Generation */*
foreach *subsection* O_{ij} ***in parallel*** **do**
    Decompose query q_{ij} into sub-queries $\{q_{ijk}\}$ using M_i;
    Initialize $L_{ij} \leftarrow \emptyset$;
    foreach *sub-query* q_{ijk} **do**
        Retrieve papers L_{ijk} using q_{ijk} and M;
        $L_{ij} \leftarrow L_{ij} \cup L_{ijk}$;
    Rerank and select top papers L_{ij}^{reranked};
    Generate content C_{ij} for O_{ij} using L_{ij}^{reranked};
Merge contents $\{C_{ij}\}$ to form draft F_{draft};
Refine F_{draft} to produce final document F;
return F;

domain knowledge, and a Survey Outline Database, which captures established structural patterns (details provided in Appendix. A.1). As shown in Algorithm 1, the framework begins with cross-database knowledge fusion, retrieving relevant papers and outlines for the given topic T from $\mathcal{D}_R$ and $\mathcal{D}_S$. This process identifies key thematic areas and their interrelations, generating the first-level outline $\mathcal{O}_i$ augmented with semantic queries Q_i that specify the scope and focus of each heading. For each section $\mathcal{O}_i$, we recursively retrieves relevant materials ($\mathcal{P}_{R_i}$, $\mathcal{P}_{S_i}$) and generates second-level outlines $\mathcal{O}_{ij}$ with sub-queries q_{ij}. Finally, these headings and their associated queries are systematically merged to construct a academically rigorous and comprehensive survey outline, serving as a foundation for subsequent content generation.

3.2 Memory-Driven Content Generation

The memory-driven content generation stage consists of two primary steps: literature retrieval and parallel content creation. These steps are performed sequentially by the proposed Scholar NAvigation Agent (SANA) and the LLM, respectively. A detailed explanation of each step is provided below.

3.2.1 SANA: Scholar Navigation Agent

To ensure that the quality and quantity of references in the generated survey papers, we propose a Scholar Navigation Agent (SANA), equipped with *memory* and *reranking* capabilities, designed to facilitate literature retrieval across various generation stages. The SANA includes three modules: Memory for Sub-query (MS), Memory for Retrieval (MR), Temporal-aware Reranking Engine (TRE).

Memory for Sub-query. Query decomposition is a common technique that involves breaking down a complex query into smaller sub-queries, thereby enabling more precise information retrieval. Existing query decomposition methods (Fan et al., 2024) are mostly achieved through naive prompts and LLMs. However, such methods require meticulous tuning of prompts to accommodate different tasks and may cause significant semantic differences between the decomposed sub-queries and the original query, which could potentially degrade the quality of the references in the AI-generated surveys. Therefore, we incorporate the memory mechanism into the query decomposition process of SANA to enhance the effectiveness of sub-queries. Specifically, as described in Sec. 3.1, when generating the first-level outline O_i, a set of literature $\mathcal{P}_{R_i}$ is retrieved by Retrieval-Augmented Generation (RAG). In the MS module, SANA takes the literature $\mathcal{P}_{R_i}$ as memory M_i, the original query consists of the titles $t_{O_{ij}}$ and descriptions $d_{O_{ij}}$ of each subsection:

$$q_{ij} = [d_{O_{ij}}, t_{O_{ij}}]. \tag{1}$$

To achieve query decomposition, q_{ij} and M_i are used together as part of the instruction to prompt the LLM to decompose q_{ij} into multiple sub-queries q_{ijk}:

$$q_{ijk} = \text{LLM}(q_{ij}, M_i). \tag{2}$$

Finally, the sub-query q_{ijk} is used in the subsequent MR module to retrieve literature related to the subsection O_{ij}.

Memory for Retrieval. The effectiveness of content generation heavily depends on the quality of retrieved information. Traditional retrieval methods (Lewis et al., 2020; Gao et al., 2023), which typically query the entire literature database $\mathcal{D}_R$, are often inefficient and lack contextual focus, particularly in generating complex, multi-section documents. These methods treat each section as an isolated unit, failing to account for the global structure and thematic coherence of the document. This results in redundant or irrelevant retrievals and limits the overall coherence of generated content.

To address these limitations, we incorporate the memory mechanism into the retrieval process of SANA to bridge the gap between the outline and content generation stages. Specifically, in the MS module, SANA takes the literature $\mathcal{P}_R$ related to the entire outline as memory M. Based on the embedding similarity between each sub-query q_{ijk} and the literature in M, the most relevant literature L_{ijk} for each sub-query of section O_{ij} is retrieved. Subsequently, the retrieved literature L_{ijk} is reranked and selected within the following TRE module for content generation.

Temporal-aware Reranking Engine. Reranking plays a important role in enhancing the quality and relevance of retrieved information. Existing methods (Glass et al., 2022; Xiao et al., 2023) typically employ advanced scoring mechanisms to measure textual relevance between queries and documents. However, these surface-level semantic matching may fall short in capturing the academic impact and quality of publications. Besides, The publication date of a paper plays a critical role in determining its influence and significance within its respective field. Consequently, analyzing papers from different time periods within the same research domain is a crucial for identifying high-quality contributions in the research field. For papers published within the same time period, there are various metrics to indicate their impact and quality, such as citation count, Essential Science Indicators (ESI), etc (Clarivate, 2024). Among these, citation count serves as a complementary quality indicator that reflects the scholarly influence and recognition of research works. To address both the limitations of pure semantic matching and the temporal bias in different quality indicators, we propose a temporal-aware reranking engine that integrates textual relevance, citation impact, and publication recency. This approach ensures not only the topical relevance but also the academic quality of the retrieved literature while maintaining a balanced representation of both established and emerging research. Specifically, the retrieved literature L_{ijk} based on embedding similarity is categorized into multiple groups $L_{ijk} = \{n_g\}_{g=1}^{G}$ according to their publication dates, with each group spanning a period of two years. For each group g, the highly cited literature is retained in a top-k manner as the final output for SANA, and the number of literature to be retained for each group is:

$$k_g = \frac{|n_g|}{|L_{ijk}|} K_{O_{ij}}, \tag{3}$$

where $K_{O_{ij}}$ is a hyper-parameter that represents the number of literature utilized for generating the content of each subsection.

3.2.2 Parallel Generation and Refinement

Due to the constraints of maximum context length and inference speed of LLMs, the content of each section is generated in parallel to reduce the generation time and ensure the length of the generated survey. However, due to the independent generation processes of each section in parallel, there may be repetition or redundancy among the contents of different section. Therefore, we employ LLMs to implement the refinement stage, which is aimed at refining the raw survey obtained by concatenating the contents of each section generated in parallel.

3.3 Multi-dimensional Evaluation Benchmark

Evaluating AI-generated surveys is challenging due to the lack of standardized benchmarks. Existing methods largely rely on automated scoring by LLMs, which face limitations: they may not adequately assess key literature coverage and depend heavily on internal model judgments without objective metrics. To address these challenges, we introduce **SurveyBench**, a comprehensive evaluation benchmark, along with SAM (Survey Assessment Metrics), a multi-dimensional evaluation series. SurveyBench consists of approximately 100 human-written survey papers across 10 distinct topics, carefully curated by doctoral-level researchers to ensure thematic consistency and academic rigor. For each topic t_i, we selected one highest-quality survey S_i^* as the reference for comparison with AI-generated surveys $\hat{S}_i$. Details of the benchmark construction process are provided in Appendix. A.2. The SAM series integrate objective metrics, expert knowledge, and multi-dimensional criteria through three core components:

SAM-R: Reference Quality Evaluation. A comprehensive and relevant bibliography is essential for a well-researched survey. Based on SurveyBench, we extract a reference set $\mathcal{R}_i$ for each topic t_i, serving as a reliable benchmark representing foundational knowledge in the field.

To measure reference quality, we define the SAM_R metric, which quantifies the overlap between the references in the AI-generated survey $\hat{S}_i$

and $\mathcal{R}_i$:

$$SAM_R(\hat{S}_i) = \frac{|R_{\hat{S}_i} \cap \mathcal{R}_i|}{|R_{\hat{S}_i}|}, \quad (4)$$

where $R_{\hat{S}_i}$ is the set of references in $\hat{S}_i$. A higher rate indicates better coverage of key literature in the topic t_i.

SAM-O: Outline Quality Evaluation. This component evaluates the structural quality of AI-generated surveys. A well-structured and logically coherent outline is crucial for content organization and readability. We assess the outline using a single comprehensive score SAM_O, ranging from 0 to 100, where higher scores indicate better quality. The evaluation is conducted by LLMs following detailed criteria described in Appendix. A.9.

SAM-C: Content Quality Evaluation. The final component measures the generated survey's quality across three dimensions: structure (SAM_C^{struct}), relevance (SAM_C^{rel}), and coverage (SAM_C^{cov}). Using the high-quality survey S_i^* as reference, we compute avg score of the overall content :

$$SAM_C^{\text{avg}} = \frac{SAM_C^{\text{struct}} + SAM_C^{\text{rel}} + SAM_C^{\text{cov}}}{3}. \quad (5)$$

Scores range from 0 to 100, with higher values indicating better performance. The LLMs assess these criteria while referencing S_i^* to ensure alignment with expert-level standards.

4 Experiment

4.1 Experimental Settings

Evaluation Dataset. To assess the performance of our proposed approach, we construct a dedicated benchmark dataset within the Computer Science (CS) domain, based on the *arXiv* repository. As mentioned in Sec. 3.3, we manually select approximately 100 human-written survey papers across 10 distinct topics, and choose one highest-quality survey for direct comparison with AI-generated surveys for each topic.

Implementation Details. To establish a baseline for comparison, we adopt AutoSurvey (Wang et al., 2024c), a state-of-the-art system for automated survey generation. Furthermore, we collect a large-scale dataset from the CS scientific field of *arXiv*, consisting of approximately 600,000 research papers and 20,000 review articles. We extract the key metadata to construct a retrieval vector database, including titles, abstracts of all papers and outlines of the review articles. To ensure a fair comparison, we align the timeline of our retrieval database with that of AutoSurvey. During the experimental evaluation, we retrieve 1,500 candidate papers for the outline generation stage and 60 relevant papers for each chapter-writing stage, following the same experimental settings as AutoSurvey.

For survey generation, we employ two LLMs independently: `Claude-3-haiku-20240307` and `GPT-4o-mini-2024-07-18`. Each model generates surveys for 10 predefined topics, with 10 independent trials conducted for each topic, resulting in a total of 100 outputs per model. The average performance across these trials is calculated to ensure stable and reliable results. In addition to the closed-source models, we have also experimented with the open source model with Deepseek-v3 (Liu et al., 2024), with impressive results, as detailed in Appendix A.5. For evaluation, we leverage more advanced models, `GPT-4o-2024-08-06` and `Claude-3.5-sonnet-20241022`, to assess both the AI-generated outlines and the content of the surveys, ensuring a robust and reliable evaluation of their quality.

4.2 Main Results

As shown in Table 1, we evaluate the performance of SURVEYFORGE across various dimensions, including reference quality, outline quality, and content quality, comparing it against the baseline AutoSurvey. The results demonstrate that SURVEYFORGE achieves significant improvements in all aspects, showcasing its potential as an advanced automated survey generation framework. Additionally, we conduct a cost analysis of the SURVEYFORGE framework, demonstrating that generating a 64k-token overview requires less than $0.50, with detailed cost breakdowns provided in Appendix A.6.

Results on Reference Quality. In terms of reference quality, SURVEYFORGE outperforms AutoSurvey on both key metrics: Input Coverage, which measures the relevance of retrieved papers, and Reference Coverage, which evaluates the alignment of the references of surveys with expert-curated benchmarks. Specifically, the Input Coverage score improves from 0.12 to 0.22 when using Claude-3-Haiku and from 0.07 to 0.20 with GPT-4o mini. Similarly, the Reference Coverage score increases from 0.23 to 0.40 and from 0.20 to 0.42 for the two respective models, indicating that SURVEYFORGE retrieves and generates references that are not only

Methods	Model	Reference Quality		Outline Quality	Content Quality			
		Input Cov.	Reference Cov.		Structure	Relevance	Coverage	Avg
Human-Written	-	-	0.6294	87.62	-	-	-	-
AutoSurvey	Claude-3-Haiku	0.1153	0.2341	82.18	72.83	76.44	72.35	73.87
SURVEYFORGE	Claude-3-Haiku	0.2231	0.3960	86.85	73.82	79.62	75.59	76.34
AutoSurvey	GPT-4o mini	0.0665	0.2035	83.10	74.66	74.16	76.33	75.05
SURVEYFORGE	GPT-4o mini	0.2018	0.4236	86.62	77.10	76.94	77.15	77.06

Table 1: Comparison of SURVEYFORGE and AutoSurvey (Wang et al., 2024c) using Survey Assessment Metrics (SAM) from three aspects: Reference (SAM-R), Outline (SAM-O) and Content quality (SAM-C). "Input Cov." means the coverage of input papers, measuring the overlap between retrieved papers and benchmark references, while "Reference Cov." means the coverage of reference, evaluating the alignment between cited references of the survey and benchmark references.

Methods	Outline Comparison			Content Comparison	
	Score Win Rate	Comparative Win Rate	Human Eval	Score Win Rate	Human Eval
AutoSurvey (Wang et al., 2024c)	27.00%	25.00%	26.00%	31.00%	30.00%
SURVEYFORGE	73.00%	75.00%	74.00%	69.00%	70.00%

Table 2: Win-rate comparison of automatic and human evaluations on outline and content quality. "Score Win Rate" reflects the win rate based on individual LLM-scores, where the LLM assigns separate score to each survey paper before determining the higher-scoring one. "Comparative Win Rate" is derived from LLM pairwise comparisons, where the LLM directly compares two articles side-by-side and decides which one is superior. "Human Eval" represents the win rate derived from expert human evaluations.

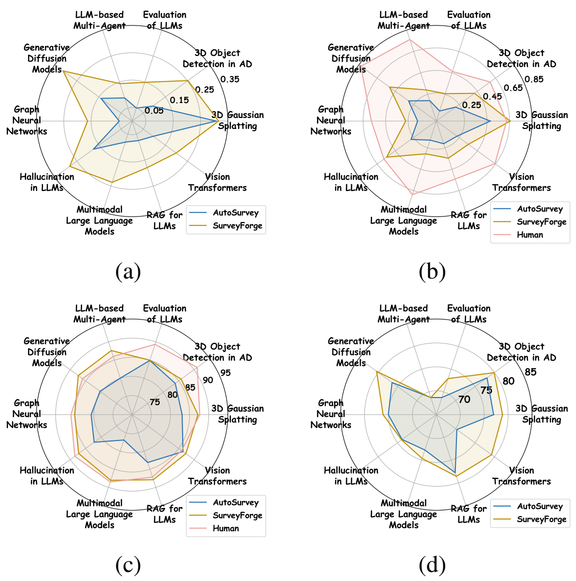

Figure 3: Evaluation results on SurveyBench. Evaluation results of (a) Input Coverage, (b) Reference Coverage, (c) Outline Quality, and (d) Content Quality.

more relevant but also more aligned with expert expectations. Notably, high-quality human-written surveys achieve a Reference Coverage score of 0.63, which further validates the reliability of our proposed reference evaluation database, which provides a robust benchmark for reference quality.

Results on Outline Quality. For outline quality, the results show that SURVEYFORGE generates outlines that are more logical, comprehensive, and closer to human-level performance compared to AutoSurvey (Wang et al., 2024c). Using Claude-3-Haiku, the outline quality score increases from 82.25 to 86.58, while GPT-4o mini achieves a similar improvement from 83.10 to 86.62. These advancements are driven by the proposed few-shot heuristic learning method, which leverages expert-curated examples from the Survey Outline Database to guide the LLMs in producing well-structured and domain-relevant outlines.

Method	Heuristic Learning	Demonstration Outline	Outline Quality
AutoSurvey	×	-	81.78
SURVEYFORGE	✓	From random surveys	84.58
SURVEYFORGE	✓	From related surveys	86.67

Table 3: Ablation study for outline generation. "Demonstration Outline" means the source of outlines used for heuristic learning.

Results on Content Quality. For content quality, SURVEYFORGE achieves consistent improvements across all three evaluation dimensions: structure, relevance, and coverage. The average content quality score increases from 73.87 to 76.34 (Claude-3-Haiku) and 75.05 to 77.06 (GPT-4o mini). These results confirm that SURVEYFORGE generates content that is better organized, more relevant, and more comprehensive, effectively addressing the critical aspects of the target domain.

As shown in Fig. 3, SURVEYFORGE demonstrates substantial improvements over the baseline AutoSurvey across all key evaluation metrics. Although not yet matching the quality of expert-crafted surveys, SURVEYFORGE significantly narrows the gap, highlighting its potential as a powerful tool for automated survey generation.

4.3 Comparison with Human Evaluation

To validate our automated evaluation system, we compare its performance with expert assess-

Components			Reference Quality	
MR	MS	TRE	Input Cov.	Reference Cov.
-	-	-	0.1119	0.2340
✓	-	-	0.1694	0.2730
✓	✓	-	0.1781	0.2984
✓	-	✓	0.1997	0.3542
✓	✓	✓	0.2224	0.3971

Table 4: Ablation study for content generation. We perform ablation on three components of SANA module: MR represents Memory for Retrieval, MS represents Memory for Sub-query, and TRE represents Temporal-aware Reranking Engine.

ments using 100 outputs from `Claude-3-haiku-20240307` across 10 topics (Please refer to Appendix A.2 and Appendix A.4 for detail information). We employ a win rate framework, presenting the anonymized results of SURVEYFORGE and AutoSurvey (Wang et al., 2024c) to 20 PhD experts in computer science field. These experts were carefully selected according to the evaluation topic and processes deep expertise in the relevant domain.

As shown in Table 2, for outline quality, the automated system achieves a Score Win Rate of 73.00% and a Comparative Win Rate of 75.00%, closely matching the human evaluation rate of 74.00%. This consistency confirms the system's robust scoring logic. For content quality, the automated system's Score Win Rate for SURVEYFORGE is 69.00%, aligning closely with the human expert rate of 70.00%. In addition, we also conduct Cohen's kappa coefficient consistency experiment, which shows a strong agreement between automated systems and human assessments, as detailed in Appendix A.4.

In summary, the automated system aligns well with human assessments for both outline and content quality, validating its effectiveness as a reliable alternative to manual evaluation.

4.4 Ablation Study

To better understand the contribution of individual components in our proposed SURVEYFORGE framework, we conduct a comprehensive ablation study. For ablation experiments, we use `Claude-3-haiku-20240307` to generate surveys on the same 10 topics, with 3 independent trials per topic to ensure statistical reliability while maintaining computational efficiency. Specifically, we analyze the memory mechanism, sub-query decomposition, and reranking strategies in the scholar navigation agent module, as well as the impact of the use of the database of survey outlines in the outline generation process. The results of the ablation experiments are presented in Table 3 and Table 4.

Analysis on Outline Generation. Table 3 highlights the impact of heuristic learning approach on outline quality. The baseline method, which generates outlines solely from retrieved research papers without structural guidance, achieves a score of 81.78. This indicates the absence of organizational cues limits the coherence and logical flow of the outlines. To address this, we first introduce a heuristic approach using outlines from random surveys. These generic outlines, representing common patterns in survey writing, improve the score to 84.58. This shows the effectiveness of structural cues, even without target-domain tailoring. Finally, we retrieve domain-specific outlines, providing both structural guidance and thematic alignment with the target domain. As a result, the outline quality score significantly rises to 86.67, showing the crucial role of domain-specific structural cues in creating coherent and relevant outlines.

Analysis on Content Generation. Based on the experimental results presented in Table 4, it can be observed that as the quality of literature obtained by SANA improves, the quality of cited references in surveys also correspondingly enhances. This observation highlights the importance of using SANA during the content generation stage to retrieve high-quality literature. Specifically, the integration of a memory mechanism into the query decomposition and retrieval processes significantly enhance the quality of literature. This improvement can be attributed to the incorporation of more comprehensive sub-query semantics and a retrieval scope better aligned with the sub-queries. Besides, the temporal-aware reranking engine ensures the selection of high-quality papers, leading to a more comprehensive and balanced reference collection.

5 Conclusion and Outlook

We have introduced SURVEYFORGE, an automated framework leveraging a heuristic outline generation and a memory-driven content generation to generate high-quality surveys. We introduce a multi-dimensional evaluation benchmark to comprehensively assess the quality of surveys. SURVEYFORGE significantly outperforms prior approaches across multiple evaluation metrics. We hope to reduce the learning curve for researchers venturing into unfamiliar fields, providing convenience and thereby promoting the integration and development of cross-disciplinary and cross-domain knowledge.

Limitations

Despite its strong performance in generating structured and high-quality surveys, SURVEYFORGE has inherent limitations, as discussed in Appendix A.3. While LLMs excel at summarizing existing literature, they face challenges in analyzing and synthesizing relationships across multiple sources, often lacking the critical thinking and originality characteristic of human-authored work, which limits their capability to reflect research trends or provide forward-looking insights. Besides, the accuracy of content and citations is also affected by the hallucination of LLMs. Future work could focus on developing methods to better capture interconnections among references to enhance the logical coherence, depth, and scholarly value of the generated content.

Ethics Statement

This work focuses on the development of an automated framework for survey generation, aiming to assist researchers in efficiently summarizing existing literature. The proposed method relies on publicly available datasets and research papers, ensuring compliance with copyright and intellectual property laws. While the framework is designed to augment human expertise, we encourage users to critically evaluate the generated outputs to ensure their alignment with ethical research practices and to mitigate any potential limitations, such as biases or incomplete summaries.

Acknowledgement

The research was supported by Shanghai Artificial Intelligence Laboratory, a locally commissioned task from the Shanghai Municipal Government, the Shanghai Municipal Science and Technology Major Project, and Shanghai Rising Star Program (Grant No. 23QD1401000).

References

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

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. URL: https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, et al. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *arXiv preprint arXiv:2411.14199*.

Assafelovic. 2023. gpt-researcher. URL: https://github.com/assafelovic/gpt-researcher.

Zheng Cai, Maosong Cao, Haojiong Chen, Kai Chen, Keyu Chen, Xin Chen, Xun Chen, Zehui Chen, Zhi Chen, Pei Chu, et al. 2024. Internlm2 technical report. *arXiv preprint arXiv:2403.17297*.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, 31(3):e4261.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024a. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Ziru Chen, Shijie Chen, Yuting Ning, Qianheng Zhang, Boshi Wang, Botao Yu, Yifei Li, Zeyi Liao, Chen Wei, Zitong Lu, et al. 2024b. Scienceagentbench: Toward rigorous assessment of language agents for data-driven scientific discovery. *arXiv preprint arXiv:2410.05080*.

Clarivate. 2024. Essential science indicators: Learn the basics. URL: https://clarivate.libguides.com/esi.

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.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, and Haofen Wang. 2023. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*.

Alireza Ghafarollahi and Markus J Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Cong Duy Vu Hoang and Min-Yen Kan. 2010. Towards automated related work summarization. In *Coling 2010: Posters*, pages 427–435.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024a. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Yue Hu and Xiaojun Wan. 2014. Automatic generation of related work sections in scientific papers: an optimization approach. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1624–1633.

Yuntong Hu, Zhuofeng Li, Zheng Zhang, Chen Ling, Raasikh Kanjiani, Boxin Zhao, and Liang Zhao. 2024b. Hireview: Hierarchical taxonomy-driven automatic literature review generation. *arXiv preprint arXiv:2410.03761*.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. 2024. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning*.

Rahul Jha, Catherine Finegan-Dollak, Ben King, Reed Coke, and Dragomir Radev. 2015. Content models for survey generation: a factoid-based evaluation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 441–450.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. 2024. Can large language models unlock novel scientific research ideas? *arXiv preprint arXiv:2409.06185*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xinxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, et al. 2024. Chain of ideas: Revolutionizing research in novel idea development with llm agents. *arXiv preprint arXiv:2410.13185*.

Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. 2023. Towards general text embeddings with multi-stage contrastive learning. *arXiv preprint arXiv:2308.03281*.

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

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The AI Scientist: Towards Fully Automated Open-Ended Scientific Discovery . *ArXiv*, abs/2408.06292.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:2410.09403*.

Xiaoping Sun and Hai Zhuge. 2019. Automatic generation of survey paper based on template tree. In *2019 15th International Conference on Semantics, Knowledge and Grids (SKG)*, pages 89–96. IEEE.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, et al. 2024a. Mineru: An open-source solution for precise document content extraction. *arXiv preprint arXiv:2409.18839*.

Wenxiao Wang, Lihui Gu, Liye Zhang, Yunxiang Luo, Yi Dai, Chen Shen, Liang Xie, Binbin Lin, Xiaofei He, and Jieping Ye. 2024b. Scipip: An llm-based scientific paper idea proposer. *arXiv preprint arXiv:2410.23166*.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, et al. 2024c. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*.

Renqiu Xia, Song Mao, Xiangchao Yan, Hongbin Zhou, Bo Zhang, Haoyang Peng, Jiahao Pi, Daocheng Fu, Wenjie Wu, Hancheng Ye, et al. 2024a. Docgenome: An open large-scale scientific document benchmark for training and testing multi-modal large language models. *arXiv preprint arXiv:2406.11633*.

Renqiu Xia, Bo Zhang, Hancheng Ye, Xiangchao Yan, Qi Liu, Hongbin Zhou, Zijun Chen, Min Dou, Botian Shi, Junchi Yan, et al. 2024b. Chartx & chartvlm: A versatile benchmark and foundation model for complicated chart reasoning. *arXiv preprint arXiv:2402.12185*.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Niklas Muennighoff. 2023. C-pack: Packaged resources to advance general chinese embedding. *Preprint*, arXiv:2309.07597.

Jiakang Yuan, Xiangchao Yan, Botian Shi, Tao Chen, Wanli Ouyang, Bo Zhang, Lei Bai, Yu Qiao, and Bowen Zhou. 2025. Dolphin: Closed-loop open-ended auto-research through thinking, practice, and feedback. *arXiv preprint arXiv:2501.03916*.

Kun Zhu, Xiaocheng Feng, Xiachong Feng, Yingsheng Wu, and Bing Qin. 2023. Hierarchical catalogue generation for literature review: A benchmark. *arXiv preprint arXiv:2304.03512*.

A Appendix

Due to the page limitation of the manuscript, we provide more details and visualizations from the following aspects:

- Sec. A.1: Database Construction.
- Sec. A.2: Details of SurveyBench.
- Sec. A.3: Discussion about Generated Surveys and Human-written Surveys.
- Sec. A.4: Details of Human Evaluation and Inter-rater Agreement.
- Sec. A.5: Additional Experiments with Open-Source Models.
- Sec. A.6: Details of Time and Economic Cost.
- Sec. A.7: Qualitative Results.
- Sec. A.8: Example of Generated Survey.
- Sec. A.9: Prompt Used.

A.1 Database Construction

To ensure the quality and relevance of the AI-generated surveys, we construct two key databases: the *Research Paper Database* and the *Survey Outline Database*, consisting of approximately 600,000 research papers and 20,000 review articles, which together serve as the foundation for content generation and structural guidance. The *Research Paper Database* comprises the titles and abstracts of research papers relevant to the survey topic, while the *Survey Outline Database* contains titles, abstracts, and outlines extracted from published survey papers.

Specifically, we utilize MinerU (Wang et al., 2024a) to extract content from a corpus of survey articles. Using rule-based extraction techniques, we isolate hierarchical outlines, including section and subsection headings. However, due to variations in formatting and structure across different papers, automatic extraction may introduce noise. To address this, we employ `Claude-3.5-sonnet-20241022` to refine and standardize the extracted outlines, ensuring consistency in structure and formatting. By leveraging the *Survey Outline Database* in this way, we provide the LLM with high-quality, expert-crafted outline examples to guide its generation process.

Additionally, we encode these documents using the `gte-large-en-v1.5` embedding model (Li et al., 2023), which captures semantic relationships and enables efficient similarity-based retrieval. This combination of structured expert examples and semantic encoding ensures a robust foundation for outline generation and content retrieval.

A.2 Details of SurveyBench

To construct SurveyBench, we select 10 trending topics in the computer science domain, as shown in Table 5. These topics span various cutting-edge areas including multimodal learning, language models, computer vision, and autonomous systems. For each topic, a set of high-quality, human-written surveys is carefully curated by *a panel of 20 researchers*. Each of these researchers holds doctoral degrees and possesses extensive expertise in the aforementioned 10 trending topics in the computer science domain. This rigorous selection process ensures strong thematic alignment and guarantees the inclusion of authoritative and relevant surveys. Besides, the development of our assessment metrics (e.g. SAM-O and SAM-C) is inspired by peer review guidelines from top-tier computer science venues. However, we observed that traditional review criteria often rely heavily on reviewers' implicit knowledge and experience, making them challenging to implement in automated evaluation systems. To address this limitation, we systematically decomposed these high-level review guidelines into more specific, measurable components that can be reliably assessed by LLMs while maintaining consistency with expert human evaluation. For example, in our outline assessment criteria, abstract concepts like "topic organization" were broken down into concrete, assessable elements such as "topic uniqueness" (checking for duplicate topics, content overlap) and "structural balance" (examining section development and proportionality). This granular approach, developed through discussions with researchers who have at least two years of reviewing experience for top CS venues, enables more consistent and reliable automated evaluation across different survey topics while preserving the essential quality standards of academic peer review.

The curated surveys, predominantly published within the last two years, are chosen to ensure both timeliness and relevance. From each selected survey, we extract the references cited to construct a dedicated reference database for each topic, resulting in comprehensive reference collections ranging

Topic	Ref Num	Selected Survey Title	Citation
Multimodal Large Language Models	912	A Survey on Multimodal Large Language Models	979
Evaluation of Large Language Models	714	A Survey on Evaluation of Large Language Models	1690
3D Object Detection in Autonomous Driving	441	3D Object Detection for Autonomous Driving: A Comprehensive Survey	172
Vision Transformers	563	A Survey of Visual Transformers	405
Hallucination in Large Language Models	500	Siren's Song in the AI Ocean: A Survey on Hallucination in Large Language Models	808
Generative Diffusion Models	994	A Survey on Generative Diffusion Models	367
3D Gaussian Splatting	330	A Survey on 3D Gaussian Splatting	128
LLM-based Multi-Agent	823	A Survey on Large Language Model Based Autonomous Agents	765
Graph Neural Networks	670	Graph Neural Networks: Taxonomy, Advances, and Trends	129
Retrieval-Augmented Generation for Large Language Models	608	Retrieval-Augmented Generation for Large Language Models: A Survey	953

Table 5: Overview of selected topics and the representative surveys in our evaluation benchmark. For each topic, we show the total number of unique references (Ref Num) collected from SurveyBench, and the citation count of selected high-quality surveys that serve as our evaluation references.

from 330 to 994 references per topic, as detailed in Table 5. Furthermore, to facilitate robust content evaluation,we identify the highest-quality survey for each topic to serve as the evaluation reference, with these selected surveys demonstrating significant impact through their citation counts (ranging from 128 to 1,690 citations). SurveyBench provides a comprehensive and reliable foundation for assessing the quality of AI-generated surveys, ensuring both reference coverage and content relevance are rigorously evaluated.

A.3 Discussion about Generated Surveys and Human-written Surveys

While our extensive evaluation of SURVEYFORGE demonstrates its effectiveness in automated survey generation, our analysis reveals several fundamental challenges that warrant further investigation. Through systematic examination of the generated surveys, we identify two primary limitations of the current system.

The first limitation lies in the depth of academic analysis. Although the system effectively extracts and organizes information from individual papers, it exhibits constraints in establishing profound connections across multiple publications. Specifically, the system's capability falls short in comparative analysis of temporal innovations and methodological evolution patterns, often defaulting to mechanical reference listing rather than providing the nuanced synthesis characteristic of expert-written surveys. This limitation stems primarily from challenges in the accurate identification of the core literature and the construction of deep logical relationships during the processing of long-form knowledge.

The second challenge concerns the accuracy of content and citation. Despite our implementation of multiple verification mechanisms, the system occasionally produces inaccurate citations or academic claims, potentially affecting the survey's reliability. This remains a critical area for improvement in automated survey generation systems.

Evaluation Pair	Aspect	κ
LLM vs. Human	Outline	0.7177
LLM vs. Human	Content	0.6462
Human Cross-Validation	Outline	0.7921
Human Cross-Validation	Content	0.7098

Table 6: Inter-rater agreement between LLM and human evaluations. κ means the Cohen's kappa coefficient.

To address these limitations, future work could focus on developing comprehensive knowledge association networks through core entity extraction and citation graph construction, which may enhance the system's capability to identify deep inter-publication connections.

A.4 Details of Human Evaluation and Inter-rater Agreement

For the human evaluation across the selected 10 topics, we recruited 20 PhD experts in computer science from various prestigious institutions, including several QS Top 50 universities and renowned research institutes within our country. The selection of these experts followed strict criteria to ensure their expertise and qualifications. All evaluators hold PhD degrees in computer science or closely related fields, and each expert has published at least one peer-reviewed paper in the specific topic they were assigned to evaluate. Moreover, all selected experts are currently active researchers in their respective fields.

To maintain evaluation quality and consistency, each expert was provided with a comprehensive evaluation guideline manual, identical to the one used in our LLM evaluation system, ensuring consistent assessment criteria across all evaluators. Before the formal evaluation, we conducted a training session to familiarize the experts with the evaluation criteria and scoring rubrics. The evaluation process was conducted in a double-blind manner to minimize potential biases. Regarding compensation, experts were paid $50 per hour, commensurate with their expertise level. The average evaluation time per survey was approximately 1-3 hours, ensuring thorough and reliable assessment.

To further verify the reliability of the evaluation system, we further conducted Cohen's kappa coefficient experiment to measure the inter-rater agreement between automatic and human evaluations and evaluations inter-rater agreement among human annotators. Specifically, as shown in Table 6, we conducted a systematic evaluation of 100 generated survey papers across 10 different research topics. We used Cohen's kappa coefficient as our evaluation metric, covering two core dimensions: outline and content.

In the outline dimension, based on the evaluation of these 100 surveys, the kappa coefficient between LLM evaluation and human evaluation reached 0.7177, indicating significant agreement between the two. Meanwhile, the cross-validation kappa coefficient between human evaluators was 0.7921. This high level of agreement not only validates the reliability of human evaluation but also supports the effectiveness of our automated evaluation method.

In the content dimension, based on the same sample size, the kappa coefficient between LLM evaluation and human evaluation was 0.6462, while the cross-validation kappa coefficient between human evaluators was 0.7098. These results demonstrate that even in the more complex task of evaluating extra-long text content, our evaluation framework still shows good consistency.

A.5 Additional Experiments with Open-Source Models

To validate the generalizability of our framework, we conduct additional experiments using DeepSeek-v3 (Liu et al., 2024), a state-of-the-art open-source language model. As shown in Table 7, the experimental results demonstrate remarkable performance across all evaluation metrics. Specifically, DeepSeek-v3 achieved an Input Coverage of 0.2554 and a Reference Coverage of 0.4553, surpassing other baseline models in literature coverage assessment. In the outline quality evaluation, DeepSeek-v3 attains a score of 87.42, which not only exceeds other models but also approaches the benchmark set by human-written surveys (87.62). Furthermore, across the three dimensions of content quality structure, relevance, and coverage, DeepSeek-v3 demonstrates exceptional performance with scores of 79.20, 80.17, and 81.07 respectively, yielding a mean score of 80.15 that outperforms other comparative models.

These empirical results not only corroborate the effectiveness of our methodology but also establish its applicability to open-source models. Notably, DeepSeek-v3 (Liu et al., 2024) exhibits superior performance at a lower operational cost ($0.37 per survey) compared to GPT-4o-mini ($0.43 per survey). Such advancement has substantial implications for the sustainable development of automated research tools and methodologies.

A.6 Details of Time and Economic Cost

The SurveyForge framework generates comprehensive survey papers with approximately 64k tokens in length, comparable to human-written surveys. The generation process requires an average input of 2.37M tokens and produces 0.13M tokens of output. Taking GPT-4-mini-2024-07-18 as an example, the economic cost amounts to merely $0.43. Regarding the temporal efficiency, the entire framework completes the generation within approximately 10 minutes (note that the actual duration may vary depending on API rate limits). These metrics demonstrate that the SurveyForge framework enables researchers to efficiently acquire domain knowledge at a remarkably low cost.

A.7 Qualitative Results

In this section, we present qualitative comparisons to demonstrate the effectiveness of our proposed framework in generating academically structured survey outlines. Specifically, we compare the outlines generated by our method with those produced by baseline approaches, as shown in Fig. 4.

The baseline outlines exhibit several notable issues. First, the logical organization of sections and subsections is often suboptimal, with limited hierarchical depth and coherence. Additionally, there is a tendency to treat individual studies or papers as standalone subsections, resulting in fragmented and overly granular structures. Furthermore, redundancy is frequently observed, with similar or overlapping topics appearing in multiple sections, which reduces clarity and disrupts the logical flow of the outline.

Methods	Model	Reference Quality		Outline Quality	Content Quality			
		Input Cov.	Reference Cov.		Structure	Relevance	Coverage	Avg
Human-Written	-	-	0.6294	87.62	-	-	-	-
SURVEYFORGE	Claude-3-Haiku	0.2231	0.3960	86.85	73.82	79.62	75.59	76.34
SURVEYFORGE	GPT-4o mini	0.2018	0.4236	86.62	77.10	76.94	77.15	77.06
SURVEYFORGE	Deepseek-v3	0.2554	0.4553	87.42	79.20	80.17	81.07	80.15

Table 7: Comparison of open source and closed source models on SurveyBench.

In contrast, the outlines generated by our framework effectively address these issues. By leveraging a heuristic learning approach and incorporating domain-specific structural patterns, our method produces well-organized outlines that align with academic writing standards. The generated outlines demonstrate clear hierarchical organization, thematic coherence, and appropriate grouping of related topics, providing a solid foundation for comprehensive and logically structured surveys.

A.8 Example of Generated Survey

As shown in Fig. 5, we have provided the example of the generated survey by SURVEYFORGE, more complete examples can be found at https://anonymous.4open.science/r/survey_example-7C37/. Specifically, by observing the generated survey paper, we found that SURVEYFORGE is not only capable of summarizing knowledge within a specific academic field based on logical structures but also excels at providing insights and recommendations for some potential research directions.

For instance, in a survey paper generated by SURVEYFORGE titled "Comprehensive Survey on Multimodal Large Language Models: Advances, Challenges, and Future Directions", Section 8 offers a detailed outlook on several potential future technological pathways for Multimodal Large Language Models (MLLMs), such as scalability enhancements, cross-modal interaction and integration, and efficient training and inference solutions. Besides, the survey paper also raises concerns about the ethical and societal implications of the excessive use of MLLMs, including their potential impact on issues such as gender, race, ethnicity, and socioeconomic status. Furthermore, SURVEYFORGE has outlined numerous application scenarios for MLLMs, including AI-driven agents, interactive systems, Augmented Reality (AR), and specialized domains such as healthcare and education. In addition, SURVEYFORGE further analyzes the challenges that need to be addressed to apply MLLMs to these practical scenarios. For instance, addressing computational limitations and tackling privacy concerns associated with systems that rely on large amounts of data, which require robust frameworks for data management and obtaining user consent.

A.9 Prompt Used

This section outlines the key prompts employed in SURVEYFORGE, covering those for outline generation, content generation, and evaluation.

The outline generation prompt incorporates two key elements: the structure of human-written survey papers and relevant literature on the topic. This prompt ensures that the generated outline adheres to academic conventions, with section titles aligned to the survey topic, maintaining logical connections between sections while avoiding redundancy. The content generation prompt guides LLMs in drafting individual sections of a survey paper. It requires the generated content to be supported by references from relevant literature and specifies length constraints to ensure clarity and precision.

For the prompts used for evaluation, we design the evaluation rules from both the outline and the content. Regarding outline evaluation, LLMs are instructed to score from the aspects of topic uniqueness, structural balance, hierarchical clarity and logical organization, with the total score for each aspect serving as the overall score for the outline. For content evaluation, the process references human-written surveys: LLMs first review such surveys on the same topic to establish context before evaluating AI-generated content. This approach grounds the evaluation in established academic writing practices, enhancing the reliability of the assessment.

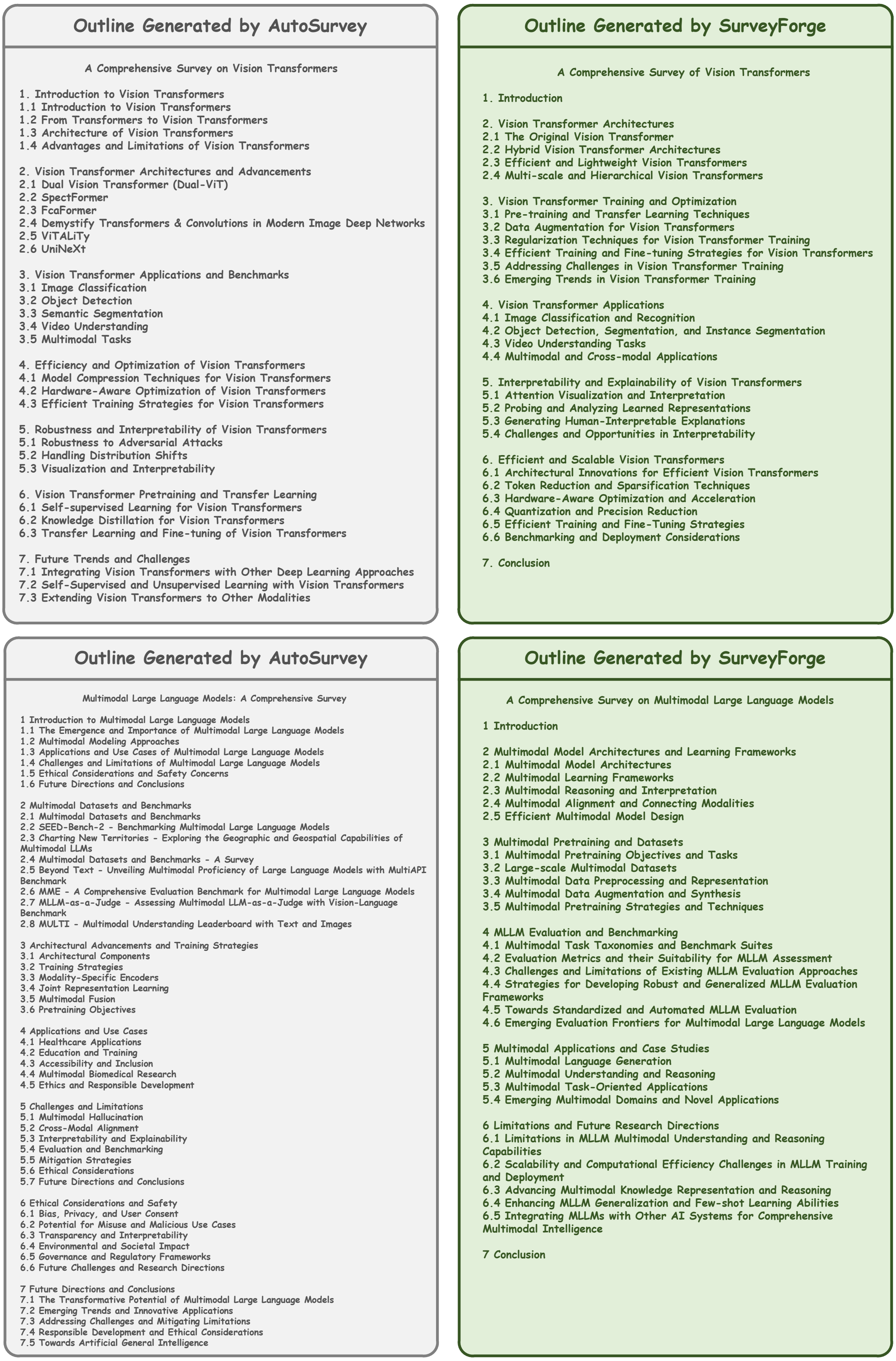

Figure 4: Comparisons of survey outlines generated by the baseline method (left) and our proposed framework (right). The baseline displays a fragmented structure, whereas our method yields a more comprehensive, systematically organized outline.

Survey Paper Generated by SurveyForge

Comprehensive Survey on Multimodal Large Language Models

1. Introduction

In recent years, the field of artificial intelligence (AI) has undergone significant transformations, largely attributed to the advancements in language models. Among these, Multimodal Large Language Models (MLLMs) stand out as a critical innovation, offering capabilities that extend beyond the confines of unimodal data processing to a more integrated and comprehensive comprehension of the world [1]. This subsection delves into the historical evolution, core significance, and transformative potential of MLLMs across varied applications, reflecting on how these models have revolutionized AI and paved the way for future innovations.

Initially, the concept of multimodality emphasized combining disparate data types such as text, images, and audio, among others, into a coherent system that could better mimic human-like understanding and reasoning [2]. Early efforts in this domain faced challenges related to data alignment and the synergistic fusion of modalities, which hindered effective cross-modal interactions [3]. However, the advent of sophisticated architectures like transformers has unlocked unprecedented potential in this area. The adoption of these architectures facilitates the seamless integration of different modalities, offering enhanced dimensionality and interaction capabilities that were previously unattainable [4].

The historical evolution of MLLMs can be traced through various developmental phases characterized by increasing model complexity and capacity for cross-modal reasoning [5]. Initially, the focus was on creating foundational models capable of handling single modalities. As research progressed, there was a significant shift towards developing models that could analyze and synthesize information from multiple sources simultaneously. This evolution was marked by seminal works that introduced frameworks for modality collaboration and integration [6]. These advancements have enabled MLLMs to excel in tasks that require holistic data interpretation, from visual question answering to complex cognitive tasks such as multimodal sentiment analysis and contextual understanding [7].

The significance of MLLMs in artificial intelligence is multifaceted. At its core, the integration of multiple data modalities within a unified framework allows for a more nuanced understanding of context, leading to better performance in multimodal tasks such as image captioning, speech recognition, and autonomous navigation [8]. For

1

instance, models that leverage textual and visual data in tandem have demonstrated the ability to perform complex reasoning tasks, such as interpreting and generating visual content based on textual prompts [9]. This capability not only enhances the accuracy of AI systems but also broadens the scope of applications to domains that require high precision and contextual awareness, such as healthcare and autonomous systems [10].

However, the transition to multimodal frameworks presents several challenges and trade-offs, particularly concerning the integration complexity and computational demands of these models. Notably, while the incorporation of multi-modal data enhances model capabilities, it also introduces vulnerabilities like a decrease in performance when one of the modalities is missing or incomplete, as explored in recent studies [11]. There is a need for robust design frameworks that can accommodate multimodal inputs seamlessly while maintaining efficiency and scalability [12]. Furthermore, developing standardized metrics and benchmarks for evaluating the performance of MLLMs across diverse modalities remains an ongoing challenge, necessitating more comprehensive approaches to assessment [13].

Looking forward, the future directions in the realm of MLLMs are poised towards enhancing scalability and interaction mechanisms, ensuring these models can efficiently manage larger datasets and more complex tasks [14]. Emerging trends suggest a promising trajectory for the integration of MLLMs with knowledge graphs and vector databases, potentially alleviating issues like hallucinations and knowledge limitations inherent in current models [15]. Ultimately, ongoing research must focus on refining the modality integration frameworks and addressing the ethical considerations surrounding MLLM deployment, particularly concerning bias and data privacy [10].

In summary, Multimodal Large Language Models represent a pivotal advancement in AI, embodying the convergence of diverse modalities to achieve more comprehensive intelligence. As these models continue to evolve, they hold the potential to redefine interactions across countless applications, driving progress towards truly intelligent systems capable of holistic reasoning and decision-making.

2. Core Architectures and Design Principles References

2.1 Architectural Paradigms
2.2 Multimodal Integration Techniques
2.3 Processing and Interaction Mechanisms
2.4 Computational Efficiency and Scalability

2

2.5 Emerging Design Innovations

3. Training Paradigms and Techniques

3.1 Foundational Training Strategies
3.2 Advanced Pre-training and Adaptation Techniques
3.3 Knowledge Transfer and Sharing Mechanisms
3.4 Optimization and Resource-Efficient Training
3.5 Challenges and Solutions in Training
3.6 Evaluation and Benchmarking in the Training Process

4. Multimodal Data Processing and Representation Learning

4.1 Representation Alignment
4.2 Multimodal Fusion Strategies
4.3 Robust Representation Learning
4.4 Self-supervised and Semi-supervised Learning
4.5 Dynamic Representation Adaptation

5. Evaluation and Benchmarking

5.1 Performance Metrics and Standard Benchmarks
5.2 Challenges in Multimodal Evaluation Practices
5.3 Advanced Benchmarking Techniques
5.4 Proposing Improvements and Innovations
5.5 Comparative Analysis of Assessment Frameworks

6. Applications and Use Cases

6.1 Healthcare and Medical Applications
6.2 Autonomous Systems and Robotics
6.3 Assistive Technologies
6.4 Content Creation and Multimedia Generation
6.5 Education and Skill Development

7. Ethical Considerations and Societal Impacts

7.1 Bias and Fairness in Multimodal Models
7.2 Privacy and Security in Multimodal Systems

3

7.3 Ethical Frameworks and Policy Implications
7.4 Societal Impacts and Responsible AI Use

8. Future Directions and Emerging Trends

8.1 Scalability Enhancements
8.2 Cross-Modal Interaction and Integration
8.3 Efficient Training and Inference Solutions
8.4 Ethical and Societal Implications
8.5 Emerging Applications and Use Cases

9. Conclusion

……

Reference

[1] A Survey on Evaluation of Multimodal Large Language Models
[2] Multimodal Machine Learning A Survey and Taxonomy
[3] Does my multimodal model learn cross-modal interactions It's harder to tell than you might think!
[4] Multimodal Deep Learning
[5] Foundations and Trends in Multimodal Machine Learning Principles, Challenges, and Open Questions
[6] mPLUG-Owl2 Revolutionizing Multi-modal Large Language Model with Modality Collaboration
[7] Multimodal Foundation Models From Specialists to General-Purpose Assistants
[8] Large Multimodal Agents A Survey
[9] How Multimodal Integration Boost the Performance of LLM for Optimization Case Study on Capacitated Vehicle Routing Problems
[10] A Comprehensive Survey of Large Language Models and Multimodal Large Language Models in Medicine
[11] Are Multimodal Transformers Robust to Missing Modality
[12] Challenges and Applications of Large Language Models
[13] MM-Vet Evaluating Large Multimodal Models for Integrated Capabilities
[14] A Survey of Resource-efficient LLM and Multimodal Foundation Models
[15] When Large Language Models Meet Vector Databases A Survey
……

4

Figure 5: Example of the survey generated by SURVEYFORGE. Please refer to https://anonymous.4open.science/r/survey_example-7C37/ for more auto-generated results.

Generation Prompt

Outline:

SUBSECTION_OUTLINE_WITH_SURVEY_PROMPT = '''
You are an expert in artificial intelligence writing a comprehensive outline of the survey about **[TOPIC]**.

You have created the following overall outline:

[OVERALL OUTLINE]

You need to enrich the section **[SECTION NAME]**, described as: **[SECTION DESCRIPTION]**

Main task:
Generate a comprehensive framework for **[SECTION NAME]** by creating an appropriate number of subsections (typically 3-6, but adjust based on content importance and complexity). Each subsection should focus on a specific aspect and be followed by a Informative description.

Resources provided:

1. **A list of [RAG NUM] relevant papers with titles, abstracts, publication dates for this section:**

   [PAPER LIST]
   ---

2. **Titles, abstracts, top-second outlines and publication dates of human-written surveys** that may be related to [TOPIC].

   [SURVEY LIST]

   Note: These surveys may not be directly about **[TOPIC]**. Only use these to understand the logical structure, style, and academic phrasing typical of academic survey papers written by humans.

How to use the provided resources:
- Use the relevant papers to identify key themes, recent developments, and important concepts within **[SECTION NAME]**.
- Refer to the human-written surveys to understand typical structures and academic phrasing, but ensure your outline is original and specifically tailored to **[TOPIC]** and **[SECTION NAME]**.
- Synthesize information from both sources to create a comprehensive and up-to-date framework for the section.
- Prioritize recent developments and emerging trends when creating your outline, while also acknowledging foundational concepts.

Guidelines:
1. **Relevance:** Each subsection must be related to **[SECTION NAME]** and align with its description.
2. **Originality:** Learn from the human-written surveys to inform your structure, but be careful to avoid plagiarism.
3. **Logical Flow:** Arrange subsections in a logical order that builds upon previous ones, ensuring a coherent progression of ideas. It is important to note that there is no overlap between subsection and its bullet points, which represent different aspects of the section.
4. **Flexibility:** The number of subsections should be determined by the content requirements of **[SECTION NAME]**. While 3-6 subsections are typical, prioritize comprehensive coverage over adhering to a strict number.
5. **Separability:** Each subsection should have **an informative description** and **several (no more than 3) sub-domain points with informative sub-description**, which do not duplicate and fit the subsection**, the number of points per subsection does not need to be consistent. You can add or subtract according to the actual scope of the section. Each bullet point should represent a key aspect or sub-domain of the section, followed by a informative description.

** Output format: **
<format>
Subsection 1: [NAME OF SUBSECTION 1]
Description 1: [INFORMATIVE DESCRIPTION OF SUBSECTION 1]
1. [Informative description of Key aspect or sub-domain 1 of SUBSECTION 1]
2. ...

Subsection 2: [NAME OF SUBSECTION 2]
Description 2: [INFORMATIVE DESCRIPTION OF SUBSECTION 2]
1. [Informative description of Key aspect or sub-domain 1 of SUBSECTION 2]
...
N. [Informative description of Key aspect or sub-domain N of SUBSECTION 2]

...

Subsection K: [NAME OF SUBSECTION K]
Description K: [INFORMATIVE DESCRIPTION OF SUBSECTION K]
1. [Informative description of Key aspect or sub-domain 1 of SUBSECTION K]
2. ...

</format>

Note: The number of subsections (K) should be appropriate for the content of **[SECTION NAME]**. Ensure descriptions are specific, contain key terminology, and provide clear guidance for detailed content creation.
Only return the outline without any other informations:
'''

Content:

You are writing the subsection "[SUBSECTION NAME]" under the section "[SECTION NAME]" for a top-tier and comprehensive survey paper on [TOPIC]. As a distinguished expert, deliver content that combines academic rigor with innovative insights.

The overall outline of your survey is as follows:\n

[OVERALL OUTLINE]

Below are a list of papers for references:\n

[PAPER LIST]

<instruction>
Now, focus on writing the content for the subsection "[SUBSECTION NAME]" under "[SECTION NAME]". The content you write must be more than [WORD NUM] words.

Subsection Focus:

[DESCRIPTION]

Core Requirements:

1. Content Structure
- Begin with a concise overview of the subsection's scope
- Maintain logical flow with clear transitions
- Conclude with synthesis and future directions
- Balance breadth and depth of coverage

2. Academic Analysis
- Provide comparative analysis of different approaches
- Evaluate strengths, limitations, and trade-offs
- Identify emerging trends and challenges
- Present technical details with precision
- Include equations/formal definitions where necessary

3. Citation Guidelines
- You should cite as many relevant paper as possible related to "[SUBSECTION NAME]".
- When writing sentences that are based on specific papers above, you cite the "paper_title" in a '[]' format to support your content.
- Note that the "paper_title" is not allowed to appear without a '[]' format. Once you mention the 'paper_title', it must be included in '[]'.
- Remember that you can only cite the paper provided above and only cite the "paper_title"!!!
- Integration: Support key claims with relevant citations
- Example: "Lin et al. [paper_title1] have shown... Further studies [paper_title2; paper_title3] confirm..."

4. Critical Insights
- Synthesize information rather than summarize
- Draw connections between different approaches
- Highlight practical implications
- Offer innovative perspectives or future directions
- Support arguments with empirical evidence
- Maintain scholarly tone throughout

Quality Markers:
- Demonstrates deep technical understanding
- Provides novel insights and analysis
- Maintains objective academic tone
- Presents coherent narrative flow
- Supports all key claims with citations

Remember, the quality of your work should reflect the standards expected in top-tier academic publications. Your analysis should be thorough, your arguments well-supported, and your insights valuable to the academic community. Approach this task as if your reputation as a leading expert in the field depends on the quality of this subsection.
</instruction>

Provide the content for subsection "[SUBSECTION NAME]" in this format:
<format>
[CONTENT OF SUBSECTION]
</format>

Only return the content more than [WORD NUM] words you write for the subsection [SUBSECTION NAME] without any other information, ensuring it provides a comprehensive, in-depth analysis that meets the high academic standards described above. Your work will be evaluated based on its scholarly merit, analytical depth, and potential contribution to the field.
Do not repeat the subsection title at the beginning of your response. Start directly with the content of the subsection.
'''

Evaluation Prompt

Outline:

Task: As a rigorous academic evaluator about {topic}, assess the quality of an AI-generated outline. You need to judge whether it can serve as an outline for a high-quality academic review paper.

Subject for Evaluation:
{ai_outline}

<Instruction>
Your job is to assess how well the outline of the generated literature review.

Evaluation Focus: **OUTLINE QUALITY ONLY**

Outline Assessment Criteria (100 points total):

1. Topic Uniqueness (30 points)
- No duplicate topics across sections/subsections
- Each section contains unique content
- No redundant future/conclusion sections
- Clear distinction between related topics

2. Structural Balance (30 points)
- Reasonably balanced number of subsections across main content chapters
- No obviously under-developed sections
- No overly detailed sections that dominate the outline
- Variations in subsection numbers should align with topic importance/complexity

3. Hierarchical Clarity (20 points)
- Clear parent-child relationships
- Appropriate topic levels for each section's role
- Logical subdivision aligned with academic conventions
- Consistent granularity where appropriate

4. Logical Organization (20 points)
- Natural topic progression following academic norms
- Clear relationships between sections
- Coherent topic grouping
- Purposeful content flow matching section functions

Score Classifications:

90-100: Exceptional
- Zero content duplication
- Perfect structural balance
- Clear hierarchy
- Logical flow

80-89: Strong
- Minimal content overlap
- Generally balanced structure
- Good hierarchical organization
- Clear progression

70-79: Adequate
- Some topic repetition
- Slightly uneven structure
- Basic hierarchy maintained
- Basic logical flow

60-69: Weak
- Notable redundancy
- Imbalanced sections
- Unclear hierarchy
- Poor topic progression

Below 60: Poor
- Extensive duplication
- Severely imbalanced
- Confused hierarchy
- No logical organization

Content Coverage:

Task: As an expert literature review evaluator, assess only the **coverage quality** of a generated literature review compared to a human-written reference on {topic}.

Note: The human-written review serves only as a reference point, not as the absolute standard.

Coverage Quality Definition:
Coverage quality refers to the comprehensiveness, depth, and balance of topic treatment within a literature review, including the breadth of relevant concepts covered and the proportional attention given to each area.

Human-Written Review (Reference):

{human_review}

Generated Review for Evaluation:

{ai_review}

Coverage Evaluation Criteria (100 points total):

1. **Topic Comprehensiveness (35 points)**
   - Range of essential topics covered
   - Inclusion of emerging areas
   - Identification of key concepts
   Scoring Guide:
   - 30-35: Comprehensive coverage with emerging topics
   - 20-29: Good coverage with minor gaps
   - 0-19: Significant omissions or major gaps

2. **Discussion Depth (35 points)**
   - Detail level of concept analysis
   - Development of key arguments
   - Thoroughness of explanations
   Scoring Guide:
   - 30-35: Exceptional depth across topics
   - 20-29: Good depth with some variation
   - 0-19: Consistently superficial treatment

3. **Content Balance (30 points)**
   - Proportional coverage of topics
   - Appropriate emphasis distribution
   - Logical allocation of space
   Scoring Guide:
   - 25-30: Well-balanced coverage throughout
   - 15-24: Generally balanced with minor issues
   - 0-14: Significant imbalance issues

Scoring Requirements:
- Prioritize accuracy over conservatism
- AVOID "safe" middle-range scores that don't reflect true quality. Score based purely on merit, not on scoring "comfort zones"
- Each score must reflect precise performance level, not range averages (e.g., 25 for 20-29 range)
- Use full scoring range (0-100)
- Base scores on objective comparison to human reference
- Acknowledge that best practices may evolve

Output Format:
Return only a single numerical score (0-100). No additional commentary.
'''

Content Relevance:

Task: As an expert literature review evaluator, assess only the **relevance quality** of a generated literature review compared to a human-written reference on {topic}.

Note: The human-written review serves only as a reference point, not as the absolute standard.

Relevance Quality Definition:
Relevance quality in a literature review refers to how well the content aligns with the stated topic, the appropriateness of included information, and the focus of the discussion on key aspects of the subject matter.

Reference Materials:
Human-Written Review (Reference):

{human_review}

Generated Review for Evaluation:

{ai_review}

Relevance Evaluation Criteria (100 points total):

1. **Topic Alignment (35 points)**
   - Coverage of core aspects
   - Alignment with research focus
   - Depth of relevant discussion
   Scoring Guide:
   - 30-35: Excellent alignment with comprehensive coverage
   - 20-29: Good alignment with minor gaps
   - 0-19: Significant misalignment or major gaps

2. **Content Appropriateness (35 points)**
   - Relevance of examples and evidence
   - Precision of discussion
   - Connection to main topic
   Scoring Guide:
   - 30-35: Highly relevant with precise discussion
   - 20-29: Generally relevant with minor inconsistencies
   - 0-19: Multiple irrelevant elements or poor precision

3. **Information Focus (30 points)**
   - Concentration on key points
   - Absence of tangential content
   - Purposeful content selection
   Scoring Guide:
   - 25-30: Sharp focus with minimal deviation
   - 15-24: Adequate focus with some tangential content
   - 0-14: Poor focus or excessive deviation

Scoring Requirements:
- Prioritize accuracy over conservatism
- AVOID "safe" middle-range scores that don't reflect true quality. Score based purely on merit, not on scoring "comfort zones"
- Each score must reflect precise performance level, not range averages (e.g., 25 for 20-29 range)
- Use full scoring range (0-100)
- Base scores on objective comparison to human reference
- Acknowledge that best practices may evolve

Output Format:
Return only a single numerical score (0-100). No additional commentary.
'''

Content Structure:

Task: As an expert literature review evaluator, assess only the **structural quality** of a generated literature review compared to a human-written reference on {topic}.

Note: The human-written review serves only as a reference point, not as the absolute standard.

Structural Quality Definition:
Structural quality in a literature review refers to the organization, logical flow, and presentation of information. It encompasses how well the review is organized, how ideas are connected and developed, and how the overall structure enhances understanding of the topic.

Reference Materials:
Human-Written Review (Reference):

{human_review}

Generated Review for Evaluation:

{ai_review}

Structural Evaluation Criteria (100 points total):

1. **Logical Flow & Organization (35 points)**
   - Progressive development of ideas
   - Effective transitions between concepts
   - Clear argumentative thread
   Scoring Guide:
   - 30-35: Exceptional logical progression with seamless transitions
   - 20-29: Generally logical with minor flow issues
   - 0-19: Significant organizational problems

2. **Hierarchical Structure (35 points)**
   - Section/subsection organization
   - Topic hierarchy clarity
   - Internal coherence
   Scoring Guide:
   - 30-35: Well-defined structure enhancing comprehension
   - 20-29: Adequate structure with some inconsistencies
   - 0-19: Poor hierarchical organization

3. **Format & Presentation (30 points)**
   - Heading/subheading usage
   - Academic formatting consistency
   - Visual organization
   Scoring Guide:
   - 25-30: Consistent, professional formatting
   - 15-24: Minor formatting inconsistencies
   - 0-14: Major formatting issues

Scoring Requirements:
- Prioritize accuracy over conservatism
- AVOID "safe" middle-range scores that don't reflect true quality. Score based purely on merit, not on scoring "comfort zones"
- Each score must reflect precise performance level, not range averages (e.g., 25 for 20-29 range)
- Use full scoring range (0-100)
- Base scores on objective comparison to human reference
- Acknowledge that best practices may evolve

Output Format:
Return only a single numerical score (0-100). No additional commentary.