A Multi-Agent LLM Framework with Hierarchical Citation Graph for Automated Survey Generation

Minh-Anh Nguyen [*1,2] **Minh-Duc Nguyen** [*1] **Ha Lan N. T.** [*1] **Kieu Hai Dang** [3] **Nguyen Tien Dong** [2,3] **Dung D. Le** [1,3]

Abstract

Large language models (LLMs) are increasingly used for automated survey generation, yet existing methods often ignore structural relationships among papers, leading to weak taxonomies and limited contextual understanding. We propose **SurveyG**, an LLM-based agent framework that integrates a hierarchical citation graph capturing both citation dependencies and semantic relatedness. The graph organizes papers into three layers: **Foundation**, **Development**, and **Frontier**, modeling research evolution from seminal works to emerging directions. By combining horizontal exploration within layers and vertical traversal across layers, SurveyG generates multi-level summaries that are consolidated into a coherent survey outline. A multi-agent validation stage further improves consistency, coverage, and factual accuracy. Experiments with human experts and LLM-as-a-judge evaluations show that SurveyG outperforms state-of-the-art methods, producing more comprehensive and better-structured surveys. Our implementation is available at `https://anonymous.4open.science/r/SurveyG-1701/`.

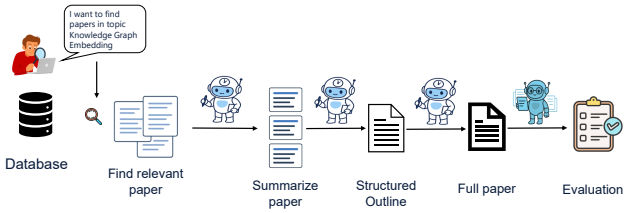

Figure 1. Standard automated survey pipeline: (1) collect relevant papers, (2) build a structured outline, and (3) write the full survey.

[*]Equal contribution [1]Center for AI Research, VinUniversity, Vietnam [2]CMC-OpenAI, Vietnam [3]CECS, VinUniversity, Vietnam. Correspondence to: Dung D. Le <dung.ld@vinuni.edu.vn>.

Accepted by ICML 2026 AI for Science Workshop.

1. Introduction

The exponential growth of research publications, particularly in rapidly evolving fields such as Artificial Intelligence (Huynh-The et al., 2023), has made it increasingly difficult for researchers to keep pace with new developments (Wang et al.; Wen et al., 2025). While survey papers serve as invaluable resources by synthesizing existing knowledge and identifying emerging trends, their manual construction is costly, time-consuming, and often unable to keep up with the overwhelming influx of literature (Liang et al., 2025). Although large language models (LLMs) offer promising text generation capabilities, they face critical limitations in handling massive reference sets, maintaining academic rigor, and providing up-to-date knowledge (Wu et al., 2025; Han et al., 2024). These challenges underscore the urgent need for an automated survey generation framework that can efficiently retrieve, organize, and synthesize literature into coherent, high-quality surveys tailored to users' research interests.

Some recent studies (Su et al., 2025; Wang et al.; Liang et al., 2025; Yan et al., 2025) have proposed autonomous survey generation frameworks based on user queries, following the basic pipeline illustrated in Figure 1. While these approaches represent promising progress, they exhibit two key limitations. **Firstly,** they neglect the relationships between papers, such as citation links, methodological connections, or subtopic dependencies, which are essential for understanding how works build upon one another, improve over foundational methods, and collectively shape research trends. **Secondly,** these frameworks employ a naive strategy for constructing structured outlines or full survey papers, simply concatenating summaries of individual papers. This not only exacerbates the long-context problem in LLMs but also fails to exploit the hierarchical organization of related works within subtopics.

To address these limitations, we propose **SurveyG**, an autonomous survey generation system that emphasizes knowledge representation and hierarchical summarization. We design an LLM-based multi-agent framework that represents knowledge using a hierarchical citation graph, where nodes correspond to papers and edges capture citation rela-

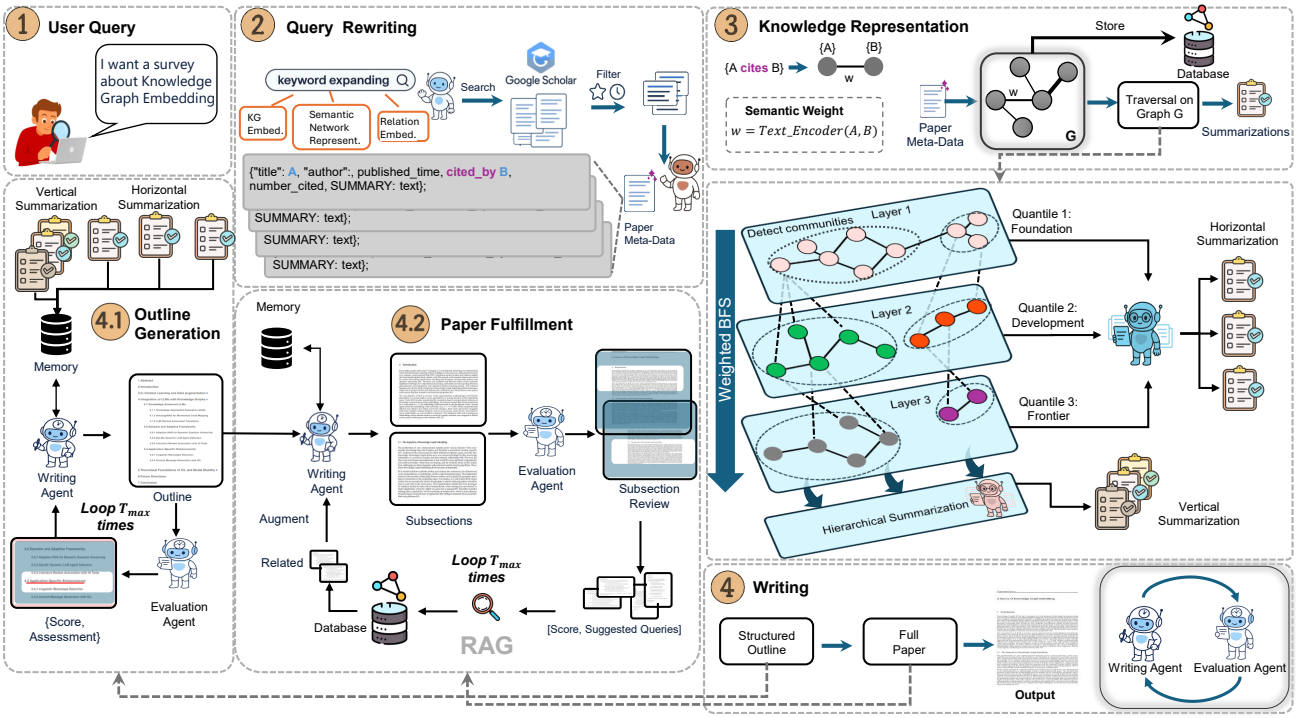

Figure 2. Starting from a user's query, SurveyG retrieves and filters relevant papers (step 1-2), builds a hierarchical citation graph, and applies horizontal and vertical traversals to produce multi-aspect summaries (step 3). A multi-agent framework then leverages these pre-built summaries to produce a structured outline and a complete survey paper (step 4).

tionships and semantic similarity. The graph is organized into three layers: **Foundation, Development**, and **Frontier**, reflecting research progression from seminal contributions to emerging directions. By combining horizontal searches within layers and vertical traversals across layers, our framework generates multi-aspect summaries that are consolidated into structured survey outlines. We evaluate SurveyG on 10 computer science topics from the SurGE benchmark (Su et al., 2025), comparing against state-of-the-art frameworks (Liang et al., 2025; Yan et al., 2025; Wang et al.) along five dimensions: *Coverage*, *Structure*, *Relevance*, *Synthesis*, and *Critical Analysis*. Our key contributions are: (1) a hierarchical citation graph representation modeling both citation and semantic relationships, (2) a graph-based traversal mechanism that produces diverse, multi-aspect summarizations capturing methodological foundations, developmental trends, and frontier directions, and (3) a multi-agent framework combining retrieval-augmented generation with hierarchical summaries as memory to construct coherent, evidence-grounded survey drafts.

2. Methodology

We introduce **SurveyG**, an automated survey generation framework with two phases: a preparation phase that retrieves and summarizes papers, builds a hierarchical citation graph, and extracts relationships, and a generation phase that creates a structured outline and composes the full survey through instruction-prompted multi-agent collaboration, detailed in Figure 2.

2.1. Preparation Phase

We represent the relationships among papers using a *hierarchical citation graph*, where nodes correspond to academic papers and edges capture both citation links and semantic similarity, each weighted by a value w. Each node is further assigned to one of three layers: **Foundation**, **Development**, or **Frontier**, which reflect the role of the paper in the progression of research. Formally, the hierarchical citation graph is defined as $G = (V, E, L)$, where V denotes the set of nodes (papers), $E \subseteq V \times V$ is the set of directed or undirected edges encoding citation or semantic relationships, and $L : V \rightarrow \{\text{Foundation}, \text{Development}, \text{Frontier}\}$ is a layer assignment function. For each node $v_i \in V$, we associate a corresponding document $d_i \in D$, where D is a database storing the complete content of all papers. In addition, each node $v_i \in V$ is equipped with attributes that include a summarization of d_i as well as metadata such as the paper's title and publication year.

2.1.1. Searching Relevant Paper

Given a user query q, our goal is to construct a hierarchical citation graph G that encompasses all relevant papers while capturing the evolutionary trends of research in the field. We first employ an LLM to expand the query into a set of diverse keywords $\{k_1, \ldots, k_n\} = \text{LLM}(q)$. Using these keywords, we retrieve candidate papers through the crawling module. After collecting the relevant papers, we establish edges between them based on citation links and quantify their semantic relatedness through weighted connections. The weight w assigned to an edge connecting papers v_i and v_j is defined as

$$w = \text{sim}\big(\text{Text_Encoder}(v_i), \text{Text_Encoder}(v_j)\big), \quad (1)$$

where $\text{sim}(\cdot)$ denotes the cosine similarity between the text embedding vectors of the two papers. We consider two embedding strategies: abstract-based embeddings, which are efficient and widely available but may be incomplete or overstated, and LLM-generated summary-based embeddings, which better capture methodological and empirical content but require full text and extra processing (Zhang et al., 2025). We use summary-based embeddings by default because our pipeline already produces content-focused summaries (as detailed in the paragraph below), enabling richer semantic representations of methods, tasks, and domain contributions, and falls back to abstract-based embeddings when full-text access is unavailable.

To better leverage the key content of each paper for survey generation, every node v_i is enriched with a summarization derived from its corresponding document d_i. We design specialized prompt templates tailored to different paper types (Liang et al., 2025) (e.g., surveys, methodological contributions, benchmarks, theoretical works) to extract salient information, including research objectives, methodological approaches, key findings, and limitations. After this phase, we obtain a flat graph $\hat{G} = (V, E)$ that encompasses the papers relevant to the user's topic along with their relationships. Each node $v_i \in V$ is associated with a set of attributes defined as

$$A(v_i) = \{\text{metadata}(v_i), \text{summary}(d_i)\},$$

where $\text{metadata}(v_i)$ contains bibliographic information such as the paper's title, authors, and publication year, and $\text{summary}(d_i)$ represents the content-based summarization of the corresponding document. These attributes are also stored in the database D to facilitate efficient retrieval and analysis.

2.1.2. Knowledge Representation

To reflect the developmental progression of research within a topic, we assign each node in the flat graph $\hat{G} = (V, E)$ to one of three hierarchical layers via a layer assignment function. (1) **Foundation Layer**: The foundation layer consists of seminal and high-impact works that form the intellectual backbone of the field. For each paper p, we define a *trending score* as

$$\text{trend}_{\text{score}}(p) = \frac{\text{citation_count}(p)}{1 + \text{year_published}(p)}, \quad (2)$$

where $\text{year_published}(p)$ denotes the number of years elapsed since the paper's publication. Papers are ranked by this score, and the top-K entries constitute the foundation set:

$$V_{\text{foundation}} = \{v_i \in V \mid \text{trend}_{\text{score}}(v_i) \leq K\}.$$

These papers are not only highly cited but also serve as conceptual anchors that establish key paradigms and problem formulations underpinning later research. (2) **Development Layer:** The development layer captures the historical evolution of the field before a time landmark T (eg, 2025), representing works that refine, extend, or challenge the foundations. Formally,

$$V_{\text{development}} = \{v_i \in V \mid \text{year}(v_i) < T,\ v_i \notin V_{\text{foundation}}\},$$

These works are often incremental yet essential: they consolidate methodological frameworks, validate empirical findings, and enable the community to mature foundational ideas into established research threads. (3) **Frontier Layer:** The frontier layer reflects the cutting edge of inquiry, consisting of recent contributions that point toward emerging trends and open challenges. It is defined as

$$V_{\text{frontier}} = \{v_i \in V \mid \text{year}(v_i) \geq T,\ v_i \notin V_{\text{foundation}}\}.$$

Unlike the development layer, frontier works are temporally close to the present and thus provide a window into the current momentum and future trajectories of the domain. After this mapping, the hierarchical citation graph is represented as

$$G = (V, E, L), \quad V = V_{\text{foundation}} \cup V_{\text{development}} \cup V_{\text{frontier}}.$$

Traversing G along horizontal (intra-layer) and vertical (inter-layer) edges then enables the generation of multi-aspect summaries covering methodologies, developmental trends, and future directions.

2.1.3. Traversal on Graph Strategy

We propose a two-stage summarization framework designed to capture both the breadth and depth of the hierarchical citation graph. In the *horizontal stage*, to capture the internal structure of each layer V_l, we partition it into communities using the Leiden algorithm (Traag et al., 2019), yielding

$$\mathcal{C}_l = \{C_{l,1}, \ldots, C_{l,m_l}\}, \quad \bigcup_{j=1}^{m_l} C_{l,j} = V_l.$$

Algorithm 1 Vertical Traversal for Multi-Summarization

Input: Citation graph $G = (V, E, L)$, foundation papers $V_{\text{foundation}}$
Output: $\{T_{\text{path}}^{(1)}, \ldots, T_{\text{path}}^{(K)}\}$, where $K = |V_{\text{foundation}}|$
for $s \in V_{\text{foundation}}$ **do**
  $P_1 \leftarrow \{\text{EXTRACT}(s)\}$
  $P_2 \leftarrow \{\text{EXTRACT}(u) \mid u \in \text{WBFS}(s, \text{Development})\}$
  $P_3 \leftarrow \{\text{EXTRACT}(w) \mid w \in \text{WBFS}(P_2, \text{Frontier})\}$
  $T_{\text{dev}} \leftarrow \text{GENERATESUMMARIZE}(P_1 \cup P_2)$
  $T_{\text{path}} \leftarrow \text{GENERATESUMMARIZE}(T_{\text{dev}}, P_3)$
  Store T_{path} as the summarization for seed s
end for
return K summarizations $\{T_{\text{path}}^{(1)}, \ldots, T_{\text{path}}^{(K)}\}$

> You are a research analyst synthesizing papers on the topic [QUERY]. <think> Explain your reasoning for clustering papers into 2-3 subgroups based on methodology, contribution, or thematic focus. </think>
> For each subgroup, summarize the shared methodological approaches, thematic contributions, and provide a concise critique comparing the works. Finally, synthesize an overall perspective highlighting how these subgroups collectively operate in the field.

Figure 3. Horizontal summarization short version prompt.

Each community $C_{l,j}$ corresponds to a coherent research direction formed by citation and semantic proximity. For every community, we query an LLM using a carefully constructed prompt that integrates Plan-and-Solve (Wang et al., 2023) strategies, along with paper-specific attributes, to generate a synthesized summary:

$$T_{l,j} = \text{LLM}(\{A(v_i) \mid v_i \in C_{l,j}\}),$$

which emphasizes the methodologies and thematic scope of the papers while capturing key relationships among them. This process uncovers sub-directions within the topic and provides a global perspective of how research clusters evolve within each layer. The detailed prompts used for information extraction are provided in Figure 3.

In the *vertical stage*, we model cross-layer dependencies by tracing the evolution of ideas from each foundation paper. For every seed $s \in V_{\text{foundation}}$, Algorithm 1 performs a weighted breadth-first search (WBFS) over the citation graph, prioritizing semantically relevant nodes via edge weights. The procedure first extracts the foundation paper (P_1), then collects Development-layer papers reachable from s (P_2), and finally expands to Frontier-layer papers from P_2 (P_3). We summarize these papers hierarchically: an intermediate summary T_{dev} is generated from $P_1 \cup P_2$, which is then used to incorporate P_3 and produce the final path summary T_{path}. This staged design leverages temporal progression to reduce long-context issues and improve insight extraction (Zhang et al., 2024). Repeating this process for all K foundation papers yields K vertical summaries, which are combined with N horizontal layer summaries for a total of $K + N$ outputs.

In contrast to earlier frameworks (Wang et al.; Wen et al., 2025; Liang et al., 2025; Yan et al., 2025) that represent papers as isolated records in a flat database and depend exclusively on RAG-based retrieval, SurveyG organizes the literature within a hierarchical citation graph G. This representation integrates both citation and semantic connections among papers, allowing the system to capture the logical progression of research topics over time. By traversing this hierarchy, SurveyG produces a series of summarizations across multiple layers, effectively revealing methodological developments, evolutionary patterns, and current research frontiers. Such a design provides a more coherent and interpretable knowledge foundation for automated survey generation.

2.2. Generation Phase

We use a multi-agent conversational framework (Wu et al., 2024) for survey generation, consisting of a Writing Agent with memory initialized from $K + N$ graph-based summaries (Sumers et al., 2023) and an Evaluation Agent that provides diversity-focused feedback using LLM reasoning. Through iterative interaction, the Writing Agent drafts content grounded in summaries, while the Evaluation Agent critiques and refines it for coherence and balance, enabling collaborative survey construction, detailed in Algorithm 2.

2.2.1. Structured Outline Construction

The Writing Agent constructs an initial structured outline by grounding each section and subsection in the $K + N$ multi-aspect summarizations, ensuring both factual grounding and thematic coherence. The Evaluation Agent then reviews the draft, assessing logical flow and suggesting refinements without altering the overall structure. After the feedback iterations, the outline converges into a coherent and evidence-supported framework. Detailed prompting for both agents is provided in Figure 6 and Appendix E. The key innovation of SurveyG lies in its ability to manage long-context survey synthesis without concatenating all reference texts (Wang et al.; Liang et al., 2025) or relying on pre-existing human-written surveys (Yan et al., 2025). Instead, it leverages hierarchical summarization from the citation graph G as structured knowledge injected into the Writing Agent.

LLM	Model	Coverage	Structure	Relevance	Synthesis	Critical Analysis
Claude	AutoSurvey	73.6 (± 3.2)	64.5 (± 4.1)	80.2 (± 2.8)	51.1 (± 5.3)	45.8 (± 4.7)
	SurveyX	74.2 (± 3.7)	71.3 (± 4.3)	82.3 (± 3.1)	62.4 (± 4.9)	52.7 (± 5.1)
	SurveyForge	81.8 (± 2.9)	78.4 (± 3.5)	89.1 (± 2.1)	75.4 (± 3.8)	70.3 (± 4.2)
	SurveyG	**88.1** (± 2.4)	**87.9** (± 2.7)	**93.6** (± 1.9)	**80.2** (± 3.2)	**77.3** (± 3.5)
GPT	AutoSurvey	90.7 (± 2.1)	86.2 (± 2.8)	89.3 (± 2.5)	84.6 (± 3.1)	82.3 (± 3.4)
	SurveyX	89.3 (± 2.6)	84.4 (± 3.2)	88.5 (± 2.8)	79.2 (± 3.9)	78.4 (± 4.1)
	SurveyForge	94.2 (± 1.8)	87.3 (± 2.5)	94.8 (± 1.6)	88.6 (± 2.7)	88.5 (± 2.9)
	SurveyG	**95.7** (± 1.5)	**88.5** (± 2.2)	**95.1** (± 1.4)	**92.2** (± 2.3)	**91.2** (± 2.6)
Deepseek	AutoSurvey	86.5 (± 2.7)	80.4 (± 3.6)	87.8 (± 2.9)	77.3 (± 4.2)	72.4 (± 4.5)
	SurveyX	85.4 (± 3.1)	81.2 (± 3.8)	89.4 (± 2.7)	78.2 (± 4.3)	75.6 (± 4.6)
	SurveyForge	**89.4** (± 2.3)	**86.5** (± 2.9)	92.5 (± 2.0)	84.1 (± 3.4)	80.7 (± 3.7)
	SurveyG	88.7 (± 2.5)	85.7 (± 3.0)	**94.2** (± 1.8)	**86.3** (± 3.1)	**82.7** (± 3.5)
Gemini	AutoSurvey	94.2 (± 1.9)	70.8 (± 5.2)	96.5 (± 1.5)	83.6 (± 3.6)	84.2 (± 3.8)
	SurveyX	90.2 (± 2.8)	71.2 (± 5.4)	95.6 (± 1.8)	82.8 (± 3.9)	84.5 (± 4.0)
	SurveyForge	94.9 (± 1.7)	**93.1** (± 2.1)	**98.7** (± 1.1)	93.5 (± 2.4)	94.2 (± 2.5)
	SurveyG	**96.2** (± 1.4)	89.4 (± 2.6)	98.6 (± 1.0)	**95.6** (± 2.1)	**96.2** (± 2.3)

Table 1. LLM-as-a-judge evaluation of generated surveys using GPT-4o-mini backbone (mean (± std) over 10 trials per topic). Each LLM evaluates four generation models across content quality dimensions.

2.2.2. Full Paper Completion

In the writing stage, the *Writing Agent* expands each subsection based on the structured outline and its memory, utilizing grounded summaries to ensure factual consistency and contextual relevance. Meanwhile, the *Evaluation Agent* provides critical feedback by offering broader perspectives and generating targeted retrieval queries to identify additional relevant papers from the database D. This iterative collaboration ensures that the final text is coherent, comprehensive, and rigorously supported by the literature. The key novelty lies in combining RAG-based retrieval, guided by the Evaluation Agent's global perspective, with pre-built hierarchical summaries that serve as localized knowledge, enabling the generation of well-balanced and contextually rich subsections.

3. Experiments

3.1. Experimental Setup

3.1.1. Baselines

We compare **SurveyG** with three state-of-the-art systems: **AutoSurvey** (Wang et al.), **SurveyX** (Liang et al., 2025), and **SurveyForge** (Yan et al., 2025). Details on each baseline are provided in Appendix D.1.

3.1.2. Dataset and Ground Truth Construction

We evaluate **SurveyG** on ten diverse computer science topics from the SurGE benchmark (Su et al., 2025), which comprises 205 human-authored surveys and over one million papers. To establish high-quality ground truth, we recruited 20 domain experts: CS Ph.D. students from QS 5-star universities with 3+ years of research experience and senior AI research engineers with publication records in top-tier venues.

The ground-truth selection followed a rigorous three-stage protocol: (1) **Topic Selection**: Ten representative research areas covering diverse subfields; (2) **Survey Selection**: For each topic, 2-3 domain-matched experts independently reviewed available surveys based on four criteria: comprehensive coverage, structural clarity, recency (within 5 years), and citation impact, with disagreements resolved through consensus; (3) **Reference Curation**: Experts identified seminal works, key methodological contributions, and recent advances essential for comprehensive coverage. The same experts served as human evaluators, assessing coherence, coverage, and factual accuracy. Complete details on topic descriptions, selection criteria, scoring rubrics, and inter-annotator agreement are in Appendix D.

3.1.3. Implementation Details

We adopt the experimental setup from prior studies (Yan et al., 2025; Wang et al.; Liang et al., 2025). Specifically, we retrieve 1,500 candidate papers for outline construction, select 60 papers per chapter, and include 300 papers in each final survey. All methods, including SurveyG and the baselines, use *GPT-4o-mini* as the main backbone model and share the same crawled paper database released with SurveyForge. To assess cross-model robustness, we additionally test *Gemini-2.5-Flash*. For evaluation, we generate ten independent runs per topic (100 surveys in total). Four state-of-the-art LLMs: *GPT-4o*, *Claude-4.5-Sonnet*, *DeepSeek-V3.2-Exp*, and *Gemini-2.5-Pro*, are used as judges. We set $T_{\text{MAX}} = 3$ and DeBERTa-v3-large as our main text encoder

Algorithm 2 SurveyG Automated Survey Generation

Input: Survey topic Q, paper database D, max iterations $T_{\max}$, summarizations $\{T_1, \ldots, T_{K+N}\}$
Output: Survey paper F
// Initialization
Create Writing Agent (WA) and Evaluation Agent (EA)
Initialize memory M for WA with $\{T_1, \ldots, T_{K+N}\}$
// Phase 1: Create Outline
WA generates initial outline $\mathcal{O}^{(0)}$ from M
for $t = 1$ **to** $T_{\max}$ **do**
  $\mathcal{O}^{(t)} \leftarrow \text{WA}(M, \text{EA}(\mathcal{O}^{(t-1)}))$
  if quality threshold met **then**
    break
  end if
end for
$\mathcal{O}^* \leftarrow \mathcal{O}^{(t)}$
// Phase 2: Write Full Paper
for subsection $O_i \in \mathcal{O}^*$ **do**
  WA produces an initial draft $O_i^{(0)}$
  for $t = 1$ **to** $T_{\max}$ **do**
    $O_i^{(t)} \leftarrow \text{WA}(M \cup R_i^{(t)}, \text{EA}(O_i^{(t-1)}))$
      where $R_i^{(t)} = \textsc{Retrieve}(Q_i^{(t)}, D)$
    if quality threshold met **then**
      break
    end if
  end for
end for
// Phase 3: Assemble Survey
$F \leftarrow \bigcup_i O_i^{(t)}$
return F

model.

3.1.4. Evaluation Metrics

We evaluate three dimensions: (1) **Outline Quality** following (Yan et al., 2025); (2) **Content Quality** using five metrics (**Coverage**, **Structure**, **Relevance**, **Synthesis**, **Critical Analysis**) rated 0-100 by LLM and human judges; (3) **Citation Quality** measuring factual consistency via NLI, reporting **Citation Recall**, **Precision**, and **F1**. Complete metric definitions and evaluation protocols are in Appendix D.3.

3.2. LLMs-as-a-judge evaluation

3.2.1. Evaluation on Content Quality

As shown in Table 1, SurveyG consistently achieves the highest scores across nearly all metrics and evaluation models, demonstrating strong generalization and robustness. SurveyG maintains low standard deviations (1.3-3.7 points) compared to baselines (1.1-5.4 points), indicating stable performance across topics and judges. Cross-judge analysis reveals remarkable consistency: SurveyG achieves top-2 performance under all four judges and ranks first in 16 out of 20 metric-judge combinations (80%). In contrast, SurveyForge shows variable performance, ranking first under Gemini but dropping to third under Claude's stricter assessment. The particularly great improvements in **Synthesis** (+2.9 to +4.6 points over SurveyForge) and **Critical Analysis** (+1.9 to +7.0 points) reflect the framework's ability to integrate information and identify research gaps through structured reasoning, achieving superior organization and analytical depth without requiring human-written survey inputs.

3.2.2. Evaluation on Ground Truth

Figure 4 reveals distinct performance patterns among the three approaches. In Synthesis, SurveyG achieves the most balanced performance, closely matching human surveys and showing more consistent scores than SurveyForge across metrics like OOD Detection and Hallucination in LLM. For Coverage, while human surveys lead with 90 scores, SurveyG demonstrates more stable cross-topic performance compared to SurveyForge's variable results, particularly in specialized areas like Knowledge Graph Embedding and RL for Language Processing. In Critical Analysis, both automated methods score 70-85, but SurveyG shows less variation between metrics, indicating more reliable quality. Overall, while SurveyForge occasionally peaks higher in individual metrics, SurveyG's consistently uniform polygon shapes across all three dimensions suggest superior robustness and generalization capability for diverse survey generation tasks.

3.2.3. Evaluation on Citation Quality

As shown in Table 3, **SurveyG** achieves superior citation quality with the highest Recall (91.40±1.8) and F1 Score (83.49±2.0), significantly outperforming all baselines. The low standard deviation (1.8-2.0) indicates consistent performance across diverse topics. Notably, SurveyG's Recall approaches the Ground Truth (92.53), demonstrating highly effective identification and linking of relevant literature. While SurveyX achieves the highest Precision (78.12±2.9), SurveyG's substantially better F1 Score confirms its superior balance between citation completeness and accuracy. This improvement stems from the hierarchical citation graph structure, which enables systematic traversal of foundational and frontier papers, ensuring comprehensive yet relevant reference coverage.

3.3. Human Evaluation

To validate our evaluation framework, we conducted win-rate comparisons between SurveyG and the strongest baseline SurveyForge across ten topics, using three metrics: *Score Win Rate* (higher absolute scores), *Comparative Win*

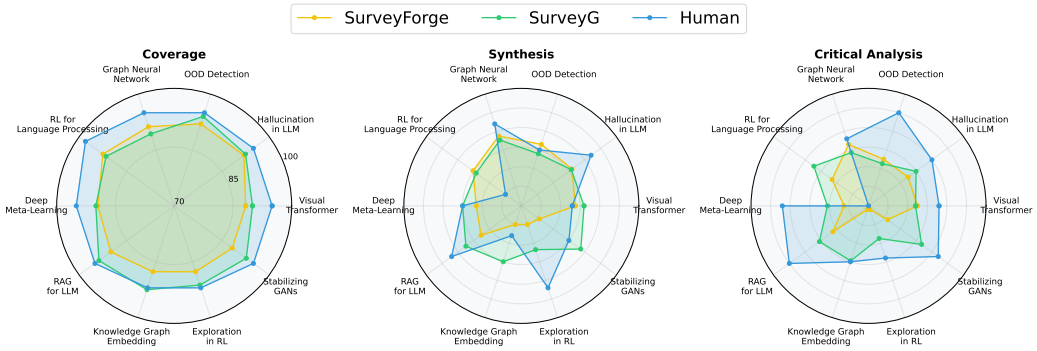

Figure 4. LLM-as-a-judge evaluation of human-written ground-truth surveys, SurveyForge, and SurveyG across ten topics using GPT-4o as the evaluator.

Evaluation Type	Model	Score Win Rate	Comparative Win Rate	Human Eval
Full Paper	SurveyForge	38.85%	27.75%	36.00%
	SurveyG (ours)	**61.15%**	**72.25%**	**64.00%**
Outline	SurveyForge	45.00%	42.00%	45.00%
	SurveyG (ours)	**55.00%**	**58.00%**	**55.00%**

Table 2. Win-rate comparison of SurveyForge and SurveyG on full-paper and outline evaluations. *Score Win Rate* indicates higher absolute evaluation scores, *Comparative Win Rate* reflects pairwise LLM preferences, and *Human Eval* denotes human judgment win rate.

Model	Recall	Precision	F1
AutoSurvey	$82.25_{\pm 2.8}$	$77.41_{\pm 3.2}$	$79.76_{\pm 2.5}$
SurveyForge	$88.34_{\pm 2.1}$	$75.92_{\pm 3.5}$	$81.66_{\pm 2.3}$
SurveyX	$85.23_{\pm 2.6}$	$\mathbf{78.12}_{\pm 2.9}$	$81.52_{\pm 2.4}$
SurveyG (ours)	$\mathbf{91.40}_{\pm 1.8}$	$77.83_{\pm 2.7}$	$\mathbf{83.49}_{\pm 2.0}$
Ground Truth	$92.53_{\pm 1.5}$	$86.42_{\pm 2.1}$	$89.34_{\pm 1.7}$

Table 3. Citation quality comparison across models.

Evaluation Pair	Aspect	Cohen κ
LLM vs. Human	Outline	0.6972
LLM vs. Human	Content	0.6062
Human Cross-Validation	Outline	0.7542
Human Cross-Validation	Content	0.7127

Table 4. Inter-rater agreement between LLMs and humans

Rate (pairwise preferences), and *Human Eval* (expert judgments). As shown in Table 2, SurveyG substantially outperforms SurveyForge in full-paper generation (61.15% vs. 38.85% score win rate, 72.25% vs. 27.75% comparative win rate, 64.00% vs. 36.00% human preference). The particularly strong comparative win rate (72.25%) indicates that judges overwhelmingly prefer SurveyG in direct side-by-side comparisons. For outline generation, SurveyG maintains consistency, demonstrating that the hierarchical citation graph provides effective structured guidance from the earliest generation stage. The strong alignment between automated and human judgments (within 3-8%) validates our LLM-based evaluation framework's reliability while maintaining scalability.

3.3.1. Details of Human Evaluation

To evaluate reliability, we assigned two domain experts to each of the ten SurGE topics. For every topic, we randomly sampled ten anonymized outputs from **SurveyG** and ten from **SurveyForge**, each obtained from independent generation runs. All outputs were fully anonymized, ensuring that neither the human experts nor the LLM judge (*GPT-4o*) was aware of their system of origin. Both experts independently rated all 20 outputs per topic using identical evaluation prompts and criteria, which covered five content metrics **Structure**, **Coverage**, **Relevance**, **Synthesis**, and **Critical Analysis** as well as an outline quality score (0 to 100). For each topic, we computed Cohen's κ to measure (i) agreement between the LLM and human raters, and (ii) agreement between human raters. Table 4 reports topic-wise and average κ values. The mean Cohen's κ for the outline metric was 0.6972 (LLM-human) versus 0.7542 (human-human), and for content metrics, 0.6062 versus 0.7127, respectively. These results demonstrate substantial inter-rater reliability and confirm that the LLM-as-a-judge evaluations align closely with expert assessments.

3.4. Ablation Studies

Table 5 presents ablation results comparing our full model against variants with specific components removed. The full model achieves the best performance across all metrics, demonstrating that each component contributes meaning-

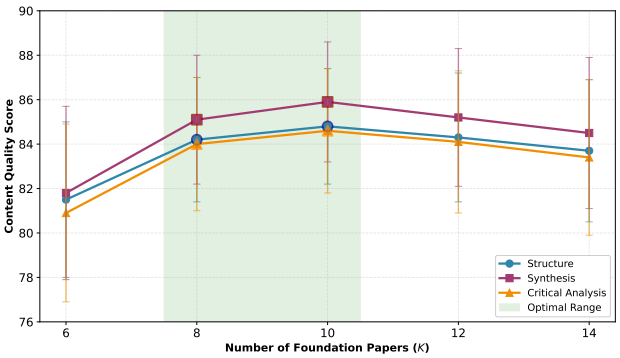

Figure 5. Ablation analysis of the number of foundation papers K.

Variant	Cov.	Str.	Rel.	Syn.	C.A
Full	**95.7**	**88.5**	**95.1**	**92.2**	**90.5**
w/o Vertical Traversal	89.8	84.8	93.9	84.8	84.6
w/o Horizontal Clustering	91.5	85.2	93.6	83.7	86.8
w/o MA	93.2	85.6	91.2	88.7	87.2

Table 5. Ablation study on key components: (1) *w/o Vertical Traversal*: horizontal clustering only; (2) *w/o Horizontal Clustering*: vertical path traversal only; (3) *w/o MA*: no multi-agent refinement. Evaluated by *GPT-4o*.

fully to survey quality. Removing vertical traversal causes the most significant degradation, particularly in Synthesis (-7.4 points) and Critical Analysis (-5.9 points), indicating that cross-layer exploration from foundational to frontier papers is crucial for integrative reasoning and identifying research gaps. The w/o horizontal clustering variant also shows substantial drops in Synthesis (-8.5 points), confirming that within-layer grouping of thematically related papers enables effective information organization. Removing the multi-agent refinement loop (w/o MA) decreases Relevance (-3.9 points) and Structure (-2.9 points), showing that iterative evaluation-agent feedback improves content focus and coherence. These results validate that the hierarchical citation graph's bidirectional traversal mechanisms and multi-agent refinement work synergistically to produce comprehensive, well-structured surveys.

Ablation on number of K. We examine how the number of foundation papers K affects vertical traversal across Structure, Synthesis, and Critical Analysis metrics (Figure 5). Performance is suboptimal at $K = 6$ (scores = 80.9-81.8, std = 3.5-4.0) due to insufficient foundational coverage, improves at $K = 8$ (scores = 84.0-85.1), and peaks at $K = 10$ (scores = 84.6-85.9, std = 2.6-2.8). Synthesis shows the highest sensitivity to K (5.0-point range), as comprehensive foundation-to-frontier paths are critical for cross-paper integration. Optimal K varies by topic breadth: broader fields benefit from $K = 10$, while focused domains achieve comparable results with $K = 8$. Beyond $K = 10$, performance plateaus or declines with increased variance. We recommend $K \in [8, 10]$ based on topic characteristics.

Cost Estimation. The SurveyG framework generates survey papers with an average length of approximately 64k tokens, comparable to expert-written surveys. Each subsection is produced using around 12k input tokens and 800 output tokens. The Evaluation Agent in the RAG loop performs one assessment per subsection, consuming approximately 3.7k input and 700 output tokens. Using GPT-4o-mini pricing ($0.150/1M input tokens, $0.600/1M output tokens), the total LLM cost for generating a full 64k-token survey is $1.5-$1.7, depending on the number of sections and revision iterations. We provide details about computational efficiency in Appendix F.

Additional ablation studies and analysis. We provide additional experiments and analyses in the appendix to further validate SurveyG. Specifically, we compare SurveyG and SurveyForge under different LLM backbones, including GPT-4o-mini and Gemini-2.5-Flash, in Appendix C.1. We analyze the impact of the number of evaluation loops in Appendix C.2. To evaluate generalizability beyond computer science, we conduct cross-domain experiments on biomedical science and AI-for-X topics in Appendix C.3. We also discuss an adaptive strategy for selecting the temporal landmark T across domains with different publication dynamics in Appendix C.4.

4. Conclusion

In this work, we introduced SurveyG, an automated framework for survey generation that leverages hierarchical knowledge representation and multi-agent collaboration to address the limitations of existing LLM-based approaches. By modeling papers through a three-layer citation-similarity graph and employing both horizontal and vertical traversal strategies, SurveyG captures the structural relationships and evolutionary progress of research, enabling the creation of coherent and well-structured outlines. On the SurGE benchmark for autonomous computer science survey generation, both LLM-as-a-judge evaluations and human expert assessments demonstrate that SurveyG outperforms state-of-the-art frameworks across multiple dimensions.

Impact Statement

This paper presents an automated survey generation framework designed to assist researchers in organizing and synthesizing rapidly growing scientific literature. However, generated surveys may still contain factual inaccuracies, incomplete coverage, or biases inherited from retrieved papers and underlying language models. Therefore, SurveyG should be used as an assistive tool rather than a replacement for expert review and critical analysis.

References

Bai, Y., Lv, X., Zhang, J., He, Y., Qi, J., Hou, L., Tang, J., Dong, Y., and Li, J. Longalign: A recipe for long context alignment of large language models, 2024. *URL https://arxiv. org/abs/2401.18058*.

Bai, Y., Lv, X., Zhang, J., Lyu, H., Tang, J., Huang, Z., Du, Z., Liu, X., Zeng, A., Hou, L., et al. Longbench: A bilingual, multitask benchmark for long context understanding. *arXiv preprint arXiv:2308.14508*, 2023.

Dong, Z., Tang, T., Li, J., Zhao, W. X., and Wen, J.-R. Bamboo: A comprehensive benchmark for evaluating long text modeling capacities of large language models. *arXiv preprint arXiv:2309.13345*, 2023.

Erera, S., Shmueli-Scheuer, M., Feigenblat, G., Nakash, O. P., Boni, O., Roitman, H., Cohen, D., Weiner, B., Mass, Y., Rivlin, O., et al. A summarization system for scientific documents. *arXiv preprint arXiv:1908.11152*, 2019.

Gao, T., Yen, H., Yu, J., and Chen, D. Enabling large language models to generate text with citations. *arXiv preprint arXiv:2305.14627*, 2023.

Han, S., Zhang, Q., Yao, Y., Jin, W., and Xu, Z. Llm multi-agent systems: Challenges and open problems. *arXiv preprint arXiv:2402.03578*, 2024.

Huynh-The, T., Pham, Q.-V., Pham, X.-Q., Nguyen, T. T., Han, Z., and Kim, D.-S. Artificial intelligence for the metaverse: A survey. *Engineering Applications of Artificial Intelligence*, 117:105581, 2023.

Li, Y., Chen, L., Liu, A., Yu, K., and Wen, L. Chatcite: Llm agent with human workflow guidance for comparative literature summary. *arXiv preprint arXiv:2403.02574*, 2024.

Liang, X., Yang, J., Wang, Y., Tang, C., Zheng, Z., Song, S., Lin, Z., Yang, Y., Niu, S., Wang, H., et al. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*, 2025.

Su, W., Xie, A., Ai, Q., Long, J., Mao, J., Ye, Z., and Liu, Y. Benchmarking computer science survey generation. *arXiv preprint arXiv:2508.15658*, 2025.

Sumers, T., Yao, S., Narasimhan, K., and Griffiths, T. Cognitive architectures for language agents. *Transactions on Machine Learning Research*, 2023.

Susnjak, T., Hwang, P., Reyes, N., Barczak, A. L., McIntosh, T., and Ranathunga, S. Automating research synthesis with domain-specific large language model fine-tuning. *ACM Transactions on Knowledge Discovery from Data*, 19(3):1–39, 2025.

Traag, V. A., Waltman, L., and Van Eck, N. J. From louvain to leiden: guaranteeing well-connected communities. *Scientific reports*, 9(1):1–12, 2019.

Wang, L., Xu, W., Lan, Y., Hu, Z., Lan, Y., Lee, R. K.-W., and Lim, E.-P. Plan-and-solve prompting: Improving zero-shot chain-of-thought reasoning by large language models. *arXiv preprint arXiv:2305.04091*, 2023.

Wang, Y., Guo, Q., Yao, W., Zhang, H., Zhang, X., Wu, Z., Zhang, M., Dai, X., Zhang, M., Wen, Q., et al. Autosurvey: Large language models can automatically write surveys, 2024. *URL https://arxiv. org/abs/2406.10252*.

Wen, Z., Cao, J., Wang, Z., Guo, B., Yang, R., and Liu, S. Interactivesurvey: An llm-based personalized and interactive survey paper generation system. *arXiv preprint arXiv:2504.08762*, 2025.

Wu, Q., Bansal, G., Zhang, J., Wu, Y., Li, B., Zhu, E., Jiang, L., Zhang, X., Zhang, S., Liu, J., et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First Conference on Language Modeling*, 2024.

Wu, Y., Sun, Z., Li, S., Welleck, S., and Yang, Y. Inference scaling laws: An empirical analysis of compute-optimal inference for llm problem-solving. In *The Thirteenth International Conference on Learning Representations*, 2025.

Xu, P., Ping, W., Wu, X., McAfee, L., Zhu, C., Liu, Z., Subramanian, S., Bakhturina, E., Shoeybi, M., and Catanzaro, B. Retrieval meets long context large language models. *arXiv preprint arXiv:2310.03025*, 2023.

Yan, X., Feng, S., Yuan, J., Xia, R., Wang, B., Zhang, B., and Bai, L. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629*, 2025.

Zhang, Y., Sun, R., Chen, Y., Pfister, T., Zhang, R., and Arik, S. Chain of agents: Large language models collaborating on long-context tasks. *Advances in Neural Information Processing Systems*, 37:132208–132237, 2024.

Zhang, Y., Yang, R., Jiao, S., Kang, S., and Han, J. Scientific paper retrieval with llm-guided semantic-based ranking. *arXiv preprint arXiv:2505.21815*, 2025.

Create a comprehensive literature review outline based on the following taxonomy summaries for three layers (Foundation, Development, Frontier) and vertical directions.
Horizontal summary: {summary_layer}.
Vertical direction: {summary_path}.
Based on the above information, create a detailed outline for a literature review paper, organizing it into sections and subsections. **Respond with the outline in JSON format with keys:**
['section_outline', 'subsection_focus', 'proof_ids'].
For each section: Put section and subsection titles in 'section_outline'; Add a paragraph in 'subsection_focus' describing the main focus of each subsection; Add 'proof_ids' from either taxonomy layer or vertical direction.

Figure 6. Structured outline creation short version prompt.

Algorithm 3 Weighted Breadth-First Search (WBFS)

Input: Start node s, target layer ℓ
Output: Set of nodes R in layer ℓ
$visited \leftarrow \{s\}$, $queue \leftarrow [s]$, $R \leftarrow \emptyset$
while $queue \neq \emptyset$ **do**
  $u \leftarrow queue.\text{DEQUEUE}()$
  for $v \in \text{SUCCESSORS}(u)$ **sorted by** $weight(u, v)$ **desc do**
    if $v \notin visited$ **then**
      $visited \leftarrow visited \cup \{v\}$
      if $v.layer = \ell$ **then**
        $R \leftarrow R \cup \{v\}$
      else
        $queue.\text{ENQUEUE}(v)$
      end if
    end if
  end for
end while
return R

A. Related works

A.1. Long-form Text Generation

LLMs have achieved remarkable progress, yet generating long-form, coherent, and logically structured documents remains a persistent challenge (Bai et al., 2023; Dong et al., 2023; Han et al., 2024). Recent works have explored different strategies to address the long-context problem. For example, Chain-of-Agents (Zhang et al., 2024) introduces a multi-agent collaboration framework where worker agents process segmented portions of text and a manager agent synthesizes them into coherent outputs, alleviating focus issues in long contexts. LongAlign (Bai et al.) proposes a recipe for long context alignment, combining instruction data construction, efficient batching, yielding strong gains on queries up to 100k tokens. Complementary to these, Xu et al. (Xu et al., 2023) systematically examine the trade-offs between retrieval-augmentation and context-window extension, showing that hybrid approaches can outperform both strategies alone. However, existing approaches often rely on raw reference texts, leading to inefficient retrieval, limited context utilization, and poor structural coherence in survey-like outputs.

A.2. Automatic Survey Generation

The automatic generation of literature reviews has been studied for over a decade, starting with multi-document summarization techniques that produced unstructured related work sections. Early systems, such as IBM Science Summarizer (Erera et al., 2019), focused on summarizing scientific articles, while more recent LLM-based methods like ChatCite (Li et al., 2024) and Susnjak et al. (Susnjak et al., 2025)'s domain-specific fine-tuning advanced the generation of comparative and knowledge-enriched reviews. Despite these advances, such methods primarily tackle summarization rather than the creation

Topic	Ground Truth Survey	Citation
Visual Transformer	A Survey of Visual Transformers	405
Hallucination in Large Language Models	Siren's Song in the AI Ocean: A Survey on Hallucination in LLMs	808
Graph Neural Networks	Graph Neural Networks: Taxonomy, Advances, and Trends	129
Deep Meta-Learning	A Survey of Deep Meta-Learning	459
Knowledge Graph Embedding	Knowledge graph embedding: A survey from the perspective of representation spaces	130
Generalized Out-of-Distribution Detection	Generalized Out-of-Distribution Detection: A Survey	1406
Reinforcement Learning for Language Processing	Survey on reinforcement learning for language processing	206
Exploration Methods in Reinforcement Learning	Exploration in Deep Reinforcement Learning: From Single-Agent to Multi-Agent Domain	194
Stabilizing Generative Adversarial Networks	Stabilizing Generative Adversarial Networks: A Survey	149
Retrieval-Augmented Generation for LLMs	Retrieval-Augmented Generation for Large Language Models: A Survey	953

Table 6. Survey Papers Overview

of fully structured survey papers. More recent systems, including AutoSurvey (Wang et al.), InteractiveSurvey (Wen et al., 2025), SurveyForge (Yan et al., 2025), and SurveyX (Liang et al., 2025), propose end-to-end pipelines integrating RAG, clustering, or multi-agent strategies to automate survey construction. These methods improve structural coherence and formatting consistency while scaling to long-form survey content. Nevertheless, most frameworks still restrict users to fixed input-output modes, overlooking relationships among papers and limiting interactivity, which often results in surveys that lack flexibility, relational awareness, and depth.

B. Details about Traversal on Graph

We provide a full algorithm of Weighted BFS in Algorithm 3.

C. More ablation study results

LLM	Model	Backbone	Coverage	Structure	Relevance	Synthesis	Critical Analysis
Claude	SurveyForge	GPT-4o-mini	81.8 (±2.9)	78.4 (±3.5)	89.1 (±2.1)	75.4 (±3.8)	70.3 (±4.2)
	SurveyForge	Gemini-2.5-Flash	82.5 (±3.1)	79.2 (±3.7)	89.8 (±2.3)	76.8 (±3.9)	72.1 (±4.3)
	SurveyG	GPT-4o-mini	**88.1** (±2.4)	87.9 (±2.7)	**93.6** (±1.9)	80.2 (±3.2)	**77.3** (±3.5)
	SurveyG	Gemini-2.5-Flash	87.5, (±2.6)	**88.6** (±2.5)	93, 2 (±2.0)	**81.4** (±3.1)	76.9 (±3.6)
GPT	SurveyForge	GPT-4o-mini	94.2 (±1.8)	87.3 (±2.5)	94.8 (±1.6)	88.6 (±2.7)	88.5 (±2.9)
	SurveyForge	Gemini-2.5-Flash	94.8 (±1.9)	87.9 (±2.6)	95.2 (±1.5)	89.4 (±2.8)	89.3 (±3.0)
	SurveyG	GPT-4o-mini	95.7 (±1.5)	**88.5** (±2.2)	95.1 (±1.4)	**92.2** (±2.3)	**91.2** (±2.6)
	SurveyG	Gemini-2.5-Flash	**96.3** (±1.4)	88.2, (±2.4)	**95.7** (±1.3)	91.8 (±2.5)	90.5 (±2.5)
Deepseek	SurveyForge	GPT-4o-mini	89.4 (±2.3)	86.5 (±2.9)	92.5 (±2.0)	84.1 (±3.4)	80.7 (±3.7)
	SurveyForge	Gemini-2.5-Flash	88.9 (±2.5)	85.8 (±3.1)	93.1 (±2.1)	85.3 (±3.5)	81.5 (±3.8)
	SurveyG	GPT-4o-mini	88.7 (±2.5)	85.7 (±3.0)	**94.2** (±1.8)	86.3 (±3.1)	**82.7** (±3.5)
	SurveyG	Gemini-2.5-Flash	**89.1** (±2.4)	**86.1** (±2.8)	93.8 (±1.9)	**86.8** (±3.0)	82.3 (±3.7)
Gemini	SurveyForge	GPT-4o-mini	94.9 (±1.7)	93.1 (±2.1)	98.7 (±1.1)	93.5 (±2.4)	94.2 (±2.5)
	SurveyForge	Gemini-2.5-Flash	95.4 (±1.6)	91.8 (±2.3)	98.3 (±1.2)	94.2 (±2.5)	95.1 (±2.4)
	SurveyG	GPT-4o-mini	96.2 (±1.4)	89.4 (±2.6)	**98.6** (±1.0)	**95.6** (±2.1)	96.2 (±2.3)
	SurveyG	Gemini-2.5-Flash	**96.5** (±1.3)	**90.1** (±2.3)	98.4 (±1.1)	95.2 (±2.3)	**96.7** (±2.1)

Table 7. Backbone comparison: SurveyForge and SurveyG with GPT-4o-mini vs. Gemini-2.5-Flash (mean (± std) over 10 trials per topic). Each LLM evaluates both models with different backbones across content quality dimensions.

C.1. Ablation on Backbone Models

Table 7 presents a systematic comparison of backbone models to isolate framework contributions from LLM capability effects. We evaluate both SurveyForge (current SOTA) and SurveyG with two different backbones: GPT-4o-mini and Gemini-2.5-Flash. The results demonstrate that backbone selection has minimal impact on relative performance rankings.

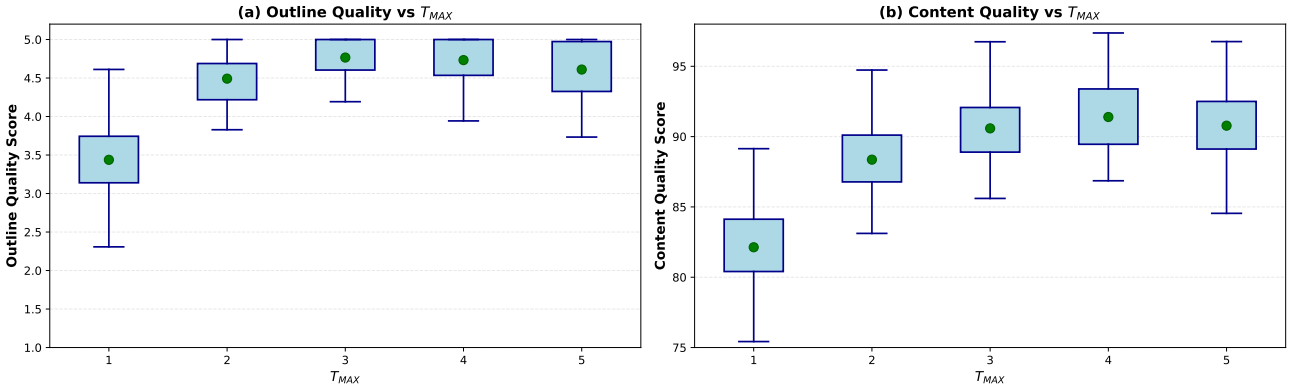

Figure 7. Ablation analysis of the number of evaluation loop T_{MAX}.

For SurveyForge, switching backbones yields marginal differences (average ±0.5-0.8 points across metrics), while SurveyG shows similarly stable performance (average variation <0.6 points). Critically, SurveyG maintains consistent superiority over SurveyForge regardless of backbone choice: under Claude's evaluation, SurveyG outperforms SurveyForge by +5.0 to +6.3 points in Coverage and +7.7 to +9.4 points in Structure across both backbones. This pattern holds across all four judges, with SurveyG achieving higher scores in 18 out of 20 model-metric combinations (90%). The consistency of performance gaps across different backbones confirms that SurveyG's improvements stem from its hierarchical citation graph and multi-level summarization framework rather than simply leveraging more capable foundation models. These findings validate that our architectural innovations, not backbone selection, drive the observed quality improvements in automated survey generation.

C.2. Ablation on the number of T_{MAX}

We analyze the effect of the maximum number of evaluation iterations, T_{MAX}, on generation quality (Figure 7; prompt details in Figure 9). For outline generation, the mean score rises from 3.5 (std = 0.6) at $T_{\text{MAX}} = 1$ to 4.8 (std = 0.30) at $T_{\text{MAX}} = 3$, indicating that iterative feedback effectively resolves structural deficiencies. A similar trend is observed for section content quality, measured as the average of five metrics, which increases from 82.5 (std = 3.2) to 90.6 (std = 2.4), reaching full-model performance at $T_{\text{MAX}} = 3$. Further increasing T_{MAX} yields only marginal gains (¡0.2 points) alongside higher variance, suggesting diminishing returns. Accordingly, we adopt $T_{\text{MAX}} = 3$ as the default setting to balance quality, robustness, and computational cost.

C.3. Cross-Domain Generalization

To further evaluate the generalizability of SurveyG beyond computer science literature, we conducted additional experiments on six topics spanning biomedical science and AI-for-X domains: *Tumor Microenvironment*, *Drug Repurposing in Cancer*, *CRISPR and Gene Editing*, *AI for Mathematics*, *AI for Clinical Trials*, and *AI for Software Development*. Importantly, all experiments used the same SurveyG pipeline and prompts without any domain-specific modifications. Following the evaluation protocol in the main paper, we performed 10 independent runs per topic and used Claude-Sonnet-4.5 with temperature $= 0$ as the judge model.

Results in Table 8 show that SurveyG generalizes effectively across diverse research domains. In several cases, SurveyG even surpasses the reference survey papers on specific dimensions. For example, on *Tumor Microenvironment*, SurveyG achieves higher scores in *Structure* and *Critical Analysis*; on *Drug Repurposing in Cancer*, it achieves higher *Synthesis*; and on *AI for Clinical Trials*, it outperforms the reference survey in both *Structure* and *Critical Analysis*. Across all six topics, SurveyG maintains relatively low standard deviations, demonstrating stable behavior despite differences in citation patterns and writing conventions across domains. These findings suggest that the hierarchical citation graph and traversal strategies capture general organizational properties of scholarly literature rather than being tailored specifically to computer science.

Ground Truth Surveys. The reference surveys used in this evaluation are listed below:

1. Arneth, Borros. "Tumor microenvironment." *Medicina* 56.1 (2019): 15.

Topic	Backbone	Coverage	Structure	Relevance	Synthesis	Critical Analysis
Tumor Microenvironment	SurveyG	79.6±2.1	78.2±1.1	86.3±2.5	74.9±3.2	70.4±3.2
	Ground Truth	80	76	85	68	65
Drug Repurposing in Cancer	SurveyG	84.7±2.6	77.6±2.2	83.4±2.6	78.4±2.5	74.9±2.1
	Ground Truth	92	88	95	85	82
CRISPR and Gene Editing	SurveyG	72.7±3.2	70.4±2.8	85.2±2.7	74.3±2.4	71.5±3.2
	Ground Truth	78	72	85	70	68
AI for Mathematics	SurveyG	76.9±2.5	72.4±2.3	85.1±3.2	73.4±2.6	70.1±3.2
	Ground Truth	85	88	92	82	86
AI for Clinical Trials	SurveyG	75.7±2.9	78.4±2.6	87.6±2.6	68.5±2.7	70.1±2.3
	Ground Truth	85	78	92	72	68
AI for Software Development	SurveyG	72.3±2.4	68.9±2.4	85.6±2.3	75.4±2.6	78.2±2.8
	Ground Truth	75	82	88	70	68

Table 8. Cross-domain evaluation of SurveyG on biomedical science and AI-for-X topics. Results are averaged over 10 independent runs.

Year	KGE	RL-LP	Hallucination	Exploration RL
2014-2019	93	-	5	25
2020	40	-	4	25
2021	55	-	8	22
2022	49	-	19	57
2023	39	34	51	28
2024	95	81	172	71
2025	16	133	68	42

Table 9. Publication year distributions for representative survey topics. Different topics exhibit substantially different temporal dynamics, motivating adaptive selection of the temporal landmark T.

2. Xia, Ying, et al. "Drug repurposing for cancer therapy." *Signal Transduction and Targeted Therapy* 9.1 (2024): 92.

3. Hirakawa, Matthew P., et al. "Gene editing and CRISPR in the clinic: current and future perspectives." *Bioscience Reports* 40.4 (2020): BSR20200127.

4. Klowden, Tanya, and Terence Tao. "Mathematical methods and human thought in the age of AI." *arXiv preprint arXiv:2603.26524* (2026).

5. Olawade, David B., et al. "Artificial intelligence in clinical trials: A comprehensive review of opportunities, challenges, and future directions." *International Journal of Medical Informatics* (2025): 106141.

6. Esposito, Matteo, et al. "Generative AI for software architecture: Applications, challenges, and future directions." *Journal of Systems and Software* (2025): 112607.

C.4. Adaptive Selection of Temporal Landmark T

SurveyG determines the time landmark T adaptively for each topic based on the publication year distribution of the retrieved papers. Specifically, we set T as the 75th percentile of the publication years in the crawled corpus. This strategy ensures that the *Frontier* layer consistently captures the most recent and emerging works relative to the historical evolution of each field, rather than relying on a manually selected calendar year.

Table 9 illustrates the publication year distributions for four representative topics. The distributions exhibit substantially different temporal dynamics. For instance, *Knowledge Graph Embedding* contains a mature body of literature distributed across 2014-2019, whereas *Hallucination in LLMs* is dominated by papers published in 2024 and 2025. Similarly, *RL for Language Processing* shows a sharp publication surge only from 2023 onward. These observations demonstrate why a fixed threshold would be inappropriate across domains. The adaptive strategy for selecting T is lightweight, requires no additional LLM calls, and generalizes naturally to topics with very different maturation rates.

D. Experimental Setting

D.1. Baseline Systems

AutoSurvey (Wang et al.) represents an early multi-stage framework for automated survey generation that combines literature retrieval with structured content synthesis. **SurveyX** (Liang et al., 2025) enhances AutoSurvey through structured knowledge extraction and outline optimization, introducing refinement mechanisms for improved coherence. **SurveyForge** (Yan et al., 2025) leverages human-written survey papers from related domains as prior knowledge for heuristic outline generation, guided by a memory-driven scholar navigation agent that retrieves high-quality references for composing new surveys.

D.2. Survey Topics and Ground Truth Construction

We compiled ten representative survey topics covering diverse research areas in machine learning and natural language processing, as summarized in Table 6. Each topic reflects an active line of inquiry, providing a strong foundation for evaluating automated survey generation.

D.2.1. Expert Recruitment and Qualifications

Our 20 domain experts comprised:

- 12 CS Ph.D. students from QS 5-star universities with 3-5 years of research experience
- 8 senior AI research engineers from leading tech companies with 2+ publications in top-tier venues (NeurIPS, ICML, ICLR, ACL, EMNLP, KDD, etc.)

Each expert was matched to topics within their domain expertise, ensuring a qualified assessment.

D.2.2. Three-Stage Ground Truth Selection Protocol

Stage 1: Topic Selection. We identified ten topics ensuring coverage across diverse AI/ML research areas: (a) **foundational architectures** (Visual Transformer, Graph Neural Networks, Generative Adversarial Networks), (b) **learning paradigms** (Deep Meta-Learning, Reinforcement Learning methods), (c) **knowledge representation and retrieval** (Knowledge Graph Embedding, Retrieval-Augmented Generation), and (d) **reliability and robustness** (Hallucination in LLMs, Generalized Out-of-Distribution Detection). This selection spans established methodologies, emerging challenges, and cross-domain techniques, ensuring comprehensive evaluation across theoretical foundations, architectural innovations, and practical applications. As shown in Table 6, selected ground-truth surveys are highly cited (mean: 484 citations), indicating strong community recognition.

Stage 2: Survey Selection Criteria. For each topic, 2-3 domain-matched experts independently evaluated candidate surveys from SurGE using four weighted criteria:

1. **Comprehensive Coverage (40%)**: Survey thoroughly discusses field evolution, seminal works, major methodological branches, and current trends
2. **Structural Clarity (25%)**: Well-organized sections with clear taxonomy, logical flow, and effective use of figures/tables
3. **Recency (20%)**: Published within the last 5 years (2019-2024) to reflect the current research landscape
4. **Citation Impact (15%)**: Highly-cited work (top 20% within topic area) indicating community recognition

Experts scored each criterion 0-10, with surveys scoring ≥ 7.0 average advancing to consensus discussion. Inter-annotator agreement (Fleiss's κ) averaged 0.73 across topics, indicating substantial agreement.

Stage 3: Reference Curation. Selected experts curated 30-50 essential papers per topic, categorized as: foundational works (seminal papers establishing the field), methodological milestones (key technical innovations), and recent advances (papers from the last 2 years). This reference set serves as ground truth for coverage evaluation.

D.3. Evaluation Metrics

D.3.1. Outline Quality

Following (Yan et al., 2025), we evaluate outline structure, logical coherence, topic coverage, and hierarchical organization on a 0-100 scale. The evaluation prompt assesses whether the outline provides a clear roadmap for comprehensive survey development.

D.3.2. Content Quality Metrics

A well-structured outline is essential for maintaining clarity and organization. Full paper evaluation serves as a comprehensive benchmark to assess academic rigor and practical utility. Following (Wang et al.; Liang et al., 2025; Yan et al., 2025; Su et al., 2025), we assess five metrics on a 0-100 scale:

Coverage measures how thoroughly the survey captures major concepts, foundational works, and emerging trends within the research area.

Structure examines logical organization, section coherence, and taxonomy quality, ensuring readers can navigate the field systematically.

Relevance assesses alignment of content with the target research topic, penalizing off-topic discussions or tangential references.

Synthesis evaluates integration of information from multiple sources into a cohesive, non-redundant narrative that reveals cross-paper insights.

Critical Analysis reflects the survey's ability to identify methodological gaps, highlight comparative trends, and articulate open research challenges.

Each metric is scored independently by both LLM judges and human experts, with final scores averaged across all evaluators.

D.3.3. Citation Quality Metrics

Following (Wang et al.; Gao et al., 2023), we measure citation accuracy and contextual relevance. We extract factual claims from generated surveys and verify support using a Natural Language Inference (NLI) model (DeBERTa-v3-large fine-tuned on ANLI).

Citation Recall: Proportion of claims correctly supported by valid references, measuring whether surveys provide adequate citation support.

Citation Precision: Proportion of cited references that truly substantiate their associated claims, measuring citation accuracy.

Citation F1: Harmonic mean of Recall and Precision, providing a balanced measure of citation quality.

For each survey, we sample 100 claim-citation pairs and compute metrics across all samples.

E. Prompt Templates

This section presents the prompt templates designed to guide each stage of automated literature review generation and evaluation. Each template specifies goals, inputs, and evaluation criteria to ensure consistency and quality across generated outputs.

E.1. Prompt to generate structured outline

We provide a short version of the prompt template (Figure 8) that instructs the model to construct a coherent, hierarchical outline that captures the logical flow of a literature review topic before detailed writing begins.

E.2. Prompt to evaluate structured outline

The prompt in Figure 9 guides the model to write complete, citation-based literature review subsections grounded in the provided focus, summaries, and development directions. The following evaluation prompt extends this process to assess

Goal: Generate a structured Literature Review Outline for: "`[QUERY]`"

INPUT SYNTHESIS DATA

- **Communities**: `[PAPER_COMMUNITIES]`
- **Directions**: `[DEVELOPMENT_DIRECTIONS]`

REQUIREMENTS & CONSTRAINTS

1. Structure:

- **Progression**: Follow Foundations → Core → Advanced → Applications → Future.
- **Mandatory Sections**: Must include Introduction, Foundational Concepts, and Conclusion.
- **Hierarchy**: Use exactly **TWO levels** (e.g., 2.1, 2.2). No deeper nesting.

2. Content & Quality:

- Create a **coherent narrative** (evolutionary story, not a list).
- Group material by **methodological families** and thematic depth.
- Include dedicated sections for Applications and Future Trends/Challenges.

3. Evidence & Output:

- **Proof IDs**: Each subsection **MUST** be grounded with 1-3 `proof_ids` (from `layer`, `community_X`, or `seed` IDs).
- **Focus Synthesis**: Provide `section_focus` (broad theme) and `subsection_focus` (specific details).
- **Format**: Return only a **JSON ARRAY** of main sections and their hierarchical subsections.

Figure 8. Generate Outline Prompt.

individual sections for depth, synthesis, and analytical quality.

E.3. Prompt to generate subsections

This prompt guides the model to write complete, citation-based literature review subsections grounded in the provided focus, summaries, and development directions (Figure 10).

E.4. Improve Section Quality

As shown in Figure 11, this prompt systematically assesses literature review sections across multiple dimensions, such as content coverage, synthesis, and critical analysis, while offering actionable feedback and retrieval suggestions for refinement.

F. Computational Efficiency

We analyze computational overhead for the hierarchical citation graph operations on a computer with setting: 1 NVIDIA RTX 4090, 24GB VRAM:

Paper Retrieval and Graph Construction: For a typical query retrieving 1,500 candidate papers, the process includes: (1) keyword expansion via LLM (2-3 seconds, $0.02), (2) Semantic Scholar API calls (15-20 seconds, free), (3) citation network construction (8-12 seconds), and (4) embedding generation using Deberta (45-60 seconds on CPU, negligible GPU cost if available). Total: ∼70-95 seconds, $0.02.

Evaluate the quality and structure of the following literature review outline. Assess whether the outline demonstrates meaningful organization of works rather than a simple concatenation of summaries. **Your feedback should include:**

- Strengths of the outline.
- Weaknesses or issues (if any).
- Specific suggestions for improvement (only if issues are found).
- Final score (1-5, with 5 being the maximum) evaluate overall organization, coherence, and coverage.

Outline to evaluate: {outline_text}

Figure 9. Prompt to evaluate structured outline

Hierarchical Layer Assignment: Organizing 300 papers into Foundation/Development/Frontier layers using citation patterns and temporal analysis requires 5-8 seconds (pure algorithmic processing, no API calls).

Multi-Aspect Summarization: Generating layer-wise and cross-layer summaries involves: (1) horizontal clustering within layers (10-15 seconds for Leiden algorithm), (2) vertical path traversal (3-5 seconds), and (3) LLM-based summarization of $\sim$15-20 paper groups ($0.80-$1.20 depending on summary length). Total: $\sim$15-25 seconds computation, $0.80-$1.20 LLM cost.

End-to-End Latency: The complete pipeline for generating one survey consists of: (1) Crawled data: 20-25 minutes, (2) Graph construction and traversal: 90-128 seconds, (2) Outline generation: 120-180 seconds, (3) Full survey writing: 5-10 minutes (parallel generation of 8-12 sections with evaluation loops). **Total end-to-end latency: 28-38 minutes per survey**, with the majority of time spent on crawled data rather than graph operations or content generation.

G. Case studies

We provided a subsection generated by **SurveyG** (Figure 12) to illustrate its ability to synthesize complex research trends in modular and agentic RAG. Overall, this subsection highlights a clear progression in the RAG landscape from simple retrieval pipelines toward multi-stage, agentic, and modular architectures. The discussed works collectively show how LLMs are evolving from passive generators to proactive reasoning agents capable of planning, coordination, and self-optimization. The emergence of meta-frameworks such as AutoRAG and FlashRAG further reflects a shift toward automated orchestration of RAG components, underscoring a broader trend toward unified, adaptive systems that integrate retrieval and reasoning for scalable knowledge synthesis.

Task: Write a comprehensive literature review subsection titled `[SUBSECTION_TITLE]` in LaTeX.
Inputs:

- **Focus:** [SUBSECTION_FOCUS]
- **Community summaries:** [COMMUNITY_SUMMARY]
- **Development directions:** [DEVELOPMENT_DIRECTION]
- **Papers (chronological):** [PAPER_INFO]

Guidelines:

- Use LaTeX format with citations (`\cite{citation_key}`).
- Minimum 400 words, no numbered subsection titles.
- Focus strictly on the subsection topic.
- Each paper: 2-3 sentences describing technical contributions.
- Connect papers by showing how later work addresses earlier limitations.
- Conclude with remaining challenges or future directions.

Avoid: sequential listing, vague critiques, unsupported claims, isolated descriptions, or ignoring contradictions.

Figure 10. Generate Subsection Prompt

Evaluate the quality of the following literature review section within the context of the overall survey outline. Your evaluation should address the following aspects, each rated from 1-5 (5 = excellent): (1) content coverage, (2) citation density, (3) academic rigor, (4) synthesis across works, (5) critical analysis, (6) coherence, (7) depth of discussion, and (8) specificity of scope.
In addition to numeric ratings, provide:
- A brief natural language summary of the section's strengths and weaknesses,
- An overall score (1-5) with justification,
- Suggestions for improvement, focusing on areas that fall short,
- A list of search queries that could retrieve additional relevant literature to strengthen the section.

Figure 11. Section Quality and Retrieval Prompt

Multi-stage and Modular RAG Frameworks

The foundational paradigm of Retrieval-Augmented Generation (RAG) typically operates on a straightforward ”retrieve-then-generate” sequence [lewis2020pwr]. However, as Large Language Models (LLMs) are increasingly tasked with complex, multi-faceted queries and dynamic information needs, this simple pipeline proves insufficient [huang2024a59, zhao2024931]. This has spurred the evolution of RAG into more sophisticated, multi-stage, and modular architectures, where the LLM transcends a passive role to become an intelligent agent capable of proactive planning, dynamic decision-making, and the orchestration of various sub-tasks [gao20238ea]. This section focuses on frameworks that empower LLMs to actively manage the information-seeking process through iterative planning, query decomposition, and the dynamic assembly of specialized modules. It is crucial to distinguish these proactive, agentic approaches from reactive or corrective mechanisms (e.g., self-correction, re-ranking) that primarily refine retrieval quality, which are discussed in detail in Section 3.

A significant advancement in modular RAG involves empowering LLMs to act as sophisticated planning agents, iteratively refining their information-seeking process and orchestrating multi-round interactions. [lee2024hif] introduced PlanRAG, which extends the popular ReAct framework by incorporating explicit ”Plan” and ”Re-plan” steps. This allows LLMs to dynamically generate and iteratively refine analytical approaches based on intermediate retrieval results, effectively acting as decision-makers for complex data analysis tasks. Similarly, [yang20243nb] presented IM-RAG, a multi-round RAG system that leverages learned inner monologues and a multi-agent reinforcement learning approach. In IM-RAG, an LLM-based ”Reasoner” dynamically switches between a ”Questioner” role (crafting queries) and an ”Answerer” role, guided by mid-step rewards from a ”Progress Tracker,” leading to flexible and interpretable multi-round information gathering. Building on the concept of autonomous interaction, [yu2024c32]’s Auto-RAG enables LLMs to engage in multi-turn dialogues with the retriever, systematically planning retrievals and refining queries until sufficient external information is gathered. This framework highlights the LLM’s powerful decision-making capabilities, autonomously adjusting iterations based on query difficulty and knowledge utility. Another approach, [wang2024zt3]’s M-RAG, proposes a multi-partition paradigm for external memories, employing a multi-agent reinforcement learning framework with an ”Agent-S” for dynamic partition selection and an ”Agent-R” for memory refinement. This enables more fine-grained and focused retrieval by orchestrating memory access across different knowledge partitions. To further optimize the interaction between these modular components, [li20243nz]’s RAG-DDR (Differentiable Data Rewards) offers an end-to-end training method that aligns data preferences between different RAG modules (agents). By collecting rewards and evaluating the impact of perturbations on the entire system, RAG-DDR optimizes agents to produce outputs that enhance overall RAG performance, particularly for smaller LLMs. These agentic frameworks collectively transform RAG into a dynamic, adaptive system capable of tackling complex, multi-hop queries that require sophisticated reasoning and iterative information synthesis.

...

In conclusion, the evolution towards multi-stage and modular RAG frameworks marks a significant advancement, transforming RAG from a simple pipeline into an intelligent, adaptive system. By enabling LLMs to engage in iterative refinement, agentic planning, and dynamic orchestration of sub-tasks, these architectures enhance robustness, reduce hallucinations, and improve the depth and faithfulness of generated responses, particularly for complex, multi-hop queries [tang2024i5r]. However, this sophistication often introduces challenges related to increased computational overhead, the complexity of orchestrating multiple modules, and the need for robust evaluation methodologies that can accurately assess the contributions of each stage and the overall system performance. Benchmarks like [friel20241ct]’s RAGBench, [krishna2024qsh]’s FRAMES, and [tang2024i5r]’s MultiHop-RAG highlight these challenges, emphasizing the need for explainable metrics and unified frameworks to evaluate the intricate interplay of retrieval, reasoning, and generation in these advanced systems. Future research will likely focus on optimizing the efficiency of these multi-stage processes, developing more autonomous and self-correcting agents, and creating more generalized frameworks that can seamlessly integrate diverse knowledge sources and reasoning paradigms while addressing the inherent trade-offs between complexity and efficiency.

Figure 12. Case studies about the result of the generated subsection.