Feedback Is The Key for Automated Survey Generation

Tianyi Xu[1*] **Zhe Zhao**[1*] **Tianshuo Wei**[1*] **Yiqun Kou**[1,2] **Liuliu Han**[3] **Ye Wei**[1†]
[1]City University of Hong Kong
[2]Huazhong University of Science and Technology
[3]Suzhou Laboratory

Abstract

The escalating demand for comprehensive literature surveys in rapidly evolving research areas makes manual writing increasingly impractical, underscoring the necessity of automation. Large Language Models (LLMs) provide a promising foundation for this task, yet guiding them to generate accurate, reliable content remains a fundamental challenge, as issues such as hallucinations and vague organization often persist. To address this, we propose **FIKSurvey**, a feedback-driven framework grounded in the idea that "***F**eedback **i**s the **k**ey for automatic **s**urvey generation*." Specifically, FIKSurvey systematically incorporates feedback across three dimensions: outline feedback for structural clarity, citation feedback for evidence validation, and content feedback for readability and analytical depth. The framework also supports optional human-in-the-loop intervention for user-specific needs. Experiments confirm that FIKSurvey substantially improves both citation and content quality, demonstrating feedback as the critical mechanism for automatic survey generation.

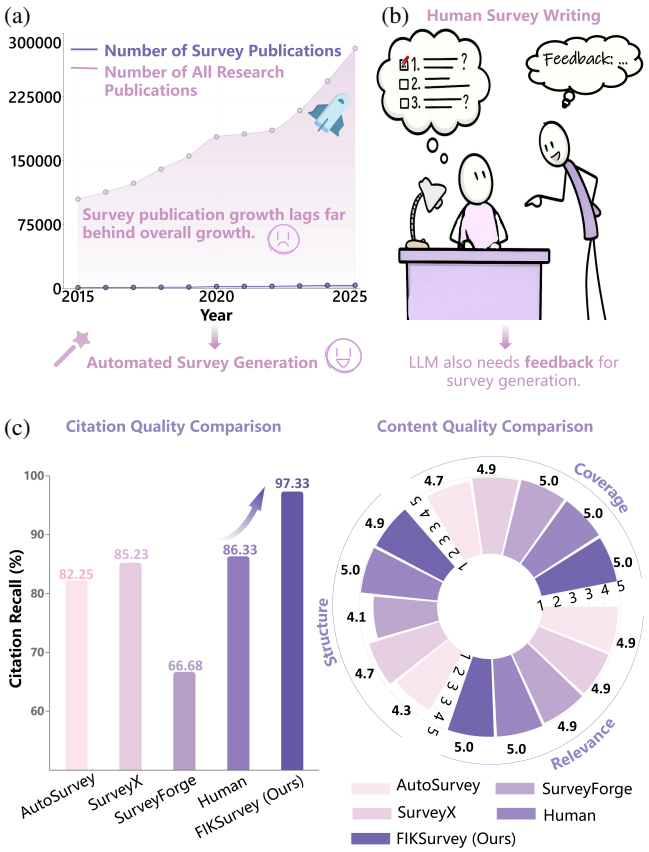

Figure 1: **Overview of FIKSurvey**. (a) illustrates the motivation to automate the survey generation. The rapid growth of scientific publications underscores the urgent need for automation in survey writing. FIKSurvey offers a practical solution with modest cost (Appendix B). (b) The motivation of introducing feedback for automated survey generation. (c) indicates the performance improvements achieved by our FIKSurvey.

1 Introduction

The rapid growth of scholarly literature has increased the demand for automated tools for knowledge synthesis (Fire and Guestrin, 2019; Bornmann et al., 2021). In particular, survey publications appear to grow much more slowly than the overall research literature (illustrated in Fig. 1(a)), which further motivates scalable approaches to survey writing. Large Language Models (LLMs) have recently been explored for survey generation (Brown et al., 2020; Raffel et al., 2020; Wei et al., 2022), with representative systems such as AutoSurvey (Wang et al., 2024b), SurveyX (Liang et al., 2025) and SurveyForge (Yan et al., 2025). AutoSurvey adopts a multi-stage workflow, while SurveyX further incorporates structured representations. Although these systems have demonstrated promising progress, they typically use a linear paradigm,treating survey generation as a one-way drafting process and do not explicitly incorporate the peer-feedback step that is common in human survey writing.

This linear paradigm contrasts with the core mechanism that ensures quality in human academic writing. Human authors rarely produce a high-

*Equal contribution.
†Corresponding author: ye.wei@cityu.edu.hk.

Findings of the Association for Computational Linguistics: ACL 2026, pages 38174–38195
July 2-7, 2026

quality survey in a single pass. Instead, drafts are revised based on feedback from peers, as illustrated in Fig. 1 (b). Motivated by this observation, we propose **FIKSurvey** (***F**eedback **i**s the **k**ey for automatic **s**urvey generation*), a feedback-driven framework that introduces an explicit "reviewer" role into survey generation.

Conceptually, this reframes survey generation from open-ended self-rewriting to a review-driven improvement loop. Unlike self-reflection or self-correction, where a single model critiques and revises its own draft, FIKSurvey externalizes critique into a dedicated reviewer role and uses survey-specific feedback organized as a review contract over *outline, content, and citations*. Moreover, citation feedback is made verifiable: for each claim–citation pair, we validate support via a sliding-window NLI procedure over retrieved evidence (Details in Sec.3). We operationalize this peer-review cycle with a dual-agent system: a Writer LLM responsible for content creation and a Helper LLM that acts as an automated "**peer reviewer**" to provide feedback. We additionally analyze 183 survey reviews collected from OpenReview, which provides empirical support for our feedback design (see Sec. 3.2). This "reviewer" provides continuous, structured feedback across three meticulously designed dimensions that mirror the core concerns of human reviewers: (i) **Outline feedback**: just as a human reviewer first assesses a paper's logical skeleton, our framework refines the survey's outline before drafting the main text, ensuring a clear, coherent, and academically sound structure and reducing major structural issues early on. (ii) **Content feedback**: the "reviewer" identifies weak or overly descriptive passages and suggests actionable revisions that encourage better quality. (iii) **Citation feedback**: the "reviewer" checks whether key claims are supported by the cited evidence and revises unsupported statements by rewriting the claim or replacing the citation when appropriate. We implement this check using a sliding-window natural language inference (NLI) procedure to match claims with relevant evidence spans.

The entire process is governed by a core principle of monotonic improvement: a revision is accepted only if it improves the survey's quality based on a quantitative rubric. Fig. 1 (c) summarizes the resulting improvements over existing systems.

To sum up, this work offers three primary contributions, reflecting a deeper vision for the future of LLM-powered academic writing:

1. **A new philosophy for LLM-powered academic writing:** We argue that the path to high-fidelity automated writing lies not in enhancing raw generative power but in automating the iterative process of critical review and revision, and we instantiate this philosophy in FIKSurvey.

2. **A novel framework with verifiable mechanisms:** We provide a robust and practical framework that translates the abstract concept of "peer review" into a set of specific, verifiable algorithms. The Sliding-Window natural language inference (NLI) mechanism for citation validation, in particular, offers a powerful new tool for ensuring the factual integrity of LLM-generated text.

3. **Empirical evidence of improved quality:** Experiments show obvious gains in quality over strong baselines (See Fig 1 (c)), highlighting the effectiveness of feedback-driven survey generation. More broadly, by supporting optional human-in-the-loop feedback, FIKSurvey sketches a workflow where where experts provide high-level guidanc and the system handles iterative refinement.

2 Related Work

Self-Reflection and Self-Correction. LLMs have show strong capabilities across a wide range of applications, yet obtaining high-quality outputs remains nontrivial. A growing line of work improves generation via critique-and-revision, using either self-reflection or self-correction (Dhuliawala et al., 2024; Madaan et al., 2023; Kamoi et al., 2024; Yang et al., 2025; Shinn et al., 2023; Wang et al., 2024b; Liang et al., 2025; Zhang et al., 2025). For example, Self-Refine (Madaan et al., 2023) prompts a single model to produce feedback and then refine its own draft repeatedly, while Reflexion (Shinn et al., 2023) records reflective notes in episodic memory to guide subsequent attempts. However, usually "it takes two to think" (Yanai and Lercher, 2024). Motivated by this, we externalize critique into an explicit reviewer role and structure feedback as a review contract tailored to surveys with rubric-gated acceptance. Crucially, our sliding-window NLI turns citation checking into a verifiable claim–evidence alignment procedure, moving beyond generic self-critique toward an auditable review process.

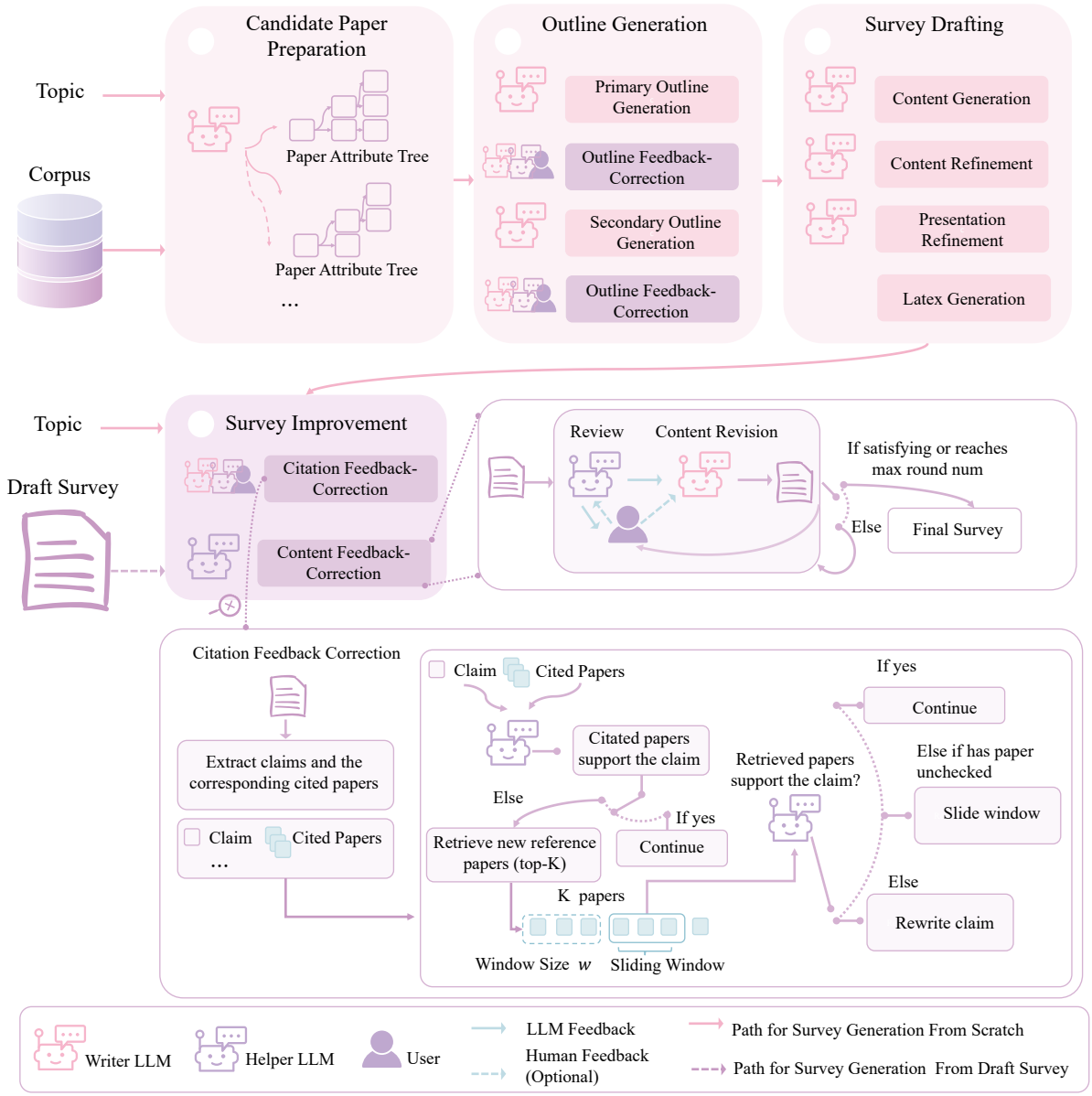

Figure 2: **Overall pipeline of FIKSurvey.** FIKSurvey supports both from-scratch and from-draft paths. It integrates automatic feedback across three dimensions: outline, content, and citation, while allowing optional lightweight human feedback. The bottom inset illustrates citation-feedback correction: for each claim–citation pair, existing references are checked; if unsupported, top-K candidate papers are retrieved and validated with a sliding window of size w. When support is found, citations are replaced, otherwise the claim can be rewritten.

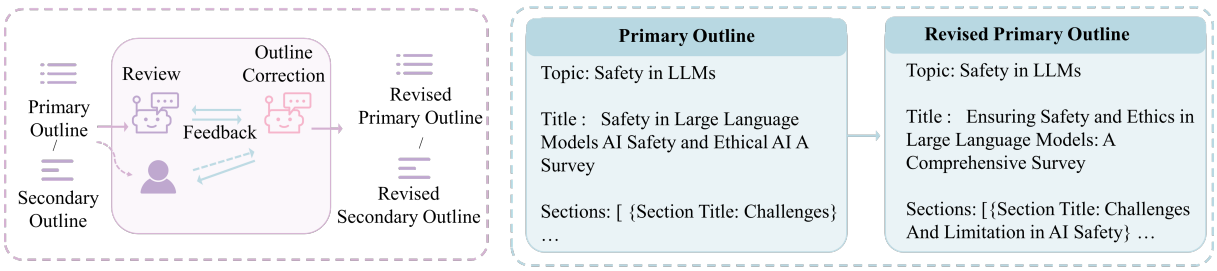

Figure 3: **Outline feedback in FiKSurvey**. (a) Illustration of outline feedback-correction: the writer LLM generates either a primary or a secondary outline, which is reviewed by the helper LLM. Feedback is incorporated iteratively until the revised outline is finalized. (b) Comparison of the primary outlines before and after outline-feedback correction for topic “Safety in LLMs".

Automatic Survey Generation. The rapid growth of scientific publications has spurred increasing interest in automatically drafting surveys to help researchers keep pace with fast-moving fields. Early work largely adopted extractive content selection and organization, which often suffers from coherence and redundancy issues (Hoang and Kan, 2010; Hu and Wan, 2014; Chen and Zhuge, 2019). Template-tree style frameworks introduce explicit hierarchical guidance, improving organization but remaining relatively rigid and sometimes requiring user-specified structure (Sun and Zhuge, 2019). The emergence of LLMs has shifted the focus to more flexible pipelines (Wang et al., 2024b; Liang et al., 2025; Yan et al., 2025; Su et al., 2025). In this line, AutoSurvey (Wang et al., 2024a) pioneers a multi-staged framework, providing hierarchical survey generation. SurveyX (Liang et al., 2025) further strengthens grounding by introducing structured reference preprocessing, such as AttributeTree, together with improved outline optimization, leading to better survey quality. However, these LLM-based pipelines do not explicitly model the feedback-driven revision process that typically shapes human survey writing.

3 Method

3.1 Pipeline Overview

Our proposed FIKSurvey supports two settings: (1) the from-scratch setting, where surveys are produced directly from a topic description, and (2) the from-draft setting, where existing drafts are iteratively improved. Both settings follow the same pipeline, with the from-draft workflow cutting off a corner. Two complementary LLMs are employed: a writer LLM that generates and revises content, and a helper LLM that provides evaluation and targeted feedback. The helper LLM incurs only a small additional overhead (see Appendix B), supporting the practicality of the framework. Optional human-in-the-loop review can be incorporated when user-specific requirements must be met, enabling tailored adjustments that go beyond generic automated feedback.

As shown in Fig. 2, FIKSurvey contains four stages: (1) candidate paper preparation (2) outline generation (3) survey drafting and (4) survey improvement. Details will be described in the subsequent paragraphs.

Candidate Paper Preparation. Given a topic and its corpus, FIKSurvey constructs attribute trees to represent candidate papers in a structured and queryable form. Following SurveyX (Liang et al., 2025), each paper is first categorized into one of four types, method, theory, benchmark, or survey. A type-specific template then prompts the LLM to extract salient dimensions from the full text (e.g., background, metrics) from the full text. The results are serialized into JSON-based attribute trees, which serve as compact yet information-rich retrieval units.

Outline Generation. Based on the attribute trees, the writer LLM first produces aprimary outline that lays out the top-level sections. The helper LLM critiques coverage, granularity, and title specificity, and issues actionable edits. The writer applies revisions until gains plateau or a cap is reached. Conditioned on the revised primary outline, the writer then generates a secondary outline with subsection-level structure and finer decomposition, followed by the same feedback–correction loop (again capped). The final secondary outline is frozen and used as the scaffold for drafting.

Survey Drafting. Using the accepted outline, the writer LLM produces a structured draft containing sections, subsections, and preliminary citations.

Survey Improvement. Drafts are refined through two dedicated modules: (1) Content feedback correction, which enhances clarity, depth; and (2) Citation feedback correction, which enforces factual consistency via claim–citation verification. Workflows of the two modules are provided in Fig. 4 and will be detailed in Sec. 3.2.

3.2 Feedback Mechanism

FIKSurvey injects feedback at three complementary levels: outline, content and citation. Outline level and content level are rubric-driven: each iteration is scored along five rubric (coverage, structure, relevance, synthesis, critical analysis). To ground the rubric in human practice, we collected 25 survey submissions from OpenReview (2021-2025) and associated reviews. After filtering out trivial comments, we retained 183 reviews and segmented them into 564 review sentences. Using keyword-based bucketing, we found that reviewer concerns align well with our rubric dimensions; for example, *Structure* (163 mentions) and *Relevance* (126 mentions) recur frequently (details in Appendix F). A revision is accepted only if the rubric score improves. Otherwise we revert to the previous version, preventing error accumulation and ensuring monotonic improvement. In contrast, citation feedback

Algorithm: Content Feedback-Correction

Inputs: draft D, rubric $\mathcal{M}$, retrieval budget K, window size w, max rounds R.
Loop (up to R rounds):

1. **Diagnose.** Helper LLM scores D on rubric dimensions, then identifies weak paragraphs and generates action plans (e.g., enhance paragraph, restructure).
2. **(If needed) Retrieval.** For actions requiring external evidence:
   (a) Retrieve candidate passages (top–K).
   (b) **Sliding window over top–K:** with window size w, move sequentially ($i = 1, 1{+}w, 1{+}2w, \dots$):
      i. Form window $\mathcal{W} = \{r_i, \dots, r_{\min(i+w-1,K)}\}$.
      ii. Apply natural language inference (NLI) to check whether any passage in $\mathcal{W}$ supports target paragraph.
      iii. Stop early if supporting evidence is found.
3. **Edit.** Writer LLM integrates the planned changes and evidence into edits.
4. **Accept/rollback.** Keep the revision only if rubric scores improve; otherwise revert. Stop after R rounds or when no further gains are found.

Output: revised survey draft D' (retaining only edits with net rubric gains).

Algorithm: Citation Feedback-Correction

Inputs: draft D, claims C, current citations $\mathcal{R}$, retrieval budget K, window size w.
For each claim $c \in C$:

1. **Check current citations.** Use NLI to verify which items in $\mathcal{R}(c)$ support c. If at least one is supported, keep only the supported subset and stop.
2. **If none support:** retrieve top–K candidate references $\{r_1, \dots, r_K\}$.
3. **Sliding window search.** With window size w, move sequentially ($i = 1, 1{+}w, 1{+}2w, \dots$):
   (a) Form window $\mathcal{W} = \{r_i, \dots, r_{\min(i+w-1,K)}\}$.
   (b) Rank candidates in $\mathcal{W}$ by NLI score with respect to c.
   (c) If a supporting reference $r^\star$ is found: replace the original citation and stop.
4. **If still unsupported:** rewrite the claim to align with available evidence.
5. **Minimal pruning.** If both supported and unsupported citations exist, prune only the unsupported ones.

Output: revised survey draft D' with verified citations.

Figure 4: **Feedback-correction in FiKSurvey**. **Left:** Content refinement loop, where weak paragraphs are diagnosed by the helper LLM, optionally supported with retrieved evidence (top–K searched via a sliding window of size w), and revised by the writer LLM. **Right:** Citation validation loop, where claims are first checked against current references. If none support, top–K candidates are scanned with a sliding window until support is found; otherwise, the claim is rewritten. The output in both cases is a revised survey draft with improved content and verified citations.

follows a verification-and-repair loop.

Outline Feedback. We cast the outline generation of FIKSurvey as a feedback loop between the writer LLM and the helper LLM. The writer LLM generates a primary and secondary outline, while the helper LLM evaluates them on rubric dimensions and produces structured feedback.

The writer then applies targeted modifications. Revisions are accepted only if scores improve, otherwise the system reverts to the best prior version with a recorded degradation analysis. As shown in Fig. 3, the title and section title of the initial primary outline can be revised into more academically appropriate forms. The process iterates up to a configurable maximum number of rounds, ensuring monotonic improvement without excessive drift. More details are provided in Appendix D.1.

Content Feedback. As shown in Fig. 4 (left), the helper LLM first evaluates the current draft survey against rubric dimensions. (More details are provided in Appendix D.2.) Based on this assessment, it identifies weak paragraphs and assigns targeted actions, and proposes targeted revision strategies, for example, enhancing clarity, restructuring the logical flow, or adding contrasting perspectives (internally mapped to action tags). If external evidence is required, FiKSurvey triggers retrieval-augmented generation (RAG): candidate passages are retrieved (top-K) and reranked using a sliding-window *natural language inference (NLI)* model to test entailment between claims and evidence. The top-ranked snippets are then supplied to the writer LLM, which performs LaTeX-aware edits. A revision is accepted only if rubric scores improve, with at most two correction rounds allowed per draft.

Citation Feedback. As illustrated in Fig. 4 (right), citation validation begins by checking whether the current references already support the claim using sliding-window NLI with window size w. If at least one citation is verified, unsupported references are pruned while valid ones are retained. If no support is found, FiKSurvey retrieves the top-K candidate references and slides the window across them to search for supporting evidence. If a valid reference is identified, it replaces the original citation. Otherwise, the system rewrites the claim to align with verifiable evidence. The process ensures

the revised survey maintains only grounded and verifiable citations (Choi et al., 2025).

Optional Human Feedback FiKSurvey is primarily designed for fully automated survey generation and refinement. Human-in-the-loop review is considered only when users have specific requirements that cannot be satisfied through automation alone. In such cases, minimal expert input can be incorporated to tailor the survey to user needs (details are provided in Appendix D.4).

4 Experiments

4.1 Setup

4.1.1 FIKSurvey Configuration

The FIKSurvey framework supports both generating surveys from scratch and improving existing drafts. It employs two LLMs: a primary writer (GPT-4o) for drafting/refining, and a helper model (GPT-4o-mini) for feedback. It incorporates feedback through three dedicated modules: *outline feedback correction*, *content feedback correction*, and *citation feedback correction*. For outline feedback correction, we allow up to three conversation rounds; for content feedback correction, we limit the process to two rounds; and for citation feedback correction, we adopt a top-K retrieval strategy with a sliding window natural language inference (NLI) to check citation validity, where $K = 10$ and the window size is set to 2.

4.1.2 Evaluation Metrics

Content Quality. Following AutoSurvey (Wang et al., 2024b), we evaluate the content quality of generated surveys across three dimensions: (1) *Coverage*: Assesses the comprehensiveness in covering the core and frontier aspects of the topic. (2) *Structure*: Evaluates the logical organization and coherence between sections. (3) *Relevance*: Measures whether the content is focused and free of irrelevant information. Each dimension is scored on a 1 - 5 scale by an evaluator LLM, using rubric-based prompts adapted from (Wang et al., 2024b).

Citation Quality. To assess factual grounding, we adopt citation *recall* and *precision* metrics. Given a set of claims extracted from the generated survey, we check whether each claim is supported by at least one cited reference (recall), and whether each cited reference is indeed relevant to the claim (precision). Formally written as follows:

$$\text{Recall} = \frac{\#\{\text{supported claims}\}}{\#\{\text{all claims}\}}$$

$$\text{Precision} = \frac{\#\{\text{supported claim-source pairs}\}}{\#\{\text{all claim-source pairs}\}}.$$

Support is determined via an entailment-based classifier that evaluates whether the core of a claim can be inferred from the cited abstract. This aligns with the implementation in SurveyX (Liang et al., 2025). Because hallucinated claims often manifest as unsupported or irrelevant citations, improvements in citation quality directly indicate reductions in hallucination, thereby increasing the reliability of the generated surveys.

4.1.3 Baselines

We compare FIKSurvey with three LLM-based baselines: AutoSurvey (Wang et al., 2024b), SurveyX (Liang et al., 2025), and SurveyForge (Yan et al., 2025), as well as a human reference consisting of published arXiv surveys.

4.2 Results and Analysis

4.2.1 Quantitative Results

We first evaluate FIKSurvey in the from-scratch generation setting. As shown in Tab. 1, FIKSurvey achieves the best overall performance among automated methods and produces results comparable to the human baseline. We further report the quantitative results of FIKSurvey in the *from-draft* setting, as shown in Tab. 2. Compared to the initial draft surveys, those refined by FIKSurvey exhibit consistent improvements across all dimensions, with particularly notable gains in citation quality. All reported scores are averaged over 20 topic-specific surveys.

4.2.2 Human Evaluation

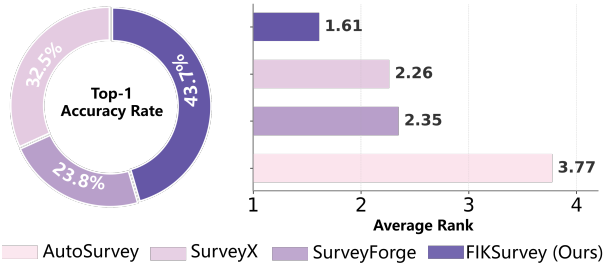

Figure 5: Results of human evaluation.

To further assess the quality of the generated surveys, we conducted a blind human evaluation comparing four LLM-based survey generation systems: AutoSurvey (Wang et al., 2024b), SurveyX (Liang

Table 1: Comparison of models on content quality and citation metrics. Best results are highlighted in **bold**.

	Content				Citation	
Method	Coverage	Structure	Relevance	Avg	Recall (%)	Precision (%)
AutoSurvey (Wang et al., 2024b)	4.73	4.33	4.86	4.64	82.25	77.41
SurveyX (Liang et al., 2025)	4.91	4.73	4.86	4.83	87.80	86.78
SurveyForge (Yan et al., 2025)	**5.00**	4.10	4.90	4.67	66.8	65.2
FIKSurvey (Ours)	4.95	**4.90**	**5.00**	**4.95**	**97.33**	**92.85**
Human	5.00	4.95	5.00	4.98	86.33	77.78

Table 2: Quantitative results of draft surveys and their improved versions refined by FIKSurvey.

	Content				Citation	
	Coverage	Structure	Relevance	Avg	Recall (%)	Precision (%)
Draft surveys	4.90	4.75	4.81	4.82	81.22	80.62
Improved by FIKSurvey	**4.95**	**4.95**	**4.95**	**4.95**	**97.22**	**95.95**

et al., 2025), SurveyForge (Yan et al., 2025), and our FIKSurvey on 20 topics. We recruited 20 participants, all graduate students with knowledge in the relevant area, and obtained their consent for data collection. For each assigned topic, participants were shown the four survey outputs in a randomized order and under anonymized labels (*i.e.*, A–D), and were asked to provide a complete best-to-worst ranking based on overall survey quality. Fig. 5 reports the Top-1 rate and average rank (lower is better) in our human evaluation. FIKSurvey achieves the highest Top-1 rate (43.7%) and the lowest average rank (1.61), indicating a consistent preference over the baselines.

4.2.3 A Closer Look at the Feedback

Herein, we take a closer look at the role of feedback in FIKSurvey. FIKSurvey incorporates feedback along three dimensions: outline, content, and citation. Each addresses distinct quality concerns, such as structural soundness and factual grounding. To understand their contributions, we conduct ablation studies in two complementary ways: (1) a step-wise ablation where feedback modules are progressively removed (Tab. 3) and (2) a module-wise ablation where each module is individually excluded while keeping the others intact (Tab. 4).

Step-Wise Ablation. Tab. 3 shows that removing the *content feedback* module alone already leads to a small drop in synthesis and critical analysis, but citation quality remains largely intact. When both *content* and *citation* feedback are removed, citation recall and precision fall sharply (from over 90% to around 60–77%), indicating the central role of citation verification in maintaining factual grounding. Finally, removing all three feedback components substantially degrades both content and citation metrics, demonstrating that their cumulative effect is essential for stable quality.

Case Study (Topic: LLM-based Agents)

Draft Survey (Excerpt).
2 Background and Definitions

- *2.4 Evolution and Advancements in LLMs*
- *2.5 Integration into AI Agent*
  "Integrating large language models (LLMs) into artificial intelligence (AI) ..."
  ...

Human Review:
"The subsection title is ambiguous. What is being integrated? Please revise the title and align the content accordingly. Be more structured."

Writer LLM Plan:
Clarify *what* is integrated; restructure by **Architecture / Functionality / Implications**. (No RAG required; structural clarity rewrite.)

Rewritten Result:
"2.5 Integration of Large Language Models into AI Agents.
Large Language Models (LLMs) have emerged as pivotal components in the development of advanced AI agents..."

Figure 6: Case study of human feedback improving draft survey quality. We show how a vague draft subsection title and content (2.5) is revised after human review, with the writer LLM producing a more clarified version.

Module-Wise Ablation. Tab. 4 provides a finer-grained view. Here, removing *content feedback* mostly impacts synthesis and critical analysis, while removing *citation feedback* causes the largest drop in citation recall and precision. By contrast, removing *outline feedback* mainly affects coverage and structure but leaves citation quality relatively stable. Taken together, these results highlight that each dimension plays a distinct and complementary role in ensuring the overall survey quality.

Human Feedback. Although FiKSurvey is de-

Table 3: Step-wise ablation of the feedback mechanism in FIKSurvey. Feedback components are progressively removed in the order Content → (Content + Citation) → (Content + Citation + Outline). Experiments are conducted on 10 randomly sampled topics. Results show that each feedback dimension contributes to both content and citation quality, and their cumulative effect is essential.

Method	Content Quality				Citation Quality	
	Coverage	Structure	Relevance	Avg	Recall (%)	Precision (%)
w/o Content Feedback	4.90	5.00	5.00	4.97	97.26	93.46
w/o Content + Citation Feedback	5.00	4.80	5.00	4.93	77.40	60.56
w/o Content + Citation + Outline Feedback	4.90	4.70	4.90	4.83	76.20	60.91
FIKSurvey (full)	4.90	5.00	5.00	4.97	97.20	93.38

Table 4: Module-wise ablation of feedback in FIKSurvey. We isolate the effect of removing each feedback dimension individually. Each dimension contributes to either content quality, citation quality, or structural soundness, and together they provide complementary improvements.

Ablation Module	Content Quality				Citation Quality	
	Coverage	Structure	Relevance	Avg	Recall (%)	Precision (%)
w/o Content Feedback	4.90	5.00	5.00	4.97	97.26	93.46
w/o Citation Feedback	5.00	4.50	5.00	4.83	80.53	65.67
w/o Outline Feedback	4.80	4.40	5.00	4.73	97.41	93.87
FIKSurvey (full)	4.90	5.00	5.00	4.97	97.20	93.38

Conversation Snapshot (Human – Helper LLM)

Human: Give me some suggestions on refining the section title.

Helper LLM: Refining section titles can significantly enhance clarity. Suggestions:

- **Background and Definitions** → *Foundations and Evolution: Defining Key Concepts in Machine Learning*
- **In-Context Learning** → *In-Context Learning: Mechanisms, Benefits, and Challenges*
  ...
- **Transfer Learning** → *Transfer Learning: Leveraging Pre-trained Models for New Horizons*
  ...

Outline Comparison (Excerpt)

Original Draft Outline.

- Introduction
- Background and Definitions
- In-Context Learning
  ...

Final Draft Survey Outline.

- *Introduction*
- *Foundations and Evolution: Defining Key Concepts in Machine Learning*
- *In-Context Learning: Mechanisms, Benefits, and Challenges*
  ...

Figure 7: Case study of human–LLM feedback for outline refinement. **Top**: conversation snapshot of helper LLM suggestions. **Bottom**: comparison of original outline and the final refined version.

signed to minimize manual intervention, optional human feedback can still offer targeted and customized improvements. As illustrated in Fig. 6, human reviewers are able to flag vague subsection titles or misaligned content, prompting the writer LLM to generate more precise and well-scoped revisions. This demonstrates how even light-touch human involvement can complement LLM feedback and enhance the final output. Further details are provided in the Appendix.

Human–LLM Feedback. Beyond purely human review, FIKSurvey also supports hybrid collaboration during outline generation. Here, human experts provide lightweight guidance (e.g., pointing out that section titles are overly broad), while the helper LLM proposes concrete refinements such as reworded titles or added subtopics. This division of labor requires only limited expert effort but yields clearer and more academically precise structures. An example of such interaction is shown in Fig. 7.

5 Conclusion and Outlook

In conclusion, grounded in observations of how humans write surveys, we use the principle that *feedback is the key* to high-quality automatic survey generation. By externalizing critique into an explicit reviewer role and enforcing gated, verifiable revisions, FIKSurvey achieves consistent improvements in survey quality over strong baselines.

We also analyze OpenReview survey reviews via multi-round LLM discussions, which offers insights for future work. For example, reviewer emphasis appears context-dependent (e.g., fast-evolving areas favor breadth and topicality,

whereas mature areas value higher-level discussion), suggesting that adaptive weighting of feedback dimensions can be beneficial (more details are described in Appendix F).

Limitation

Our current feedback design equally treats content criteria, but the review analysis mentioned in Sec. 5 suggests that the relative importance of reviewer criteria may vary across topics and venues. Incorporating adaptive weighting, therefore, can be a natural extension for future work. Additionally, occasional divergences between reviewer critiques and LLM judgments indicate that explicitly integrating uncertainty-aware feedback may improve robustness in when facing ambiguous cases.

Ethics Statement

This paper presents FIKSurvey, a feedback-driven framework for automated survey generation. We report the details of our methodology and experiments to support transparency and reproducibility. We report methodological and experimental details to support transparency and reproducibility. We have considered potential societal impacts, avoided privacy-sensitive data collection, and are committed to making a positive contribution to the research community.

References

Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. 2021. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Humanities and Social Sciences Communications*, 8(1):1–15.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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

Juhwan Choi, JungMin Yun, Changhun Kim, and YoungBin Kim. 2025. Position paper: How should we responsibly adopt llms in the peer review process? *Submitted to ACL Rolling Review-July*, 2025.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the association for computational linguistics: ACL 2024*, pages 3563–3578.

Michael Fire and Carlos Guestrin. 2019. Over-optimization of academic publishing metrics: observing goodhart's law in action. *GigaScience*, 8(6):giz053.

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

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.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

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

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, and 1 others. 2023. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36:46534–46594.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.

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

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.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024a. Autosurvey: Large language models can automatically

write surveys. *Advances in neural information processing systems*, 37:115119–115145.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Min Zhang, Qingsong Wen, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Autosurvey: Large language models can automatically write surveys. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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

Itai Yanai and Martin J. Lercher. 2024. It takes two to think. *Nature Biotechnology*, 42(1):18–19.

Zhe Yang, Yichang Zhang, Yudong Wang, Ziyao Xu, Junyang Lin, and Zhifang Sui. 2025. Confidence vs critique: A decomposition of self-correction capability for llms. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3998–4014.

Jiebin Zhang, J Yu Eugene, Qinyu Chen, Chenhao Xiong, Dawei Zhu, Han Qian, Mingbo Song, Xiaoguang Li, Qun Liu, and Sujian Li. 2024. Retrieval-based full-length wikipedia generation for emergent events. *Preprint*.

Ziyan Zhang, Yang Hou, Chen Gong, and Zhenghua Li. 2025. Self-correction makes LLMs better parsers. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 6749–6762, Suzhou, China. Association for Computational Linguistics.

Appendices

We first present the survey topics in Appendix A. Then we provide the cost analysis in Appendix B, details of the implementation in Appendix C and the feedback mechanism in Appendix D. Case studies and further investigations are presented in Appendix E and Appendix F, respectively. Finally, the use of LLM is elaborated in Appendix G and the evaluation prompts are presented in Appendix H.

A Survey Topics

Tab. 5 lists the 20 survey topics used in our main evaluation experiments. For reference, we also provide the titles of the corresponding surveys authored by human experts.

B Cost Analysis

We report the cost and runtime breakdown of FIKSurvey by agent role (*ie* Writer LLM and Helper LLM) in Tab. 6. The average cost is modest ($4.90), suggesting that the framework is economically feasible in practice, and the Helper LLM incurs only a small overhead, making the feedback-driven gains achievable at low additional expense.

C Implementation Details

In FIKSurvey, all RAG components use `bge-base-en-v1.5` (Zhang et al., 2024) as the embedding model. For both outline- and content-feedback correction, the framework supports iterative multi-round refinement, with the maximum number of rounds set to 3 and 2, respectively. Since our focus is not on literature collection itself, for the selected 20 topics we directly adopt the corpus provided by SurveyX (Liang et al., 2025) and follow its procedure to construct attribute trees for candidate papers.

D Details of Feedback Mechanism

D.1 Details of Outline Feedback

We treat survey outline construction as a closed-loop control problem. The helper LLM produces structured, machine-readable feedback, then the writer LLM applies targeted revisions under explicit constraints. An acceptance gate decides whether to keep or rollback based on quality deltas. This design converts vague reviewer-style "comments" into actionable signals that drive monotonic improvement.

D.1.1 Workflow of Outline-Feedback Mechanism in FIKSurvey

Steps. Our pipeline proceeds in seven steps:

1. Primary outline generation (writer LLM).
2. Primary outline review (helper LLM).
3. Primary modification loop (writer LLM $\leftrightarrow$ helper LLM, rollback if degraded).
4. Paper mounting and clue extraction for secondary outline generation.
5. Secondary outline generation (writer LLM).

Table 5: Survey topics and corresponding human-written survey titles.

Topic	Survey Title
In-context Learning	A Survey for In-context Learning
LLMs for Recommendation	A Survey on Large Language Models for Recommendation
LLM-Generated Texts Detection	A Survey of Detecting LLM-Generated Texts
Explainability for LLMs	Explainability for Large Language Models
Evaluation of LLMs	A Survey on Evaluation of Large Language Models
LLMs-based Agents	A Survey on Large Language Model based Autonomous Agents
LLMs in Medicine	A Survey of Large Language Models in Medicine
Domain Specialization of LLMs	Domain Specialization as the Key to Make Large Language Models Disruptive
Challenges of LLMs in Education	Practical and Ethical Challenges of Large Language Models in Education
Alignment of LLMs	Aligning Large Language Models with Human
ChatGPT	A Survey on ChatGPT and Beyond
Instruction Tuning for LLMs	Instruction Tuning for Large Language Models
LLMs for Information Retrieval	Large Language Models for Information Retrieval
Safety in LLMs	Towards Safer Generative Language Models: Safety Risks, Evaluations, and Improvements
Chain of Thought	A Survey of Chain of Thought Reasoning
Hallucination in LLMs	A Survey on Hallucination in Large Language Models
Bias and Fairness in LLMs	Bias and Fairness in Large Language Models
Large Multi-Modal Language Models	Large-scale Multi-Modal Pre-trained Models
Acceleration for LLMs	A Survey on Model Compression and Acceleration for Pretrained Language Models
LLMs for Software Engineering	Large Language Models for Software Engineering

Table 6: Cost of FIKSurvey.

Writer LLM		Helper LLM		Total	
Cost (USD)	Time (s)	Cost (USD)	Time (s)	Cost (USD)	Time (s)
4.80	3386	0.10	1066	4.90	4452

6. Secondary outline review and modification loop (helper LLM $\leftrightarrow$ writer LLM, rollback).
7. Finalization and logging (best version, JSON export).

Paper mounting is a preparatory step that grounds the secondary outline in concrete evidence. Each candidate paper is assigned to a primary outline section by matching its title or abstract, and concise clues are extracted (e.g., datasets, benchmarks, method families, limitations). These clues provide the Writer LLM with domain-grounded anchors when expanding the primary sections into more detailed secondary subsections, thereby improving coverage and reducing superficial enumeration.

Feedback Artifacts. The loop exchanges four types of signals:

- Review signals (Primary Outline): lenient 1–5 scores across five dimensions, accompanied by short `notes` (one–two sentence comments clarifying why a score was given) and `quick_wins` (low-cost, high-impact edits such as renaming an ambiguous section title).
- Review signals (Secondary Outline): similar output to review signals for primary outline.
- Prescriptive signals: modification instructions that translate evaluation into concrete edits, while respecting stage-specific constraints.
- Diagnostic signals: degradation analysis that explains why a revision reduced quality and suggests alternative strategies for the next iteration.

Control Logic. We accept a revision if its average score improves or meets a threshold; otherwise we rollback to the best-known version. Primary outlines preserve the number and order of sections. Secondary outlines allow flexible restructuring of subsections. This separation stabilizes global structure while enabling local refinement.

D.2 Details of Content Feedback

We extend the closed-loop feedback principle from outlines to full survey drafts. In this setting, a helper LLM produces structured, machine-readable review signals (scores, notes, and action tags) at the paragraph level, while the writer LLM applies targeted revisions under explicit constraints. An acceptance gate decides whether to accept or rollback changes depending on quality deltas. When external support is required, retrieval-augmented generation (RAG) with sliding-window NLI ensures that only evidence-entailing passages are integrated, reducing hallucination risk. This design transforms vague reviewer-style comments into actionable, localized edits that improve clarity, logical flow, and critical depth.

D.2.1 Workflow of Content-Feedback Mechanism in FIKSurvey

The content-feedback pipeline proceeds in four steps:

1. Content review (helper LLM). Score the draft on five rubric dimensions; provide short notes and assign action tags to weak paragraphs.
2. Modification instructions (helper LLM). Translate action tags into concrete editing suggestions.
3. Revision (writer LLM). Apply paragraph-level edits in LaTeX, optionally using retrieved evidence (top-K snippets verified via NLI).
4. Acceptance gate. Re-evaluate the revised draft; accept if scores improve, rollback otherwise. A maximum of two correction rounds are allowed per draft.

D.2.2 Prompts for Content Feedback

In FIKSurvey, content refinement relies on structured prompts to ensure that evaluations, modification plans, and RAG-enhanced revisions remain machine-readable and auditable. Below we reproduce the core prompts.

D.3 Details of Citation Feedback

D.3.1 Workflow of Citation-Feedback Mechanism in FIKSurvey

The process of citation-feedback mechanism can be summarized in the following steps:

1. Locate cited sentences. Identify sentences containing citations and split them into the claim text and its citation keys.
2. Check support. For each claim–citation pair, retrieve the cited paper and test whether the claim is supported using a natural language inference (NLI) model.
3. Handle supported claims. If support is found, optionally prune away unsupported references and keep only the valid ones.
4. Diagnose unsupported claims. If no support is found, a helper LLM generates a short diagnosis explaining why the claim fails.
5. Propose fixes. The helper LLM suggests a factual rewrite, style notes, and alternative phrasings in a structured format.
6. Retrieve candidate references. Search for new candidate citations from the corpus and verify them with the NLI model.
7. Apply repair strategies. Depending on configuration, repair may prioritize adjusting references, rewriting the claim, or both. A revision is accepted only if it is NLI-supported.
8. Finalize and log. The corrected Latex file and a JSON record of diagnostics, fixes, and chosen actions are saved for transparency.

This workflow turns judgments of the citations into explicit, machine-verifiable feedback, enabling automatic citation repair in a transparent and auditable way.

D.3.2 Illustrative Case Study of Citation Repair

To further illustrate how the citation-feedback module repairs unsupported or weakly supported references, we provide representative examples in Fig. 15 from the topic *Domain Specialization of LLMs*.

D.4 Details of Human Feedback

In this module, the writer LLM serves as the central actor: it receives human opinions, structured plans, and optional evidence, and is responsible for carrying out edits. The workflow unfolds as following steps:

1. Opinion Collection. Human feedback items are gathered (ID, category, priority, text). These define the editing objectives.
2. Structure Analysis. The writer LLM summarizes the current draft into JSON, providing contextual grounding for later editing.
3. Global Modification Plan. The writer LLM integrates all opinions into a structured plan specifying execution order, dependencies, and evidence needs.
4. Evidence Provision. For opinions requiring external support, candidate passages are retrieved, filtered via NLI entailment, and supplied as verified evidence.
5. Stepwise Modification. Guided by the plan, the Writer applies LaTeX-aware edits using the draft, the given opinion, and optional evidence, while preserving global structure.

Prompt: Review (Coverage)

You are a reviewer evaluating an academic survey draft.
Dimension: COVERAGE
Task: Assess how well the draft survey covers the breadth of the field. Does it include central methods, datasets, benchmarks, and peripheral topics? Identify major gaps or missing areas.
Scoring (1–5): 1 = very poor coverage; 5 = very comprehensive.
Return JSON only:

```
{
  "dimension": "coverage",
  "score": <1-5>,
  "reasoning": "...",
  "improvement_plan": "..."
}
```

Figure 8: Review prompt for helper LLM.

Prompt: Review (Structure)

You are a reviewer evaluating an academic survey draft.
Dimension: STRUCTURE
Task: Assess the logical organization of the survey. Is the hierarchy clear and consistent? Does the flow make sense (e.g., Problem/Scope → Landscape → Methods → Synthesis → Future)?
Scoring (1–5): 1 = very disorganized; 5 = clear and coherent.
Return JSON only:

```
{
  "dimension": "structure",
  "score": <1-5>,
  "reasoning": "...",
  "improvement_plan": "..."
}
```

Figure 9: Review prompt for helper LLM.

Prompt: Review (Relevance)

You are a reviewer evaluating an academic survey draft.
Dimension: RELEVANCE
Task: Assess whether the survey content is on-topic and aligned with the intended scope. Are tangential or irrelevant areas included? Are exclusions or boundaries clearly stated?
Scoring (1–5): 1 = mostly irrelevant; 5 = highly focused and on-topic.
Return JSON only:

```
{
  "dimension": "relevance",
  "score": <1-5>,
  "reasoning": "...",
  "improvement_plan": "..."
}
```

Figure 10: Review prompt for helper LLM.

Prompt: Review (Synthesis)

You are a reviewer evaluating an academic survey draft.
Dimension: SYNTHESIS
Task: Assess whether the survey synthesizes and compares works, rather than just listing them. Does it provide unifying frameworks, axes, or comparative insights?
Scoring (1–5): 1 = no synthesis, just enumeration; 5 = strong unifying synthesis.
Return JSON only:

```
{
  "dimension": "synthesis",
  "score": <1-5>,
  "reasoning": "...",
  "improvement_plan": "..."
}
```

Figure 11: Review prompt for helper LLM.

Prompt: Review (Critical Analysis)

You are a reviewer evaluating an academic survey draft.
Dimension: CRITICAL ANALYSIS
Task: Assess whether the survey includes critical perspectives. Does it discuss limitations, open problems, threats to validity, or future challenges?
Scoring (1–5): 1 = no critical analysis; 5 = strong and insightful critique.
Return JSON only:

```
{
  "dimension": "critical_analysis",
  "score": <1-5>,
  "reasoning": "...",
  "improvement_plan": "..."
}
```

Figure 12: Review prompt for helper LLM.

Prompt: Parse Improvement Plan

You are a technical editor analyzing survey improvement suggestions.
Parse the following improvement plan and extract specific, actionable modifications in JSON format.
Dimension: {dimension} Improvement Plan: {improvement_plan}
For each suggested modification, return JSON with this structure:

```
{
    "type": "add_section|insert_subsection|
    add_table|add_timeline|enhance_paragraph",
    "target_location": "section_name or general_location",
    "description": "specific description of what to do",
    "content_keywords": ["key", "terms", "to", "search"],
}
```

Return only a JSON array of modification objects.

Figure 13: Prompt for parsing improvement plan.

Prompt: Generate Enhanced Content with RAG

You are a technical writer creating content for an academic survey.
Based on the following action description and reference materials, generate appropriate LaTeX content.
Action Type: {action.action_type}
Action Description: {action.description}
Reference Materials: {references_text}
Requirements:
1. Generate content appropriate for the action type:
- add_section: Full section with \section{} command and content
- insert_subsection: Full subsection with \subsection{} command and content
- add_table: LaTeX table with \begin{table} environment
- add_timeline: Structured chronological content or table
- enhance_paragraph: Detailed paragraph content
2. Include proper LaTeX citations using \cite{} for the references
3. Use academic writing style appropriate for surveys
4. Make content substantive and well-structured
5. Ensure content is directly relevant to the action description
Generated LaTeX content.

Figure 14: Prompt for generating enhanced content with RAG.

Case Study (Topic: Domain Specialization of LLMs)

Example 1
Sentence: "Ethical concerns in embodied conversational agents (ECAs) highlight the importance of inclusivity and fairness in AI systems. . . "
Original citation: *It Takes a Village: Multidisciplinarity and Collaboration for the Development of ECAs*
Repaired citation: *Intersectional Bias in Causal Language Models*
Rationale: The original reference primarily discusses multidisciplinary collaboration in ECA development, whereas the repaired reference directly addresses bias- and fairness-related issues in AI systems.

Example 2
Sentence: "The generative AI paradox, balancing creativity and control in AI-generated content, underscores the necessity for robust ethical guidelines. . . "
Original citation: *Intersectional Bias in Causal Language Models*
Repaired citations:

- *The Generative AI Paradox: "What It Can Create, It May Not Understand"*
- *ChatGPT for Good? On Opportunities and Challenges of Large Language Models for Education*

Rationale: The claim combines multiple aspects, including the generative-AI paradox, the tension between creativity and control, and the broader ethical framing. A single citation focused on bias is insufficient; the repaired set provides more complementary support for these different facets of the statement.

Figure 15: Illustrative case study of citation repair.

6. Acceptance Gate And Rollback. Each revision is re-scored on five dimensions. If quality degrades, the system reverts to the prior version and logs a degradation note.

7. Finalization And Reporting. The Writer's best-performing version is exported as the final draft, together with a structured log of all modifications.

E Case Study of Optional Human Feedback

In this section, we provide example of human feedback for content refinement in Fig. 16 and example of human-LLM feedback for outline refinement in Fig. 17.

F Further Investigation on Building Feedback

We collected 25 survey submissions from OpenReview (2021–2025) and their associated reviews. After filtering out trivial comments, we retained 183 reviews and segmented them into 564 sentences. Using keyword-based bucketing, we counted sentence-level mentions of each criterion in the reviews: *Structure* was mentioned 163 times, followed by *Relevance* (126) and *Coverage* (119). Representative snippets were then supplied to multiple large language models acting as distinct roles (methodologist, domain expert, statistician). All the roles are implemented by GPT-4o. Through multi-round dialogue, these roles critiqued and refined the rubric; finally, an adjudicator LLM merged the proposals into a unified rubric with normalized weights. This procedure ensured that the resulting framework was not only theoretically motivated but also grounded in real reviewer concerns.

Insights across dimensions. As shown in Fig. 18, the analysis confirms that the five dimensions are strongly supported by recurring reviewer critiques. For example, under *Coverage*, reviewers frequently stressed omissions of recent work or inadequate search protocols. *Structure* was highlighted when papers lacked logical flow or signposting. On *Relevance*, reviewers questioned topical fit and timeliness. For *Synthesis*, concerns centered on superficial taxonomies. Finally, *Critical Analysis* was repeatedly mentioned when surveys lacked balanced critique or articulation of open problems. These consistent patterns suggest that our dimension design is well aligned with human judgment.

Future directions. Three directions arise from the review evidence. First, while our current rubric assigns equal weights to the five dimensions, reviewer emphasis clearly varies across domains: Coverage and Relevance dominate in fast-moving fields, whereas Synthesis and Critical Analysis matter more in mature topics. This indicates the need for **adaptive weighting**, conditioned on topic maturity, venue norms, or temporal context. Second, reviewers often noted interdependencies (e.g., weak coverage undermining synthesis). Future feedback mechanisms should therefore **model interactions between dimensions** rather than treating them independently. Third, discrepancies between reviewer critiques and LLM assessments suggest the value of **uncertainty-aware feedback**: when signals are weak or conflicting, the system could modulate its interventions or escalate to optional human feedback. Overall, our analysis demonstrates that the five-dimension rubric is not arbitrary but arises from consistent reviewer expectations. Future work should build on this foundation to design **intelligent, context-sensitive feedback**, extending FIK-Survey from lightweight corrections toward adaptive, reasoning-driven interventions that mirror and augment human reviewing practice.

G Statement on the Use of AI Assistants

During manuscript preparation, AI Assistants were used only for general writing assistance (proofreading, stylistic polishing, and limited rephrasing, with all such text subsequently revised by the authors. The research conceptualization, framework design, experimental design, and result interpretation were conducted exclusively by the human authors, who assume responsibility for the manuscript.

H Evaluation Prompt

In this section, we provide the prompt we use for evaluation in our experiments.

Case Study (Topic: LLM-based Agents)

Draft Survey (Excerpt).
1 Introduction

- *1.1 Significance of LLMs-based Agents*
  The emergence of large language models (LLMs) has significantly transformed artificial intelligence, particularly in natural language processing (NLP), by enhancing AI agents ...
- *1.2 Objectives of the Survey*
  This survey provides a comprehensive examination of LLM-based agents in natural language processing, focusing on several key objectives. Firstly, it explores the potential benefits and challenges of deploying large language ...
  ...

Human Review.
"The introduction lacks sufficient background on recent developments in large language models and their role in autonomous agents."

Rewritten Result.
1 Introduction
Recent developments in large language models (LLMs) have revolutionized the landscape of Artificial Intelligence and natural language processing (NLP). These models, such as GPT-4 and others, have demonstrated unprecedented performance across an array of language tasks, culminating in human-like capabilities in understanding and generating text. Notably, the advancements in training techniques, scaling parameters, and architectural refinements have accelerated their integration into autonomous agents.
LLMs are now foundational in creating autonomous agents capable of performing complex interactions in real-time environments. Their ability to process and generate natural language allows for sophisticated dialogues, making them adept in various applications ranging from customer service to academic support. The emergence of multimodal frameworks further showcases LLMs' capacity to bridge language and visual data, offering new dimensions to autonomous systems.
Over the past few years, the significance of LLMs in AI has grown substantially, enhancing tasks such as sentiment analysis, information retrieval, and content creation. This paper explores the transformative impact of LLMs-based agents on NLP, emphasizing their recent developments, capabilities, and the challenges that lie ahead.

- *1.1 Significance of LLMs-based Agents*
  The emergence of large language models (LLMs) has significantly transformed artificial intelligence, particularly in natural language processing (NLP), by enhancing AI agents ...
- *1.2 Objectives of the Survey*
  This survey provides a comprehensive examination of LLM-based agents in natural language processing, focusing on several key objectives. Firstly, it explores the potential benefits and challenges of deploying large language ...
  ...

Figure 16: Example of human feedback for content refinement.

Conversation Snapshot between Human And Helper LLM (Topic: In-Context Learning)

Human: Give me some suggestions on refining the section title.

Helper LLM: Refining section titles can significantly enhance clarity. Suggestions:

- **Background and Definitions** → *Foundations and Evolution: Defining Key Concepts in Machine Learning*
  Rationale: Emphasizes both the foundational definitions and the historical evolution of the concepts.
- **In-Context Learning** → *In-Context Learning: Mechanisms, Benefits, and Challenges*
  Rationale: Clearly outlines the focus on mechanisms, benefits, and challenges, making it easier for readers to know what to expect.
- **Few-Shot Learning** → *Few-Shot Learning: Bridging the Gap with Minimal Data*
  Rationale: Highlights the essence of few-shot learning, which is achieving results with minimal data.
- **Contextual Understanding** → *Contextual Understanding: Enhancing Model Comprehension and Interpretation*
  Rationale: Focuses on the enhancement of comprehension and interpretation, which are key aspects of contextual understanding.
- **Transfer Learning** → *Transfer Learning: Leveraging Pre-trained Models for New Horizons*
  Rationale: Emphasizes the use of pre-trained models to explore new tasks, capturing the essence of transfer learning.
- **Prompt Engineering** → *Prompt Engineering: Crafting Inputs for Optimal Model Performance*
  Rationale: Focuses on the crafting aspect of prompt engineering, which is crucial for optimizing model outputs.

Outline Comparison (Topic: In-Context Learning)

Original Draft Outline.

- Introduction
- Background and Definitions
- In-Context Learning
- Few-Shot Learning
- Contextual Understanding
- Transfer Learning
- Prompt Engineering
- Conclusion

Final Draft Survey Outline.

- *Introduction*
- *Foundations and Evolution: Defining Key Concepts in Machine Learning*
- *In-Context Learning: Mechanisms, Benefits, and Challenges*
- *Few-Shot Learning: Bridging the Gap with Minimal Data*
- *Contextual Understanding: Enhancing Model Comprehension and Interpretation*
- *Transfer Learning: Leveraging Pre-trained Models for New Horizons*
- *Prompt Engineering: Crafting Inputs for Optimal Model Performance*
- Conclusion

Figure 17: Example of human–LLM feedback for outline refinement.

Report Generated by The Multi-Agent Discussion (Excerpt)

Equal Weighting vs. Adaptive Weighting
The current rubric employs equal weighting for all dimensions, reflecting their collective importance in producing high-quality survey papers. However, this approach may not always capture the nuanced needs of different topics or venues. For instance, some papers may require more emphasis on Synthesis if they aim to introduce new frameworks, while others might prioritize Critical Analysis to challenge existing methodologies. Adaptive weighting, which adjusts the importance of each dimension based on the paper's focus or the venue's expectations, could provide a more tailored evaluation. This approach would allow for a dynamic assessment that better aligns with the specific goals and context of each survey paper.
Proposing Smarter Constructs
To enhance the evaluation framework, several smarter constructs could be considered:
1. **Topic/Venue-Adaptive Weights**: Implementing adaptive weights that vary based on the topic or venue could ensure that the evaluation criteria are aligned with the specific objectives and standards of different research areas.
2. **Uncertainty Measures**: Incorporating uncertainty measures could provide insights into the confidence of assessments, particularly in dimensions like Relevance and Critical Analysis where subjective judgments play a significant role.
3. **Reviewer–LLM Agreement**: Analyzing the agreement between human reviewers and language models (LLMs) could offer a novel perspective on the consistency and reliability of evaluations, potentially identifying areas where human judgment diverges from automated assessments.
4. **Interaction Effects**: Exploring interactions between dimensions, such as how Structure influences the effectiveness of Synthesis, could uncover deeper insights into the interdependencies that contribute to a paper's overall quality.
Research Directions
Several concrete research directions emerge from this discussion:
- **Developing Adaptive Weighting Systems**: Future research could focus on creating algorithms that dynamically adjust dimension weights based on the paper's topic and venue. This would involve analyzing historical data to identify patterns and preferences specific to different research communities.
- **Integrating Uncertainty Metrics**: Research could explore the development of uncertainty metrics for each dimension, providing evaluators with tools to express confidence levels in their assessments and identify areas of ambiguity.
- **Automating Agreement Analysis**: Investigating the alignment between human and LLM evaluations could lead to the development of automated systems that flag discrepancies and suggest areas for further review.
- **Studying Dimension Interactions**: Conducting empirical studies to examine the interaction effects between dimensions could yield insights into how different aspects of a paper contribute to its perceived quality, informing more holistic evaluation strategies.
In conclusion, while the current rubric provides a solid foundation for evaluating survey papers, there is potential for significant enhancements through adaptive weighting, uncertainty measures, and the exploration of dimension interactions. These advancements could lead to more nuanced and context-sensitive evaluations, ultimately improving the quality and impact of survey papers in the field of machine learning.

Figure 18: Insights reveals by the multi-agent discussion.

Evaluation Prompt: Content – Coverage

Here is an academic survey about the topic "{topic}":

```
---
{content}
---
```

<instruction> Please evaluate this survey about the topic {topic} based on the criterion provided below, and give a score from 1 to 5 according to the score descriptions.

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

Score Descriptions:

- **Score 1:** The survey has very limited coverage, only touching on a small portion of the topic and lacking discussion on key areas.
- **Score 2:** The survey covers some parts of the topic but has noticeable omissions, with significant areas either underrepresented or missing.
- **Score 3:** The survey is generally comprehensive in coverage but still misses a few key points that are not fully discussed.
- **Score 4:** The survey covers most key areas of the topic comprehensively, with only very minor topics left out.
- **Score 5:** The survey comprehensively covers all key and peripheral topics, providing detailed discussions and extensive information.

Return the score without any other information.

Figure 19: Evaluation prompt for content coverage.

Evaluation Prompt: Content – Structure

Here is an academic survey about the topic "{topic}":

```
---
{content}
---
```

<instruction> Please evaluate this survey about the topic {topic} based on the criterion provided below, and give a score from 1 to 5 according to the score descriptions.

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

Score Descriptions:

- **Score 1:** The survey lacks logic, with no clear connections between sections, making it difficult to understand the overall framework.
- **Score 2:** The survey has weak logical flow with some content arranged in a disordered or unreasonable manner.
- **Score 3:** The survey has a generally reasonable logical structure, with most content arranged orderly, though some links and transitions could be improved (e.g., repeated subsections).
- **Score 4:** The survey has good logical consistency, with content well arranged and natural transitions, only slightly rigid in a few parts.
- **Score 5:** The survey is tightly structured and logically clear, with all sections and content arranged most reasonably, and transitions between adjacent sections smooth without redundancy.

Return the score without any other information.

Figure 20: Evaluation prompt for content structure.

Evaluation Prompt: Content – Relevance

Here is an academic survey about the topic "{topic}":

```
---
{content}
---
```

<instruction> Please evaluate this survey about the topic {topic} based on the criterion provided below, and give a score from 1 to 5 according to the score descriptions.

Criterion Description (Relevance): Relevance measures how well the content of the survey aligns with the research topic and maintains a clear focus.

Score Descriptions:

- **Score 1:** The content is outdated or unrelated to the field it purports to review, offering no alignment with the topic.
- **Score 2:** The survey is somewhat on topic but with several digressions; the core subject is evident but not consistently adhered to.
- **Score 3:** The survey is generally on topic, despite a few unrelated details.
- **Score 4:** The survey is mostly on topic and focused; the narrative has a consistent relevance to the core subject with infrequent digressions.
- **Score 5:** The survey is exceptionally focused and entirely on topic; the article is tightly centered on the subject, with every piece of information contributing to a comprehensive understanding of the topic.

Return the score without any other information.

Figure 21: Evaluation prompt for content relevance.

Evaluation Prompt: Citation

– Claim: {claim} —
Source: {source} —
Claim: {claim} —
Is the Claim faithful to the Source? A Claim is faithful to the Source if the core part in the Claim can be supported by the Source.
Only reply with 'Yes' or 'No'.

Figure 22: Evaluation prompt for citation.