ARISE: Agentic Rubric-Guided Iterative Survey Engine for Automated Scholarly Paper Generation

Zi Wang[1], **Xingqiao Wang**[1], **Sangah Lee**[2], **Xiaowei Xu**[1*]

[1] University of Arkansas at Little Rock
[2]Seoul National University
zwang@ualr.edu, xwang1@ualr.edu, sanalee@snu.ac.kr, xwxu@ualr.edu

Abstract

The rapid expansion of scholarly literature presents significant challenges in synthesizing comprehensive, high-quality academic surveys. Recent advancements in agentic systems offer considerable promise for automating tasks that traditionally require human expertise, including literature review, synthesis, and iterative refinement. However, existing automated survey-generation solutions often suffer from inadequate quality control, poor formatting, and limited adaptability to iterative feedback—core elements intrinsic to scholarly writing.

To address these limitations, we introduce ARISE, an Agentic Rubric-guided Iterative Survey Engine designed for automated generation and continuous refinement of academic survey papers. ARISE employs a modular architecture composed of specialized large language model agents, each mirroring distinct scholarly roles, such as topic expansion, citation curation, literature summarization, manuscript drafting, and peer-review-based evaluation. Central to ARISE is a rubric-guided iterative refinement loop where multiple reviewer agents independently assess manuscript drafts using a structured, behaviorally anchored rubric, systematically enhancing the content through synthesized feedback.

Evaluating ARISE against state-of-the-art automated systems and recent human-written surveys, our experimental results demonstrate superior performance, achieving an average rubric-aligned quality score of 92.48. ARISE consistently surpasses baseline methods across metrics of comprehensiveness, accuracy, formatting, and overall scholarly rigor. All code, evaluation rubrics, and generated outputs are provided openly at https://github.com/ziwang11112/ARISE.

Introduction

Recent advances in agentic AI have demonstrated the potential of large language model (LLM) agents to collaborate, reason, and solve complex tasks by mirroring human workflows. By orchestrating multiple specialized agents in a modular, feedback-driven manner, agentic systems offer a powerful paradigm for automating labor-intensive processes that traditionally require expert human collaboration, and have already proven effective in domains such as code generation, multi-step planning, and academic research systems.

*Corresponding author.

Survey paper writing is a promising application for agentic systems, as a subdomain of academic research systems, it demands coordinated, expert-level reasoning and reflects the complexity of real-world scholarly workflows. With scientific literature expanding rapidly (Conde et al. 2024), synthesizing new developments into high-quality surveys has become increasingly challenging. Recent studies have explored automating survey paper generation with LLMs, combining retrieval, structuring, and generation in end-to-end pipelines. Notable approaches such as *AutoSurvey*(Wang et al. 2024), *SurveyX*(Liang et al. 2025), and *SurveyForge* (Yan et al. 2025) demonstrate promising results by leveraging LLM-assisted writing and hybrid retrieval techniques.

While recent literature review automation methods offer promising performance (Wu et al. 2025), they still have notable limitations. For example, these approaches commonly rely on preprint-heavy sources, generate survey paper within one pass running and lacking efficient evaluation standards. As a result, ensuring high-quality, reliable references is challenging, as retrieval agents often surface outdated, non-peer-reviewed, or low-quality sources. In addition, real-world survey writing is inherently iterative, requiring multiple cycles of review and revision—a process not well captured by existing single-pass systems. Finally, although benchmarking and evaluation protocols are improving, comprehensive rubrics and peer-review-style feedback remain rare, limiting the interpretability, transparency, and reproducibility of automated outputs.

To address these limitations, we propose **ARISE**, an agentic system that decomposes the academic survey writing process into specialized LLM-powered agents, each mirroring a distinct human role. It employs a *citation-first, article-level* retrieval pipeline that prioritizes references from reputable, peer-reviewed journals and conferences, and produces fully editable LaTeX manuscripts with structured bibliographies tailored to target publication venues. Dedicated validation agents at every stage ensure factual accuracy and minimize hallucinations. ARISE further incorporates a rubric-guided, multi-agent iterative refinement loop: powerful LLMs serve as distinct reviewer agents, each applying a shared evaluation rubric to assess system outputs and generate structured feedback. This feedback is synthesized by the refinement module, driving systematic and interpretable improve-

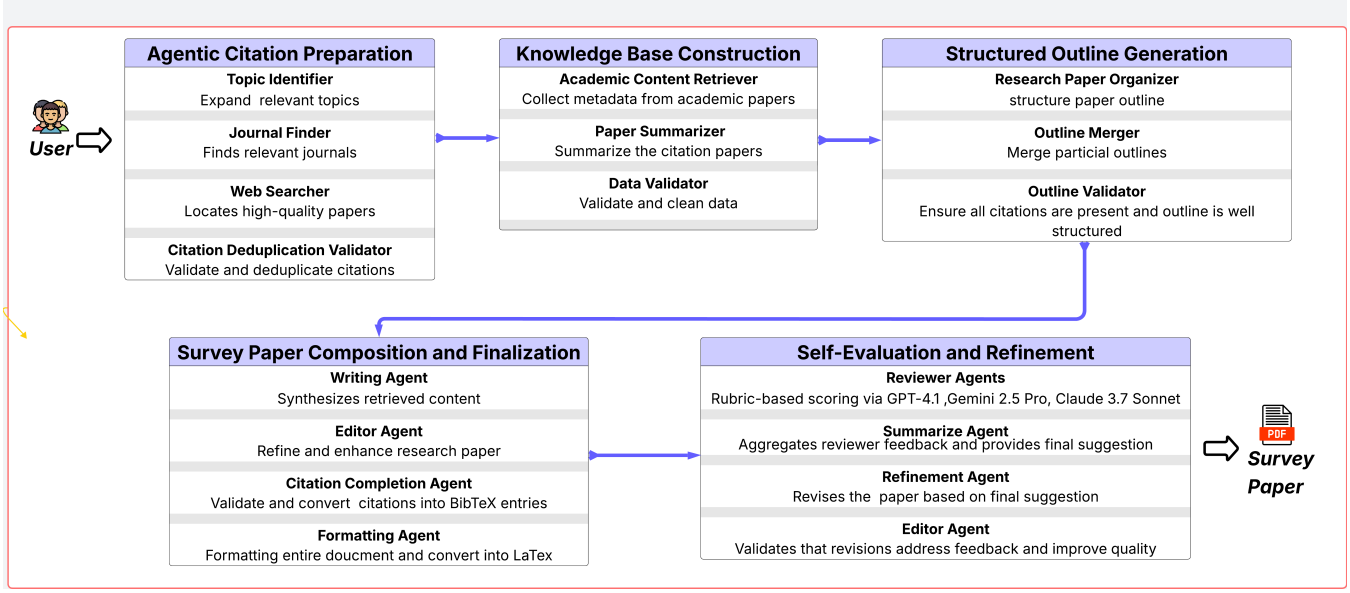

Figure 1: Agent roles and functionality in ARISE. Each agent is assigned a specific function in the modular survey generation and refinement pipeline. All non-reviewer agents run on GPT-4.1 by default. Reviewer agents use a cross-family judge pool (GPT-4.1(optional), Gemini 2.5 Pro, and Claude 3.7 Sonnet).

ments through multiple revision cycles while ensuring transparency and reproducibility.

Our main contributions are:

- We present **ARISE**, an agentic system that automates end-to-end survey generation and peer-review, supporting full-cycle scholarly workflows with modular, specialized LLM agents.
- We introduce a citation-first, rubric-guided, multi-agent iterative refinement framework that produces template flexible outputs, enabling systematic and transparent quality improvement through structured reviewer feedback.
- We introduce an extensible, behaviorally anchored rubric and evaluation framework that mirrors human peer review, enabling transparent, reproducible, and customizable assessment.

Related Work

Agentic System Collaborative Frameworks

Recent advances in large language models have spurred the development of frameworks for orchestrating agentic systems, including CrewAI (Contributors 2024b), AutoGen (Wu et al. 2023), and LangGraph (Contributors 2024a). These toolkits provide abstractions for designing and coordinating teams of LLM-powered agents in hierarchical, collaborative, and conversational workflows.

Further progress includes frameworks for emergent behavior and specialized capabilities, such as HuggingGPT (Shen et al. 2023), MM-Agent (Liu et al. 2025), AgentVerse (Chen et al. 2023), and Any-Agent (Ye et al. 2025), enabling dynamic tool use, web-based interaction, and multimodal collaboration. Additional research explores communicative agent systems for software engineering (Qian, Wang et al. 2024; He, Treude, and Lo 2025), collective reasoning (Du et al. 2023), and open challenges in cooperative AI (Dafoe et al. 2021), as well as scientific discovery (Ghafarollahi and Buehler 2025).

These advances have made it increasingly feasible to design, prototype, and benchmark multi-agent LLM systems for tasks such as collaborative writing, complex reasoning, tool use, and structured document generation. Our work builds on this foundation, leveraging agentic orchestration and modular task decomposition for robust and extensible academic survey generation.

LLM-Based Survey Generation

Recent systems have explored automating survey paper generation with large language models (LLMs), combining retrieval, structuring, and generation in end-to-end pipelines. *AutoSurvey* (Wang et al. 2024) follows a four-phase pipeline with arXiv-based retrieval and LLM-assisted writing, but relies solely on preprints and offers limited source curation. *SurveyX* (Liang et al. 2025) enhances retrieval via hybrid keyword expansion and semantic filtering with an *AttributeTree* citation structure, though it could be improved in modularity and user control. *SurveyForge* (Yan et al. 2025) uses heuristic templates and a memory-driven Scholar Navigation Agent (SANA), and introduces the SurveyBench benchmark for holistic evaluation, but remains constrained by a static architecture and preprint-heavy sourcing.

LLM Evaluation and Rubric Design

Evaluation of LLM-generated outputs remains a major challenge (Chang et al. 2023; Guo et al. 2023; Laskar et al. 2024), especially for complex tasks like survey writing, due

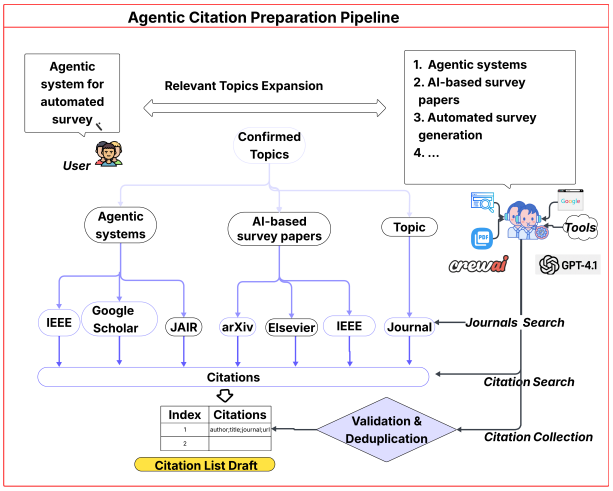

Figure 2: Agentic citation preparation pipeline. Users shape topics and guide domain scoping; agents handle retrieval, filtering, and validation.

to subjectivity, inconsistency, and ambiguity in assessment criteria. Several recent works have addressed these issues by proposing more comprehensive, behaviorally anchored rubrics and checklists (Messer et al. 2024; Lee et al. 2025). Peer-review guidelines from organizations such as IEEE and ACL (IEEE 2020; ACL Rolling Review 2025) provide additional best practices for rubric construction and reviewer alignment.

Methodology

System Overview

Figure 1 presents the modular architecture of ARISE, designed for automated survey generation and iterative self-improvement. We target a systematic mapping review with article-level screening and structured rubric-guided synthesis. Each module consists of one or more dedicated LLM agents, assigned to perform specific scholarly tasks such as topic expansion, citation discovery and validation, literature summarization, outline drafting, manuscript generation, and peer review. In total, ARISE orchestrates 22 specialized agents, including subagents for title and abstract generation, section-level summarization, and citation completion. Some roles (e.g., "Summarizer") are instantiated with module-specific prompts and objectives, reflecting the unique requirements of each pipeline stage. A complete roster of agents, along with their task descriptions and configurations, is provided in our supplementary material to support transparency and reproducibility.

Citation Preparation

Figure 2 illustrates the agentic citation preparation pipeline in ARISE. The process begins with the user specifying an initial survey theme (e.g., "agentic systems for automated survey generation"). An expansion agent then proposes semantically related subtopics, which the user can further refine or approve.This interactive expansion step determines the conceptual scope of the survey and ensures both thematic breadth and alignment with user goals.

Once topics are confirmed, a *domain-scoping* agent suggests publication venues that are *topically appropriate* for each subtopic, including publisher portals (e.g., IEEE Xplore, Elsevier/ScienceDirect), academic search/indexing services (e.g., Google Scholar, Crossref, Semantic Scholar), and open-access repositories (e.g., arXiv). The goal of this step is to avoid obviously unrelated domains (e.g., business or biology journals when the topic is AI/ML), not to use venue prestige as a proxy for article quality. In other words, venues act as *filters on field*, while inclusion and ranking decisions remain at the *article level*.

For each (topic, source) pair, a citation retrieval agent gathers candidate references enriched with metadata such as authors, title, venue, year, and URL. Candidates retrieved from different sources are unified and normalized, then passed through automatic de-duplication and formatting validation to remove near-duplicates and malformed entries. The resulting curated citation list forms the basis for downstream knowledge base construction and outline generation.

Structured Knowledge Base Construction

Figure 3 presents the structured knowledge base construction pipeline in ARISE. The process begins with the curated citation list from the previous phase or user-provided references. For each citation, the system first attempts direct retrieval via stored URLs. If a direct link fails due to paywalls or missing resources, a fallback metadata search is performed using author and title information on platforms such as Google Scholar or arXiv.

If full or partial content (such as the abstract or introduction) is successfully retrieved, an agent summarizes the text into a concise, contribution-focused entry. If all retrieval attempts fail, the citation is recorded in an *Error List* for later review and possible reprocessing.

All retrieved summaries are then deduplicated and validated before being indexed into a persistent knowledge base, organized by citation index `refN`. This knowledge base forms the factual backbone for downstream modules, ensuring that the subsequent outline and paper composition phases are grounded in coherent, context-rich information.

Citation-Keyed Memory (CKM). Beyond serving as a passive repository, the knowledge base functions as a *citation-keyed memory*: each entry is stored as a key–value pair ($\texttt{refN} \rightarrow$ summary). During drafting and refinement, sections query CKM only with the citation keys they already use (i.e., $\texttt{cite}(S)$ for a section S), and the system injects just those summaries into the prompt. This design minimizes irrelevant context and interference while preserving traceability from generated text back to its evidence sources.

Structured Outline Generation

To support coherent and citation-grounded survey writing, the system uses a team of agents to build a structured outline based on the knowledge base. As shown in Figure 4, the process starts from a knowledge base containing N cleaned citation summaries. These are divided into mini-batches, and

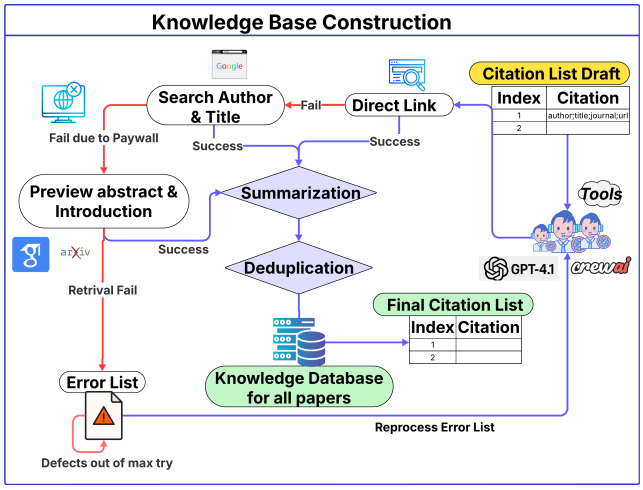

Figure 3: Structured Knowledge Base Construction. Each citation is processed via direct or fallback retrieval, summarized, deduplicated, and indexed in a persistent database aligned with the citation index.

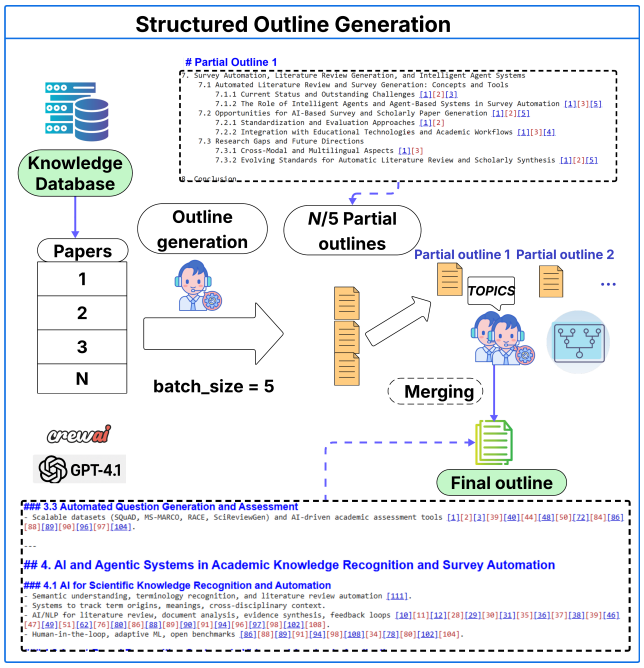

Figure 4: Structured outline generation component. Citation summaries are grouped, outlined, and iteratively merged to form a thematically coherent, citation-preserving global structure.

each batch is processed by a Writing Agent, which generates a partial outline with sections, subsections, and explicit citation index references (e.g., [1][2][3]).

The partial outlines are then passed to a Merging Agent, which combines them in pairs to form progressively larger outlines. After each merge, a Validation Agent checks that the merged outline is coherent and well organized, and that no newly introduced redundancies or obvious gaps appear.

Citation-Preserving Outline Synthesis (CPOS). Beyond semantic coherence, we enforce a *citation-preserving invariant* during merging: for any pair of outlines A and B with citation index sets $\mathrm{cite}(A)$ and $\mathrm{cite}(B)$, the merged outline C must satisfy $\mathrm{cite}(C) = \mathrm{cite}(A) \cup \mathrm{cite}(B)$. After each merge, the Validation Agent compares citation sets and identifies any missing indices; a backfilling step then re-inserts the corresponding references and local context. This merging and validation cycle continues until a single, citation-complete Final Outline is created, and a final check ensures that all original references from the curated citation index are present.

Survey Paper Composition and Finalization

This component completes the initial drafting and formatting stages of the system. It transforms the structured outline and curated knowledge base into a citation-grounded, academically formatted survey.

As shown in Figure 5, the process begins by decomposing the final outline into section-level prompts based on the document hierarchy established during outline generation. For each section S, the system retrieves the relevant citation indices $\mathrm{cite}(S)$ and their associated summaries from the citation-keyed memory (CKM). A Writing Agent then synthesizes these section-specific summaries into coherent, thematically organized prose, ensuring that every claim is anchored in the cited works and that traceability to original sources is preserved. The resulting drafts are passed to an Editor Agent, which improves logical flow and clarity, resolves local redundancies, and inserts placeholders for tables or figures where appropriate. These refined sections are finally merged into a unified draft that respects the outline structure and maintains citation alignment.

Citation and Formatting Hygiene. Figure 6 illustrates the citation completion and LaTeX formatting pipeline. Starting from a citation list that may contain incomplete bibliographic information, a Citation Completion Agent validates each entry and resolves missing metadata fields (e.g., DOIs, venues, years) by querying trusted academic databases and search engines. The agent emits standardized BibTeX entries, which are compiled into a structured bibliography.

A Formatting Agent then enforces *hygiene normalization* on the LaTeX manuscript: it standardizes citation commands (e.g., transforming raw numeric brackets into consistent `\cite{refN}` calls), harmonizes section and table environments, and removes residual artifacts from earlier stages of the pipeline. To ensure professional presentation and reliable compilation, the agent applies structured table environments and consistent style conventions throughout the document. The cleaned LaTeX source can then be compiled into a camera-ready PDF under the target conference or journal template.

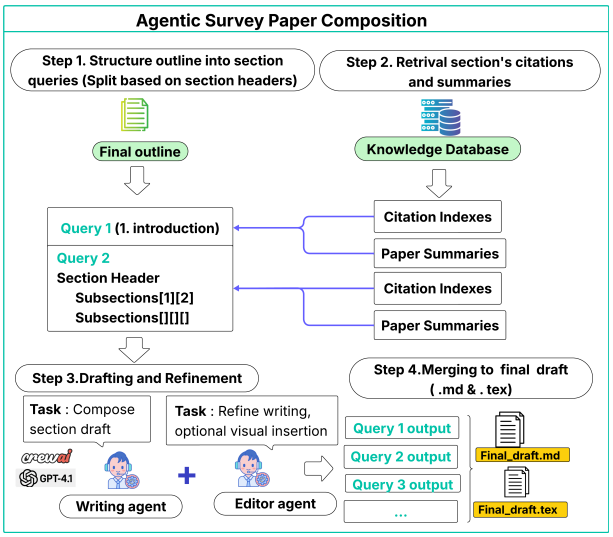

Figure 5: Agentic Survey Paper Composition. Each section query is matched with relevant citations and summaries. A writing agent drafts the content, which is refined by an editor agent before merging into intermediate content outputs.

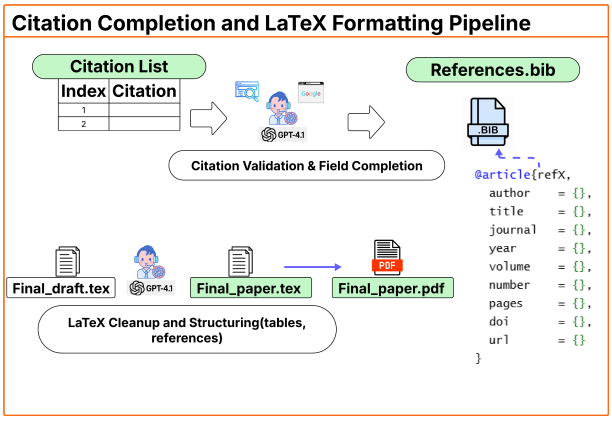

Figure 6: Citation and Formatting Pipeline. The system completes citation metadata, generates BibTeX entries, and standardizes the LaTeX document to produce a clean, structured PDF ready for academic use.

Agentic Rubric-Guided Iterative Refinement Framework

Framework Overview. To mirror real-world peer review, ARISE treats refinement as a *multi-agent control loop*. At each iteration t, the current draft D_t is segmented into contiguous page chunks and independently evaluated by a set of reviewer agents $\mathcal{R}$, each applying the shared rubric $\mathcal{B}$. Reviewer i returns a rubric-based score s_t^i and structured feedback f_t^i. Scores are then aggregated to compute an average quality estimate:

$$\bar{s}_t = \frac{1}{|\mathcal{R}|} \sum_{i \in \mathcal{R}} s_t^i, \tag{1}$$

where $|\mathcal{R}| = N$ is the number of reviewer agents.

If the aggregated score $\bar{s}_t$ meets or exceeds the target threshold τ, the draft is accepted as the final output D^*. Otherwise, a summary agent plays the role of a meta-reviewer: it synthesizes the individual feedback items $\{f_t^i\}$ into an actionable revision plan $\hat{f}_t$. This plan specifies *which sections and issues* to address (e.g., missing related work, weak analysis, unclear structure). A refinement agent and an editor agent then apply the plan to obtain an improved draft D_{t+1}, and the loop repeats until the threshold or another stopping condition (e.g., max rounds) is reached. Appendix 1 shows concrete examples of reviewer feedback and the resulting revision plans.

Evidence-Locked Targeted Revision. Crucially, ARISE couples this control loop with *evidence-locked, section-level revision*. For a given section S, we first recover its set of cited keys $\mathrm{cite}(S)$ (i.e., the `refN` indices that appear in the text) and query the citation-keyed memory (CKM) to obtain only the corresponding summaries (`refN` $\rightarrow$ summary). The refinement agent is instructed to:

- modify *only* the sections that are explicitly targeted in $\hat{f}_t$ (no global free-form rewriting), and
- ground new wording exclusively in the CKM entries for $\mathrm{cite}(S)$, without introducing new, uncited claims or references.

This evidence-locked sectional refinement (ELSR) minimizes hallucination and scope drift by design, while preserving traceability from every revision back to the underlying sources.

Bias Controls and Early Stopping. To reduce LLM-as-judge bias, we use a cross-family reviewer pool and report both a *tri-judge* setting (GPT-4.1, Gemini 2.5 Pro, Claude 3.7 Sonnet) and a setting where the generator family is excluded from the judge set. We also vary the number of reviewers N, the number of refinement rounds t, and the acceptance threshold τ to probe quality–cost trade-offs. In all cases, we log chunk-level scores and aggregated trajectories $\{\bar{s}_t\}_{t=0}^{T}$, enabling analysis of convergence behavior and diminishing returns.

Rubric Construction and Evaluation Rationale. Our shared rubric $\mathcal{B}$ is constructed by synthesizing best practices from established peer-review guidelines—specifically, the IEEE (IEEE 2020) and ACL Rolling Review (ACL Rolling Review 2025)—as well as recent advances in automated, rubric-based evaluation (Messer et al. 2024). We further incorporate insights from frameworks such as CheckEval (Lee et al. 2025), which emphasize subdividing evaluation criteria into explicit, behaviorally anchored sub-aspects to improve reliability and inter-rater agreement in LLM-based assessment.

To ensure consistent and interpretable scoring, $\mathcal{B}$ operationalizes seven core dimensions and twenty subcategories, each with explicit, detailed criteria for every score from 1 to 5 (Table 1). This structure minimizes ambiguity and subjectivity, supporting transparent and reproducible evaluation for both human- and LLM-generated surveys. All criteria and

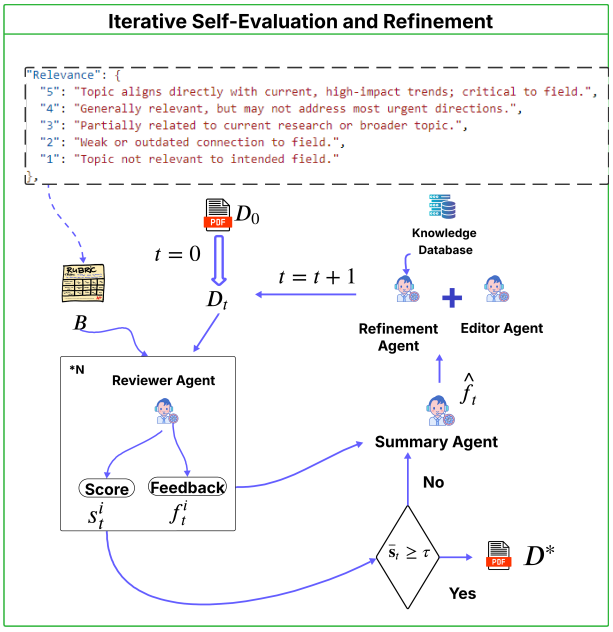

Figure 7: Agentic rubric-guided iterative refinement. At each iteration t, reviewer agents independently score draft D_t using rubric $\mathcal{B}$, producing scores s_t^i and feedback f_t^i. The average score $\bar{s}_t$ is computed; if $\bar{s}_t \geq \tau$, the draft is accepted as D^*. Otherwise, feedback is synthesized into a revision plan that drives targeted updates to produce D_{t+1}. The process repeats until the threshold is met or a stopping condition is reached.

scoring anchors are fully documented (Appendix 2), providing a robust foundation for benchmarking, actionable feedback, and community adaptation to new domains.

Experiment Setup

Benchmarking and Comparison Framework

Human-Written Baselines We benchmark ARISE against ten recently published, human-authored survey papers, selected for topical diversity, publication recency (2023–2025), and high visibility in either arXiv or leading peer-reviewed venues. These baselines were chosen to represent a spectrum of research areas—including LLM reasoning, evaluation, multimodality, time-series modeling, and human-agent interaction—ensuring relevance to our target domains. See Appendix 3 for the full summary table of baseline survey papers. For direct comparison, we use ARISE to generate survey papers on the same research areas as the human-written baselines.

Automated Survey Generation Systems We also evaluate ARISE against prior automated systems, including *SurveyForge* (Yan et al. 2025), *SurveyX* (Liang et al. 2025), and *AutoSurvey* (Wang et al. 2024). Due to availability constraints, we include 10 papers each from SurveyForge and SurveyX, and 3 from AutoSurvey.

Dimension	Subcategories
Scope	Objectives, Relevance, Audience
Literature	Comprehensiveness, Balance, Currency
Analysis	Depth, Integration, Gaps
Originality	Novelty, Advancement, Redundancy Avoidance
Organization	Logical Flow, Section Clarity, Summarization
Presentation	Language, Visuals, Formatting
References	Accuracy, Appropriateness

Table 1: Structure of the shared evaluation rubric: seven dimensions, twenty subcategories, each scored from 1 to 5 (total 100 points).

Experimental Design

ARISE is implemented using CrewAI (Contributors 2024b), with API keys for GPT-4.1 (OpenAI 2023), Gemini 2.5 Pro (Google 2025), and Claude 3.7 Sonnet (Anthropic 2024) managed via environment variables. All system agents run on GPT-4.1 by default. Our pipeline also integrates the Serper API (Serper 2024) for web search and source scraping(see Appendix 5 for cost and time details). The rubric, shown in Table 1, defines twenty subcategories across seven core dimensions, and all reported scores are computed using the aggregate formula given in Eq. (1).

To ensure full-document coverage and minimize positional bias, we segment each paper into contiguous 3-page chunks, encouraging balanced attention across sections and ensuring each chunk fits within the LLMs' context windows. Chunk-level reviews produce localized rubric scores, enabling fine-grained analysis of writing quality and document structure. All outputs from ARISE, competing automated systems, and human-written baselines are evaluated using the same domain-agnostic rubric, chunking strategy, and tri-model reviewer setup.

Beyond automated evaluation, we include a human expert study to validate perceived quality improvements after iterative refinement. We also perform a model-based ablation, comparing end-to-end system performance with both small (`gpt-4.1-mini`) and large (`gpt-4.1`) language models as pipeline agents. Finally, we conduct a citation traceability audit to assess reference reliability and report inter-rater agreement using Krippendorff's Alpha.

Results and Analysis

Overall Performance.

Our system, **ARISE**, achieves the highest overall quality among all five evaluated systems. As shown in Table 2, On our agentic rubric-based evaluation framework, ARISE's outputs receive higher scores than all baselines across each individual reviewer—Claude 3.7, Gemini 2.5 Pro, and GPT-4.1—with a tri-judge average of **92.48** (Table 2). To check for potential self-judging bias, we also report a bi-judge setting that *excludes* GPT-4.1 from the reviewer pool (Bi-judge (G+C) column); ARISE remains the top-performing system (92.43 vs. the next best 87.58), indicating that our conclusions do not depend on using the generator family as a judge.

System	GPT-4.1	Gemini	Claude	Tri-judge Avg	Bi-judge Avg
ARISE	**92.57**	**92.59**	**92.27**	**92.48**	**92.43**
AutoSurvey	81.87	81.74	83.76	82.46	82.75
Baseline	86.58	85.81	86.06	86.15	85.94
SurveyForge	87.88	87.51	87.64	87.68	87.58
SurveyX	81.13	81.99	81.82	81.65	81.91

Table 2: Mean TOTAL scores by system and reviewer. "Tri-judge Avg" uses all three reviewers (GPT-4.1, Gemini 2.5 Pro, Claude 3.7 Sonnet). "Bi-judge Avg (G+C)" excludes the generator family and averages only Gemini 2.5 Pro and Claude 3.7 Sonnet.

Rubric-Level Superiority

ARISE demonstrates consistently strong rubric-level performance, leading in all seven categories evaluated (Table 3). In critical dimensions such as *Literature* (4.95), *Presentation* (4.84), *References* (4.98), and *Organization* (4.82), ARISE surpasses both human-written baselines and recent automated systems. These gains are attributed to our agentic refinement strategy and explicit modular writing design.

Category	**ARISE**	Autosurvey	Baseline	SurveyForge	SurveyX
Analysis	**4.56**	3.94	3.74	4.45	3.64
Literature	**4.95**	4.53	4.70	4.79	4.29
Organization	**4.82**	4.26	4.52	4.55	4.41
Originality	**4.44**	3.88	3.97	4.10	3.68
Presentation	**4.84**	3.57	4.35	3.87	4.22
References	**4.98**	4.81	4.94	4.85	4.45
Scope	**4.30**	4.10	4.14	4.23	4.00

Table 3: Mean reviewer scores by rubric category and system

ARISE's consistently high scores across all rubric dimensions confirm the effectiveness of our modular agent design and rubric-guided refinement strategy. Inter-rater reliability among the reviewers was consistently high across all systems, with Krippendorff's Alpha (α) exceeding **0.966** in all cases and reaching up to **0.987**—see Appendix 4 for full agreement scores.

These results validate our core hypothesis: rubric-guided iterative refinement enables transparent, interpretable, and high-quality survey generation. By integrating modular agent roles, structured evaluation criteria, and multi-round refinement, ARISE achieves higher rubric scores than baseline and other generation systems, and establishes a reproducible foundation for self-improving academic writing.

Validating the Refinement Process with a Domain-Specific Example

To illustrate the impact of ARISE's rubric-guided iterative refinement loop, we present a representative refinement trajectory for a system-generated survey paper in the domain of *LLM reasoning and replication*. This example tracks how quality evolves over successive refinement iterations based on rubric-guided feedback from three independent reviewer agents.

Round	**GPT-4.1**	**Gemini 2.5 Pro**	**Claude 3.7 Sonnet**	**Average**
0	86.8	87.9	86.2	87.0
1	90.5	89.8	90.0	90.1
2	93.8	89.2	91.3	91.4
3	92.3	92.8	92.9	92.7

Table 4: A case of average review score progression across rubric-guided refinement rounds for a generated survey. The target average score is 92.0.

As shown in Table 4, the case begins with an average reviewer score of 87.0. By Round 3, the average score exceeds the threshold, reaching 92.7. This trajectory demonstrates ARISE's ability to iteratively elevate content quality through modular, reviewer-guided revision. See the supplementary material for results from all other topic domains.

Human Evaluation

To assess the effectiveness of our agentic refinement process, we conducted a human evaluation with four experts (two professors, one postdoc, one PhD student). Each expert independently reviewed five system-generated survey papers, both *before* and *after* iterative refinement, using the same 20-subcategory rubric as in our automated evaluation.

Results. Table 5 summarizes the scores. On average, the total score increased from 70.2 to 83.7 out of 100 , and the mean subcategory rating rose from 3.51 to 4.18 out of 5. All topics and reviewers observed substantial, consistent improvement. After refinement, system outputs consistently achieved "strong" ($\geq$4.0) expert ratings.

Topic	**Before** (Total)	**After** (Total)	**Before** (Avg)	**After** (Avg)
ASG LitRev	69.8	80.5	3.49	4.03
GenAI in Manufacturing	68.8	82.8	3.44	4.14
LLM Reasoning	66.3	86.0	3.31	4.30
Clustering Indexing	70.5	82.0	3.53	4.10
Retrieval-Aug. Gen.	75.8	87.0	3.79	4.35
Average	70.2	83.7	3.51	4.18
Improvement (%)	–	**19.2%**	–	**19.2%**

Table 5: Human evaluation scores for five system-generated surveys (N=4 experts per paper). "Total" is out of 100; "Avg" is per-rubric average out of 5.

These results confirm that rubric-guided refinement substantially and reliably improves survey quality as perceived by domain experts. Additional human evaluations are detailed extensively in Appendix 6.

Model-Based Ablation Study

To assess how agent capacity influences end-to-end survey generation quality, we conducted an ablation study using two full system configurations: one with `gpt-4.1-mini`

as the primary agent in all modules, and another with the larger `gpt-4.1`. For each configuration, we generated complete survey papers and evaluated them across refinement rounds using the same rubric-guided iterative process.

As shown in Table 6, both configurations demonstrate quality improvements through the iterative refinement loop. The system powered by the larger model achieved a higher final rubric score and greater improvement, reflecting the advantages of increased model capacity for both generation and refinement tasks. Even with the smaller model configuration, ARISE achieves an average final rubric score of **88.04**, which remains slightly higher than those of prior automated systems and human-written baselines, as shown in Table 2.

System Model	Initial	Final	Improvement (%)
`gpt-4.1-mini`	83.09	88.04	**5.96%**
`gpt-4.1`	86.53	92.48	**6.88%**

Table 6: Average per-reviewer rubric score across refinement rounds for system-generated papers using small and large models.

Reference Reliability

To validate the reliability of ARISE's citation preparation process, we conducted a traceability audit of the final system-generated references. We define the *Expanded Citation Traceability Rate* (eCTR) as:

$$\text{eCTR} = \frac{V}{T}, \quad \text{Hallucination Rate} = 1 - \text{eCTR} \qquad (2)$$

where V is the number of verifiable citations successfully matched to external databases, and T is the total number of citations extracted from the system-generated PDF.

We applied layout-aware reference extraction (via PyMuPDF) to final PDF outputs and matched each citation against CrossRef, Semantic Scholar, and arXiv using public APIs. Across all evaluated ARISE outputs, we observed a perfect mean eCTR of **1.00**, corresponding to a hallucination rate of **0.00**. This result demonstrates the robustness of our citation-first pipeline in producing factually grounded scholarly references.

Conclusion

We present ARISE, a modular, agentic system for automated academic survey generation and iterative refinement. By decomposing the survey writing and peer review process into specialized LLM-powered agents, ARISE delivers transparent, reproducible, and high-quality scholarly outputs, and effectively addressing longstanding challenges in quality control, formatting, and iterative improvement. Our experiments show that ARISE achieves higher rubric scores than both human-written baselines and state-of-the-art automated survey generation systems across a comprehensive, behaviorally anchored evaluation rubric. The system demonstrates near-perfect reference reliability and substantial quality improvements through rubric-guided, multi-agent iterative refinement. Further discussion of broader impact, ethical statement, and limitations is provided in Appendix 5.

Appendix

1. Rubric-Guided Iterative Refinement Sample

Each round of refinement in our framework is guided by structured feedback from multiple reviewer agents, who independently evaluate each manuscript draft using a detailed, standardized rubric (see Table A4). Agents provide both quantitative scores and qualitative suggestions; representative reviewer outputs and sample revision recommendations are compiled in Tables A1–A3. These records illustrate the diversity and depth of agentic assessment throughout the iterative improvement process described in the main text.

For full transparency and reproducibility, all generated survey drafts, reviewer feedback, meta-review synthesis, and the codebase for executing rubric-driven evaluation and revision are provided in the supplementary materials (see each topic's subfolder `review_output`).

Generated Survey Sample

Our ARISE system assists academic researchers by generating high-quality PDF drafts using the full `TeX Live` library, which supports a wide variety of LaTeX classes for major journals and conferences (e.g., AAAI, ACM, IEEE, Springer, Elsevier, and more). The title and abstract for each draft are autonomously generated by dedicated agents based on the complete manuscript content, ensuring that each paper features a contextually relevant and well-structured summary. The modular formatting and finalization module enables users to specify their preferred style, automatically adapting the manuscript structure, citation format, and page layout to the conventions of the target venue. ARISE manages all technical aspects of LaTeX formatting, including title, abstract, metadata, tables, and references, allowing researchers to focus on the intellectual content rather than formatting details. This workflow streamlines the preparation of survey drafts in LaTeX, saving substantial time and effort when initiating new research projects. An example output generated by ARISE is illustrated in Figure A1.

Reference Sample

To illustrate the professional formatting quality and citation currency of ARISE-generated manuscripts, we provide a reference sample generated by our system on the topic of *Clustering, Indexing, and Data Structures for High-Dimensional and Categorical Data* (see Figure A2).

ARISE emphasizes citations from peer-reviewed journals and established venues when available, while still incorporating relevant preprints for timeliness. This approach enhances the credibility, accuracy, and academic rigor of the generated surveys compared to systems predominantly reliant on preprints.

Reasoning, Replicability, and Benchmarking in Large Language and Foundation Models: Methodologies, Challenges, and Pathways Toward Trustworthy, Interpretable, and Inclusive AI

Abstract

This survey provides a comprehensive synthesis of recent advances, methodologies, and enduring challenges in the development, evaluation, and responsible deployment of large language models (LLMs) and foundation models. Motivated by the transformative impact of LLMs across natural language processing, scientific discovery, and real-world applications, the paper critically examines the evolution from symbolic and neural paradigms through contemporary transformer-driven and neurosymbolic architectures, highlighting emergent reasoning capabilities and the drive towards human-like abstraction. The review systematically analyzes benchmarking ecosystems, probing frameworks, and evaluation metrics, emphasizing the limitations of prevailing practices in capturing semantic faithfulness, compositionality, and real-world reasoning, particularly on multistep, cross-modal, and domain-specific tasks. Key contributions include a structured taxonomy of model architectures and fusion strategies, an assessment of hybrid approaches integrating neural, symbolic, and graph-based reasoning, and comparative analyses of benchmark methodologies across linguistic, reasoning, and multimodal domains.

The survey underscores persistent gaps in robustness, interpretability, fairness, and reproducibility—drawing attention to vulnerabilities in adversarial and out-of-distribution scenarios, challenges in auditability and demographic inclusion, and the ongoing reproducibility crisis stemming from inadequate reporting and opaque "language-models-as-a-service" paradigms. It highlights advances in adaptive prompting, modular workflow orchestration, and explainability, while advocating for open science, FAIR data practices, and transparent, community-driven benchmarking. Strategic recommendations target holistic evaluation protocols, enhanced benchmarking diversity, rigorous auditing, responsible design, and the institutionalization of modular, reproducible workflows. The paper concludes that future progress in LLM research and deployment is contingent upon sustaining openness, modularity, explainability, reproducibility, and ethical responsibility, thus ensuring trustworthy, equitable, and societally beneficial language technologies.

ACM Reference Format:

. 2025. Reasoning, Replicability, and Benchmarking in Large Language and Foundation Models: Methodologies, Challenges, and Pathways Toward Trustworthy, Interpretable, and Inclusive AI. In . ACM, New York, NY, USA, 22 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

1 Introduction

1.1 Overview of Large Language and Foundation Models (LLMs)

The evolution of artificial intelligence (AI) has been profoundly shaped by advances in language understanding and generation. The trajectory spans from symbolic, rule-based systems—characterized by explicit grammatical rules and formal symbolic manipulation—to statistical methods, neural, and deep learning architectures. Early symbolic approaches excelled in interpretability but were hindered by a lack of scalability and the brittleness of handcrafted rules. The advent of statistical models, and subsequently neural network architectures, marked a paradigm shift by enabling data-driven learning of complex linguistic patterns. This progression culminates in large-scale Transformer-based models, wherein pre-trained language models (PLMs), especially large language models (LLMs), distinguish themselves through scale and the emergence of novel capabilities.

Distinctive behaviors—such as in-context learning and abstract reasoning—emerge in LLMs due not solely to increased parameter counts, but also to innovations in model design, architecture, and training paradigms. Key developments include:

- The adoption of large-scale, unsupervised pre-training;
- Attention mechanisms, as popularized by the Transformer architecture;
- Alignment of model objectives with downstream utility.

The launch and societal integration of models such as ChatGPT exemplify LLMs' transformative effect on not only conventional natural language processing (NLP) tasks, but also on digital interaction, information retrieval, content creation, and scientific discovery. Concomitantly, there has been renewed interest in hybrid algorithmic-neural approaches and neural-symbolic (NeSy) systems. These are motivated by enduring challenges—particularly in reasoning and interpretability—where pure neural architectures, despite their success, fall short [69, 91, 95, 105]. A move toward models exhibiting compositionality and explicit knowledge manipulation reflects the AI community's recognition that human-like reasoning and adaptability may require synthesizing symbolic and subsymbolic learning, an imperative for ongoing advancements toward artificial general intelligence (AGI).

Figure A1: Example first page of a survey paper generated by ARISE, formatted in ACM style. This sample illustrates the structured abstract, clear academic formatting, and section organization achieved by the pipeline. Full outputs for additional topics are provided in the supplementary materials.

[31] Simeon Emanuilov and Aleksandar Dimov. 2024. Billion-scale Similarity Search Using a Hybrid Indexing Approach with Advanced Filtering. *Cybernetics and Information Technologies* 24, 4 (2024), 45–58. doi:10.2478/cait-2024-0035

[32] Johannes Fischer, Tomohiro I, Dominik Köppl, and Kunihiko Sadakane. 2018. Lempel-Ziv Factorization Powered by Space Efficient Suffix Trees. *Algorithmica* 80, 7 (2018), 2048–2081. doi:10.1007/s00453-017-0354-y

[33] T. Gagie, A. Hartikainen, K. Karhu, J. Kärkkäinen, G. Navarro, S. J. Puglisi, and J. Sirén. 2017. Document retrieval on repetitive string collections. *Information Retrieval Journal* 20 (2017), 273–303. doi:10.1007/s10791-017-9297-7

[34] A. J. Gallego, J. R. Rico-Juan, and J. J. Valero-Mas. 2022. Efficient k-nearest neighbor search based on clustering and adaptive k values. *Pattern Recognition* 122 (2022), 108356. doi:10.1016/j.patcog.2021.108356

[35] Z. Gniazdowski. 2024. New Approach to Clustering Random Attributes. *Zeszyty Naukowe WWSI* 19, 31 (2024), 41–90. doi:10.48550/arXiv.2412.09748

[36] E. Gorstein, R. Aghdam, and C. Solís-Lemus. 2025. HighDimMixedModels.jl: Robust high-dimensional mixed-effects models across omics data. *PLOS Computational Biology* 21, 1 (2025), e1012143. doi:10.1371/journal.pcbi.1012143

[37] Ralf Hartmut Güting, Suvam Kumar Das, Fabio Valdés, and Suprio Ray. 2025. Exact Trajectory Similarity Search With N-tree: An Efficient Metric Index for kNN and Range Queries. *ACM Transactions on Spatial Algorithms and Systems* 11, 1 (2025), 5:1–5:54. doi:10.1145/3716825

[38] R. K. Halder, M. N. Uddin, M. A. Uddin, S. Aryal, and A. Khraisat. 2024. Enhancing K-nearest neighbor algorithm: a comprehensive review and performance analysis of modifications. *Journal of Big Data* 11, 1, Article 113 (2024). https://journalofbigdata.springeropen.com/articles/10.1186/s40537-024-00973-y

[39] S. W. Harrar and X. Kong. 2022. Recent developments in high-dimensional inference for multivariate data: Parametric, semiparametric and nonparametric approaches. *Journal of Multivariate Analysis* 188 (2022), Article 104855. doi:10.1016/j.jmva.2021.104855

[40] Muhammad Umair Hassan, Xiuyang Zhao, Raheem Sarwar, Naif R. Aljohani, S. M. M. Rahman, K. Muhammad, and M. A. Raza. 2024. SODRet: Instance retrieval using salient object detection for self-service shopping. *Machine Learning with Applications* 15 (2024), 100523. https://www.sciencedirect.com/science/article/pii/S2666827023000762

[41] Majid Hojati, Rob Feick, Steven Roberts, Carson Farmer, and Colin Robertson. 2023. Distributed spatial data sharing: a new model for data ownership and access control. *Journal of Spatial Information Science* 2023, 27 (2023), 1–26. doi:10.5311/JOSIS.2023.27.220

[42] Zainab Iftikhar, Adeel Anjum, Abid Khan, Munam Ali Shah, and Gwanggil Jeon. 2023. Privacy preservation in the internet of vehicles using local differential privacy and IOTA ledger. *Cluster Computing* 26 (2023), 3361–3377. doi:10.1007/s10586-023-04002-0

[43] F. Iglesias, T. Zseby, and A. Zimek. 2020. Absolute Cluster Validity. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 42, 9 (2020), 2096–2112. https://ieeexplore.ieee.org/document/8695871

[44] Bharathi B. K. and K. Jaganathan. 2022. The Intrinsic Structure of High-Dimensional Data According to Principal Graphs. *Mathematics* 10, 20 (2022), 3894. https://www.mdpi.com/2227-7390/10/20/3894

Nearest Neighbors Search. *arXiv preprint arXiv:2412.02448* (2024). https://arxiv.org/abs/2412.02448

[54] J. Lin and A. Trotman. 2017. The role of index compression in score-at-a-time query evaluation. *Information Retrieval Journal* 20 (2017), 274–314. doi:10.1007/s10791-016-9291-5

[55] J. Liu and M. Vinck. 2022. Improved visualization of high-dimensional data using the distance-of-distance transformation. *PLOS Computational Biology* 18, 12 (2022), e1010764. doi:10.1371/journal.pcbi.1010764

[56] Y. Liu, J. Ding, H. Wang, and Y. Du. 2025. A Clustering Algorithm Based on the Detection of Density Peaks and the Interaction Degree Between Clusters. *Applied Sciences* 15, 7 (2025), 1–19. doi:10.3390/app15073612

[57] S. Loodtoy and V. Yalagandula. 2021. Bibliometric Analysis of International Journal of Information Management. *International Journal of Information Management* (2021). http://repo.lib.jfn.ac.lk/ujrr/bitstream/123456789/4718/2/Bibliometric%20Analysis%20of%20International%20Journal%20of%20Information%20Management.pdf

[58] S. Lu, W. Martens, M. Niewerth, and Y. Tao. 2023. Partial Order Multiway Search. *ACM Transactions on Database Systems* 48, 4 (2023), 1–31. doi:10.1145/3626956

[59] Qing Mai, Xin Zhang, Yuqing Pan, and Kai Deng. 2022. A Doubly Enhanced EM Algorithm for Model-Based Tensor Clustering. *J. Amer. Statist. Assoc.* 117, 540 (2022), 2120–2134. doi:10.1080/01621459.2021.1904959

[60] A.-A. Mamun, H. Wu, Q. He, J. Wang, and W. G. Aref. 2024. A Survey of Learned Indexes for the Multi-dimensional Space. *arXiv preprint arXiv:2403.06456* (2024). https://arxiv.org/abs/2403.06456

[61] Magdalen Dobson Manohar, Taekseung Kim, and Guy E. Blelloch. 2025. Range Retrieval with Graph-Based Indices. *arXiv preprint arXiv:2502.13245* (2025). https://arxiv.org/abs/2502.13245

[62] N. Marco, D. Şentürk, S. Jeste, C. C. DiStefano, A. Dickinson, and D. Telesca. 2024. Flexible regularized estimation in high-dimensional mixed membership models. *Computational Statistics & Data Analysis* 194 (2024), 107931. doi:10.1016/j.csda.2024.107931

[63] Jorge Martinez-Gil. 2022. Evaluation of Code Similarity Search Strategies in Large-Scale Codebases. *Machine Learning with Applications* 10 (2022), 100423. https://www.sciencedirect.com/science/article/pii/S2666827022000868

[64] A. Michalopoulos, D. Tsitsigkos, P. Bouros, N. Mamoulis, and M. Terrovitis. 2025. Efficient Distance Queries on Non-point Data. *ACM Transactions on Spatial Algorithms and Systems* 11, 1 (2025), 1:1–1:37. doi:10.1145/3698194

[65] Xiangbo Mo and Hao Chen. 2024. A new classification framework for high-dimensional data. *arXiv preprint arXiv:2306.15199* (2024). https://arxiv.org/abs/2306.15199

[66] Neda Dousti Mousavi, S. Mostafa Hosseini, and Mahdi Mahmoudi. 2023. Categorical Data Analysis for High-Dimensional Sparse Covariates with Multinomial Responses: An RNA-Seq Cancer Application. *Mathematics* 11, 14 (2023), 3202. https://www.mdpi.com/2227-7390/11/14/3202

[67] Abhinav Natarajan, Maria De Iorio, Andreas Heinecke, Emanuel Mayer, and Simon Glenn. 2024. Cohesion and Repulsion in Bayesian Distance Clustering. *J. Amer. Statist. Assoc.* 119, 546 (2024), 1374–1384. doi:10.1080/01621459.2023.2191821

Figure A2: Reference sample generated by ARISE on the topic of *Clustering, Indexing, and Data Structures for High-Dimensional and Categorical Data*, highlighting its prioritization of peer-reviewed journal and conference papers to ensure citation quality and credibility.

Model	Section	Suggestions
gpt-4.1	Abstract, Introduction, and Historical and Foundational Landscape	Expand literature coverage, especially in the benchmarking and reasoning evaluation literature.; Ensure all referenced tables and visuals are present, clear, and visually improve synthesis.; Replace generic numbered citations with a full reference list in the final version for traceability.
gemini-2.5	Abstract, Introduction, Historical and Foundational Landscape (through start of Benchmarking)	Ensure referenced figures/tables (e.g., Table 1) are included in the final version.; Verify full references section for accuracy and formatting.; Strengthen in-chunk summarization using inline tables or boxes if possible to reinforce key comparative points.
claude-3.7	Introduction, Historical and Foundational Landscape	Broaden and deepen the engagement with competing or alternative views where appropriate (e.g., critiques of hybrid models or transformer approaches).; Ensure that figures, tables, and diagrams are present and directly support claims when referenced.; Replace placeholder citation markers with complete bibliography for full submission.; Consider summarizing key takeaways at the end of major sections more explicitly.
gpt-4.1	3.2–4.3 Benchmark Evaluation and Probing Sections	Add an explicit restatement or recap of the overall survey objectives when introducing new major sub-sections.; Provide a few concrete examples where benchmarking volatility misled the field (to deepen critical analysis).; Highlight implications or actionable guidance for benchmark and metric developers.; Consider briefly summarizing emerging benchmarks from late 2023 or 2024, if possible, for currency.
gemini-2.5	3. Benchmarking, Evaluation, and Comparative Analysis	Provide a brief, explicit statement of objectives at the start or end of the section.; Consider enhancing section summaries or explicitly restating takeaways after major analyses.; Integrate more conceptual figures to complement the empirical tables.; Check for correct and non-redundant formatting in citation numbering.
claude-3.7	Benchmarking and Evaluation Paradigms; Probing, Reasoning, and Linguistic Benchmarks	Add an explicit section-level objective statement or overview at the start.; Improve section transitions or provide mini-introductions to major subsections.; Standardize citation formatting in text and ensure consistent reference styling.; Consider including workflow diagrams, paradigm maps, or conceptual illustrations.; Make audience/who-will-benefit aspects clear in introductory text.

Table A1: Rubric-Guided Iterative Refinement Sample

2. Rubric Structure and Scoring Anchors.

To operationalize consistent, high-fidelity review, Table A4 presents the complete scoring rubric $\mathcal{B}$ used throughout our agentic evaluation pipeline. Each dimension and subcategory is defined with explicit, behaviorally anchored criteria for every possible score from 1 (lowest) to 5 (highest). This granular structure—covering scope, literature review, analysis, originality, organization, presentation, and references—translates general peer-review standards into actionable, reproducible evaluation guidance for both human and automated assessment. By making all criteria and scoring anchors fully explicit, the rubric supports transparent benchmarking, facilitates community adaptation, and underpins robust feedback cycles in our iterative refinement framework.

3. Baseline Survey Papers Used for Benchmarking

To ensure fair and meaningful benchmarking, the selection of baseline research domains was constrained by the availability of comparable sample outputs from all three automated survey generation systems evaluated in this study. Within these feasible domains, we aimed for a balanced representation by including five preprint surveys and five peer-reviewed surveys. Table A5 details each baseline's research area, full title with citation, and publication venue or year, providing a transparent reference set for direct comparison against ARISE and prior automated systems.

4. Reliability Evaluation

To assess consistency among reviewer agents, we compute **Krippendorff's Alpha** (α), a standard inter-rater reliability

Model	Chunk	ChunkIndex	Category	Sub-Criterion	Score	Comment (shared for category)
gpt-4.1	Pages 1–3	0	Literature	Comprehensiveness	4	The section references many key works across subfields and uses over 30 citations (references numbering up to 114 are mentioned), but the specific breadth cannot be fully verified as most citation details are numeric placeholders in this excerpt. Coverage appears current, with references to recent methods and issues, but explicit years or full coverage cannot be absolutely confirmed. Discussion is balanced and cites limitations and open problems.
gpt-4.1	Pages 1–3	0	Literature	Balance	5	
gpt-4.1	Pages 1–3	0	Literature	Currency	4	
gemini-2.5	Pages 1–3	0	Literature	Comprehensiveness	5	The section displays strong comprehensiveness, referencing foundational and very recent works (with $\geq$ 30 citations), covering multiple domains (survey generation, LLMs, hybrid workflows). Balance is maintained by discussing benefits and challenges, though the critique could be more granular. Currency is excellent, with up-to-date references.
gemini-2.5	Pages 1–3	0	Literature	Balance	4	
gemini-2.5	Pages 1–3	0	Literature	Currency	5	
claude-3.7	Pages 1–3	0	Literature	Comprehensiveness	5	References are numerous and span many subfields, often citing multiple recent papers in each subtopic ($\geq$30 citations appear, many recent, referencing preprints/conferences). Strengths and weaknesses are discussed, including presenting trade-offs (e.g., hybrid vs. black-box). Multiple perspectives are mentioned, but direct debates between competing schools are limited.
claude-3.7	Pages 1–3	0	Literature	Balance	4	
claude-3.7	Pages 1–3	0	Literature	Currency	5	

Table A2: Detailed rubric scores sample of Literature category for reviewer agents (Pages 1–3, Chunk 0), grouped by agent with shared comments.

metric. It is defined as:

$$\alpha = 1 - \frac{D_o}{D_e}$$

where D_o denotes observed disagreement and D_e denotes expected disagreement by chance. Values range from $-\infty$ to 1, with $\alpha = 1$ indicating perfect agreement.

We compute α under the interval-level setting using rubric scores from three reviewers (GPT-4.1, Gemini 2.5, Claude 3.7) across four systems and a published baseline. All evaluations use the same rubric and chunking protocol.

As shown in Table A6, all systems—including ARISE, the baselines, and prior automated methods—achieved exceptionally high inter-rater reliability, with Krippendorff's Alpha (α) values exceeding 0.96 across the board. This reflects strong consistency in scoring among the independent reviewer agents and further supports the robustness and interpretability of the rubric-guided evaluation protocol adopted in our study. High agreement at both the system and model levels suggests that our rubric design and agent workflow produce reproducible, reliable assessments suitable for benchmarking survey generation quality.

5. Limitations, Broader Impact, and Ethical Statement

Limitations

While ARISE achieves strong empirical performance in generating and evaluating survey papers, several limitations should be acknowledged.

First, our evaluation framework relies entirely on large language models (LLMs) as reviewer agents—specifically GPT-4.1, Gemini 2.5 Pro, and Claude 3.7 Sonnet. Although we adopt a detailed and standardized rubric to promote consistency, we were unable to involve human reviewers due to time and resource constraints. As such, the evaluation may reflect alignment patterns and blind spots specific to current LLMs.

Second, Our system relies on commercial LLM APIs with tiered pricing and computational constraints, such as request

rate limits, context window caps, and quota exhaustion. These factors can intermittently affect pipeline stability or trigger retry logic. We employ both large-scale models (e.g., GPT-4.1, Claude 3.7, Gemini 2.0 Pro) and smaller, cost-efficient alternatives (e.g., GPT-4.1 Mini, Gemini Flash), balancing performance and budget. As shown in Figure A7, larger models tend to offer stronger reasoning and editing capabilities but at significantly higher costs—up to 5–10× per million tokens compared to compact variants. In addition to LLM usage, ARISE integrates paid external APIs such as Serper for web search and scraping, which incurs $0.01 per query beyond the free tier (100/month), making it affordable at an academic scale. For our experiments, total Serper usage cost remained under $200.

In practice, a single paper generation cycle typically costs $10–$20 and takes approximately 3.5 hours, depending on the topics. As topic complexity and breadth grow, additional time is required for citation expansion, data collection, and outline refinement. Notably, our refinement module—which includes agentic evaluation and iterative rewriting—accounts for 30-40 percent of total processing time, especially when multi-round improvement is needed for quality assurance.

Despite these limitations, ARISE maintains a modular, auditable, and reproducible architecture. Future work may address these constraints through lightweight model fine-tuning, asynchronous feedback loops with human-in-the-loop reviewers, or more cost-efficient batching strategies.

Broader Impact

ARISE aims to improve the scalability, structure, and factual consistency of academic survey writing by automating key tasks such as citation preparation, outline construction, and LaTeX formatting through modular agent workflows. Its intended users include researchers, educators, and academic writers seeking assistance in synthesizing large volumes of literature.

Beyond survey writing, ARISE's modular architecture and role-specialized agents can be readily adapted for broader document generation tasks, including grant proposals, technical reports, and educational materials. The rubric-guided feedback loop and structured refinement process generalize well to any domain requiring high-fidelity, structured writing.

The broader impact of ARISE is twofold. Positively, it lowers the barrier to entry for producing well-organized, citation-grounded scholarly outputs. This is especially beneficial in under-resourced research communities or interdisciplinary areas where manual literature synthesis is prohibitively time-consuming. Furthermore, ARISE emphasizes traceable references, rubric-based evaluation, and workflow transparency, promoting responsible deployment and downstream auditing.

However, certain risks must be considered. Over-reliance on automated systems may diminish critical thinking or perpetuate biases inherent in training data. If deployed without expert oversight, ARISE could contribute to derivative content or fail to capture diverse perspectives. To mitigate these risks, ARISE is explicitly designed as an assistive tool—never a replacement for human authorship. All final outputs must be reviewed and approved by domain experts.

We encourage future research to explore participatory reviewer integration, adaptive learning mechanisms, and safeguards to ensure originality, diversity, and attribution fidelity.

Ethical Statement

We used large language models, including ChatGPT, solely to polish the manuscript's language, including grammar and phrasing. All substantive content, system design, and experimental decisions were authored and verified by the human research team.

ARISE is designed to assist researchers by automating structured writing tasks such as citation collection, summarization, and LaTeX formatting. It operates on verified academic inputs and uses low-temperature generation to minimize hallucinations. Most observed failure cases stem from external API disruptions (e.g., quota exhaustion) rather than model unpredictability. ARISE is not intended to replace human authorship or scholarly judgment. It produces draft materials for human review, and all final responsibility and intellectual ownership remain with the user.

6. Additional Human Evaluation with GPT-4.1-Mini

To assess whether our rubric-guided refinement remains effective with smaller models, we conducted an additional human evaluation using GPT-4.1-mini. As before, four expert reviewers (two professors, one postdoc, one PhD student) assessed five survey topics using the same detailed rubric.

As shown in Table A8, the system achieved an average total score improvement from 66.45 to 80.70 (**+21.45%**) and a subcategory average increase from 3.32 to 4.03 (**+21.45%**). These gains are even higher than the +19.20% improvement observed with GPT-4.1 (Table 5), highlighting that rubric-guided refinement remains robust and impactful even when starting from weaker initial drafts. The smaller model benefits more due to its lower baseline quality, while the full GPT-4.1 starts from a higher base and thus shows slightly smaller relative gains. This comparison underscores the generalizability of our framework across model sizes.

7. Ablation Study: Small-Model Pipeline (GPT-4.1 Mini)

To assess the trade-offs between cost and performance, we conducted an ablation study using the GPT-4.1 Mini model throughout the agentic pipeline. As shown in Table A9, the smaller model achieved slightly lower initial and final rubric scores compared to the default (full-size) ARISE pipeline, but remained competitive with other systems and published baselines. Notably, we observed that the primary source of performance loss stemmed from the small model's weaker instruction-following and formatting ability (e.g., outputting raw metadata, markdown, or incomplete LaTeX environments). Despite this, the 4.1 Mini pipeline still delivered

high-quality survey drafts at significantly reduced computational cost, supporting its use in resource-constrained settings or for rapid iteration.

References

Abdullahi, S.; Danyaro, K. U.; Zakari, A.; Aziz, I. A.; Zawawi, N. A. W. A.; and Adamu, S. 2025. Time-Series Large Language Models: A Systematic Review of State-of-the-Art. *IEEE Access*, 13: 30235–30261.

ACL Rolling Review. 2025. Reviewer Guidelines. https://aclrollingreview.org/reviewerguidelines. Accessed May 2025.

Anthropic. 2024. Claude 3 Model Family. https://www.anthropic.com/api.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2023. A Survey on Evaluation of Large Language Models. arXiv:2307.03109.

Chen, W.; Su, Y.; Zuo, J.; Yang, C.; Yuan, C.; Chan, C.-M.; Yu, H.; Lu, Y.; Hung, Y.-H.; Qian, C.; Qin, Y.; Cong, X.; Xie, R.; Liu, Z.; Sun, M.; and Zhou, J. 2023. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. arXiv:2308.10848.

Conde, J.; Reviriego, P.; Salvachúa, J.; Martínez, G.; Hernández, J. A.; and Lombardi, F. 2024. Understanding the Impact of Artificial Intelligence in Academic Writing: Metadata to the Rescue. *Computer*, 57(1): 85–88.

Contributors, L. 2024a. LangGraph: Composable Multi-Agent Workflows for LLM Applications. https://www.langchain.com/langgraph. Accessed: 2025-07-14.

Contributors, T. C. 2024b. CrewAI: Fast and Flexible Multi-Agent Automation Framework. https://github.com/joaomdmoura/crewai. Accessed: 2025-05-16.

Dafoe, A.; Hughes, E.; Bachrach, Y.; Collins, T.; McKee, K. R.; Leibo, J. Z.; Larson, K.; and Graepel, T. 2021. Open Problems in Cooperative AI. *arXiv preprint arXiv:2012.08630*.

Dong, W.; Zhao, Y.; Sun, Z.; Liu, Y.; Peng, Z.; Zheng, J.; Zhang, Z.; Zhang, Z.; Wu, J.; Wang, R.; Xu, S.; Huang, X.; and He, X. 2025. Humanizing LLMs: A Survey of Psychological Measurements with Tools, Datasets, and Human-Agent Applications. arXiv:2505.00049.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325.

Gao, Y.; Xiong, Y.; Gao, X.; Jia, K.; Pan, J.; Bi, Y.; Dai, Y.; Sun, J.; Wang, M.; and Wang, H. 2024. Retrieval-Augmented Generation for Large Language Models: A Survey. arXiv:2312.10997.

Ghafarollahi, A.; and Buehler, M. J. 2025. SciAgents: Automating Scientific Discovery Through Bioinspired Multi-Agent Intelligent Graph Reasoning. *Advanced Materials*, 37(22): 2413523.

Google, G. T. 2025. Gemini: A Family of Highly Capable Multimodal Models. arXiv:2312.11805.

Guo, Z.; Jin, R.; Liu, C.; Huang, Y.; Shi, D.; Supryadi; Yu, L.; Liu, Y.; Li, J.; Xiong, B.; and Xiong, D. 2023. Evaluating Large Language Models: A Comprehensive Survey. arXiv:2310.19736.

He, J.; Treude, C.; and Lo, D. 2025. LLM-Based Multi-Agent Systems for Software Engineering: Literature Review, Vision and the Road Ahead. arXiv:2404.04834.

IEEE. 2020. IEEE Reviewer Guidelines. https://ieeeauthorcenter.ieee.org/wp-content/uploads/ieee-reviewer-guidelines.pdf. Accessed: 2025-05-18.

Karapantelakis, A.; Alizadeh, P.; Alabassi, A.; et al. 2024. Generative AI in mobile networks: a survey. *Annals of Telecommunications*, 79: 15–33.

Laskar, M. T. R.; Alqahtani, S.; Bari, M. S.; Rahman, M.; Khan, M. A. M.; Khan, H.; Jahan, I.; Bhuiyan, A.; Tan, C. W.; Parvez, M. R.; Hoque, E.; Joty, S.; and Huang, J. 2024. A Systematic Survey and Critical Review on Evaluating Large Language Models: Challenges, Limitations, and Recommendations. arXiv:2407.04069.

Lee, S. K.; and Ko, H. 2025. Generative Machine Learning in Adaptive Control of Dynamic Manufacturing Processes: A Review. arXiv:2505.00210.

Lee, Y.; Kim, J.; Kim, J.; Cho, H.; Kang, J.; Kang, P.; and Kim, N. 2025. CheckEval: A reliable LLM-as-a-Judge framework for evaluating text generation using checklists. arXiv:2403.18771.

Liang, X.; Yang, J.; Wang, Y.; Tang, C.; Zheng, Z.; Song, S.; Lin, Z.; Yang, Y.; Niu, S.; Wang, H.; Tang, B.; Xiong, F.; Mao, K.; and Li, Z. 2025. SurveyX: Academic Survey Automation via Large Language Models. *arXiv preprint arXiv:2502.14776*.

Liu, F.; Yang, Z.; Liu, C.; Song, T.; Gao, X.; and Liu, H. 2025. MM-Agent: LLM as Agents for Real-world Mathematical Modeling Problem. arXiv:2505.14148.

Luo, J.; and Rao, H. 2025. A Survey on the Topology of Fractal Squares. arXiv:2505.00309.

Messer, M.; Brown, N. C. C.; Kölling, M.; and Shi, M. 2024. Automated Grading and Feedback Tools for Programming Education: A Systematic Review. *ACM Trans. Comput. Educ.*, 24(1).

Miah, A. S. M.; taro Suzuki; and Shin, J. 2025. A Methodological and Structural Review of Parkinsons Disease Detection Across Diverse Data Modalities. arXiv:2505.00525.

OpenAI. 2023. GPT-4 Technical Report. https://openai.com/research/gpt-4.

Qian, J.; Wang, X.; et al. 2024. Communicative Agents for Software Development. In *ICLR*.

Serper. 2024. Serper API Documentation. https://serper.dev/.

Shen, Y.; Song, K.; Tan, X.; Li, D.; Lu, W.; and Zhuang, Y. 2023. HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in Hugging Face. arXiv:2303.17580.

Wang, Y.; Guo, Q.; Yao, W.; Zhang, H.; Zhang, X.; Wu, Z.; Zhang, M.; Dai, X.; Zhang, M.; Wen, Q.; Ye, W.; Zhang, S.; and Zhang, Y. 2024. AutoSurvey: Large Language

Models Can Automatically Write Surveys. *arXiv preprint arXiv:2406.10252*.

Wu, Q.; Li, C.; Peng, B.; and et al. 2023. AutoGen: Enabling Next-Generation Large Language Model Applications via Multi-Agent Conversation Framework. https://microsoft.github.io/autogen/.

Wu, S.; Ma, X.; Luo, D.; Li, L.; Shi, X.; Chang, X.; Lin, X.; Luo, R.; Pei, C.; Du, C.; Zhao, Z.-J.; and Gong, J. 2025. Automated Review Generation Method Based on Large Language Models. *National Science Review*. ArXiv:2407.20906.

Yan, X.; Feng, S.; Yuan, J.; Xia, R.; Wang, B.; Zhang, B.; and Bai, L. 2025. SurveyForge: On the Outline Heuristics, Memory-Driven Generation, and Multi-dimensional Evaluation for Automated Survey Writing. *arXiv preprint arXiv:2503.04629 [cs.CL]*.

Ye, R.; Liu, X.; Wu, Q.; Pang, X.; Yin, Z.; Bai, L.; and Chen, S. 2025. X-MAS: Towards Building Multi-Agent Systems with Heterogeneous LLMs. *arXiv preprint arXiv:2505.16997*.

Yin, S.; Fu, C.; Zhao, S.; Li, K.; Sun, X.; Xu, T.; and Chen, E. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Zhang, C.; Deng, Y.; Lin, X.; Wang, B.; Ng, D.; Ye, H.; Li, X.; Xiao, Y.; Mo, Z.; Zhang, Q.; and Bing, L. 2025. 100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models. arXiv:2505.00551.

Model	Chunk	Summary	Major Strengths	Major Weaknesses	Recommendations	Final Score	Dec.
gpt-4.1	Pages 1–3	This section is exemplary for an academic survey, setting a strong foundation for both technical insight and methodological rigor. It weaves together the latest advances in agentic systems, AI-driven survey automation, and academic workflow innovation, making substantial contributions via new taxonomies, thorough discussion of standards, and explicit identification of unresolved challenges.	Explicit objectives and robust relevance to urgent AI and academia intersections; Integrated, up-to-date, and balanced literature coverage; Thorough analysis, new frameworks, and clearly presented challenges/gaps; Excellent organization and professional presentation.	Minor gaps in visual diversity—additional figures would be valuable; Clustered citations could be better detailed in the references section.	Add one or more schematic diagrams or flowcharts of workflow architectures; Maintain detailed, accessible reference lists for full citation clarity.	4.9	accept
gemini-2.5	Pages 1–3	This section of the survey provides a comprehensive, state-of-the-art review of AI-driven academic automation. Theoretical and workflow foundations are robust, and major research challenges are well-delineated via succinct tables. The taxonomy of GAEs and focus on explainability, standardization, and equity are notable advances.	Clear objectives and targeted audience; Thorough, recent, and balanced literature review; Insightful theoretical foundation and comparative analysis; Succinct, effective summarization via tables	Minor overlap with prior automation/LLM surveys; Occasional imbalance in strength vs weakness discussion; Visuals could extend beyond tables; citation formatting is numeric-only	Minor revision: emphasize weaknesses/limitations alongside strengths more uniformly; Introduce more diverse visuals (e.g., workflow diagrams); Include reference list or mapping in the supplementary material for citation traceability	4.7	minor revision
claude-3.7	Pages 1–3	This is an ambitious and well-written survey that provides a timely, comprehensive, and insightful overview of automated survey generation, literature review, and agentic academic AI systems. The writing is clear and professional, objectives are explicit, and the structure is logical and effective. Integration of tables reinforces important points and summarizes complex ideas. The taxonomy of GAEs is novel and enhances the field. Citations and literature engagement are mature but could be marginally more explicit or current in places.	Explicit objectives and well-defined scope; Addresses cutting-edge trends and pressing academic needs; Original taxonomy/framework and clear advancement for the field; Balanced, analyzed, and critical discussion; Effective summarization and organization	Some integration could be strengthened between technical and methodological examples; Citation formatting and clarity could be made more consistent; Minor need for additional visuals to complement tables	Strengthen interdisciplinary connections and integration across subfields; Standardize and clarify citation format throughout the section; Incorporate additional figures or diagrams for workflow illustration; Elaborate on specific research gaps with suggested future directions	4.6	minor revision

Table A3: Grouped review summary scores for each section (Pages 1–3 and 4–6) and agent.

Category	Criterion	1	2	3	4	5
Scope	Objectives	No objectives stated or inferred	Unclear or implicit; requires inference	Vague or generic; lacks focus	Clear in one section; lacks precision	Clearly stated in abstract and intro; scoped and measurable
	Relevance	Not relevant to the field	Weak or outdated connection	Partially related to broader topic	Generally relevant, not urgent	Directly aligns with high-impact trends
	Audience	No discernible audience	Confusing or poorly targeted	Somewhat unclear	Generally appropriate tone	Clear academic or interdisciplinary targeting
Literature	Comprehensiveness	Sparse or incomplete coverage	Major omissions	Some omissions or limited domain	Mostly complete with minor gaps	$\geq$ 30 citations, across subfields, up-to-date
	Balance	Highly biased or promotional	One-sided view	Somewhat unbalanced	Balanced with minor bias	Discusses strengths/weaknesses and perspectives
	Currency	Ignores recent developments	Mostly dated content	Some outdated dominance	Mostly recent with few older works	Up-to-date including preprints and conferences
Analysis	Depth	No meaningful analysis	Minimal or weak analysis	Descriptive only	Moderate depth	Theoretical rigor, layered insight
	Integration	Disjointed and fragmented	Mostly disconnected ideas	Partial, siloed integration	Good integration	Seamless integration of multiple perspectives
	Gaps	Ignores all research gaps	Barely addresses open questions	Surface-level mention	Mentions some gaps	Clearly identifies open challenges
Originality	Novelty	No original contribution	Mostly derivative	Slightly original	Novel combination of ideas	New taxonomy, framework, or domain
	Advancement	No advancement	Minimal progress	Incremental value	Moderate contribution	Strong guidance for future research
	Redundancy Avoidance	Highly repetitive	Largely redundant	Moderate overlap	Mostly unique	Clearly distinct from prior surveys
Organization	Logical Flow	Chaotic and disorganized	Poor transitions	Basic structure with issues	Mostly clear flow	Excellent transitions and structure
	Section Clarity	No clear structure	Unclear or unlabeled	Confusing or too long	Mostly clear	Well-labeled and crystal clear
	Summarization	No summary or synthesis	Almost none	Minimal synthesis	Some synthesis and structure	Effective use of summaries and visuals
Presentation	Language	Unreadable or ungrammatical	Poor grammar or clarity	Clumsy tone	Mostly well-written	Clear academic language throughout
	Visuals	No meaningful visuals	Irrelevant or low-quality	Basic, not integrated	Good visuals with minor issues	Strong figures/tables supporting content
	Formatting	Disorganized formatting	Distracting issues	Inconsistent formatting	Minor format problems	Clean, consistent styles
References	Accuracy	Unreliable or incorrect citations	Multiple citation errors	Some mismatched or incomplete	Minor format issues	Accurate, traceable, properly formatted
	Appropriateness	Poor citation quality	Many low-quality sources	Some irrelevant or filler	Mostly appropriate	Highly relevant, current and foundational

Table A4: Evaluation Rubric for Survey Paper Quality (Scores 1–5)

Research Area	Paper Title and Reference	Venue / Year
LLM reasoning and replication	*100 Days After DeepSeek-R1: A Survey on Replication Studies...* (Zhang et al. 2025)	arXiv 2025
LLM evaluation metrics	*A Survey on Evaluation of Large Language Models* (Chang et al. 2023)	ACM 2024
Retrieval-augmented generation	*Retrieval-Augmented Generation for LLMs: A Survey* (Gao et al. 2024)	arXiv 2024
Multimodal large language models	*A Survey on Multimodal Large Language Models* (Yin et al. 2024)	NSR 2024
Time-series modeling	*Time-Series Large Language Models: A Systematic Review of State-of-the-Ar* (Abdullahi et al. 2025)	IEEE 2025
Generative AI in manufacturing	*Generative Machine Learning in Adaptive Control of Dynamic Manufacturing Processes: A Review* (Lee and Ko 2025)	arXiv 2025
Topology of fractal squares	*A Survey on the Topology of Fractal Squares* (Luo and Rao 2025)	arXiv 2025
Human-agent interaction	*Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications* (Dong et al. 2025)	arXiv 2025
Disease detection across modalities	*A Methodological and Structural Review of Parkinson's Disease Detection Across Diverse Data Modalitie* (Miah, taro Suzuki, and Shin 2025)	IEEE 2025
Generative AI in mobile networks	*Generative AI in Mobile Networks: A Survey* (Karapantelakis et al. 2024)	AoT 2024

Table A5: Selected Baseline Survey Papers with Research Areas and Publication Venues

Type	System / Model Pair	Krippendorff's Alpha (α)
System-Level Agreement		
	SurveyX	0.974
	SurveyForge	0.977
	Baseline (Published)	0.966
	Autosurvey	0.987
	ARISE	0.973
Model-Level Agreement (All Systems)		
	Claude 3.7 vs Gemini 2.5	0.974
	Claude 3.7 vs GPT-4.1	0.977
	Gemini 2.5 vs GPT-4.1	0.973

Table A6: Krippendorff's Alpha scores for system-level and inter-model agreement. Computed using interval-scale rubric ratings across reviewer agents.

Model	Type	Input	Output	Context Caching
GPT-4.1	Premium	$2.00	$8.00	$0.50
GPT-4.1 mini	Balanced	$0.40	$1.60	$0.10
Claude 3.7 Sonnet	Premium	$3.00	$15.00	$0.30 (read), $3.75 (write)
Gemini 2.5 Pro	Premium	$1.25 / $2.50	$10.00 / $15.00	$0.31 / $0.625
Gemini 2.0 Flash	Affordable	$0.10 (text/image), $0.70 (audio)	$0.40	$0.025

Table A7: LLM Pricing Comparison (USD per 1M tokens)

Topic	Before (Total Score)	After (Total Score)	Before (Subcat. Avg)	After (Subcat. Avg)
Disease Detection Across Modalities	69.00	82.25	3.45	4.11
Automated Survey Generation	70.00	86.25	3.50	4.31
Retrieval-Augmented Generation	65.25	79.25	3.26	3.96
Time-Series Modeling	68.00	81.00	3.40	4.05
Topology of Fractal Squares	60.00	74.75	3.00	3.74
Average	**66.40**	**80.70**	**3.32**	**4.03**
Improvement (%)		**21.45%**		**21.45%**

Table A8: Human evaluation results under GPT-4.1-mini. Despite reduced model capacity, the agentic refinement pipeline yields substantial improvements across five diverse topics, as verified by four expert reviewers.

Folder / Topic	Initial Score	Final Score	Avg Improvement per Round
Evaluation of Large Language Models	85.90	91.17	1.76
Retrieval-Augmented Generation	79.90	90.90	11.00
Time-Series Modeling	86.93	90.70	3.77
Clustering, Indexing and Range Searching	88.67	90.53	1.87
Disease Detection Across Modalities	80.10	89.67	4.78
Integrating Multimodal Fusion, PLMs	81.33	87.10	2.88
Telecommunication	82.50	86.60	1.03
Generative AI in Mobile Networks	82.50	86.60	1.03
Generative AI in Manufacturing	82.00	86.07	1.36
Automated Survey Generation Lit. Review	84.43	84.87	0.22
Topology of Fractal Squares	79.73	84.27	2.27
Average	**83.09**	**88.04**	**2.91**

Table A9: Ablation study: Initial and final rubric scores and average improvement per round using the GPT-4.1 Mini model. All scores are averaged across reviewer agents and rubric categories.