AURORA: An Agentic Writing System Leveraging Rubric-Guided Reinforcement Learning

Anonymous ACL submission

Abstract

We present AURORA, a modular agentic system framework for automated academic survey generation and iterative refinement. At its core is Agentic Reinforcement Learning (ARL), where multiple reviewer agents evaluate drafts using a shared rubric, producing structured feedback that guides a fixed-policy refinement agent across successive revisions. The system comprises five coordinated components: citation preparation, knowledge base construction, outline generation, paper composition, and self-evaluation—each designed for modularity, reproducibility, and interoperability. To evaluate AURORA's effectiveness as a survey generation system, we compare its outputs against two baselines: (1) ten recent (2023–2025) human-written survey papers across diverse domains from arXiv and peer-reviewed venues, and (2) outputs from state-of-the-art automatic survey generation approaches. Experimental results show that AURORA outperforms both, achieving an average rubric-aligned score of 92.48. This score is derived from a 100-point evaluation rubric grounded in professional peer-review standards, covering seven dimensions and twenty subcategories such as clarity, originality, relevance, and literature coverage. These findings validate the effectiveness of AURORA's agentic refinement loop and rubric-as-reward framework in generating high-quality, transparent, and academically rigorous survey papers.

1 Introduction

The exponential growth of scientific literature has made it increasingly difficult for researchers to synthesize developments and produce high-quality survey papers. Traditional approaches to literature review remain labor-intensive, error-prone, and inconsistent—often lacking transparency, standardization, and scalability (Conde et al., 2024).

Recent advances in large language models (LLMs) and agentic system offer new opportunities to rethink scholarly writing. By decomposing writing workflows into specialized agents—responsible for retrieval, drafting, validation, and critique—LLM-driven systems enable scalable, structured academic composition.

We introduce **AURORA**, an agentic system for generating and refining citation-grounded academic surveys (Figure 1). Beyond automating content creation, AURORA improves output quality through structured feedback and revision. AURORA's pipeline spans five modular phases—from citation preparation and knowledge base construction to outline generation, LaTeX formatting, and rubric-based evaluation—each runnable independently or as part of a full workflow. Reviewer agents (GPT-4.1, Gemini 2.5 Pro, Claude 3.7) assess drafts using a shared rubric, and their feedback guides a refinement agent that iteratively improves the output until a quality threshold is reached.

Our key contributions are threefold: (1) a rubric-guided multi-agent review loop that provides consistent, interpretable, and goal-aligned feedback; (2) a stable and general-purpose Agentic Reinforcement Learning (ARL) framework that leverages a Rubric-as-Reward (RaR) signal for iterative refinement without relying on stochastic learning; and (3) a calibrated evaluation rubric grounded in peer-review standards, supporting modular integration of alternate reviewers or strategies. Together, these components enable transparent, high-quality, and extensible survey generation.

Building on these contributions, the remainder of this paper is organized as follows. Section 2 situates AURORA within the landscape of automated survey generation and reinforcement learning framework for text refinement. Section 3 provides a detailed breakdown of AURORA's architecture and agentic workflow, including its ARL and RaR mechanisms. We then describe our experimental setup and evaluation protocol in Section 4, followed by results and analysis in Section 5.Sec-

tion 6 presents our conclusions. This is followed by sections discussing the limitations, broader impact, ethical statement, and acknowledgements. The paper concludes with the references and appendix.

2 Related Work

2.1 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 lacks source curation and relies solely on preprints. *SurveyX* (Liang et al., 2025) enhances retrieval via hybrid keyword expansion and semantic filtering with an *AttributeTree* citation structure, though it lacks modularity and user control. *SurveyForge* (Yan et al., 2025) uses heuristic templates and a memory-driven citation agent (SANA), and introduces the SurveyBench benchmark for holistic evaluation, but remains constrained by a static architecture and preprint-heavy sourcing.

2.2 Reinforcement Learning for Structured Text Refinement

Reinforcement learning (RL) has been applied to improve language models using preference-based rewards. In summarization, Stiennon et al. (2020) train a reward model from human feedback to fine-tune policies, outperforming reference-based baselines. Other work applies RL recursively to refine long-form summaries (Wu et al., 2021).

Ramamurthy et al. (2023) generalize this paradigm with `RL4LMs`, a modular framework for applying RL to text generation, along with GRUE, a benchmark guided by automated reward functions. While promising, these approaches often rely on opaque scalar rewards and monolithic architectures.

2.3 Agentic Reinforcement for Text Improvement with Rubric-Based Feedback

We introduce *ARL*, a modular framework for iterative text refinement guided by rubric-based feedback. In each round, reviewer agents independently evaluate a draft using a shared rubric, producing structured, dimension-wise scores. These are aggregated into a reward signal that guides a fixed-policy refinement agent.

ARL emphasizes interpretability, modularity, and reproducibility by embedding explicit rubric evaluation into the generation loop. This enables scalable, self-improving workflows for academic writing, summarization, and structured content generation.

3 Methodology

3.1 System Overview

Figure 1 presents the high-level architecture of AURORA, our agentic system framework for automated academic survey generation and iterative self-improvement. The system is composed of five core components: (1) *Citation Preparation*, (2) *Structured Knowledge Base Construction*, (3) *Structured Outline Generation*, (4) *Survey Paper Composition and Finalization*, and (5) *Agentic Reinforcement Learning for Self-Evaluation and Refinement*.

The figure lists all agents, their tasks, and goals, orchestrated via CrewAI (Grossmann and the CrewAI Contributors, 2024). This modular design supports plug-and-play extensibility, enabling each agent to operate independently while contributing to a coordinated, self-correcting generation loop. Through this architecture, AURORA enables scalable, transparent, and rubric-grounded automation of academic writing workflows.

3.2 Agentic Reinforcement Learning Framework for AURORA

In this study *ARL* integrates structured rubric scoring into a reinforcement-style feedback loop. Unlike black-box optimization or static fine-tuning, ARL uses interpretable rubric-based feedback from multiple reviewer agents to guide iterative refinement by a dedicated revision agent.

We model this process using a reinforcement learning-style tuple (S, A, Q), where S denotes the current draft state D_t, A is a fixed revision policy π that generates a revised draft D_{t+1} based on reviewer feedback, and $Q(S, A)$ represents the vector of rubric scores assigned by the reviewer agents after applying action A to state S.

ARL does not rely on a scalar reward function or policy learning. The GPT-4.1-based refinement agent operates under a static policy that deterministically incorporates reviewer feedback into its revision strategy.

Visual Overview. Figure 2 presents a detailed view of the ARL architecture. The process begins

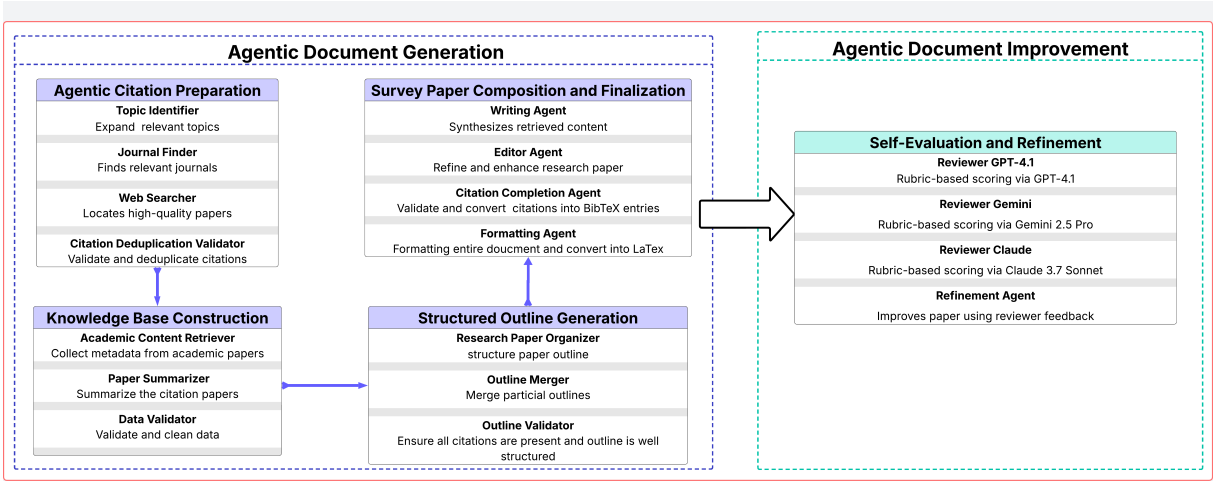

Figure 1: Agent Roles and Functionality in AURORA. Each agent is assigned a specific function in the modular survey generation and refinement pipeline. All agents—excluding the reviewer agents (Gemini and Claude)—are implemented using GPT-4.1.

with a draft document D_0 evaluated by three reviewer agents. Each agent scores the draft according to a shared rubric $\mathcal{B}$ composed of well-defined criteria (e.g., `Relevance`, `Comprehensiveness`, `Formatting`), each anchored with calibrated definitions along a 1–5 scale. An example of such a rubric is illustrated at the top of the figure, showing score anchors for the `Relevance` dimension. A score of "5" in `Relevance` indicates alignment with "current, high-impact trends," while "1" flags a topic as "not relevant to the intended field."

Feedback Aggregation and Decision. The ARL cycle is defined in Algorithm 1. In each iteration t, the current draft D_t is evaluated by multiple reviewer agents $\mathcal{A}_m$, each applying the shared rubric $\mathcal{B}$. The result of each evaluation is a rubric-based feedback vector $S^{(m)} = [s_t^{(m,1)}, s_t^{(m,2)}, \ldots, s_t^{(m,N)}]$, where $s_t^{(m,i)}$ denotes the score assigned by agent $\mathcal{A}_m$ to the i-th rubric subcategory. These vectors collectively represent fine-grained assessments of D_t and are used to compute the total reward:

$$Q_t = \sum_{m \in M} \sum_{i=1}^{N} s_t^{(m,i)} \tag{1}$$

where $M = \{\text{GPT-4.1}, \text{Gemini 2.5}, \text{Claude 3.7}\}$ is the set of reviewer agents, and N is the number of rubric sub-criteria. If the aggregated reward $Q_t \geq \tau$, the draft is accepted as final. Otherwise, the refinement agent $\mathcal{R}$ produces a revised version of the draft, using all feedback vectors $\{S^{(m)}\}_{m \in M}$. This process continues iteratively until the quality threshold is met or a stopping condition is reached.

Algorithm 1 Agentic Reinforcement Learning (ARL)

Require: Initial draft D_0, rubric $\mathcal{B}$, reviewer agents $\mathcal{A}_m$, refinement agent $\mathcal{R}$, threshold τ
Ensure: Final output D^*

$t \leftarrow 0$
repeat
  for all $m \in M$ **do**
    $S^{(m)} \leftarrow \mathcal{A}_m(D_t, \mathcal{B})$
  end for
  $Q_t \leftarrow \sum_{m \in M} \sum_{i=1}^{N} s_t^{(m,i)}$
  if $Q_t \geq \tau$ **then**
    return $D^* \leftarrow D_t$ ▷ Accept final draft
  else
    $D_{t+1} \leftarrow \mathcal{R}(D_t, \{S^{(m)}\}_{m \in M})$
    $t \leftarrow t + 1$
  end if
until max iterations reached

Rubric-as-Reward (RaR). ARL is powered by the principle of Rubric-as-Reward (RaR). Rather than relying on opaque or task-specific heuristics, RaR transforms standardized rubric dimensions into a numerical reward signal. Each dimension includes detailed anchors that promote scoring consistency across agents. This structure allows both LLM reviewers and human evaluators to score drafts with high agreement, enhancing transparency and reproducibility. RaR is organized into seven dimensions and twenty subcategories, each scored out of 5 points for a total of 100, as shown in Table 1. The rubric draws structural inspiration from professional peer-review guidelines, includ-

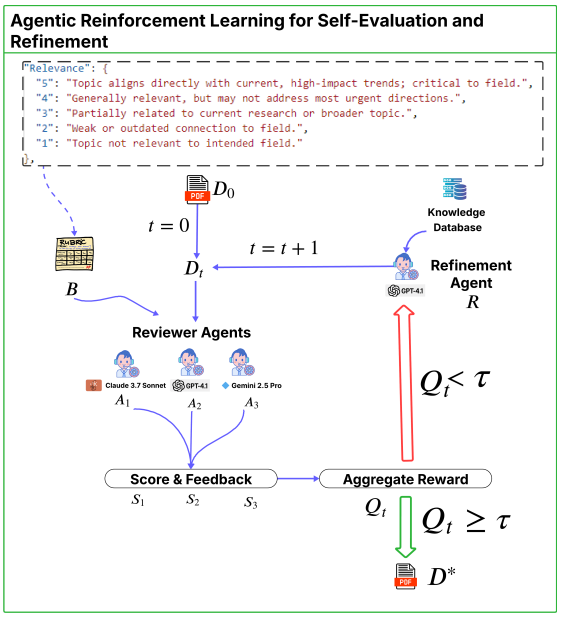

Figure 2: Agentic Reinforcement Learning (ARL) framework. Reviewer agents independently score draft D_t using a shared rubric $\mathcal{B}$ (sample shown for `Relevance`). Their scores are aggregated into a scalar reward Q_t. If $Q_t \geq \tau$, the draft is accepted as D^*. Otherwise, a refinement agent generates an improved version, and the loop continues.

Table 1: Rubric-as-Reward (RaR) Structure: 7 dimensions, 20 subcategories, each scored out of 5 (total 100 points).

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

ing those of IEEE (IEEE, 2020) and ACL Rolling Review (ACL Rolling Review, 2025), which emphasize evaluation along dimensions such as originality, literature, clarity, and relevance. The complete rubric definitions are provided in Appendix B.

Reinforcement Loop Dynamics. Conceptually, ARL reframes the editing cycle as a reinforcement learning problem: the draft D_t is the environment state, reviewer feedback defines the reward, and the refinement agent acts as the policy learner. Each round of scoring and refinement moves the document toward a reward-maximizing state—one that fully satisfies the rubric. The loop terminates when the draft meets the quality threshold or plateaus in improvement.See Appendix C for a detailed example of reviewer feedback synthesized through the ARL loop.

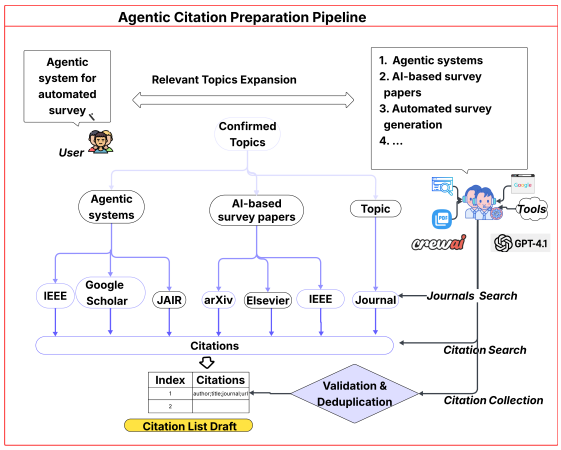

Figure 3: Agentic citation preparation pipeline. Users shape topics and guide journal selection; agents handle retrieval, filtering, and validation.

3.3 Agentic Document Generation

3.3.1 Citation Preparation

As illustrated in Figure 3, the citation preparation module collects high-quality references through agent-led retrieval and human-in-the-loop oversight. The pipeline ensures relevance, academic credibility, and alignment with user intent.

Interactive Topic Expansion Starting from user-defined themes, an expansion agent proposes semantically related subtopics for user approval. This iterative refinement balances thematic breadth with goal alignment.

Journal-Aware Retrieval A journal agent suggests k reputable venues per subtopic, spanning peer-reviewed journals, open-access archives, and vetted preprints. This journal-first filter enhances source quality and publication relevance.

Citation Collection and Deduplication For each topic–journal pair, a retrieval agent collects metadata-rich citations. Post-processing includes deduplication and formatting validation, yielding a curated reference set for downstream synthesis.

3.3.2 Structured Knowledge Base Construction

To support generation, the system builds a structured knowledge base from the retrieved papers. As shown in Figure 4, documents are fetched via direct URLs or fallback metadata searches on Google Scholar or arXiv. If partial content (e.g., abstract) is

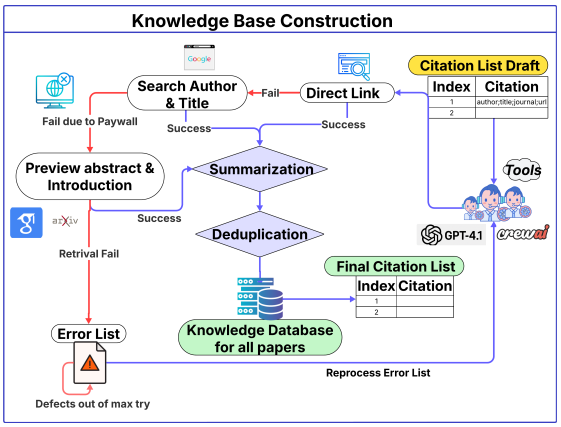

Figure 4: Structured Knowledge Base Construction. For each citation, the system attempts direct or fallback retrieval, extracts summaries, deduplicates content, and stores the result in a structured database aligned with the citation index.

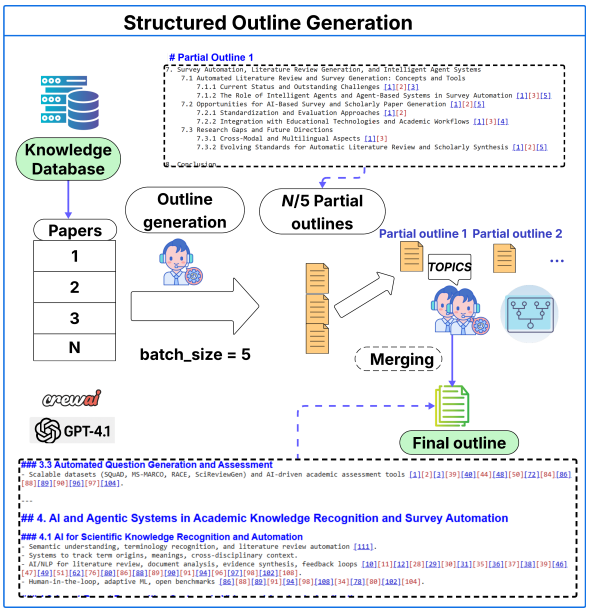

Figure 5: Structured outline generation component. Citation summaries are grouped, outlined, and iteratively merged to form a thematically coherent, citation-preserving global structure.

found, it is extracted; otherwise, the entry is logged in an *Error List*. Retrieved texts are summarized by a GPT-4.1 agent into concise, contribution-focused entries. Outputs are deduplicated and indexed into a persistent *Knowledge Base* aligned with the citation index. This knowledge base forms the factual substrate for later modules, enabling grounded and context-aware survey generation.

3.3.3 Structured Outline Generation

To support coherent and citation-grounded survey writing, the system uses a team of agents to build a structured outline from citation summaries. This process ensures that the final output is modular, traceable, and thematically organized.

As shown in Figure 5, the process starts with a **Knowledge Base** containing N cleaned citation summaries. These are divided into mini-batches, and each batch is processed by a **Writing Agent**, which generates a partial outline with sections, subsections, and 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 a larger outline. After each merge, a **Validation Agent** checks that all citation references are preserved, the content flows logically, and there are no redundancies or gaps.

This merging and validation cycle continues until a single, citation-complete **Final Outline** is created. A final validation step ensures that all original references are included by comparing the result with the citation index.

The final outline provides a clear, structured foundation for the next phase—drafting survey text that is well-organized and tightly connected to the original sources.

3.3.4 Survey Paper Composition and Finalization

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

As shown in Figure 6, the process begins by decomposing the outline into section-level prompts based on the document hierarchy established during outline generation. For each prompt, the system retrieves the relevant citation indices and their associated summaries from the structured knowledge base.

A **Writing Agent**, powered by GPT-4.1, synthesizes the retrieved content into coherent, thematically organized prose. Each section incorporates cited works appropriately, maintains traceability to original sources, and conforms to academic writing conventions. The resulting drafts are passed to an **Editor Agent**, which enhances logical flow, improves clarity, corrects formatting inconsistencies, and inserts placeholders for tables where necessary. These refined sections are then merged into a unified draft that preserves the original outline

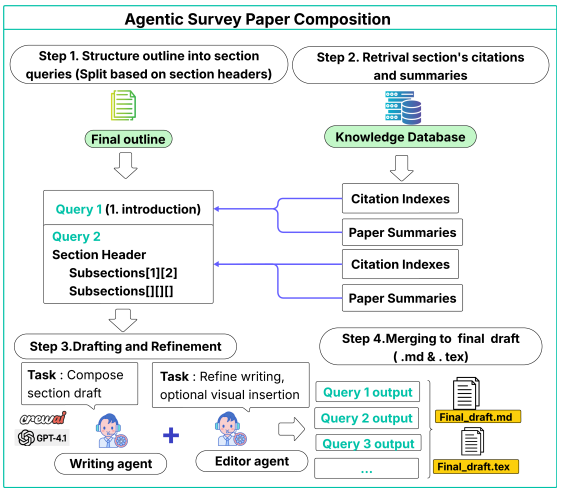

Figure 6: Agentic Survey Paper Composition. Each section query is matched with relevant citations and summaries. A writing agent drafts the content, which is refined by an editor agent before merging into intermediate content outputs.

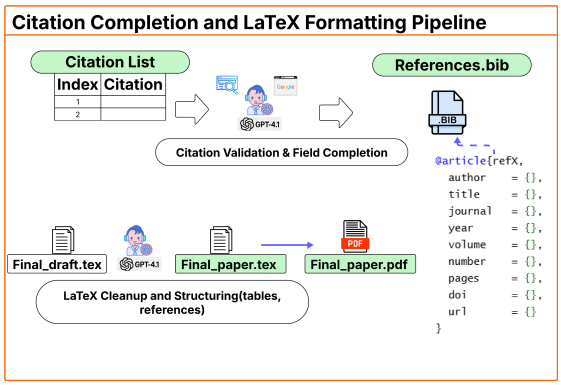

Figure 7: Citation and Formatting Pipeline. The system completes citation metadata, generates BibTeX entries, and standardizes the LaTeX document to produce a clean, structured PDF ready for academic use.

structure and citation alignment.

Next, a **Citation Completion Agent** resolves missing bibliographic metadata such as author names, publication venues, DOIs, and URLs by querying trusted databases and repositories. The completed entries are converted into BibTeX format and compiled into a structured bibliography.

A **Formatting Agent** performs a comprehensive pass over the entire LaTeX document, standardizing citation commands, harmonizing section and table styles, and cleaning up residual artifacts from earlier processing stages. It applies structured environments such as `adjustbox` and `booktabs` to enhance the presentation of tabular content and ensure stylistic consistency. By default, survey papers are compiled using the ACM `sigconf` class with a standard LaTeX toolchain based on `TeX Live`. The formatting setup is modular and supports alternate styles such as `acl_pub`, `ieeeconf`, or `arxiv`, depending on venue requirements.

4 Experiments

4.1 Evaluation Setup

We adopt a multi-agent, rubric-aligned evaluation protocol designed for precision, reproducibility, and cross-model consistency. The rubric, shown in Table 1, defines twenty subcategories across seven core dimensions. Our primary goal is to assess the quality of system-generated survey papers relative to human-written baselines. A secondary goal is to demonstrate that structured rubric grounding yields stable, interpretable evaluation results across large language models.

4.2 Chunked Review for Depth and Coverage

To ensure full-document coverage and minimize positional bias, we segment each paper into contiguous 3-page chunks. This design encourages balanced attention across sections and ensures 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.

4.3 Baseline Papers

We benchmark AURORA against ten recently published, human-authored survey papers selected for topical diversity, recency (2023–2025), and visibility across arXiv and peer-reviewed venues. Table 2 lists the selected papers and their research areas. These serve as real-world baselines to evaluate structure, citation practices, and thematic coherence.

4.4 Controlled Research Area Matching and Evaluation Fairness

To enable direct comparison, we use AURORA to generate survey papers covering the same ten research areas as the human-written baselines. Each was produced using the full AURORA pipeline—from citation collection to rubric-based refinement—ensuring alignment in research area and structural intent.

We also evaluate AURORA against prior automated systems, including *SurveyForge* (Yan et al., 2025), *SurveyX* (Liang et al., 2025), and *AutoSurvey* (Wang et al., 2024). Due to availability

Table 2: Selected Baseline Survey Papers with Research Areas

Research Area	Paper Title and Reference
LLM reasoning and replication	*100 Days After DeepSeek-R1: A Survey on Replication Studies and More Directions for Reasoning Language Models* (Zhang et al., 2025)
LLM evaluation metrics	*A Survey on Evaluation of Large Language Models* (Chang et al., 2023)
Retrieval-augmented generation	*Retrieval-Augmented Generation for Large Language Models: A Survey* (Gao et al., 2024)
Multimodal large language models	*A Survey on Multimodal Large Language Models* (Yin et al., 2024)
Time-series modeling	*Time-Series Large Language Models: A Systematic Review of State-of-the-Art* (Abdullahi et al., 2025)
Generative AI in manufacturing	*Generative Machine Learning in Adaptive Control of Dynamic Manufacturing Processes: A Review* (Lee and Ko, 2025)
Topology of fractal spaces	*A Survey on the Topology of Fractal Squares* (Luo and Rao, 2025)
Human-agent interaction	*Humanizing LLMs: A Survey of Psychological Measurements with Tools, Datasets, and Human-Agent Applications* (Dong et al., 2025)
Disease detection across modalities	*A Methodological and Structural Review of Parkinson's Disease Detection Across Diverse Data Modalities* (Miah et al., 2025)
Generative AI in mobile networks	*Generative AI in Mobile Networks: A Survey* (Karapantelakis et al., 2024)

constraints, we include 10 papers each from SurveyForge and SurveyX, and 3 from AutoSurvey. Unlike AURORA, which is explicitly configured to align with predefined research areas, the available outputs from these systems are only partially aligned and include topics outside the targeted set of ten.

Nonetheless, all output and baseline papers are evaluated using the same domain-agnostic rubric, chunking strategy, and tri-model reviewer setup. This ensures fair and consistent scoring across structurally diverse content, enabling valid comparative analysis.

Although AURORA-generated surveys are intentionally aligned with the baselines' research areas, strict topic matching is not required for valid evaluation. Our domain-agnostic rubric emphasizes structure, citation integrity, clarity, and originality—enabling consistent comparisons even when third-party systems differ in subject matter.

5 Results and Analysis

5.1 Overall Performance.

Our system, **AURORA**, achieves the highest overall quality among all five evaluated systems. As shown in Table 3, AURORA outperforms every baseline across all reviewers—Claude 3.7, Gemini 2.5 Pro, and GPT-4.1—with an average total score of **92.48**, substantially higher than SurveyForge (87.68), Baseline (86.15), Autosurvey (82.46), and SurveyX (81.65). The consistent superiority across all reviewers confirms AURORA's robustness and general reliability.

Table 3: Mean TOTAL scores by system and reviewer

System	GPT-4.1	Gemini 2.5 Pro	Claude 3.7 Sonnet	Mean of Agents
AURORA	**92.57**	**92.59**	**92.27**	**92.48**
Autosurvey	81.87	81.74	83.76	82.46
Baseline	86.58	85.81	86.06	86.15
SurveyForge	87.88	87.51	87.64	87.68
SurveyX	81.13	81.99	81.82	81.65

5.2 Rubric-Level Superiority

AURORA demonstrates consistently strong rubric-level performance, leading in **all seven categories** evaluated (Table 4). In critical dimensions such as *Literature* (4.95), *Presentation* (4.84), *References* (4.98), and *Organization* (4.82), AURORA surpasses both human-written baselines and recent automated systems. These gains are attributed to our agentic refinement strategy and explicit modular writing design.

Table 4: Mean reviewer scores by rubric category and system

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

As visualized in Figure 8, AURORA's radar profile forms a wide and balanced polygon, dominating each axis. Competing systems exhibit narrow or irregular profiles, indicating gaps in either structure, originality, or presentation fluency. SurveyX and Autosurvey, in particular, show significant underperformance in analytical depth and clarity. 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 A for full agreement scores.

These results validate our core hypothesis: **rubric-guided agentic reinforcement enables transparent, interpretable, and high-quality survey generation**. By integrating modular agent roles, structured evaluation criteria, and multi-round refinement, AURORA not only outperforms traditional baselines but also establishes a reproducible foundation for self-improving academic writing systems.

5.3 Validating the ARL Process with a Domain-Specific Refinement Example

To illustrate the impact of AURORA's Agentic Reinforcement Learning (ARL) loop, we present a representative refinement trajectory for a system-

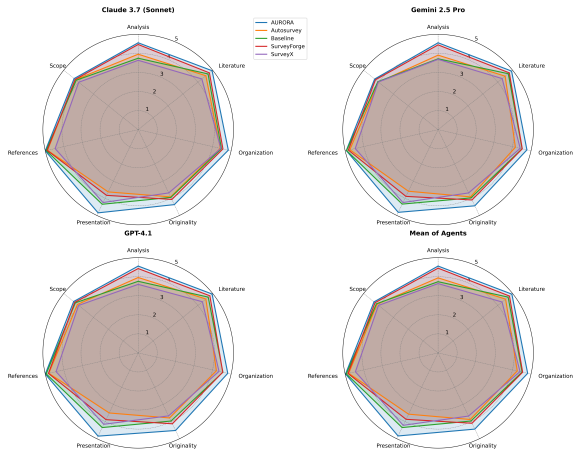

Figure 8: Radar charts comparing rubric scores across systems. Each subplot corresponds to one reviewer (Claude 3.7 Sonnet, Gemini 2.5 Pro, GPT-4.1) or the averaged mean.

Table 5: Score progression over ARL refinement rounds for the LLM Reasoning and Replication survey. The target combined score is 276 out of 300 (92%).

Round	GPT-4.1	Gemini 2.5 Pro	Claude 3.7 Sonnet	Combined Total
0	86.8	87.9	86.2	260.9
1	90.5	89.8	90.0	270.3
2	93.8	89.2	91.3	274.3
3	92.3	92.8	92.9	277.9

generated survey paper in the domain of *LLM reasoning and replication*, aligned with the first baseline listed in Table 2. This example tracks how quality evolves over successive ARL iterations based on rubric-guided feedback from three independent reviewer agents.

As shown in Table 5, the survey begins with a combined total score of 260.9, reflecting strong but improvable quality. Following each review cycle, the refinement agent incorporates structured feedback across 20 rubric subcategories and regenerates the draft. By Round 3, the score exceeds the 92% threshold, reaching 277.9 out of 300. Notably, this improvement is achieved without model updates or retraining—underscoring the power of structured, interpretable feedback loops in agentic systems. This trajectory demonstrates AURORA's ability to iteratively elevate content quality through modular, reviewer-guided revision.

5.4 Reference Reliability

To validate the reliability of AURORA's citation preparation process (Section 3.3.1), we conducted a traceability audit of the final system-generated references. We define the *Expanded Citation Traceability Rate* (eCTR) as:

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

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 AURORA 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.

6 Conclusion

We presented AURORA, a modular agentic system for automated academic survey generation and iterative refinement. At its core is the Agentic Reinforcement Learning (ARL) framework, where multiple reviewer agents independently score drafts using a shared rubric. Their feedback is aggregated via a Rubric-as-Reward (RaR) mechanism to guide a fixed-policy refinement agent in a structured, interpretable loop—without requiring gradient-based updates.

The full pipeline spans five coordinated phases: citation preparation, knowledge base construction, outline generation, paper composition, and rubric-driven self-evaluation. Empirical results show that AURORA outperforms human-written baselines and recent automated systems, achieving a mean score of 92.48 across GPT-4.1, Gemini 2.5 Pro, and Claude 3.7 Sonnet reviewers. In detailed rubric-level analysis, AURORA led all systems across seven core dimensions—including *Literature*, *Presentation*, and *References*. Its radar profile demonstrated balanced strength across analytical depth, originality, and structure.

These findings validate ARL and RaR as effective strategies for transparent, high-quality text refinement. Future work will extend the framework to new document genres, multimodal generation, and interactive human-in-the-loop revision. System outputs, agent prompts and reviewer evaluations are available at: https://anonymous.4open.science/r/AURORA-EE33/README.md.

Limitations

While AURORA 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 depends on commercial LLM APIs that are subject to request limitations, rate throttling (e.g., RPM/QPM), context window caps, and quota exhaustion. These factors occasionally interrupt long document processing, delay pipeline execution, or require retry logic. Moreover, the overall generation and evaluation process is computationally expensive—each paper costs approximately $35–$40 and requires 3.5 hours to complete.

Third, in the Agentic Reinforcement Learning (ARL) framework, the revision policy remains fixed. While the refinement agent applies rubric-grounded edits, it is not updated dynamically through learning. Consequently, AURORA's ability to improve over time depends solely on the performance and reasoning consistency of the underlying LLMs.

Despite these limitations, AURORA 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

AURORA aims to improve the scalability, structure, and factual consistency of academic survey writing by automating citation preparation, outline construction, and LaTeX formatting through modular agent workflows. Its intended audience includes researchers, educators, and academic writers seeking assistance in synthesizing large volumes of literature.

The broader impact of this work is twofold. On the positive side, AURORA lowers the barrier to entry for producing well-organized, citation-grounded scholarly outputs. This could be especially beneficial in under-resourced research communities or interdisciplinary fields where manual literature review is prohibitively time-consuming. Additionally, our emphasis on traceable references, rubric-based evaluation, and modular transparency supports responsible deployment and downstream auditing.

However, risks remain. Over-reliance on automated survey generation may discourage critical thinking or reinforce biases encoded in training data. If deployed naively, AURORA could contribute to the proliferation of derivative content or fail to surface underrepresented research perspectives. To mitigate these risks, AURORA is designed to assist—not replace—human authorship, and all final outputs must be reviewed and approved by domain experts.

We encourage future work to explore participatory integration of human reviewers, adaptive learning mechanisms, and safeguards for originality, diversity, and attribution fidelity.

Ethical Statement

AURORA is designed to assist researchers by automating structured writing tasks such as citation collection, summarization, and LaTeX formatting. Because it operates on verified academic inputs and uses low-temperature summarization, the risk of hallucination is minimal. Most failure cases arise from external API issues (e.g., quota exhaustion), not model instability. AURORA does not replace human authorship or creativity; instead, it generates ACL- or ACM-compatible drafts to streamline academic workflows. Final responsibility and intellectual authorship remain fully with the user.

Acknowledgements

We used generative AI tools in a limited capacity for non-substantive assistance. Specifically, large language models were employed to help with language polishing and paraphrasing of originally written content, comparable to grammar or style checking tools. Additionally, AI-assisted search tools were used during the literature review phase to identify relevant prior work. All content, including citations, was thoroughly reviewed and verified by the authors. The authors remain solely responsible for the integrity and correctness of the paper's content.

References

Shamsu Abdullahi, Kamaluddeen Usman Danyaro, Abubakar Zakari, Izzatdin Abdul Aziz, Noor Amila Wan Abdullah Zawawi, and Shamsuddeen Adamu. 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.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arXiv:2307.03109.

J. Conde, P. Reviriego, J. Salvachúa, G. Martínez, J. A. Hernández, and F. Lombardi. 2024. Understanding the impact of artificial intelligence in academic writing: Metadata to the rescue. *Computer*, 57(1):85–88.

Wenhan Dong, Yuemeng Zhao, Zhen Sun, Yule Liu, Zifan Peng, Jingyi Zheng, Zongmin Zhang, Ziyi Zhang, Jun Wu, Ruiming Wang, Shengmin Xu, Xinyi Huang, and Xinlei He. 2025. Humanizing llms: A survey of psychological measurements with tools, datasets, and human-agent applications. *Preprint*, arXiv:2505.00049.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. 2024. Retrieval-augmented generation for large language models: A survey. *Preprint*, arXiv:2312.10997.

Rafael Grossmann and the CrewAI Contributors. 2024. Crewai: A framework for agentic workflows with llms. https://github.com/joaomdmoura/crewai. Accessed: 2025-05-16.

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

Angelos Karapantelakis, Payam Alizadeh, Ammar Alabassi, and 1 others. 2024. Generative ai in mobile networks: a survey. *Annals of Telecommunications*, 79:15–33.

Suk Ki Lee and Hyunwoong Ko. 2025. Generative machine learning in adaptive control of dynamic manufacturing processes: A review. *Preprint*, arXiv:2505.00210.

Xun Liang, Jiawei Yang, Yezhaohui Wang, Chen Tang, Zifan Zheng, Shichao Song, Zehao Lin, Yebin Yang, Simin Niu, Hanyu Wang, Bo Tang, Feiyu Xiong, Keming Mao, and Zhiyu Li. 2025. Surveyx: Academic survey automation via large language models. *arXiv preprint arXiv:2502.14776*.

Jun Luo and Hui Rao. 2025. A survey on the topology of fractal squares. *Preprint*, arXiv:2505.00309.

Abu Saleh Musa Miah, taro Suzuki, and Jungpil Shin. 2025. A methodological and structural review of parkinsons disease detection across diverse data modalities. *Preprint*, arXiv:2505.00525.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization. *Preprint*, arXiv:2210.01241.

Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. 2020. Learning to summarize with human feedback. In *Advances in Neural Information Processing Systems*, volume 33, pages 3008–3021. Curran Associates, Inc.

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. 2024. Autosurvey: Large language models can automatically write surveys. *arXiv preprint arXiv:2406.10252*.

Jeff Wu, Long Ouyang, Daniel M. Ziegler, Nisan Stiennon, Ryan Lowe, Jan Leike, and Paul Christiano. 2021. Recursively summarizing books with human feedback. *Preprint*, arXiv:2109.10862.

Xiangchao Yan, Shiyang Feng, Jiakang Yuan, Renqiu Xia, Bin Wang, Bo Zhang, and Lei Bai. 2025. Surveyforge: On the outline heuristics, memory-driven generation, and multi-dimensional evaluation for automated survey writing. *arXiv preprint arXiv:2503.04629 [cs.CL]*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12).

Chong Zhang, Yue Deng, Xiang Lin, Bin Wang, Dianwen Ng, Hai Ye, Xingxuan Li, Yao Xiao, Zhanfeng Mo, Qi Zhang, and Lidong Bing. 2025. 100 days after deepseek-r1: A survey on replication studies and more directions for reasoning language models. *Preprint*, arXiv:2505.00551.

Appendix

A Reliability Evaluation

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

Type	System / Model Pair	Krippendorff's Alpha (α)
System-Level Agreement		
	SurveyX	0.974
	SurveyForge	0.977
	Baseline (Published)	0.966
	Autosurvey	0.987
	AURORA	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 6: Krippendorff's Alpha scores for system-level and inter-model agreement. Computed using interval-scale rubric ratings across reviewer agents.

B Evaluation Rubric

Table 7 presents the full evaluation rubric used to assess survey paper quality. Each sub-criterion is rated on a 1-5 scale, where 5 represents the strongest performance.

Table 7: Evaluation Rubric for Survey Paper Quality (Scores 1–5)

Category	Criterion	Score 1	Score 2	Score 3	Score 4	Score 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
	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

C Sample Agentic Reviewer Feedback (ARL Loop)

Table 8: Reviewer Suggestions on Survey Sections (ARL Feedback Sample)

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 subsections.; 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.