OUTLINEFORGE: Hierarchical Reinforcement Learning with Explicit States for Scientific Writing

Abstract

Scientific paper generation requires document-level planning and factual grounding, but current large language models, despite their strong local fluency, often fail in global structure, input coverage, and citation consistency. We present a reinforcement learning framework that casts scientific outline construction as a long-horizon planning problem over hierarchical document structures. Our approach models edit evolving outlines through structured actions, enabling the system to incrementally build a complete scientific manuscript. To support effective and stabilize learning,we introduce a two-stage optimization procedure consisting of (i) backward outline reconstruction from partial plans to enforce global structural consistency, and (ii) forward value-guided reinforcement learning with rewards explicitly modeling scientific correctness, discourse coherence, and citation fidelity. In addition, We further introduce a benchmark for scientific paper generation that evaluates document planning, input utilization, reference faithfulness, outline organization, and content-level factual accuracy. Our results show consistent improvements over strong neural and LLM baselines, particularly in long-range structural coherence and citation reliability.

1 Introduction

The scientific paper generation task is the intersection of long-form text generation,document-level planning, and verifiable scientific reasoning. Compared with short form tasks, it requires models not only to write fluently, but also to maintain coherent global structure, cover relevant inputs, and use citations in a faithful and consistent manner.

Over the past two years, large language models (LLMs) have evolved from passive text generators into autonomous, agentic systems capable of structured reasoning, planning, and tool use. Driven by advances in architectures and post-training, LLMs are now viewed as dynamic cognitive engines rather than static predictors, enabling them to handle complex scientific and engineering workflows.

Parallel to these developments, artificial intelligence has become deeply embedded across the scientific research pipeline. Surveys such as AI4Research (Chen et al., 2025) and The Evolving Role of Large Language Models in Scientific Innovation (Zhang et al., 2025a) document how AI systems now operate as literature reviewers, experiment planners, hypothesis evaluators, and even autonomous research collaborators. Within this emerging paradigm, LLMs are no longer just tools that accelerate human workflows; they increasingly take on structured roles in knowledge discovery, evaluation, and scientific creativity.

A complementary line of work focuses on LLM-based agentic reasoning frameworks, which formalize how language models can decompose tasks, call external tools, iteratively refine outputs, and use feedback signals to improve their reasoning trajectories. Recent surveys, including LLM-based Agentic Reasoning Frameworks: A Survey from Methods to Scenarios (Zhao et al., 2025), outline the architectural patterns behind contemporary agent systems such as ReAct (Yao et al., 2022), Tree-of-Thoughts (Yao et al., 2023), Chain-of-Thought prompting (Wei et al., 2022), and open-ended agents like Voyager (Wang et al., 2023). At the same time, the emergence of scientific LLMs, summarized in A Survey of Scientific Large Language Models (Hu et al., 2025), demonstrates a clear trend toward specialized models trained on high-quality scientific corpora, domain-specific reasoning tasks, and structured multimodal inputs. Together, these developments indicate a convergence: scientific

LLMs are becoming more agentic, and agentic LLM frameworks are increasingly being adapted to scientific research.

Despite these advances, systematic mechanisms for improving agentic scientific reasoning remain underdeveloped, particularly in tasks involving long-horizon decision making such as scientific writing, experimental design, and multi-step argument construction. Existing approaches typically rely on supervised fine-tuning or heuristic feedback rather than grounded optimization principles. Furthermore, progress in this field is hindered by the lack of standardized benchmarks that evaluate the quality, structure, and correctness of scientific text generated by LLMs.

To address these gaps, this work makes three contributions. First, we propose a reinforcement learning and value-based training paradigm for scientific paper generation, where the model learns to construct coherent, verifiable, multi-section research manuscripts under explicit value functions. Unlike prior methods that optimize for surface-level textual quality, our framework introduces a structured credit-assignment mechanism tailored to scientific writing. Second, we build a new benchmark for evaluating scientific paper generation, designed around realistic multi-section writing tasks, citation consistency, argumentative structure, and factual validity. This benchmark enables reproducible evaluation and comparative analysis for future work on scientific LLMs and agentic scientific reasoning. Third, our experiments demonstrate that the proposed value-guided generation framework significantly improves global narrative coherence, scientific correctness, and cross-section consistency over strong supervised baselines.

2 Related Works

2.1 AI for Scientific Research and Automated Writing Systems

Artificial intelligence has become increasingly involved across the scientific research pipeline. Surveys such as AI4Research (Chen et al., 2025) and the evolving role of LLMs in scientific innovation (Zhang et al., 2025a) document how modern systems now assist with literature review, experiment planning, hypothesis evaluation, and autonomous scientific inquiry. In parallel, research on agentic reasoning explores how LLMs can plan, decompose tasks, use tools, and refine solutions through iterative feedback. The survey by Zhao et al. (Zhao et al., 2025) organizes this emerging paradigm, and foundational frameworks such as ReAct (Yao et al., 2022), Tree-of-Thoughts (Yao et al., 2023), Chain-of-Thought prompting (Wei et al., 2022), and Voyager (Wang et al., 2023) illustrate how structured reasoning and action execution can be combined. Multi-agent systems for scientific writing, including SurveyForge (Yan et al., 2025) and SciSage (Shi et al., 2025), further demonstrate that outline planning, memory-driven generation, and coordinated agent roles improve complex writing tasks. Our work extends this line of research by introducing principled value-based optimization for long-horizon scientific writing.

A growing body of work focuses on scientific large language models trained on domain-specific corpora or tailored for scientific reasoning. Hu et al. (Hu et al., 2025) provide a recent overview, covering data foundations and agentic extensions. Representative models such as Galactica (Taylor et al., 2022), SciGPT (She et al., 2025), and SciBERT (Beltagy et al., 2019) aim to support scientific retrieval, knowledge grounding, and citation reasoning. Complementing these models, several systems target automatic survey or scientific writing. AutoSurvey (Wang et al., 2024), SurveyX (Liang et al., 2025), and SGSimEval (Guo et al., 2025) explore various strategies for survey drafting and multi-dimensional evaluation. More advanced approaches such as SurveyForge (Yan et al., 2025) and SciSage (Shi et al., 2025) employ outline-based heuristics, memory mechanisms, or multi-agent coordination to improve structure and content quality. Recent iterative frameworks for literature analysis and survey automation (Yiran et al., 2025; Zhang et al., 2025b) highlight the importance of structured intermediate representations and multi-aspect reasoning. Although these methods show strong local improvements, they typically lack a unified optimization principle for controlling global manuscript quality. Our work addresses this limitation by introducing reinforcement learning with explicit value functions to guide multi-section scientific paper generation.

2.2 Citation Accuracy, Evaluation Benchmarks, and Scientific Text Structure

Accurate citations are essential in scientific writing and have become a central focus of recent LLM

research. ScholarCoPilot (Wang et al., 2025) enhances academic writing through citation-aware triggers and retrieval supervision, achieving substantial improvements in retrieval accuracy and human-judged citation quality. CiteGuard (Choi et al., 2025) complements generation-side methods with retrieval-augmented validation to assess whether model-generated citations faithfully match the claims they support. Beyond citation, a number of benchmarks evaluate scientific text quality from multiple dimensions. SGSimEval (Guo et al., 2025) proposes similarity-enhanced metrics for measuring outline quality, content fidelity, and citation behavior. Additional work on scientific text structuring (Bolaños et al., 2025; Zhu et al., 2025) provides annotated schemas and multi-aspect taxonomies for analyzing scientific discourse. These efforts highlight that scientific writing quality is multi-dimensional, involving structure, factual grounding, and citation faithfulness. Our proposed benchmark draws on these insights, integrating structural, argumentative, and citation-aware criteria while our value-based training framework incorporates them directly into optimization.

2.3 Reinforcement Learning and Value Modeling for Long-Form Language Generation

Reinforcement learning has become an influential paradigm for controlling high-level behaviors in large language models, particularly in tasks requiring long-horizon coherence, preference alignment, or structured decision making. Early work on reinforcement learning from human feedback (RLHF) demonstrated that reward-based optimization can significantly improve factuality, helpfulness, and reasoning capabilities in LLMs. Notably, the InstructGPT framework (Ouyang et al., 2022) formalized the modern RLHF pipeline, showing that preference-based optimization can robustly guide model behavior through human demonstrations and comparisons. The underlying optimization is often implemented using Proximal Policy Optimization (PPO) (Schulman et al., 2017), which has become the standard policy-gradient backbone for RLHF systems. More recent extensions, such as reinforcement learning from AI feedback (RLAIF) (Lee et al., 2023) and direct preference optimization (DPO) (Rafailov et al., 2023), further expand the feasibility of applying value modeling, preference learning, and policy optimization to long-form generation.

3 Method

To more systematically characterize the intermediate reasoning structures underlying scientific paper generation, we reformulate the task from conventional token-level sequence generation into a multi-step evolution process over a structured intermediate state space.

3.1 Formulation of Scientific Writing

We formulate scientific paper generation as a structured, long-horizon conditional generation task. Given a set of heterogeneous inputs, the model is required to produce a coherent, well structured scientific article that not only achieves high conditional likelihood, but also satisfies a collection of structural, semantic, and factual constraints implicit in academic writing.

3.1.1 Input Space

Let the input be

$$X = (T, C, R), \tag{1}$$

where

- $T \in \mathcal{T}$ denotes the topic or high level writing intent.
- $C = \{c_1, c_2, \ldots, c_m\} \in \mathcal{C}$ represents background materials, documents, or other contextual evidence.
- $R = \{r_1, r_2, \ldots, r_k\} \in \mathcal{R}$ is the set of external references available for citation.

This formulation is agnostic to how C and R are obtained; in practice, they can come from retrieval modules, pre-defined reading lists, or memory-like components.

3.1.2 Output Space

A scientific article is represented as a sequence of textual units:

$$Y = (y_1, y_2, \ldots, y_L), \tag{2}$$

where each $y_i \in \mathcal{U}$ may be a token, a phrase, or an element from a domain-specific markup language (e.g., Markdown or LaTeX). The full output space is defined as

$$\mathcal{Y} = \left\{ Y = (y_1, \ldots, y_L) \;\middle|\; \begin{array}{l} L \in \mathbb{N}, \\ y_i \in \mathcal{U}, \\ \mathrm{ValidArticle}(Y) \end{array} \right\}. \tag{3}$$

where $\mathrm{ValidArticle}(Y)$ denotes that Y conforms to the structural and semantic constraints of a scientific article.

During generation, the model parameterized by θ defines a conditional distribution over $\mathcal{U}$, typically with an autoregressive factorization:

$$\pi_\theta(Y \mid X) = \prod_{i=1}^{L} \pi_\theta(y_i \mid y_{<i}, X), \quad \pi_\theta(y_i \mid y_{<i}, X) \in \Delta(\mathcal{U}). \tag{4}$$

3.1.3 Basic Conditional Generation Objective

The most general form of the paper-generation task is the maximization of the conditional likelihood:

$$Y^* = \arg\max_{Y \in \mathcal{Y}} \pi_\theta(Y \mid X), \tag{5}$$

where $\pi_\theta(Y \mid X)$ denotes the joint likelihood of the entire article under the model. This view recovers standard maximum likelihood training and decoding when no additional constraints are imposed.

3.1.4 Information Coverage and Structural Constraints

A scientific article must not only be fluent and coherent but also faithfully utilize information provided in the input materials. Let $I(\cdot)$ denote a semantic information mapping (for example, extracted facts, key claims, or important concepts). The coverage of contextual information is defined as

$$\mathrm{Coverage}(Y, C) = \frac{|I(Y) \cap I(C)|}{|I(C)|}. \tag{6}$$

Most prior work implicitly or explicitly seeks to maximize this quantity, ensuring that the generated paper incorporates essential content from the input, rather than hallucinating unsupported statements or ignoring core evidence.

The overall objective can be written as a composite optimization problem:

$$Y^* = \arg\max_{Y \in \mathcal{Y}} \Big[\pi_\theta(Y \mid X) + \sum_I \lambda_I \cdot \mathrm{Coverage}(Y, C)\Big], \tag{7}$$

where λ_I are tunable coefficients balancing likelihood and coverage-based criteria. This regularized objective captures the intuition that a good scientific article should both be likely under the model and align well with the provided evidence.

3.2 Reinforcement Learning

Reinforcement learning (RL) provides an alternative perspective to long-form scientific writing by modeling the generation process as a sequential decision-making problem. Formally, we consider an episodic Markov decision process (MDP)

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, r, \gamma), \tag{8}$$

where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, P is the transition kernel, r is the reward function, and $\gamma \in (0, 1]$ is the discount factor. At each step t, the model is in state $s_t \in \mathcal{S}$ and emits an action $a_t \in \mathcal{A}$ that corresponds to producing the next textual unit y_t, followed by a transition to s_{t+1} and receipt of a scalar reward r_t.

In this view, the writing policy $\pi_\theta(a_t \mid s_t)$ defines a distribution over textual actions conditioned on the current partial document. The objective is to maximize the expected return

$$J(\theta) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t=1}^{L} \gamma^{t-1} r_t\Big], \tag{9}$$

where rewards can encode high-level writing desiderata, such as semantic coverage or citation correctness. This formulation decouples training from token-level likelihood and allows integration of non-differentiable constraints through reward shaping.

Existing RL approaches for long-form generation often rely on preference modeling, reward models, or heuristic scoring functions. However, directly applying RL to scientific writing is challenging due to extremely long horizons, sparse rewards, and complex structural constraints. These limitations motivate the introduction of hierarchical abstractions and intermediate reasoning states, enabling the model to plan at multiple timescales and reduce the difficulty of credit assignment.

3.3 Overview and Schema

Specifically, we represent a paper outline as an editable hierarchical structure, and model the generation process as a sequence of discrete state transitions over outline states $s_{\text{outline}} \in \mathcal{S}$. Each transition is realized through an explicit editing action, denoted as $a_{\text{diff}} \in \mathcal{A}$. Borrowing terminology from code version control systems, we refer to these editing operations as *diffs*. Under this formulation, paper generation is no longer

treated as a single-pass sampling of a token sequence, but rather as a reasoning trajectory that progressively refines an intermediate structure: starting from a coarse or empty outline, a sequence of local edits gradually converges to a high-quality outline, which in turn induces the final document generation. The intermediate state space $\mathcal{S}$ can therefore be interpreted as the set of all valid outline structures.

To capture human editing preferences, we sample trajectories of outline-editing behaviors from real scientific papers, yielding preference data of the form $(s_{\text{outline}}, a_{\text{diff}}) \sim \mathcal{T}(\mathcal{D}_{\text{Paper}})$. Here, $\mathcal{D}_{\text{Paper}}$ denotes a corpus of raw scientific papers crawled from the arXiv API, and $\mathcal{T}$ is a transcription module that converts document revision histories into state–diff pairs over outline structures. We train a policy $\pi_\theta(a_{\text{diff},t} \mid s_{\text{outline},t})$ to align with these human preferences; for brevity, we subsequently write $\pi_\theta(a_t \mid s_t)$.

Depending on the intent of an editing diff, some actions (e.g., node reordering, deletion, or structural relocation) can be executed deterministically, while others (e.g., semantic refinement or content expansion) inherently require generative capabilities. For the latter class, the system constructs a query conditioned on the editing intent and delegates execution to a language model. Formally, this process is described as

$$\begin{cases} q_t = f(a_t), \\ s_{t+1} \sim \text{LLM}(\cdot \mid s_t, q_t), \end{cases} \tag{10}$$

where f maps an action to a model query.

Under this formulation, the outline generation process follows $s_{t+1} \sim P(\cdot \mid s_t, a_t)$ and can be formalized as a Markov Decision Process (MDP) $\mathcal{M} = \langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$. A complete generation trajectory is defined as $\tau = (s_0, a_0, s_1, a_1, \ldots, s_T)$, where the initial state s_0 corresponds to an empty outline. Subsequent states are generated through the alternating interaction between the policy π_θ and the transition operator P:

$$a_t \sim \pi_\theta(\cdot \mid s_t), \quad s_{t+1} \sim P(\cdot \mid s_t, a_t). \tag{11}$$

Human preference data can thus be interpreted as trajectories induced by an unknown optimal policy π^*, i.e., $\tau \sim p_{\pi^*}(\tau)$. We learn a parameterized policy π_θ via maximum likelihood estimation or preference alignment objectives to approximate this distribution, enabling the model to reproduce human editing decisions within the structured outline-editing space. Importantly, the learned policy not only produces high-quality outline structures, but also explicitly exposes intermediate reasoning steps and structural evolution trajectories, yielding an interpretable and controllable representation.

3.4 Methodology for Benchmarking

Our benchmarking data are constructed by systematically parsing arXiv papers. We sample 1,500 articles across six high-level domains and eleven subcategories. While experiments are conducted on a fixed subset for reproducibility, the pipeline is fully scalable and can be extended to arbitrarily large corpora for continuous data generation.

Inspired by abstract syntax tree (AST) processing in code tasks, we extract hierarchical document structures from arXiv HTML, including sections, subsections, and paragraphs. Based on this representation, an augmentation model proposes modification operations over the content.

Specifically, we sample individual paragraphs and apply stochastic perturbations (e.g., deletion or modification). The corrupted paragraph forms a problem instance, and the target is a coherent modification plan to restore or improve it. This reframes document evolution as a reasoning task. Each instance yields pre- and post-edit states with corresponding prompts, mirroring the inverse of our generation pipeline and enabling the model to learn how local edits lead to global improvements.

Figure 1 summarizes key statistical properties of the constructed dataset. Specifically, it reports the distribution of individual paragraph lengths and the citation density per paragraph across major research categories. A notable observation is that the citation distribution is highly skewed rather than uniform. This characteristic helps explain the limitations of prior approaches, many of which rely on retrieving a fixed number of references at each step via retrieval-augmented generation (RAG).

Figure 2 analyzes the expected number of steps required to fully decompose a complete article into an empty outline, which serves as a proxy for estimating the planning horizon in document evolution. The x-axis represents the proportion of update oper-

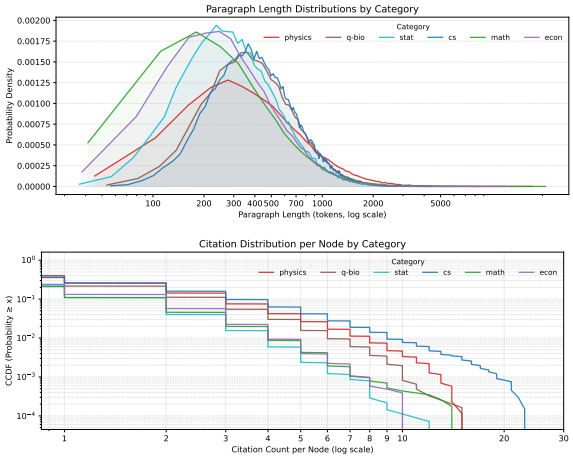


Figure 1: Paragraph length and citation density distributions by category.

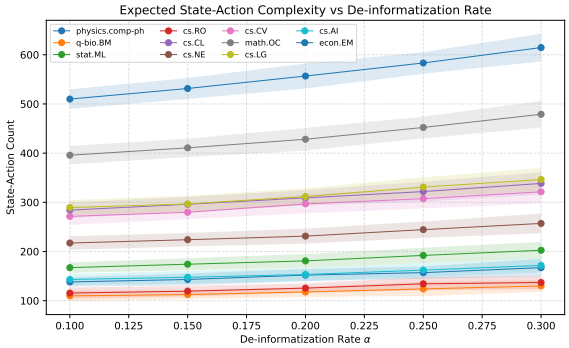


Figure 2: State-action complexity under varying de-informatiation rates.

ations, which preserve the document structure but are nevertheless crucial for producing high-quality articles. Empirically, we find that for the majority of research domains, a complete article can be generated within 200 to 300 steps. Based on this observation, we adopt these two step budgets as the primary experimental settings in subsequent evaluations.

4 Experiments

We evaluate our approach on a survey generation task, which is closely aligned with prior systems such as SurveyForge, SciSage, and AutoSurvey. Unless otherwise specified, all experiments are conducted under two inference budgets, 200 steps and 300 steps, which are selected based on the empirical analysis presented in Section 3.

We fine-tune three compact models, namely Gemma2-2B, Qwen3-1.8B, and Phi-3.8B, using the training data generated by our proposed pipeline. In addition, we evaluate several external large language models in a zero- or few-shot manner, including GPT-4o-mini, Claude-3.5-Haiku, and Llama-3.1-Instruct-70B. This setup allows us to examine both the effect of model scale and the benefit of task-specific fine-tuning under our data construction paradigm.

The quantitative results in Table 1 show that prior knowledge is critical for survey generation. Fine-tuning yields limited gains for models below 2B parameters due to insufficient intrinsic knowledge, whereas larger models consistently improve and even outperform strong general-purpose baselines. Additionally, while SurveyForge and AutoSurvey rely on a single-pass paradigm rather than a step-based formulation, including them as reference baselines highlights the clear advantage of modeling scientific writing as an iterative, sequential decision process over one-shot generation.

To further analyze the dynamics of the evolution process, we conduct an additional experiment illustrated in Fig. 3. The first three plots report the averaged results over multiple survey generation runs, showing that reference retrieval quickly saturates within the first 50 to 150 steps. Beyond this point, the number and relevance of retrieved citations exhibit marginal improvement, indicating that the document structure has largely stabilized.

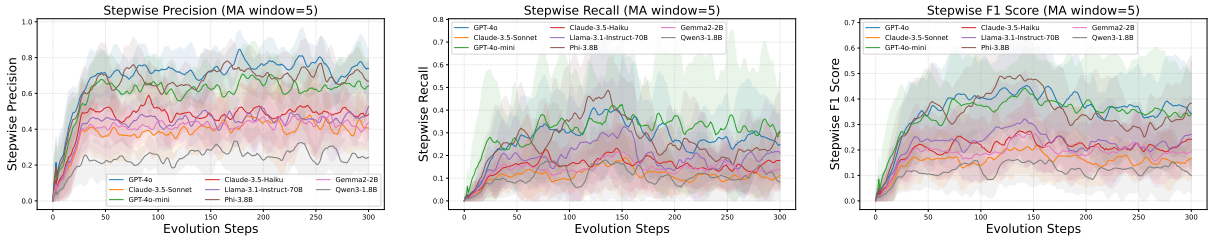
The remaining three plots in Fig. 4 present evaluations of structural completeness, citation relevance, and information density, as assessed by GPT-4o. Despite the plateau in citation-related metrics, these higher-level quality measures continue to improve in later stages of generation, suggesting that subsequent steps primarily focus on refinement rather than structural expansion. Due to the high cost of automated evaluation, these metrics are computed at a coarse granularity of once every 50 steps.

4.1 LLM as a judge

We recruited a panel of seven human evaluators with diverse academic backgrounds, including two undergraduate students, three master's students, and two PhD candidates specializing in Computer Science and related fields. The evaluation was

Model	200 steps			300 steps		
	Precision	Recall	F1	Precision	Recall	F1
Ours (GPT-4o-mini)	0.612±0.089	0.264±0.105	0.352±0.083	0.668±0.103	0.312±0.151	0.397±0.164
SurveyForge	0.346±0.028	0.261±0.113	0.285±0.083	—	—	—
AutoSurvey	0.295±0.028	0.181±0.080	0.213±0.062	—	—	—
Ours (Claude-3.5-Haiku)	0.533±0.094	0.173±0.071	0.249±0.082	0.524±0.069	0.186±0.129	0.246±0.119
SurveyForge	0.267±0.027	0.265±0.162	0.243±0.082	—	—	—
AutoSurvey	0.228±0.027	0.172±0.073	0.187±0.050	—	—	—
Ours (Phi-3.8B, Finetuned)	0.678±0.081	0.317±0.116	0.422±0.117	0.717±0.126	0.311±0.225	0.381±0.206
SurveyForge	0.359±0.028	0.345±0.274	0.313±0.129	—	—	—
AutoSurvey	0.296±0.029	0.306±0.198	0.274±0.101	—	—	—
Ours (Llama-3.1-Instruct-70B)	0.446±0.064	0.161±0.078	0.223±0.076	0.512±0.130	0.196±0.224	0.210±0.153
Ours (Gemma2-2B, Finetuned)	0.399±0.108	0.166±0.134	0.201±0.112	0.406±0.109	0.251±0.282	0.247±0.175
Ours (Qwen3-1.8B, Finetuned)	0.244±0.084	0.095±0.071	0.120±0.070	0.298±0.126	0.074±0.052	0.111±0.070

Table 1: erformance under different prompt ratios and inference steps. All metrics are reported as percentages.



(a) Precision (b) Recall (c) F1 Score

Figure 3: Stepwise performance metrics

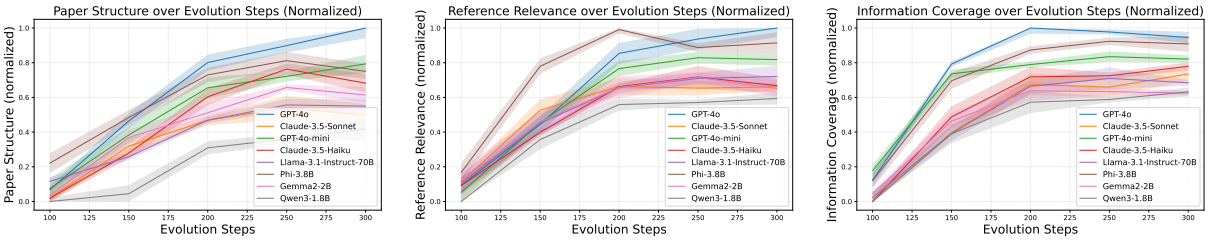
conducted under a **single-blind** protocol, where evaluators were presented with anonymized survey outlines without knowledge of the generating system. Each evaluator was tasked with scoring the outputs on a Likert scale from 1 to 5 based on four critical dimensions: **Coherence**, which measures the logical flow and transitions between sections; **Structural Completeness**, which assesses whether the outline covers the necessary breadth of the research topic; **Citation Relevance**, evaluating the appropriateness of references within their specific contexts; and **Logic & Soundness**, which examines the technical correctness of the hierarchical organization.

In parallel, we employ two state-of-the-art models, `gemini-3.1-pro` and `claude-3-5-sonnet`, to act as automated judges. Following established protocols for evaluating complex generative tasks, we provide these models with a detailed system prompt encompassing the same scoring rubrics as the human evaluators. The LLM judges were required to provide both a numerical score and a brief qualitative justification for each dimension to ensure the reliability of the automated assessment and to minimize potential positional or length biases.

The evaluation results, summarized in Table 2, demonstrate that our iterative evolution framework consistently outperforms *SurveyForge* and *AutoSurvey* across all metrics. We observe a high degree of Spearman correlation between human expert scores and the LLM-based evaluations, particularly with `claude-3-5-sonnet`, which exhibits a more nuanced understanding of structural logic.

5 Limitations

Despite its promising results, our work has several limitations. First, evaluating long-form scientific writing is inherently challenging; while common, relying on LLM-based automatic judges may introduce biases aligned with the models' priors, potentially diverging from human expert judgment. Second, our validation focuses primarily on survey generation. While representative, this does not fully cover the spectrum of academic writing, such as original research articles or interdisciplinary works, leaving the approach's generality to be further tested.



(a) Paper Structure (b) Reference Relevance (c) Information Coverage

Figure 4: Normalized evaluation scores

Judge	**Metric**	**Ours**	SurveyForge	AutoSurvey
Human (Avg.)	Coherence	**4.2**	3.5	3.1
	Structural Completeness	**4.4**	3.8	3.4
	Citation Relevance	**4.3**	3.4	2.9
	Logic & Soundness	**4.1**	3.2	2.8
Gemini-3.1-Pro	Coherence	**4.0**	3.1	2.9
	Structural Completeness	**4.3**	3.5	3.3
	Citation Relevance	**4.2**	3.6	3.1
	Logic & Soundness	**4.4**	3.9	3.5
Claude-3.5-Sonnet	Coherence	**4.3**	3.7	3.2
	Structural Completeness	**4.4**	4.0	3.4
	Citation Relevance	**4.5**	3.6	3.1
	Logic & Soundness	**4.2**	3.3	3.0

Table 2: Comparison of Human and LLM-based evaluation scores (1-5) across four key dimensions.

6 Conclusion

In this work, we propose a state-action-based framework for modeling long-horizon scientific writing as an iterative planning and reasoning process. By inverting document edits into structured supervision signals, we transform document evolution into a tractable reasoning problem that can be used to generate scalable training data. Our analysis of arXiv-derived statistics further reveals structural properties of scientific writing, such as highly skewed citation distributions, that are often overlooked by existing methods.

Experimental results on survey generation demonstrate that our approach consistently outperforms strong baselines and enables smaller, fine-tuned models to surpass larger general-purpose models under comparable settings. Beyond survey generation, our framework naturally supports flexible editing objectives and non-survey scientific writing, highlighting its potential as a general paradigm for controllable long-form text generation. We hope this work will inspire future research on self-evolving systems and structured reasoning for complex generative tasks.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Francisco Bolaños, Angelo Salatino, Francesco Osborne, and Enrico Motta. 2025. Modelling and classifying the components of a literature review. *arXiv preprint arXiv:2508.04337*.

Qiguang Chen, Mingda Yang, Libo Qin, Jinhao Liu, Zheng Yan, Jiannan Guan, Dengyun Peng, Yiyan Ji, Hanjing Li, Mengkang Hu, Yimeng Zhang, Yihao Liang, Yuhang Zhou, Jiaqi Wang, Zhi Chen, and Wanxiang Che. 2025. Ai4research: A survey of artificial intelligence for scientific research. *Preprint*, arXiv:2507.01903.

Yee Man Choi, Xuehang Guo, Qingyun Wang, and 1 others. 2025. Citeguard: Faithful citation attribution for llms via retrieval-augmented validation. *arXiv preprint arXiv:2510.17853*.

Beichen Guo, Zhiyuan Wen, Yu Yang, Peng Gao, Ruosong Yang, and Jiaxing Shen. 2025. Sgsimeval: A

comprehensive multifaceted and similarity-enhanced benchmark for automatic survey generation systems. In *International Conference on Advanced Data Mining and Applications*, pages 393–407. Springer.

Ming Hu, Chenglong Ma, Wei Li, Wanghan Xu, Jiamin Wu, Jucheng Hu, Tianbin Li, Guohang Zhuang, Jiaqi Liu, Yingzhou Lu, and 1 others. 2025. A survey of scientific large language models: From data foundations to agent frontiers. *arXiv preprint arXiv:2508.21148*.

Harrison Lee, Samrat Phatale, Hassan Mansoor, Kellie Ren Lu, Thomas Mesnard, Johan Ferret, Colton Bishop, Ethan Hall, Victor Carbune, and Abhinav Rastogi. 2023. Rlaif: Scaling reinforcement learning from human feedback with ai feedback.

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

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. *Advances in neural information processing systems*, 36:53728–53741.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*.

Fengyu She, Nan Wang, Hongfei Wu, Ziyi Wan, Jingmian Wang, and Chang Wang. 2025. Scigpt: A large language model for scientific literature understanding and knowledge discovery. *arXiv preprint arXiv:2509.08032*.

Xiaofeng Shi, Qian Kou, Yuduo Li, Ning Tang, Jinxin Xie, Longbin Yu, Songjing Wang, and Hua Zhou. 2025. Scisage: A multi-agent framework for high-quality scientific survey generation. *arXiv preprint arXiv:2506.12689*.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*.

Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. 2023. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*.

Yidong Wang, Qi Guo, Wenjin Yao, Hongbo Zhang, Xin Zhang, Zhen Wu, Meishan Zhang, Xinyu Dai, Qingsong Wen, Wei Ye, and 1 others. 2024. Autosurvey: Large language models can automatically write surveys. *Advances in neural information processing systems*, 37:115119–115145.

Yubo Wang, Xueguang Ma, Ping Nie, Huaye Zeng, Zhiheng Lyu, Yuxuan Zhang, Benjamin Schneider, Yi Lu, Xiang Yue, and Wenhu Chen. 2025. Scholarcopilot: Training large language models for academic writing with accurate citations. *arXiv preprint arXiv:2504.00824*.

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

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

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2023. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2022. React: Synergizing reasoning and acting in language models. In *The eleventh international conference on learning representations*.

CAI Yiran, HU Zhengyin, and LIU Chunjiang. 2025. Analysis of progress in data mining of scientific literature using large language models. *Journal of Library & Information Science in Agriculture*, 37(2).

Haoxuan Zhang, Ruochi Li, Yang Zhang, Ting Xiao, Jiangping Chen, Junhua Ding, and Haihua Chen. 2025a. The evolving role of large language models in scientific innovation: Evaluator, collaborator, and scientist. *arXiv preprint arXiv:2507.11810*.

Hongbo Zhang, Han Cui, Yidong Wang, Yijian Tian, Qi Guo, Cunxiang Wang, Jian Wu, Chiyu Song, and Yue Zhang. 2025b. Deep literature survey automation with an iterative workflow. *arXiv preprint arXiv:2510.21900*.

Bingxi Zhao, Lin Geng Foo, Ping Hu, Christian Theobalt, Hossein Rahmani, and Jun Liu. 2025. Llm-based agentic reasoning frameworks: A survey from methods to scenarios. *arXiv preprint arXiv:2508.17692*.

Kun Zhu, Lizi Liao, Yuxuan Gu, Lei Huang, Xiaocheng Feng, and Bing Qin. 2025. Context-aware hierarchical taxonomy generation for scientific papers via llm-guided multi-aspect clustering. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 15627–15645.