Closed-Loop Self-Improving Scientific Paper Generation: An Iterative Framework Combining AI Generation, Automated Review, and Targeted Optimization

Xucheng Yu[1] and Ruojia Tao[2] and Haohan Wang[1]

[1]University of Illinois Urbana-Champaign
[2]Cornell University

{xy63, haohanw}@illinois.edu, rt456@cornell.edu

Abstract

Recent advances in large language models (LLMs) have created new opportunities to support and partially automate multiple stages of the scientific research workflow, including literature synthesis, writing, and review. While end-to-end LLM-based systems can autonomously generate full research manuscripts, output quality varies substantially. To improve quality and consistency, we propose a self-improving framework that integrates iterative, review-augmented scientific writing into an autonomous paper-generation pipeline. Experiments on 15 generated papers show that review-guided revision improves quality from near-rejection (rating: 1.0) to borderline acceptance (rating: 4.7, max: 6.0), with improvements recognized by multiple independent evaluators (Claude: 100%, LLaMA: 86.7%, ChatGPT: 81.3%). These findings suggest a promising direction for more reliable autonomous scientific writing.

1 Introduction

Scientific research plays a central role in advancing human knowledge by systematically exploring unanswered questions across diverse fields(Solandt, 1957). Research articles are a primary vehicle for disseminating findings, providing a structured and archival record of research questions, methods, results, and interpretations. However, producing a high-quality manuscript remains complex and time-consuming.

Recent advances in large language models (LLMs) have created new opportunities to streamline parts of the scientific workflow. Researchers have begun applying them across multiple stages of the research and writing pipeline to explore how the LLM could improve each separate part of research, including research idea generation and hypothesis discovery, agentic systems that design and execute experiments and draft manuscripts, scientific writing and citation checking, and automated or LLM-assisted peer-review feedback generation (Luo et al., 2025; Yang et al., 2024; Baek et al., 2024; Si et al., 2024; Weng et al., 2025; Idahl and Ahmadi, 2025).

Building on these component-level advances, recent systems have further attempted fully automated open-ended scientific discovery; however, output quality is not consistently guaranteed(Lu et al., 2024). Common failure modes include incomplete or unsuccessful experiment execution, incorrect implementation of proposed ideas, incorrect citations, and resource constraints that prevent the generated manuscript from meeting the standards of a typical ML conference paper.

To address these issues, we introduce a self-improving framework that uses a separate model to generate peer-review–style feedback and actionable revision suggestions, which are then used to iteratively refine the generated manuscript and improve its quality and consistency. In our evaluation, this process can substantially raise manuscript quality—often moving drafts from a near-rejection level to borderline acceptability—suggesting a promising direction for more reliable fully automated paper generation.

2 Related Work

Recent work on autonomous scientific writing includes multi-agent systems like Data-to-paper (Ifargan et al., 2024) and end-to-end frameworks like AI-Scientist (Lu et al., 2024), which automate idea generation, experimentation, and manuscript writing. While demonstrating feasibility, these systems exhibit reliability issues including failed experiments and citation errors. For peer-review feedback, OpenReviewer (Idahl and Ahmadi, 2025) uses fine-tuned models trained on expert reviews, producing more critical and aligned feedback than general-purpose LLMs. In contrast to multi-agent approaches, we focus on review-based feedback

as a closed-loop self-improvement mechanism, using an independent evaluator to iteratively refine generated manuscripts.

3 Method

Our framework integrates an independent reviewer into the AI-Scientist pipeline through four stages: (1) **Paper Generation**: AI-Scientist generates complete manuscripts from executed experiments. (2) **Independent Review**: OpenReviewer, fine-tuned on conference reviews, evaluates manuscripts across novelty, soundness, presentation, contribution, and overall quality. (3) **Feedback Extraction**: Reviews are parsed into structured components (strengths, weaknesses, scores). (4) **Review-Guided Revision**: AI-Scientist revises manuscripts based on feedback, constrained to improving clarity without introducing new experiments. This process iterates until quality thresholds are met or iteration limits are reached.

4 Experimental Setup

We evaluate our review-augmented framework through two experiments: (1) measuring optimization effectiveness within the AI-Scientist–OpenReviewer pipeline, and (2) validating improvements through independent third-party evaluators.

4.1 Paper Generation and Optimization Protocol

We conduct 15 experimental runs using AI-Scientist across 2D diffusion modeling (8 papers) and nanoGPT language modeling (7 papers). Each paper is generated from executed experiments (6–8 pages, 3–5 metrics). We apply 5–8 optimization rounds per paper: (1) Convert PDF to Markdown; (2) OpenReviewer generates structured reviews with numerical scores (Novelty, Soundness, Presentation, Contribution: 1–4; Overall: 1/3/5/6/8/10) and textual feedback; (3) Parse to JSON; (4) AI-Scientist revises LaTeX without modifying findings; (5) Recompile and evaluate. Optimization terminates at 8/10 or after 8 iterations (all terminated at limits; none reached 8/10).

4.2 Cross-Model Evaluation

To validate robustness, we evaluate all paper pairs using four independent models: LLaMA-3-70B, ChatGPT (GPT-4), Claude-3.5-Sonnet, and Gemini-1.5-Pro.

Paper	I-Ovr	F-Nov	F-Snd	F-Pre	F-Con	F-Ovr	Iter
P1	1	2	3	3	2	5	6
P2	1	2	2	2	2	3	4
P3	1	2	2	2	2	5	5
P4	1	2	2	2	2	5	6
P5	1	2	3	3	2	5	6
P6	1	2	2	2	2	5	7
P7	1	2	3	3	2	5	5
P8	1	2	2	2	2	3	8
P9	1	3	3	3	2	5	5
P10	1	2	2	2	2	5	4
P11	1	3	2	3	2	6	6
P12	1	2	2	3	2	5	6
P13	1	2	2	3	2	5	6
P14	1	2	2	2	2	5	4
P15	1	2	2	2	2	3	8

Table 1: Initial and final scores for 15 papers. I-Ovr=Initial Overall, F-Nov/Snd/Pre/Con/Ovr=Final Novelty/Soundness/Presentation/Contribution/Overall, Iter=Iterations. All start at (1,1,1,1,1).

For each of the 15 papers, both versions are presented in PDF format with randomized order (A/B). Each evaluator receives a standardized prompt requesting binary preference with justification across five dimensions:

- **Factual accuracy**: Are claims supported by evidence?
- **Logical coherence**: Do arguments follow logically?
- **Structural organization**: Are sections well-organized?
- **Clarity**: Is language clear and precise?
- **Overall quality**: Which paper is better overall?

Evaluators provide forced binary choices (A or B) to ensure decisiveness. This yields 75 comparisons per model (15 papers $\times$ 5 dimensions).

5 Results

5.1 Review-Guided Optimization

All papers begin with minimum scores (1,1,1,1,1), indicating that autonomous generation without review guidance produces manuscripts far below acceptance standards. After 5–8 optimization rounds guided by OpenReviewer feedback, substantial improvements are observed across all metrics.

Table 2 shows substantial improvements across all metrics. Presentation shows the largest gain (+1.5), followed by Soundness (+1.3) and Novelty (+1.1). Contribution improves least (+1.0), consistent with our no-new-experiments constraint. Overall rating increases by +3.7 points on average.

Metric	Initial	Final	Peak	Avg Δ	Max Δ
Novelty	1.0	2.1	2.1	+1.1	+2
Soundness	1.0	2.3	2.5	+1.3	+2
Presentation	1.0	2.5	2.7	+1.5	+2
Contribution	1.0	2.0	2.1	+1.0	+1
Overall Rating	1.0	4.7	4.7	+3.7	+5

Table 2: Average score improvements across all metrics (n=15 papers). Initial: Starting score, Final: Score after last iteration, Peak: Highest score achieved during optimization (may be unstable), Avg Δ: Average improvement, Max Δ: Maximum improvement observed. Presentation shows the largest improvement, while contribution shows the smallest, consistent with our revision-only approach.

Iteration	Paper 1	Paper 4	Paper 9	Paper 11
0	1	1	1	1
1	3	5	3	1
2	3	3	1	3
3	3	5	3	3
4	5	3	3	3
5	5	3	5	3
6	5	5	—	6
Pattern	Steady	Unstable	Early peak	Best score

Table 3: Representative optimization trajectories showing overall rating across iterations. Paper 4 demonstrates high instability (5→3→5→3), Paper 11 achieves the highest final score (6), Paper 9 shows early peak followed by stabilization, and Paper 1 exhibits steady improvement. These patterns illustrate the diverse and non-monotonic nature of the optimization process.

5.1.1 Optimization Instability Analysis

Table 3 shows representative patterns: Paper 11 achieves the highest score (6), Paper 4 exhibits instability (1→5→3→5→3→3→5), and Paper 9 shows early peaks. Optimization is non-monotonic with frequent score fluctuations, stabilizing around iteration 5. Most papers reach 3–5, consistent with revision limitations, though many intermittently achieve 5–6.

Score improvements exhibit diminishing returns: the largest gains occur in rounds 1–3, with saturation around iteration 5. Most papers stabilize at rating 3–5, consistent with our hypothesis that revision cannot fundamentally alter research novelty or soundness without introducing new experiments. However, most papers achieve ratings of 5–6 intermittently across iterations, though unstably, corresponding to borderline acceptance through improved motivation, limitations discussion, and related work integration.

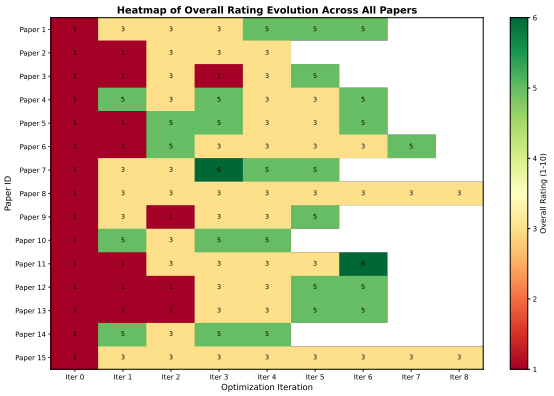

Figure 1: Heatmap visualization of overall rating evolution across all 15 papers and optimization iterations. White cells indicate iterations not performed. The visualization reveals clustering of final scores around 3 and 5, with Paper 11 (bottom cluster) uniquely achieving 6. The color gradient from red (low scores) to green (high scores) illustrates the general trend of improvement while highlighting the instability of intermediate iterations.

Final Overall Rating	Count (%)	Papers
3 (Reject)	3 (20.0%)	P2, P8, P15
5 (Marginally below)	11 (73.3%)	P1, P3–P7, P9, P10, P12–P14
6 (Marginally above)	1 (6.7%)	P11
8+ (Accept)	0 (0.0%)	—
Total	**15 (100%)**	**All**

Table 4: Distribution of final overall ratings after optimization. Most papers (73.3%) achieve a rating of 5 (marginally below acceptance), with only one paper reaching 6 (marginally above acceptance). No paper achieved the termination criterion of 8/10, indicating fundamental limitations of revision-only optimization.

5.1.2 Final Score Distribution

The majority of papers (73.3%, n=11) achieve a rating of 5, corresponding to "marginally below the acceptance threshold" according to our review rubric. Three papers (20.0%) finish at rating 3 ("reject, not good enough"), and only one paper (6.7%, Paper 11) reaches rating 6 ("marginally above the acceptance threshold"). Critically, no paper achieved ratings of 8 or higher, which would indicate clear acceptance quality.

This distribution supports our central claim: review-guided revision can elevate papers from near-rejection (1.0) to borderline acceptance levels (3–5), but consistent acceptance likely requires new empirical contributions beyond what revision alone can provide. The scarcity of rating-6 papers and complete absence of rating-8+ papers suggests a ceiling effect where presentation improvements

Model	Prefer Optimized	Agreement Level
Claude-3.5-Sonnet	75/75 (100.0%)	Very High
LLaMA-3-70B	65/75 (86.7%)	High
ChatGPT (GPT-4)	61/75 (81.3%)	High
Gemini-1.5-Pro	41/75 (54.7%)	Low

Table 5: Cross-model evaluation across 75 comparisons (15 papers $\times$ 5 dimensions). Claude demonstrates perfect consistency, while Gemini shows near-chance performance, revealing significant differences in evaluator sensitivity to paper quality improvements.

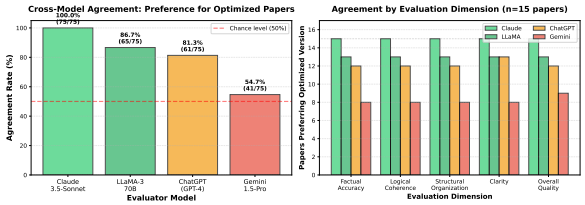

Figure 2: Cross-model evaluation: (left) agreement rates, (right) breakdown by dimension. Claude: 100%, LLaMA: 86.7%, ChatGPT: 81.3%, Gemini: 54.7% (near-chance).

cannot compensate for limitations in novelty and experimental soundness.

5.2 Cross-Model Evaluation

Across model validation results, we observe substantial variation in evaluator sensitivity. Claude demonstrates perfect agreement (100%), consistently identifying optimized versions across all 75 comparisons. LLaMA-3 and ChatGPT show strong preferences (86.7% and 81.3% respectively), while Gemini exhibits near-chance performance (54.7%).

Claude achieves perfect agreement (75/75) while Gemini performs near-chance (41/75). The Claude-Gemini divergence (100% vs 54.7%) reveals that Claude aligns with structural improvements, while Gemini may prioritize experimental novelty. Improvements are most consistently recognized in structural organization and clarity.

A representative example: Paper 1 initially scored 1.0. Through 6 iterations, revisions restructured the introduction, expanded related work, and standardized notation, achieving rating 5.0. Claude, LLaMA, and ChatGPT recognized improvements across all dimensions; Gemini preferred the optimized version in only 3 of 5 dimensions.

6 Discussion

Our results demonstrate that independent, reviewer-aligned feedback effectively improves

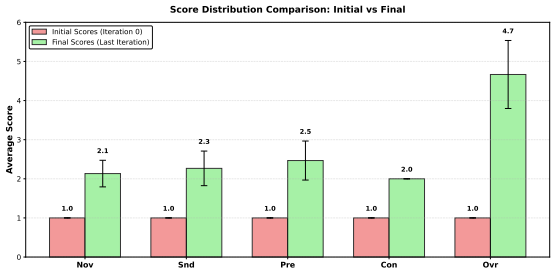

Figure 3: Comparison of score distributions before and after optimization across five metrics. Initial scores (red bars) and final scores (green bars) are shown side-by-side for direct comparison. Error bars represent standard deviation. All metrics show significant improvements, with presentation (Pre) and soundness (Snd) exhibiting the largest gains. The overall rating (Ovr) increases from 1.0 to 4.7 on average, though with high variance reflecting the diverse final outcomes across papers.

autonomously generated papers without modifying experimental content. Score saturation around 3–5 reflects inherent limitations: novelty and soundness remain constrained by initial experiments, though presentation improvements can elevate papers to borderline acceptance.

Cross-model evaluation reveals significant LLM heterogeneity. While Claude (100%), LLaMA (86.7%), and ChatGPT (81.3%) show strong consensus, Gemini's near-chance performance (54.7%) indicates improvements are not universally recognized, likely reflecting differing evaluator priorities.

Limitations Our evaluation relies on LLM judgments without human validation. We test only two domains, limiting generalizability. Score instability indicates revision-only approaches cannot consistently achieve acceptance standards.

7 Conclusion

We present a review-augmented autonomous scientific writing framework integrating an independent reviewer into the AI-Scientist pipeline. Experiments on 15 papers show review-guided revision improves quality from near-rejection (1.0) to borderline acceptance (4.7), without introducing new experiments. Cross-model validation reveals improvements are consistently recognized by Claude, LLaMA, and ChatGPT, though Gemini shows near-chance agreement (54.7%). These findings suggest a promising direction for autonomous research frameworks while highlighting the importance of evaluator selection.

References

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. 2024. Researchagent: Iterative research idea generation over scientific literature with large language models. *ArXiv*, abs/2404.07738.

Maximilian Idahl and Zahra Ahmadi. 2025. Openreviewer: A specialized large language model for generating critical scientific paper reviews. *Preprint*, arXiv:2412.11948.

Tal Ifargan, Lukas Hafner, Maor Kern, Ori Alcalay, and Roy Kishony. 2024. Autonomous llm-driven research from data to human-verifiable research papers. *Preprint*, arXiv:2404.17605.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *Preprint*, arXiv:2408.06292.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. Llm4sr: A survey on large language models for scientific research. *Preprint*, arXiv:2501.04306.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2024. Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. *Preprint*, arXiv:2409.04109.

Omond Solandt. 1957. Scientific research in the modern world.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. *Preprint*, arXiv:2411.00816.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. 2024. Large language models for automated open-domain scientific hypotheses discovery. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13545–13565, Bangkok, Thailand. Association for Computational Linguistics.

A Implementation Details

A.1 Review Generation

OpenReviewer generates structured peer reviews from papers provided in PDF format. Each paper is first converted into a Markdown representation using PyMuPDF, preserving textual content and section structure. Based on the extracted Markdown, the system produces a review following a standardized template inspired by OpenReview-style conference reviews.

The generated review consists of the following components:

- **Summary**: A concise and neutral overview of the paper's problem setting, proposed approach, and main contributions.
- **Strengths**: Key positive aspects of the work, presented as a small number of concrete points supported by evidence from the paper.
- **Weaknesses**: Major limitations or concerns affecting validity, clarity, or impact, formulated as specific and actionable points.
- **Questions**: Clarification questions for the authors, whose answers could potentially influence the overall assessment.
- **Scores**: Numerical ratings for *Novelty*, *Soundness*, *Presentation*, and *Contribution*, each on a 1–4 scale, accompanied by dedicated explanatory justifications, as well as an *Overall Rating* on a 1/3/5/6/8/10 scale.

During post-processing, the questions section is additionally interpreted as actionable feedback to guide subsequent paper revisions.

A.2 Cross-Model Evaluation Prompt

The standardized prompt used for all four evaluators (Claude, LLaMA, ChatGPT, Gemini):

> *You are evaluating two versions of a research paper (Paper A and Paper B). Compare them across the following dimensions and select which is better:*
>
> - *Factual accuracy: Are claims supported by evidence?*
> - *Logical coherence: Do arguments follow logically?*
> - *Structural organization: Are sections well-organized?*
> - *Clarity: Is language clear and precise?*
> - *Overall quality: Which paper is better overall?*
>
> *For each dimension, respond with either "A" or "B" and provide brief justification.*

A.3 Example Review

Below is a representative initial review generated by OpenReviewer, demonstrating the detailed and critical feedback that drives the optimization process. This review assigns the minimum scores on

all sub-scales (Novelty, Soundness, Presentation, Contribution: 1 on a 1–4 scale; Overall Rating: 1 on a 1/3/5/6/8/10 scale), illustrating the low quality of autonomous generation before optimization:

Summary: This paper proposes a simple method to accelerate grokking by assigning different learning rates to different layers.

Weaknesses:

- Lacks novelty—merely assigns different learning rates without theoretical justification
- Experiments limited to small-scale models and toy tasks
- No comparison with existing adaptive learning rate methods
- Learning rate choices appear arbitrary

Questions:

- What is the rationale behind the learning rate choices?
- Is there theoretical or empirical evidence for the acceleration claim?
- How does this compare to existing methods?

Scores: Novelty: 1, Soundness: 1, Presentation: 1, Contribution: 1, Overall: 1

After 6 optimization iterations addressing these weaknesses (added theoretical motivation, expanded related work, clarified hyperparameters), the final version achieved: Novelty: 2, Soundness: 3, Presentation: 3, Contribution: 2, Overall: 5.