ReviewRL: Towards Automated Scientific Review with RL

Sihang Zeng[2]*, Kai Tian[1]*, Kaiyan Zhang[1]*, Yuru Wang[1]
Junqi Gao[3], Runze Liu[1,3], Sa Yang[4], Jingxuan Li [5]
Xinwei Long[1], Jiaheng Ma[6], Biqing Qi[3†], Bowen Zhou[1,3†]
[1]Tsinghua University, [2]University of Washington, [3]Shanghai AI Laboratory
[4]Peking University, [5]Harbin Engineering University, [6]Beijing Institute of Technology
qibiqing@pjlab.org.cn, zhoubowen@tsinghua.edu.cn

Abstract

Peer review is essential for scientific progress but faces growing challenges due to increasing submission volumes and reviewer fatigue. Existing automated review approaches struggle with factual accuracy, rating consistency, and analytical depth, often generating superficial or generic feedback lacking the insights characteristic of high-quality human reviews. We introduce ReviewRL, a reinforcement learning framework for generating comprehensive and factually grounded scientific paper reviews. Our approach combines: (1) an ArXiv-MCP retrieval-augmented context generation pipeline that incorporates relevant scientific literature, (2) supervised fine-tuning that establishes foundational reviewing capabilities, and (3) a reinforcement learning procedure with a composite reward function that jointly enhances review quality and rating accuracy. Experiments on ICLR 2025 papers demonstrate that ReviewRL significantly outperforms existing methods across both rule-based metrics and model-based quality assessments. ReviewRL establishes a foundational framework for RL-driven automatic critique generation in scientific discovery, demonstrating promising potential for future development in this domain. The implementation of ReviewRL will be released at GitHub.

1 Introduction

Peer review is critical for scientific progress, ensuring that published research meets rigorous standards of quality, validity, and significance. However, the growing volume of submissions to academic conferences and journals has created unsustainable pressure on the review system, leading to reviewer fatigue, inconsistent evaluations, and increasingly long review cycles (Hosseini and Horbach, 2023; Kim et al., 2025). For instance, top-tier
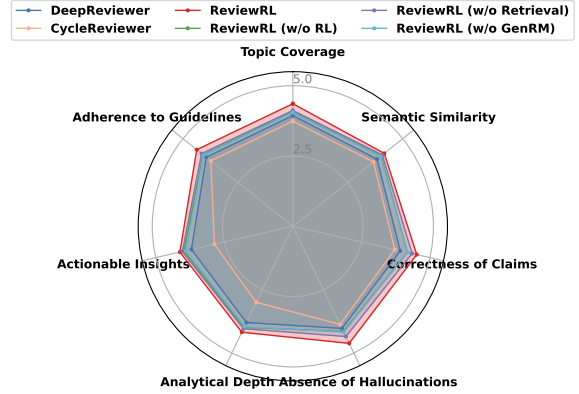

Figure 1: Evaluation results of ReviewRL under the criteria of ReviewEval (Kirtani et al., 2025)

AI conferences like NeurIPS and ICLR now process thousands of submissions annually, requiring tens of thousands of reviews (Kim et al., 2025). This explosion in scientific output has amplified the need for automated tools to assist or augment the peer review process.

Recent advances in large language models (LLMs) have created promising opportunities for AI-assisted scientific assessment. These models can analyze complex scientific texts, identify methodological strengths and weaknesses, and generate structured feedback at scale (Weng et al., 2024; Lu et al., 2024; Zhu et al., 2025; Qi et al., 2024; Zhang and Abernethy, 2025). However, existing approaches to automated paper reviewing face three significant challenges. First, they often struggle to maintain factual accuracy and provide evidence-based critiques that connect the paper to relevant prior work (Zhou et al., 2024). Second, they tend to overestimate paper quality, assigning ratings that are inconsistently aligned with human judgment (Yu et al., 2025). Third, they frequently generate superficial or generic reviews lacking the analytical depth and actionable insights characteristic of human reviews (Shin et al., 2025).

---

*Equal contribution.
†Corresponding author.

16931

Recent research has demonstrated the effectiveness of reinforcement learning (RL) in enhancing LLMs' reasoning capabilities. Models like DeepSeek-R1 (Guo et al., 2025) have achieved impressive performance improvements through carefully designed RL training regimes, while innovations such as Group Relative Policy Optimization (Shao et al., 2024) and Reinforce++ (Hu, 2025) have made RL more efficient and stable for LLM training. Concurrently, the Model Context Protocol (MCP) has emerged as a standardized communication framework that enables LLMs to interact seamlessly with external knowledge sources (Hou et al., 2025), facilitating accurate information retrieval and reducing hallucinations. The combination of enhanced reasoning through RL and factual grounding via MCP-based retrieval offers a promising approach to addressing the limitations of current automated review systems.

In this paper, we introduce ReviewRL, a reinforcement learning framework for generating comprehensive, factually grounded, and constructively critical scientific paper reviews. Our approach combines three key components: (1) a ArXiv-MCP based retrieval-augmented context generation pipeline that identifies and incorporates relevant scientific literature to support factual assessments, (2) a supervised fine-tuning (SFT) phase that establishes foundational reviewing capabilities and initial rating alignment, and (3) a RL optimization procedure that jointly enhances review quality and rating accuracy. Through this integrated approach, ReviewRL produces reviews that not only emulate human-like analytical depth but also provide ratings that consistently align with human judgments. Our experiments demonstrate that ReviewRL significantly outperforms existing approaches across both rule-based metrics and model-based assessments of review quality. We further examine the importance of each component through comprehensive ablation studies, revealing that both retrieval augmentation and our composite reward formulation contribute substantially to ReviewRL's superior performance. To our knowledge, this represents the first successful application of reinforcement learning to enhance both the quality and rating consistency of automated scientific peer reviews. Our contributions include as follows:

1) We introduce ReviewRL, a novel framework that integrates RL for automatic paper review generation. ReviewRL comprises three key components: ArxivMCP, context-aware fine-tuning, and composed reward RL training.

2) Unlike previous approaches such as DeepSeek-R1, which rely on rule-based rewards, we find such rewards insufficient for review generation, where structural coherence and content quality are paramount. To address this, we design a comprehensive reward system incorporating both rule-based metrics and judge-model-based evaluations, effectively mitigating the limitations of purely rule-based methods.

3) Compared to prior work, ReviewRL achieves superior performance in context-awareness, factual consistency, and review depth. This framework represents a preliminary step toward RL-driven automatic critique generation in scientific discovery.

2 Related Works

LLM for Paper Review   Recent advancements have explored the use of LLMs to automate and enhance the academic peer review process. Early efforts, such as PeerRead (Kang et al., 2018) and NLPeer (Dycke et al., 2022), provided foundational datasets and benchmarks for review generation and analysis. Building upon these resources, systems like Reviewer2 (Gao et al., 2024) proposed a two-stage framework involving aspect prompt generation and review generation to improve the specificity and coverage of generated reviews. CycleResearcher (Weng et al., 2024) and The AI Scientist (Lu et al., 2024) have introduced end-to-end frameworks that simulate the entire research lifecycle, including manuscript drafting and iterative peer review, where their reviewer modules are trained via supervised fine-tuning or operate through agentic inference. More recently, DeepReviewer (Zhu et al., 2025) is trained through SFT using long chain-of-thought (CoT) data to enhance its reasoning ability. Despite these advancements, challenges remain in ensuring the factualness, reasoning depth and rating consistency of LLM-generated reviews.

Reinforcement Learning for LLMs   Reinforcement Learning (RL) (Sutton et al., 1998) plays a pivotal role in enhancing the instruction-following capabilities of Large Language Models (LLMs), particularly through approaches like Reinforcement Learning from Human Feedback (RLHF) (Ouyang et al., 2022). RLHF aligns foundation models with human preferences, typically leveraging algorithms such as Proximal Policy Optimization (PPO) (Schulman et al., 2017) or Direct Preference Optimization (Rafailov et al., 2023),
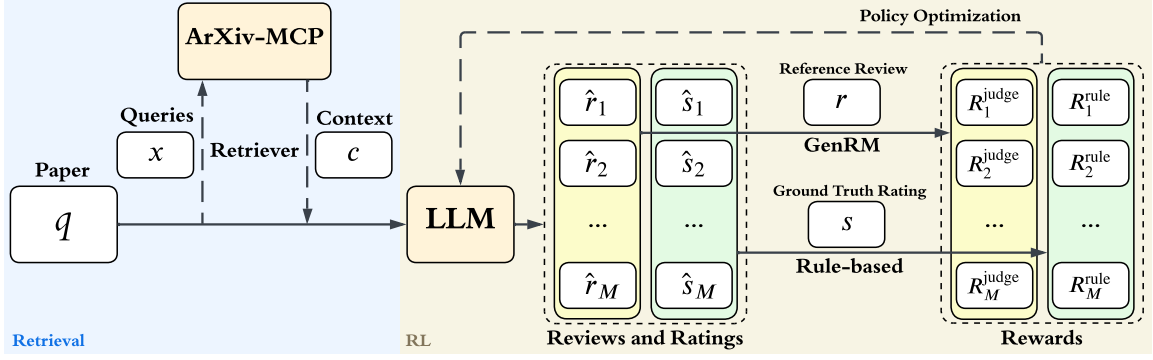
Figure 2: Overview of ReviewRL, including Arxiv-MCP, SFT warm up, and RL optimization.

where precise preference modeling is essential. Recent advancements have demonstrated RL's effectiveness in improving reasoning abilities in Large Reasoning Models (LRMs), such as DeepSeek-R1 (Guo et al., 2025), through rule-based reward mechanisms, as exemplified by GRPO (Shao et al., 2024). Unlike RLHF, which is generally applied to open-domain instructions, GRPO is specifically designed to promote extended Chain-of-Thought (CoT) (Wei et al., 2022) reasoning, particularly in mathematical problem-solving scenarios. Benefiting from its simplicity and effectiveness, GRPO has been successfully applied across diverse domains, including vision understanding and generation (Liu et al., 2025; Team et al., 2025; Xue et al., 2025), agentic search and planning (Li et al., 2025; Jin et al., 2025; Zhang et al., 2025), and beyond. However, the potential of RL methods like GRPO to enhance review and critique generation (Whitehouse et al., 2025) still need more exploration.

3 Methodology

3.1 Task Formulation

Given a target paper q, the automated scientific review task is defined as generating a comprehensive review r, including the paper's strengths and weaknesses, and a rating s. To ensure high-quality review generation, we formulate ReviewRL's workflow as a retrieval-augmented generation (RAG) (Lewis et al., 2020) and a LLM reasoning process. This process mimics the cognitive steps of human reviewers—retrieving relevant domain knowledge, analyzing the paper in context, and making evaluative judgments. Specifically, a retriever model R generates three queries x and identifies a set of relevant contextual papers c through searching, for-

mulated as $q \xrightarrow{R} x, c$. An LLM-based reviewer π then reasons over the paper and the retrieved context to produce an intermediate thinking process z, represented as $(q, c) \xrightarrow{\pi} z$. Finally, the LLM generates the review and rating based on the paper, the retrieved context, and the reasoning trace, i.e., $(q, c, z) \xrightarrow{\pi} (r, s)$.

In the following sections, we present the components of ReviewRL as shown in Figure 2. We first introduce the RAG pipeline (Section 3.2) that accurately identifies contextually relevant literature given the target paper. We then describe our training strategy for ReviewRL, which combines SFT (Section 3.3) and RL (Section 3.4) to enhance reasoning capabilities.

3.2 Context Retrieval

For each paper q, we retrieve relevant contextual information c using a retrieval pipeline R. Following a two-step approach inspired by the novelty verification in DeepReviewer (Zhu et al., 2025), we first generate query questions x and then retrieve relevant contexts c from ArXiv. This method leverages an LLM agentic workflow and integrates the Model Context Protocol (MCP) for efficient context retrieval and generation.

Specifically, we employ Qwen3-8B (Yang et al., 2025) to analyze the target paper q and generate three query questions x that probe the paper's novelty, methodology, and relationship to prior work. These queries are formulated as natural language questions rather than keywords, allowing for more nuanced retrieval of relevant literature. For example, a query might ask "What recent papers have proposed reinforcement learning for LLM-based paper reviews?" rather than simply searching for "reinforcement learning LLM reviews."

We implement the retrieval functionality through ArXiv-MCP[1], an open-source Model Context Protocol server that provides LLMs with standardized access to the arXiv repository. ArXiv-MCP enables efficient paper search, filtering, and full-text retrieval without requiring low-level API implementation. The server processes the generated queries and returns structured information including paper metadata, abstracts, and relevant full-text excerpts.

The retrieval execution is orchestrated by Qwen-Agent[2], an agent framework built on the Qwen model family that provides function calling and tool orchestration capabilities. Qwen-Agent sequentially routes each query to ArXiv-MCP and manages the retrieval results. The retrieved contexts c undergo post-processing to remove tool invocation artifacts and are consolidated into a coherent format that includes: (1) query-response pairs, (2) bibliographic information of retrieved papers, and (3) relevance-ranked excerpts from these papers. This processed context is then concatenated with the original paper representation to form the input for the review generation model.

3.3 Supervised Finetuning

Given a paper q and its retrieved context c, the next step is to generate the review r and the corresponding rating s using a policy model π. While our goal is to enhance this process using RL, direct RL application on base models presents challenges. These models typically overestimate paper quality compared to human reviewers (Yu et al., 2025), leading to uninformative trajectories with weak reward signals and unstable RL training. Empirically, we observe that without proper initialization, RL training exhibits a cold-start problem characterized by training collapse and performance degradation (e.g., generated ratings collapse around 6).

To mitigate this, we first apply SFT on long CoT data to initialize the RL policy with essential review-writing capabilities. This strategy is inspired by similar practice in DeepSeek-R1 (Guo et al., 2025) and Kimi-k1.5 (Team et al., 2025). We leverage data derived from DeepReview-13k (Zhu et al., 2025), a high-quality dataset comprising long CoT reviews and accurate rating annotations, as the cold-start data for training the model. Specifically, we use the ICLR 2024 portion of the dataset and preprocess it to fit our task definition. We include

their novelty verification results and the corresponding queries from the best mode in the input, and use the final meta review as the output, with intermediate analysis regarded as the long CoT thinking process. We train for 2 epochs on top of the model Qwen2.5-7B-Instruct (Team, 2024). Different from previous work without RL, in our framework, SFT serves two primary goals: (1) to equip the policy model with foundational reasoning ability to perform structured and reasoned peer reviews, and (2) to align predicted scores with human ratings, thereby stabilizing downstream RL training and preventing early-stage collapse.

3.4 Reinforcement Learning

Following the SFT phase, we conduct large-scale reinforcement learning (RL) to further enhance the reasoning capabilities of the LLM reviewer. Paper reviewing is a non-verifiable problem with a partially verifiable outcome—the numerical rating—where both the *review quality* and *rating consistency* with human judgments are essential. Prior work has demonstrated the effectiveness of rule-based outcome rewards in improving LLM reasoning (Guo et al., 2025). However, our experiments show that relying solely on a rating consistency reward leads to overly generic reviews lacking analytical depth and actionable insights, indicating insufficient reasoning ability.

To jointly optimize the review quality and rating consistency, we design a composite reward that integrates rule-based rewards with a generative reward model (GenRM) (Zhang et al., 2024), which prioritizes reviews with high rating consistency, format adherence, and strong analytical depth.

Rule-Based Rewards   We define two rule-based reward components: rating consistency reward and format reward. The rating consistency reward R_{rc} is computed using a Gaussian kernel:

$$R_{rc} = \exp\left(-\frac{(s - \hat{s})^2}{2\sigma^2}\right) \quad (1)$$

where s denotes the ground-truth rating, obtained by averaging human-assigned scores for the given paper, and $\hat{s}$ is the rating predicted by the model. The format adherence reward R_f penalizes outputs that omit essential structural components. Let $\mathcal{S}$ be the set of required elements, including a reasoning block (delimited by <think> and </think>), summary, strengths, and weaknesses:

$$R_f = -\sum_{s \in \mathcal{S}} \mathbb{1}(s \text{ is missing}) \qquad (2)$$

The final rule-based reward is given by:

$$R^{\text{rule}} = \text{clip}\left(\alpha \cdot R_{rc} + \beta \cdot R_f, 0, 1\right) \qquad (3)$$

where α and β are hyperparameters that balance the importance of rating consistency and format completeness. This reward formulation encourages the generation of outputs that are both aligned with human ratings and structurally well-formed.

GenRM-based Rewards   Following prior work (Seed et al., 2025; Hogan, 2024), we employ a GenRM π_{judge} to evaluate the quality of the LLM-generated review $\hat{r}$ against a reference r. The reward is derived from the win rate, based on the agreement that LLM-as-a-judge can reliably assess relative response quality (Zheng et al., 2023). In our framework, π_{judge} evaluates reviews across multiple dimensions: factual accuracy, completeness, level of detail, comparison with related work, constructiveness, and clarity. The GenRM reward R^{judge} is defined as:

$$R^{\text{judge}} = \begin{cases} 1 & \text{if } \hat{r} \text{ is preferred} \\ 0 & \text{if } r \text{ is preferred} \end{cases} \qquad (4)$$

The final reward signal is computed as a weighted combination of the rule-based reward and the GenRM reward:

$$R^{\text{final}} = \gamma R^{\text{rule}} + (1 - \gamma)R^{\text{judge}} \qquad (5)$$

RL Training Data   We construct the RL training dataset using papers from top-tier machine learning conferences, such as ICLR and ACL, sourced from the raw data of Reviewer2 (Gao et al., 2024) and the best mode split of DeepReview-13k (Zhu et al., 2025). For each paper q, we retrieve the context c using the method described in Section 3.2. Ratings from different conferences are normalized to a common scale of 1–10, and the ground truth rating s is computed as the average of scores from multiple human reviewers. Reference reviews r are derived as follows: for each paper from Reviewer2, we summarize multiple human reviews using DeepSeek-R1-Distill-Qwen-32B into a single formatted review; for DeepReview-13k, we use the meta-reviews from the best mode split. ICLR 2025

data are excluded to avoid data leakage. Dataset statistics are reported in Table 1. Because a large proportion of ground truth ratings fall between 5 and 6, we apply a *balancing* preprocessing step that downsamples papers with mid-range ratings (5–6) and upsamples those with more extreme ratings. This strategy emphasizes papers with highly positive or negative assessments, which tend to be more informative for learning, and helps prevent the RL model from collapsing to generic, non-discriminative ratings around the middle range.

	ICLR	NeurIPS	ARR	COLING	CONLL	ACL
Year	2017-2024	2021-2022	2022	2020	2016	2017
Count	13312	3994	336	82	22	131
Avg. #Token	9854	10275	9153	8138	7888	8571

Table 1: RL training data statistics

Therefore, the RL training data comprises tuples of (q, c, s, r), without access to the intermediate reasoning steps that lead to the review r and rating s. This setup encourages the policy model to explore its own reasoning trajectories that produce high-quality reviews and ratings consistent with human judgments.

RL Training Setting   The policy model π is initialized from the supervised finetuned model π_{sft} to ensure stable learning and prevent cold-start collapse. We adopt the Reinforce++ algorithm (Hu, 2025). π_{judge} is a Qwen2.5-14B-Instruct model. Training details are shown in Appendix C.

4   Experiments

4.1   Evaluation Data

We construct the evaluation set by sampling 472 papers from the ICLR 2025 review corpus. For each paper, the ground truth rating is computed as the average of scores assigned by all human reviewers. To ensure fair evaluation across the full rating spectrum, we sample papers such that the distribution of average ratings is approximately uniform across the rating scale.

4.2   Evaluation Metrics

We employ two families of quantitative metrics: **(i) rule-based** metrics, and **(ii) model-based** metrics.

4.2.1   Rule-based Quantitative Metrics

We evaluate the model using both rating-level and pairwise-level metrics. For score prediction, we compute the mean squared error (MSE) and Spearman rank correlation between predicted scores and

ground truth ratings. For pairwise paper evaluation, following JudgeLRM (Chen et al., 2025), we assess the model's ability to rank paper quality using three pairwise metrics: *relation*, *absolute*, and *confidence*. These respectively measure directional consistency with human rankings, score closeness to ground truth, and discriminative strength in differentiating papers of varying quality. Concordance index is also reported as a global ranking metric. Formal definitions of the pairwise metrics are provided in Appendix B.

4.2.2 Model-based Quantitative Metrics

While rule-based metrics focus on the accuracy of the generated ratings, it is equally important to assess whether the generated reviews emulate human-written reviews and provide constructive, content-rich feedback. To this end, we adopt an LLM-as-a-judge framework (Zheng et al., 2023) inspired by the ReviewEval benchmark (Kirtani et al., 2025), evaluating review quality across seven dimensions. Each dimension is rated on a 1-5 scale and aims to capture a distinct aspect of human-aligned peer reviewing:

- **Topic Coverage:** Does the AI-generated review comprehensively address the main topics and arguments of the paper? Does it cover aspects typically emphasized by human reviewers?

- **Semantic Similarity:** Does the review capture the core critique and suggestions of a plausible human review, even if phrased differently?

- **Correctness of Claims:** Are the statements in the review factually accurate with respect to the paper's content? Does the review avoid misinterpretations or incorrect representations of the methodology, results, or conclusions?

- **Absence of Hallucinations:** Does the review refrain from introducing information or claims not supported by the paper?

- **Analytical Depth:** Does the review demonstrate deep engagement with the research? This includes evaluating methodological rigor, identifying logical gaps, interpreting results, and contextualizing contributions within related work.

- **Actionable Insights:** Does the review provide specific, constructive suggestions for improving the paper? Are the recommendations practical and clearly articulated?

- **Adherence to Guidelines:** Does the review follow standard academic review criteria such as originality, significance, methodological soundness, clarity, and ethical compliance (if applicable)?

This evaluation framework enables a comprehensive assessment of the model's ability to perform nuanced and human-aligned paper reviewing beyond surface-level metrics. Llama-3.3-70B-Instruct is used as the judge model.

4.3 Baselines

We compare against three classes of baselines. **Open-source *instruction*** models (e.g., Qwen-2.5-Instruct) and **Open-source *reasoning*** models (e.g., Qwen 3) provides baseline paper review performance with basic instruction following ability and enhanced reasoning capabilities. Additionally, **SFT** models trained on public peer-review datasets, such as CycleReviewer-8B, are included to highlight the performance gain achieved by our RL-enhanced model over purely supervised approaches.

5 Results

5.1 Rule-based Evaluation Results

Table 2 presents rule-based evaluation results. Open-source *instruct* models perform weakest overall, showing poor rating accuracy and limited ranking capability, even at larger scales. Open-source *reasoning* models improve upon pairwise metrics, particularly in discriminative strength (Pair-Confidence), but still lag in MSE. SFT models trained on peer review datasets demonstrate significant gains in the alignment with human ratings—e.g., DeepReviewer-7B achieves the best Pair-Relation score. Our proposed ReviewRL model achieves the strongest performance across the board. The performance gap between ReviewRL and its SFT-only counterpart highlights the effectiveness of reinforcement learning in optimizing rating consistency.

5.2 Model-based Evaluation Results

Figure 1 shows the model-based evaluation results across seven review quality dimensions. ReviewRL consistently outperforms all baselines, particularly in dimensions that require deeper reasoning and reliability, such as analytical depth. Compared to its supervised-only counterpart, ReviewRL exhibits clear improvements in all dimensions, highlighting the benefits of RL in refining review generation.

Table 2: Rule-based evaluation results. ReviewRL consistently outperforms baseline methods.

Model	MSE ↓	Spearman ↑	Pair-Relation ↑	Pair-Absolute ↑	Pair-Confidence ↑	Concordance ↑
Open Source Instruct						
Qwen2.5-7B-Instruct	12.024	0.158	0.514	0.051	0.138	0.668
Qwen2.5-32B-Instruct	9.847	0.147	0.538	0.055	0.345	0.575
Qwen2.5-72B-Instruct	9.418	0.325	0.529	0.074	0.318	0.705
Llama-3.3-70B-Instruct	9.839	0.285	0.539	0.061	0.286	0.687
Open Source Resoning						
DeepSeek-R1-Distill-Qwen-7B	9.247	0.062	0.512	0.063	0.399	0.527
DeepSeek-R1-Distill-Qwen-14B	10.064	0.271	0.525	0.065	0.281	0.683
DeepSeek-R1-Distill-Qwen-32B	6.463	0.341	0.569	0.097	0.414	0.677
DeepSeek-R1-Distill-Llama-70B	9.021	0.389	0.539	0.065	0.336	0.747
QwQ-32B	5.440	0.425	0.585	0.128	0.402	0.743
Qwen3-8B	4.852	0.237	0.567	0.157	0.294	0.649
Qwen3-14B	8.348	0.371	0.534	0.072	0.391	0.706
Qwen3-32B	9.613	0.415	0.528	0.182	**0.462**	0.753
SFT Training						
CycleReviewer-ML-Llama3.1-8B	4.409	0.482	0.495	0.192	0.355	0.743
DeepReviewer-7B	3.445	0.539	**0.639**	0.245	0.245	0.710
ReviewRL-7B (w/o RL)	2.829	0.335	0.528	0.135	0.260	0.644
RL Training						
ReviewRL-7B	**2.585**	**0.634**	0.579	**0.249**	0.360	**0.806**

These results further demonstrate that RL leads to more informative, faithful, and constructive reviews.

5.3 RL Training Dynamics

We analyze the RL training dynamics of ReviewRL to provide insights for future training recipes in LLM reviewer models.

Training Curves Figure 3 illustrates the training dynamics of ReviewRL across three key metrics. As training progresses, the training reward steadily increases, indicating that the policy is effectively optimizing for the reward function. Simultaneously, the response length grows in the earlier stages and stabilizes after approximately step 10, suggesting that the model learns to generate more detailed outputs. The evaluation MSE decreases consistently over training steps, confirming that the learned policy generalizes better to held-out data and produces more accurate review ratings. These trends collectively demonstrate the effectiveness of the RL training procedure.

Cold-Start Phase We observe a cold-start issue in ReviewRL when RL training is initiated without proper policy initialization. As shown in Figure 4, training from scratch leads to rating collapse, where the model predominantly outputs generic scores around 6 and fails to differentiate between input papers. Our data balancing strategy for RL training, which upweights examples with extreme

ground-truth ratings, partially mitigates this issue but remains insufficient alone. In contrast, combining SFT with data balancing enables ReviewRL to produce ratings across the full spectrum, exhibiting stronger discrimination and alignment with ground-truth annotations.

Reward Shaping We conduct an ablation study where only the rule-based reward is used during RL training. As shown in Figure 1, removing the GenRM reward leads to no significant improvement over the SFT baseline across model-based evaluation metrics, with slightly lower actionable insights. This highlights the critical role of GenRM in guiding the policy model to generate high-quality reviews with sufficient details and reliable reasoning. The result aligns with recent findings that emphasize the importance of GenRM or judge models for learning in RL settings involving non-verifiable tasks.

5.4 Context Retrieval

We evaluate the context retrieval module from two perspectives: the quality of the retrieved context c and its impact on review generation.

Quality of Retrieved Context For each generated query x, we compare two responses: an ArXiv-MCP retrieval-augmented answer c and a vanilla answer c_0 generated without external search. Three independent LLM judges assess each pair based on three criteria: (1) *Factual Accu-*
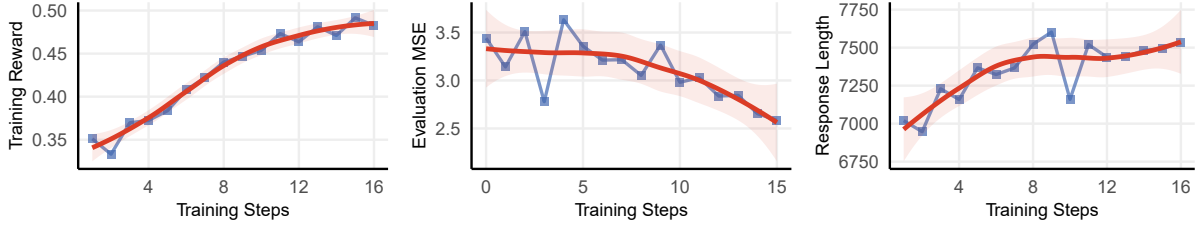
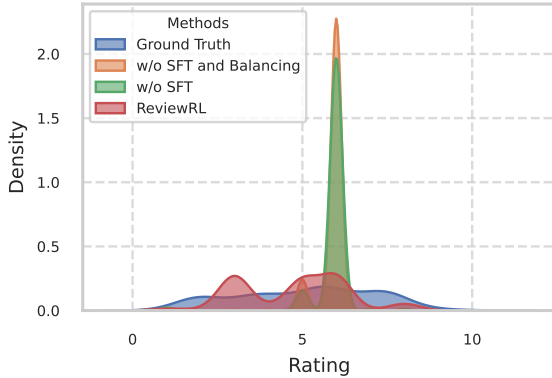Figure 3: Training Dynamics of RL
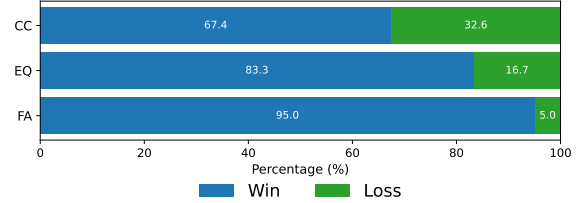

Figure 4: Rating Distributions

Figure 5: Win/Loss percentages of the retrieval answer (Win = blue) versus the non-retrieval answer (Loss = green) across three evaluation dimensions.

age also shrinks, potentially because comparison between the paper and related works may not be effectively conducted without retrieval.

6 Conclusion

In this paper, we introduced ReviewRL, a reinforcement learning framework for automating scientific paper reviews. Our approach integrates context retrieval, supervised fine-tuning, and reinforcement learning to generate high-quality, human-aligned paper reviews with accurate ratings. Experimental results on ICLR 2025 papers demonstrate that ReviewRL significantly outperforms existing methods across both rule-based and model-based evaluation metrics. We established a principled methodology for combining SFT and RL in non-verifiable reasoning tasks, showing that properly initialized policy models can effectively learn from composite rewards without experiencing cold-start issues. Additionally, we demonstrated the critical role of retrieved context in enhancing review factuality and analytical depth, substantiating the effectiveness of our ArXiv-MCP retrieval pipeline.

Limitations

Our framework relies on the accessibility and comprehensiveness of ArXiv as the primary knowledge source, which may provide insufficient context for papers exploring emerging research directions or highly specialized domains with limited representation in the repository. Additionally, although our

racy—correctness and alignment with real-world facts, (2) *Evidence Quality*—sufficiency and relevance of supporting evidence, and (3) *Clarity & Coherence*—readability, organization, and logical flow. We report the win rate where the retrieval-augmented answer is better than the vanilla answer.

As shown in Figure 5, retrieval-augmented responses outperform vanilla responses across all criteria. In terms of factual accuracy, 95.0% of comparisons favor the retrieval variant, indicating a substantial reduction in hallucinations. Evidence quality shows an 83.3% win rate, suggesting effective integration of retrieved citations. While the gain in clarity and coherence is smaller (67.4%), retrieval-augmented responses are still preferred in the majority of cases, implying that additional evidence does not hinder readability. Overall, retrieval consistently enhances context quality, with the most pronounced effect on factual accuracy.

Impact on Review Generation We conduct inference on ReviewRL under the setting where no retrieved context is provided as input (ReviewRL w/o Retrieval). As shown in Figure 1, without retrieval, we observe performance degradation across all metrics, especially for the factualness metrics including correctness of claims and absence of hallucinations. The analytical depth and topic cover-

composite reward function effectively balances rating consistency and review quality, it remains challenging to fully capture the nuanced aspects of human peer review that extend beyond our seven evaluation dimensions. Domain-specific criteria and conference-specific review expectations, which often involve implicit knowledge and norms within academic communities, may not be adequately represented in our current reward formulation, potentially limiting ReviewRL's adaptation to specialized venues or interdisciplinary research areas.

Ethical Considerations

To ensure the ethical development and use of the ReviewRL system, a multifaceted approach has been implemented. During training, we have carefully curated training data and designed the system's reward function to prioritize factual accuracy, analytical depth, and rating consistency, thereby reducing unintended biases and risks. Crucially, ReviewRL is intended to support, not replace, human reviewers, with its outputs serving as drafts for expert evaluation and refinement. For transparency, we will open-source the system, accompanied by detailed documentation on its architecture and training. We will require the users to disclose their affiliation and intended use, fostering accountability and a feedback mechanism for continuous improvement. Furthermore, the system's context retrieval component is designed to maximize coverage and minimize citation bias, and its evaluation metrics consider diverse aspects of review quality beyond simple accuracy, aiming to harness ReviewRL's benefits while proactively mitigating potential harms in the peer review process.

Acknowledgements

References

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. Judgelrm: Large reasoning models as a judge. *arXiv preprint arXiv:2504.00050*.

Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2022. Nlpeer: A unified resource for the computational study of peer review. *arXiv preprint arXiv:2211.06651*.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. Reviewer2: Optimizing review generation through prompt generation. *arXiv preprint arXiv:2402.10886*.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, and 1 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Brendan Hogan. 2024. Debate framework for language model training.

Mohammad Hosseini and Serge PJM Horbach. 2023. Fighting reviewer fatigue or amplifying bias? considerations and recommendations for use of chatgpt and other large language models in scholarly peer review. *Research integrity and peer review*, 8(1):4.

Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. 2025. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*.

Jian Hu. 2025. Reinforce++: A simple and efficient approach for aligning large language models. *arXiv preprint arXiv:2501.03262*.

Bowen Jin, Hansi Zeng, Zhenrui Yue, Jinsung Yoon, Sercan Arik, Dong Wang, Hamed Zamani, and Jiawei Han. 2025. Search-r1: Training llms to reason and leverage search engines with reinforcement learning. *arXiv preprint arXiv:2503.09516*.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine Van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (peerread): Collection, insights and nlp applications. *arXiv preprint arXiv:1804.09635*.

Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025. Position: The ai conference peer review crisis demands author feedback and reviewer rewards. *arXiv preprint arXiv:2505.04966*.

Chhavi Kirtani, Madhav Krishan Garg, Tejash Prasad, Tanmay Singhal, Murari Mandal, and Dhruv Kumar. 2025. Reviewval: An evaluation framework for ai-generated reviews. *arXiv preprint arXiv:2502.11736*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474.

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. 2025. Webthinker: Empowering large reasoning models with deep research capability. *arXiv preprint arXiv:2504.21776*.

Ziyu Liu, Zeyi Sun, Yuhang Zang, Xiaoyi Dong, Yuhang Cao, Haodong Duan, Dahua Lin, and Jiaqi Wang. 2025. Visual-rft: Visual reinforcement fine-tuning. *arXiv preprint arXiv:2503.01785*.

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. *arXiv preprint arXiv:2408.06292*.

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.

Biqing Qi, Kaiyan Zhang, Kai Tian, Haoxiang Li, Zhang-Ren Chen, Sihang Zeng, Ermo Hua, Hu Jinfang, and Bowen Zhou. 2024. Large language models as biomedical hypothesis generators: A comprehensive evaluation. *Preprint*, arXiv:2407.08940.

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*.

ByteDance Seed, Yufeng Yuan, Yu Yue, Mingxuan Wang, Xiaochen Zuo, Jiaze Chen, Lin Yan, Wenyuan Xu, Chi Zhang, Xin Liu, and 1 others. 2025. Seed-thinking-v1. 5: Advancing superb reasoning models with reinforcement learning. *arXiv preprint arXiv:2504.13914*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Automatically evaluating the paper reviewing capability of large language models. *arXiv preprint arXiv:2502.17086*.

Richard S Sutton, Andrew G Barto, and 1 others. 1998. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge.

Kimi Team, Angang Du, Bofei Gao, Bowei Xing, Changjiu Jiang, Cheng Chen, Cheng Li, Chenjun Xiao, Chenzhuang Du, Chonghua Liao, and 1 others. 2025. Kimi k1. 5: Scaling reinforcement learning with llms. *arXiv preprint arXiv:2501.12599*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

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.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2024. Cycleresearcher: Improving automated research via automated review. *arXiv preprint arXiv:2411.00816*.

Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. 2025. J1: Incentivizing thinking in llm-as-a-judge via reinforcement learning. *arXiv preprint arXiv:2505.10320*.

Zeyue Xue, Jie Wu, Yu Gao, Fangyuan Kong, Lingting Zhu, Mengzhao Chen, Zhiheng Liu, Wei Liu, Qiushan Guo, Weilin Huang, and 1 others. 2025. Dancegrpo: Unleashing grpo on visual generation. *arXiv preprint arXiv:2505.07818*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sungduk Yu, Man Luo, Avinash Madusu, Vasudev Lal, and Phillip Howard. 2025. Is your paper being reviewed by an llm? a new benchmark dataset and approach for detecting ai text in peer review. *arXiv preprint arXiv:2502.19614*.

Lunjun Zhang, Arian Hosseini, Hritik Bansal, Mehran Kazemi, Aviral Kumar, and Rishabh Agarwal. 2024. Generative verifiers: Reward modeling as next-token prediction. *arXiv preprint arXiv:2408.15240*.

Tianmai M. Zhang and Neil F. Abernethy. 2025. Reviewing scientific papers for critical problems with reasoning llms: Baseline approaches and automatic evaluation. *Preprint*, arXiv:2505.23824.

Yuxiang Zhang, Yuqi Yang, Jiangming Shu, Xinyan Wen, and Jitao Sang. 2025. Agent models: Internalizing chain-of-action generation into reasoning models. *arXiv preprint arXiv:2503.06580*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. 2025. Deepreview: Improving llm-based paper review with human-like deep thinking process. *arXiv preprint arXiv:2503.08569*.

A   Prompts

The prompts for both the Generation, Evaluation, and GenRM are presented in Tables 6, 7, 8, 9 and 10.

B   Pairwise Metrics

$$P_{\text{relation}} = \begin{cases} 1.0, & \text{if } \operatorname{sgn}(s_1 - s_2) = \operatorname{sgn}(s_1^* - s_2^*), \\ 0, & \text{otherwise.} \end{cases} \quad (6)$$

$$P_{\text{absolute}} = \begin{cases} 1.0, & \text{if } |s_1 - s_1^*| + |s_2 - s_2^*| = 0, \\ 0.6, & |s_1 - s_1^*| + |s_2 - s_2^*| \leq 2, \\ 0, & \text{otherwise.} \end{cases} \quad (7)$$

$$P_{\text{confidence}} = \begin{cases} 1.0, & |s_1 - s_2| \geq |s_1^* - s_2^*|, \\ 0, & \text{otherwise.} \end{cases} \quad (8)$$

In this formula, s_1 and s_2 represent the model's output review ratings, while s_1^* and s_2^* represent the corresponding ground truth values. **Relation** assesses *directional consistency* with human reviewers. **Absolute** measures the *score proximity* to human reviewers. **Confidence** examines differences in *discriminative strength*.

Llama-3.1-8B-Instruct	MSE ↓
SFT	3.01
Reinforce++	2.80

Table 3: MSE of *Llama-3.1-8B-Instruct* after SFT vs. Reinforce++ training (lower is better)

Qwen2.5-7B-Instruct	MSE ↓
SFT	2.83
Reinforce++ (ReviewRL)	2.59
PPO	2.69
GRPO	2.63

Table 4: Comparison of RL algorithms on the *Qwen* backbone (lower MSE is better)

C   Training Settings

C.1   RL Training

DeepSpeed ZeRO-3 and Ray were employed for distributed reinforcement learning on dual $8 \times$A800 GPU clusters. Configuration: micro-batch size of 1, global batch size of 128, and 8-sample rollouts per prompt. Reference and actor models were colocated, with 6 GPUs allocated to the vLLM Engine and 2 GPUs to the GenRM. The composite reward used a balancing coefficient $\gamma = 0.5$. Training completed in 15 optimization steps over 48 hours.

C.2   SFT Training

Supervised fine-tuning utilized DeepSpeed ZeRO-3 with a batch size of 8 and learning rate of 5e-6.

C.3   Model-based Evaluation

Table 5 presents the quantitative results from Figure 1 to enable a precise comparison of system performance. ReviewRL achieves the highest score across all evaluated dimensions, with particularly significant gains in analytical depth and factuality. We observe a strong positive correlation across all dimensions, indicating that systems excelling in one metric tend to excel in others. This finding suggests that generating high-quality scientific reviews is a multifaceted task that requires a comprehensive set of integrated capabilities rather than proficiency in isolated skills.

C.4   Comparison across RL Algorithms

To explore the robustness of REVIEWRL to alternative RL algorithms, we train the model with **PPO** and **GRPO** in addition to our default **Reinforce++**. As reported in Table 4, RL models consistently outperform the SFT baseline under all three algorithms, validating the robustness of our composite reward.

C.5   Model Generality across Architectures

To assess architectural generality, we apply the same training recipe on **Llama-3.1-8B-Instruct**. Table 3 shows that the RL model again reduces MSE relative to its SFT counterpart, mirroring the improvements observed for the Qwen backbone. These findings confirm that the REVIEWRL training recipe generalizes across model families.

	Topic Cov.	Sem.Sim.	Cor. of Claims	Abs. of Hal.	Ana. Depth	Act.Ins.	Adh. to Guide.
DeepReviewer	3.94	3.83	3.92	4.03	3.80	3.70	3.94
CycleReviewer	3.74	3.67	3.72	3.87	3.00	2.86	3.73
ReviewRL	**4.36**	**4.16**	**4.52**	**4.62**	**4.18**	**4.12**	**4.37**
ReviewRL (w/o RL)	4.07	4.01	4.18	4.15	3.99	3.97	4.08
ReviewRL (w/o Retrieval)	4.12	4.07	4.35	4.35	4.04	4.03	4.15
ReviewRL(w/o GenRM)	4.05	3.99	4.17	4.18	3.96	3.90	4.06

Table 5: Model-based evaluation scores on seven quality dimensions for baselines, ablation variants, and our proposed REVIEWRL system (higher is better). This table contains the same quantitative results visualised in Figure 1.

GENERATE QUERIES PROMPT

You are now an academic paper review expert capable of conducting thorough analyses of research papers to provide the most reliable review results. You are now allowed to use the search tool to obtain background information on the paper—please provide three different questions. I will assist you with the search. Please present the three questions in the following format:
1.xxx
2.xxx
3.xxx
Do not include any additional content.

Here is a research paper:
{paper}

Table 6: Prompt for Generate queries prompt

RETRIEVAL SYSTEM PROMPT

You are an academic expert who specializes in answering questions by retrieving information from arXiv.

Table 7: Retrieval system prompt

OPEN SOURCE MODEL EVALUATION PROMPT

Here is a research paper:
{paper}
You are a senior reviewer for top-tier AI conferences (NeurIPS/ICML/CVPR/ACL).
You must be strict and professional enough.
Read the Paper Carefully:
    Analyze each paragraph of each section critically.
    Identify any logical flaws, technical inconsistencies, missing citations, or unclear explanations.
Detailed Paragraph-by-Paragraph Review:
    Provide a detailed critique of each paragraph in every section.
    If a paragraph contains multiple issues, list them separately.
    Highlight strengths, but be critical of weaknesses.
    Use <think> </think> tags to document your detailed thought process during the review.
Comprehensive Structured Review:
After the detailed paragraph-by-paragraph critique, provide a structured review using the following format:
    ## Summary
    (3–5 sentences: core contribution + methodology)
    ## Strengths
    - Bullet points focusing on: Technical merit | Novelty | Empirical validation
    ## Weaknesses
    - Bullet points labeled [Major] or [Minor]: Methodology flaws | Experimental issues | Presentation problems
    ## Rating
    One integer from: [1, 3, 5, 6, 8, 10] (10=Strong Accept; 8=Accept; 6=Borderline Accept; 5=Borderline Reject; 3=Reject; 1=Strong Reject)

Table 8: Open source model evaluation prompt

RETRIEVAL EFFECTIVENESS EVALUATION PROMPT

Factual Accuracy:
You are an extremely meticulous domain expert.

Task: Compare Answer-A (which uses retrieval) with Answer-B (which does not) **only on factual accuracy / faithfulness**.

Scoring rule
- If Answer-A is fully correct or clearly more accurate than Answer-B → output 0
- If Answer-B is clearly more accurate → output 1
- If both are equally correct but Answer-A supplies extra verifiable details, still treat Answer-A as better → output 0

Output format: a single character 0 or 1—nothing else.

Evidence Quality:
You are an academic reviewer. Judge the two answers solely on the quality and usefulness of their evidence or citations.

Decision rule
0 = Answer-A provides stronger or clearer evidence / citations.
1 = Answer-B provides stronger or clearer evidence / citations.
If both contain little or equivalent evidence, but Answer-A supplies extra verifiable details → output 0

Return **only** the single digit 0 or 1. Any extra text is invalid.

Clarity & Coherence:
You are a senior instructor. Evaluate which answer demonstrates better clarity and coherence.

Consider
- Is the writing easy to follow and well-organized?
- Are ideas presented in a logical order with smooth transitions?
- Is terminology explained and jargon minimized?
- Does the answer avoid unnecessary repetition or ambiguity?

If Answer-A is better clear/coherent than Answer-B → output 0; otherwise output 1.

Output must be exactly one character: 0 or 1.

Table 9: Retrieval effectiveness evaluation prompt

GENRM PROMPT

You are an expert academic peer reviewer. You will be shown the abstract/content of a research paper and two peer reviews for that paper. Your task is to determine which peer review is of higher quality based on the following criteria:

1. Factual Accuracy & Soundness: Does the review accurately understand the paper's contributions and limitations? Is the critique based on sound reasoning?
2. Completeness & Coverage: Does the review address the core aspects of the paper (e.g., methodology, results, significance)?
3. Level of Detail & Specificity: Does the review provide specific examples and detailed comments rather than vague statements?
4. Comparison with Existing Work: Does the review appropriately contextualize the paper within the existing literature and compare it to relevant methods?
5. Constructiveness: Is the feedback helpful for the authors to improve the paper? Is the tone professional and constructive?
6. Clarity & Organization: Is the review well-structured and easy to understand?

Paper Context (Abstract/Content): {paper_context}
Review 1: {review1}
Review 2: {review2}

Which peer review is of higher quality based on the criteria above? Respond with **EXACTLY** one of these options:
- REVIEW_1_BETTER
- REVIEW_2_BETTER

YOU MUST CHOOSE A BETTER REVIEW. A TIE IS NOT ALLOWED.

Table 10: GenRM prompt.