# ReviewRL：迈向使用强化学习实现科学审查自动化

**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

同行评审对于科学进步至关重要，但由于投稿量的增加和审稿人疲劳而面临越来越大的挑战。现有的自动化审稿方法在事实准确性、评分一致性和分析深度方面表现不佳，往往生成的反馈肤浅或通用，缺乏高质量人类评审所特有的洞察力。我们介绍了 ReviewRL，一个用于生成全面且事实扎实的科学论文评论的强化学习框架。我们的方法结合了：(1) 一个增添 ArXiv-MCP 检索的上下文生成管道，融入相关的科学文献，(2) 建立基础评审能力的监督微调，以及 (3) 通过复合奖励函数的强化学习过程，联合提升评论质量和评分准确性。在 ICLR 2025 论文上的实验表明，ReviewRL 在基于规则的指标和基于模型的质量评估方面都显著优于现有方法。ReviewRL 为基于 RL 的科学发现中的自动评论生成奠定了基础框架，展示了在此领域未来发展的巨大潜力。ReviewRL 的实现将在 GitHub 上发布。

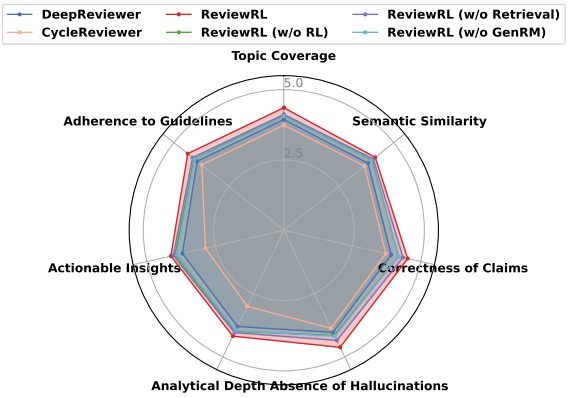


Figure 1: 在 ReviewEval (Kirtani et al., 2025) 标准下 ReviewRL 的评估结果

## 1 介绍

同行评审对科学进步至关重要，确保已发表的研究达到严格的质量、有效性和重要性标准。然而，递交到学术会议和期刊的稿件数量不断增加，对评审系统造成了不可持续的压力，导致评审员疲惫不堪、评估不一致以及评审周期越来越长 (Hosseini and Horbach, 2023; Kim et al., 2025) 。例如，顶级 AI 会议如 NeurIPS 和 ICLR 现在每年处理数千份提交，需要数万次评审 (Kim et al., 2025) 。这种科学产出的激增加剧了对自动化工具的需求，以辅助或增强同行评审过程。

最近在大型语言模型（LLMs）方面的进展为人工智能辅助的科学评估创造了有前景的机会。这些模型可以分析复杂的科学文本，识别方法上的优点和缺点，并大规模生成结构化反馈 (Weng et al., 2024; Lu et al., 2024; Zhu et al., 2025; Qi et al., 2024) 。然而，现有的自动化论文审查方法面临三个显著挑战。首先，它们常常难以保持事实的准确性，并提供基于证据的批评，使论文与相关的前期工作相联系 (Zhou et al., 2024) 。其次，它们倾向于高估论文质量，给出的评级与人类判断不一致 (Yu et al., 2025) 。第三，它们往往生成浅显或通用的评论，缺乏人类评论所特有的分析深度和可操作的见解 (Shin et al., 2025) 。

最近的研究表明，强化学习（RL）在增强大型语言模型（LLM）的推理能力方面具有有效性。像 DeepSeek-R1 (Guo et al., 2025) 这样的模型通过精心设计的 RL 训练方案实现了令人印象深刻的性能提升，而诸如群体相对策略优化 (Shao et al., 2024) 和 Reinforce++ (Hu, 2025) 等创新使得 RL 在 LLM 训练中更加高效和稳定。同时，模型上下文协议（MCP）作为一种标准化的通讯框架出现，使得 LLM 能够无缝地与外部知识源 (Hou et al., 2025) 进行交互，促进了准确的信息检索并减少幻觉。通过 RL 增强的推理能力和通过基于 MCP 的检索实现事实基础的结合提供了一个有前景的方法来解决当前自动化审查系统的限制。

在本文中，我们介绍了 ReviewRL，这是一种用于生成全面、事实依据和建设性审查的科学论文评论的强化学习框架。我们的方法结合了三个关键组件：(1) 一个基于 ArXiv-MCP 的

*Equal contribution.
†Corresponding author.

检索增强上下文生成流程，识别并整合相关科学文献以支持实事求是的评估，（2）一个有监督的微调（SFT）阶段，确立基础审查能力和初始评级对齐，以及（3）一个 RL 优化程序，联合提升审查质量和评级准确性。通过这种综合方法，ReviewRL 生成的评论不仅模拟了人类的分析深度，还提供了与人类判断一致的评级。我们的实验表明，ReviewRL 在规则基础指标和基于模型的审查质量评估中均显著优于现有方法。我们进一步通过全面的消融研究检查每个组件的重要性，揭示了检索增强和我们的复合奖励公式对 ReviewRL 优越性能的显著贡献。据我们所知，这是首次成功应用强化学习来同时提高自动化科学同行评审的质量和评级一致性。我们的贡献包括以下几点：

1) 我们介绍了 ReviewRL，一种将强化学习（RL）集成用于自动论文评审生成的新框架。ReviewRL 包含三个关键组件：ArxivMCP、上下文感知的微调和复合奖励的 RL 训练。

2) 不同于以前的方法，例如 DeepSeek-R1，它依赖于基于规则的奖励，我们发现这样的奖励不足以用于评论生成，其中结构连贯性和内容质量至关重要。为了解决这个问题，我们设计了一个综合奖励系统，结合了基于规则的指标和基于评判模型的评估，有效缓解了纯粹基于规则的方法的限制。

3）与之前的工作相比，ReviewRL 在上下文感知、事实一致性和评论深度方面表现优越。该框架代表了基于强化学习驱动的科学发现自动评论生成的初步步骤。

## 2 相关研究

最近的进展探索了使用大语言模型（LLMs）来自动化和增强学术同行评审过程。早期的努力比如 PeerRead (Kang et al., 2018) 和 NLPeer (Dycke et al., 2022)，提供了为评审生成和分析的基础数据集和基准。在这些资源的基础上，诸如 Reviewer2 (Gao et al., 2024) 这样的系统提出了一个两阶段框架，涉及方面提示生成和评审生成，以提高生成评审的具体性和覆盖面。CycleResearcher (Weng et al., 2024) 和 The AI Scientist (Lu et al., 2024) 引入了端到端框架，模拟整个研究生命周期，包括手稿起草和迭代的同行评审，其中他们的评审模块通过监督微调进行训练或通过代理推理操作。更近的 DeepReviewer (Zhu et al., 2025) 通过使用长链思维（CoT）数据进行 SFT 训练，以增强其推理能力。尽管这些进展，确保 LLM 生成的评审的真实性、推理深度和评级一致性仍然面临挑战。

**用于大语言模型的强化学习** 强化学习（RL）(Sutton et al., 1998) 在增强大型语言模型（LLMs）的指令跟随能力方面发挥了关键作用，特别是通过像从人类反馈中强化学习（RLHF）(Ouyang et al., 2022) 这样的方法。RLHF 使基础模型与人类偏好对齐，通常利用诸如近端策略优化（PPO） (Schulman et al., 2017) 或直接偏好优化 (Rafailov et al., 2023) 之类的算法，其中精确的偏好建模至关重要。最近的进展显示了 RL 在提升大型推理模型（LRMs）如 DeepSeek-R1 (Guo et al., 2025) 的推理能力方面的有效性，通过基于规则的奖励机制，以 GRPO (Shao et al., 2024) 为例。与通常用于开放域指令的 RLHF 不同，GRPO 专门设计用于促进延长的思维链（CoT） (Wei et al., 2022) 推理，特别是在数学问题解决情境中。由于其简单性和有效性，GRPO 已成功应用于各种领域，包括视觉理解和生成 (Liu et al., 2025; Team et al., 2025; Xue et al., 2025)、代理搜索和规划 (Li et al., 2025; Jin et al., 2025; Zhang et al., 2025) 等。然而，像 GRPO 这样的 RL 方法在增强评论和批评生成 (Whitehouse et al., 2025) 的潜力仍需更多探索。

## 3 方法论

### 3.1 任务定义

给定一个目标论文 $q$，自动化科学评论任务被定义为生成一个综合评审 $r$，包括论文的优点和缺点，以及一个评分 $s$。为了确保高质量的评审生成，我们将 ReviewRL 的工作流程表述为一个检索增强生成 (RAG) (Lewis et al., 2020) 和一个 LLM 推理过程。这个过程模仿人类评审者的认知步骤——检索相关领域知识，分析上下文中的论文，并进行评价判断。具体而言，检索模型 $R$ 生成三个查询 $x$ 并通过搜索识别一组相关的上下文论文 $c$，表述为 $q \xrightarrow{R} x, c$。基于 LLM 的评审者 $\pi$ 然后推理论文和检索到的上下文以生成中间思考过程 $z$，表示为 $(q, c) \xrightarrow{\pi} z$。最后，LLM 基于论文、检索到的上下文和推理轨迹生成评审和评分，即 $(q, c, z) \xrightarrow{\pi} (r, s)$。

在接下来的部分中，我们展示了如图 2 所示的 ReviewRL 的组成部分。我们首先介绍 RAG 流程（第 3.2 节），该流程可以为目标论文准确识别上下文相关的文献。随后，我们描述了 ReviewRL 的训练策略，它结合了 SFT（第 3.3 节）和 RL（第 3.4 节）以增强推理能力。

### 3.2 上下文检索

对于每篇论文 $q$，我们使用检索管道 $R$ 检索相关的上下文信息 $c$。参考 DeepReviewer (Zhu

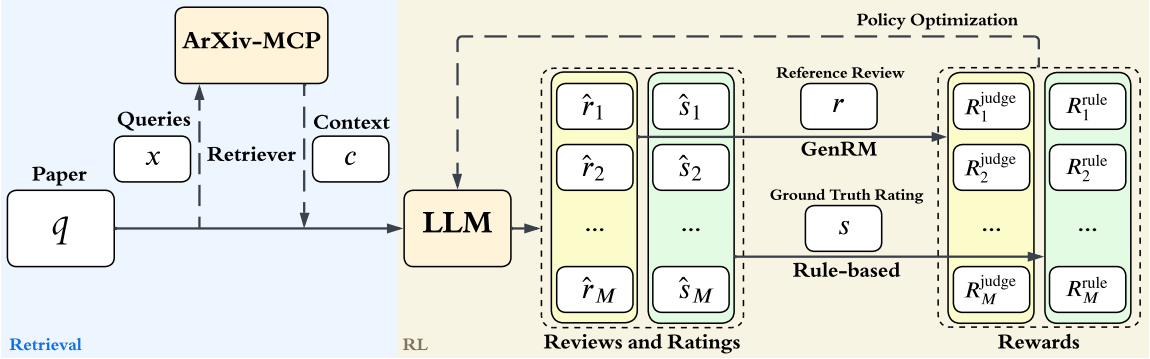


Figure 2: ReviewRL 概述，包括 Arxiv-MCP、SFT 预热和 RL 优化。

et al., 2025) 中的新颖性验证，采用两步方法，我们首先生成查询问题 $x$，然后从 ArXiv 检索相关的上下文 $c$。该方法利用了 LLM 代理化工作流，并整合了模型上下文协议（MCP）以实现高效的上下文检索和生成。

具体来说，我们采用 Qwen3-8B (Yang et al., 2025) 来分析目标论文 $q$，并生成三个查询问题 $x$，这些问题探讨论文的新颖性、方法论以及与前人工作的关系。这些查询被表述为自然语言问题而不是关键词，从而能够更细致地检索相关文献。例如，一个查询可能会问“哪些近期的论文提出了基于 LLM 的论文审查的强化学习方法?”而不是仅仅搜索“强化学习 LLM 审查”。

我们通过 ArXiv-MCP [1] 实现检索功能，这是一个开源的模型上下文协议服务器，为大型语言模型提供对 arXiv 存储库的标准化访问。ArXiv-MCP 能够实现高效的论文搜索、筛选和全文检索，而无需底层 API 的实现。该服务器处理生成的查询，返回包括论文元数据、摘要和相关全文摘录在内的结构化信息。

检索执行由 Qwen-Agent [2] 编排，这是一个基于 Qwen 模型家族构建的代理框架，提供函数调用和工具编排能力。Qwen-Agent 按顺序将每个查询路由到 ArXiv-MCP 并管理检索结果。检索到的上下文 $c$ 经过后处理以去除工具调用的痕迹，并被整合成一个连贯的格式，包括：(1) 查询-响应对，(2) 检索到的论文的书目信息，以及 (3) 这些论文中按相关性排序的摘录。然后，这个处理过的上下文与原始论文表示结合，形成用于生成评论模型的输入。

[1] https://github.com/blazickjp/arxiv-mcp-server

[2] https://github.com/QwenLM/Qwen-Agent

### 3.3 监督微调

给定一篇论文 $q$ 及其检索到的上下文 $c$，下一步是通过策略模型 $\pi$ 生成评论 $r$ 和相应的评分 $s$。虽然我们的目标是使用强化学习（RL）来增强这个过程，但在基础模型上直接应用 RL 存在挑战。这些模型通常会高估论文质量，相比于人类评审者 (Yu et al., 2025)，导致轨迹信息不清晰，奖励信号较弱以及 RL 训练不稳定。根据经验，我们观察到没有适当的初始化，RL 训练表现出冷启动问题，其特征是训练崩溃和性能下降（例如，生成的评分集中在 6 附近）。

为了缓解这个问题，我们首先在长段 CoT 数据上应用 SFT，以具备基本的评论撰写能力来初始化 RL 策略。这一策略的灵感来源于 DeepSeek-R1 (Guo et al., 2025) 和 Kimi-k1.5 (Team et al., 2025) 中的类似实践。我们利用 DeepReview-13k (Zhu et al., 2025) 中的数据，这是一套包含长段 CoT 评论和准确评分注释的高质量数据集，作为训练模型的冷启动数据。具体来说，我们使用数据集中的 ICLR 2024 部分，并对其进行预处理以适应我们的任务定义。我们在输入中包含他们的新颖性验证结果和最佳模式的相应查询，并将最终的元评论作为输出，中间的分析被视为长段 CoT 的思维过程。我们在 Qwen2.5-7B-Instruct (Team, 2024) 模型之上训练 2 个 epoch。与没有 RL 的以往工作不同，在我们的框架中，SFT 起到了两个主要作用：（1）为策略模型提供基本的推理能力，以执行有结构和有理据的同行评审；（2）使预测分数与人类评分对齐，从而稳定下游 RL 训练，防止早期崩溃。

### 3.4 强化学习

在 SFT 阶段之后，我们进行大规模的强化学习（RL），以进一步增强 LLM 审稿人的推理能力。论文审阅是一个不可验证的问题，具有部分可验证的结果——即数值评分，其中审阅质量和

评分与人类判断的一致性都至关重要。之前的工作已证明基于规则的结果奖励在改善 LLM 推理能力方面的有效性 (Guo et al., 2025) 。然而，我们的实验表明，仅依赖评分一致性奖励会导致审阅过于泛泛，缺乏分析深度和可操作的见解，这表明推理能力不足。

为了共同优化评论质量和评分一致性，我们设计了一种复合奖励机制，将基于规则的奖励与生成奖励模型（GenRM）(Zhang et al., 2024) 相结合，该模型优先考虑评分一致性高、格式遵从性好以及具有强分析深度的评论。

我们定义了两个基于规则的奖励组件：评分一致性奖励和格式奖励。评分一致性奖励 $R_{rc}$ 是使用高斯核计算的：

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

其中 $s$ 表示真实评分，通过对给定论文的人为分配分数进行平均得到，$\hat{s}$ 是模型预测的评分。格式遵从奖励 $R_f$ 会对遗漏关键结构组成部分的输出进行惩罚。设 $\mathcal{S}$ 为所需元素集，包括一个推理块（由 <think> 和 </think> 界定）、总结、优点和缺点：

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

最终的基于规则的奖励由以下公式给出：

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

其中，$\alpha$ 和 $\beta$ 是超参数，用于平衡评分一致性和格式完整性的重要性。此奖励公式鼓励生成的输出既符合人类评分又结构良好。

根据之前的工作 (Seed et al., 2025; Hogan, 2024) ，我们使用一个 GenRM $\pi_{\text{judge}}$ 来评价由 LLM 生成的评论 $\hat{r}$ 相对于参考 $r$ 的质量。奖励来源于获胜率，这基于一种共识，即 LLM 作为裁判可以可靠地评估相对的响应质量 (Zheng et al., 2023) 。在我们的框架中，$\pi_{\text{judge}}$ 在多个维度上评价评论：事实准确性、完整性、细节层次、与相关工作的比较、建设性和清晰度。GenRM 奖励 $R^{\text{judge}}$ 定义为：

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

最终的奖励信号是基于规则的奖励和 GenRM 奖励的加权组合：

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

我们使用来自顶级机器学习会议（例如 ICLR 和 ACL）的论文构建 RL 训练数据集，该数据集来源于 Reviewer2 的原始数据以及 DeepReview-13k 的最佳模式拆分。对于每篇论文 $q$ ，我们使用第 3.2 节描述的方法检索上下文 $c$ 。不同会议的评分被规范化为 1–10 的通用尺度，真实评分 $s$ 计算为多个人工审稿人评分的平均值。参考评论 $r$ 的生成方法如下：对于 Reviewer2 的每篇论文，我们使用 DeepSeek-R1-Distill-Qwen-32B 将多个人工评论总结为单个格式化评论；对于 DeepReview-13k，我们使用最佳模式拆分中的元评论。ICLR 2025 的数据被排除以避免数据泄露。数据集统计在表 1 中报告。由于大比例的真实评分落在 5 到 6 之间，我们应用了一种平衡预处理步骤，减少评分在中等范围（5-6）的论文，增加那些评分更极端的论文。此策略强调了具有高度积极或消极评估的论文，这些论文通常对学习更有价值，并有助于防止 RL 模型趋向于中等范围内的通用、非区分性评分。

| | 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: 强化学习训练数据统计

因此，RL 训练数据包含 $(q, c, s, r)$ 元组，而无法获取导致评论 $r$ 和评分 $s$ 的中间推理步骤。该设置鼓励策略模型探索其自身的推理轨迹，以生成与人类判断一致的高质量评论和评分。

策略模型 $\pi$ 从监督微调后的模型 $\pi_{\text{sft}}$ 初始化，以确保稳定学习并防止冷启动崩溃。我们采用 Reinforce++ 算法 (Hu, 2025) 。$\pi_{\text{judge}}$ 是一个 Qwen2.5-14B-Instruct 模型。训练细节显示在附录 C 中。

## 4 实验

我们通过从 ICLR 2025 审稿集里抽取 472 篇论文来构建评估集。对于每篇论文，真实评分被计算为所有人类审稿人给出的分数的平均值。为了确保在完整评分范围内的公平评估，我们抽取论文使得平均评分的分布在评分尺度上大致均匀。

### 4.1 评估指标

我们采用两类定量指标：(i) 基于规则的指标，和 (ii) 基于模型的指标。

我们使用评分级别和成对级别的指标来评估模型。对于评分预测，我们计算预测分数与真实评分之间的均方误差（MSE）和斯皮尔曼等级相关系数。在成对论文评估中，按照

JudgeLRM (Chen et al., 2025) 的方式，我们使用三个成对指标来评估模型对论文质量排序的能力：关系、绝对值和置信度。这些指标分别测量与人工排序的一致性、评分与真实值的接近程度以及区分不同质量论文的辨别强度。我们还报告了一致性指数作为全局排名指标。成对指标的正式定义在附录 B 中提供。

#### 4.1.1 基于模型的定量指标

虽然基于规则的指标注重生成评分的准确性，但评估生成的评论是否模仿人类撰写的评论并提供建设性、富含内容的反馈同样重要。为此，我们采用一个基于 LLM 的评审框架 (Zheng et al., 2023) , 该框架受 ReviewEval 基准 (Kirtani et al., 2025) 的启发，从七个维度评估评论质量。每个维度的评分范围为 1-5，旨在捕捉与人类一致的同行评审的不同方面：

- 主题覆盖度：AI 生成的评论是否全面涵盖了论文的主要主题和论点？它是否涉及了通常由人类审稿人强调的方面？
- 语义相似性：该评论是否捕捉到一个合理的人类评论的核心批评和建议，即使措辞有所不同？
- 主张正确性：评论中的陈述在论文内容方面是否准确？评论是否避免了对方法、结果或结论的误解或错误表述？
- 没有幻觉：评论是否避免引入论文未支持的信息或论点？
- 分析深度：评审是否展示了对研究的深入参与？这包括评估方法的严谨性，识别逻辑漏洞，解释结果，以及在相关工作中对贡献进行背景化。
- 可操作的见解：该审稿意见是否提供了具体的、建设性的建议来改善论文？这些建议是否实用且表达清晰？
- 遵循指南：该审查是否遵循标准的学术审查标准，如创新性、重要性、方法论上的严谨性、清晰性以及伦理合规性（如适用）？

这个评估框架可以全面评估模型在进行超越表层指标的细致和与人类一致的论文评审能力。Llama-3.3-70B-Instruct 被用作评审模型。

### 4.2 基线

我们针对三类基线进行比较。开源指令模型（例如 Qwen-2.5-Instruct）和开源推理模型（例如 Qwen 3），提供了具有基本指令跟随能力和增强推理能力的基线论文审查性能。此外，还包括在公共同行评审数据集上训练的 SFT 模型，如 CycleReviewer-8B，以突出显示我们通过 RL 增强模型在纯监督方法上实现的性能提升。

## 5 结果

### 5.1 基于规则的评估结果

表格 2 展示了基于规则的评估结果。开源指令模型整体表现最弱，表现出较差的评级准确性和有限的排名能力，即使在较大规模时也是如此。开源推理模型在成对指标上有所改进，特别是在判别强度（Pair-Confidence）方面，但仍然落后于均方误差（MSE）。在同行评审数据集上训练的 SFT 模型在人类评级的对齐度方面表现出了显著的提升——例如，DeepReviewer-7B 实现了最佳的成对关系评分。我们提出的 ReviewRL 模型在各个方面都表现出最强的性能。ReviewRL 与其仅包含 SFT 的对照模型之间的表现差距突出显示了增强学习在优化评分一致性方面的有效性。

### 5.2 基于模型的评估结果

图 1 展示了基于模型的评估结果在七个评论质量维度上的表现。ReviewRL 始终优于所有基准，尤其是在需要更深入推理和可靠性的维度，例如分析深度。与仅监督的对应方法相比，ReviewRL 在所有维度上表现出显著的改进，突出了 RL 在优化评论生成中的优势。这些结果进一步证明了 RL 能够生成更具信息性、真实且富有建设性的评论。

我们分析了 ReviewRL 的 RL 训练动态，以为未来在大型语言模型审稿人模型中的训练方案提供见解。图 3 展示了 ReviewRL 在三个关键指标上的训练动态。随着训练的进行，训练奖励稳步增加，表明策略在有效优化奖励函数。同时，响应长度在早期阶段增长，并在大约第 10 步后稳定下来，这表明模型学习生成更详细的输出。评估 MSE 在训练步骤中持续下降，证实了所学策略对保留数据的泛化能力更好，并产生更准确的评审评分。这些趋势总体上展示了 RL 训练过程的有效性。

我们观察到在没有适当策略初始化的情况下启动 RL 训练时，ReviewRL 存在冷启动问题。如图 4 所示，从头开始的训练导致评分崩溃，模型主要输出约为 6 的通用分数，无法区分输入论文。我们的 RL 训练数据平衡策略，通过加权具有极端真实评分的示例，部分缓解了这个问题，但单独使用仍然不足。相比之下，将 SFT 与数据平衡结合，使 ReviewRL 能够在整个评分范围内生成评分，表现出更强的区分能力和与真实标注的对齐性。我们进行了一项消融研究，仅在强化学习训练期间使用基于规则的奖励。如图 1 所示，移除 GenRM 奖励在基于

Table 2: 基于规则的评估结果。ReviewRL 持续优于基线方法。

| 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 Reasoning | | | | | | |
| 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 |

模型的评估指标中未能带来对 SFT 基线的显著改进，并且可操作的见解略有减少。这强调了 GenRM 在指导策略模型生成具有充分细节和可靠推理的高质量评论时的关键作用。该结果与最近的研究结果一致，这些研究强调了在涉及不可验证任务的强化学习环境中，GenRM 或判评分模型的重要性。

我们从两个角度评价上下文检索模块：检索到的上下文 $c$ 的质量及其对评论生成的影响。

**检索上下文的质量** 对于每个生成的查询 $x$，我们比较两个回答：一个是通过 ArXiv-MCP 检索增强得到的回答 $c$，另一个是没有使用外部搜索生成的普通回答 $c_0$。三个独立的大型语言模型评审员根据三个标准对每对回答进行评估：（1）事实准确性—正确性和与现实世界事实的一致性，（2）证据质量—支持证据的充分性和相关性，以及（3）清晰度和连贯性—可读性、组织性及逻辑流畅性。我们报告检索增强回答优于普通回答的胜率。

如图 5 所示，在所有标准上，检索增强型回答的表现均优于普通回答。在事实准确性方面，95.0 % 的比较结果倾向于检索变体，这表明幻觉显著减少。证据质量显示 83.3 % 的胜率，建议有效地整合了检索到的引用。尽管在清晰度和连贯性方面的提升较小（67.4 % ），在大多数情况下仍然更偏向于检索增强型回答，这意味着额外的证据不会妨碍可读性。总的来说，检索始终提高了语境质量，对事实准确性的影响最为显著。

**对评论生成的影响** 我们在没有检索到的上下文作为输入的情况下，对 ReviewRL 进行推理（ReviewRL 无检索）。如图 1 所示，在没有检索的情况下，我们观察到所有指标上的性能下降，特别是在事实性指标上，包括论点的正确性和幻觉的缺乏。分析深度和主题覆盖范围也有所缩减，这可能是因为在没有检索的情况下，无法有效地进行论文与相关工作的比较。

在这篇论文中，我们介绍了 ReviewRL，一种用于自动化科学论文审阅的强化学习框架。我们的方法整合了上下文检索、监督微调和强化学习，以生成高质量、与人类对齐的论文评论并提供准确的评分。在对 ICLR 2025 论文的实验结果表明，ReviewRL 在基于规则和基于模型的评估指标两方面都显著优于现有方法。我们建立了一个有原则的方法论来结合 SFT 和 RL 在不可验证的推理任务中，展示了适当初始化的策略模型可以有效地从复合奖励中学习而不会遇到冷启动问题。此外，我们展示了所检索的上下文在增强评论的事实性和分析深度方面的关键作用，证实了我们 ArXiv-MCP 检索流程的有效性。

## 6

局限性 我们的框架依赖于 ArXiv 的可访问性和全面性作为主要的知识来源，这可能会为探索新兴研究方向或在该资源库中代表有限的高专业领域的论文提供不充分的背景。此外，尽管我们的综合奖励函数有效平衡了评分一致性和审稿质量，但仍然难以完全捕捉到超出我们

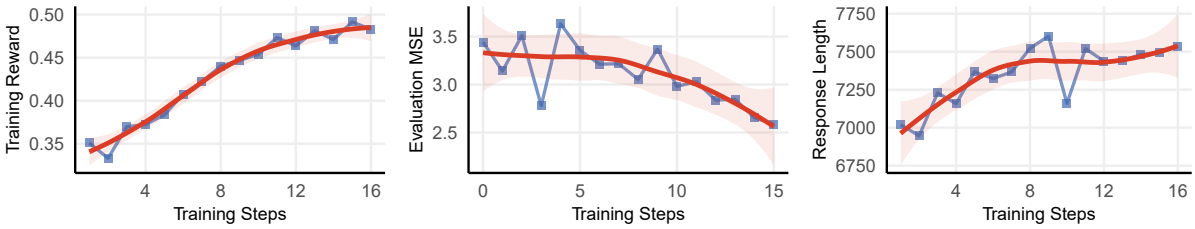


Figure 3: 强化学习的训练动态

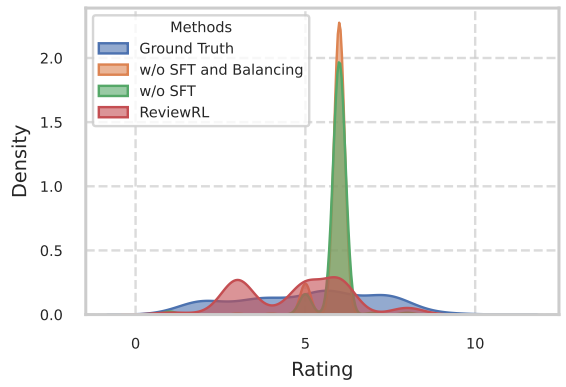


Figure 4: 评分分布

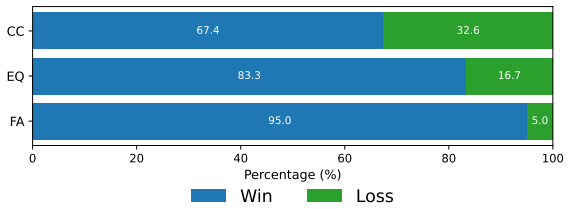


Figure 5: 在三个评估维度上，检索答案（胜 = 蓝色）与非检索答案（负 = 绿色）的胜负百分比。

七个评估维度的人类同行评审的细微方面。特定领域的标准和特定会议的审稿期望，通常涉及学术界内的隐性知识和规范，可能未在我们当前的奖励公式中得到充分体现，这可能会限制 ReviewRL 对专业场所或跨学科研究领域的适应性。

为了确保 ReviewRL 系统的伦理开发和使用，我们实施了多方面的方法。在训练过程中，我们精心选取了训练数据，并设计了系统的奖励函数，以优先考虑事实准确性、分析深度和评价一致性，从而减少无意的偏见和风险。重要的是，ReviewRL 旨在支持而非取代人类评审员，其输出作为专家评估和完善的草稿。为了透明化，我们将开源该系统，并附以详细的架构和训练文档。我们将要求用户披露其所属机构和使用目的，以促进问责制和持续改进的反馈机制。此外，系统的上下文检索组件旨在最大化覆盖率并最小化引用偏差，其评估指标考虑了审查质量的多样性方面，而不仅仅是简单的准确性，旨在利用 ReviewRL 的优势，同时积极缓解同行评审过程中的潜在危害。

## 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. Revieweval: 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.

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 提示

生成、评估和 GenRM 的提示展示在表格 6 、7 、8 、9 和 10 中。

## B 成对指标

$$\mathrm{P}_{\text{relation}} = \begin{cases} 1.0, & \text{if } \operatorname{sgn}(s_1 - s_2) = \operatorname{sgn}(s_1^* - s_2^*), \\ [4pt]0, & \text{otherwise.} \end{cases} \tag{6}$$

$$\mathrm{P}_{\text{absolute}} = \begin{cases} 1.0, & \text{if } |s_1 - s_1^*| + |s_2 - s_2^*| = 0, \\ [4pt]0.6, & |s_1 - s_1^*| + |s_2 - s_2^*| \leq 2, \\ [4pt]0, & \text{otherwise.} \end{cases} \tag{7}$$

$$\mathrm{P}_{\text{confidence}} = \begin{cases} 1.0, & |s_1 - s_2| \geq |s_1^* - s_2^*|, \\ [4pt]0, & \text{otherwise.} \end{cases} \tag{8}$$

在这个公式中，$s_1$ 和 $s_2$ 代表模型的输出评分，而 $s_1^*$ 和 $s_2^*$ 代表对应的真实值。Relation 评价与人类评审一致的方向性。Absolute 测量与人类评审分数的接近程度。Confidence 审查辨别能力的差异。

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

Table 3: 经过 SFT 的 Llama3.18BInstruct 与 Reinforce++ 训练后的 MSE（越低越好）

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

Table 4: 在 Qwen 骨干上的 RL 算法比较（MSE 越低越好）

## C 训练设置

### C.1 强化学习训练

DeepSpeed ZeRO-3 和 Ray 被用于双 8 × A800 GPU 集群上的分布式强化学习。配置：微批大小为 1，全球批大小为 128，每个提示进行 8 样本回合。参考模型和演员模型共存，其中 6 个 GPU 分配给 vLLM 引擎，2 个 GPU 分配给 GenRM。复合奖励使用了平衡系数 $\gamma = 0.5$ 。训练在 48 小时内完成了 15 次优化步骤。

### C.2 SFT 训练

监督微调使用 DeepSpeed ZeRO-3，批大小为 8，学习率为 5e-6。

### C.3 基于模型的评估

表 5 提供了来自图 1 的定量结果，以便精确比较系统性能。ReviewRL 在所有评估维度上都获得了最高分，特别是在分析深度和事实性上取得了显著增益。我们观察到所有维度之间存在强正相关，表明在一个指标上表现优秀的系统往往在其他指标上也表现出色。这个发现表明，生成高质量的科学评论是一项多方面的任务，需要一整套综合能力，而不是单一技能的熟练掌握。

### C.4 RL 算法比较

为了探索 REVIEWRL 对不同 RL 算法的鲁棒性，除了使用默认的 Reinforce++，我们还用 PPO 和 GRPO 训练模型。正如表 4 所报告的，RL 模型在所有三种算法下始终优于 SFT 基线，验证了我们复合奖励的鲁棒性。

### C.5 模型在不同架构间的通用性

为了评估架构的普适性，我们在 Llama3.18BInstruct 上应用相同的训练方案。表 3 显示，RL 模型相对于其 SFT 对

| | 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: 基于模型的评分在七个质量维度上对基线、消融变体以及我们提出的 REVIEWRL 系统进行评价（分数越高越好）。该表包含了与图 1 中呈现的相同的定量结果。

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: 生成查询提示的提示

应模型再次减少了 MSE ，反映了在 Qwen 骨干模型上观察到的改进。这些发现证实，REVIEWRL 训练方案在不同的模型系列中具有普适性。

| RETRIEVAL SYSTEM PROMPT |
|---|
| You are an academic expert who specializes in answering questions by retrieving information from arXiv. |

Table 7: 检索系统提示

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

| RETRIEVAL EFFECTIVENESS EVALUATION PROMPT |
|---|
| Factual Accuracy:<br>You are an extremely meticulous domain expert.<br>Task: Compare Answer-A (which uses retrieval) with Answer-B (which does not) only on factual accuracy / faithfulness .<br>Scoring rule<br>• If Answer-A is fully correct or clearly more accurate than Answer-B → output 0<br>• If Answer-B is clearly more accurate → output 1<br>• If both are equally correct but Answer-A supplies extra verifiable details, still treat Answer-A as better → output 0<br>Output format: a single character 0 or 1—nothing else. |
| Evidence Quality:<br>You are an academic reviewer. Judge the two answers solely on the quality and usefulness of their evidence or citations.<br>Decision rule<br>0 = Answer-A provides stronger or clearer evidence / citations.<br>1 = Answer-B provides stronger or clearer evidence / citations.<br>If both contain little or equivalent evidence, but Answer-A supplies extra verifiable details → output 0<br>Return only the single digit 0 or 1. Any extra text is invalid. |
| Clarity & Coherence:<br>You are a senior instructor. Evaluate which answer demonstrates better clarity and coherence.<br>Consider<br>• Is the writing easy to follow and well-organized?<br>• Are ideas presented in a logical order with smooth transitions?<br>• Is terminology explained and jargon minimized?<br>• Does the answer avoid unnecessary repetition or ambiguity?<br>If Answer-A is better clear/coherent than Answer-B → output 0; otherwise output 1.<br>Output must be exactly one character: 0 or 1. |

Table 9: 检索效果评估提示

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 提示。