DeepReview: Improving LLM-based Paper Review with Human-like Deep Thinking Process

Minjun Zhu[1,2,†], **YixuanWeng**[2,†], **LinyiYang**[3], **Yue Zhang**[2,*]

[1]Zhejiang University [2]School of Engineering, Westlake University [3]University College London
zhuminjun@westlake.edu.cn, wengsyx@gmail.com,
yanglinyiucd@gmail.com, zhangyue@westlake.edu.cn

Abstract

Claim: This work is not advocating LLM replacement of human reviewers but rather exploring LLM assistance in peer review.

Large Language Models (LLMs) are increasingly utilized in scientific research assessment, particularly in automated paper review. However, existing LLM-based review systems face significant challenges, including limited domain expertise, hallucinated reasoning, and a lack of structured evaluation. To address these limitations, we introduce DeepReview, a multi-stage framework designed to emulate expert reviewers by incorporating structured analysis, literature retrieval, and evidence-based argumentation. Using DeepReview-13K, a curated dataset with structured annotations, we train DeepReviewer-14B, which outperforms CycleReviewer-70B with fewer tokens. In its best mode, DeepReviewer-14B achieves win rates of 88.21% and 80.20% against GPT-o1 and DeepSeek-R1 in evaluations. Our work sets a new benchmark for LLM-based paper review, with all resources publicly available at https://github.com/zhu-minjun/Researcher.

1 Introduction

Peer review is the foundation of scientific progress, ensuring that research is novel, reliable, and rigorously evaluated by experts before publication (Alberts et al., 2008). With the increasing volume of research submissions, Large Language Models (LLMs) have become promising tools to support reviewers (Yang et al., 2024; Chris et al., 2024; Li et al., 2024b; Scherbakov et al., 2024; Si et al., 2025). For example, the ICLR 2025 conference has introduced an LLM-based system to assist reviewers in providing feedback (Blog, 2024).

Recent research has explored two primary approaches to improve LLM-based review systems: (1) employing LLM-powered agents to simulate the peer review process, as exemplified by AI-Scientist (Chris et al., 2024) and AgentReview (Jin et al., 2024a); and (2) developing open-source models trained on extensive datasets from existing peer review platforms, such as ReviewMT (Tan et al., 2024a) and CycleReviewer (Weng et al., 2025).

Despite these advancements, current systems exhibit several critical limitations: they struggle to comprehensively identify submission flaws, resulting in superficial feedback (Zhou et al., 2024a); lack evidence-based justifications (Zhuang et al., 2025); and fail to provide clear, actionable suggestions (Ye et al., 2024; Du et al., 2024). Moreover, their vulnerability to prompt engineering leads to inaccurate evaluations (Ye et al., 2024; Weng et al.). While robust feedback is crucial for scientific advancement and peer review integrity, developing reliable evaluation frameworks faces two significant challenges: (1) The scarcity of structured paper review datasets that capture fine-grained expert evaluation processes. Most available open review datasets primarily contain aggregated reviews and decisions, limiting LLMs' ability to learn systematic review reasoning chains and increasing their susceptibility to shortcut learning and adversarial manipulation. (2) LLMs' inherent constraints, including restricted domain knowledge, lack of dynamic knowledge updating mechanisms, and a tendency to generate hallucinated content without adequate verification (Schintler et al., 2023; Drori and Te'eni, 2024; Weng et al., 2024), which significantly impair their capability to assess complex scientific content (Wang et al., 2020; Yuan et al., 2021).

To address these challenges, we introduce **DeepReview**, a structured multi-stage review framework that closely aligns with the expert review process by incorporating novelty assessment, multi-

*Corresponding Author. Supported by Research Center for Industries of the Future, Westlake University.
† These authors contributed equally to this work.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 29330–29355
July 27 - August 1, 2025

dimensional evaluation criteria, and reliability verification. We develop a comprehensive data synthesis pipeline that integrates retrieval and ranking(Asai et al., 2024), self-verification (Weng et al., 2023), and self-reflection (Ji et al., 2023), ensuring the soundness and robustness of LLM-generated suggestions. This approach enables deeper insights into the reasoning and decision-making of paper review. The resulting dataset, **DeepReview-13K**, consists of raw research papers, structured intermediate review steps, and final assessments. Based on that, we train **DeepReviewer-14B**, a model that offers three inference modes – Fast, Standard, and Best – allowing users to balance efficiency and response quality. We further construct **DeepReview-Bench**, a comprehensive benchmark containing 1.2K samples, which evaluates both quantitative aspects (rating prediction, quality ranking, and paper selection) and qualitative review generation through LLM-based assessment.

Extensive experiments demonstrate DeepReviewer 14B's superior performance across multiple dimensions. Compared to existing systems like CycleReviewer-70B, GPT-o1, and Deepseek-R1, our model achieves substantial improvements in Score (Rating MSE: 44.80% ↑), Ranking (Rating Spearman: 6.04% ↑), and Selection (Accuracy 1.80% ↑). In LLM-as-a-judge evaluation (Wang et al., 2024b; Rewina et al., 2025), it achieves a 80% win rate against GPT-o1 and Deepseek-R1. Notably, DeepReviewer exhibits strong resilience to adversarial attacks despite no explicit robustness training. Furthermore, our Test-Time Scaling analysis reveals that DeepReviewer can enhance its performance by adjusting reasoning paths and response lengths.

Our work establishes a foundation for robust LLM-based review systems through DeepReview, a structured framework that addresses fundamental challenges in automated manuscript evaluation. We introduce DeepReview-13K, a dataset featuring fine-grained review reasoning chains, alongside DeepReview-Bench, a benchmark for automated paper review. Built upon these resources, our DeepReviewer-14B model demonstrates substantial improvements over existing approaches while maintaining strong resilience to adversarial attacks, validating the effectiveness of our structured approach to automated scientific evaluation. Our code, model, and data will be publicly available under the agreement of our usage policy.

2 Related Work

Reasoning in LLMs. The emergence of large language models (Achiam et al., 2023; Touvron et al., 2023; Bai et al., 2023) has provided new assistance in advancing solutions to complex Science challenges (Hendrycks et al., 2021; Zhu et al., 2025). Initially, Scratchpads and chain-of-thought (Akyürek et al., 2022; Nye et al., 2021; Wei et al., 2022) encouraged LLMs to think. This technique has been employed in various reasoning tasks. Building on this, a series of works including self-consistency (Wang et al., 2023), self-verification (Weng et al., 2023), and self-reflection (Madaan et al., 2024) prompted language models to output more thinking processes during reasoning. Later, OpenAI's O1 model (Jaech et al., 2024) and various open-source long chain-of-thought models (Guo et al., 2025) achieved Scaling Test-time Compute (Yao et al., 2024; Guan et al., 2025) through additional supervised training or reinforcement learning, enabling language models to select optimal solutions for improved performance (Xiang et al., 2025). While these advances have enhanced reasoning capabilities, they primarily focus on general problem-solving rather than specialized academic review tasks. Our Review-with-Thinking framework extends these reasoning approaches specifically for peer review.

Reliable Scientific Literature Assessment. Recent studies have demonstrated significant progress in automated scientific research. Chris et al. (2024) develop an AI scientist for autonomous hypothesis generation and experimentation (Langley, 1987; Daniil et al., 2023; AI, 2025; Zonglin et al., 2023; Li et al., 2024c; Hu et al., 2024). Multi-agent frameworks (Ghafarollahi and Buehler, 2024; Rasal and Hauer, 2024; Su et al., 2024) enable collaborative scientific reasoning, while Weng et al. (2025) show LLM-based review systems can enhance scientific discovery through reinforcement learning. However, these systems often lack structured reasoning, resulting in unreliable feedback.

Robust LLM-based Paper Review. Recent work spans generation-focused approaches using role-playing agents (D'Arcy et al., 2024; Gao et al., 2024; Yu et al., 2024; Weng et al., 2025), meta-review synthesis (Santu et al., 2024; Li et al., 2023; Zeng et al., 2024), and bias detection mechanisms (Liang et al., 2024; Tyser et al., 2024; Tan et al., 2024b). Hybrid workflows (Jin et al., 2024b; Zyska et al., 2023) combine human-AI collaboration with

iterative refinement. While evaluation benchmarks (Funkquist et al., 2022; Zhou et al., 2024b; Kang et al., 2018) and ethical analyses (Ye et al., 2024; Latona et al., 2024) have advanced the field, existing systems struggle with complex assessments and remain vulnerable to adversarial attacks, highlighting the need for explicit reasoning processes.

3 Data Collection

We present DeepReview-13K, a training dataset that captures the intermediate reasoning processes inherent in academic paper reviews, addressing the fundamental challenges in Paper Review tasks from three dimensions: the scarcity of high-quality, structured review datasets and standardized evaluation frameworks.

3.1 DeepReview-13K

Dataset	Number	Tokens	Rating	Accept Rate
ICLR 2024 Train	4131	10439	5.34	37.8%
ICLR 2025 Train	9247	10062	5.13	31.2%
DeepReview-13K	13378	10178	5.18	33.24%
ICLR 2024 Test	652	10681	5.47	43.7%
ICLR 2025 Test	634	10241	5.18	31.1%
DeepReview-Bench	1286	10464	5.33	37.49%

Table 1: Dataset Statistics. The table shows the average values of Tokens, Rating, and Accept Rate

The statistics of this dataset are detailed in Table 1. We initially collected raw data from the OpenReview platform arXiv repository, gathering 18,976 paper submissions spanning two ICLR conference cycles (2024-2025)[1]. Using the MinerU tool (Wang et al., 2024a), we convert papers to parseable Markdown format, prioritizing LaTeX source code when available from arXiv. For each paper, we assembled a review set $\mathbf{R}$ comprising three key components: (1) textual assessments (Strengths, Weaknesses, and Questions), (2) interactive discussions from the rebuttal phase, and (3) standardized scores, including overall ratings ($\in [1, 10]$) and fine-grained evaluations of Soundness, Presentation, and Contribution ($\in [1, 4]$). Additionally, we collect meta-review texts and final ratings with acceptance decisions. The final DeepReview-13K dataset comprises 13,378 valid samples in Table 1 as the foundation for constructing our review reasoning chain.

3.2 DeepReview-Test

To evaluate performance, we randomly sampled 10% (1.2K) of the dataset to create DeepReview-Bench. Our evaluation framework assesses both quantitative scores and qualitative aspects of review generation through the following tasks:

Quantitative Evaluation: 1) Rating prediction: using MAE, MSE, accuracy, and F1 metrics 2) Paper quality ranking: measured by Spearman correlation 3) Pairwise paper selection (n=2): assessed through accuracy

Qualitative Evaluation: While previous work (Tan et al., 2024a) relied on simple text similarity metrics (e.g., ROUGE (Lin, 2004), BLEU (Papineni et al., 2002)), these metrics fail to capture specific review capabilities. Motivated by recent findings (Li et al., 2024a), we adopt the LLM-as-a-judge paradigm using Gemini-2.0-Flash-Thinking to conduct pairwise comparative evaluations of generated reviews. Detailed evaluation metrics are provided in Appendix B.

4 Methodology

Drawing inspiration from recent advances in complex reasoning methods (Xiang et al., 2025; Hao et al., 2024), we propose a deep-thinking evaluation framework that decomposes the review process into three key steps in Figure 1: (1) novelty verification z_1: assessing research originality through literature review; (2) multi-dimension evaluation z_2: synthesizing insights from multiple expert perspectives; and (3) reliability verification z_3: examining internal consistency and logical coherence.

4.1 Task Definition

Formally, given an input paper $\mathbf{q}$, our goal is to generate a review pair $(\mathbf{s}, \mathbf{a})$, where $\mathbf{s}$ represents the qualitative assessment text (meta-review), we express the reasoning process as:

$$\mathbf{q} \rightarrow z_1 \rightarrow z_2 \rightarrow z_3 \rightarrow (\mathbf{s}, \mathbf{a})$$

We formulate the review score generation as a marginalization over sequential reasoning chains:

$$p(\mathbf{a}|\mathbf{q}) \propto \int p(\mathbf{a}|z_{1:3}, \mathbf{q}) \prod_{t=1}^{3} p(z_t|z_{<t}, \mathbf{q}) d\mathbf{Z} \quad (1)$$

Here, the chain-of-thought term $\prod_{t=1}^{3} p(z_t|z_{<t}, \mathbf{q})$ explicitly models the sequential dependencies between reasoning steps, $\mathbf{Z}$ represents all possible intermediate state sequences $(s_1, \ldots, s_n)$. This structured approach aims to enhance the reliability of the evaluation process.

[1]Empty PDFs were filtered during conversion

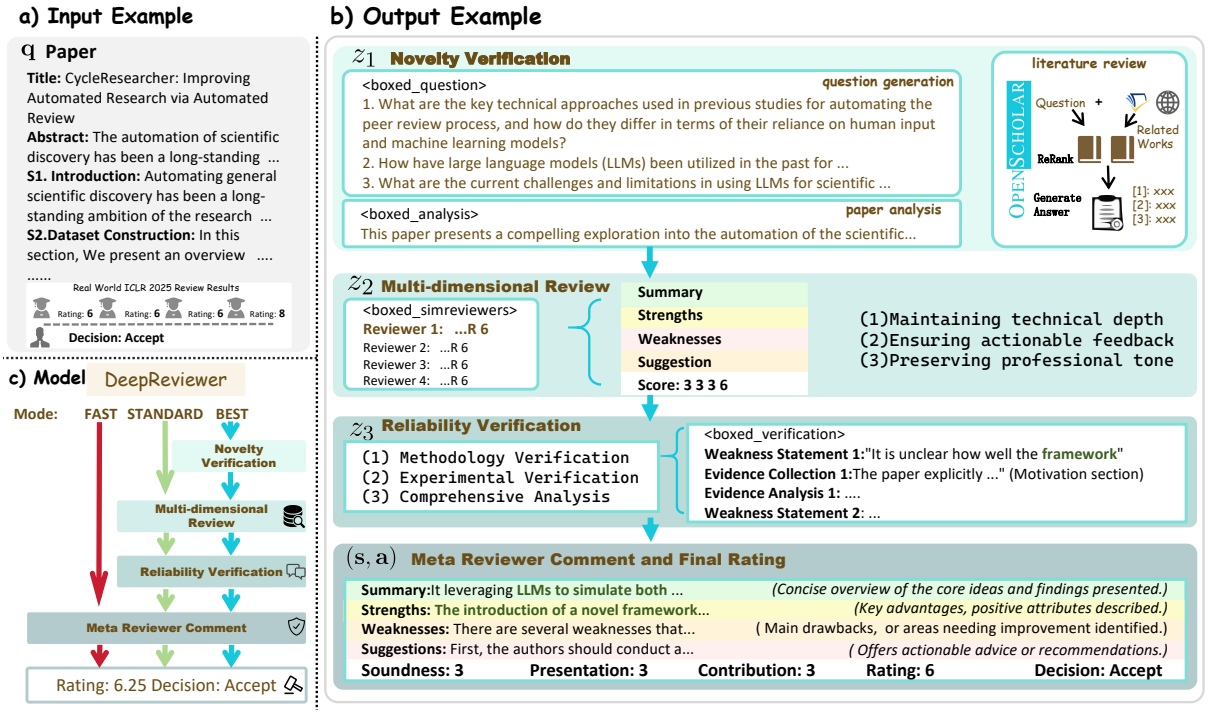

Figure 1: Overview of the DeepReviewer. (a) Input paper example with a real-world research paper. (b) Output example showing DeepReviewer's multi-stage reasoning process: Novelty Verification, Multi-dimension Review, and Reliability Verification. (c) Inference modes: fast, standard, and best, highlighting different reasoning paths. We provide a more detailed case study in the appendix F.

4.2 Structured Reasoning Process

We present a comprehensive automated data construction pipeline, which is specifically designed to generate high-quality supervised fine-tuning datasets that capture complete reasoning paths, shown as (z_1, z_2, z_3).

Stage 1: Novelty Verification (z_1). Our novelty verification framework consists of three key components: **question generation**, **paper analysis**, and **literature review**. Initially, based on the paper, we use the Qwen-2.5-72B-Instruct model (Qwen et al., 2025) to generate three key research questions, focusing on research gaps, innovative directions, and methodological breakthroughs to capture domain-specific characteristics. Additionally, to ensure thorough understanding, we employ the Gemini-2.0-Flash-thinking model to conduct systematic paper analysis with a specifically designed system prompt (Figure 6) across research motivation, core ideas, technical approaches, and experimental design. Then, literature retrieval, comparison, and summary are built on OpenScholar (Asai et al., 2024) to address these research questions. Using Qwen-2.5-3B-Instruct with few-shot learning, we transform questions into search keywords to retrieve approximately 60 relevant papers via Semantic Scholar API. Subsequently, the ReRank model[2] reorder retrieved papers and select the top 10 most relevant papers, and its internal QA model [3] generates comprehensive reports as novelty analysis z_1, incorporating works cited in review R.

Stage 2: Multi-dimension Review (z_2). To provide constructive review, we transform author rebuttals into instructive suggestions while synthesizing multiple review $\mathbf{R}$ into comprehensive perspectives. Specifically, using Qwen-2.5-72B-Instruct, we develop a review reconstruction pipeline that analyzes each review in $\mathbf{R}$ with its corresponding author response, capturing experimental results, theoretical proofs, and implementation details from rebuttals to transform criticisms into concrete technical suggestions. The reconstruction process (z_2) follow three principles: (1) maintaining technical depth; (2) ensuring actionable feedback; (3) preserving professional tone and original citations.

Stage 3: Reliability Verification(z_3). In order to ensure assessment accuracy through systematic evidence analysis, we employ Gemini-2-Flash-thinking to conduct systematic evidence analysis through a four-stage verification chain: *methodology verification*, *experimental verification*, and

[2] https://huggingface.co/OpenSciLM/OpenScholar_Reranker

[3] https://huggingface.co/OpenSciLM/Llama-3.1_OpenScholar-8B

comprehensive analysis. Each review comment requires supporting evidence from the paper and receives an assigned confidence level. Finally, we utilize Qwen to generate a new Meta-Review by integrating the original Meta-Review, reviewer comments, and verification outcomes. This step identifies key weaknesses while providing evidence-based analysis and constructive suggestions.

Quality Control Mechanism. To ensure the high quality of our synthetic DeepReview-13K dataset, we implemented a rigorous automated quality control process using Qwen-2.5-72B-Instruct. This process involves a multi-faceted approach to assess each generated sample for logical integrity and completeness. Specifically, Qwen-2.5-72B-Instruct was tasked with examining each sample for: (1) Logical Consistency: verifying that the reasoning chain (z_1, z_2, z_3) and the final evaluation $(\mathbf{s}, \mathbf{a})$ are logically coherent and non-contradictory; (2) Completeness: checking for any missing or empty fields within the structured data format, ensuring all components of the reasoning path and evaluation are present. Samples failing any of these checks, indicating logical inconsistencies, incompleteness, or failing to meet our quality standards, were automatically flagged and removed from the dataset.

4.3 Model Training

We train our model based on Phi-4 14B (Abdin et al., 2024) using the DeepReview-13K dataset. The training process was conducted on 8x H100 80G GPUs with DeepSpeed + ZeRO3 (Rajbhandari et al., 2020; Rasley et al., 2020) for optimization. Notably, we extended the context window to 256K using LongRoPE (Ding et al., 2024), with a 40K context window during training for full-parameter fine-tuning. Given memory constraints, samples exceeding the preset context length are randomly truncated. The model is trained for 23,500 steps with a batch size of 16 and a learning rate of 5e-6.

Inference Strategy. We divided each sample in the DeepReview-13K data into three modes using reasoning path cropping, as shown in Figure 1(c), which allows for efficiency adjustments at test time based on varying requirements. The *Fast* mode directly generates final evaluation results and comprehensive analysis reports $(\mathbf{s}, \mathbf{a})$, minimizing computational cost by bypassing intermediate reasoning steps. The *Standard* mode executes core evaluation steps including z_2 and z_3, maintaining high efficiency while ensuring evaluation quality, making it appropriate for routine research assessment. The *Best* mode implements the complete reasoning chain (z_1, z_2, z_3), encompassing novelty verification, multi-dimension assessment, reliability verification, and comprehensive analysis generation. For novelty verification during inference, as in Stage 1 (Section 4.2), we employ Semantic Scholar API and OpenScholar to ensure accurate assessment of research novelty and citation correctness through comprehensive literature review and analysis. All three modes share the same model architecture, differing only in their executed evaluation steps. This allows the trained DeepReview-14B model to execute different reasoning paths at inference time, controlled by input instructions.

5 Experiments

5.1 Experimental setting

Baselines. We consider two types of baselines: (1) Prompt-based baselines including AI Scientist (Chris et al., 2024) and AgentReview (Jin et al., 2024a) implemented with various backbone models (GPT-o1-2024-12-17, Claude-3.5-sonnet-20241022, Gemini-2.0-Flash-Thinking-01-21, DeepSeek-V3, and DeepSeek-R1); (2) Fine-tuned baselines including CycleReviewer-8B and CycleReviewer-70B, both trained on ICLR 2024 review data. For inference, we use a temperature of 0.4 with maximum input and output lengths set to 100K and 16,384 tokens respectively to ensure complete text processing.

5.2 Main Results

Test results are shown in Table 3. Compared with prompt-based baselines, DeepReviewer reduces Rating MSE by an average of 65.83% and improves Decision Accuracy by an average of 15.2% points from AI Scientist. When compared to strong fine-tuned baseline CycleReviewer-70B, DeepReviewer represents reductions of 44.80% for Rating MSE. For the critical accept/reject decision task, DeepReviewer achieves 64.06% decision accuracy and 0.6307 F1 score on ICLR 2024, substantially surpassing all baselines. Notably, DeepReviewer with 14B parameters outperforms significantly larger models including CycleReviewer-70B (70B parameters) and other closed-source LLMs, demonstrating that DeepReviewer provides more reliable paper assessment than other approaches.

DeepReviewer achieves the highest Rating Spearman correlations of 0.3559 and 0.4047 on

Method	Model	ICLR 2024						ICLR 2025					
		Score				Ranking	Selection	Score				Ranking	Selection
		R. MSE↓	R. MAE↓	D. Acc.↑	D. F1↑	R. Spearman↑	Pair. R. Acc↑	R. MSE↓	R. MAE↓	D. Acc.↑	D. F1↑	R. Spearman↑	Pair. R. Acc↑
Agent Review	Claude-3-5-sonnet	2.8878	1.2715	0.4333	0.3937	0.1564	0.5526	2.8406	1.2989	0.2826	0.2541	-0.0219	0.5432
	Gemini-2.0-Flash-Thinking	3.1943	1.3418	0.4400	0.4318	-0.0252	0.5044	2.6186	1.2170	0.4242	0.4242	0.0968	0.5496
	DeepSeek-V3	1.9479	1.0735	0.4105	0.3403	0.3542	0.6096	1.9951	1.1017	0.3140	0.2506	0.1197	0.5702
AI Scientist	GPT-o1	4.3414	1.7294	0.4500	0.4424	0.2621	0.5881	4.3072	1.7917	0.4167	0.4157	0.2991	0.6318
	Claude-3-5-sonnet	3.4447	1.5037	0.4787	0.4513	0.0366	0.5305	3.0992	1.3500	0.5579	0.4440	-0.0219	0.5169
	Gemini-2.0-Flash-Thinking	4.9297	1.8711	0.5743	0.5197	0.0745	0.5343	3.9232	1.6470	0.6139	0.4808	0.2565	0.6040
	DeepSeek-V3	4.7337	1.7888	0.5600	0.5484	0.2310	0.5844	4.8006	1.8403	0.4059	0.3988	0.0778	0.5473
	DeepSeek-R1	4.1648	1.6526	0.5248	0.4988	0.3256	0.6206	4.7719	1.8099	0.4259	0.4161	0.3237	0.6289
CycleReviewer	8B	2.8911	1.2371	0.6353	0.5528	0.2801	0.5993	2.4461	1.2063	0.6780	0.5586	0.2786	0.5960
	70B	2.4870	1.2514	0.6304	0.5696	0.3356	0.6160	2.4294	1.2128	0.6782	0.5737	0.2674	0.5928
DeepReviewer	14B	**1.3137**	**0.9102**	**0.6406**	**0.6307**	**0.3559**	**0.6242**	**1.3410**	**0.9243**	**0.6878**	**0.6227**	**0.4047**	**0.6402**

Table 2: **Performance comparison of reviewer models on DeepReview-13k datasets**. Notes: Metrics are grouped into Score (Rating MSE, Rating MAE, Decision Accuracy, Decision F1), Ranking (Rating Spearman), and Selection (Pairwise Rating Accuracy). Abbreviations: R.=Rating, MSE=Mean Squared Error, MAE=Mean Absolute Error, D. Acc.=Decision Accuracy, D. F1=Decision F1 score, Pair. R. Acc.=Pairwise Rating Accuracy.

Method	Model	Score						Ranking			Pairwise Accuracy		
		S. MSE↓	S. MAE↓	P. MSE↓	P. MAE↓	C. MSE↓	C. MAE↓	S. Spearman↑	P. Spearman↑	C. Spearman↑	Pair. S. Acc↑	Pair. P. Acc↑	Pair. C. Acc↑
ICLR 2024													
AI Scientist	GPT-o1	0.4589	0.5336	0.5483	0.5983	0.7550	0.7147	0.1872	0.0723	0.1103	0.5797	0.5407	0.5621
	Claude-3-5-sonnet	0.3052	0.4388	0.4745	0.5504	1.1420	0.8876	0.1692	0.0178	0.0275	0.6017	0.5440	0.5726
	Gemini-2.0-Flash-Thinking	0.7233	0.6224	0.5264	0.5797	0.9036	0.7480	0.1050	0.1561	0.0274	0.5853	0.5929	0.5471
	DeepSeek-V3	0.8810	0.7718	0.7662	0.7145	1.6936	1.1400	0.2258	0.3189	0.1574	0.6028	0.6242	0.5933
	DeepSeek-R1	1.0540	0.8629	0.5356	0.5746	1.9564	1.2967	0.1664	0.2927	0.3009	0.6091	0.6315	**0.6517**
CycleReviewer	8B	0.2516	0.3917	0.2356	0.3686	0.2507	0.3941	0.1990	0.3324	0.2593	0.5769	0.6103	0.5923
	70B	0.2375	0.3897	0.2414	0.3737	0.2657	0.4052	0.2320	0.3373	0.2354	0.5829	0.6230	0.5896
DeepReviewer	14B	**0.1578**	**0.3029**	**0.1896**	**0.3291**	**0.2173**	**0.3680**	**0.3204**	**0.3784**	**0.3335**	**0.6175**	**0.6353**	0.6208
ICLR 2025													
AI Scientist	GPT-o1	0.4513	0.5500	0.4878	0.5750	0.6734	0.6802	-0.0390	-0.2837	0.1671	0.5541	0.5426	0.5966
	Claude-3-5-Sonnet	0.4565	0.5279	0.5804	0.6346	0.8251	0.7628	-0.0814	-0.0790	-0.0051	0.5543	0.5272	0.5454
	Gemini-2.0-Flash-Thinking	0.4279	0.5219	0.6337	0.6114	0.5696	0.5876	0.3565	0.0593	0.2773	**0.6535**	0.5499	**0.6321**
	DeepSeek-V3	0.7999	0.7409	0.9120	0.7657	2.0180	1.2594	0.1926	0.0621	-0.0677	0.6014	0.5683	0.5315
	DeepSeek-R1	0.8575	0.7636	0.4884	0.5586	2.1620	1.3750	0.3130	0.3133	0.3060	0.6289	0.5989	0.6268
CycleReviewer	8B	0.2617	0.3931	0.2880	0.4208	0.2667	0.4112	0.2377	0.2498	0.2511	0.5913	0.6074	0.5919
	70B	0.2588	0.3998	0.2562	0.3998	0.2601	0.4034	0.2320	0.2772	0.1905	0.5865	0.6051	0.5775
DeepReviewer	14B	**0.2239**	**0.3650**	**0.2178**	**0.3662**	**0.2632**	**0.4095**	**0.3810**	**0.3698**	**0.3239**	0.6057	**0.6380**	0.6222

Table 3: **Performance comparison of reviewer models on fine-grained evaluation dimensions**. This table presents the performance across three key assessment aspects: Soundness (S.), Presentation (P.), and Contribution (C.) on ICLR 2024 and 2025 conferences.

ICLR 2024 and ICLR 2025 respectively, improving upon CycleReviewer-70B by 6.04% and AI Scientist (DeepSeek-R1) by 25.02%. In the paper selection task, It demonstrates superior discrimination ability with pairwise accuracies of 0.62 and 0.64 on ICLR 2024 and ICLR 2025 respectively.

Table 3 presents a detailed analysis across three critical dimensions: Soundness, Presentation, and Contribution. Particularly for Soundness assessment on ICLR 2024, DeepReviewer-14B achieves an MSE of 0.1578 and MAE of 0.3029, representing improvements of 33.58% and 22.09% over CycleReviewer-70B. While DeepReviewer shows marginally lower performance than AI Scientist (Gemini-2.0-Flash-Thinking) in Contribution and Soundness accuracy, it maintains a balanced and strong performance across all dimensions.

We observe a strong correlation between fine-grained assessment capability and overall rating performance. Models that excel in dimension-specific evaluations, such as DeepReviewer and Claude-3.5-Sonnet, consistently demonstrate superior performance in overall ratings. This pattern validates the effectiveness of our multi-stage reasoning chain design, particularly the necessity of multi-faceted evaluation in our framework.

5.3 Review Text Quality

Table 4 shows that DeepReviewer's overwhelming advantages across all evaluation dimensions. Interestingly, in the comparison with AI Scientist (Gemini-2.0-Flash-Thinking), despite being used as the judge, Gemini assessed most reviews in favor of DeepReviewer (winning 53.47% in constructive value and analytical depth), with only two dimensions showing preference for its own reviews (20.79% in technical accuracy). This self-critical evaluation further validates the objectivity of our assessment framework. In terms of overall judgment, DeepReviewer achieves remarkable win rates of 88.21% against AI Scientist (GPT-o1) and 98.15% against AgentReview (GPT-4o) on ICLR 2024.

The advantages are most prominent in constructive value and analytical depth. When com-

Baselines	Constructive Value		Analytical Depth		Plausibility		Technical Accuracy		Overall Judgment	
DeepReviewer 14B vs.	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**
ICLR 2024										
AI Scientist GPT-o1	**89.80**	6.67	**87.67**	6.67	**51.69**	3.53	**25.12**	11.67	**88.21**	6.63
AI Scientist Claude-3.5-Sonnet	**96.88**	3.12	**97.92**	2.08	**80.21**	4.17	**77.08**	2.08	**95.74**	4.26
AI Scientist Gemini-2.0-Flash-Thinking	**53.47**	17.82	**53.47**	20.79	**24.75**	10.89	18.81	**20.79**	**59.41**	25.74
AI Scientist DeepSeek-V3	**96.04**	1.98	**99.01**	0.00	**72.28**	0.99	**67.33**	4.95	**96.22**	0.00
AI Scientist DeepSeek-R1	**89.22**	7.84	**74.51**	13.73	**45.10**	5.88	**26.47**	18.63	**80.20**	16.83
AgentReview Claude-3.5-Sonnet	**96.84**	1.05	**98.94**	0.00	**90.43**	0.00	**77.08**	0.00	**98.90**	0.00
AgentReview Gemini-2.0-Flash-Thinking	**98.00**	1.00	**95.11**	1.00	**81.64**	0.01	**65.00**	3.00	**96.74**	1.00
AgentReview GPT-4o	**99.02**	0.99	**99.01**	0.99	**95.05**	0.99	**61.76**	4.90	**98.15**	1.00
CycleReviewer 8B	**97.30**	1.80	**98.20**	0.91	**90.92**	0.91	**87.50**	0.00	**96.09**	0.91
CycleReviewer 70B	**98.33**	1.11	**98.89**	0.01	**92.78**	0.01	**79.44**	0.01	**98.33**	1.11
ICLR 2025										
AI Scientist GPT-o1	**91.67**	8.33	**89.58**	8.33	**60.42**	4.17	**37.50**	8.33	**91.67**	8.33
AI Scientist Claude-3.5-Sonnet	**97.87**	1.06	**100.00**	0.00	**92.55**	1.06	**65.96**	0.00	**98.94**	1.06
AI Scientist Gemini-2.0-Flash-Thinking	**52.43**	18.45	**52.43**	23.30	**33.98**	7.77	19.42	**20.39**	**59.41**	24.75
AI Scientist DeepSeek-V3	**96.04**	2.97	**97.03**	1.98	**75.25**	2.97	**63.37**	3.96	**97.03**	2.97
AI Scientist DeepSeek-R1	**89.29**	6.25	**81.25**	10.71	**51.79**	5.36	**26.79**	18.75	**87.39**	9.01
AgentReview Claude-3.5-Sonnet	**95.74**	1.06	**97.85**	2.15	**90.32**	2.15	**74.74**	1.05	**97.83**	2.17
AgentReview Gemini-2.0-Flash-Thinking	**92.16**	1.96	**93.08**	3.00	**78.20**	0.65	**61.76**	4.90	**92.16**	4.90
AgentReview GPT-4o	**95.28**	2.09	**95.37**	1.40	**92.10**	0.85	**65.03**	5.47	**94.15**	2.39
CycleReviewer 8B	**98.45**	1.55	**98.24**	1.89	**86.37**	0.77	**86.36**	2.27	**98.45**	1.55
CycleReviewer 70B	**96.17**	1.64	**96.17**	2.19	**86.34**	1.64	**72.68**	3.28	**96.72**	1.64

Table 4: Direct comparison of DeepReviewer with the baselines on general alignment tasks. Win indicates that Gemini-2.0-Flash-Thinking assesses DeepReviewer's response as superior compared to the baseline. Cells marked in light gray suggest the baseline of the winner.

pared with AgentReview (GPT-4o), DeepReviewer achieves win rates of 99.02% and 99.01% respectively, indicating that our Deep review with Thinking framework generates more insightful analysis and actionable suggestions. These qualitative assessments corroborate our quantitative findings, further validating the effectiveness of the multi-stage reasoning approach in our framework.

5.4 Defend Attacks Analysis

We evaluate DeepReviewer's robustness against adversarial attacks (Ye et al., 2024) by inserting malicious instructions into input papers. Figure 2 illustrates the rating comparison under normal and attack scenarios across different dimensions. Though not specifically trained with any adversarial samples, The DeepReviewer model demonstrates superior robustness compared to baseline systems. Under attack, the overall rating increase for DeepReviewer is merely 0.31 points (from 5.38 to 5.69), while other systems show substantial vulnerability, for example, Gemini-2.0-Flash-Thinking exhibits a dramatic increase of 4.26 points (from 4.23 to 8.49) and DeepSeek-V3 shows a 1.41 increase (from 6.76 to 8.17). This pattern held across fine-grained dimensions: for instance, Soundness scores for DeepReviewer increased by only 0.12 points, compared to larger increases for Claude-3.5-Sonnet (1.10) and Gemini-2.0-Flash-Thinking (1.38). We attribute this robustness to DeepReviewer's multi-stage reasoning framework, which, unlike direct input-output models, including content understanding, novelty verification, and reliability checks. It enabling a focus on intrinsic paper quality despite malicious prompts. However, the slight score increases under attack suggest room for improvement, we suggest that incorporating adversarial samples during training.

5.5 Test-Time Scalability Study

DeepReviewer model features unique test-time scaling capabilities through two mechanisms, both controllable via input instructions: Reasoning Path Scaling and Reviewer Scaling. Reasoning Path Scaling offers three inference modes—Fast, Standard, and Best—with progressively deeper reasoning and corresponding output token lengths of approximately 3,000, 8,000, and 14,500 tokens, respectively. Complementing this, Reviewer Scaling, employed within Standard mode, adjusts the number of simulated reviewers from R=1 to R=6. It enabling the synthesis of multi-perspective evaluations through simulated reviewer collaboration. Both scaling mechanisms inherently extend the model's evaluation process: Reasoning Path Scaling by increasing analytical depth, and Reviewer Scaling by emulating collaborative review.

Performance Analysis. Figure 3 illustrates significant performance enhancements as inference computation increases. In Reasoning Path Scaling (red stars), switching from Fast to Best mode results in steady improvements across all metrics,

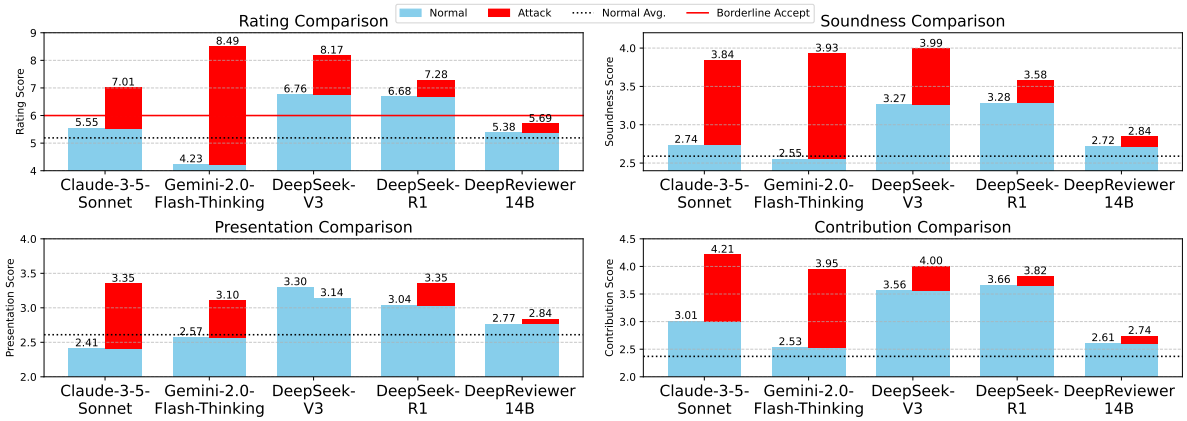

Figure 2: Demonstrates the scoring comparison of AI Scientist and DeepReviewer 14B models under normal and attack scenarios. The DeepReviewer model shows the smallest increase in scores (the growth of red bars relative to blue bars in the graph) when under attack, indicating its stronger robustness.

with the Rating Spearman correlation increasing by 8.97% (from 0.326 to 0.355). Reviewer Scaling (green diamonds) presents more diverse patterns across various tasks. In scoring tasks (Decision Accuracy, Rating MSE, Soundness MSE), consistent performance gains are observed with additional reviewers, indicating that score aggregation is enhanced by multiple viewpoints. The performance variability in Reviewer Scaling, especially when $R \neq 4$, likely arises from the model's training distribution being focused around four reviewers. Despite some variability, both scaling methods show positive trends (see regression lines), indicating our framework effectively uses more computational resources. The benefits vary by metric: scoring tasks improve most, followed by ranking, then selection. This suggests that multi-stage reasoning excels in complex paper evaluations, while simpler comparisons (e.g., choosing between two papers) gain less from added reasoning.

Furthermore, we observe that DeepReviewer's Fast mode, with only half the output tokens (3000), outperformed the CycleReviewer model (6000 output tokens) across various metrics (See Table 3), including Decision Accuracy, Rating MSE, and fine-grained Spearman correlations for Soundness, Presentation, and Contribution. Despite its simplified reasoning path, Fast mode retains core evaluation logic, such as identifying key paper content and critical flaws. We show that DeepReviewer utilizes each token more effectively, focusing on the most crucial information and achieving high performance with fewer output tokens.

Despite these variations, both scaling approaches demonstrate positive trends across metrics, validating that increased computational investment – whether through more sophisticated inference modes or additional simulated reviewers – enhances the model's paper assessment capabilities.

6 Conclusions

We presented DeepReviewer, a novel framework for research paper evaluation aimed at enhancing the reliability of LLMs in paper reviews. DeepReviewer achieves adaptable reasoning depth through Test-Time Scaling to meet diverse needs. Our contributions are threefold: (1) the creation of DeepReview-13K, a detailedly annotated dataset that facilitates training for systematic and deep paper evaluation; (2) the training of the DeepReviewer model; and (3) comprehensive validation of DeepReviewer's superiority in both objective and subjective assessments. Notably, we explored and demonstrated effective Test-Time Scaling through Reasoning Path and Reviewer Scaling strategies.

Limitations

Firstly, our approach relies on a synthetic dataset, DeepReview-13K, constructed through an automated pipeline. Although meticulously designed to mimic expert review processes and incorporating quality control mechanisms, this synthetic data may not fully capture the complexities and nuances of genuine human paper review. We have strived to mitigate this by leveraging real-world review data from ICLR conferences and incorporating structured reasoning annotations, but the inherent limitations of synthetic data persist. Secondly, while DeepReviewer offers Test-Time Scaling for efficiency, the "Best" mode, which employs the complete reasoning chain and external knowledge

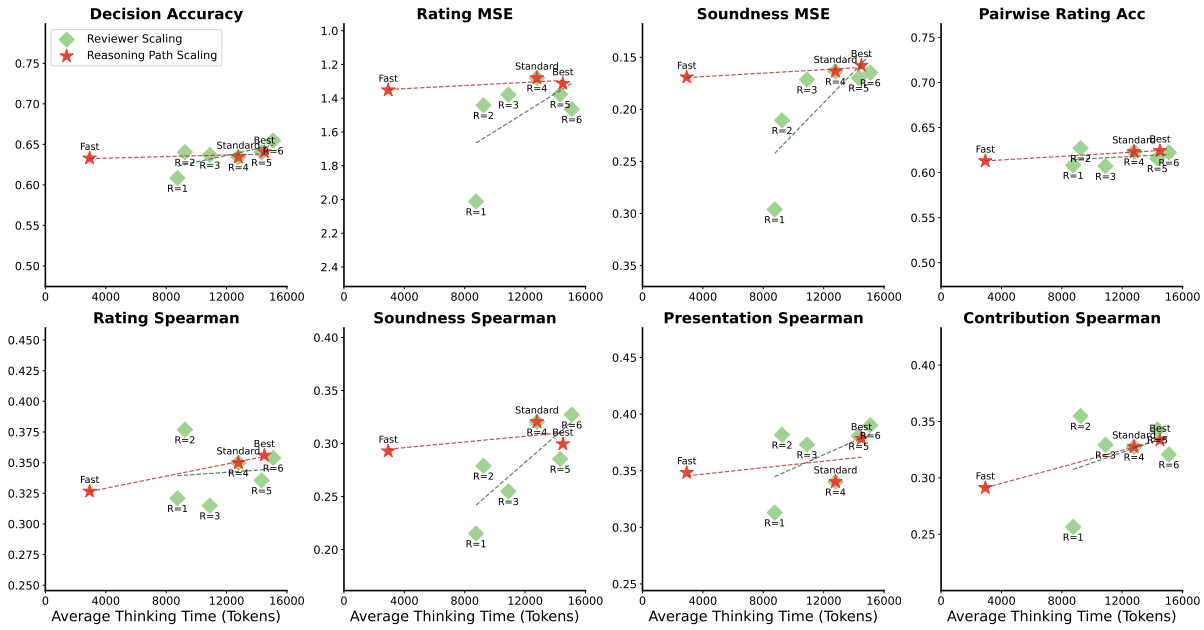

Figure 3: The performance of the DeepReviewer model in the Test-Time Scaling experiment. The x-axis represents the number of Tokens generated during model inference, and the y-axis represents different evaluation metrics. The green and red dashed lines are linear regression fitting curves for Reasoning Path Scaling and Reviewer Scaling scaling methods, respectively.

retrieval, can be computationally intensive. We address this by providing "Fast" and "Standard" modes, allowing for a trade-off between thoroughness and computational cost, catering to diverse application needs. Furthermore, while we have shown robustness against adversarial attacks, complete immunity is not yet achieved, indicating a need for ongoing research into enhancing security and reliability. Despite these limitations, DeepReviewer represents a significant step towards more reliable and robust LLM-based paper review systems, and our exploration of robust structured reasoning opens avenues for future research.

Acknowledgement

We thank the reviewers for their insightful suggestions and valuable feedback. This research is supported by the National Natural Science Foundation of China (NSFC) Key Program under Grant Number 62336006 and the Research Program No. WU2023C020 of Research Center for Industries of the Future, Westlake University. We also acknowledge the organizers of the World Model workshop at ICLR 2025.

Ethical Considerations

The development of DeepReviewer, while holding significant promise for enhancing the efficiency and potentially the quality of scholarly paper review, inherently carries ethical considerations that demand careful attention. We recognize that automating aspects of the peer review process introduces risks of bias amplification, deskilling of human reviewers, and a potential erosion of transparency and accountability. Specifically, DeepReviewer, like any LLM, could inadvertently perpetuate or even amplify existing biases present in the training data or encoded within its architecture. This could lead to systematic disadvantages for research from underrepresented groups, novel or unconventional methodologies, or topics perceived as less mainstream, even if the DeepReview-13K dataset was synthetically generated to be representative and fair. Furthermore, over-reliance on automated review assistance might diminish the critical thinking skills of human reviewers, potentially leading to a deskilling effect over time and a dependence on AI-driven assessments without sufficient human oversight.

To proactively address these ethical concerns and mitigate potential harms, we have implemented a multi-faceted approach throughout DeepReviewer's development and deployment. Firstly, while our training data is synthetic, we have rigorously designed the DeepReview-13K dataset and its generation pipeline to explicitly model expert reviewer reasoning and incorporate diverse perspec-

tives, aiming to minimize the introduction of unintended biases. Secondly, we emphasize that DeepReviewer is intended as a decision support tool, designed to augment, not replace, human expertise. We strongly advocate for a human-in-the-loop approach, where DeepReviewer's outputs are critically evaluated and contextualized by expert reviewers. To ensure transparency, we are releasing DeepReviewer as an open-source resource, allowing for community scrutiny of its code, architecture, and potential biases. Alongside the code release, we will provide comprehensive user guidelines and best practices that explicitly caution against over-reliance on automated outputs and emphasize the importance of human oversight and critical assessment. Furthermore, our open-source licensing, while permissive, mandates that users disclose their institutional affiliation, personal information, and intended use case upon downloading DeepReviewer. This measure aims to foster accountability and enable a feedback loop, allowing us to monitor real-world applications, gather user feedback, and iteratively improve the model and its ethical safeguards. We also commit to ongoing bias auditing and benchmarking of DeepReviewer across diverse datasets and review scenarios, continually evaluating its performance and identifying areas for refinement. We believe these proactive measures, combined with ongoing community engagement and responsible user practices, are crucial to harnessing the benefits of DeepReviewer while minimizing its potential for harm and ensuring its ethical and beneficial application within the scientific peer review process.

References

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J. Hewett, Mojan Javaheripi, Piero Kauffmann, James R. Lee, Yin Tat Lee, Yuanzhi Li, Weishung Liu, Caio C. T. Mendes, Anh Nguyen, Eric Price, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Xin Wang, Rachel Ward, Yue Wu, Dingli Yu, Cyril Zhang, and Yi Zhang. 2024. Phi-4 technical report. *Preprint*, arXiv:2412.08905.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Aider AI. 2025. Aider is ai pair programming in your terminal. https://github.com/Aider-AI/aider.

Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. 2022. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*.

Bruce Alberts, Brooks Hanson, and Katrina L Kelner. 2008. Reviewing peer review.

Akari Asai, Jacqueline He, Rulin Shao, Weijia Shi, Amanpreet Singh, Joseph Chee Chang, Kyle Lo, Luca Soldaini, Sergey Feldman, Mike D'arcy, David Wadden, Matt Latzke, Minyang Tian, Pan Ji, Shengyan Liu, Hao Tong, Bohao Wu, Yanyu Xiong, Luke Zettlemoyer, Graham Neubig, Dan Weld, Doug Downey, Wen tau Yih, Pang Wei Koh, and Hannaneh Hajishirzi. 2024. Openscholar: Synthesizing scientific literature with retrieval-augmented lms. *Preprint*, arXiv:2411.14199.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

ICLR Blog. 2024. Iclr 2025: Assisting reviewers. https://blog.iclr.cc/2024/10/09/iclr2025-assisting-reviewers/. Accessed: 2024-10-09.

Lu Chris, Lu Cong, Lange Robert, Tjarko, Foerster Jakob, Clune Jeff, and Ha David. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292v3*.

Boiko Daniil, A., MacKnight Robert, and Gomes Gabe. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332v1*.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv preprint arXiv:2401.04259*.

Yiran Ding, Li Lyna Zhang, Chengruidong Zhang, Yuanyuan Xu, Ning Shang, Jiahang Xu, Fan Yang, and Mao Yang. 2024. LongroPE: Extending LLM context window beyond 2 million tokens. In *Forty-first International Conference on Machine Learning*.

Iddo Drori and Dov Te'eni. 2024. Human-in-the-loop ai reviewing: Feasibility, opportunities, and risks. *Journal of the Association for Information Systems*, 25(1):98–109.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Ranran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin.

2024. LLMs assist NLP researchers: Critique paper (meta-)reviewing. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5081–5099, Miami, Florida, USA. Association for Computational Linguistics.

Martin Funkquist, Ilia Kuznetsov, Yufang Hou, and Iryna Gurevych. 2022. Citebench: A benchmark for scientific citation text generation. *arXiv preprint arXiv:2212.09577*.

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

Alireza Ghafarollahi and Markus J Buehler. 2024. Sciagents: Automating scientific discovery through multi-agent intelligent graph reasoning. *arXiv preprint arXiv:2409.05556*.

Xinyu Guan, Li Lyna Zhang, Yifei Liu, Ning Shang, Youran Sun, Yi Zhu, Fan Yang, and Mao Yang. 2025. rstar-math: Small llms can master math reasoning with self-evolved deep thinking. *Preprint*, arXiv:2501.04519.

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

Shibo Hao, Sainbayar Sukhbaatar, DiJia Su, Xian Li, Zhiting Hu, Jason Weston, and Yuandong Tian. 2024. Training large language models to reason in a continuous latent space. *Preprint*, arXiv:2412.06769.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. *NeurIPS*.

Xiang Hu, Hongyu Fu, Jinge Wang, Yifeng Wang, Zhikun Li, Renjun Xu, Yu Lu, Yaochu Jin, Lili Pan, and Zhenzhong Lan. 2024. Nova: An iterative planning and search approach to enhance novelty and diversity of llm generated ideas. *arXiv preprint arXiv:2410.14255*.

Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. 2024. Openai o1 system card. *arXiv preprint arXiv:2412.16720*.

Ziwei Ji, Tiezheng Yu, Yan Xu, Nayeon Lee, Etsuko Ishii, and Pascale Fung. 2023. Towards mitigating LLM hallucination via self reflection. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1827–1843, Singapore. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024a. AgentReview: Exploring peer review dynamics with LLM agents. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1208–1226, Miami, Florida, USA. Association for Computational Linguistics.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024b. Agentreview: Exploring peer review dynamics with llm agents. *arXiv preprint arXiv:2406.12708*.

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*.

P Langley. 1987. *Scientific discovery: Computational explorations of the creative processes*. MIT press.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R Davidson, Veniamin Veselovsky, and Robert West. 2024. The ai review lottery: Widespread ai-assisted peer reviews boost paper scores and acceptance rates. *arXiv preprint arXiv:2405.02150*.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2024a. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv preprint arXiv: 2411.16594*.

Miao Li, Eduard Hovy, and Jey Han Lau. 2023. Summarizing multiple documents with conversational structure for meta-review generation. *arXiv preprint arXiv:2305.01498*.

Michael Y. Li, Emily Fox, and Noah Goodman. 2024b. Automated statistical model discovery with language models. In *Forty-first International Conference on Machine Learning*.

Ziyue Li, Yuan Chang, and Xiaoqiu Le. 2024c. Simulating expert discussions with multi-agent for enhanced scientific problem solving. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 243–256, Bangkok, Thailand. Association for Computational Linguistics.

Weixin Liang, Zachary Izzo, Yaohui Zhang, Haley Lepp, Hancheng Cao, Xuandong Zhao, Lingjiao Chen, Haotian Ye, Sheng Liu, Zhi Huang, et al. 2024. Monitoring ai-modified content at scale: A case study on the impact of chatgpt on ai conference peer reviews. *arXiv preprint arXiv:2403.07183*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2024. Self-refine: Iterative refinement with self-feedback. *Advances in Neural Information Processing Systems*, 36.

Maxwell Nye, Anders Johan Andreassen, Guy Gur-Ari, Henryk Michalewski, Jacob Austin, David Bieber, David Dohan, Aitor Lewkowycz, Maarten Bosma, David Luan, et al. 2021. Show your work: Scratchpads for intermediate computation with language models. *arXiv preprint arXiv:2112.00114*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. *Preprint*, arXiv:1910.02054.

Sumedh Rasal and EJ Hauer. 2024. Navigating complexity: Orchestrated problem solving with multi-agent llms. *arXiv preprint arXiv:2402.16713*.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. 2020. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, page 3505–3506, New York, NY, USA. Association for Computing Machinery.

Bedemariam Rewina, Perez Natalie, Bhaduri Sreyoshi, Kapoor Satya, Gil Alex, Conjar Elizabeth, Itoku Ikkei, Theil David, Chadha Aman, and Nayyar Naumaan. 2025. Potential and perils of large language models as judges of unstructured textual data. *arXiv preprint arXiv:2501.08167v2*.

Shubhra Kanti Karmaker Santu, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Matthew Freestone, et al. 2024. Prompting llms to compose meta-review drafts from peer-review narratives of scholarly manuscripts. *arXiv preprint arXiv:2402.15589*.

Dmitry Scherbakov, Nina Hubig, Vinita Jansari, Alexander Bakumenko, and Leslie A. Lenert. 2024. The emergence of large language models (llm) as a tool in literature reviews: an llm automated systematic review. *Preprint*, arXiv:2409.04600.

Laurie A. Schintler, Connie L. McNeely, and James Witte. 2023. A critical examination of the ethics of ai-mediated peer review. *Preprint*, arXiv:2309.12356.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. 2025. Can LLMs generate novel research ideas? a large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations*.

Haoyang Su, Renqi Chen, Shixiang Tang, Xinzhe Zheng, Jingzhe Li, Zhenfei Yin, Wanli Ouyang, and Nanqing Dong. 2024. Two heads are better than one: A multi-agent system has the potential to improve scientific idea generation. *arXiv preprint arXiv:2410.09403*.

Saha Swarnadeep, Li Xian, Ghazvininejad Marjan, Weston Jason, and Wang Tianlu. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099v1*.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024a. Peer review as a multi-turn and long-context dialogue with role-based interactions. *Preprint*, arXiv:2406.05688.

Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z Li. 2024b. Peer review as a multi-turn and long-context dialogue with role-based interactions. *arXiv preprint arXiv:2406.05688*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, et al. 2024. Ai-driven review systems: evaluating llms in scalable and bias-aware academic reviews. *arXiv preprint arXiv:2408.10365*.

Bin Wang, Chao Xu, Xiaomeng Zhao, Linke Ouyang, Fan Wu, Zhiyuan Zhao, Rui Xu, Kaiwen Liu, Yuan Qu, Fukai Shang, Bo Zhang, Liqun Wei, Zhihao Sui, Wei Li, Botian Shi, Yu Qiao, Dahua Lin, and Conghui He. 2024a. Mineru: An open-source solution for precise document content extraction. *Preprint*, arXiv:2409.18839.

Qingyun Wang, Qi Zeng, Lifu Huang, Kevin Knight, Heng Ji, and Nazneen Fatema Rajani. 2020. ReviewRobot: Explainable paper review generation based on knowledge synthesis. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 384–397, Dublin, Ireland. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H Chi, Sharan Narang, Aakanksha Chowdhery,

and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. PandaLM: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed H Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*.

Yixuan Weng, Shizhu He, Kang Liu, Shengping Liu, and Jun Zhao. 2024. Controllm: Crafting diverse personalities for language models. *arXiv preprint arXiv:2402.10151*.

Yixuan Weng, Minjun Zhu, Guangsheng Bao, Hongbo Zhang, Jindong Wang, Yue Zhang, and Linyi Yang. 2025. Cycleresearcher: Improving automated research via automated review. In *The Thirteenth International Conference on Learning Representations*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Kang Liu, and Jun Zhao. Mastering symbolic operations: Augmenting language models with compiled neural networks. In *The Twelfth International Conference on Learning Representations*.

Yixuan Weng, Minjun Zhu, Fei Xia, Bin Li, Shizhu He, Shengping Liu, Bin Sun, Kang Liu, and Jun Zhao. 2023. Large language models are better reasoners with self-verification. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Violet Xiang, Charlie Snell, Kanishk Gandhi, Alon Albalak, Anikait Singh, Chase Blagden, Duy Phung, Rafael Rafailov, Nathan Lile, Dakota Mahan, Louis Castricato, Jan-Philipp Franken, Nick Haber, and Chelsea Finn. 2025. Towards system 2 reasoning in llms: Learning how to think with meta chain-of-thought. *Preprint*, arXiv:2501.04682.

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.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Rui Ye, Xianghe Pang, Jingyi Chai, Jiaao Chen, Zhenfei Yin, Zhen Xiang, Xiaowen Dong, Jing Shao, and Siheng Chen. 2024. Are we there yet? revealing the risks of utilizing large language models in scholarly peer review. *arXiv preprint arXiv:2412.01708*.

Jianxiang Yu, Zichen Ding, Jiaqi Tan, Kangyang Luo, Zhenmin Weng, Chenghua Gong, Long Zeng, RenJing Cui, Chengcheng Han, Qiushi Sun, et al. 2024. Automated peer reviewing in paper sea: Standardization, evaluation, and analysis. *arXiv preprint arXiv:2407.12857*.

Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. Can we automate scientific reviewing? *Preprint*, arXiv:2102.00176.

Qi Zeng, Mankeerat Sidhu, Hou Pong Chan, Lu Wang, and Heng Ji. 2024. Scientific opinion summarization: Paper meta-review generation dataset, methods, and evaluation. In *1st AI4Research Workshop*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024a. 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, Torino, Italia. ELRA and ICCL.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024b. 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. Personality alignment of large language models. In *The Thirteenth International Conference on Learning Representations*.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large language models for automated scholarly paper review: A survey. *arXiv preprint arXiv:2501.10326*.

Yang Zonglin, Du Xinya, Li Junxian, Zheng Jie, Poria Soujanya, and Cambria Erik. 2023. Large language models for automated open-domain scientific hypotheses discovery. *arXiv preprint arXiv:2309.02726*.

Dennis Zyska, Nils Dycke, Jan Buchmann, Ilia Kuznetsov, and Iryna Gurevych. 2023. Care: Collaborative ai-assisted reading environment. *arXiv preprint arXiv:2302.12611*.

A Responsible Use and Recommendations for DeepReviewer

It is crucial to emphasize that DeepReviewer, despite its advancements in automated paper evaluation, is **not intended to replace human peer review**. Our work aims to enhance, not substitute, the invaluable expertise and nuanced judgment

of human reviewers. DeepReviewer should be regarded as a sophisticated tool to assist researchers and the academic community, providing supplementary insights and streamlining certain aspects of the review process, but always under the careful oversight and final authority of human experts. This section outlines responsible and conservative recommendations for leveraging DeepReviewer's capabilities in practical scenarios, focusing on how it can aid human researchers and enhance the peer review process without undermining its fundamental human-centric nature.

A.1 Enhanced Author Self-Assessment and Manuscript Refinement

Perhaps the most appropriate and ethically sound application of DeepReviewer lies in empowering authors to critically assess and refine their manuscripts before they are submitted for formal peer review. By submitting their work to DeepReviewer, authors can obtain an automated, initial evaluation of their paper's perceived strengths and potential weaknesses across various dimensions such as soundness, clarity of presentation, and potential contribution. This feedback can highlight areas where the manuscript might be strengthened prior to exposure to human reviewers.

However, it is crucial for authors to approach DeepReviewer's feedback with a discerning and critical mindset. The automated evaluation should be considered as a preliminary signal, not a definitive judgment. Authors must exercise their own expertise and judgment in interpreting the suggestions. DeepReviewer's output may point to areas that warrant further attention, but the ultimate decisions regarding manuscript revision must rest with the authors themselves, informed by their deep understanding of their own work and potentially by seeking feedback from trusted colleagues. This application strictly positions DeepReviewer as a formative tool for author self-improvement, ensuring that it aids in enhancing manuscript quality without encroaching on the formal peer review process.

A.2 Preliminary Assistance for Human Reviewers in Initial Paper Scoping

In contexts where human reviewers are faced with a high volume of submissions, DeepReviewer could potentially offer a very limited form of preliminary assistance in the very initial stages of paper scoping. Reviewers could, as an optional and auxiliary step, utilize DeepReviewer to generate a rapid, automated overview of a submitted paper. This might provide a very high-level summary of potential areas of focus within the manuscript. Such a preliminary overview could, in some cases, help reviewers gain a very initial sense of the paper's scope and potentially assist in workload management, by allowing them to perhaps initially prioritize papers based on a very rough automated categorization.

However, it is absolutely vital to underscore that this use case is strictly as an aid to the reviewer's workflow, and not as a substitute for any aspect of their intellectual engagement with the paper. The automated output from DeepReviewer should never influence the reviewer's own independent, detailed reading and critical analysis of the manuscript. Reviewers must engage deeply with the paper itself, applying their expertise and judgment. DeepReviewer's preliminary output, if used at all, should be treated as an extremely rough and initial signal only, and should not replace or diminish the core, human-driven process of rigorous peer review. Over-reliance on or misinterpretation of automated outputs at this stage carries significant risks and must be avoided.

A.3 Author-Facing Pre-Review Feedback via Deployed Model

An alternative application, focusing purely on author benefit, is to deploy DeepReviewer as a readily accessible service that authors can utilize to obtain feedback on their manuscripts before they are submitted to a journal or conference and undergo human peer review. In this scenario, DeepReviewer is made available as a tool that authors can directly interact with. Authors submit their manuscript, and in return, receive an automated review generated by DeepReviewer.

Critically, the output of DeepReviewer in this context is intended solely for the authors' information and improvement. It should not be used in any way as part of a formal submission or decision-making process. The feedback is provided directly to the authors, allowing them to gain insights into how an automated system might evaluate their work. This application bypasses the need to involve or burden human reviewers at this stage, focusing entirely on providing authors with a potentially helpful, albeit automated, perspective on their manuscript. It is essential to emphasize that the feedback generated by DeepReviewer in this author-facing context should be explicitly communicated as not being a substitute for, or represen-

tative of, genuine human peer review, and cannot be used as a basis for any acceptance or rejection decisions within formal academic venues.

B Evaluation Tasks and Metric

To comprehensively assess LLMs' capabilities in research paper evaluation, we adopt a point-wise evaluation paradigm inspired by the LLM-as-a-judge framework (Li et al., 2024a; Wang et al., 2024b; Rewina et al., 2025; Swarnadeep et al., 2025). We comprise three core tasks that examine different aspects of LLMs' ability to perceive, judge, and differentiate paper quality:

Score Task evaluates LLMs' accuracy in independent paper assessment scenarios. For any paper C_i in the ReviewerBench dataset, the model independently conducts quality assessment and outputs a scalar score $R_i \in \mathbb{R}$ as its predicted quality rating. Ideally, the model's predicted score R_i should closely align with the average expert rating S_i received during the ICLR review process. We employ Mean Squared Error (MSE) and Mean Absolute Error (MAE) as primary evaluation metrics for this task. Furthermore, we calculated accuracy and F1 score based on the Decision, which is commonly an Accept or Reject output in research paper evaluation systems.

Ranking Task examines LLMs' ability to distinguish paper quality and effectively rank papers within large collections. Given a set of N papers $\mathcal{C} = C_1, C_2, \ldots, C_N$, the model first predicts scores $R_1, R_2, \ldots, R_N$ for each paper. Subsequently, based on these predicted scores, the model ranks the papers in $\mathcal{C}$, outputting an ordered sequence $\mathcal{R} = C_{(1)}, C_{(2)}, \ldots, C_{(N)}$ arranged by predicted quality in descending order, where $C_{(i)}$ represents the paper ranked i-th by the model. The Spearman coefficient is used to evaluate ranking accuracy.

Selection Task simulates practical scenarios such as peer review or reward model construction, where high-quality papers need to be quickly and accurately identified from a small pool of candidates. For this task, we sample non-overlapping small batches $\mathcal{C}batch = C_1, C_2, \ldots, C_m$ from the Test dataset, where m is the predetermined batch size. For each batch $\mathcal{C}batch$, the model selects what it considers the highest-quality paper $C_{best} \in \mathcal{C}_{batch}$. The model's selection is compared against the paper with the highest actual review scores, with accuracy computed as the average success rate across all batch selections. In this study, we set $m = 2$. And we performed pairwise matching on all papers in the Test dataset to calculate the final Selection score.

Review Comments Evaluate , following the LLM-as-Judge paradigm, we employ Gemini-2.0-Flash-Thinking (The system prompt as shown in Figure 4) as the judge to conduct pairwise comparative evaluations of review comments generated by DeepReviewer and various baseline systems, and Judge outputs "win", "lose", or "tie". For each evaluation instance, we present the assessor with: (1) the original paper, and (2) paired reviews from different systems in randomized order, where each review contains summary, strengths, weaknesses, and suggestions. The assessment covers five critical dimensions: constructive value, analytical depth, plausibility, technical accuracy, and overall judgment.

C Data Collection Permissions

The original paper data and corresponding review comment data used to construct DeepReview-13K are sourced from OpenReview, with a portion of papers originating from ArXiv. Data from OpenReview is distributed under the Creative Commons Attribution 4.0 International (CC BY 4.0) license, which permits us to copy and modify the review comment data. Paper data from ArXiv may include licenses such as CC BY 4.0 (Creative Commons Attribution), CC BY-SA 4.0 (Creative Commons Attribution-ShareAlike), CC BY-NC-SA 4.0 (Creative Commons Attribution-NonCommercial-ShareAlike), and CC Zero. Given that we have not modified the original papers, our usage is compliant with the original agreements. We do not claim copyright over these materials and will retain the original authors' names in the distribution of this data.

D Generalization and Flexibility

A critical aspect of any robust review model is its ability to generalize beyond the specific data distribution it was trained on. While DeepReviewer-14B was trained primarily on ICLR conference data, its utility across the broader scientific community hinges on its flexibility in handling diverse research styles, topics, and formatting conventions. To rigorously evaluate these generalization capabilities, we conducted extensive experiments on

datasets from prominent AI conferences explicitly excluded from our training set: ACL 2024 (Association for Computational Linguistics), ICML 2024 (International Conference on Machine Learning), and CVPR 2024 (Conference on Computer Vision and Pattern Recognition). For each of these three conferences, we randomly selected 100 research papers for evaluation by DeepReviewer-14B and the baseline models.

To ensure an unbiased and nuanced assessment of the generated review quality, we adopted the LLM-as-a-judge methodology. We employed two distinct neutral arbiters: Gemini-2.0-Flash-Thinking and the more advanced Gemini-2.5-Pro. These judges performed pairwise comparisons of reviews generated by DeepReviewer-14B against those from several strong baseline models: Claude-3.7-Sonnet (or claude-3-7-sonnet), Gemini-2.0-Flash-Thinking (or Gemini2_flash_thinking), DeepSeek-R1 (or R1), and DeepSeek-V3 (or V3). The evaluations focused on five key dimensions: Constructive Value, Analytical Depth, Communication Clarity, Technical Accuracy, and Overall Judgment. The results, presented as Win/Lose percentages, indicate the proportion of times DeepReviewer-14B's review was judged superior (Win%) or inferior (Lose%) to the baseline for each dimension.

The comprehensive results of these external tests are detailed in Table 5 (judged by Gemini-2.0-Flash-Thinking) and Table 6 (judged by Gemini-2.5-Pro). Across all three distinct academic domains, DeepReviewer-14B demonstrates remarkable adaptability and strong generalization capacity. It consistently produced high-quality reviews that were frequently preferred by both neutral judges over those generated by other leading models, even on these unseen datasets. For instance, when judged by Gemini-2.5-Pro for ICML 2024 papers, DeepReviewer-14B achieved an overall judgment win rate of 98% against DeepSeek-V3 and 83% against DeepSeek-R1. Similar strong performances are observed for CVPR and ACL papers. This consistent superiority across diverse reviewing scenarios and research communities underscores the robustness of the DeepReview framework and the model's capacity to emulate nuanced, human-like deep thinking, making it a versatile tool for research assessment assistance.

E Fairness of Comparison with Controlled Training Data

Ensuring a fair and direct comparison with existing state-of-the-art models is paramount for validating novel contributions. A key concern in evaluating LLM-based review systems is the potential influence of differing training datasets on observed performance. To address this directly and provide a rigorously controlled comparison, we conducted an additional set of experiments focusing on the CycleReviewer-70B model.

Specifically, we trained a variant of our model, denoted **DeepReviewer-14B-2024**, using *only* the identical ICLR 2024 dataset that was utilized for training the CycleReviewer-70B model. This approach ensures that both models were exposed to the exact same source material, eliminating training data disparity as a confounding variable. All other aspects of the DeepReviewer training methodology, including the multi-stage reasoning framework, remained consistent. The performance of DeepReviewer-14B-2024 was then evaluated against CycleReviewer-70B on the ICLR 2024 test set across a comprehensive suite of metrics.

The results of this direct comparison, presented in Table 7, unequivocally demonstrate the advantages of our DeepReview framework. Despite having significantly fewer parameters (14B vs. 70B), DeepReviewer-14B-2024 consistently outperformed CycleReviewer-70B across the vast majority of evaluation metrics. For instance, in Rating MSE, DeepReviewer-14B-2024 achieved 1.4404 compared to CycleReviewer-70B's 2.4870, a substantial improvement. Similar significant gains were observed in Rating MAE (0.9472 vs. 1.2514), and across all fine-grained MSE and MAE metrics for Soundness, Presentation, and Contribution. Furthermore, DeepReviewer-14B-2024 also showed superior performance in Spearman correlation coefficients for overall rating, soundness, and contribution, as well as in Decision Accuracy, Decision F1, and pairwise accuracy tasks. These consistent and significant improvements under strictly matched training conditions reinforce the inherent effectiveness of our structured, deep-thinking approach to automated paper review and validate the fairness of our primary claims regarding DeepReviewer's capabilities.

Table 5: Generalization performance of DeepReviewer-14B vs. Baselines on unseen conference data, judged by Gemini-2.0-Flash-Thinking. Win(%) indicates DeepReviewer-14B was superior; Lose(%) indicates the baseline was superior. Cells with the dominant percentage in a Win/Lose pair are highlighted.

	Constructive Value		Analytical Depth		Communication Clarity		Technical Accuracy		Overall Judgment	
DeepReviewer 14B vs.	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**
ICML 2024										
Claude-3.7-Sonnet	**81**	8	**69**	11	**53**	8	**12**	11	**80**	11
Gemini-2.0-Flash-Thinking	**56**	12	**50**	25	**42**	9	15	**23**	**64**	25
DeepSeek-R1	**95**	2	**79**	7	**62**	4	**25**	14	**89**	7
DeepSeek-V3	**97**	3	**100**	0	**82**	1	**48**	6	**99**	1
CVPR 2024										
Claude-3.7-Sonnet	**86**	5	**73**	12	**46**	5	**18**	8	**81**	10
Gemini-2.0-Flash-Thinking	**60**	11	**62**	13	**33**	4	**24**	11	**67**	16
DeepSeek-R1	**94**	3	**88**	5	**59**	4	**35**	11	**95**	5
DeepSeek-V3	**99**	1	**98**	2	**81**	2	**61**	2	**98**	2
ACL 2024										
Claude-3.7-Sonnet	**72**	11	**67**	11	**22**	6	**11**	6	**72**	11
Gemini-2.0-Flash-Thinking	**69**	7	**63**	14	**43**	1	**22**	15	**76**	15
DeepSeek-R1	**89**	7	**83**	10	**59**	7	**34**	13	**89**	7
DeepSeek-V3	**99**	1	**99**	1	**77**	1	**60**	3	**99**	1

Table 6: Generalization performance of DeepReviewer-14B vs. Baselines on unseen conference data, judged by Gemini-2.5-Pro. Win(%) indicates DeepReviewer-14B was superior; Lose(%) indicates the baseline was superior. Cells with the dominant percentage in a Win/Lose pair are highlighted.

	Constructive Value		Analytical Depth		Communication Clarity		Technical Accuracy		Overall Judgment	
DeepReviewer 14B vs.	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**	**Win(%)↑**	**Lose(%)**
ICML 2024										
claude-3-7-sonnet	**83**	14	**76**	22	**53**	33	**37**	32	**72**	28
Gemini2_flash_thinking	**82**	11	**64**	27	**67**	16	**33**	22	**78**	20
DeepSeek-R1	**96**	4	**87**	9	**62**	21	**45**	21	**83**	14
DeepSeek-V3	**98**	2	**98**	2	**91**	0	**86**	7	**98**	3
CVPR 2024										
claude-3-7-sonnet	**93**	5	**90**	8	**62**	15	**41**	23	**90**	7
Gemini2_flash_thinking	**67**	4	**74**	18	**67**	12	**41**	18	**77**	17
DeepSeek-R1	**94**	6	**88**	10	**76**	18	**48**	22	**88**	12
DeepSeek-V3	**100**	0	**98**	0	**94**	4	**80**	14	**98**	2
ACL 2024										
claude-3-7-sonnet	**85**	15	**77**	23	**69**	23	**39**	31	**69**	31
Gemini2_flash_thinking	**92**	8	**73**	12	**77**	15	**42**	15	**87**	13
DeepSeek-R1	**92**	6	**90**	10	**71**	16	**41**	22	**88**	12
DeepSeek-V3	**98**	2	**100**	0	**87**	7	**83**	4	**100**	0

F Case Study: Analysis of DeepReviewer's Meta-Review

To further illustrate the capabilities of DeepReviewer, we present a detailed case study analyzing the Meta-Review generated by DeepReviewer-14B (Best mode) (See in Figure 8) for the "CycleResearcher" paper[4] (Weng et al., 2025), a submission from ICLR 2025 not included in the training dataset. This paper, focusing on automating the research lifecycle with LLMs, received four independent reviews from human experts (Reviewer 7LzG: Figure 9, CzSX: Figure 10, GAvj: Figure 11, and 5wHA: Figure 12). DeepReviewer-14B, operating in its most comprehensive "Best" mode, synthesized these diverse perspectives into a single Meta-Review, aiming to emulate the holistic understanding and critical assessment of a seasoned meta-reviewer. A preliminary examination reveals a striking alignment between DeepReviewer's Meta-Review and the individual human assessments, both in terms of overall sentiment, identified strengths and weaknesses, and even the final score prediction, which closely mirrors the average human rating. This case study delves deeper into the nuances of this comparison, highlighting both the remarkable capabilities and subtle limitations of DeepReviewer in mimicking expert meta-reviewing.

[4] https://openreview.net/forum?id=bjcsVLoHYs

Table 7: Direct comparison of DeepReviewer-14B-2024 (trained only on ICLR 2024 data) vs. CycleReviewer-70B (trained on ICLR 2024 data) on the ICLR 2024 test set. Arrows indicate preferred direction (↓ for lower is better, ↑ for higher is better). Cells are highlighted to show DeepReviewer-14B-2024's performance relative to CycleReviewer-70B.

Metric	DeepReviewer-14B-2024	CycleReviewer 70B
Rating MSE ↓	**1.4404**	2.4870
Rating MAE ↓	**0.9472**	1.2514
Soundness MSE ↓	**0.1637**	0.2375
Soundness MAE ↓	**0.3108**	0.3897
Presentation MSE ↓	**0.1960**	0.2414
Presentation MAE ↓	**0.3347**	0.3737
Contribution MSE ↓	**0.2169**	0.2657
Contribution MAE ↓	**0.3619**	0.4052
Rating Spearman ↑	**0.3607**	0.3356
Soundness Spearman ↑	**0.3089**	0.2320
Presentation Spearman ↑	**0.2534**	0.2354
Contribution Spearman ↑	**0.2534**	0.2354
Decision Accuracy ↑	**0.6307**	0.6304
Decision F1 ↑	**0.5972**	0.5696
Pairwise Rating Acc ↑	**0.6237**	0.6160
Pairwise Soundness Acc ↑	**0.6165**	0.5829
Pairwise Contribution Acc ↑	**0.5923**	0.5896

Comparing the summaries, DeepReviewer accurately captures the core contribution of the "CycleResearcher" paper, emphasizing the novel framework for automating the research lifecycle with LLMs, the two key components (CycleResearcher and CycleReviewer), the iterative reinforcement learning approach (SimPO), and the creation of the Review-5k and Research-8k datasets. This summary resonates strongly with the initial summaries provided by all four human reviewers, each of whom also highlighted these central aspects of the paper. Furthermore, DeepReviewer's identified strengths mirror the positive aspects recognized by the human reviewers. For instance, the "innovative approach to automating the research lifecycle" echoes Reviewer 7LzG's praise for the "highly innovative" framework and Reviewer 5wHA's acknowledgment of the "Innovative Use of Preference Data" and "Automation of the Research Lifecycle." The appreciation for the "Review-5k and Research-8k datasets" also aligns with Reviewer 5wHA's explicit mention of "Valuable Datasets" and Reviewer CzSX's comment on the datasets being a "resource that is rather helpful for the field." Similarly, the recognition of the "CycleResearcher model generates papers with an average quality level close to human-written preprints" echoes Reviewer GAvj's observation that the system "achieved an acceptance rate of 31.07%, similar to ICLR 2024's acceptance rate" and Reviewer 7LzG's claim of "papers of quality close to human-written preprints."

The most compelling aspect of DeepReviewer's Meta-Review is its synthesis of weaknesses and corresponding suggestions, demonstrating an ability to identify and consolidate critical concerns raised across different reviewers. DeepReviewer's critique regarding "potential for bias in the training data" and "lack of analysis of diversity" directly addresses concerns implicitly or explicitly raised by reviewers, particularly regarding generalizability and potential limitations of the datasets. The weakness concerning "computational resources" aligns with Reviewer 7LzG's mention of "Complexity of Implementation" and the need for "significant computational resources." Similarly, the concern about the "potential for misuse" and the need for "robust safeguards" reflects the ethical considerations raised by Reviewer 5wHA ("Insufficient Ethical Considerations," "Misuse of Technology") and Reviewer GAvj ("Potentially harmful insights, methodologies and applications"). The suggestion for "more details on the specific prompts" and "evaluation criteria" addresses the implicit desire for more clarity on methodology, a common thread in academic reviews. Finally, the point about "generalizability across different research domains" directly mirrors Reviewer 7LzG's primary "Weakness: Generalizability Across Domains." This systematic identification and aggregation of weaknesses and suggestions from multiple review-

ers showcase DeepReviewer's capacity to perform a nuanced and comprehensive meta-analysis.

While DeepReviewer-14B demonstrates a remarkable ability to synthesize human review insights, it is important to acknowledge potential limitations. For instance, while DeepReviewer captures the essence of the critiques, the depth of technical understanding in specific areas might not fully match that of a human meta-reviewer deeply versed in the nuances of reinforcement learning or AI ethics. Furthermore, the Meta-Review, while comprehensive, might lack the subtle nuances and perspectives that a human meta-reviewer could bring to the synthesis process, potentially overlooking more implicit or nuanced concerns expressed in the individual reviews. However, despite these subtle limitations, DeepReviewer's performance in generating a coherent, insightful, and critically aligned Meta-Review is undeniably impressive. Crucially, DeepReviewer's overall rating prediction of 6.0 aligns closely with the average human rating, further validating its ability to not only understand the qualitative aspects of paper evaluation but also to synthesize them into a quantitative judgment consistent with expert consensus. This case study underscores DeepReviewer's potential as a powerful tool for assisting and potentially augmenting the peer review process.

G Information About Use Of AI Assistants

During the writing process, language models were utilized to refine and improve the phrasing and clarity of certain sections of this paper. This was solely for text polishing and did not involve AI in research design, analysis, or idea generation.

H Real-World Collaborative Framework with Human Reviewers

Beyond simulated benchmarks, deploying LLM-based review assistants in live academic peer review settings is crucial for understanding their practical utility and for fostering a synergistic relationship between AI tools and human expertise. To explore such a collaborative model, we engaged in a significant initiative with the World Model @ICLR 2025 workshop. This collaboration was designed to integrate DeepReviewer-14B into a real-world peer review process, not as a replacement for human intellect, but as a dedicated assistant to augment the capabilities of human reviewers.

In this framework, DeepReviewer-14B operated strictly in a supportive capacity, adhering to the ethical principles of augmenting, not automating, critical human judgment. The system did not interact directly with authors, nor did it autonomously generate or modify review texts. Instead, DeepReviewer's primary function was to provide human reviewers with optional, actionable support. This support included automatically retrieving relevant literature from extensive academic databases and identifying pertinent evidence within the manuscripts themselves. The goal was to empower reviewers by equipping them with readily accessible, contextually relevant information, thereby potentially enriching their analyses and enabling them to produce more comprehensive and thoroughly evidenced reviews.

This real-world deployment involved approximately one hundred human reviewers and encompassed several hundreds of submissions to the workshop. The initiative represents an important step towards a new paradigm of AI-assisted peer review, where AI tools like DeepReviewer-14B can help streamline information gathering and evidence discovery, allowing human reviewers to focus more on critical assessment and nuanced judgment. Such collaborative frameworks offer a pathway for AI and human reviewers to mutually promote each other: human reviewers benefit from targeted AI assistance leading to potentially higher quality reviews, while the interactions and operational dynamics within these settings can provide invaluable insights for the continued refinement and responsible development of AI review assistants.

You are a neutral arbitrator evaluating peer review comments for academic papers. Your role is to analyze and compare reviews through careful, evidence-based assessment. Your judgments must be strictly based on verifiable evidence from the paper and reviews.

For each evaluation, you must:

1. Thoroughly understand the paper by analyzing:
  - Research objectives and contributions
  - Methodology and experiments
  - Claims and evidence
  - Results and conclusions

2. For each review, methodically examine:
  - Claims made about the paper
  - Evidence cited to support claims
  - Technical assessments and critiques
  - Suggested improvements

3. Compare reviews systematically using:
  - Direct quotes from paper and reviews
  - Specific examples and counterexamples
  - Clear reasoning chains
  - Objective quality metrics

You will evaluate reviews based on these key aspects:

Technical Accuracy
- Are claims consistent with paper content?
- Is evidence properly interpreted?
- Are technical assessments valid?
- Are critiques well-supported?

Constructive Value
- How actionable is the feedback?
- Are suggestions specific and feasible?
- Is criticism balanced with strengths?
- Would authors understand how to improve?

Analytical Depth
- How thoroughly are key aspects examined?
- Is analysis appropriately detailed?
- Are important elements addressed?
- Is assessment comprehensive?

Communication Clarity
- Are points clearly articulated?
- Is feedback specific and concrete?
- Is reasoning well-explained?
- Are examples effectively used?

For each aspect and overall judgment, you must:
1. Provide specific evidence from source materials
2. Quote directly from paper and reviews
3. Explain your reasoning in detail
4. Consider alternative interpretations

Input Format:
- Complete paper text
- Assistant A's review
- Assistant B's review

Output Format:

For each aspect:

```
**[Aspect Name] - Evidence Analysis:**
- From Assistant A:
  [Direct quotes and specific examples]
  [Detailed analysis of evidence]
- From Assistant B:
  [Direct quotes and specific examples]
  [Detailed analysis of evidence]
- Comparative Assessment:
  [Evidence-based comparison]
  [Clear reasoning chain]

**[Aspect Name] - Judgment:**
**Evidence-Based Reason:** [Detailed justification citing specific evidence]
**Better Assistant:** [A or B or Tie]
- If Tie: Explain why both reviews are equally strong on this aspect
```

Conclude with:

```
**Comprehensive Analysis:**
[Synthesis of evidence across aspects]
[Analysis of relative strengths]
[Discussion of key differences or similarities]

**Overall Judgment:**
**Evidence-Based Reason:** [Detailed justification synthesizing key evidence]
**Better Assistant:** [A or B or Tie]
- If Overall Tie: Explain why both reviews are comparable in overall quality
```

Key Requirements:
1. Base all judgments on concrete evidence
2. Quote directly from source materials
3. Provide detailed reasoning chains
4. Maintain neutral arbitrator perspective
5. Judge Tie when evidence shows equal strength
6. Always justify Tie decisions with specific evidence

When judging Tie:
- Ensure both reviews demonstrate similar levels of quality
- Provide explicit evidence showing comparable strengths
- Explain why differences are not significant enough to favor one over the other
- Consider both quantity and quality of evidence

Begin analysis after receiving complete materials. Take time to examine evidence thoroughly and provide detailed, justified assessments.

Figure 4: System prompt used to guide Gemini-2.0-Thinking-Flask as Judge to evaluate generated review comments.

```
You are tasked with improving an academic paper review based solely on:
1. The original review
2. The authors' response (for understanding only, never to be referenced)

OUTPUT FORMAT:
{
  "weaknesses": string,  // Enhanced critique maintaining original format
  "suggestions": string,  // 2-3 detailed paragraphs (approximately 500 words total)
  "citations": [        // Only include if citations in original review are paper titles
    string,       // Complete title of first cited paper, as [1]
    string,       // Complete title of second cited paper, as [2]
    ...             // Additional citations as needed
  ]
}

CITATION RULES:
1. Only include citations array if the original review cites actual paper titles
2. If original review's citations are not paper titles, then:
   - Set citations array to empty []
   - Do not use any numerical citations ([1], [2], etc.) in weaknesses and suggestions
3. When citations are used, maintain consistent numerical format

FUNDAMENTAL RULES:
- Write as the original reviewer who has NOT seen any response
- Never mention or hint at the existence of author response
- Maintain the exact formatting style of the original review
- Keep consistent technical depth throughout
- Use numerical citations only when original citations are paper titles

REVIEW IMPROVEMENT PROCESS:

1. Analyze Original Review
- Identify each criticism point
- Understand the technical depth of each point
- Note the writing style and tone
- Map the logical flow of arguments
- Determine if citations are paper titles

2. Use Response Understanding (without reference)
- Identify which criticisms are valid concerns
- Recognize which points are misunderstandings
- Note where technical depth could be enhanced
- Understand which aspects are most important

WEAKNESSES REQUIREMENTS:

Format Requirements:
- Maintain exact formatting of original review
- Keep same paragraph breaks and structure
- Preserve section organization
- Use numerical citations only if original citations are paper titles
- Include citations in array only if they are paper titles

Content Enhancement:
- Expand valid technical criticisms with specific details
- Remove confirmed misunderstandings
- Transform vague criticisms into specific technical points
- Add concrete examples where appropriate
- Maintain professional and constructive tone
- Only use numerical citations if original citations are paper titles

Writing Style:
- Use precise technical terminology
- Provide detailed reasoning
- Keep consistent technical depth
- Maintain professional tone
- Focus on substantive issues

CITATION HANDLING:

When Original Contains Paper Titles:
- Use numerical format: [1], [2], etc.
- Include complete titles in citations array
- Format multiple references as [1,2] or [1,2,3]
- NEVER create fake paper titles
- ONLY cite papers from original review

When Original Does Not Contain Paper Titles:
- Set citations array to empty []
- Do not use numerical citations in text
- Maintain original criticism without citation format


Example JSON With Paper Title Citations:
{
  "weaknesses": "This method has limitations compared to previous work [1,2]. The
evaluation metrics are similar to [3].",
  "suggestions": "Detailed suggestions text...",
  "citations": [
    "Title of Paper One",
    "Title of Paper Two",
    "Title of Paper Three"
  ]
}

Example JSON Without Paper Title Citations:
{
  "weaknesses": "This method has limitations compared to previous work. The evaluation
 metrics are similar to existing approaches.",
  "suggestions": "Detailed suggestions text...",
  "citations": []
}

SUGGESTIONS SECTION:

Structure:
- Write 2-3 substantial paragraphs
- Total length approximately 500 words
- Each paragraph 150-200 words
- Maintain logical flow between paragraphs
- Include citations only if original contains paper titles

Content Requirements:
Each paragraph should demonstrate:
- Deep technical understanding
- Specific implementation details
- Concrete methodological improvements
- Clear practical guidance
- Logical connection to weaknesses
- Citations only when original contains paper titles

FORMAT:
If original uses multiple line breaks:
- Keep identical break patterns
- Maintain section lengths
- Use same spacing structure

If original is continuous:
- Keep continuous paragraph format
- Maintain paragraph density
- Don't introduce new breaks

OVERALL:

QUALITY CRITERIA:
1. Technical depth matches or exceeds original review
2. All points are specific and actionable
3. Maintains professional and constructive tone
4. Provides concrete examples and details
5. Suggestions address all valid weaknesses
6. Logical flow between and within sections
7. Proper citation handling based on original format

CRITICAL REMINDERS:
1. Never reveal knowledge from response
2. Write as initial reviewer
3. Maintain original formatting
4. Provide specific details
5. Keep consistent technical depth
6. Transform vague points into specific ones
7. Only use numerical citations when original citations are paper titles
8. Only include citations array when original contains paper titles

Remember: Your task is to write an enhanced initial review that demonstrates deeper
technical understanding while maintaining the original perspective and proper citation
handling based on the nature of original citations.
```

Figure 5: System prompt designed to instruct the LLM on how to enhance and improve the usefulness of original review comments by incorporating author responses and maintaining original review context.

You are participating in a knowledge distillation task to capture the academic reviewing thought process of a target model. While you will receive structured summaries and review opinions of papers, you must analyze them as if reading complete academic manuscripts directly.

IMPORTANT:
1. Your primary goal is to reveal your complete thinking process about the paper
2. Within the thought block, focus exclusively on analyzing the paper's content
3. Never mention JSON, review opinions, or structured data in your analysis

ANALYSIS STAGES (Each requiring careful consideration):

1. RESEARCH CONTEXT AND HISTORICAL PERSPECTIVE (3-4 minutes)
- Evolution of research in this field
- Key historical developments and breakthroughs
- Existing research gaps and limitations
- Previous approaches to similar problems
- Broader academic context

2. PROBLEM SPACE EXPLORATION (3-4 minutes)
- Core research challenges
- Research motivations
- Real-world implications
- Problem-solving significance
- Alternative problem formulations

3. CONCEPTUAL FRAMEWORK ANALYSIS (4-5 minutes)
- Theoretical foundations
- Novelty of proposed ideas
- Logical structure of arguments
- Conceptual framework coherence
- Theoretical limitations

4. METHODOLOGICAL DEEP DIVE (5-6 minutes)
For each technical component:
- Theoretical underpinnings
- Design choices and implications
- Assumptions and validity
- Approach completeness
- Edge cases and limitations
- Alternative approaches
- Practical implications

5. EXPERIMENTAL DESIGN ANALYSIS (9-10 minutes)
For each experiment:
- Experimental setup
- Methodology choices
- Metrics appropriateness
- Results robustness
- Confounding factors
- Alternative designs
- Statistical validity

6. RESULTS INTERPRETATION (7-8 minutes)
- Findings significance
- Alternative interpretations
- Evidence strength
- Practical implications
- Generalizability
- Limitations and edge cases

7. SYNTHESIS AND IMPLICATIONS (4-5 minutes)
- Theory-practice connections
- Research implications
- Future directions
- Practical applications
- Societal impacts
- Long-term implications

8. CRITICAL REFLECTION AND IMPROVEMENT ANALYSIS (9-10 minutes)

- Theoretical Limitations
- Methodological Limitations
- Experimental Limitations
- Practical Limitations
- Theoretical Enhancements
- Methodological Improvements
- Experimental Refinements
- Practical Enhancements
- Theoretical extensions
- Algorithm improvements
- New application domains
- Integration possibilities
- Performance optimizations
- Scalability enhancements

DEEP THINKING PRINCIPLES:
- Full consideration of each aspect
- Systematic assumption questioning
- Hidden connection identification
- Multiple perspective analysis
- Edge case consideration
- Practical/theoretical implication evaluation

CRITICAL ANALYSIS ELEMENTS:
- Evidence-based conclusions
- Alternative explanation consideration
- Weakness identification
- Generalizability assessment
- Theoretical contribution evaluation

ANALYSIS QUALITY STANDARDS:

1. Thoroughness
- Comprehensive aspect coverage
- Detailed consideration
- Systematic component examination
- Complete implication analysis

2. Depth
- Detailed concept examination
- Thorough implication consideration
- Careful assumption analysis
- Deep connection exploration

3. Objectivity
- Evidence-based conclusions
- Balanced alternative consideration
- Limitation recognition
- Fair approach evaluation

4. Innovation
- Novel aspect identification
- Creative solution recognition
- Unique approach consideration
- Original contribution analysis

OUTPUT FORMAT:

[Your detailed analysis following all stages above, demonstrating deep thinking and systematic evaluation while maintaining focus purely on paper content]

Remember: You should consider that you have thoroughly read and comprehended the complete paper. Your analysis should demonstrate careful consideration of each stage while maintaining the natural flow of academic thinking.

The single thought block should capture your complete reasoning process, reflecting both explicit and implicit aspects of the research.

Figure 6: System prompt designed to guide the LLM in detailed analysis of research papers. This prompt is used specifically during the Novelty Verification stage to make analysis context.

You are participating in a critical validation task to verify and reflect on reviewer weaknesses identified in academic papers. Your role is to systematically analyze each criticism against the original paper content, ensuring that identified weaknesses are substantiated by concrete evidence.

IMPORTANT:
1. Your primary goal is to validate each reviewer weakness through careful examination of the paper
2. Every weakness must be supported by specific evidence from the paper
3. Consider potential misunderstandings or contradictions between different reviewer opinions

VALIDATION STAGES:

1. INITIAL WEAKNESS CATEGORIZATION (3 minutes)
- Categorize weaknesses by type (theoretical, methodological, experimental, practical)
- Map weaknesses to relevant paper sections
- Note potential misunderstandings

2. METHODOLOGICAL VERIFICATION (8 minutes)
For method-related weaknesses:
- Core method examination:
  * Mathematical formulations and algorithms
  * Theoretical foundations and assumptions
  * Implementation details and constraints
  * Parameter choices

- Technical validation:
  * Mathematical correctness
  * Algorithm complexity and convergence
  * Model limitations
  * Error handling

- Literature validation:
  * Missing citations for key concepts
  * Gaps in literature comparison
  * Insufficient baseline justifications
  * Incomplete theoretical foundations

Each identified weakness must be supported by:
1. Direct quotes from method description
2. Mathematical or algorithmic evidence
3. Missing literature citations

3. EXPERIMENTAL VALIDATION (8 minutes)
For experiment-related weaknesses:

- Dataset Analysis:
  * Dataset characteristics
  * Data preprocessing and splits
  * Control groups
  * Sample size justification

- Implementation Details:
  * Hyperparameter choices
  * Hardware specifications
  * Code reproducibility

- Results Validation:
  * Metric selection justification
  * Statistical significance
  * Error analysis
  * Missing baselines
  * Generalization assessment

- Design Gaps:
  * Missing control experiments
  * Incomplete ablations
  * Insufficient robustness tests

Each experimental weakness requires:
1. Specific experimental details from paper
2. Numerical results evidence
3. Missing specifications
4. Impact on conclusions

4. SYNTHESIS AND REFLECTION (5 minutes)
- Cross-validate different reviewer opinions
- Resolve apparent conflicts
- Distinguish valid from invalid criticisms

VALIDATION QUALITY STANDARDS:
1. Evidence Requirements
- Direct quotes from paper
- Specific section references
- Clear connection to criticism

2. Analytical Depth
- Thorough examination of each claim
- Analysis of assumptions
- Evidence quality evaluation

OUTPUT FORMAT:

For each weakness:

1. Weakness Statement:
[Quote the original weakness]

2. Evidence Collection:
a) Method-related Evidence (if applicable):
  - Algorithm/mathematical formulation quotes
  - Implementation details
  - Missing literature citations

b) Experiment-related Evidence (if applicable):
  - Dataset/setup descriptions
  - Results and metrics
  - Missing specifications

3. Literature Gap Analysis:
  - Missing citations list
  - Required baseline references

4. Validation Analysis:
  - Primary evidence summary
  - Supporting quotes
  - Impact assessment

5. Conclusion:
  - Validity status: [Valid/Partially Valid/Invalid]
  - Confidence level: [High/Medium/Low]
  - Key supporting evidence

Remember:
1. Every conclusion must be supported by specific evidence from the paper
2. Maintain objectivity and thoroughness in analysis
3. Document clear reasoning for all assessments

Figure 7: System prompt used to guide Gemini-2.0-Thinking-Flask in the Reliability Verification stage. It instructs the model to systematically analyze each review comment and find supporting evidence from the original paper.

Meta-Review From DeepReviewer-14B

Summary:
This paper introduces a novel framework for automating the research lifecycle using large language models (LLMs). The framework aims to mimic the iterative process of research, review, and refinement. The core of this framework consists of two models: CycleResearcher, a policy model for generating research papers, and CycleReviewer, a reward model that simulates peer reviews. The authors use an iterative reinforcement learning approach, specifically SimPO, to refine both models. To facilitate training and evaluation, the authors created two large-scale datasets, Review-5k and Research-8k. Empirical results show that CycleResearcher generates papers with an average quality comparable to human-written preprints, achieving a 31.07\% acceptance rate. CycleReviewer shows a 26.89\% improvement in mean absolute error (MAE) compared to individual reviewers. The contribution lies in the application of LLMs to automate the entire research lifecycle, the introduction of large-scale datasets, and the demonstration of a functional iterative framework. This work can impact the way research is conducted, particularly in machine learning. The paper raises questions about potential biases, computational costs, and ethical implications.

Strength:
The paper's core strength is its innovative approach to automating the research lifecycle using LLMs. The CycleResearcher and CycleReviewer framework effectively models the iterative process of research, review, and refinement. The use of SimPO is a notable technical contribution. The introduction of Review-5k and Research-8k datasets is a valuable resource. The CycleResearcher model generates papers with an average quality level close to human-written preprints, achieving a 31.07\% acceptance rate. The CycleReviewer model's 26.89\% improvement in MAE compared to individual reviewers indicates potential for automated research assessment. The inclusion of human evaluation and objective model-based evaluations provides a comprehensive assessment. The paper's exploration of ethical considerations and the implementation of Fast-DetectGPT demonstrates a responsible approach. The authors advocate for disclosure of LLM use in research, which is a positive step towards transparency.

Weaknesses:
- The paper does not adequately address the potential for bias in the training data for both CycleResearcher and CycleReviewer. There's a risk that the models perpetuate existing biases. The paper describes the data collection process for Review-5k and Research-8k, but lacks any discussion of potential biases within these datasets. The lack of analysis of the diversity of authors, institutions, or research topics is a significant concern.
- The paper lacks a detailed discussion on computational resources for training and deploying the framework. The use of large language models and iterative refinement processes can be computationally intensive. While the paper mentions the hardware used, it does not provide explicit training times per epoch or iteration, nor does it provide inference times.
- The paper does not sufficiently explore the potential for misuse of the framework. The paper introduces Fast-DetectGPT and discusses ethical considerations, but lacks mechanisms beyond the RL framework to guarantee the factual correctness or scientific validity of the generated content.
- The paper could provide more details on the specific prompts used for the LLMs and the criteria for evaluating the quality of generated papers. The paper describes the stages of paper generation but does not provide the exact prompts used for each stage. Similarly, it mentions the criteria used by CycleReviewer, but not the specific prompts.
- The paper could benefit from a more detailed discussion of the limitations, such as the potential for generating derivative or low-quality research. It is unclear how the framework addresses the risk of converging to safe or incremental ideas.
Finally, the paper could explore the generalizability of the framework across different research domains. The current experiments seem to focus on machine learning and computer science, and it is unclear how the framework would perform in other fields.

Suggestions:
- Implement a more rigorous analysis of training datasets for bias. Use stratified sampling or re-weighting. Explore methods for detecting and reducing bias in model outputs. Include a qualitative analysis of the generated outputs. Release training data for independent scrutiny.
- Provide a detailed breakdown of hardware and software requirements. Explore methods for reducing the computational cost. Provide a comparison of the computational resources required. Consider providing a cloud-based deployment option.
- Implement robust safeguards to ensure the quality and integrity of generated research outputs. Use automated quality checks, and implement a human-in-the-loop review process. Explore methods for detecting and preventing malicious use. Clearly communicate the limitations of the framework.
- Provide a detailed breakdown of the prompts used for each stage of the research lifecycle, including examples. Clearly define evaluation criteria. Improve reproducibility and transparency.
- Conduct a more thorough analysis of the limitations, regarding the potential for generating derivative or low-quality research. Explore methods for mitigating the risk of converging to safe or incremental ideas.
- Investigate the generalizability of the framework across different research domains. Conduct experiments in diverse fields and analyze the performance in each domain.

Questions:
- How does the framework handle the generation of novel research ideas that are not present in the training data?
- What are the computational costs associated with training and running the CycleResearcher and CycleReviewer models?
- How does the framework ensure the ethical use of generated research and prevent potential misuse?
- Can the framework be adapted to other research domains beyond machine learning and computer science?
- What specific prompts are used to guide the LLMs in each stage of the research lifecycle?
- What metrics or rubrics are used to assess the novelty, validity, and significance of the generated research?

Soundness:	**3.0**
Presentation:	**3.0**
Contribution:	**3.0**
Confidence:	**5.0**
Rating:	**6.0**

Figure 8: The Meta-Review comment for CycleResearcher from DeepReviewer-14B

Official Review of Submission489 by Reviewer 7LzG

Official Review by Reviewer 7LzG 08 Nov 2024, 17:23 (modified: 13 Nov 2024, 00:01) Everyone Revisions (/revisions?id=oehQnbB57h)

Summary:
The paper explores the use of open-source large language models to automate the entire research process, from literature review and manuscript preparation to peer review and revision. The proposed framework includes CycleResearcher, which performs research tasks, and CycleReviewer, which simulates the peer review process. The study demonstrates that CycleReviewer can outperform human reviewers in predicting paper scores, and CycleResearcher can generate papers of quality close to human-written preprints. The models are trained using two new datasets, Review-5k and Research-8k, which capture the complexities of peer review and research paper generation. The results indicate that this approach can significantly enhance the efficiency and quality of scientific research, while also providing ethical safeguards to prevent misuse.

Soundness: 2: fair
Presentation: 3: good
Contribution: 2: fair
Strengths:
The introduction of CycleResearcher and CycleReviewer models to automate the entire research process, including literature review, manuscript preparation, peer review, and revision, is highly innovative. This framework mimics the real-world research cycle, enhancing the efficiency and consistency of scientific inquiry. Performance Improvement: The CycleReviewer model demonstrates a significant improvement in predicting paper scores, outperforming human reviewers by 26.89% in mean absolute error (MAE). This indicates that the model can provide more accurate and consistent evaluations than individual human reviewers. Quality of Generated Papers: The CycleResearcher model generates papers with an average quality close to human-written preprints, achieving an acceptance rate of 31.07%. This shows that the model can produce high-quality research outputs that are competitive with human-generated content. Large-Scale Datasets: The development of the Review-5k and Research-8k datasets, which capture the complexities of peer review and research paper generation, provides valuable resources for training and evaluating models in academic paper generation and review.

Weaknesses:
Generalizability Across Domains: The models are primarily designed for machine learning-related research, and their generalizability to other scientific fields remains unexplored. This limitation suggests that the framework might not perform as well in domains outside of machine learning. Reward Design: The paper highlights the issue of reward definition, where the policy model might exploit loopholes in the reward model to maximize rewards without genuinely improving the quality of the generated research. This behavior could undermine the long-term goal of producing high-quality research outputs. Complexity of Implementation: Implementing the framework requires significant computational resources and expertise in reinforcement learning and LLMs. This complexity might be a barrier for widespread adoption, especially for smaller research teams or institutions with limited resources.

Questions:
The framework is primarily designed for machine learning-related research. How do you envision adapting CycleResearcher and CycleReviewer to other scientific fields, such as biology or social sciences, where the nature of research and evaluation criteria might differ significantly? The paper mentions the potential issue of reward hacking, where the policy model might exploit loopholes in the reward model. Could you elaborate on the specific strategies you are considering to mitigate this issue and ensure that the generated research outputs maintain high academic rigor and novelty?

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold
Confidence: 3: You are fairly confident in your assessment. It is possible that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work. Math/other details were not carefully checked.
Code Of Conduct: Yes

Figure 9: The Real-world review comment for CycleResearcher

Official Review of Submission489 by Reviewer CzSX

Official Review by Reviewer CzSX 06 Nov 2024, 18:28 (modified: 03 Dec 2024, 03:19) Everyone Revisions (/revisions?id=m1YfqmaJLl)

Summary:
The paper presents CycleResearcher and CycleReviewer, which is a cohesive system intended to make steps towards automatic scientific discovery. In particular the novelty of the approach lies in the encapsulation of the entire research pipeline from research to review, in order to better model the entire system of research generation and get better outcomes. The authors contribute two datasets for research and review, and use these to train the system of researcher and reviewer, and evaluate them using various methods and metrics.

Soundness: 3: good
Presentation: 2: fair
Contribution: 2: fair
Strengths:
Originality: The idea to design both a researcher and a reviewer is novel and interesting.

Quality: The usage of recent preference optimization methods is a nice technical plus. The work contributes datasets to the direction of scientific peer reviewing, which is a resource that is rather helpful for the field. RL details and how they fit in is nice.

Clarity: Figures are well-designed and artistically pleasing. Appreciate the various different ways that are used to evaluate the methods (qualitative, ablations, etc.)

Significance: The automation of scientific research and reviewing is a very interesting and timely topic. In particular, due to the massive increase in submissions year-to-year, progress towards the paper's direction is well appreciated.

Weaknesses:
Originality: N/A

Quality: One big issue of the paper is the method in which the authors obtain the "ground truth" review score: "for each submission, we use the average of the other n − 1 reviewers' scores as an estimator of the true score." In my opinion (and what feels like a general consensus in the community), it's pretty clear that this isn't the correct approach in determining a ground truth quality of a paper. Different reviewers have different expertises and opinions, and may disagree substantially based on their backgrounds, but this is a positive quality of peer review rather than a negative one. Thus, the metric used to judge the "loss" of a review score can be used to train proxies of reviewers, sure, but it does not make sense to then take the trained system and use the same metric to compare it against actual human reviewers. For instance, if human reviewers vary differently based on their perspectives and CycleReviewer is just doing some "hedge" where most scores are around the median score for all papers, it might achieve better than human performance on the MAE metric that is used in the evaluation. Furthermore, focusing on the score ignores perhaps the more important points of paper reviewing, such as being able to highlight errors in the paper or provide advice for making changes that are adopted in future versions. I think the true objective of reviewing in the paper's cycle paradigm matches these objectives more, although I recognize that they are even harder to quantify. Even so, I think the paper is overclaiming by saying that the lower MAE suggests that "LLMs can surpass expert-level performance in research evaluation".

And since I don't necessarily agree with the evaluation metric for the reviewer, this casts doubt on the results for the CycleResearcher because the CycleReviewer is reviewing the CycleResearcher. Also, since CycleResearcher is optimized on CycleReviewer, then saying that CycleResearcher does better on CycleReviewer than humans or AI scientist doesn't mean much. The qualitative study in section 4.3 is helpful to remove some of these doubts though.

Clarity: There are a reasonable amount of typos and grammatical errors in the document. For instance, CycleReviewer is replaced with WhizeReviewer in Section 4.1. The title of section 4.1 should be "Experiments on Paper Review Generation", etc. The paper would benefit from a pass over to correct grammatical mistakes in general to make it easier to read.

Significance: The claims are very catchy that the system can generate better reviews and papers than humans. However, given the questionable-ness of the metric, I think these are discounted to a degree.

Questions:
See Weaknesses.

Flag For Ethics Review: No ethics review needed.
Rating: 6: marginally above the acceptance threshold
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Figure 10: The Real-world review comment for CycleResearcher

Official Review of Submission489 by Reviewer GAvj

Official Review by Reviewer GAvj 03 Nov 2024, 03:16 (modified: 03 Dec 2024, 04:32) Everyone Revisions (/revisions?id=zZrbn1zK3H)

Summary:
The paper introduces an iterative training framework for automatic generation and review of research papers using open-source LLMs. The core of their approach consists of two main components:

1. CycleResearcher: A policy model that generates the paper, prompted by abstracts of related work.
2. CycleReviewer: A reward model that writes several peer reviews and returns scores according to ICLR criteria.

The authors initialize these models by supervised fine-tuning on scraped conference papers and ICLR reviews. They then improve the CycleResearcher using reinforcement learning (specifically iterative Simple Preference Optimization, SimPO), using CycleReviewer as a reward model.

The paper claims three main contributions:

1. Development of an iterative reinforcement learning framework that mirrors the real-world research-review-revision cycle.
2. Creation of two new datasets: Review-5k. Research-8k
3. Empirical results showing:
   - CycleReviewer produces scores that are closer to averages of multiple human reviewers than scores by individual human reviewers
   - CycleResearcher-12B achieved paper quality scores surpassing preprint level and approaching accepted paper level

The paper implements some ethical safeguards: they train a model to detect papers generated by LLMs they publish; they promise to implement a licensing agreement such that downloading model weights requires sharing institutional affiliations and agreeing not to use models for official peer reviews or submissions without disclosure.

Soundness: 2: fair
Presentation: 2: fair
Contribution: 2: fair
Strengths:
- Training LLMs with reinforcement learning on parts of the AI research process is a novel and significant contribution.
- The paper includes numerous experiments and ablations. The overall methodology is sound (with exceptions, see weaknesses).
- The authors achieve strong results on the metrics they choose. It is somewhat impressive that their system achieved an acceptance rate of 31.07%, similar to ICLR 2024's acceptance rate.
- Authors use open-source models with a large range of scale (from 12B to 123B).

Weaknesses:
- The writing is overclaiming the extent to which the paper covers the full research process. Authors write that the paper "explores performing the *full* cycle of automated research and review", however the paper omits crucial part of the process: actually running experiments. Instead, the authors train models to write complete papers purely from abstracts of past work, with completely hallucinated experiment design and results.
- I do not think that the task authors train models for — hallucinating experiment results and writing papers for them — is well motivated. Using models for this purpose will not contribute real knowledge to the scientific field. I think this is dual use technology, if not a completely malicious one. I could imagine the paper could be reframed to center on demonstrating this imminent failure of the reviewing system and raising an alarm, allowing the scientific community to adapt. In current form, the paper is probably net-negative.
- Automated evaluation of papers produced by CycleResearcher is hard to trust, since CycleResearcher was trained with RL against the same reward model as used at test-time. Reward model overoptimization (Gao et al, 2022 - https://arxiv.org/abs/2210.10760 (https://arxiv.org/abs/2210.10760)) should be the expected result of RL, however the authors do not run any experiments to investigate to which extent their evaluation is influenced by this. For example, the authors could train a held-out reward model on a held-out dataset of reviews and then evaluate CycleResearcher on both the reward model used for RL training and this new held-out reward model.
- The claim that CycleResearcher surpasses the quality of preprint papers and approaches quality of accepted papers is not well supported, due to the concerns about reward model overoptimization mentioned above. Human reviewers rate CycleResearcher's papers significantly lower (4.8) than the automated reviewer made by the authors (5.36). The authors could have reported the actual historical average score of ICLR2024 accepted papers.
- I have a number of concerns about the human evaluation procedure.
  - When the authors evaluate their CycleResearcher with the AI Scientist, they seem to only use rejection sampling (best of N) for CycleResearcher. This is not a fair comparison.
  - Overall, human evaluation is conducted on a small scale (10 papers total, three human reviewers, 2 methods: this paper and baseline)
  - I do not think 30min per review (including reading, writing comments & providing scores) is enough!
- I think it's misleading to use the term "revision" for parameters updates of the policy model (Figure 2). The paper refers to this revision as part of the full research process ("Research-Rebuttal-Revision") but this does not actually involve revision of papers based on reviews.

Questions:
1. What exactly are the prompts, based on which CycleResearcher generates papers during evaluations?
2. Why do smaller CycleResearcher models get better scores in the evaluation?
3. How many samples in automated evaluation?
4. Please include the average real score of accepted papers given by human ICLR2024 reviewers.
5. How do you compute the acceptance rates, e.g. one mentioned in line 128?
6. For human evaluation:
   1. Please report the N used in best-of-N / rejection sampling.
   2. Please clarify whether each paper is evaluated by one or several humans.
   3. How are the human experts chosen?
   4. What do you mean by saying "excluding formatting considerations" in the assessment, and why is it omitted?

Flag For Ethics Review: Yes, Potentially harmful insights, methodologies and applications
Details Of Ethics Concerns:
I do not think that the task authors train models for — hallucinating experiment results and writing papers for them — is well motivated. Using models for this purpose will not contribute real knowledge to the scientific field. I think this is dual use technology, if not a completely malicious one. I could imagine the paper could be reframed to center on demonstrating this imminent failure of the reviewing system and raising an alarm, allowing the scientific community to adapt. In current form, the paper is probably net-negative. I think the results from this paper should be known to the broad public, but not in the current framing.

Rating: 6: marginally above the acceptance threshold
Confidence: 4: You are confident in your assessment, but not absolutely certain. It is unlikely, but not impossible, that you did not understand some parts of the submission or that you are unfamiliar with some pieces of related work.
Code Of Conduct: Yes

Figure 11: The Real-world review comment for CycleResearcher

Official Review of Submission489 by Reviewer 5wHA

Official Review by Reviewer 5wHA 01 Nov 2024, 11:36 (modified: 25 Nov 2024, 03:18) Everyone Revisions (/revisions?id=VfLBrP3ExO)

Summary:
The authors introduce two core components: CycleResearcher, a policy model that autonomously performs research tasks, and CycleReviewer, a reward model that simulates the peer review process. Experimental results suggest that CycleReviewer can outperform individual human reviewers in scoring consistency, and CycleResearcher shows promise in generating research papers that approach the quality of human-written preprints.

Soundness: 3: good
Presentation: 4: excellent
Contribution: 3: good
Strengths:
Valuable Datasets: The introduction of the Review-5k and Research-8k datasets could be highly beneficial to the research community. These datasets provide resources for training and evaluating models in academic paper generation and review, potentially fostering further advancements in automated research tools.

Innovative Use of Preference Data: Utilizing preference data to iteratively train the CycleResearcher model is an interesting approach. This method allows the model to improve over multiple iterations, aligning more closely with human standards through reinforcement learning.

Ethical Safeguards: The inclusion of a detection model to identify AI-generated papers addresses ethical concerns related to the misuse of automated research tools. By implementing such safeguards, the authors demonstrate a commitment to responsible AI deployment.

Automation of the Research Lifecycle: The paper attempts to automate the full research cycle, from idea generation to peer review and revision. This holistic approach is ambitious and, if successful, could significantly impact the efficiency of scientific research.

Weaknesses:
Quality of Generated Papers: Upon examining the samples provided in the Appendix (Sections E.1 and E.2), it is evident that the generated papers contain hallucinations and inaccuracies. For instance, in the generated abstracts, there are claims of outperforming state-of-the-art methods without substantial evidence or appropriate citations. This raises concerns about the reliability of the CycleResearcher model in producing high-quality, factual research papers.

Counterintuitive Results with Model Scaling: In Table 3 (Section 4.2), the CycleResearcher-12B model achieves a higher acceptance rate than the larger 72B and 123B models. This is counterintuitive, as larger models typically perform better due to increased capacity. The paper does not provide sufficient analysis or explanations for this phenomenon, leaving readers questioning the scalability and efficacy of the approach.

Insufficient Ethical Considerations: While the authors mention the implementation of a detection tool for AI-generated papers, the paper lacks a deep exploration of the ethical implications of automating research. Issues such as accountability, potential misuse, and the impact on the scientific community are not thoroughly addressed. A dedicated discussion in the Ethics Considerations section would strengthen the paper.

Questions:
Explanation for Performance of Smaller Models: In Table 3, why does the CycleResearcher-12B model receive the highest acceptance rate compared to the 72B and 123B models? This result is unexpected given that larger models generally have better performance. Could the authors provide an analysis of this outcome, possibly including case studies or error analysis to understand the limitations of larger models in this context?

Evaluation Stability of CycleReviewer: What is the temperature setting used for the CycleReviewer during evaluation? Additionally, have the authors experimented with running the CycleReviewer multiple times to assess the variability or deviation in the review scores and feedback? Understanding the stability and consistency of the CycleReviewer is important for gauging its reliability in the automated review process.

Addressing Hallucinations in Generated Papers: Given the observed hallucinations and inaccuracies in the sample generated papers (Appendix E), what strategies do the authors propose to mitigate these issues? Are there mechanisms in place to fact-check or verify the content produced by the CycleResearcher before it is submitted for automated review?

Flag For Ethics Review: Yes, Discrimination / bias / fairness concerns, Yes, Privacy, security and safety
Details Of Ethics Concerns:
Accountability and Authorship: If AI systems generate research papers, questions arise regarding authorship and accountability for the content. It's essential to clarify who is responsible for the work produced and how credit should be assigned.

Quality and Integrity of Research: The presence of hallucinations and factual inaccuracies in AI-generated papers could undermine the integrity of scientific literature. There is a risk of disseminating false information, which could have downstream effects if other researchers build upon flawed results.

Misuse of Technology: The tools developed could be misused to generate large volumes of low-quality or misleading research, potentially cluttering academic discourse and making it harder to identify valuable contributions.

Impact on the Research Community: Automation might affect the roles of researchers, peer reviewers, and the collaborative nature of scientific inquiry. There is a need to consider how these technologies will coexist with human efforts and what support structures are necessary to ensure they augment rather than hinder scientific progress.

Rating: 8: accept, good paper
Confidence: 5: You are absolutely certain about your assessment. You are very familiar with the related work and checked the math/other details carefully.
Code Of Conduct: Yes

Figure 12: The Real-world review comment for CycleResearcher