

# ReviewAgents: Bridging the Gap Between Human and AI-Generated Paper Reviews

Xian Gao<sup>1</sup>, Jiacheng Ruan<sup>1</sup>, Zongyun Zhang<sup>1</sup>, Jingsheng Gao<sup>1</sup>, Ting Liu<sup>1</sup>, Yuzhuo Fu<sup>1</sup>

<sup>1</sup>Shanghai Jiao Tong University

## Abstract

Academic paper review is a critical yet time-consuming task within the research community. With the increasing volume of academic publications, automating the review process has become a significant challenge. The primary issue lies in generating comprehensive, accurate, and reasoning-consistent review comments that align with human reviewers' judgments. In this paper, we address this challenge by proposing **ReviewAgents**, a framework that leverages large language models (LLMs) to generate academic paper reviews. We first introduce a novel dataset, **Review-CoT**, consisting of 142k review comments, designed for training LLM agents. This dataset emulates the structured reasoning process of human reviewers—summarizing the paper, referencing relevant works, identifying strengths and weaknesses, and generating a review conclusion. Building upon this, we train LLM reviewer agents capable of structured reasoning using a relevant-paper-aware training method. Furthermore, we construct **ReviewAgents**, a multi-role, multi-LLM agent review framework, to enhance the review comment generation process. Additionally, we propose **ReviewBench**, a benchmark for evaluating the review comments generated by LLMs. Our experimental results on ReviewBench demonstrate that while existing LLMs exhibit a certain degree of potential for automating the review process, there remains a gap when compared to human-generated reviews. Moreover, our ReviewAgents framework further narrows this gap, outperforming advanced LLMs in generating review comments.

## 1 Introduction

Peer review of academic papers is a crucial component of the scholarly publishing system, as it ensures the quality of scientific research and facilitates the improvement of academic writing. However, with the rapid increase in the number of aca-

ademic paper submissions, the traditional peer review process is confronted with challenges such as inefficiency and a limited number of reviewers, resulting in restricted feedback for authors and hindering the timely acquisition of review comments.

The reasoning capabilities of LLMs for complex tasks render them potentially valuable for providing feedback on academic papers (Liu and Shah, 2023; Zhao et al., 2024; Zhuang et al., 2025). This, to some extent, alleviates the pressure on reviewers and allows authors to independently use LLMs to review and revise their papers prior to formal submission. However, existing research and applications primarily focus on prompting LLMs to directly generate comments on the submitted papers (Zhou et al., 2024; D'Arcy et al., 2024), which greatly simplifies the review process. This approach does not align with the cognitive process of human reviewers in real-world peer review, nor with the multi-step, multi-role nature of the peer review process. As a result, comments generated directly by LLMs are often difficult to align with the review comments provided by human reviewers. In fact, when human reviewers write their reviews, they are typically required to first summarize the main content of the paper, followed by enumerating its strengths and weaknesses, and ultimately reaching a review conclusion. After multiple reviewers have written their comments, the area chair (AC) synthesizes the feedback from all reviewers and provides a final summary as a meta-review. This process ensures that each reviewer fully understands the paper's content, guaranteeing the accuracy, objectivity, and fairness of the review outcomes while minimizing the influence of individual reviewer biases on the final review comments.

To bridge the gap between LLM-based reviews and human review behavior, we have developed a multi-agent review framework, named **ReviewAgents**, which emulates the multi-step reasoning pro-

cess of human reviewers. To train our agents, we collected a substantial amount of review comments from publicly available review platforms and structurally transcribed them, constructing the Review-CoT dataset, which guides the training of both reviewer agents and area chair agents. Specifically, Review-CoT contains 37,403 papers and 142,324 corresponding review comments and meta-reviews, collected from open peer review platforms. We utilized state-of-the-art large language models to transcribe the review comments into a stepwise reasoning format, aligning them with the cognitive process of human reviewers. Considering the timeliness of novelty assessments in review comments, we incorporated references to the most relevant papers, published up to the point of submission, into the dataset, and employed a relevant-paper-aware training method. This approach mirrors the process by which human reviewers evaluate novelty by retrieving relevant papers. The dataset is used to train both the reviewer and area chair agents. Upon completion of training, we employ ReviewAgents, a multi-agent framework, to simulate the review process by human reviewers and area chairs, utilizing multi-agent, multi-step review procedures to generate the final review comments. Unlike previous approaches, our ReviewAgents framework leverages the knowledge of relevant papers and fully emulating the paper evaluation process conducted by human experts.

To comprehensively assess the quality of generated review comments and their alignment with human reviewers’ feedback, we propose ReviewBench, a benchmark specifically designed to evaluate the quality of automatically generated review comments. ReviewBench includes the latest papers and review comments from open platforms, ensuring that its data is not included in the pretraining datasets of most existing LLMs. We evaluate the capabilities of large language models in generating paper review comments across four dimensions.

The main contributions of our work are summarized as follows:

1. We introduce the Review-CoT dataset, which, to the best of our knowledge, is the first and largest dataset to simulate the human review thought process and includes relevant papers.
2. We propose ReviewBench, a benchmark for the quantitative evaluation of review comments generated by LLMs.

3. By employing a relevant-paper-aware training method on the Review-CoT dataset, we construct ReviewAgents, a multi-step, multi-role framework for paper review that simulates the human review process. Experimental results demonstrate that our approach generates review comments closely aligned with those of human reviewers.

## 2 Related Work

### 2.1 Large Language Models for Paper Review

Large language models (LLMs) have demonstrated significant potential in reviewing and comprehending complex texts (Liu and Shah, 2023; Zhao et al., 2024; Zhuang et al., 2025; Zhu et al., 2024). Initial studies, which compared reviews generated by human reviewers and LLMs for academic papers, indicated that LLMs can produce comments with substantial overlap with those of human reviewers, contributing effectively to the peer review process (Robertson, 2023; Liang et al., 2023). However, further research has shown that, despite the advanced capabilities of LLMs such as GPT-3.5 and GPT-4 in scoring input paper texts, the generated reviews still fall short of fully meeting human expectations (Zhou et al., 2024). Researchers have fine-tuned LLMs to accommodate review requirements by constructing datasets from publicly available review comments (Kang et al., 2018; Yuan et al., 2021; Shen et al., 2022; Dycke et al., 2023; Gao et al., 2024; Du et al., 2024). Additionally, some studies have employed multi-turn dialogues (Tan et al., 2024) or prompted multiple LLMs to construct review systems (D’Arcy et al., 2024), providing comprehensive feedback on the entire text. AgentReview (Jin et al., 2024) attempts to simulate the interactions among different roles involved in the peer review process, generating review comments and discussion dialogues for each role. However, such interactions do not necessarily occur in real-world review scenarios, and its approach to generating review comments does not adhere to the cognitive patterns typically followed by human reviewers. In contrast to existing work, our approach aims to reconstruct the complete reasoning and evaluation process throughout the peer review workflow. We emphasize the in-depth reflection and multi-step, multi-role nature of real-world peer reviewing, thereby mitigating biases inherent in single-LLM-generated reviews and achieving better alignment with human reviewer expectations.

## 2.2 Chain-of-thought Reasoning in Large Language Models

Through chain-of-thought prompting (Wei et al., 2022), LLMs are guided with a step-by-step reasoning process, enabling them to break down complex problems into a series of incremental reasoning steps (Chu et al., 2024). This decomposition enhances LLMs’ performance in reasoning tasks, advancing structured prompting techniques (Qiao et al., 2024; Cesista, 2025). Recent studies highlight that the systematization and structuring of reasoning significantly affect LLMs’ performance, particularly their ability to perform multi-stage reasoning and identify key tasks at each stage. This can be achieved through independent language reasoning systems (Zhong et al., 2024) or supervised fine-tuning (Xu et al., 2025). In this paper, we introduce structured reasoning to automated peer review, aligning LLM behavior with human review practices by dividing the review process into a three-stage structured procedure, improving review generation outcomes.

## 3 Dataset and Benchmark

### 3.1 Structured Thinking in Review

Given that paper review requires models to thoroughly read the manuscript, integrate external knowledge, and conduct an in-depth analysis of the paper’s strengths and weaknesses before assigning scores, a detailed stepwise reasoning process is essential. Inspired by the cognitive process of human reviewers, we decompose the review process of the intelligent review agent into three structured stages: **Summarization, Analysis, and Conclusion**, thereby aligning the automated review procedure with human-like reasoning. In the summarization stage, the agent is provided with the full text of the paper and tasked with concisely summarizing its main contributions and methodologies. During the analysis stage, the agent examines the strengths and weaknesses of the paper based on its content and relevant literature within the same domain, serving as the foundation for subsequent evaluation. In the conclusion stage, drawing upon the outcomes of the previous two stages, the agent formulates a final review verdict for the paper.

To facilitate this structured stepwise reasoning, we employ dedicated tags (e.g., `<SUMMARY>...</SUMMARY>`) to explicitly mark the beginning and end of each stage. These tags enable the model to maintain clarity throughout the entire

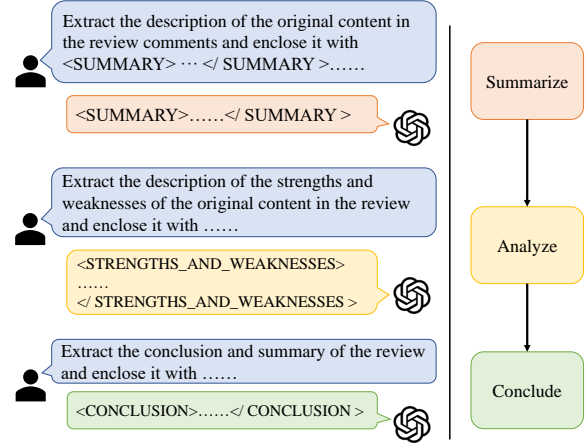


Figure 1: The process of transcribing the collected review comments using advanced LLMs according to a pre-defined structured thinking process

reasoning process.

### 3.2 Dataset Preparation

We have defined a standardized data format for Review-CoT. Each data entry contains the following key information: a unique identifier, paper title, abstract, full-text content, submission conference, reviewer comments, Program Chair’s remarks, and the decision regarding paper acceptance. Additionally, the data records the titles and abstracts of the most relevant papers in the field up to the point of submission. With this information, the dataset samples we construct can be aligned as closely as possible with the information typically accessed by human reviewers during the review process, enabling the trained models to generate conclusions that are more consistent with those of human reviewers.

#### 3.2.1 Metadata Collection

We crawled ICLR papers from 2017 to 2024 and NeurIPS papers from 2016 to 2024 from OpenReview<sup>1</sup> and NeurIPS Proceedings<sup>2</sup>. For each paper, we adhere to the standardized data format we have defined, retaining as much metadata information as possible. Each paper includes the full text as well as the corresponding review comments. For the paper content, we use Scipdf\_Parser<sup>3</sup> to parse the PDFs into structured JSON files.

<sup>1</sup><https://openreview.net/>

<sup>2</sup><https://papers.nips.cc/>

<sup>3</sup>[https://github.com/titipata/scipdf\\_parser](https://github.com/titipata/scipdf_parser)

### 3.2.2 Transcription

An examination of review comments within the collected metadata revealed that recent comments already follow a predefined structured review process, including a paper summary, analysis of strengths and weaknesses, and a final conclusion. While earlier comments contained these elements, they lacked explicit structural divisions. To address this, we employed advanced LLMs to transcribe the review comments according to the structured reasoning process (Figure 1). This transcription reorganizes the paragraphs while preserving the original content and semantics, adding dedicated tags (e.g., <SUMMARY>...</SUMMARY>). This process transforms the review comments into a unified format suitable for model training.

### 3.2.3 Relevant Paper Retrieval

Human reviewers typically reference relevant literature to understand the field’s current state to assess a paper’s novelty, ensuring the timeliness of their review comments. The review comments in Review-CoT dataset, which evaluate content and novelty, reflect the reviewer’s judgments during the review process. These judgments must be compared with the latest research to maintain timeliness, as outdated reviews cannot be generalized across time without considering recent findings. To enhance the reviewer’s thought process and ensure timely comments, we use the Semantics Scholar API<sup>4</sup> to retrieve papers in the same research domain as the reviewed paper, up to the submission date. We extract titles and abstracts from these papers and incorporate them into the dataset as references to assess the paper’s novelty. Both the paper itself and the titles and abstracts of the retrieved relevant papers are encoded as vectors, and their similarity is computed. We retain the two papers with the highest similarity as the most relevant references for the reviewer during the review process.

### 3.2.4 Dataset Details

We have collected a total of 142,324 review comments for 37,403 papers from publicly available review platforms. Each review comment was transcribed into a structured format with appropriate delimiters added. For each paper, we included the titles and abstracts of the two most relevant papers, published up to the submission date, as references for the review process.

	# Papers	# Reviews	Meta Review	Review Process	Related Papers
PeerRead (Kang et al., 2018)	3,006	10,770	✗	✗	✗
ASAP-Review (Yuan et al., 2021)	8,877	28,119	✗	✗	✗
MReD (Shen et al., 2022)	7,894	30,764	✓	✗	✗
NLPeer (Dycke et al., 2023)	5,672	11,515	✗	✗	✗
ReviewCritique (Du et al., 2024)	120	440	✗	✓	✗
Review2 (Gao et al., 2024)	27,805	99,729	✓	✗	✗
Review-CoT (ours)	<b>37,403</b>	<b>142,324</b>	✓	✓	✓

Table 1: Comparison between Review-CoT and existing review comment datasets

For the training dataset of the Review Agent, the model’s input consists of the original paper and the two relevant papers, with the output being the formatted review comments. For the Area Chair Agent’s training dataset, the model’s input is the output from the Review Agent, and the output is the paper’s meta-review.

Table 1 presents a comparison between our dataset and existing review comment datasets, showing that our dataset outperforms the existing ones in both data volume and diversity.

## 3.3 ReviewBench

Given that the evaluation of LLM-generated review comments is still limited to model scoring and human ratings, with a lack of standardized quantitative assessment criteria, we propose ReviewBench, a benchmark for evaluating LLM-generated review comments across multiple dimensions. Table 2 presents the evaluation dimensions and corresponding metrics included in ReviewBench. Our benchmark encompasses assessments of LLM-generated review comments in terms of language quality, semantic consistency, and sentiment consistency, as well as a Review Arena task to measure the model’s ability to generate review comments. Further details on the evaluation metrics can be found in Appendix A.

### 3.3.1 Language Diversity

The language diversity dimension is evaluated using Distinct (Li et al., 2016) and Inverse Self-BLEU as metrics. **Distinct** measures the proportion of non-redundant n-grams in the generated text relative to the total number of n-grams, reflecting the degree of repetition in the text and serving as an indicator of its diversity. Based on Self-BLEU (Zhu et al., 2018), **Inverse Self-BLEU** is inversely cor-

<sup>4</sup><https://www.semanticscholar.org/product/api>



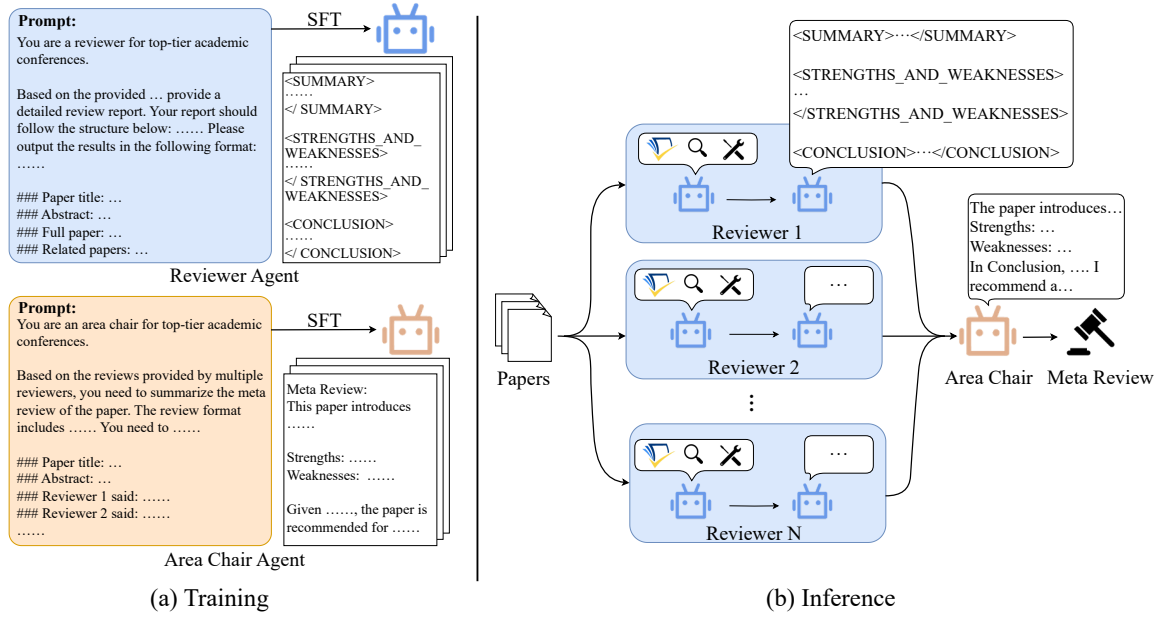


Figure 2: An overview of the ReviewAgents framework

related with the BLEU score between n-grams in the text. The higher the Inverse Self-BLEU score, the more diverse the text generation, facilitating the construction of a composite score in conjunction with other metrics. The calculation formula is given by:  $\text{Inverse Self-BLEU} = 2 - \text{Self-BLEU}$ .

### 3.3.2 Semantic Consistency

Semantic consistency is employed to assess the degree of semantic alignment between the review comments generated by the model and those written by human reviewers, using the ROUGE-1, ROUGE-L, and SPICE metrics. ROUGE (Lin, 2004) is a widely used metric for evaluating the overlap between generated text and reference text. **ROUGE-1** measures the overlap of individual words between the generated and reference texts, while **ROUGE-L** evaluates the longest common subsequence (LCS) between them. The **SPICE** metric (Anderson et al., 2016), originally designed to evaluate the consistency of image captions, measures semantic consistency by comparing part-of-speech and semantic structures, assessing whether the generated text aligns semantically with the reference text.

### 3.3.3 Sentiment Consistency

Sentiment consistency is employed to measure the overall alignment of sentiment and attitudinal orientation between the review comments generated by the model and those written by human reviewers, specifically assessing whether both are predomi-

nantly positive or negative (Wang and Wan, 2018). The sentiment tendencies of both model-generated and human review comments are extracted using sentiment analysis models based on BERT (Devlin et al., 2019) and VADER (Hutto and Gilbert, 2014), respectively. The distance between these tendencies is then evaluated as the sentiment consistency scores, **BERTScore** and **VADER Score**.

### 3.3.4 Review Arena

The open nature of peer review presents significant challenges for automated evaluation. Previous studies have shown that direct scoring methods using LLMs do not align well with human preferences (Liang et al., 2023). In contrast, LLMs perform better in ranking tasks (Li et al., 2024). To obtain reliable evaluations, we propose Review Arena, which employs a tournament-style pairwise evaluation system to compare pairs of review comments. To mitigate the influence of hallucination, we use human review comments as a reference and require LLM judges to rank a pair of model-generated review comments based on their similarity in content and form to the human review comments. Each pair of review comments undergoes two evaluations through reverse sorting to minimize positional bias.

### 3.3.5 Statistical Information

We selected 100 papers from ICLR 2024 and NeurIPS 2024 as the test set for our benchmark. ICLR 2024 and NeurIPS 2024 were published in

Dimensions	Metrics
Language Diversity	Distinct <sub>4</sub>
	Inverse Self-BLEU@4
Semantic Consistency	ROUGE-1
	ROUGE-L SPICE
Sentiment Consistency	BERTScore
	VADER Score
Review Arena	Win Rate

Table 2: Evaluation dimensions and corresponding metrics in ReviewBench

February and September of 2024, respectively, both of which postdate the knowledge cutoff times of most LLMs’ pretraining datasets<sup>5</sup>, ensuring that the test data was not used during the pretraining phase. These papers were removed from the Review-CoT training dataset. To better simulate the real review process, the selected benchmark follows an acceptance-to-rejection ratio of 3:7, which closely approximates the acceptance rate of these two conferences.

## 4 Methodology

### 4.1 Framework Overview

Figure 2 illustrates an overview of the ReviewAgents framework we propose. During the training phase, we employ a training strategy that incorporates awareness of relevant papers and a structured reasoning approach to train the reviewer agent, while using meta-review to train the area chair agent, enabling the model to simulate the human peer review process. In the inference phase, the reviewer, trained with awareness of relevant papers, generates review comments by considering relevant papers retrieved from the literature database. The area chair then synthesizes multiple review comments to form the final review conclusion.

### 4.2 Relevant-paper-aware Training

The timeliness of review comments is reflected in the relevant papers that have been published up until the time of the review. In the Review-CoT training dataset, relevant papers are provided as references for the review process. During training, the model learns how to integrate these relevant papers into the generation of review comments, thereby incorporating timeliness information and closely replicating the reasoning process employed by human reviewers when crafting their feedback.

<sup>5</sup>The knowledge cutoff time for the GPT-4o and Llama-3.1 is October 2023 and December 2023, respectively.

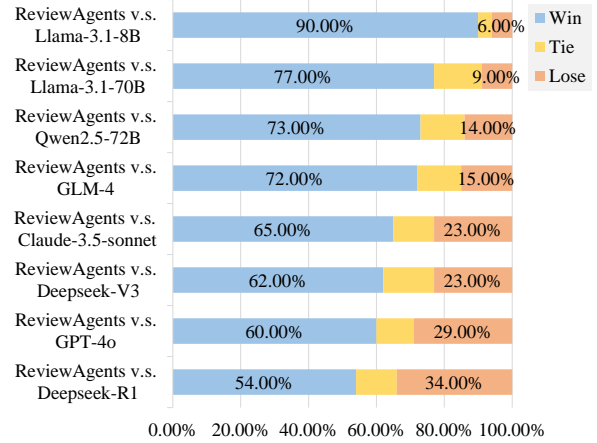


Figure 3: Win rates of ReviewAgents in Review Arena

To ensure that the model accesses the relevant paper information without exceeding the context window, we input the titles and abstracts of the relevant papers as prompts during training.

### 4.3 Multi-step Multi-role Review Generation

After training LLM agents for different roles, we construct a review framework that emulates the human review process. The paper is initially assigned to multiple distinct reviewers, each of whom utilizes a retrieval API to search for relevant papers in the literature database as references for their review. Each reviewer then follows a structured reasoning process in sequence—SUMMARY, ANALYZE, and CONCLUDE—to generate their review comments. Subsequently, the review comments are submitted to the area chair agent for aggregation and synthesis, eliminating any biases present in the individual reviewer’s feedback. The final review comments, provided by the area chair, serve as the ultimate output of the review process.

## 5 Experiments

### 5.1 Experimental Setup

We evaluated our method on ReviewBench, selecting advanced open-source models and closed-source models for comparison, evaluating the meta-reviews generated by different models. The evaluation metrics, Distinct and Inverse Self-BLEU, are computed using 4-grams, denoted as Distinct<sub>4</sub> and Inverse Self-BLEU@4, respectively. For ROUGE-1 and ROUGE-L, their F1-scores are computed. For the diversity and consistency metrics, including Language Diversity, Semantic Consistency, and Sentiment Consistency, we compute their overall scores. Detailed implementation specifics and

Model	Language Diversity ↑		Semantic Consistency ↑			Sentiment Consistency ↑		Overall ↑
	Distinct <sub>4</sub>	Inverse Self-BLEU@4	ROUGE-1	ROUGE-L	SPICE	BERT Score	VADER Score	
Human	98.88	89.84	100.00	100.00	100.00	100.00	100.00	98.39
<i>API-based LLMs</i>								
Claude-3.5-sonnet	98.50	<b>84.39</b>	39.93	16.26	14.78	33.65	86.29	53.40
Deepseek-R1	<b>99.37</b>	77.64	37.22	14.40	13.57	<b>48.13</b>	83.09	53.35
GPT-4o	99.26	82.48	41.35	16.62	14.64	30.55	84.32	52.75
Deepseek-V3	95.51	69.36	38.06	16.57	14.58	30.20	85.70	50.00
GLM-4	89.80	64.33	40.69	<u>18.29</u>	<u>15.33</u>	30.98	<b>86.79</b>	49.46
<i>Open-source LLMs</i>								
Llama-3.1-70B-Instruct	87.58	60.94	40.53	17.83	15.24	35.56	84.02	48.81
Qwen-2.5-72B-Chat	94.72	64.93	29.97	15.56	15.18	26.42	84.66	47.35
Llama-3.1-8B-Instruct	82.98	55.70	30.39	13.90	12.95	26.23	81.00	43.31
Ours	96.57	77.60	<b>42.88</b>	<b>19.27</b>	<b>15.75</b>	<u>44.71</u>	86.26	<b>54.72</b>

Table 3: Comparison of the diversity and consistency metric scores

prompt templates are provided in Appendix B and D.

## 5.2 Main Results

### 5.2.1 Diversity and Consistency

Table 3 presents a comparison of the diversity and consistency metric scores of our ReviewAgents with those of other models on ReviewBench. It is evident that, despite being constrained by the model’s parameter size, the proposed ReviewAgents method exhibits slight limitations in language diversity. However, the review comments generated by ReviewAgents outperform larger parameter open-source models in both semantic consistency and viewpoint alignment, achieving results comparable to or exceeding those of advanced closed-source models. In terms of overall metrics, our approach achieves state-of-the-art results, validating the comprehensiveness and effectiveness of our method.

### 5.2.2 Review Arena Results

Figure 3 illustrates the win rates of ReviewAgents compared to other models in the Review Arena. ReviewAgents achieved a higher win rate than closed-source LLMs and significantly outperformed open-source LLMs, further validating the effectiveness of our approach in generating review comments that align closely with human feedback.

### 5.2.3 Human Evaluation Results

We engaged five AI researchers with submission and reviewing experience at top-tier AI conferences such as ICLR, NeurIPS, and \*CL to serve as evaluators. They assessed and scored the review comments generated by different LLM-based methods. The evaluation dimensions included *Soundness &*

*Validity (SV)*, *Clarity & Organization (CO)*, *Constructiveness (C)*, *Consistency with Human Review (CHR)*, and *Position Clarity (PC)*. Each dimension was rated on a scale from 1 to 5. Detailed descriptions of each evaluation dimension can be found in Appendix C.

Table 5 presents the results of the human evaluation. ReviewAgents achieved higher average scores compared to proprietary LLMs, with particularly notable improvements in the dimensions of *Clarity & Organization (CO)* and *Consistency with Human Review (CHR)*. These findings further substantiate that the reviews generated by our approach align more closely with the expectations of human reviewers.

### 5.3 Ablation Study on Number of Reviewer Agents

The number of agents in multi-agent systems directly affects the final outcome (Zhang et al., 2024). To determine the optimal number of reviewers within the ReviewAgents framework, we examined the performance of review comments generated with different reviewer counts. Table 4 shows the impact of varying values of  $N$  on the generated review comments. The results indicate that when  $N = 3$ , the generated review comments align most closely with human reviewers’ feedback. Further experiments reveal that fewer reviewers yield more diverse comments, while an excessive number of reviewers does not improve alignment with human opinions. This may be due to the fact that, in the training dataset, the average number of review comments per paper is between 3 and 4. As a result, too many review comments during inference introduce opinion discrepancies and extended context, negatively affecting the area chair agent’s judgment.

N of Reviewers	Language Diversity ↑		Semantic Consistency ↑			Sentiment Consistency ↑		Overall
	Distinct4	Inverse Self-BLEU@4	ROUGE-1	ROUGE-L	SPICE	BERT Score	VADER Score	
1	<b>97.78</b>	<b>82.29</b>	<b>43.20</b>	<b>19.41</b>	15.66	31.67	84.83	53.55
2	<u>96.69</u>	<u>78.08</u>	<u>42.53</u>	<u>19.37</u>	<b>15.82</b>	44.49	84.31	54.47
3	<u>96.57</u>	<u>77.60</u>	<u>42.88</u>	<u>19.27</u>	<u>15.75</u>	<b>44.71</b>	<b>86.26</b>	<b>54.72</b>
4	<u>96.49</u>	<u>76.33</u>	<u>43.13</u>	<u>19.27</u>	<u>15.71</u>	<u>44.52</u>	<u>86.03</u>	<u>54.50</u>
5	<u>96.16</u>	<u>75.95</u>	<u>42.23</u>	<u>19.04</u>	<u>15.55</u>	<u>42.98</u>	<u>85.85</u>	<u>53.97</u>
6	<u>95.81</u>	<u>75.24</u>	<u>42.25</u>	<u>19.08</u>	<u>15.67</u>	<u>43.39</u>	<u>85.91</u>	<u>53.91</u>

Table 4: Ablation study on the number of reviewer agents

	SV	CO	C	CHR	PC	Average
ReviewAgents	3.46	<b>4.40</b>	3.43	<b>3.99</b>	<u>3.85</u>	<b>3.83</b>
Deepseek-R1	3.65	4.06	3.94	3.94	3.45	3.81
Deepseek-V3	3.61	<u>3.85</u>	3.62	3.82	<b>3.92</b>	3.76
GPT-4o	<b>3.80</b>	3.96	<b>4.00</b>	3.84	2.83	3.69
Claude-3.5-sonnet	3.58	3.82	3.78	3.09	3.72	3.60
Qwen-2.5-72B-Chat	3.48	3.79	3.46	2.76	3.11	3.32
GLM-4	2.80	3.57	3.77	3.07	2.75	3.19
Llama-3.1-70B-Instruct	<u>3.66</u>	3.50	3.23	2.53	3.00	3.18
Llama-3.1-8B-Instruct	2.68	3.32	2.85	2.02	2.57	2.69

Table 5: The results of the human evaluation

## 5.4 Discussion and Case Study

As shown in Table 3, the improvements of ReviewAgents over other methods primarily stem from enhanced semantic and sentiment similarity with human review comments. This improvement is primarily due to the structured generation process of review comments, which mirrors the stages of human peer review. As a result, each LLM stage generates review comments through a reasoning process closely aligned with human reviewers. Additionally, the final meta-review is derived from aggregating comments from multiple reviewer agents, reducing individual reviewer biases.

Figure 4 presents a case study. Compared to comments from advanced LLMs, ReviewAgents’ comments show greater overlap with human reviewers’ comments. For example, both address issues such as insufficient comparison and citation, highlighting the effectiveness of our relevant-paper-aware training approach. While GPT-4o tends to generate more positive comments, ReviewAgents, through stepwise analysis of strengths and weaknesses, produces rejection comments aligned with human reviewers’ assessments. We provide a complete case of the peer review comment generation process in the Appendix E.

## 6 Conclusion

In this paper, we aim to enhance the ability of LLMs to generate review comments, thereby advancing the automation of academic paper review.

<b>Meta review from Human</b> The authors suggest a significant advancement in the field of..... Strength: - the authors report consistent improvements over prior work ..... Weakness: - the evaluation of ..... Additionally, <b>there should be runs with a large number of PGD steps</b> and a large number of restarts to ensure that ..... - the main contribution of the paper is ..... However, <b>there is no comparison (not citation) of a large number of works</b> ..... - Regarding..... <b>All these papers are not cited and not compared to in the submission</b> ..... In total, this papers suggests..... However, <b>the authors missed major prior works</b> ..... Thus <b>this paper is not ready yet for publication</b> .....	
<b>Meta review from GPT-4o:</b> ## Summary The paper presents..... ## Strengths 1. <b>**Innovative Approach**</b> : ..... 2. <b>**Strong Theoretical Basis**</b> : ..... ## Weaknesses 1. <b>**Complexity</b> and Reusability*: ..... 2. <b>**Dependency on Pre-trained Models**</b> : ..... ## Recommendation Overall, the paper presents a noteworthy contribution..... <b>The paper is recommendable for acceptance</b> .....	<b>Meta review from ReviewAgents:</b> The paper ..... introduces a novel approach for ..... Strengths: Novelty and Innovation:..... Robust Theoretical Framework: ..... Weaknesses: <b>Computational Efficiency</b> Concerns: ..... <b>Limited Comparisons with Related Work</b> : ..... Unexplored Theoretical Limitations: ..... In conclusion, the paper presents a significant advancement in the field of ..... However, <b>the complexities inherent in the methodology, computational demands, and the inadequate exploration of related works</b> and theoretical limitations hinder the paper's readiness for acceptance. <b>I recommend a rejection</b> .....

Figure 4: A Case Study. The green highlighted sections represent matching opinions, while the red highlighted sections represent opposing decisions.

To achieve this objective, we collected papers and review comments and constructed the Review-CoT dataset, which includes the original papers, review comments with structured thinking processes, area chair comments, and relevant papers. Building on this, we employed a relevant-paper-aware method for fine-tuning LLMs and introduced ReviewAgents, a multi-agent review system that simulates the human review process, incorporating relevant paper retrieval, structured thinking, and a multi-role, multi-step approach. To comprehensively evaluate the capacity of LLMs to generate review comments, we developed ReviewBench, a benchmark comprising four testing dimensions. Experimental results demonstrate that our ReviewAgents surpass advanced LLMs, such as GPT-4o and Deepseek-R1.



## 7 Limitations

In this section, we discuss some of the limitations of our work.

### 7.1 Limited Research Area

Thanks to the openness of paper reviews in the AI field, our current work constructs datasets using review comments from AI conference papers. As a result, our approach to generating review comments is currently limited to the AI domain. In the future, we plan to collect papers and review comments from additional fields, thereby constructing datasets that encompass a broader range of disciplines, which will enhance the generalizability of our method across diverse research areas.

### 7.2 Disjoint Process for Review Generation

Our approach generates the final review comments through a multi-step generation process. During training, the reviewer agent and area chair agent are trained separately, which results in a disconnection between the training process and the inference phase. This may cause the input distribution for the area chair agent during the inference phase to differ from that during training. A potential improvement could involve performing joint training after separate supervised fine-tuning in both stages.

### 7.3 Limited Context Length

Due to the limitations imposed by the context window of large language models, ReviewAgents is constrained by a maximum paper length. Papers that exceed this length may experience truncation, leading to potential biases in the final review comments. Expanding the context window could result in increased training and inference costs. A potential solution to this issue is the batching of papers for processing.

### Ethical Statement

This study constructs our dataset and benchmark using publicly available research paper data. All data used in this research is sourced from openly accessible academic repositories and databases, and no personal or sensitive information of authors or research participants is involved. Data processing and analysis adhered to the relevant terms of use of these data sources, and no privacy or confidentiality issues arose during the data collection process. The authors declare no conflicts of interest, and all analyses were conducted using automated methods,

with complete transparency in the algorithms and data processing procedures. This research follows the ethical guidelines for the responsible use of data in academic research.

The potential risk of our research is the abuse of automatic peer review, which could lead to a decline in the quality of review comments and irresponsible feedback. We emphasize that our work is not intended to replace human reviewers. Instead, our approach aims to enhance the efficiency of the review process while maintaining scientific rigor and alignment with human evaluation standards, rather than replacing human reviewers. The proposed method is designed to serve as an auxiliary tool—integrating LLM-generated reviews as a component of the overall review pipeline—to offer complementary perspectives to traditional expert assessments. In essence, we are exploring how LLMs can augment, rather than supplant, the irreplaceable expertise and discernment of human reviewers. Additionally, our work can serve as a valuable tool for authors to self-edit their papers and as an aid for human reviewers. The results produced by the model should not be misconstrued as definitive and genuine comments on the respective papers.

All annotators were carefully recruited and reasonably compensated for their contributions. They received substantial remuneration to ensure that their time and expertise were fairly valued. Annotators were fully informed of the research objectives and the intended use of the data. All annotators operated under strict institutional guidelines to ensure that data handling adhered to ethical standards, and confidentiality was maintained throughout the annotation and evaluation processes.

## References

- Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. [SPICE: Semantic Propositional Image Caption Evaluation](#). *Preprint*, arXiv:1607.08822.
- Franz Louis Cesista. 2025. [Multimodal Structured Generation: CVPR’s 2nd MMFM Challenge Technical Report](#). *Preprint*, arXiv:2406.11403.
- Zheng Chu, Jingchang Chen, Qianglong Chen, Weijiang Yu, Tao He, Haotian Wang, Weihua Peng, Ming Liu, Bing Qin, and Ting Liu. 2024. [Navigate through Enigmatic Labyrinth A Survey of Chain of Thought Reasoning: Advances, Frontiers and Future](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long*

- Papers*), pages 1173–1203, Bangkok, Thailand. Association for Computational Linguistics.
- Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. [MARG: Multi-Agent Review Generation for Scientific Papers](#). *Preprint*, arXiv:2401.04259.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. [BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding](#). In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.
- 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, Jiayang Cheng, 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](#). *Preprint*, arXiv:2406.16253.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. [The llama 3 herd of models](#). *Preprint*, arXiv:2407.21783.
- Nils Dycke, Ilia Kuznetsov, and Iryna Gurevych. 2023. [NLPeer: A Unified Resource for the Computational Study of Peer Review](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5049–5073, Toronto, Canada. Association for Computational Linguistics.
- Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. 2024. [Reviewer2: Optimizing Review Generation Through Prompt Generation](#). *Preprint*, arXiv:2402.10886.
- 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*.
- C. Hutto and Eric Gilbert. 2014. [VADER: A Parsimonious Rule-Based Model for Sentiment Analysis of Social Media Text](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1):216–225.
- Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. 2024. [AgentReview: Exploring Peer Review Dynamics with LLM Agents](#). *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](#). In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.
- Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. [A Diversity-Promoting Objective Function for Neural Conversation Models](#). In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.
- Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. 2024. [Chain of Ideas: Revolutionizing Research Via Novel Idea Development with LLM Agents](#). *Preprint*, arXiv:2410.13185.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. 2023. [Can large language models provide useful feedback on research papers? A large-scale empirical analysis](#). *Preprint*, arXiv:2310.01783.
- 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.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Ryan Liu and Nihar B. Shah. 2023. [ReviewerGPT? An Exploratory Study on Using Large Language Models for Paper Reviewing](#). *Preprint*, arXiv:2306.00622.
- Yuxuan Qiao, Haodong Duan, Xinyu Fang, Junming Yang, Lin Chen, Songyang Zhang, Jiaqi Wang, Dahua Lin, and Kai Chen. 2024. [Prism: A Framework for Decoupling and Assessing the Capabilities of VLMs](#). *Preprint*, arXiv:2406.14544.
- 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.
- Zachary Robertson. 2023. [GPT4 is Slightly Helpful for Peer-Review Assistance: A Pilot Study](#). *Preprint*, arXiv:2307.05492.
- Chenhui Shen, Liying Cheng, Ran Zhou, Lidong Bing, Yang You, and Luo Si. 2022. [MReD: A Meta-Review Dataset for Structure-Controllable Text Generation](#). In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2521–2535, Dublin, Ireland. Association for Computational Linguistics.
- Cheng Tan, Dongxin Lyu, Siyuan Li, Zhangyang Gao, Jingxuan Wei, Siqi Ma, Zicheng Liu, and Stan Z. Li. 2024. [Peer Review as A Multi-Turn and Long-Context Dialogue with Role-Based Interactions](#). *Preprint*, arXiv:2406.05688.
- Ke Wang and Xiaojun Wan. 2018. [Sentiment Analysis of Peer Review Texts for Scholarly Papers](#). In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR '18*, pages 175–184, New York, NY, USA. Association for Computing Machinery.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V. Le, and Denny Zhou. 2022. [Chain-of-Thought Prompting Elicits Reasoning in Large Language Models](#). *Advances in Neural Information Processing Systems*, 35:24824–24837.
- Guowei Xu, Peng Jin, Hao Li, Yibing Song, Lichao Sun, and Li Yuan. 2025. [LLaVA-CoT: Let Vision Language Models Reason Step-by-Step](#). *Preprint*, arXiv:2411.10440.
- Weizhe Yuan, Pengfei Liu, and Graham Neubig. 2021. [Can We Automate Scientific Reviewing?](#) *Preprint*, arXiv:2102.00176.
- Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2024. [Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14544–14607, Bangkok, Thailand. Association for Computational Linguistics.
- Penghai Zhao, Qinghua Xing, Kairan Dou, Jinyu Tian, Ying Tai, Jian Yang, Ming-Ming Cheng, and Xiang Li. 2024. [From Words to Worth: Newborn Article Impact Prediction with LLM](#). *Preprint*, arXiv:2408.03934.
- Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyang Luo, Zhangchi Feng, and Yongqiang Ma. 2024. [LlamaFactory: Unified Efficient Fine-Tuning of 100+ Language Models](#). *Preprint*, arXiv:2403.13372.
- Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, Chao Cao, Hanqi Jiang, Hanxu Chen, Yiwei Li, Junhao Chen, Huawen Hu, Yihen Liu, Huaqin Zhao, Shaochen Xu, Haixing Dai, Lin Zhao, Ruidong Zhang, Wei Zhao, Zhenyuan Yang, Jingyuan Chen, Peilong Wang, Wei Ruan, Hui Wang, Huan Zhao, Jing Zhang, Yiming Ren, Shihuan Qin, Tong Chen, Jiaxi Li, Arif Hassan Zidan, Afrar Jahin, Minheng Chen, Sichen Xia, Jason Holmes, Yan Zhuang, Jiaqi Wang, Bochen Xu, Weiran Xia, Jichao Yu, Kaibo Tang, Yaxuan Yang, Bolun Sun, Tao Yang, Guoyu Lu, Xianqiao Wang, Lilong Chai, He Li, Jin Lu, Lichao Sun, Xin Zhang, Bao Ge, Xintao Hu, Lian Zhang, Hua Zhou, Lu Zhang, Shu Zhang, Ninghao Liu, Bei Jiang, Linglong Kong, Zhen Xiang, Yudan Ren, Jun Liu, Xi Jiang, Yu Bao, Wei Zhang, Xiang Li, Gang Li, Wei Liu, Dinggang Shen, Andrea Sikora, Xiaoming Zhai, Dajiang Zhu, and Tianming Liu. 2024. [Evaluation of OpenAI o1: Opportunities and Challenges of AGI](#). *Preprint*, arXiv:2409.18486.
- Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. [Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks](#). In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, Torino, Italia. ELRA and ICCL.
- Tong Zhu, Xiaoye Qu, Daize Dong, Jiacheng Ruan, Jingqi Tong, Conghui He, and Yu Cheng. 2024. [Llama-moe: Building mixture-of-experts from llama with continual pre-training](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 15913–15923.
- Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. 2018. [Txygen: A Benchmarking Platform for Text Generation Models](#). *Preprint*, arXiv:1802.01886.
- Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. [Large language models for automated scholarly paper review: A survey](#). *Preprint*, arXiv:2501.10326.

Research Domain	# of papers
Machine Learning and Deep Learning	25
NLP and Generative Models	22
Reinforcement Learning	16
Privacy and Security	14
Computer Vision	9
Graph Neural Networks	6
Mathematics and Optimization	5
Application	3

Table 6: The research domains and the number of the papers in ReviewBench

## A More Details on ReviewBench and Evaluation Metrics

### A.1 Research Areas in ReviewBench

Table 6 presents the research domains of the papers in ReviewBench and their corresponding quantities.

### A.2 More Details on Evaluation Metrics

#### Distinct

Distinct measures the diversity of the generated text by calculating the proportion of unique n-grams in the generated text. A higher Distinct score indicates more variety.

$$\text{Distinct}_n = \frac{\# \text{ of unique n-grams}}{\text{Total \# of n-grams}}$$

#### Self-BLEU

Self-BLEU measures the diversity of the generated text by computing the BLEU score between the generated text and its multiple variants. A higher Self-BLEU score indicates less diversity.

$$\text{Self-BLEU} = \frac{1}{N} \sum_{i=1}^N \text{BLEU}(r_i, \hat{r}_i)$$

Where:

- $r_i$  is the i-th variant of the generated text.
- $\hat{r}_i$  is another variant with which the i-th variant is compared.
- $N$  is the total number of generated texts.

In order to compute the overall score, we employ **Inverse Self-BLEU** in this paper, where a higher score indicates greater diversity.

## ROUGE-1

ROUGE-1 measures the overlap of unigrams (single words) between the generated text and reference text.

$$\text{Precision} = \frac{\# \text{ of matching unigrams}}{\text{Total \# of unigrams in generated text}}$$

$$\text{Recall} = \frac{\# \text{ of matching unigrams}}{\text{Total \# of unigrams in reference text}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

## ROUGE-L

ROUGE-L measures the overlap of the longest common subsequence (LCS) between the generated text and the reference text.

$$\text{Precision} = \frac{\text{LCS length}}{\text{Length of generated text}}$$

$$\text{Recall} = \frac{\text{LCS length}}{\text{Length of reference text}}$$

$$\text{F1-score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

In our work, the calculation of the ROUGE series metrics is done using the rouge\_score library in Python.

## B Implementation Details

Our framework utilizes the Llama-3.1-8B-Instruct model trained on NVIDIA A100 GPUs. When generating review comments, we set the number of reviewers,  $N$ , to 3. The judge for the Review Arena is GPT-4o. We selected advanced open-source models such as Llama-3.1 (Dubey et al., 2024), Qwen2.5 (Qwen et al., 2025), and closed-source models including GPT-4o, Claude-3.5, and Deepseek (Liu et al., 2024; Guo et al., 2025) for comparison. We performed supervised fine-tuning on the base model Llama-3.1-8B-Instruct implemented using Llama-Factory (Zheng et al., 2024), setting the number of training epochs to 3, gradient accumulation steps to 32, and warm-up steps to 100. The learning rate was set to 1e-5, with training acceleration implemented using liger-kernel and memory scheduling managed through the Zero3 strategy. The total training cost was approximately 1000 GPU hours.



## C Human Evaluation Questionnaire

Table 7 presents the questionnaire used for human evaluation, detailing the meaning of each evaluation dimension and instructing evaluators to rate the LLM-generated review comments on a scale from 1 to 5. Human-written reviews were also provided alongside as a reference for assessing the relevance and alignment of the generated reviews.

## D Prompt Templates

- Transcription: Figure 10
- Review generation: Figure 11
- Meta review generation: Figure 12
- Review Arena: Figure 13

## E A Case Study of a Complete Process

In Figures 5 to 8, we present the output of a complete review process, providing genuine human review comments for reference (Figure 9). The results indicate that the area chair agent in the ReviewAgents framework effectively synthesizes feedback from different reviewers and generates a well-rounded conclusion. The review conclusions produced by the agent show a high degree of overlap with those made by human reviewers, offering similar feedback on aspects such as the novelty, clarity, lack of theoretical depth, and limited scope of the paper. The final recommendation, similarly, results in an "accept" decision.

Dimension	Description	Scoring Guide
Soundness & Validity	Does the review provide reasonable, objective evaluations and suggestions based on the content of the paper?	1 = Groundless or incorrect judgments; 5 = Rigorous and well-supported analysis
Clarity & Organization	Is the review well-structured, logically organized, and easy to understand?	1 = Disorganized and unclear; 5 = Well-structured and clearly articulated
Constructiveness	Does the review offer concrete and valuable suggestions for improvement, rather than just identifying problems?	1 = Vague or unhelpful feedback; 5 = Clear, specific, and helpful feedback
Consistency with Human Review	How consistent is the review with real human expert reviews in terms of conclusions, focus points, and suggestions?	1 = Significantly inconsistent; 5 = Highly consistent
Position Clarity	Does the review clearly state a stance?	1 = No clear stance or vague position; 5 = Clear and reasonable stance

Table 7: The questionnaire used for human evaluation

<p>&lt;SUMMARY&gt;</p> <p>Summary of the paper: The paper introduces a new framework for decision-time planning (DTP) called the Update-Equivalence Framework. This framework is designed to address the scalability issues associated with existing DTP methods that rely on solving subgames, particularly when dealing with games with large amounts of non-public information. The framework focuses on replicating the updates of last-iterate algorithms, which do not require public information. The paper derives two search algorithms from this framework: Mirror Descent Search (MDS) for fully cooperative games and Magnetic Mirror Descent Search (MMDS) for adversarial games. The authors validate the performance of these algorithms in Hanabi, a benchmark for search in fully cooperative imperfect-information games, and in 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe, adversarial games with minimal public information. The results show that MDS outperforms or matches the performance of public information-based search while using significantly less search time, and MMDS reduces the approximate exploitability of the blueprint policies. The paper's background includes the formalism of finite-horizon partially observable stochastic games (POSGs) and introduces the concept of update equivalence, which establishes a relationship between last-iterate algorithms and DTP algorithms. The paper then provides a detailed explanation of how to derive DTP algorithms using mirror descent and magnetic mirror descent, including the theoretical guarantees and empirical validation through experiments.</p> <p>&lt;/SUMMARY&gt;</p> <p>&lt;STRENGTHS_AND_WEAKNESSES&gt;</p> <p>Strengths and weaknesses of the paper:</p> <p>Strengths: 1. Novel Framework: The Update-Equivalence Framework provides a novel approach to DTP that is scalable to games with large amounts of non-public information. 2. Theoretical Soundness: The paper provides rigorous theoretical guarantees for the algorithms derived from the framework, including convergence properties. 3. Empirical Validation: The paper presents compelling empirical results demonstrating the effectiveness of the proposed algorithms compared to state-of-the-art methods. 4. Clarity and Organization: The paper is well-organized and clearly written, making the complex concepts and algorithms easy to understand.</p> <p>Weaknesses: 1. Limited Scope: The paper focuses on specific games and may not generalize well to other types of games. 2. Limited Comparison: While the paper compares the proposed algorithms to state-of-the-art methods, a more comprehensive comparison with a wider range of algorithms would be beneficial. 3. Implementation Details: The paper lacks detailed information about the implementation of the algorithms, which could be useful for other researchers attempting to replicate the results. 4. Theoretical Challenges: The theoretical analysis of the algorithms may be challenging for some readers to follow, especially those not familiar with optimization and game theory.</p> <p>&lt;/STRENGTHS_AND_WEAKNESSES&gt;</p> <p>&lt;CONCLUSION&gt;</p> <p>Review conclusion: The paper presents a significant contribution to the field of decision-time planning by introducing a novel and scalable framework for handling imperfect-information games. The theoretical guarantees and empirical results provided in the paper are strong, and the algorithms have the potential to outperform existing methods in certain settings. Given the paper's clear writing and well-organized structure, it is recommended for acceptance. However, the authors should consider addressing the limitations mentioned above, such as providing more comprehensive comparisons and detailed implementation information. Additionally, the theoretical analysis could be made more accessible to a broader audience. With these revisions, the paper would be an excellent addition to the top-tier academic conferences in artificial intelligence and machine learning.</p> <p>&lt;/CONCLUSION&gt;</p>
---

Figure 5: Review from reviewer agent 1

<SUMMARY>  
Summary of the paper: The paper introduces the Update-Equivalence Framework for Decision-Time Planning (DTP), which is a novel approach to DTP that does not rely on solving subgames, but rather on update equivalence. This framework is motivated by the limitations of current DTP methods, particularly in games with large amounts of non-public information. The authors derive a provably sound search algorithm for fully cooperative games based on mirror descent and a search algorithm for adversarial games based on magnetic mirror descent. The performance of these algorithms is validated in cooperative and adversarial domains, notably in Hanabi, the standard benchmark for search in fully cooperative imperfect-information games. The mirror descent approach exceeds or matches the performance of public information-based search while using two orders of magnitude less search time, making it the first non-public-information-based algorithm to outperform public-information-based approaches in this domain. The paper's background and notation section provides a detailed formalism of finite-horizon partially observable stochastic games (POSGs) and defines the necessary concepts for understanding the proposed framework and algorithms. The framework of update equivalence is then presented, followed by the derivation of action-value-based planners using mirror descent and magnetic mirror descent. The paper also discusses beyond action-value-based planners and provides empirical evidence for the soundness of the proposed algorithms in Hanabi and 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe.  
</SUMMARY>

<STRENGTHS\_AND\_WEAKNESSES>  
Strengths and weaknesses of the paper:  
Strengths: 1. Novelty: The Update-Equivalence Framework provides a fresh perspective on DTP that addresses limitations of existing methods. 2. Scalability: The proposed algorithms are shown to be scalable to games with large amounts of non-public information. 3. Performance: The mirror descent approach outperforms or matches the performance of public information-based search in the Hanabi benchmark. 4. Soundness: The algorithms are proved to be sound and effective in cooperative and adversarial domains. 5. Empirical evidence: The paper provides extensive empirical evidence supporting the effectiveness of the proposed algorithms.  
Weaknesses: 1. Limited comparison: The paper primarily compares the proposed algorithms to public information-based search methods, and a more comprehensive comparison with other DTP approaches would be beneficial. 2. Lack of theoretical analysis: While the paper provides some theoretical guarantees for the proposed algorithms, a more in-depth theoretical analysis would strengthen the paper's contribution. 3. Practical considerations: The paper does not discuss practical aspects such as computational complexity and implementation details. 4. Limited discussion of related work: The paper could benefit from a more thorough discussion of related work, particularly in the context of alternative approaches to DTP for common-payoff games and 2p0s games.  
</STRENGTHS\_AND\_WEAKNESSES>

<CONCLUSION>  
Review conclusion: The paper presents a novel and promising approach to DTP, the Update-Equivalence Framework, which offers scalability and improved performance in games with large amounts of non-public information. The theoretical guarantees and empirical evidence provided in the paper are strong, and the proposed algorithms demonstrate competitive performance in benchmark domains. However, the paper could benefit from a more comprehensive comparison with other DTP approaches, a deeper theoretical analysis, discussion of practical considerations, and a more thorough discussion of related work. Based on the current state of the paper, I recommend acceptance with the following suggestions for revisions: 1. Expand the comparison with other DTP approaches to provide a more comprehensive evaluation of the proposed algorithms. 2. Provide a more in-depth theoretical analysis of the proposed algorithms. 3. Discuss practical aspects such as computational complexity and implementation details. 4. Include a more thorough discussion of related work, particularly in the context of alternative approaches to DTP for common-payoff games and 2p0s games.  
</CONCLUSION>

Figure 6: Review from reviewer agent 2

<SUMMARY>  
Summary of the paper: The paper introduces the Update-Equivalence Framework for Decision-Time Planning (DTP), which is a novel approach to DTP in imperfect-information games. The framework avoids the limitations of traditional subgame-solving methods that rely on public belief states (PBS), which become impractical with large amounts of non-public information. Instead, it focuses on replicating the updates of last-iterate algorithms, which can be applied without PBSs. The paper derives two search algorithms from this framework: Mirror Descent Search (MDS) for fully cooperative games and Magnetic Mirror Descent Search (MMDS) for adversarial games. The authors validate these algorithms on the Hanabi card game and the 3x3 Abrupt Dark Hex and Phantom Tic-Tac-Toe games, demonstrating that MDS outperforms or matches the performance of PBS-based search algorithms while using significantly less search time. This is the first time a non-PBS-based algorithm has outperformed PBS-based approaches in a domain where they have historically dominated.  
</SUMMARY>

<STRENGTHS\_AND\_WEAKNESSES>  
Strengths and weaknesses of the paper:  
Strengths: 1. Innovative framework: The Update-Equivalence Framework offers a novel and potentially scalable approach to DTP in imperfect-information games. 2. Theoretical foundation: The paper provides a solid theoretical foundation for the proposed algorithms, with rigorous proofs of their properties. - Empirical validation: The paper presents strong empirical results, demonstrating the effectiveness of the proposed algorithms on challenging benchmark games. 3. Practical implications: The proposed algorithms could have significant practical implications for DTP in various domains, including games and real-world decision-making problems.  
Weaknesses: 1. Limited scope: The paper focuses on fully cooperative and adversarial games, which may not be applicable to all types of imperfect-information games. 2. Complexity: The proposed algorithms may be computationally more complex than traditional PBS-based methods, which could be a limitation in some practical applications. 3. Comparison with other methods: While the paper compares the proposed algorithms with PBS-based methods, it would be beneficial to include comparisons with other state-of-the-art DTP algorithms in the same domain.  
</STRENGTHS\_AND\_WEAKNESSES>

<CONCLUSION>  
Review conclusion: The paper presents a significant contribution to the field of decision-time planning in imperfect-information games. The Update-Equivalence Framework is innovative and has the potential to improve the scalability and performance of DTP algorithms. The theoretical foundation and empirical validation provided in the paper are strong, and the practical implications of the proposed algorithms are promising. However, the paper could benefit from a broader scope, including other types of imperfect-information games, and a more comprehensive comparison with other DTP algorithms. Overall, I recommend acceptance of this paper with minor revisions to address the mentioned weaknesses.  
</CONCLUSION>

Figure 7: Review from reviewer agent 3

Meta Review Summary

SUMMARY: The paper "The Update-Equivalence Framework for Decision-Time Planning" introduces a novel approach to decision-time planning (DTP) in imperfect-information games, specifically addressing scalability issues in games with large amounts of non-public information. The proposed Update-Equivalence Framework avoids subgame-solving methods and instead relies on replicating the updates of last-iterate algorithms, which do not require public information. The framework is validated through two search algorithms, Mirror Descent Search (MDS) for fully cooperative games and Magnetic Mirror Descent Search (MMDS) for adversarial games. These algorithms are demonstrated to be effective in domains such as Hanabi, where MDS achieves performance comparable to or better than public information-based search while using significantly less search time.

STRENGTHS AND WEAKNESSES

Strengths:

Novel Framework: The Update-Equivalence Framework is innovative and addresses a significant challenge in DTP, scalability in games with extensive non-public information.

Theoretical Soundness: The paper provides rigorous theoretical guarantees for the algorithms, including convergence properties.

Empirical Validation: Strong empirical results are presented, showing the effectiveness of the proposed algorithms against state-of-the-art methods.- Clarity and Organization: The paper is well-organized and clearly written, facilitating understanding of complex concepts and algorithms.

Weaknesses:

Limited Scope: The framework and algorithms are specifically designed for fully cooperative and adversarial games, limiting their applicability.

Limited Comparison: While comparisons with public information-based search methods are provided, a more comprehensive comparison with other DTP approaches would enhance the evaluation.

Implementation Details: The paper lacks detailed implementation information, which could hinder replication and further development of the algorithms.- Theoretical Analysis: The theoretical analysis is challenging for some readers, and a more accessible presentation could broaden the audience.

Lack of Theoretical Depth: Some reviewers suggest that the theoretical analysis could be more in-depth to strengthen the contribution.

Practical Considerations: The paper does not discuss computational complexity or implementation details, which are important for practical applications.

Limited Discussion of Related Work: The discussion of related work is not comprehensive, and a more thorough exploration of alternative approaches would be beneficial.

CONCLUSION:

Given the strengths and weaknesses identified, the meta review recommends the paper for acceptance with certain revisions. The proposed Update-Equivalence Framework represents a significant contribution to the field of DTP and has the potential to impact future research and applications. To enhance the paper's impact and accessibility, the authors are encouraged to address the following: 1. Expand the scope to include a broader range of imperfect-information games, not limited to fully cooperative and adversarial games. 2. Conduct a more comprehensive comparison with other DTP approaches to demonstrate the relative advantages and disadvantages of the proposed algorithms. 3. Provide detailed implementation information to facilitate replication and further research. 4. Simplify the theoretical analysis to make it more accessible to a wider audience. 5. Include a thorough discussion of related work, focusing on alternative approaches to DTP for various types of games and decision-making problems. 6. Discuss practical aspects such as computational complexity and potential challenges in implementing the proposed algorithms. With these revisions, the paper will be a valuable addition to top-tier academic conferences in artificial intelligence and machine learning.

Figure 8: Meta review from area chair agent

Summarize the scientific claims and findings of the paper: The paper introduces a novel framework for decision-time planning (DTP) in imperfect information games, emphasizing the concept of "update equivalence." This framework is designed to replicate updates of global policy learners, offering an alternative to traditional subgame-based approaches. The authors propose two algorithms under this framework: Mirror Descent Update Equivalent Search (MD-UES) and Magnetic Mirror Descent Update Equivalent Search (MMD-UES). These algorithms are evaluated in various games, including Hanabi, showcasing their effectiveness compared to state-of-the-art methods. The authors claim that the proposed algorithms are simpler, computationally less expensive, and more scalable while delivering comparable or superior performance.

Strengths of the paper:

Novelty of Update Equivalence (Reviewer kQSj, Reviewer W84D): The introduction of the concept of update equivalence is a novel contribution. Despite its conceptual simplicity, this approach had surprisingly been overlooked in the imperfect-information game literature. The relaxation of the size of non-public information makes it applicable to a broader range of imperfect information games.

Competitive Performance with Less Search Time (Reviewer kQSj, Reviewer W84D): The proposed algorithms show superior performance compared to existing state-of-the-art methods while requiring significantly less search time.

Applicability to Imperfect Information Games (Reviewer ipeG, Reviewer W84D): The paper looks at the more challenging setting of imperfect information games, different from conventional Public Belief State (PBS)-based planning.

Clarity and Intuitiveness of Methods (Reviewer W84D): The method is straightforward and intuitive. The fact that mirror descent leads to update equivalence seems elegant and useful.

Weaknesses of the paper:

Presentation and Clarity (Reviewer ipeG, Reviewer Vf98): Reviewers pointed out the presentation of the paper was poor. Notations were reportedly undefined or unclear, and the core idea was hard to follow. The explanation of how the algorithms convert a global policy learner to a decision-time planner was specifically cited as unclear.

Lack of Comprehensive Theoretical Characterization (Reviewer W84D): The theoretical results, such as Theorem 3.3, are seen as relatively basic and lacking a comprehensive characterization of global convergence.

Limited Scope of Empirical Evaluation (Reviewer W84D): While the Hanabi results are impressive, demonstrating the algorithm on a broader set of imperfect information games would strengthen the paper.

Recommendation: Accept

Figure 9: Meta review from human



Given the following review comments, please extract the {CONTENT} in the review and enclose it with {<{TAG}>... </{TAG}>}.

### Notes (!!IMPORTANT)

\*\*1.\*\* You must strictly maintain the original content and meaning of the review, only dividing the content and removing unnecessary parts. No modifications, simplifications, or additions of your own are allowed.

\*\*2.\*\* You must strictly follow the specified format, replacing all the required fields. Use the provided format template for output, and do not alter the output format template or add any extra fields.

\*\*3.\*\* <{TAG}></{TAG}> are special tags that must not be modified, replaced, or omitted.

{review comments}

Figure 10: Prompt for transcription

You are an area chair for top-tier academic conferences.

Based on the reviews provided by multiple reviewers, you need to summarize the meta review of the paper.

The review format includes three sections:

<SUMMARY>Summary of the paper</SUMMARY>  
<STRENGTHS\_AND\_WEAKNESSES>Strengths and weaknesses of the paper</STRENGTHS\_AND\_WEAKNESSES>  
<CONCLUSION>Review conclusion</CONCLUSION>

You need to:

Extract and synthesize the information from multiple reviewers' comments.

Provide a coherent and professional summary, detailing the paper's strengths, weaknesses, and overall evaluation.

Include a final recommendation and clear rationale for that recommendation."

### Paper title: {paper\_title}

### Abstract: {paper\_abstract}

### Reviews:

#### Reviewer 1 said: {review 1}

#### Reviewer 2 said: {review 2}

.....

Figure 12: Prompt for meta review generation

You are a reviewer for top-tier academic conferences.

Based on the provided paper's title, abstract, full paper, and titles and abstracts of related papers, provide a detailed review report.

Your report should follow the structure below:

1. Summary of the paper: Briefly summarize the research background, research questions, methods, results, and contributions of the paper.
2. Strengths and weaknesses of the paper: List the strengths and weaknesses of the paper.
3. Review conclusion: Provide your overall evaluation of the paper, including whether you recommend acceptance, any suggestions for revisions, etc.

Please output the results in the following format:

<SUMMARY>Summary of the paper</SUMMARY>  
<STRENGTHS\_AND\_WEAKNESSES>Strengths and weaknesses of the paper  
</STRENGTHS\_AND\_WEAKNESSES>  
<CONCLUSION>Review conclusion</CONCLUSION>

### Paper title: {paper\_title}

### Abstract: {paper\_abstract}

### Full paper: {full\_paper}

### Related papers:

#### Titles: {related\_paper\_title}

#### Abstracts: {related\_paper\_abstract}

Figure 11: Prompt for review generation for Review Agents

You are the judge in a competition, and you need to decide which review comment is better, i.e., closer to the comment from human.

### The paper title is: {title}

### Review comment 0: {review\_0}

### Review comment 1: {review\_1}

### The standard reference review comment from human is: {review\_human}

### Here are your evaluation criteria:

Consistency of Viewpoint: Whether the viewpoints in the review comments are consistent with the human reviewer's perspective on accepting or rejecting the paper.

Consistency of Content and Ideas: The content of the review should be highly consistent with standard reference reviews, especially in aspects such as scientific contribution, innovation, and experimental results. The review should closely follow the language style and structure of the reference review to ensure consistency and completeness of the arguments.

Conciseness and Clarity: The review should be concise and clear, avoiding lengthy expressions while clearly conveying the key strengths and weaknesses for easy understanding.

### Your task is:

If you think review comment 0 is better than review comment 1, output 0.

If you think review comment 1 is better than review comment 0, output 1.

If you think review comment 0 and review comment 1 are equally good, output 2.

Please directly output your choice without any explanations.

Figure 13: Prompt used in Review Arena