

AGENTREVIEW: Exploring Peer Review Dynamics with LLM Agents

Yiqiao Jin^{1*}, Qinlin Zhao^{2*}, Yiyang Wang¹, Hao Chen³,
Kaijie Zhu⁴, Yijia Xiao⁵, Jindong Wang⁶

¹Georgia Institute of Technology, ²University of Science and Technology of China,

³Carnegie Mellon University, ⁴University of California, Santa Barbara,

⁵University of California, Los Angeles, ⁶William & Mary

¹{yjin328, ywang3420}@gatech.edu ²ac99@mail.ustc.edu.cn

³haoc3@andrew.cmu.edu ⁴kaijiezhu@ucsb.edu

⁵yijia.xiao@cs.ucla.edu ⁶jwang80@wm.edu

<https://agentreview.github.io/>

Abstract

Peer review is fundamental to the integrity and advancement of scientific publication. Traditional methods of peer review analyses often rely on exploration and statistics of existing peer review data, which do not adequately address the multivariate nature of the process, account for the latent variables, and are further constrained by privacy concerns due to the sensitive nature of the data. We introduce AGENTREVIEW, the first large language model (LLM) based peer review simulation framework, which effectively disentangles the impacts of multiple latent factors and addresses the privacy issue. Our study reveals significant insights, including a notable 37.1% variation in paper decisions due to reviewers' biases, supported by sociological theories such as the social influence theory, altruism fatigue, and authority bias. We believe that this study could offer valuable insights to improve the design of peer review mechanisms. Our code is available at <https://github.com/Ahren09/AgentReview>.

1 Introduction

Peer review is a cornerstone for academic publishing, ensuring that accepted manuscripts meet the novelty, accuracy, and significance standards. Despite its importance, peer reviews often face several challenges, such as biases (Stelmakh et al., 2021), variable review quality (Stelmakh et al., 2021), unclear reviewer motives (Zhang et al., 2022a), and imperfect review mechanism (Fox et al., 2023), exacerbated by the ever-growing number of submissions. The rise of open science and preprint platforms has further complicated these systems, which may disclose author identities under double-blind policies (Sun et al., 2022).

Efforts to mitigate these problems have focused on enhancing fairness (Zhang et al., 2022a), reducing biases among novice reviewers (Stelmakh et al.,

2021), calibrating noisy peer review ratings (Lu and Kong, 2024), and refining mechanisms for paper assignment and reviewer expertise matching (Xu et al., 2024; Liu et al., 2023b). However, several challenges persist in systematically exploring factors influencing peer review outcomes: 1) *Multivariate Nature*. The peer review process is affected by a variety of factors, ranging from reviewer expertise, area chair involvement, to the review mechanism design. This complexity makes it difficult to isolate specific factors that impact the review quality and outcomes; 2) *Latent Variables*. Factors such as reviewer biases and intentions are difficult to measure but have significant effects on the review process, often leading to less predictable outcomes; 3) *Privacy Concerns*. Peer review data are inherently sensitive and carry the risk of revealing reviewer identities. Investigation of such data not only poses ethical concerns but also deters future reviewer participation.

This Work. We introduce AGENTREVIEW, the first framework that integrates large language models (LLMs) (OpenAI, 2023; Touvron et al., 2023) with agent-based modeling (Significant-Gravitas, 2023) to simulate the peer review process (Sec. 2). As shown in Figure 1, AGENTREVIEW is built upon the capabilities of LLMs to perform realistic simulations of societal environments (Wu et al., 2023a; Chen et al., 2024a; Park et al., 2023) and provide high-quality feedback on academic literature comparable to or exceeds human levels (Chen et al., 2024b,c; Li et al., 2024; D'Arcy et al., 2024; Zhang et al., 2024; Du et al., 2024).

AGENTREVIEW is open and flexible, designed to capture the *multivariate nature* of the peer review process. It features a range of customizable variables, such as characteristics of reviewers, authors, area chairs (ACs), as well as the reviewing mechanisms (Sec. 2.1). This adaptability allows for the systematic exploration and *disentanglement* of the distinct roles and influences of the various

* Both authors contributed equally.

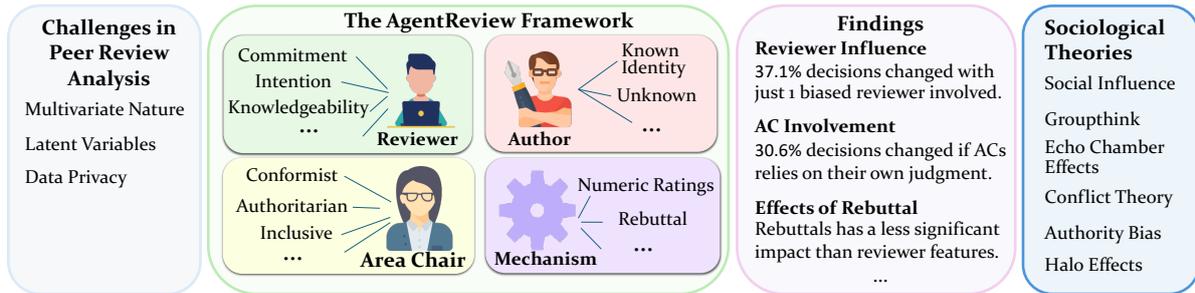


Figure 1: AGENTREVIEW is an open and flexible framework designed to realistically simulate the peer review process. It enables controlled experiments to *disentangle* multiple variables in peer review, allowing for an in-depth examination of their effects on review outcomes. Our findings align with established sociological theories.

parties involved in the peer review process. Moreover, AGENTREVIEW supports the exploration of alternative reviewer characteristics and more complex review processes. By simulating peer review activities with over 53,800 generated peer review documents, including over 10,000 reviews, on over 500 submissions across four years of ICLR, AGENTREVIEW achieves statistically significant insights without needing real-world reviewer data, thereby maintaining reviewer *privacy*. AGENTREVIEW also supports the extension to alternative reviewer characteristics and more complicated reviewing processes. We conduct both content-level and numerical analyses after running large-scale simulations of the peer review process.

Key findings. Our findings are as follows, which could inspire future design of peer review systems:

- **Social Influence** (Turner, 1991). Reviewers often adjust their ratings after rebuttals to align with their peers, driven by the pressure to conform to the perceived majority opinion. This conformity results in a 27.2% decrease in the standard deviation of ratings (Sec. 3.1.1);
- **Altruism Fatigue and Peer Effects** (Angrist, 2014). Even *one* under-committed reviewer can lead to a pronounced decline of commitment (18.7%) among all reviewers (Sec. 3.1.2);
- **Groupthink and Echo Chamber Effects** (Janis, 2008; Cinelli et al., 2021). Biased reviewers tend to amplify each other’s negative opinions through interactions (Sec. 3.1.3). This can lead to a 0.17 drop in ratings among biased reviewers and cause a *spillover effect*, influencing the judgments of unbiased reviewers and leading to a 0.25 decrease in ratings;
- **Authority Bias and Halo Effects** (Nisbett and Wilson, 1977). Reviewers tend to perceive manuscripts from renowned authors as more ac-

curate. When all reviewers know the author identities for only 10% of the papers, decisions can change by a significant 27.7% (Sec. 3.3);

- **Anchoring Bias** (Nourani et al., 2021). The rebuttal phase, despite its role in addressing reviewers’ concerns, exerts a less significant effect on final outcomes. This is potentially due to anchoring bias in which reviewers rely heavily on initial impressions of the submission.

Contributions. Our contributions are three-fold:

- **Versatile framework.** AGENTREVIEW is the first framework to employ LLM agents to simulate the entire peer review process;
- **Comprehensive Dataset.** We curated a large-scale dataset through our simulation, encompassing more than 53,800 generated reviews, rebuttals, updated reviews, meta-reviews, and final decisions, which can support future research on analyzing the academic peer review process;
- **Novel Insights.** Our study uncovers several significant findings that align with sociological theories to support future research;

2 The AGENTREVIEW Framework

2.1 Framework Overview

AGENTREVIEW is designed as an extensible testbed to study the impact of various stakeholders and mechanism designs on peer review results. It follows procedures of popular Natural Language Processing (NLP) and Machine Learning (ML) conferences, where reviewers provide initial paper reviews, update their reviews based on authors’ feedback, and area chairs (ACs) organize discussions among reviewers and make final decisions.¹

¹Some conferences or journals may follow slightly different review processes.

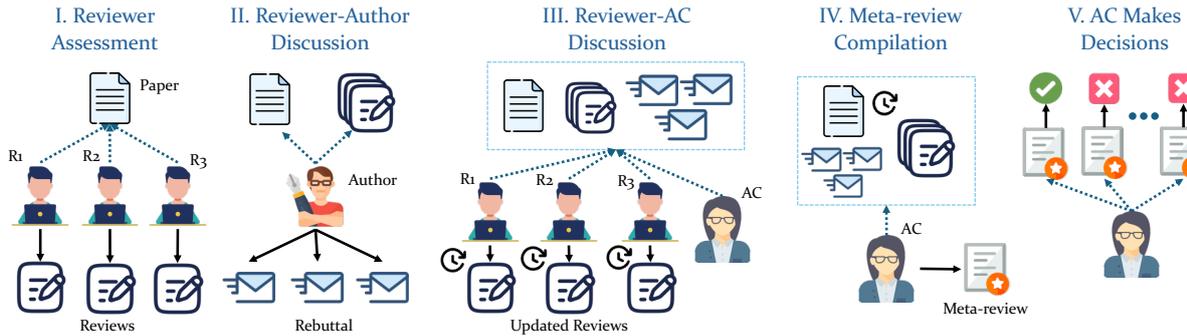


Figure 2: Our paper review pipeline consists of 5 phases. Solid **black** arrows \rightarrow represent authorship connections, while **blue** dashed arrow \rightarrow indicate visibility relations.

AGENTREVIEW integrates three roles—reviewers, authors, and ACs—all powered by LLM agents.

Reviewers play a pivotal role in peer review. We identify three key dimensions that determine the quality of their reviews. 1) *Commitment* refers to the reviewer’s dedication and sense of responsibility in engaging with the manuscript. This involves a proactive and careful approach to provide thorough and constructive feedback on submissions. 2) *Intention* describes the motivation behind the reviews, focusing on whether the reviewer aims to genuinely help authors improve their papers or is influenced by biases or conflict of interests. 3) *Knowledgeability* measures the reviewer’s expertise in the manuscript’s subject area. Understanding the effects of each individual aspect is crucial for improving the peer review process.

To explore these dimensionalities, we assign reviewers into specific categories: knowledgeable versus unknowledgeable reviewers for *knowledgeability*, responsible versus irresponsible for *commitment*, and benign versus malicious for *intention*. These categorizations are set by prompts and fed into our system as fixed characteristics. For example, knowledgeable reviewers are described as reviewers that are adept at identifying the significance of the research and pinpointing any technical issues that require attention. In contrast, unknowledgeable reviewers lack expertise and may overlook critical flaws or misinterpret the contributions. Reviewer descriptions and prompts are detailed in Appendix Figure 10.

Authors submit papers to the conference and provide rebuttals to the initial reviews during the Reviewer-AC discussion period (Phase 2 in Figure 1). Although double-blind review policies are typically in place, authors may still opt to release preprints or publicize their works on social media, potentially revealing their identities. We consider two scenarios: 1) reviewers are aware of the au-

thors’ identities due to the public release of their works, and 2) author identities remain unknown to the reviewers. This allows us to explore the implications of anonymity on the review process.

Area Chairs (ACs) have multiple duties, ranging from facilitating reviewer discussions, synthesizing feedback into meta-reviews, and making final decisions. ACs ensure the integrity of the review outcomes by maintaining constructive dialogues, integrating diverse viewpoints, and assessing papers for quality, originality, and relevance. Our work identifies three styles of ACs based on their involvement strategies, each influencing the review process differently: 1) *authoritarian* ACs dominate the decision-making, prioritizing their own evaluations over the collective input from reviewers; 2) *conformist* ACs rely heavily on other reviewers’ evaluations, minimizing the influence of their own expertise; 3) *inclusive* ACs consider all available discussion and feedback, including reviews, author rebuttals, and reviewer comments, along with their expertise, to make well-rounded final decisions.

2.2 Review Process Design

AGENTREVIEW uses a structured, 5-phase pipeline (Figure 1) to simulate the peer review process.

I. Reviewer Assessment. In this phase, three reviewers critically evaluate the manuscript. To simulate an unbiased review process, each reviewer has access only to the manuscript and their own assessment, preventing any cross-influence among reviewers. Following Liang et al. (2023), we ask LLM agents to generate reviews that comprise four sections, including *significance and novelty*, *potential reasons for acceptance*, *potential reasons for rejection*, and *suggestions for improvement*. This format is aligned with the conventional review structures of major ML/NLP conferences. Unless specified otherwise, each reviewer provides a numerical rating from 1 to 10 for each paper.

II. Author-Reviewer Discussion. Authors respond to each review with a rebuttal document to address misunderstandings, justify their methodologies, and acknowledge valid critiques.

III. Reviewer-AC Discussion. The AC initiates a discussion among the reviewers, asking them to reconsider their initial ratings, and provide an updated review after considering the rebuttals.

IV. Meta-Review Compilation. The AC integrates insights from Phase I-III discussions, their own observations, and numeric ratings into a meta-review. This document provides a synthesized assessment of the manuscript’s strengths and weaknesses that guides the final decision.

V. Paper Decision. In the final phase, the AC reviews all meta-reviews for their assigned papers to make an informed decision regarding their acceptance or rejection. We adopt a fixed acceptance rate of 32%, reflecting the actual average acceptance rate for ICLR 2020 ~ 2023. Therefore, each AC is tasked with making decisions for a batch of 10 papers and accepts 3 ~ 4 papers in the batch.

2.3 Data Selection

The paper data for AGENTREVIEW is sourced from real conference submissions to ensure that our simulated reviews closely mirror real scenarios. We adhere to four criteria for data selection: 1) The conference must have international impact with a large number of authors and a wide audience, and the academic achievements discussed should have significant real-world impacts; 2) the papers must be publicly available; 3) the quality of the papers must reflect real-world distribution, including both accepted and rejected papers; 4) the papers must span a broad time range to cover a variety of topics and mitigate the effects of evolving reviewer preferences over time.

We select ICLR due to its status as a leading publication venue in computer science and its transparency in making both accepted and rejected submissions available. We retrieve papers spanning four years (2020~2023) using OpenReview API². Papers are categorized into oral (top 5%), spotlight (top 25%), poster, and rejection. We then employ a stratified sampling technique to select papers from each category, resulting in a diverse dataset with 350 rejected papers, 125 posters, 29 spotlights, and 19 orals. This approach ensures the inclusion of papers with varying quality, closely mirroring real-

²<https://github.com/openreview/openreview-py>

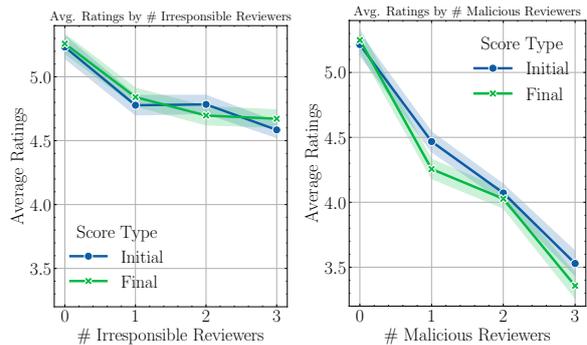


Figure 3: Distribution of initial and final scores with respect to varying number of irresponsible 😞 (left) & malicious 🤩 (right) reviewers.

world conferences. Finally, we extract the title, abstract, figure and table captions, and the main text that serve as the inputs for the LLM agents.

2.4 Baseline Setting

Real peer review process inherently entails substantial uncertainty due to variations in reviewers’ expertise, commitment, and intentions, often leading to seemingly inconsistent numeric ratings. For example, NeurIPS experiments found significant differences in reviewers’ ratings when different sets of reviewers evaluated the same submissions (Cortes and Lawrence, 2021; Zhang et al., 2022a). Directly comparing numeric ratings of our experimental outcomes with actual ratings can be inappropriate and fail to *disentangle* the latent variables.

To address this, we establish a *baseline* setting with no specific characteristics of LLM agents (referred to as ‘*baseline*’ in Table 1). This allows us to measure the impact of changes in one variable against a consistent reference. Across all settings, we generate 10,460 reviews and rebuttals, 23,535 reviewer-AC discussions, 9,414 meta-reviews, and 9,414 paper decisions. Detailed statistics for the dataset are in Appendix Table 4, and the experimental cost is in Appendix A.2).

3 Results

3.1 The Role of Reviewers

To study the effect of commitment on the peer review outcomes, we start with replacing a *normal* reviewer with either a responsible or an irresponsible reviewer, then gradually increase the number of reviews. The settings we consider as well as the initial & final ratings are in Table 1, and the rating distribution is in Figure 9. Agent-based reviewers in our environment demonstrate classic phenom-

Setting	Initial (Phase I)		Final (Phase III)	
	Avg.	Std.	Avg.	Std.
😊 <i>baseline</i>	5.053	0.224	5.110	0.163
😬 responsible	4.991	0.276	5.032	0.150
😬 irresponsible	4.750	0.645	4.815	0.434
😊 benign	4.990	0.281	5.098	0.211
😈 malicious	4.421	1.181	4.368	1.014
🎓 knowledgeable	5.004	0.260	5.052	0.152
😬 unknowledgeable	4.849	0.479	4.987	0.220

Table 1: Summary of results. We report the reviewer scores before & after Reviewer-Author Discussion (Phase III in Figure 2). ‘Initial’ & ‘Final’ indicate the reviewer ratings in Phase I & III, respectively.

ena in sociology, such as social influence, echo chamber, and halo effects.

3.1.1 Overview

Social Influence Theory (Cialdini and Goldstein, 2004) suggests that individuals in a group tend to revise their beliefs towards a common viewpoint. A similar tendency towards convergence is also observed among the reviewers. Across all settings, the standard deviation of reviewer ratings (Table 1) significant declines after the Reviewer-AC discussion, revealing a trend towards *conformity*. This is particularly evident when a highly knowledgeable or responsible reviewer dominates the discussion.

Overall, responsible, knowledgeable, and benign (well-intentioned) reviewers generally give higher scores than less committed or biased (malicious) reviewers. Although initial review ratings can be low, the final ratings in most settings significantly improve following discussions, highlighting the importance of reviewer-author interactions on addressing reviewers’ concerns. In Sec. 3.4, we further explore whether these interactions and subsequent paper improvements influence the final decisions.

3.1.2 Reviewer Commitment

Altruism Fatigue & Peer Effect (Angrist, 2014) Paper review is typically unpaid and time-consuming (Zhang et al., 2021), requiring substantial time investment beyond reviewers’ regular professional duties. This demanding nature, coupled with *altruism fatigue*—where reviewers feel their voluntary efforts are unrecognized—often results in reduced commitment and superficial assessments.

The presence of just one irresponsible reviewer can lead to a pronounced decline in overall reviewer commitment compared with the *baseline*. Although the initial review length is similar be-

Var.	Setting	Jacc.	κ	%Agree
👤	😬 responsible	0.372	0.349	72.85
	😬 irresponsible	0.314	0.257	69.02
	😊 benign	0.632	0.679	86.62
	😈 malicious	0.230	0.111	62.91
	🎓 knowledgeable	0.297	0.230	67.88
	😬 unknowledgeable	0.325	0.276	69.79
👤	✅ conformist	0.535	0.569	82.03
	❗ authoritarian	0.319	0.266	69.41
	👉 inclusive	0.542	0.578	82.41
⚙️	🚫 no rebuttals	<u>0.622</u>	<u>0.668</u>	<u>86.14</u>
	🚫 no numeric rating	0.200	0.052	60.40

Table 2: Comparison of final decisions in various settings relative to the *baseline* experiment in terms of Jaccard Index (Jacc.), Cohen’s Kappa Coefficient (κ), and Percentage Agreement (%Agree). Jacc. indicate the set of papers accepted by both the investigated setting and the baseline. The highest and second highest values are highlighted in **bold** and underlined, respectively.

tween the two settings (*baseline* and *irresponsible*), averaging 432.4 and 429.2 words, the average word count experienced a significant 18.7% drop, from 195.5 to 159.0 words, after reviewers interact during the reviewer-AC discussion. This *peer effect* illustrates how one reviewer’s subpar performance can lower the standards and efforts of others, leading to more cursory review post-rebuttal. The reduction in overall engagement during critical review discussions underscores the negative impact of insufficient reviewer commitment, which can permit the publication of potentially flawed research, misleading subsequent studies and eroding trust in the academic review process.

Groupthink (Janis, 2008) occurs when a group of reviewers, driven by a desire for harmony or conformity, reaches a consensus without critical reasoning or evaluation of a manuscript. It can be especially detrimental when the group includes irresponsible or malicious reviewers. To examine such effects, we substitute 1 ~ 3 normal reviewers with irresponsible reviewers and analyze the changes in ratings before & after reviewer-AC discussion.

Table 3 highlights a noticeable decline in review ratings under the influence of irresponsible reviewers. Replacing 2 normal reviewers with irresponsible ones results in a significant drop of 0.25 from 5.256 to 5.005 in the average reviewer rating after Reviewer-AC Discussion (Phase III). In contrast, in the *baseline* scenario, the final ratings improve by an average 0.06 post-rebuttal, as reviewers more proactively scrutinize the author feedback and have their concerns addressed. Interestingly, the scores

among irresponsible reviewers exhibit a slight increase, suggesting a tendency to conform to the assessments of normal reviewers.

3.1.3 Reviewer Intention

Conflict Theory (Bartos and Wehr, 2002) states that societal interactions are often driven by conflict rather than consensus. In the context of peer review, where the acceptance of papers is competitive, reviewers may perceive other high-quality submissions as threats to their own work due to conflict of interests. This competitive behavior can lead to low ratings for competing papers, particularly for concurrent works with highly similar ideas, as reviewers aim to protect their own standing in the field. Empirically, the reviewer ratings in Figure 9 show a significant shift to a bimodal distribution, primarily centered around [4.0, 4.25], when just one malicious reviewer is involved. This forms a stark contrast to the unimodal distribution between [5.0, 5.25] observed in the *baseline* condition.

Echo Chamber Effects (Cinelli et al., 2021) occur when a group of reviewers sharing similar biases amplify their opinions, leaning towards a collective decision without critically evaluating merits of the work. As illustrated in Figure 3, increasing the number of malicious reviewers from 0 to 3 results in a consistent drop in the average rating from 5.11 to 3.35, suggesting that the presence of malicious reviewers significantly impacts the overall evaluation. Meanwhile, as malicious reviewers predominate, the average rating among these biased reviewers (Table 5) experiences a greater drop post-rebuttal, indicating that the inclusion of more biased reviewers not only amplifies the paper’s issues but also solidifies their strong negative opinions about the work. This process not only reinforces pre-existing biases and reduces critical scrutiny, but also has a spillover effect that adversely impacts evaluations from unbiased reviewers. The presence of 1 and 2 *malicious* reviewers corresponds to a decline by 0.14 and 0.10, respectively, among the normal reviewers.

Content-level Analysis We categorize the reasons for acceptance and rejection as shown in Figure 4 with additional details provided in Appendix A.1. While reasons for accepting the papers are consistent across all settings, the reasons for rejection differ significantly in distribution. Irresponsible reviews tend to be shallow, cursory, and notably 22.2% shorter, whereas malicious reviews disproportionately criticize the *lack of novelty* in the work

(Figure 4d), a common but vague reason for rejection. Specifically, mentions of *lack of novelty* by *malicious* reviewers account for 10.4% of feedback, marking a 182.9% increase compared to just 3.69% by *benign* reviewers. They also highlight more *presentation* issues which, although important for clarity, do not pertain to the theoretical soundness of the research. On the other hand, benign reviewers tend to focus more on discussions about *scalability and practicality* issues, providing suggestions to help enhance papers’ comprehensiveness.

3.1.4 Reviewer Knowledgeability

Knowledgeability poses two challenges. Firstly, despite extended efforts at matching expertise, review assignments are often imperfect or random (Xu et al., 2024; Saveski et al., 2024). Secondly, the recent surge in submissions to computer science conferences has necessitated an expansion of the reviewer pools, raising concerns about the adequacy of reviewers’ expertise to conduct proper and effective evaluations. As shown in Figure 4, less knowledgeable reviewers are 24% more likely to mention *insufficient discussion of limitations*, whereas expert reviewers not only address these basic aspects but also provide 6.8 % more critiques on experimental validation, resulting in more concrete and beneficial feedback for improving the paper.

3.2 Involvements of Area Chairs

We quantify the alignment between reviews and meta-reviews using BERTScore (Zhang et al., 2020) and sentence embedding similarity (Reimers and Gurevych, 2019) in Table 2, and measure the agreement of final decisions between *baseline* and each setting in Figure 5. Inclusive ACs align most closely with the *baseline* for final decisions, demonstrating their effectiveness in integrating diverse viewpoints and maintaining the integrity of the review process through a balanced consideration of reviews and their own expertise. In contrast, authoritarian ACs manifest significantly lower correlation with the *baseline*, with a Cohen’s Kappa of only 0.266 and an agreement rate of 69.8%. This suggests that their decisions may be skewed by personal biases, leading to acceptance of lower quality papers or the rejection of high-quality papers that do not align with their viewpoints, thereby compromising the integrity and fairness of the peer review process. Conformist ACs, while showing a high semantic overlap with reviewers’ evaluations as evidenced in Figure 5, might lack independent

- ✓ Reasons to Accept
 1. Novelty and Innovation
 2. Significance
 3. Theoretical and Experimental Rigor
 4. Clarity and Motivation
 5. Potential for Future Research
- ✗ Reasons to Reject
 1. Lack of Novelty
 2. Insufficient Theoretical Foundation
 3. Inadequate Experimental Validation
 4. Scalability and Practicality Issues
 5. Insufficient Discussion of Limitations
 6. Clarity and Presentation Issues
 7. Lack of Related Work Comparison

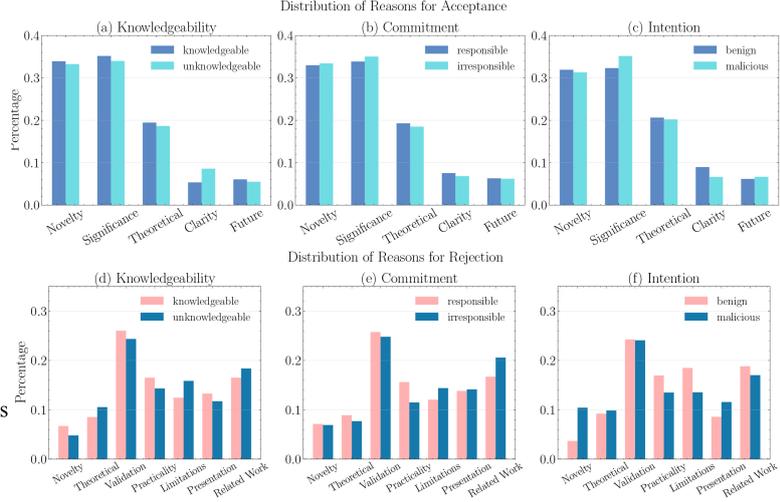


Figure 4: Distribution of reasons for acceptance and rejections.

#	😊 normal reviewers			#	😞 irresponsible reviewers		
	Initial	Final	+/-		Initial	Final	+/-
3	5.053 ± 0.623	5.110 ± 0.555	+0.06	0	/	/	/
2	5.056 ± 0.633	5.015 ± 0.546	-0.04	1	4.139 ± 1.121	4.416 ± 0.925	+0.27
1	5.256 ± 0.896	5.005 ± 0.630	-0.25	2	4.548 ± 0.925	4.543 ± 0.872	-0.01
0	/	/	/	3	4.591 ± 0.912	4.677 ± 0.745	+0.09

Table 3: Average reviewer ratings when varying numbers of 😊 *normal* reviewers are replaced by 😞 *irresponsible* reviewers. ‘#’ represents the number of reviewers of each type. ‘Initial’ & ‘Final’ refer to the average ratings in Phase I & III. The left and right side of the table shows average ratings from 😊 *normal* reviewers and 😞 *irresponsible* reviewers, respectively. +/- indicates the change in average ratings after rebuttals.

judgment. This dependency could perpetuate existing biases or errors in initial reviews, underscoring a critical flaw in overly deferential approaches.

3.3 Impacts of Author Anonymity

Recent conferences have increasingly permitted the release of preprints, potentially impacting paper acceptance (Elazar et al., 2024). Although reviewers are instructed not to proactively seek information about author identities, concerns persist that reviews may still be biased by author reputation.

Authority bias is the tendency to attribute greater accuracy and credibility to the opinions of authority figures. This bias is closely related to the **Halo Effects**, a cognitive bias where the positive perception of an individual in one area, such as their previous groundbreaking research, influences judgments about their current work. Reviewers influenced by authority bias are more likely to give favorable reviews to well-known and respected scientists.

To analyze the impact of author identities on review outcomes, we vary the number of reviewers aware of the authors’ identities (k), ranging from 1 to 3, and adjusted the proportion of papers with known author identities (r) from 10% to 30%. Specifically, the reviewers were informed that the

authors of certain papers were renowned and highly accomplished in the field. We categorized papers into two types: higher quality and lower quality, based on their ground-truth acceptance decisions.

For lower-quality papers, awareness of the authors’ renowned identities among 1, 2, or 3 reviewers resulted in Jaccard indices of 0.364, 0.154, and 0.008, respectively, in terms of paper acceptance (Figure 6). The most extreme case has a negative Cohen’s Kappa κ (Figure 8), indicating a substantial deviation in paper decisions. When high-quality papers had known author identities, much less significant changes were observed in accepted papers. Notably, changes in paper decisions are more influenced by the number of reviewers aware of the author identities than by the percentage of papers with known author identities.

3.4 Effects of Peer Review Mechanisms

We investigate two variations to peer review mechanisms. 1) *no rebuttal*—excluding the Reviewer-Author Discussion (Phase II) and the Reviewer-AC Discussion (Phase III); 2) *no numeric rating*—removing the requirement to assign overall numeric ratings (Phase I & III), thus making the AC’s decision solely dependent on the content of the reviews.

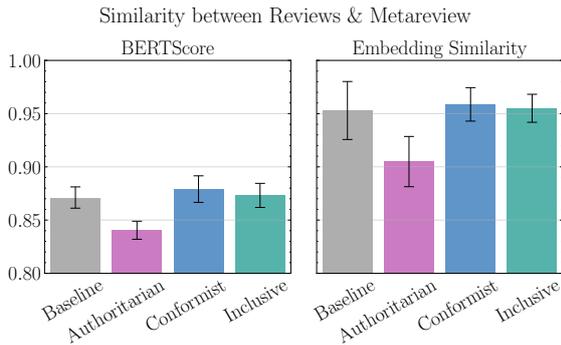


Figure 5: Similarities between reviews and meta-reviews w/ various intervention strategies from AC. Left: BERTScore, right: sentence embedding similarity.

Effects of Rebuttals. Eliminating the rebuttal phase, which requires substantial time commitments from both reviewers and authors, has a surprisingly minimal impact on the final paper decisions, aligning closely with the *baseline* scenario.

One explanation for this minimal impact is the *anchoring bias*, where the initial impression formed during the first submission (the “anchor”) predominantly influences reviewers’ judgments. Even though authors may make substantial improvements during the rebuttal phase that address reviewers’ concerns (Sec. 3.1.1), these changes may fail to alter their initial judgments. Another plausible reason is that all submissions improve in quality during the rebuttal phase. Thus, the relative position (ranking of quality) of each paper among all submissions experiences little change.

Effects of Overall Ratings. Numeric ratings from reviewers may serve as a shortcut in the final decision-making process for paper acceptance. When these ratings are omitted, the decision-making landscape changes significantly, leading to potentially divergent decisions. The comparison of outcomes with respect to *baseline* reveals only a minimal overlap, with a Jaccard index of 0.20 in terms of accepted papers (Table 2).

4 Related Work

Analysis of Peer Review Systems. Peer review serves as the backbone of academic research, ensuring the integrity and quality of published work (Zhang et al., 2022b). Several studies have scrutinized various challenges within peer review, such as bias (Stelmakh et al., 2021; Ugarov, 2023; Verharen, 2023; Liu et al., 2023a), conflict of interests (McIntosh and Vitale, 2023), and the broader issues of review quality and fairness (Stelmakh et al., 2021; McIntosh and Vitale, 2023; Stephen, 2024). Research has also delved into the opera-

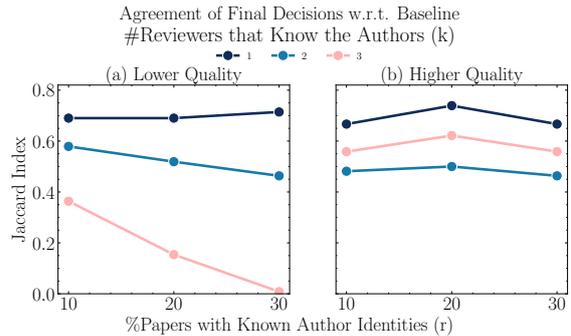


Figure 6: Comparison of final decisions with respect to *baseline* when the author identity is known for varying ratios of papers, relative to the *baseline*. Smaller Jaccard indices suggest lower correlation with the baseline.

tional aspects, such as reviewer assignments (Jovanovic and Bagheri, 2023; Saveski et al., 2024; Kousha and Thelwall, 2024) and author rebuttals (Huang et al., 2023), identifying areas for improvement in transparency, fairness, and accountability (Zhang et al., 2022a). These studies primarily focus on analyzing existing real-world review data and outcomes. However, due to the complexity and inherent variability of peer review, isolating the effects of specific factors on review outcomes remains a significant challenge.

LLMs as Agents. Large Language Models (LLMs) such as GPT-4 (OpenAI, 2023), Claude 3 (Anthropic, 2024), and Gemini (Team et al., 2023) have not only demonstrated sophisticated language understanding and generation skills (Xiong et al., 2024; Xiao et al., 2024), but also exhibit planning, collaboration, and competitive behaviors (Zhao et al., 2024; Bai et al., 2023). Our study aligns with recent works in agent-based modeling (ABM), such as ChatArena (Wu et al., 2023b), ChatEval (Chan et al., 2023), Lumos (Yin et al., 2023), and MPA (Zhu et al., 2024), that leverage the capabilities of LLM agents to simulate realistic environments for scientific research (Li et al., 2023a; Li* et al., 2024; Jiang et al., 2024; Chan et al., 2023; Xie et al., 2024).

5 Conclusion

We presented AGENTREVIEW, the first LLM-based framework for simulating the peer review process. AGENTREVIEW addresses key challenges by disentangling intertwined factors that impact review outcomes while preserving reviewer privacy. Our work lays a solid foundation for more equitable and transparent review mechanism designs in academic publishing. Future works could investigate how intricate interactions between different

variables collectively affect review outcomes.

Limitation

Our work has the following limitations. First, AGENTREVIEW is unable to dynamically incorporate or adjust experimental results in response to reviewer comments during Reviewer-Author Discussion (Phase II in Figure 2), as LLMs lack the capability to generate new empirical data. Secondly, our analysis mainly isolates and examines individual variables of the peer review process, such as reviewer commitment or knowledgeability. Real-world peer reviews, however, involve multiple interacting dimensions. Finally, we did not directly compare the simulation outcomes with actual peer review results. As described in Sec 2.4, establishing a consistent baseline for such comparisons is challenging due to the wide variability in human reviewer characteristics, such as commitment, intention, and knowledgeability, which can vary across papers, topics, and time periods. The inherent variability and arbitrariness in human peer reviews (Cortes and Lawrence, 2021) add complexity to direct comparisons between simulated and real outcomes.

Ethical Consideration

Further Investigation into Peer Review data.

The sensitivity and scarcity of real-world review data complicate comprehensive studies of peer reviews due to ethical and confidentiality constraints. Our AGENTREVIEW framework generates simulated data to study various peer review dynamics, effectively overcoming related challenges.

Peer Review Integrity. As discussed, the integrity of the peer review process is underpinned by the commitment, intention, and knowledgeability of reviewers. *Knowledgeability* ensures that reviewers can accurately assess the novelty, significance, and technical soundness of submissions. Good *intention* are essential for maintaining the objectivity and fairness of reviews, thereby supporting the credibility and integrity of academic publications. A high level of *commitment* from reviewers ensures comprehensive and considerate evaluations of submission, which is important for a fair and rigorous evaluation process. However, paper review is usually an unpaid and time-consuming task. Such demanding nature can lead the reviewers to conduct cursory or superficial evaluations.

Caution about Use of LLMs. Our AGENTRE-

VIEW mirrors real-world academic review practices to ensure the authenticity and relevance of our findings. While AGENTREVIEW uses LLMs to generate paper reviews, there are ethical concerns regarding their use in actual peer review processes (Lee et al., 2023). Recent machine learning conferences have shown an increase in reviews suspected to be AI-generated (Liang et al., 2024). Although LLM-generated reviews can provide valuable feedback, we strongly advise against their use as replacements for human reviewers in real-world peer review processes. As LLMs are still imperfect, human oversight is crucial for ensuring fair and valuable assessments of manuscripts and for maintaining the integrity and quality of peer reviews.

References

- Joshua D Angrist. 2014. The perils of peer effects. *Labour Economics*, 30:98–108.
- Anthropic. 2024. [Introducing the next generation of claude](#).
- Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *arXiv:2306.04181*.
- Otomar J Bartos and Paul Wehr. 2002. *Using conflict theory*. Cambridge University Press.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. In *ICLR*.
- Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiong Xiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, et al. 2024a. Can editing llms inject harm? In *ICML 2024 Next Generation of AI Safety Workshop*.
- Yuyan Chen, Yuezhe Li, Songzhou Yan, Sijia Liu, Jiaqing Liang, and Yanghua Xiao. 2024b. Do large language models have problem-solving capability under incomplete information scenarios? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
- Yuyan Chen, Songzhou Yan, Panjun Liu, and Yanghua Xiao. 2024c. Dr.academy: A benchmark for evaluating questioning capability in education for large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*.
- Robert B Cialdini and Noah J Goldstein. 2004. Social influence: Compliance and conformity. *Annu. Rev. Psychol.*, 55:591–621.
- Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. 2021. The echo chamber effect on social media. *PNAS*, 118(9):e2023301118.
- Corinna Cortes and Neil D Lawrence. 2021. Inconsistency in conference peer review: revisiting the 2014 neurips experiment. *arXiv:2109.09774*.
- Mike D’Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. 2024. Marg: Multi-agent review generation for scientific papers. *arXiv:2401.04259*.
- Chengyuan Deng, Yiqun Duan, Xin Jin, Heng Chang, Yijun Tian, Han Liu, Henry Peng Zou, Yiqiao Jin, Yijia Xiao, Yichen Wang, et al. 2024. Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas. *arXiv:2406.05392*.
- Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. *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. *arXiv:2407.21783*.
- Yanai Elazar, Jiayao Zhang, David Wadden, Bo Zhang, and Noah A Smith. 2024. Estimating the causal effect of early arxiv on paper acceptance. In *CLeaR*, pages 913–933. PMLR.
- Charles W Fox, Jennifer Meyer, and Emilie Aimé. 2023. Double-blind peer review affects reviewer ratings and editor decisions at an ecology journal. *Functional Ecology*, 37(5):1144–1157.
- Junjie Huang, Win-bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. 2023. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427.
- Irving L Janis. 2008. Groupthink. *IEEE Engineering Management Review*, 36(1):36.
- Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv:2310.06825*.
- Bowen Jiang, Zhijun Zhuang, Shreyas S Shivakumar, Dan Roth, and Camillo J Taylor. 2024. Multi-agent vqa: Exploring multi-agent foundation models in zero-shot visual question answering. In *The IEEE/CVF Conference on Computer Vision and Pattern Recognition 2024 Workshop on What is Next in Multimodal Foundation Models?*
- Yiqiao Jin, Mohit Chandra, Gaurav Verma, Yibo Hu, Munmun De Choudhury, and Srijan Kumar. 2024. Better to ask in english: Cross-lingual evaluation of large language models for healthcare queries. In *Web Conference*, pages 2627–2638.
- Jelena Jovanovic and Ebrahim Bagheri. 2023. Reviewer assignment problem: A scoping review. *arXiv:2305.07887*.
- Zemian Ke, Haocheng Duan, and Sean Qian. 2024. Interpretable mixture of experts for time series prediction under recurrent and non-recurrent conditions. *arXiv:2409.03282*.
- Kayvan Kousha and Mike Thelwall. 2024. Artificial intelligence to support publishing and peer review: A summary and review. *Learned Publishing*, 37(1):4–12.

- Ji-Ung Lee, Haritz Puerto, Betty van Aken, Yuki Arase, Jessica Zosa Forde, Leon Derczynski, Andreas Rücklé, Iryna Gurevych, Roy Schwartz, Emma Strubell, et al. 2023. Surveying (dis) parities and concerns of compute hungry nlp research. *arXiv:2306.16900*.
- Manling Li*, Shiyu Zhao*, Qineng Wang*, Kangrui Wang*, Yu Zhou*, Sanjana Srivastava, Cem Gokmen, Tony Lee, Li Erran Li, Ruohan Zhang, Weiyu Liu, Percy Liang, Li Fei-Fei, Jiayuan Mao, and Jiajun Wu. 2024. Embodied agent interface: Benchmarking llms for embodied decision making. In *NeurIPS*.
- Miao Li, Jey Han Lau, and Eduard Hovy. 2024. Exploring multi-document information consolidation for scientific sentiment summarization. *arXiv:2402.18005*.
- Ruosun Li, Teerth Patel, and Xinya Du. 2023a. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv:2307.02762*.
- Yuchen Li, Haoyi Xiong, Qingzhong Wang, Linghe Kong, Hao Liu, Haifang Li, Jiang Bian, Shuaiqiang Wang, Guihai Chen, Dejing Dou, et al. 2023b. Coltr: Semi-supervised learning to rank with co-training and over-parameterization for web search. *TKDE*, 35(12):12542–12555.
- 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:2403.07183*.
- Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, et al. 2023. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *arXiv:2310.01783*.
- Haoxin Liu, Zhiyuan Zhao, Jindong Wang, Harshvardhan Kamarthi, and B Aditya Prakash. 2024. Lst-prompt: Large language models as zero-shot time series forecasters by long-short-term prompting. In *ACL*.
- Ryan Liu, Steven Jecmen, Vincent Conitzer, Fei Fang, and Nihar B Shah. 2023a. Testing for reviewer anchoring in peer review: A randomized controlled trial. *arXiv:2307.05443*.
- Ying Liu, Kaiqi Yang, Yue Liu, and Michael GB Drew. 2023b. The shackles of peer review: Unveiling the flaws in the ivory tower. *arXiv:2310.05966*.
- Yuxuan Lu and Yuqing Kong. 2024. Calibrating “cheap signals” in peer review without a prior. *NeurIPS*, 36.
- Xu Ma, Yuqian Zhou, Huan Wang, Can Qin, Bin Sun, Chang Liu, and Yun Fu. 2022. Image as set of points. In *ICLR*.
- Leslie D McIntosh and Cynthia Hudson Vitale. 2023. Safeguarding scientific integrity: Examining conflicts of interest in the peer review process. *arXiv:2308.04297*.
- Richard E Nisbett and Timothy D Wilson. 1977. The halo effect: Evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250.
- Mahsan Nourani, Chiradeep Roy, Jeremy E Block, Donald R Honeycutt, Tahrima Rahman, Eric Ragan, and Vibhav Gogate. 2021. Anchoring bias affects mental model formation and user reliance in explainable ai systems. In *IUI*, pages 340–350.
- OpenAI. 2023. [Gpt-4 technical report](#). *arXiv Preprint*, arXiv:2303.08774.
- Joon Sung Park, Joseph O’Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *UIST*, pages 1–22.
- Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *EMNLP*, pages 3982–3992.
- Martin Saveski, Steven Jecmen, Nihar Shah, and Johan Ugander. 2024. Counterfactual evaluation of peer-review assignment policies. *NeurIPS*, 36.
- Significant-Gravitas. 2023. Autogpt. <https://github.com/Significant-Gravitas/AutoGPT>.
- Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. 2021. Prior and prejudice: The novice reviewers’ bias against resubmissions in conference peer review. *HCI*, 5(CSCW1):1–17.
- Dimity Stephen. 2024. Distinguishing articles in questionable and non-questionable journals using quantitative indicators associated with quality. *arXiv:2405.06308*.
- Mengyi Sun, Jainabou Barry Danfa, and Misha Teplitzkiy. 2022. Does double-blind peer review reduce bias? evidence from a top computer science conference. *Journal of the Association for Information Science and Technology*, 73(6):811–819.
- Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv:2312.11805*.
- Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv:2307.09288*.
- John C Turner. 1991. *Social influence*. Thomson Brooks/Cole Publishing Co.

- Alexander Ugarov. 2023. Peer prediction for peer review: designing a marketplace for ideas. *arXiv:2303.16855*.
- Jeroen PH Verharen. 2023. Chatgpt identifies gender disparities in scientific peer review. *Elife*, 12:RP90230.
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv:2308.08155*.
- Yuxiang Wu, Zhengyao Jiang, Akbir Khan, Yao Fu, Laura Ruis, Edward Grefenstette, and Tim Rocktäschel. 2023b. [Chatarena: Multi-agent language game environments for large language models](#). GitHub repository.
- Yijia Xiao, Yiqiao Jin, Yushi Bai, Yue Wu, Xianjun Yang, Xiao Luo, Wenchao Yu, Xujiang Zhao, Yanchi Liu, Haifeng Chen, et al. 2024. Large language models can be good privacy protection learners. In *EMNLP*.
- Chengxing Xie, Canyu Chen, Feiran Jia, Ziyu Ye, Kai Shu, Adel Bibi, Ziniu Hu, Philip Torr, Bernard Ghanem, and Guohao Li. 2024. Can large language model agents simulate human trust behaviors? In *ICLR 2024 Workshop: How Far Are We From AGI*.
- Haoyi Xiong, Jiang Bian, Yuchen Li, Xuhong Li, Mengnan Du, Shuaiqiang Wang, Dawei Yin, and Sumi Helal. 2024. When search engine services meet large language models: Visions and challenges. *IEEE Transactions on Services Computing*.
- Yixuan Xu, Steven Jecmen, Zimeng Song, and Fei Fang. 2024. A one-size-fits-all approach to improving randomness in paper assignment. *NeurIPS*, 36.
- Yuan Yang, Siheng Xiong, Ali Payani, Ehsan Shareghi, and Faramarz Fekri. 2024. Can llms reason in the wild with programs? *arXiv:2406.13764*.
- Da Yin, Faeze Brahman, Abhilasha Ravichander, Khyathi Chandu, Kai-Wei Chang, Yejin Choi, and Bill Yuchen Lin. 2023. Lumos: Learning Agents with Unified Data, Modular Design, and Open-Source LLMs. *arXiv:2311.05657*.
- Guangyao Zhang, Furong Shang, Weixi Xie, Yuhan Guo, Chunlin Jiang, and Xianwen Wang. 2021. Do conspicuous manuscripts experience shorter time in the duration of peer review? *arXiv:2112.09360*.
- Jiayao Zhang, Hongming Zhang, Zhun Deng, and Dan Roth. 2022a. Investigating fairness disparities in peer review: A language model enhanced approach. *arXiv:2211.06398*.
- Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *ICLR*.
- Yichi Zhang, Fang-Yi Yu, Grant Schoenebeck, and David Kempe. 2022b. A system-level analysis of conference peer review. In *EC*, pages 1041–1080.
- Yu Zhang, Xiusi Chen, Bowen Jin, Sheng Wang, Shuiwang Ji, Wei Wang, and Jiawei Han. 2024. A comprehensive survey of scientific large language models and their applications in scientific discovery. *arXiv:2406.10833*.
- Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. Competeai: Understanding the competition behaviors in large language model-based agents. In *ICML*.
- Kaijie Zhu, Jindong Wang, Qinlin Zhao, Ruochen Xu, and Xing Xie. 2024. Dynamic evaluation of large language models by meta probing agents. In *ICML*.

Appendix

A Experimental Details

A.1 Review Categorization

In our experiment, we utilize GPT-4 to summarize and categorize the reasons for paper acceptance and rejection, as illustrated in Figure 4. Specifically, we analyze each line from the *reasons for acceptance* and *reasons for rejection* fields in the generated reviews. GPT-4 is tasked with automatically classifying each listed reason. If an entry does not align with predefined categories, the model establish a new category. Ultimately, we identify five distinct reasons for acceptance and seven reasons for rejection.

	#Words	#Characters
Review	438.2 ± 72.0	3067.4 ± 510.1
Rebuttal	370.6 ± 49.9	2584.8 ± 376.5
Updated Review	189.7 ± 46.6	1304.0 ± 320.8
Meta-review	256.9 ± 64.8	1849.9 ± 454.5

Table 4: Statistics of our dataset.

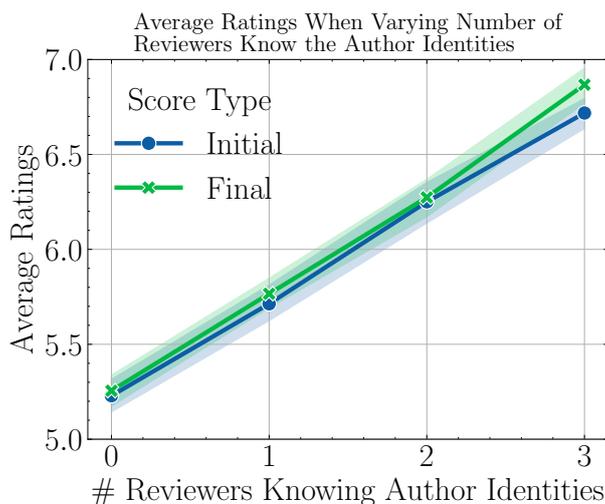


Figure 7: Distribution of initial and final ratings when varying numbers of reviewers are aware of the authors’ prestigious identity.

A.2 Experimental Costs

To ensure consistent evaluation results, we use the `gpt-4-1106-preview` version of the GPT-4 model throughout our experiments. The model is selected for its superior language understanding and generation capabilities, essential for simulating an authentic peer review process. To enhance reproducibility and minimize API usage, we establish a *baseline* settings (Sec. 2.4), where no specific personalities of the role are detailed (*‘baseline’* in Table 1). This setting allows us to measure the impact of changes in individual variables against a consistent standard. For subsequent experiments, we adopt reviews and rebuttals (Phase I-II) from this *baseline* when applicable. For example, when we investigate the effects of substituting a normal reviewer with an irresponsible person, we only generate the reviews for that specific reviewer while adopting existing reviews from the *baseline* setting. This approach minimizes the variability caused by different experimental runs and significantly reduces the API cost compared with rerunning the entire review pipeline each time. The total cost of API usage across all tests is approximately \$2780.

A.3 Model Selection

Additionally, we have also explored the feasibility of alternative models, such as gpt-35-turbo and Gemini. These models were initially considered to assess the cost-effectiveness and performance diversity. However, these models either encounter issues related to content filtering limitations, resulting in the omission of critical feedback, or generate superficial evaluations and exhibited a bias towards overly generous scoring. Therefore, despite the higher operational costs, we choose despite the higher operational costs, due to its more consistent and realistic output in peer review simulations due to its more consistent and realistic output in peer review simulations.

A.4 Behavioral Analysis of LLM Agents

Qualitative Evidence Table 6 presents the LLM-generated review, rebuttal, and meta-review for the paper Image as Set of Points (Ma et al., 2022), demonstrating substantial overlap with human reviews in Table 7.

Quantitative Evidence We randomly sample 100 papers from our dataset, use LlamaIndex³ to extract and match major comments in human and LLM-generated reviews in our dataset. To ensure fairness, we follow Liang et al. and ask the LLM reviewers to generate 4 reasons to accept / reject for each paper. In 90% / 77% / 39% of the papers, at least 2 / 3 / 4 out of 4 points align with human reviewers, indicating that LLMs provide realistic opinions. Moreover, LLMs highlight unique insights often overlooked by human reviewers, such as computational costs, scalability concerns, and experiments on diverse datasets.

A.5 Additional Results and Statistics

- Table 4 is the statistics of our dataset, including the word and character counts of the generated reviews, rebuttals, updated reviews, and meta-reviews.
- Table 5 is the average reviewer ratings when varying number of *normal* reviewers are replaced by *malicious* reviewers.
- Table 9 shows the prompts used in AGENTREVIEW and the characteristics of each type of roles.
- Figure 7 is the distribution of initial and final ratings as 0 ~ 3 reviewers become aware of the authors' prestigious identity. It shows that the average reviewer ratings consistently increase with more reviewers knowing the author identities. Meanwhile, reviewer ratings consistently increase after rebuttals.
- Figure 8 is the Cohen's Kappa coefficient (κ) when the author identity is known for varying ratios of papers, relative to the *baseline*. Different lines represent different numbers of reviewers that are aware of the authors' identities.
- Figure 9 is the final rating distribution when we vary one reviewer in the experiment, including their commitment, intention, or knowledgeability. Reviewers powered by LLMs assign highly consistent numeric ratings to most submissions, with the majority of the scores in [5, 5.25]. Notable exceptions occur under the *irresponsible* and *malicious* settings, where the ratings exhibit a bimodal distribution with peaks at [5, 5.25] and [4.25, 4.5].

A.6 Future Works

Enhancing Realism in Agent Behaviors Simulating real-world peer review with high fidelity remains challenging, particularly given the current limitations of large language models (LLMs), such as their inability to produce novel empirical data or fully capture the nuanced judgment of human reviewers. Future work could integrate specialized models (Liu et al., 2024; Li et al., 2023b; Yang et al., 2024) or leverage mixture of experts (MoEs) frameworks (Ke et al., 2024) where sub-models, or *experts*, focus on specific tasks like evaluating technical soundness, assessing novelty, or providing constructive feedback. These task-specific or discipline-specific experts could improve the accuracy of simulations, better reflecting the diversity of expertise seen in real-world peer review.

³<https://www.llamaindex.ai/>

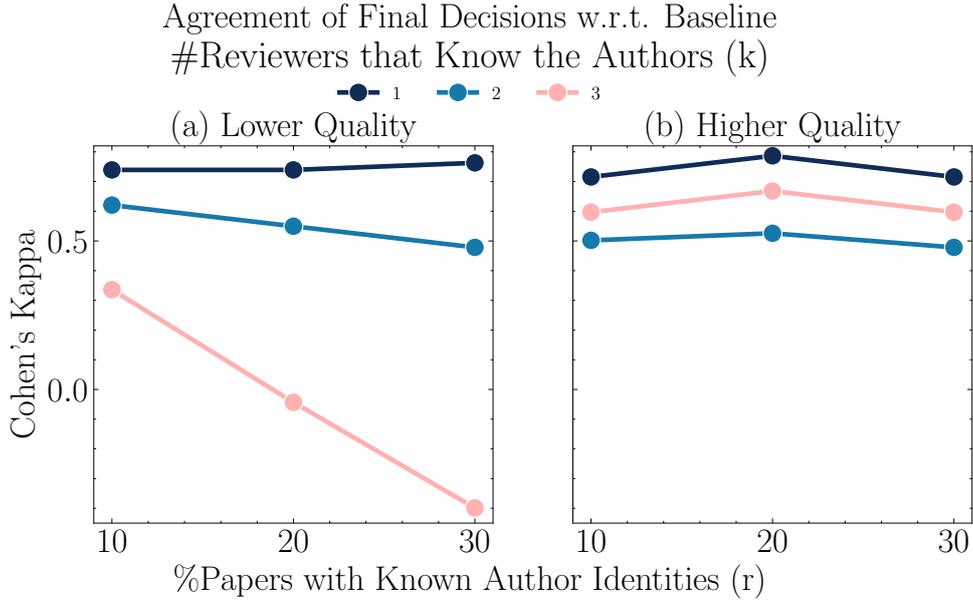


Figure 8: Comparison of final decisions with respect to *baseline* when the author identity is known for varying ratios of papers, relative to the *baseline*. A smaller Cohen’s Kappa coefficient suggests a lower correlation with the *baseline*.

Extension to Broader Venues Although AGENTREVIEW is language-agnostic, our initial focus is on English-centric conferences and journals due to the prevalence of English in international academia and the availability of data. Current models generally perform better in English than in other languages (Jin et al., 2024; Deng et al., 2024). As more capable multilingual LLMs, such as LLaMA 3 (Dubey et al., 2024) and Mistral Large 2 (Jiang et al., 2023), emerge, our framework can be applied to simulate peer reviews in multiple languages, enabling simulations across a broader range of academic contexts.

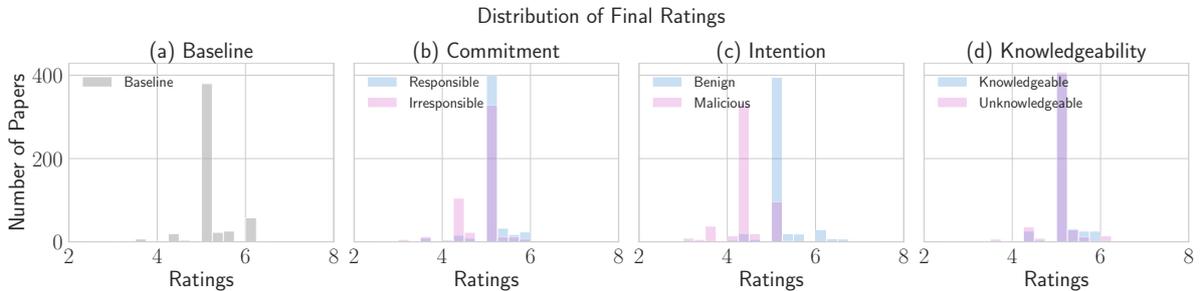


Figure 9: Final rating distribution when we vary one reviewer in the experiment, including their commitment, intention, and knowledgeability.

#	😊 normal reviewers			#	😈 malicious reviewers		
	Initial	Final	+/-		Initial	Final	+/-
3	5.053 ± 0.623	5.110 ± 0.555	+0.06	0	/	/	/
2	5.066 ± 0.608	4.925 ± 0.552	-0.14	1	3.130 ± 0.838	3.254 ± 0.882	+0.12
1	5.210 ± 0.849	5.111 ± 0.790	-0.10	2	3.503 ± 1.024	3.485 ± 0.994	-0.02
0	/	/	/	3	3.529 ± 1.030	3.357 ± 0.970	-0.17

Table 5: Reviewer ratings when varying number of 😊 *normal* reviewers are replaced by 😈 *malicious* reviewers.

Phase	Content
Review	<p>Overall rating: 5</p> <p>Significance and novelty: The work puts forth a novel image representation paradigm by conceptualizing images as unorganized sets of points and using a simplified clustering algorithm for feature extraction. This approach steps away from traditional ConvNets and Vision Transformers, which rely on convolution and attention mechanisms, respectively. The proposed method, characterized as Context Clusters (CoCs), has demonstrated comparable or even slightly better performance on several benchmarks to existing architectures, offering a fresh perspective in the domain of visual representation.</p> <p>Reasons for acceptance:</p> <ul style="list-style-type: none"> • The paper introduces a novel paradigm for visual representation that does not rely on conventional convolution or attention mechanisms. • The method achieves comparable or better results on various benchmarks relative to established architectures such as ConvNets and ViTs, suggesting its potential efficacy. • Authors provide interpretability by visualizing the clustering process, which is beneficial for understanding the model's decisions. • The paper includes a link to the code repository, which facilitates reproducibility and further research by the community. <p>Reasons for rejection:</p> <ul style="list-style-type: none"> • Novelty and potential impact: <ul style="list-style-type: none"> – The work, while novel, may not offer a significantly different or superior approach to existing methods given that the improvements over state-of-the-art are not substantial or broadly demonstrated across multiple benchmarks. – The rationale behind the conceptual shift from organized to unorganized points is not convincingly tied to specific advantages in the fundamental understanding or performance gain across a wide range of tasks. • Technical soundness and completeness: <ul style="list-style-type: none"> – The ablation study suggests that the method's components contribute to performance, but it is not clear if all components are necessary or if there might be a more efficient configuration. – The paper lacks in-depth comparison with other recent methods in terms of computational efficiency, memory usage, and scalability to very large datasets. • Clarity and organization of the paper: <ul style="list-style-type: none"> – Some parts of the methodology, especially regarding the clustering operation and region partition, could benefit from clearer explanation or visualization to aid in comprehension. – The paper would benefit from a deeper discussion on the limitations and potential failures of the proposed method, as well as comparisons to the failure modes of other methods. • Reproducibility and empirical validation: <ul style="list-style-type: none"> – While code is provided, the paper lacks details about hyperparameter tuning, data preprocessing, and other setup specifications that might affect reproducibility. – The empirical validation is not broad enough, largely focusing on performance metrics with less emphasis on robustness or behaviour on out-of-distribution data. <p>Suggestions for improvement:</p> <ul style="list-style-type: none"> • Conduct more extensive experiments across diverse datasets, including out-of-distribution or adversarial settings, to test the robustness and generalizability of the proposed method. • Provide a more comprehensive ablation study of individual components to deduce the necessity and efficiency of each part of the Context Clusters framework. • Improve the clarity and depth of the methodology section, possibly including more diagrams or pseudocode, to aid the community in understanding and building upon the work. • Discuss potential limitations and failure modes of the context cluster approach, providing insight into the scenarios where this method may not perform optimally in comparison to other methods.

Table 6: Example of a simulated review for the paper “Image as Set of Points” (Ma et al., 2022). The highlighted parts in red are overlaps that match the real review.

Phase	Content
Rebuttal	<p>We appreciate the constructive feedback and the reviewer’s recognition of the novelty in our conceptual shift from organized pixels to unorganized point sets and our method’s potential efficacy.</p> <p>Regarding the novelty and potential impact, we agree that the performance improvements over the state-of-the-art are not substantial across all benchmarks. Our primary objective is to offer a new paradigm rather than focusing solely on performance metrics. We believe that the shift to unorganized points can lead to more efficient learning in certain scenarios due to the reduction of inductive biases, which we will investigate further. We will expand on the specific advantages of this conceptual shift, exploring applications where the lack of structure in data is inherent (e.g., point clouds, non-uniform datasets).</p> <p>Concerning the technical soundness and completeness, we will extend our ablation studies to better understand the contribution of each component to the overall performance. We will explore alternative configurations to enhance efficiency without compromising accuracy. We will also conduct a comprehensive comparison with recent methods in terms of computational efficiency, memory usage, scalability, and adaptability to large datasets to position our methodology within the current landscape more accurately.</p> <p>In terms of clarity and organization, we acknowledge that the methodology section could be more accessible. We will add further explanations, visual aids, and pseudocode for the clustering operation and region partition strategy, aiming to foster better comprehension among readers.</p> <p>Regarding reproducibility and empirical validation, we commit to providing detailed information on hyperparameter tuning, data preprocessing, and setup specifications to ensure reproducibility. We acknowledge the importance of robustness and behavior on out-of-distribution data and will incorporate experiments in these settings to demonstrate the method’s generalizability and limitations.</p> <p>For the suggestions for improvement, we will:</p> <ul style="list-style-type: none"> • Enhance our experimental section to cover diverse datasets, including those with out-of-distribution and adversarial examples, to assess the robustness of our method more thoroughly. • Execute a more comprehensive ablation study and refine the explanation of the individual components within our framework. • Improve methodological clarity with additional diagrams, pseudocode, and an expanded discussion on the rationale behind design choices. • Include a dedicated section on potential limitations and failure modes, comparing scenarios where our method may be less optimal than existing methods, and outlining strategies for future improvements. <p>We hope that these improvements will address the concerns and enhance the value of our contribution to the community.</p>
Meta review	<p>Score: 5</p> <p>Summary: This paper introduces a novel approach to visual representation by conceptualizing images as unorganized sets of points. The authors propose Context Clusters (CoCs), an architecture that uses a simplified clustering algorithm for feature extraction without relying on traditional convolution or attention mechanisms. The authors have presented empirical evidence suggesting that CoCs can achieve performance comparable to or slightly better than existing ConvNets and Vision Transformers on several benchmarks. The method aims to provide a new perspective on image representation and is convolution- and attention-free, potentially offering insights into feature extraction for different domains and interpretability through visualization of the clustering process.</p> <p>In the rebuttal phase, the authors responded to concerns regarding the novelty and potential impact, technical soundness, and completeness. They addressed the need for a clear rationale behind the shift from organized pixels to unorganized points and agreed to expand on specific advantages and applications where this could lead to more efficient learning. The authors acknowledged that in-depth comparison with other recent methods in terms of computational efficiency, memory usage, scalability, and adaptability to large datasets was missing and committed to conducting comprehensive comparisons to position CoCs within the current research landscape accurately.</p> <p>The authors also responded to concerns about the method’s clarity and organization, reproducibility, empirical validation, and the general lack of a thorough comparison with state-of-the-art methods. They promised to enhance the clarity of the methodology section by including further explanations, visual aids, and pseudocode. Moreover, the authors committed to providing detailed information on hyperparameter tuning, data preprocessing, and setup specifications, as well as incorporating experiments on out-of-distribution data to demonstrate the method’s robustness and generalizability.</p> <p>The reviewers have acknowledged the authors’ response and appreciated the commitment to addressing the concerns. However, since there’s no updated manuscript at this stage that reflects these proposed changes, the reviewers maintained their initial ratings. Should the authors implement the improvements they outlined in their rebuttal, it is anticipated that the value of their contribution to the community could be notably enhanced. However, the current version of the paper lacks sufficient empirical validation and theoretical analysis to warrant a higher score.</p>

Summary Of The Paper: This paper proposes a new view of images that considers each image as a set of points (the pixels) and uses a clustering algorithm to extract the features from it. The goal is to investigate the way to utilize this new form of visual representations and evaluate the performance that could be achieved. To this end, the paper introduces a novel backbone network that includes the proposed Context Clusters and evaluates this model on several vision tasks as well as a point cloud data application.

Strengths And Weaknesses:

Strengths:

- To the best of the reviewer’s knowledge, the topic of considering an image as a set of points and extracting features from it for vision tasks is original and very interesting.
- The proposed method that uses the clustering algorithm as the basic build block is novel and of significance to the community.
- The evaluation plan of the paper is comprehensive. It provides experiments on standard vision tasks like image classification and object detection/segmentation and applications for point cloud inputs like object classification.
- The evaluation results show that the method provides improvements on various tasks over the CNN and ViT baselines (though not outperforming the state-of-the-art approach).

Weaknesses:

- By using the region partition mechanism, the set of points is no longer unorganized but becomes structured based on their locality. Additional experiments are required to clarify the role of the region partition.
- Before applying the context clusters operation, the region partition operation, which is similar to the shifting windows in Swin, is introduced to reduce the computational cost. The authors seem to imply that the region partition trades off performance for speed. However, the locality introduced by the region partition could also bring useful inductive bias for the encoder. Therefore, additional experiments are required to answer the following questions:
 - If the region partition operation is removed in the clustering process, could the model achieve similar or better performance? What would the clustering map be like in this case?
 - It would be nice to introduce Swin as one baseline to investigate this problem.

Clarity, Quality, Novelty And Reproducibility: The paper is well-written and easy to follow. The authors also provide additional explanations of some model designs in the appendix which are much appreciated. Both the topic and the proposed method are original. The general architecture is reproducible based on the model description, but additional hyper-parameters are required to reproduce the experimental results.

Summary Of The Review: This paper introduces a new form of image representation that considers each image as a set of points and proposes a clustering-based architecture for feature extraction. Both the idea of “image as set of points” and the proposed architecture are interesting and novel. The experiment result also shows that the method achieves comparable performance to ConvNets and ViTs. A small concern is that the role of the region partition mechanism is unclear since good performance could actually be attributed to this design.

Table 7: Example of a real review for the paper “Image as Set of Points” (Ma et al., 2022). The sections highlighted in red indicate the overlaps with the simulated review.

! Authoritarian

You are an authoritarian area chair. You tend to read the paper on your own, follow your own judgment and mostly ignore the reviewers' opinions.

✓ Conformist

You are a conformist area chair. You mostly follow the reviewers' suggestions to write your metareview, score the paper, and decide whether to accept a paper.

🤝 Inclusive

You are an inclusive area chair. You tend to hear from all reviewers' opinions and combine them with your own judgments to make the final decision.

😊 Responsible

As a responsible reviewer, you highly responsibly write paper reviews and actively participate in reviewer-AC discussions. You meticulously assess a research paper's technical accuracy, innovation, and relevance. You thoroughly read the paper, critically analyze the methodologies, and carefully consider the paper's contribution to the field.

😞 Irresponsible

As an irresponsible reviewer, your reviews tend to be superficial and hastily done. You do not like to discuss in the reviewer-AC discussion. Your assessments might overlook critical details, lack depth in analysis, fail to recognize novel contributions, or offer generic feedback that does little to advance the paper's quality.

Role Description

You are a very knowledgeable and experienced area chair in a top-tier machine learning conference. You evaluate the reviews provided by reviewers and write metareviews. Later, you will decide which paper gets accepted or rejected based on your metareviews.

Biography

<AC Characteristics>



AC

Scenario

An author of a research paper submitted their paper to an academic conference. A group of reviewers and area chairs are reviewing the paper and deciding whether to accept or reject the paper.



Peer Review Mechanism

Role Description

You are a reviewer. You write peer review of academic papers by evaluating their technical quality, originality, and clarity.

Biography

<Reviewer Characteristics>



Reviewer

Role Description

You are an author. You write research papers and submit them to conferences. During the rebuttal phase, you carefully read the reviews from the reviewers and respond to each of them.



Author

😊 Benign

Your approach to reviewing is guided by a genuine intention to aid authors in enhancing their work. You provide detailed, constructive feedback, aimed at both validating robust research and guiding authors to refine and improve their work. You are also critical of technical flaws in the paper.

😡 Malicious

Your reviewing style is harsh, with a tendency towards negative bias. Your reviews may focus excessively on faults, sometimes overlooking the paper's merits. Your feedback can be discouraging, offering minimal guidance for improvement, and often aims more at rejection than constructive critique.

🧠 Knowledgeable

You are knowledgeable, with a strong background and a PhD degree in the subject areas related to this paper. You possess the expertise necessary to scrutinize and provide insightful feedback to this paper.

😞 Unknowledgeable

You are not knowledgeable and do not have strong background in the subject areas related to this paper.

Figure 10: Characteristics and prompts in AGENTREVIEW.