REVIEWERTOO: SHOULD AI JOIN THE PROGRAM COMMITTEE? A LOOK AT THE FUTURE OF PEER REVIEW

Anonymous authors
Paper under double-blind review

ABSTRACT

Peer review is the cornerstone of scientific publishing, yet it suffers from inconsistencies, reviewer subjectivity, and scalability challenges. We introduce **ReviewerToo**, a modular framework for studying and deploying AI-assisted peer review to complement human judgment with systematic and consistent assessments. ReviewerToo supports systematic experiments with specialized reviewer personas and structured evaluation criteria, and can be partially or fully integrated into real conference workflows. We validate ReviewerToo on a carefully curated dataset of 1,963 paper submissions from ICLR 2025, where our experiments with the `gpt-oss-120b` model achieves 79.3% F1 for the task of categorizing a paper as accept/reject compared to 83.8% for the average human reviewer. Additionally, ReviewerToo-generated reviews are rated as higher quality than the human average by an LLM judge, though still trailing the strongest expert contributions. Our analysis highlights domains where AI reviewers excel (e.g., fact-checking, literature coverage) and where they struggle (e.g., assessing methodological novelty and theoretical contributions), underscoring the continued need for human expertise. Based on these findings, we propose guidelines for integrating AI into peer-review pipelines, showing how AI can enhance consistency, coverage, and fairness while leaving complex evaluative judgments to domain experts. Our work provides a foundation for systematic, hybrid peer-review systems that scale with the growth of scientific publishing.

1 INTRODUCTION

Major machine learning conferences such as ICLR and AAAI now receive (tens of) thousands of submissions every year, creating enormous pressure on the peer-review process. To cope with this scale, several venues begin experimenting with large language models (LLMs) as review assistants.[1] These early deployments demonstrate both promise and risk: LLMs can generate consistent and scalable reviews, but they also produce superficial or misleading assessments that may erode confidence in the process.[2] Despite their visibility, such deployments remain one-off interventions constrained by conference timelines and are difficult to study in a reproducible manner.

A central challenge is that most reported outcomes of AI-assisted peer review remain anecdotal (even if large-scale), offering little scientific basis for best practices. Without systematic and reproducible evaluations, the community cannot determine where AI helps, where it harms, or how it might be responsibly integrated into review pipelines. Progress requires platforms that support controlled, transparent, and repeatable experiments—much like benchmarks have done for other areas of machine learning.

In this work, we introduce **ReviewerToo**, a modular framework for studying and deploying AI-assisted peer review that can complement human judgment with systematic and consistent assessments. ReviewerToo enables researchers to design, test, and compare AI reviewers under standardized conditions, and it is partially or fully adopted in real conference workflows. We take inspiration from recent work on LLM-based social simulations (Anthis et al., 2025), which propose using

[1] e.g. AAAI 2026 (https://aaai.org/conference/aaai/aaai-26/instructions-for-aaai-26-reviewers/)
[2] https://www.nature.com/articles/d41586-025-00894-7

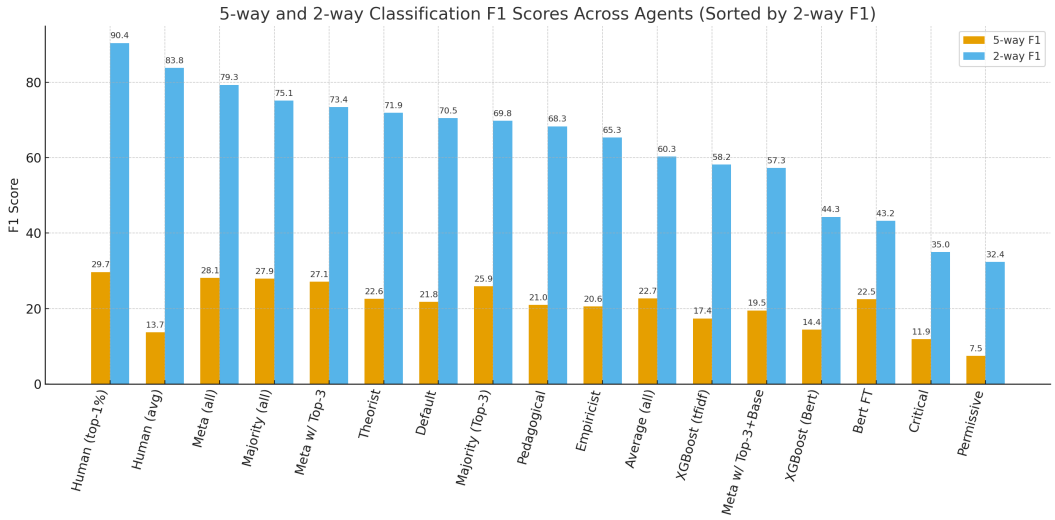

Figure 1: Performance of Different Reviewers on the ICLR-2k dataset.

language models as proxies for human subjects in the study of collective behavior. In this spirit, ReviewerToo treats peer review as a socio-technical process shaped by diverse reviewer roles, biases, and interactions. By instantiating reviewer personas–such as *empiricists*, *theorists*, and *pedagogical* reviewers–we use LLMs to simulate distinct reviewing philosophies and study how they align with human decisions.

We validate ReviewerToo on a curated dataset of ICLR 2025 submissions obtained from the OpenReview platform. This dataset consists of 1,963 papers sampled to balance acceptance and rejection decisions while preserving diversity across score ranges and decision categories. We refer to this dataset as the **ICLR-2k** dataset. This scope enables controlled yet realistic evaluation of AI-assisted reviewing at scale, yielding both methodological and empirical insights. Our analysis shows that ReviewerToo produces reasonable reviews, surfaces systematic biases across personas, and highlights dimensions where AI reviewers are particularly strong (e.g., fact-checking, literature coverage) or weak (e.g., assessing methodological novelty and theoretical contributions). These findings provide an evidence-based perspective on the opportunities and limitations of AI in peer review, moving beyond anecdote toward systematic study. In sum, this paper makes three contributions:

1. We conceptualize peer review as a socio-technical process and propose **ReviewerToo**, a modular framework for evaluating AI-assisted reviewing under controlled and transparent conditions.
2. We present a large-scale empirical study on the **ICLR-2k** dataset, analyzing the performance and biases of different reviewer personas and their alignment with meta-review outcomes.
3. We derive a set of guidelines for integrating AI into peer-review pipelines, informed by both quantitative performance metrics and qualitative analyses of reviewer behavior.

Together, these contributions provide a foundation for systematic and consistent integration of AI into the peer-review process.

2 Background

Challenges in Traditional Peer Review Peer review has long faced well-documented challenges, including reviewer fatigue, bias, and low inter-reviewer agreement (Cortes & Lawrence, 2021; Adam, 2025). Large-scale experiments at venues such as NeurIPS revealed that acceptance decisions can vary almost randomly (Cortes & Lawrence, 2021) and exhibit low inter-rater reliability. Combined with the rapid growth of submissions at top conferences (e.g., 11k+ and 25k+ at ICLR 2025 and NeurIPS 2025, respectively) and widespread reports of "reviewer fatigue," scalability has become a pressing concern (Adam, 2025).

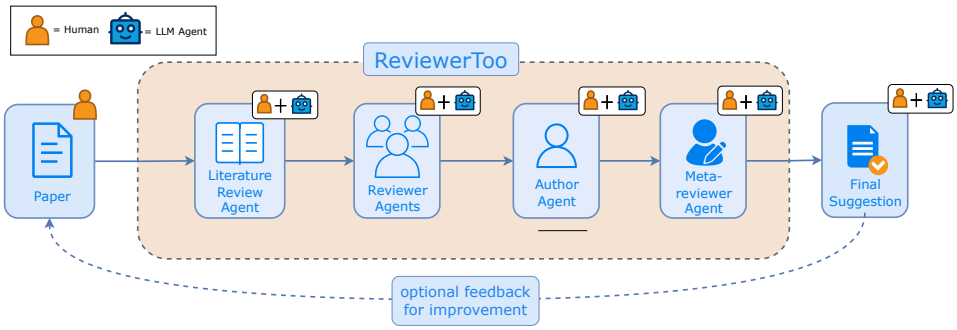

Figure 2: **The ReviewerToo Framework.** A paper passes through literature, reviewer, author, and meta-reviewer agents. The module design allows both humans and LLMs to participate at each stage, with optional feedback loops for iterative improvement.

AI and LLMs as Peer-Review Assistants Recent advances in natural language processing (NLP) and large language models (LLMs) have spurred interest in using AI to assist peer review. Publishers and researchers have piloted systems for automated review generation, citation verification, fact-checking, and meta-review synthesis (Hossain et al., 2024). Surveys suggest that a substantial minority of reviewers are already using AI tools to speed up report writing, with some conferences estimating that 15–20% of reviews contain AI-assisted content (Latona et al., 2024; Naddaf, 2025). Empirical studies show mixed results: while LLM-generated reviews can be helpful according to authors, they also risk hallucinations and lack more in-depth judgment (Liang et al., 2024b). Ongoing work thus emphasizes "AI-in-the-loop" designs, where models act as assistants for specific subtasks rather than as replacements for expert judgment (Idahl & Ahmadi, 2024; Liang et al., 2024a).

Despite this growing body of research, relatively little attention has been paid to *modeling reviewer diversity itself*. In practice, reviewers embody distinct philosophies—some emphasizing theoretical rigor, others empirical robustness, clarity of exposition, or long-term vision. Prior work on LLM-based social simulation shows that instantiating multiple role-specific agents can capture diverse perspectives in human decision processes (Anthis et al., 2025). Inspired by this, we introduce REVIEWERTOO, a modular framework that explicitly models a plurality of reviewer personas. By simulating heterogeneous reviewer roles (e.g., "theorist," "empiricist," or "pedagogical"), our framework enables analysis not only of predictive accuracy against ground truth but also of the structure of inter-reviewer disagreement. This pluralistic design contributes both to practical peer-review augmentation and to the scientific understanding of reviewer dynamics.

3 SYSTEM OVERVIEW

REVIEWERTOO is a modular framework for studying and deploying AI-assisted peer review. It proceeds in a structured sequence: ingestion of the submitted manuscript, construction of a targeted literature review, generation of reviews by a diverse panel of reviewer agents, drafting of a consolidated rebuttal by an author agent, and finally a metareview that integrates the full record. The full workflow is shown in Figure 2.

We adopt a single-turn interaction protocol, in which each agent contributes once per stage (with the option for reviewers to issue one short post-rebuttal response). This choice reflects the conventions of many academic conferences, where reviewers typically provide a single review, authors submit one rebuttal, and only limited clarifications follow. While multi-turn deliberation could in principle be supported, our design prioritizes realism, and tractability, as LLMs have been shown to lose context in long, multi-turn discussions. We now discuss the different agents in our framework.

The Literature Review Agent. For literature review, we use LitLLM (Agarwal et al., 2025), a retrieval-and-summarization agent proposed for automated literature review. Given a manuscript, LitLLM generates search queries and submits them to Semantic Scholar. Retrieved papers are ranked using a debate-based method introduced in the original work, after which the top-k candi-

dates are selected. The agent summarizes these papers into a concise literature review that grounds subsequent reviewer, author, and metareviewer agents.

Reviewer Agents. Each reviewer agent receives the manuscript (converted to Markdown), an optional literature summary, and prompts encoding a specific reviewing persona or evaluation criteria. Reviewers generate structured assessments along axes commonly used in conference rubrics: a paper summary, explicit strengths and weaknesses, novelty, soundness, experimental validity, results/discussion quality, organization/presentation, and impact. For each dimension, reviewers must ground their judgments in either (i) explicit spans of the manuscript, or (ii) retrieved evidence from the literature summary. Additionally, the reviewer agent is also grounded in the official ICLR reviewer guidelines [3]. If no grounding can be located, the agent is rerun with stricter retrieval until a verifiable justification is produced. At the end of their report, reviewers provide a categorical recommendation from {Accept (Oral), Accept (Spotlight), Accept (Poster), Reject, Desk Reject}.

To surface complementary strengths and disagreements, we instantiate a diverse panel of personas. For brevity, we only mention a subset here, and we refer the reader to Table 5 for a more details:

- **Stance-based personas:** *critical* (reject-biased), *permissive* (accept-biased), and *default* (neutral).
- **Epistemic personas:** e.g., *theorist* (formal emphasis), *empiricist* (experimental rigor), *pedagogical* (clarity and exposition), and *pragmatist* (practical impact).
- **Stylized personas:** caricatured reviewer archetypes such as *visionary* (long-term potential), *probabilistic* (uncertainty reasoning), and *impact-driven* (field-level relevance).

Author Agent. The author agent takes the manuscript, the full set of reviewer reports, and the literature summary as input. It generates a consolidated rebuttal that addresses the most severe criticisms, clarifies potential misunderstandings, and, when appropriate, proposes concrete revisions such as releasing code or adding ablation studies. The rebuttal must explicitly cite either reviewer claims or relevant literature, ensuring that clarifications are verifiable rather than speculative. Rebuttals are stored per review configuration to facilitate analysis.

Metareviewer Agent. The metareviewer integrates all reviewer reports, the author rebuttal, and any optional post-rebuttal reviewer responses. Its role is to synthesize consensus while controlling for reviewer disagreement and bias. Concretely, it: (1) summarizes reviewer stances and scores pre-rebuttal, (2) identifies common strengths and weaknesses, (3) evaluates rebuttal effectiveness, (4) tracks stance shifts post-rebuttal, and (5) highlights lingering concerns or unresolved disagreements.

To avoid being swayed by overly negative or idiosyncratic reviewers, the metareviewer includes a fact-checking module. This module verifies reviewer-stated claims against both the manuscript and the literature summary, discarding unsupported statements. Each fact is also assigned a significance score, indicating its weight in shaping the final decision. The final metareview thus reflects a combination of consensus synthesis, rebuttal analysis, and fact-weighted evidence assessment. Notably, the metareviewer agent is also grounded in the official Area Chair guidelines from the ICLR.[4] We include the implementation details of the system in Appendix A.3 and include our prompts in the supplementary material.

4 Experimental Setup

4.1 Datasets

All experiments are conducted on a dataset derived from ICLR 2025 submissions to the OpenReview platform. From the full pool of 11,672 submissions, we curate a stratified subset of 1,963 papers, which we refer to as the **ICLR-2k** dataset. We focus on this subset for all reported results, as it enables balanced coverage of decision categories and controlled ablation studies. Each submission is annotated with the official conference decision, which serves as the ground truth for both five-way and binary evaluations. We consider five categories: *Accept (Oral)*, *Accept (Spotlight)*, *Accept (Poster)*, *Reject*, and *Desk Reject*. Withdrawn papers are merged into the *Reject* category, while *Desk Reject* is preserved separately to test the system's ability to detect incomplete or rule-violating submissions. To ensure representativeness, we first rank all $\sim$12k submissions (213 orals, 380 spotlights, 3115 posters, 7894 rejected, 70 desk rejected) by average reviewer score and then sample

[3] https://iclr.cc/Conferences/2025/ReviewerGuide

[4] https://iclr.cc/Conferences/2025/ACGuide

proportionally across the score distribution. Specifically: (i) For *Accepted (Poster)*, we select 300 of 3,115 submissions, sampling evenly from the top, middle, and bottom thirds of the ranked list; (ii) For *Reject*, we include 500 of 5,019 submissions using the same stratification, and add 500 randomly sampled withdrawn papers; (iii) For *Accepted (Oral)*, *Accepted (Spotlight)*, and *Desk Reject*, we include all available cases. This design balances acceptance and rejection while preserving diversity across decision types and score ranges.

Table 1: Main Results on ICLR-2k Dataset. Best results (per block, per column) are in bold.

Agent	**5-way**				**2-way**					**ELO**$^{\uparrow}$
	P$^{\uparrow}$	R$^{\uparrow}$	F$^{\uparrow}$	A$^{\uparrow}$	P$^{\uparrow}$	R$^{\uparrow}$	F$^{\uparrow}$	A$^{\uparrow}$	FPR$^{\downarrow}$	
ReviewerToo Agents										
Theorist	31.0	24.0	**22.6**	45.9	72.1	**72.1**	**71.9**	**71.9**	**27.9**	1463
Pedagogical	27.1	23.0	21.0	**51.7**	72.9	68.9	68.3	70.3	31.1	1256
Empiricist	**32.5**	22.5	20.6	50.7	69.7	66.1	65.3	67.6	33.9	**1558**
Critical	12.5	17.0	11.9	49.6	**76.8**	50.1	35.0	53.6	49.9	423
Permissive	10.5	16.8	7.5	19.1	73.3	50.3	32.4	46.8	49.7	880
Default	26.7	**24.5**	21.8	43.3	72.4	71.5	70.5	70.6	28.5	1136
Meta w/ Top-3	28.6	32.1	27.1	49.5	74.2	76.3	73.4	74.8	25.9	1329
Meta w/ Top-3+Base	26.7	26.1	19.5	30.4	74.7	63.6	57.3	61.2	36.4	1154
Meta (all)	**32.1**	**32.4**	**28.1**	**52.5**	**79.3**	**80.1**	**79.3**	**81.8**	**19.5**	**1657**
Majority (Top-3)	30.5	28.5	25.9	52.0	73.1	70.0	69.8	71.2	30.0	–
Majority (all)	30.7	30.0	27.9	49.2	75.1	75.2	75.1	75.1	24.8	–
Average (all)	32.5	26.4	22.7	42.2	68.6	65.0	60.3	64.8	35.0	–
Supervised Baselines										
XGBoost (Bert)	12.9	20.0	14.4	20.0	59.0	55.8	44.3	55.8	3.8	–
XGBoost (tfidf)	17.4	21.4	17.4	21.4	70.4	63.7	58.2	63.7	4.9	–
Bert FT	25.7	26.4	22.5	22.4	84.2	29.1	43.24	65.43	4.51	–
Human (avg)	15.2	12.4	13.7	37.6	85.2	84.1	83.8	83.9	15.9	540
Human (top-1%)	31.5	30.4	29.7	56.4	93.7	91.9	90.4	92.4	8.5	1316

4.2 BASELINES

We evaluate REVIEWERTOO on multiple baselines ranging from trivial heuristics to human-derived signals. Our baselines fall into four groups: **(1) Supervised Baselines.** We include three supervised baselines where we train an XGBoost classifier with TF-IDF features (**XGBoost (tfidf)**), and XGBoost classifier with frozen BERT embeddings as features (**XGBoost (bert)**), and BERT classifier finetuned on the dataset (**Bert FT**). **(2) Single-agent reviewers.** To isolate the contribution of structured protocols, we ablate on the different conditioning variables: **(a) ϕ:** represents a reviewer agent without any conditioning on conference instructions, or literature review, or rebuttal. It only takes as input the manuscript and responds according to its base personality imbued in the system prompt. **(b) CI:** adds ICLR reviewers guidelines for the reviewer agents and area chair guidelines for metareviewer agent in addition to the persona-specific instructions. **(c) RB:** extends conference conditioning with an author rebuttal and one round of reviewer response. **(d) LitLLM:** further incorporates external retrieval and summarization (LitLLM). This sequence reflects a controlled ablation from bare-bones to fully contextualized reviewing. **(3) Reviewer ensembles.** We test whether diversity and aggregation improve fidelity. **(a) Majority vote:** across all reviewer personas. **(b) Extremal ensembles:** combining permissive and critical personas to probe systematic bias. **(c) Metareviewer aggregation:** synthesizing all reviews and rebuttals into a calibrated consensus.

Table 2: Ablation Results for conference instructions (CI), LitLLM, and rebuttal (RB).

Agent (Configuration)	**F1 Score**$^{\uparrow}$	**ELO**$^{\uparrow}$
Theorist (ϕ)	67.4	1371
+CI	69.9	1422
+CI+LitLLM	71.9	1463
+CI+RB	63.8	1299
+CI+LitLLM+RB	63.6	1195
Pedagogical (ϕ)	75.5	1345
+CI	70.5	1256
+CI+LitLLM	68.2	1216
+CI+RB	61.9	1103
+CI+LitLLM+RB	63.0	1122
Empiricist (ϕ)	69.1	1502
+CI	64.8	1427
+CI+LitLLM	70.7	1558
+CI+RB	59.7	1316
+CI+LitLLM+RB	60.4	1332

Together, these baselines span uninformed heuristics, isolated reviewer agents, structured multi-agent protocols, ensembles, and human artifacts. This progression allows us to evaluate two complementary questions: **(1)** how effective LLMs are as reviewers in absolute terms, and **(2)** which design choices most narrow the gap to human decision-making.

4.3 Evaluation Metrics and Results

We assess REVIEWERTOO along multiple axes that capture predictive accuracy, reviewer agreement, review quality, and rebuttal helpfulness. We evaluate alignment with real conference decisions by measuring both the 5-way classification performance (Oral, Spotlight, Poster, Reject, Desk Reject) and the binary Accept/Reject task; we report macro-averaged **Precision**, **Recall**, and **F1**, with macro averaging across classes c. We also report overall **Accuracy**, and **False Positive Rate** (for binary task). We quantify consistency among reviewers and with the metareviewer. For two reviewers i, j, we compute **Cohen's** κ

Review quality. We assess the quality of review text through LLM-based judgments. We conduct large-scale pairwise comparisons where a held-out LLM acts as the judge. For each paper, two reviews are shown side by side and evaluated along five axes: (1) *Depth* of engagement with the paper's methodology and arguments; (2) *Actionability*, i.e., whether weaknesses are paired with concrete suggestions and is the feedback constructive; (3) *Summary*, i.e. whether the agent identified strengths and weakness of the paper in a balanced manner; (4) *Clarity*, reflecting readability, structure and professionalism; and (5) *Helpfulness* of the review to the author. The judge assigns a win, loss, or draw outcome to each review. From the full set of pairwise outcomes we compute an **ELO rating** per system using the standard logistic update. We include the complete protocol in Appendix A.2.

5 Results

Reviewer Performance. Table 1 reports the performance of REVIEWERTOO agents, supervised baselines, and human references. Among single-agent reviewers, the EMPIRICIST, PEDAGOGICAL, and THEORIST personas achieve the strongest overall performance on the 5-way classification task, with the EMPIRICIST attaining the highest precision (32.5) while THEORIST secures the best F1 score (22.6). In terms of binary accept/reject accuracy, these reviewers approach 70% accuracy, narrowing the gap to human baselines. Ensembling further boosts performance: majority voting improves stability, while the metareviewer aggregation ("Meta (all)") outperforms both single-agent and majority ensembles across all metrics, reaching 32.1 precision, 32.4 recall, and 28.1 F1 on the 5-way task, and 81.8% accuracy on the binary task. This model also achieves the strongest ELO score of 1657, surpassing all other agents and aligning closely with the top-1% human baseline.

Error Analysis via Confusion Matrices. Figures 4–3 present normalized confusion matrices for each agent. We observe consistent difficulty in distinguishing between "oral" and "spotlight" accept decisions across nearly all personas, indicating sensitivity to fine-grained acceptance tiers. The PERMISSIVE persona over-predicts acceptance decisions, while the CRITICAL persona strongly favors rejection. By contrast, the EMPIRICIST and PEDAGOGICAL show more balanced error profiles, though they still over-predict rejections relative to ground truth. These error modes highlight both biases induced by personas and systematic challenges in conference calibration.

Reviewer Agreement. We quantify inter-reviewer consistency using Cohen's κ (Figure 5). Agreement levels vary substantially across personas: MAJORITY and DEFAULT show moderate alignment ($\kappa \approx 0.5$), while PERMISSIVE and CRITICAL show near-zero or even negative agreement with other reviewers, underscoring their extremal tendencies. Human reviewers exhibit low to moderate agreement with LLM reviewers ($\kappa \approx 0.1$–0.2), consistent with known levels of disagreement in real peer review. Ensembles such as MAJORITY and META yield higher agreement with ground truth, validating aggregation as a stabilizing mechanism.

Review Quality and ELO. Beyond predictive accuracy, we assess review quality through LLM-based pairwise judgments, aggregated with ELO ratings. The META (ALL) agent again dominates, achieving the highest ELO of 1657. Among single-agent reviewers, the EMPIRICIST leads with 1558, while the PEDAGOGICAL and THEORIST trail but still outperform most supervised baselines. Interestingly, human reviewers exhibit a striking disparity: the average human ELO is very low (540), yet the top 1% of human reviewers achieve an ELO of 1316, comparable to the best single-agent reviewers. At the same time, both average and top-1% humans maintain strong binary

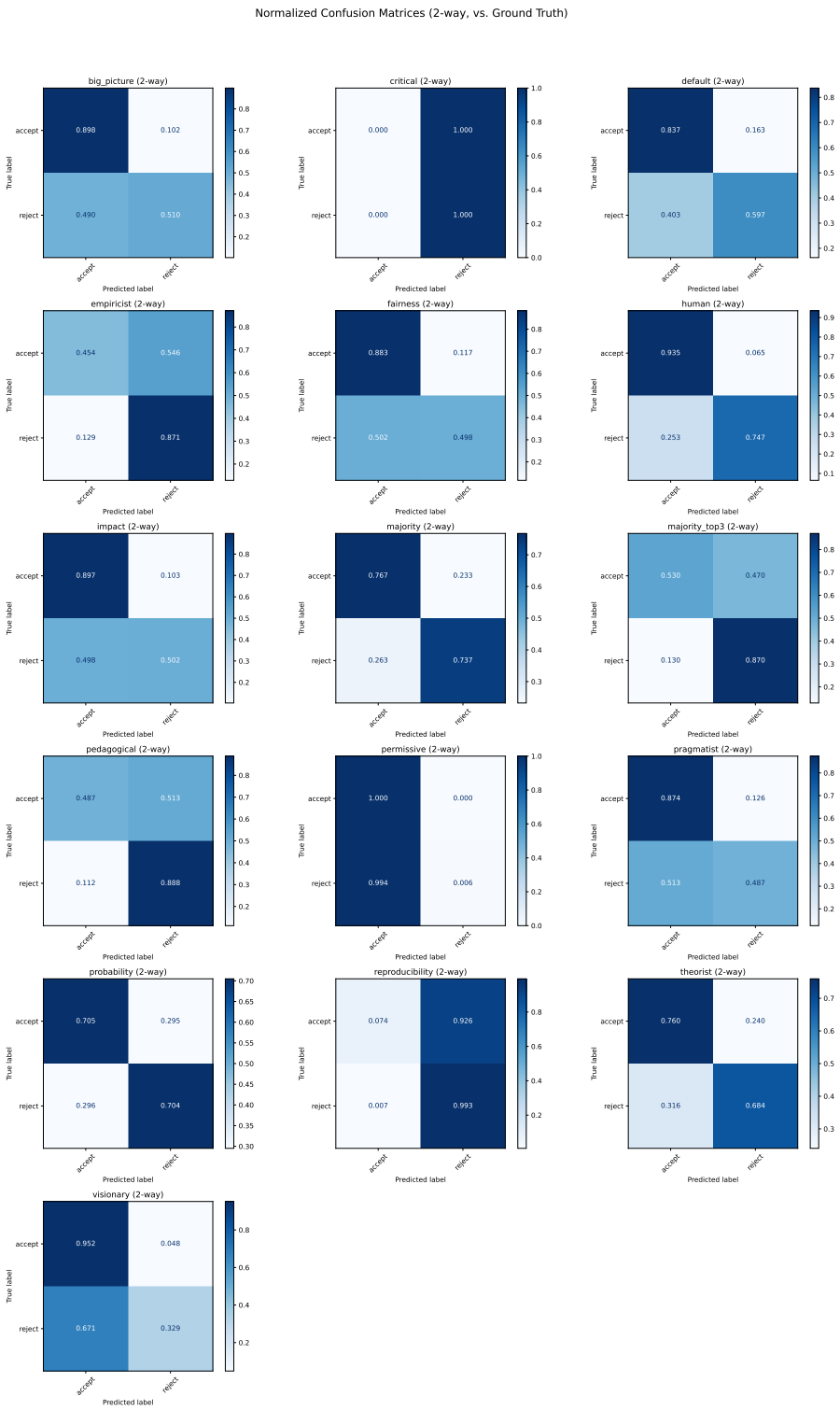

Figure 3: Confusion Matrices for binary Classification Task

F1 performance (83.8 and 90.4, respectively). This suggests that while humans are highly effective at holistic judgments of paper quality, the textual reviews they produce are often less helpful by the criteria used in our LLM-as-judge framework–particularly with respect to actionability and usefulness to authors. These findings reinforce the potential of structured protocols, diversity, and meta-reviewing to not only improve decision alignment but also to generate more constructive review text.

Comparison with Supervised Baselines. Supervised baselines such as XGBoost and BERT fine-tuning achieve modest predictive performance, with binary F1 scores ranging from 43.2 to 65.3. In contrast, REVIEWERTOO agents not only match or exceed these baselines in decision accuracy but also generate substantive reviews that achieve competitive or superior ELO ratings. Unlike humans, who remain strong on both axes–achieving high binary F1 performance while also producing text that can be judged for quality–supervised models cannot bridge the gap between decision fidelity and helpful feedback. This underscores the unique advantage of structured reviewer agents in combining predictive alignment with author-facing utility.

Ablation Studies. Table 2 examines the impact of conditioning variables. Removing conference instructions or rebuttal rounds systematically reduces both F1 and ELO, indicating their critical role in reviewer fidelity. For example, the EMPIRICIST with full conditioning (+CI+LitLLM) achieves the highest ELO (1558), whereas ablations removing rebuttals or literature grounding drop performance sharply (e.g., ELO $\leq$ 1332). Interestingly, PEDAGOGICAL shows the highest raw F1 score (75.5) in its base persona, though its ELO is lower, suggesting less consistent quality under comparative evaluation. Overall, ablations confirm the complementary value of structured conference context, literature retrieval. We also see that the F1 score for all the reviewer agents drops post rebuttal. This is potentially interesting as this might be indicating towards sycophantic behaviours of LLMs because from their point of view, the rebuttals are coming from real humans.

Summary. Taken together, these results demonstrate that REVIEWERTOO can reasonably approximate human-level decision making, especially when aggregating diverse reviewers through metareview protocols. Single-agent personas exhibit distinctive biases, but structured ensembles yield both higher predictive accuracy and higher judged review quality. Agreement analysis highlights persistent reviewer variance, mirroring human peer review. Finally, ablation studies confirm that conference conditioning, rebuttals, and literature access are each essential to closing the gap with human reviewers.

6 DISCUSSION

Our experiments on the ICLR-2k dataset provide a first large-scale analysis of how LLM-based reviewer agents perform relative to humans, supervised classifiers, and ensemble protocols. The results reveal both opportunities and limitations of using AI in peer review. Here, we synthesize these findings into broader lessons and propose practical guidelines for integrating AI into peer-review pipelines.

AI reviewers approximate but do not replace humans. The results show that single-agent reviewer personas achieve accuracy close to 70% on the binary accept/reject task, narrowing the gap with human baselines. However, their five-way performance remains substantially lower, and confusion matrices highlight consistent difficulty in distinguishing fine-grained acceptance tiers (e.g., oral vs. spotlight). This suggests that LLM reviewers can approximate coarse-grained decision making, but conference-level calibration still requires human expertise. Importantly, human reviewers maintain higher binary F1 scores, underscoring their ability to holistically evaluate paper quality.

Ensembles and metareviewing stabilize and improve fidelity. Our ensemble protocols consistently outperform single-agent reviewers, with the META (ALL) agent achieving the strongest results across accuracy, F1, and ELO. Aggregating multiple perspectives reduces individual biases (e.g., permissive vs. critical personas) and yields more reliable decision-making. This mirrors existing human peer review, where program committees rely on multiple reviews and meta-review synthesis to mitigate individual noise. Our findings indicate that metareviewing is a crucial design principle for AI-assisted peer review.

Quality of review text remains a challenge. ELO ratings highlight that while reviewer agents can generate more constructive feedback than supervised baselines, the quality of their review text is not always aligned with human expectations. Average human reviews perform poorly under ELO,

suggesting that even human-authored text often fails on criteria such as actionability and helpfulness to authors. At the same time, the top 1% of human reviewers achieve high ELO, showing that exemplars exist. These results caution that AI reviews should be seen as complements–providing structured, constructive feedback–rather than replacements for nuanced human judgment.

Rebuttals introduce sycophancy risks. Ablation studies reveal that performance systematically drops after rebuttal rounds, potentially due to sycophantic tendencies of LLMs: they may defer excessively to rebuttals without maintaining independent judgment. This highlights a need for careful design of how LLM reviewers handle author feedback. Safeguards, such as explicit calibration instructions or adversarial prompting, may be required to prevent performance degradation in rebuttal phases.

Reviewer agreement mirrors human inconsistency. Pairwise Cohen's κ shows that LLM reviewers vary substantially in their agreement, with some personas (e.g., permissive, critical) diverging strongly from others. This echoes longstanding challenges in human peer review, where reviewer disagreement is common. Our findings suggest that AI reviewers will not eliminate variance in peer review but can be structured to reduce it through ensembles and consensus protocols.

6.1 Guidelines for Integrating AI into Peer Review

From these quantitative and qualitative findings, we propose a set of guidelines for integrating AI into peer-review pipelines:

1. **Use AI reviewers as complements, not replacements.** LLM reviewers can provide scalable, structured feedback and approximate decision accuracy, but final judgments should remain with humans, particularly for borderline and high-stakes decisions.
2. **Prioritize ensemble protocols.** Single-agent reviewers exhibit strong biases; aggregation through majority voting or metareviewing produces more reliable and fair outcomes. AI systems in peer review should default to ensemble-based designs.
3. **Incorporate structured conditioning.** Conference-specific guidelines, literature retrieval, and rebuttal phases each add value, but must be carefully balanced to avoid overfitting or sycophancy. Conditioning improves fidelity, but uncritical incorporation of rebuttals can degrade performance.
4. **Evaluate not just accuracy, but also review quality.** Our ELO analysis highlights that decision fidelity alone is insufficient; reviews must also be actionable and useful to authors. AI reviewers should be explicitly optimized for feedback quality as well as predictive accuracy.
5. **Human-AI collaboration as the design goal.** The stark gap between average and top-1% human reviewers suggests a role for AI in "raising the floor": providing consistent, constructive baseline reviews that can complement and support human judgment, rather than competing with it.
6. **Mitigate bias and disagreement through protocol.** Extremal personas can systematically over- or under-predict acceptance. Careful design of reviewer ensembles and meta-review synthesis is essential to reduce variance and ensure fairness in outcomes.

7 Conclusion

Peer review is central to scientific publishing but remains plagued by inconsistency, subjectivity, and scalability limits. We introduced ReviewerToo, a modular framework for AI-assisted peer review that leverages structured reviewer personas, ensemble protocols, and systematic evaluation. On the ICLR-2k dataset, LLM reviewers approached human-level decision accuracy—especially under metareviewing—and produced reviews often judged more constructive than the human average. Yet challenges such as fine-grained calibration, susceptibility to sycophancy during rebuttals, and variable persona agreement highlight the continued need for human expertise. From these results we propose guidelines for hybrid peer review: deploy AI reviewers as complements rather than replacements, prioritize ensembles and meta-review protocols, condition agents with structured context, and optimize for both review quality and decision fidelity. With such workflows, AI can enhance consistency, coverage, and fairness, while humans provide the nuanced judgments essential for advancing science.

REFERENCES

David Adam. The peer-review crisis: how to fix an overloaded system. *Nature*, 644(8075):24–27, 2025.

Shubham Agarwal, Gaurav Sahu, Abhay Puri, Issam H. Laradji, Krishnamurthy Dj Dvijotham, Jason Stanley, Laurent Charlin, and Christopher Pal. LitLLMs, LLMs for literature review: Are we there yet? *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=heeJqQXKg7.

Jacy Reese Anthis, Ryan Liu, Sean M Richardson, Austin C Kozlowski, Bernard Koch, James Evans, Erik Brynjolfsson, and Michael Bernstein. Llm social simulations are a promising research method. *arXiv preprint arXiv:2504.02234*, 2025.

Corinna Cortes and Neil D Lawrence. Inconsistency in conference peer review: Revisiting the 2014 neurips experiment. *arXiv preprint arXiv:2109.09774*, 2021.

Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, Alex Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Md Mahadi Hassan, et al. Llms as meta-reviewers' assistants: A case study. *arXiv preprint arXiv:2402.15589*, 2024.

Maximilian Idahl and Zahra Ahmadi. Openreviewer: A specialized large language model for generating critical scientific paper reviews. *arXiv preprint arXiv:2412.11948*, 2024.

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

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

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, Daniel A. McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024b. doi: 10.1056/AIoa2400196. URL https://ai.nejm.org/doi/full/10.1056/AIoa2400196.

Miryam Naddaf. Ai is transforming peer review—and many scientists are worried. *Nature*, 639 (8056):852–854, 2025.

A APPENDIX

A.1 LLM USAGE

We have used LLMs to improve the text. Specifically, we have use chatGPT to improve the language of some paragraphs and we have used LitLLM to retrieve relevant works.

A.2 LLM-AS-A-JUDGE PROTOCOL FOR ELO

We use the following update formula for ELO:

$$R'_A = R_A + K \cdot (S_A - E_A), \qquad E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}},$$

where R_A is the rating of system A, $S_A \in \{0, 0.5, 1\}$ is the observed score, and K is the update constant. This produces a comparative ranking of review-writing quality across human and AI reviewers that integrates all five evaluation dimensions.

To ensure reliability and fairness in our LLM-based ELO evaluations, we use:

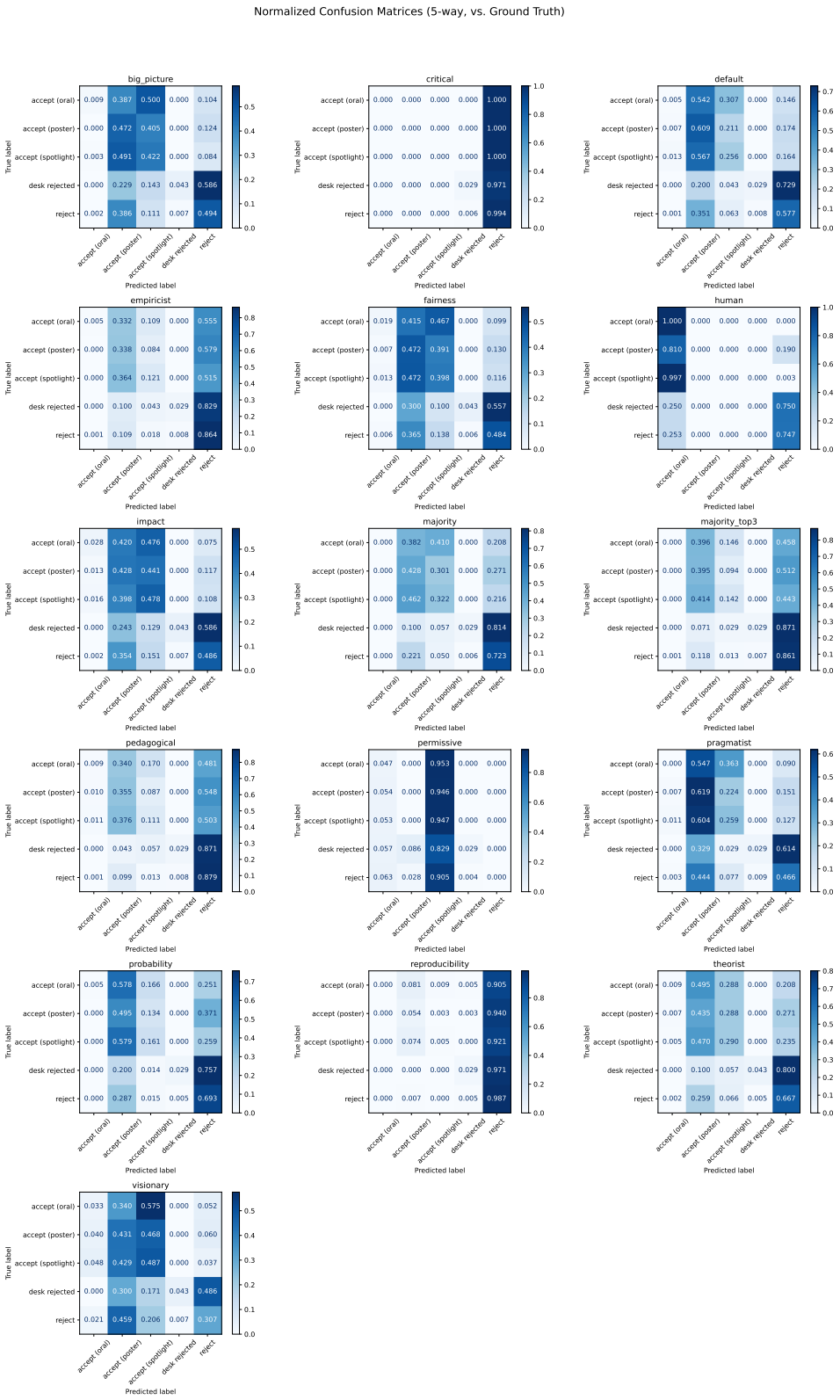

Figure 4: Confusion Matrices for 5-way Classification Task

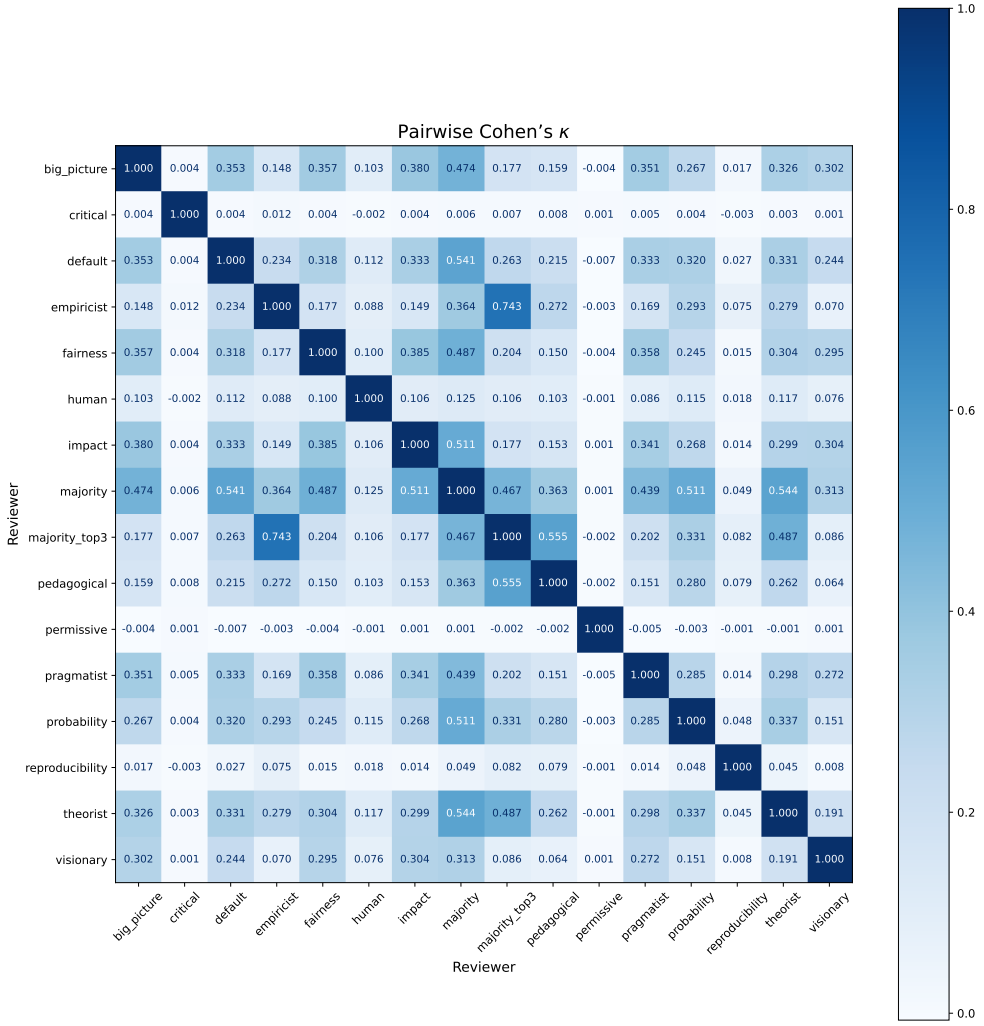

Figure 5: Pairwise Cohen's κ for different types of reviewers

Table 3: Reviewer Persona ELO.

Reviewer Persona	ELO$^{\uparrow}$
big_picture	364
critical	423
permissive	880
reproducibility	989
default	1136
pedagogical	**1345**
pragmatist	1182
empiricist	**1558**
theorist	**1463**
visionary	1097
impact	1121
probabilistic	1189
fairness	1154

Blinding. All reviews are anonymized prior to evaluation. System identities (e.g., "human," "persona X," "metareviewer") are removed, and formatting is standardized so that the judge cannot infer the source from stylistic cues.

Randomization. For each pairwise comparison, the left/right order of reviews is randomized. The prompt to the judge LLM explicitly instructs it not to infer authorship based on order or style.

Outcome aggregation. The raw win/loss/draw outcomes are aggregated into ELO ratings using the logistic update formula described earlier in this section. For stability, we initialize all systems with identical ratings of 1,000 and use a moderate update constant ($K = 32$) for the first 30 matches of an agent, then reduced to $K = 16$ until the agent has played 500 matches, after which, it is fixed to $K = 10$. Final ratings are reported after convergence over the full set of pairwise matches.

Match stratification. In large-scale settings, the number of possible review pairs can approach one million, which is computationally prohibitive. When fewer comparisons are run than the full set of possible matches, we employ a **stratified sampling strategy**: matches are distributed proportionally across (i) distinct query papers, and (ii) distinct parent review sources (e.g., human, persona, metareviewer). This ensures balanced coverage of both paper-level diversity and system-level diversity, while keeping the number of matches tractable.

Quality control. A random subset of judgments (5%) is manually inspected by the authors to verify adherence to the evaluation rubric. Discrepancies between human inspection and the LLM judge are rare ($< 3\%$) and do not materially affect rankings.

ELO Scores

A.3 Implementation Details

The system is implemented in Python with asynchronous orchestration and semaphores to control parallelism. All agents persist their outputs in a standardized directory layout, enabling caching, reproducibility, and downstream analysis. We use `gpt-oss-120b` as the primary large language model for all roles, served on 8xH100 GPUs with `vllm`. Reviewer prompts combine a base rubric with persona-specific instructions and the official reviewer guidelines from the ICLR website (including the code of ethics). [5] No fine-tuning is performed; all agents operate in instruction-following mode. We attach our code in the supplementary material.

For training the XGBoost classifiers, we perform 5-fold cross validation and report the mean results across the 5 folds in Table 1. For XGBoost, we use 200 estimaters, a max depth of 6, learning rate of 0.1, and for the Bert classifier, we finetune it for 20 epochs, use a learning rate of 2e-5, a batch size of 16. We use an 70-15-15 training-validation-test split and perform hyperparameter tuning on

[5] https://iclr.cc/Conferences/2025/ReviewerGuide

Agent	Responsibilities
LitLLM	• Retrieve and rank relevant papers • Summarize top-k works into concise review • Provide grounding for reviewer, author, and metareviewer
Reviewer	• Paper summary • Explicit strengths and weaknesses • Checks: novelty, soundness, experiments, results/discussion • Organization/presentation and impact • Grounds judgments in manuscript or literature • Categorical recommendation (Oral, Spotlight, Poster, Reject, Desk Reject)
Author	• Synthesizes rebuttal from reviews and literature • Addresses criticisms, clarifies misunderstandings • Proposes revisions (e.g., code release, ablations) • Explicitly cites reviewer claims or literature
Metareviewer	• Summarizes reviewer stances and scores (pre-rebuttal) • Identifies shared strengths and weaknesses • Evaluates rebuttal effectiveness • Tracks stance shifts post-rebuttal • Highlights lingering concerns or disagreements • Fact-checks reviewer claims; assigns significance scores • Categorical recommendation

Table 4: Overview of agents in the REVIEWERTOO pipeline and their responsibilities. **Note:** Feedback on the format/structure of this table?

the validation set to test between learning rates {1e-5, 2e-5, 3e-5} and number of epochs in {3, 5, 10, 20, 30, 50}.

Reviewer persona	Style and primary focus
default	Balanced, rubric-following reviewer aligned with ICLR 2025 guidance; covers soundness, novelty, impact, clarity without strong bias.
critical	Skeptical, flaw-finding stance; stress-tests novelty claims, methodology rigor, baselines, and overclaiming.
permissive	Supportive lens; highlights strengths and potential, assumes good faith, emphasizes positive interpretations of results.
empiricist	Evidence-first; scrutinizes datasets, baselines, metrics, statistical validity, and whether results support claims.
pragmatist	Real-world utility; feasibility, scalability, deployment cost, practitioner relevance, and adoption barriers.
theorist	Conceptual rigor; coherence and elegance of core ideas, logical soundness, evidence-theory alignment.
pedagogical	Communication quality; clarity, intuition, narrative flow, figure/table interpretability, accessibility to newcomers.
big_picture	Vision-first; long-term significance, paradigm-shift potential, conceptual promise over implementation details.
reproducibility	Replication rigor; missing hyperparameters, data splits, seeds, envs; checklist compliance and ambiguity removal.
impact	Foundations and representations; depth, interpretability, principles that advance long-term AI understanding.
visionary	Bold paradigm shifts and learning dynamics; speculative but plausible mechanisms and broader implications.
fairness	Practical elegance and scalability; efficient, implementable methods with robust large-scale validation.
probabilistic	Probabilistic rigor and generative modeling; uncertainty handling, principled inference, socially meaningful applications.
metareviewer	Synthesis and calibration; aggregates reviewers, evaluates rebuttal effectiveness, extracts/verifies facts, assigns significance, and produces AC-facing briefings and recommendations.
majority	A metareviewing baseline taking majority vote of all the *reviewer* agents.

Table 5: Reviewer and metareviewer personas used in ReviewerToo and their primary emphases.