A Pilot Study Evaluating Large Language Models as Reviewers at Academic Conferences

Anonymous authors
Paper under double-blind review

Abstract

This paper presents a new system for academic peer review that is more objective, efficient, and community-guided. Our system incorporates *author-assisted evaluation* (Author-AAE) and *community-guided review* (CGR) into the peer review of AI conferences. This is in contrast to existing approaches that prioritize alternative systems that only address some of these challenges. Our evaluation uses data from three major AI conferences that used our system and from a survey of reviewers. Their feedback indicates that our system's reviews are superior to single-LLM-based reviews due to their reduced subjectivity and enhanced quality. The reviewers' scores for our system's reviews were significantly higher than for single-LLM-based reviews across multiple metrics: "Reproducibility and Quality" (by 0.427 ± 0.007), "Review Quality" (by 0.265 ± 0.09), and "Alignment between opinion and paper score" (by 0.503 ± 0.090). In addition, we discovered that single-LLM-based reviews are more likely to be rejected by the program committee after author major revisions (on average by 0.182 ± 0.103) and are much more likely to be rejected overall (on average by 0.300 ± 0.124), compared to our system's reviews. These results suggest that our system performs better in reducing the arbitrary nature of the current peer review system and can serve as an inspiration for the scientific community to explore new review systems.

1 Introduction

The academic peer review system is a core component of scientific advancement and quality control, particularly in the field of computer science, which encompasses areas like artificial intelligence (AI), data science, and information science. In 2024, the Artificial Intelligence Index from the Stanford Institute for Human-Centered Artificial Intelligence notes that AI, data science, and information science are expected to be the areas with the highest growth and job opportunities in the near future (Maslej et al., 2024). Therefore, the performance of review system for these areas is important. However, many experts argue that "there was no valid reason to believe that *the whole peer review system is a good idea*". For example, in the field of medicine, there has been a 25-year long debate about whether to save the current system or to change it (Van Dalen and Henkens, 2012). In the area of computer science, many experts believe that the current system has a "fundamental flaw" (Shah, 2022; Nicholson and Alperin, 2016; Tran et al., 2020). Our work has no intention to join either side of this debate, but rather proposes to the community a working prototype with positive evidence from reviewers.

Specifically, the current system is time-consuming for both authors and reviewers (Stelmakh et al., 2021). It is also open to abuse by reviewers (Russo, 2021; Nicholson and Alperin, 2016) and authors (Wu et al., 2021; Jecmen et al., 2024). In addition, there are concerns of bias and discrimination (Hargittai, 2020), and the subjectivity of research expertise and subject knowledge is often a subject of debate (Mimno and McCallum, 2007; Zhang et al., 2022). Finally, reviews are often arbitrary, even when an author assigns higher value to one over the other, as, empirically demonstrated in multiple works (Beygelzimer et al., 2023; Cortes and Lawrence, 2021).

Existing approaches to these shortcomings can be organized on a spectrum. On one end are approaches that do not change the fundamental nature of the system, but rather address its limitations. On the other end, there are approaches that are fully revolutionary. For example, rather than force reviewers to be “blind” to author information, we can perform various computations to assist them (Kuznetsov et al., 2024). Alternatively, we can create a system in which reviews can be performed by a single large language model (LLM), but other processes need to remain (Liu and Shah, 2023).

Zhou et al. (2024) suggest that LLMs could be a reliable reviewer. Specifically, they cite evidence that LLMs can perform as well as humans on several different tasks, including (1) scoring, (2) commenting, and (3) rating content (Zhou et al., 2024). Previous work has already explored the potential of using LLMs for generating high-quality feedback or summaries, even if their use as a reviewer is left as future work (Liu and Shah, 2023; Liang et al., 2024).

Our method, see Algorithm 1, incorporates Author-Assisted Evaluation (AAE) (Su, 2021) and Community-Guided Review (CGR) (Tran et al., 2020) into LLM-based review system. Concretely, our method is designed to improve the quality of the peer review process from multiple perspectives. First, we reduce the subjectivity of the review system by allowing authors to provide their own scores for their own work, which are considered by LLM-based reviewers. In addition, we greatly improve review quality by directly incorporating community feedback to assist the LLM reviewer. We now formally define the review system:

Definition 1 (Review System) *We say that a review system is composed of (1) an author evaluation process, and (2) a LLM-based reviewer that receives the evaluation from step 1, and the paper to be reviewed. The author evaluation process, which we refer to as AAE, is to ask the author to score their own work, and return their score, their name, and the reasons for their beliefs. In AAE, their name is useful to the LLM for two reasons. First, under some review systems, the LLM can choose to ignore some reviews if the reviewer is blind to the author information. Second, some researchers have suggested that we should “blind” the authors to their own opinions during the revision process, so that they do not work against their own reviews. We propose that rather than blind them, we can use their names to better guide them to focus more on improvements to their work. The CGR is a process that allows reviewers to review the overall quality of the paper, on a scale of 1 to 5, prior to the LLM reviewer beginning to write the standard review.*

Author-based collusion (e.g. author and reviewer are the same person) is a major issue in peer review. Our AAE mechanism both provides additional information to reviewers, and potentially reduces collusion in the first place. In the case of authors, the information gathered can be treated as evidence of self-cheating by the system, and we can choose to ignore reviews with high self-reported scores. Similarly, in the case of reviewers, we can choose to ignore reviews from authors who receive high scores for their own work. We illustrate our approach to AAE in Algorithm 2, where a natural language response from the LLM is parsed to extract the score, owner name, and reason for the belief.

In this paper, we do not compare our approach to the traditional peer review process used by the community and rather justify the necessity of our new approach. In particular, we compare our approach to a simple rule-based review that parses a paper and asks an LLM to peer review without using AAE or CGR (we refer to it as ”single-LLM”, see Algorithm 1). This rule-based process is an iteration of the review process that was explored in past literature (Liu and Shah, 2023). Note that even though they mention the use of LLMs only as reviewers, they can also be combined with other processes which improve the quality of the paper such as author rebuttal matching, paper assignment, and other things. Our work improves all of these processes, but we focus on improving the quality of the LLM reviewer itself. In this paper, we compare our method of reviewing with two rule-based review: (1) the single-LLM review process shown in Algorithm 1, and (2) the ReviewerGPT review process (Liu and Shah, 2023). The first process is a straightforward rule-based approach that provides an apple-to-apple comparison with our work (i.e. no other changes to the peer review system are involved). Our work is the first to evaluate the impact of the rule-based reviews over multiple dimensions, including author-opinions and overall quality, as evaluated by the PC. We use data from *three* major conferences: AAAI 2024, CoRLL 2024, and ICML 2023. Our contributions include: (i) A novel peer review system that incorporates author-assisted evaluation and community-guided reviews. (ii) A three-dimensional evaluation of our approach, considering both apple-to-apple comparisons and overall review quality. (iii) We show evidence that the reviewers think our system’s reviews are better, on average, than single-LLM-based reviews. (iv) We show evidence that reviews from single-LLM reviewers that are received lower scores, both from the PC and the authors.

Our results are summarized as follows. We find positive evidence that the reviewers think our system's reviews are better, on average, than single-LLM-based reviews. We introduced the notion of ”Reproducibility and Quality”, a proxy for whether the work is robust and methodical. It also evaluates whether reviewers can reproduce the findings in the paper, which may require running the code provided with the paper. The reviewers' scores for our system's reviews were significantly higher (p-value = 0.02) than for single-LLM-based reviews across this metric: 0.427 ± 0.007. In addition, we find that reviewers think our system's reviews score were significantly higher (p-value = 0.03) than single-LLM-based reviews across the ”Review Quality” metric: 0.265 ± 0.09. Finally, we find positive evidence that our system's reviews were better aligned with the reviewers' own opinions than single-LLM-based reviews: 0.503 ± 0.090.

In summary, the data we present in this paper positively justifies that our system is an improvement over the single-LLM-based review process.

Algorithm 1 Author-Assisted Evaluation (AAE) and Community-Guided Review (CGR) of a Paper

```
Require: Author name, Author's paper file, Anonymous version of paper
Step 1: Author-Assisted Evaluation (AAE)
Extract author name from paper metadata.
Ask the author to provide:
        • A self-assigned score for the paper.
        • A justification for the score.
        • Their name (for record linkage).
if the reviewer is blind to the author information then
    Skip to reviewer evaluation phase.
else
    Provide the LLM with the author's provided score and justification.
end if
Step 2: Community-Guided Review (CGR)
Invite community members (review group) to rate the paper on a scale of 1–5.
Request the LLM to also provide a paper score on a scale of 1–5.
Compute the average of all scores (human and LLM) without revealing any rater identities.
Ensure: Anonymous paper, CGR score, AAE score, and author name =0
```

Algorithm 2: Author-Assisted Evaluation (AAE)

```
Input: Paper, Author Name
Ask for a rating from the author, a justification,
and the Author Name
Extrapolate from an LLM response:
        1. The rating that the author would give
           to the paper,
        2. The author name,
        3. The author's justification
Output: Rating, Justification, Author Name =0
```

Rule-Based Review (*singlellm*)

```
Input: Paper, Author
For each section s in paper do:
Generate Review Section based on rules
Output: Full Review
```

Algorithm 3: ReviewerGPT (Liu and Shah, 2023)

```
Input: Paper, Reviewer Name
Ask for a rating from the reviewer, a justification, and the Reviewer Name
Create a rule-based review and provide it to the LLM
Extrapolate from an LLM response:
    1. the rating that the reviewer would give to the paper,
    2. the reviewer name,
    3. the justification for their belief
Output: Rating, Justification, Reviewer Name
=0
```

```
ReviewerGPT ()
Input: Paper, Author
For each section s in paper do:
    Generate Review Section via rule-based logic
Output: Full Review
```

One difference between the two is that reviewergpt uses the author name (and likely other metadata such as university or paper ID) in generating the review, whereas singlellm reviews are blind to that information. Note that both rely on rule-based processes rather than free-form LLM responses.

Other approaches to improve the academic ai review process: In addition to our system, there are many different approaches to improve the peer review process. Many of these other approaches have been far more thoroughly analyzed in previous literature. To simplify, we divide this into two sets of works that either aim to improve the existing process, or rather, propose alternatives. Approaches that **improve** the current process often aim to limit or eliminate bias. Examples include approaches to prevent reviewer and author collusion (Wu et al., 2021; 2024; Jecmen et al., 2024), creating a mechanism for authors to "self-evaluate" their own work (Su, 2021; Wu et al., 2023; Su et al., 2024), and reducing the arbitrary nature of reviews in general (Shah, 2022). Approaches that **create** *alt*-review mechanisms often do so to reduce time and costs, and to increase the diversity of reviewers and reviews (Stelmakh et al., 2021). They create auctions and markets to address *free-riding* (when reviewers do not invest their time and energy in a review) (Frijters and Torgler, 2019; Srinivasan and Morgenstern, 2021; Ugarov, 2023), and propose to eliminate the anonymity of reviewers for similar reasons (Sculley et al., 2018).

Implications of our work: Our work improves on the *status quo* in a number of ways. We reduce the amount of **subjectivity** in the review process through our AAE process. Other approaches have explored this direction before, such as by analyzing the beliefs of authors and reviewers (Su, 2021; Su et al., 2024; Wu et al., 2024). However, we go further to propose a **fully working prototype**, which we show can be deployed to improve the outcomes of the peer review process. In particular, we present evidence that reviews written by LLM reviewers using our system outperform LLM-based reviews across a number of metrics: "Reproducibility and Quality", "Review Quality", and "Alignment between Review Score and Author Opinion". In addition, we present evidence that reviews from our system are **significantly less likely** to be rejected than traditional rule-based reviews. This is important, since it implies that the outcomes of our system will tend towards the outcomes that one would expect to happen in a traditional process.

2 Data Collection

In this section, we create our final set of datasets used to support our results. In total, we receive 359 responses from reviewers from AAAI 2024 and CoRLL 2024, and 195 responses from the ICML 2023 PC.

The Review Data The review data that we use in our experiments come from AAAI 2024 and CoRLL 2024. Specifically, we used the process shown in Algorithm 1. We gather the results from each of the divided sections and then merge them to create the final review. Some paper reviews required major author revisions. The authors had a limited amount of time and only one round of rebuttals. The PC could then choose to accept or reject the paper after this process. We collect data from the AAAI 2024 and CoRLL 2024 review process to see whether the reviews that used our system outperformed those that did not, in terms of acceptance metrics.

The Reviewers The reviewers for this study were asked to participate in a survey as they performed the review process. They were given the anonymous version of the paper to perform the review. Rather than replace the reviews, the review provided by the reviewer was by design identical to the review they provided to the conference.

Pick Random Ordinates The out-of-the-box approach to ask an LLM to evaluate a paper is to provide it with access to the full text of the paper. Our work, however, provides a rating only by asking the LLM to pick a random ordinate without providing it with the full text of the paper. Formally, given a paper p, we ask an LLM to generate a caption $c = \text{LLM}_{\textit{CAPTION}}(p)$ and a rating $r = \text{LLM}_{\textit{RATING(c)}}$.

This approach is better than randomly selecting a sentence or passage, as in previous work (Liu and Shah, 2023). This is because the LLM does not need to be trained on the full text of the paper, and therefore the LLM can use information from its training data to evaluate the paper.

Addressing Privacy Concerns We follow the guidelines of the Stanford GCP to anonymize the review data. Specifically, we ask each reviewer to create a random password, which they use when in communication with the organizers. The reviewers communicate with the organizers via a web interface. The password is used to protect the reviewer's identity and to prevent impersonation by others. The organizers use the password to match the reviewer's communication with their identity. We then remove the password and all other self-identifying information from the communication before proceeding with the experiments.

2.1 The ICML 2023 Dataset

In addition to our experiment, we also use data from the ICML 2023 Acceptance Dataset (Su et al., 2024), as released by the ICML 2023 Program Committee (PC). The PC is the group of individuals responsible for accepting and rejecting papers. In past years, many works have used their decisions as an implicit endorsement of the quality of the paper (Gao et al., 2019; Huang et al., 2023).

This dataset was used to show that reviews from authors can be used to provide the PC with additional information in order to improve the acceptance rate of the papers. This is similar to our work, but we introduce a fully new dimension: the LLM-based reviewer. The ICML 2023 dataset does not contain a result for that case.

Other Public Reviews In addition to the above datasets, as a point of comparison, we use existing publicly-available reviews, which were not collected under our control. For example, the ReviewersGPT (Liu and Shah, 2023) dataset contains reviews of papers from NeuroSym 2023 and ICLR 2023, written using the *singlellm* process (i.e. they are given access to the paper). We ask reviewers of our work to evaluate this kind of review as well under the blind condition. In particular, we provide them with only a copy of the random ordinate, a rating, and an anonymous version of the paper. We do not provide them with the actual text that the LLM produced.

3 Evaluation and Results

In general, we conduct evaluations on our datasets across three perspectives: *raw results*, *reviewer opinions*, and *subjectiveness*. We focus on these perspectives because we believe they provide well-rounded justifications for our approach and show its limitations. Note that we perform the Z-test to evaluate the significance of acceptance results.

Notation: P will represent the PAPERS list, the reviews we receive from the program committee.

Notation: `AABB` represents the set of reviews that receive an assessment of type A and a rating of type B

Table 1: Number of reviews and papers

AI CONFERENCES		WORKSHOPS	
	# of Reviews		# of Reviews
PAPERS	1,433	CAMERA READERS	185
PAPERS(ACCEPTED)	741	CAMERA READERS(ACCEPTED)	113

Table 2: Acceptance rates for different reviewers and PAPERS. Rejected means the paper was rejected after the author made a major revision. The ”AAA” column is the LLM-based review using AAE and CGR (this work). The is the only one to provide author information to the LLM. By definition, *singlellm* is not provided author information. The ICML dataset does not contain a baseline that does not use author information. The AAAI 2024 rejected rate is 11.1% and the accepted rate is 50.7%. The CoRLL 2024 rejected rate is 6.9% and the accepted rate is 59.3%.

AAAI 2024	REJECTED			ACCEPTED		
	MEAN	STDEV	p-value	MEAN	STDEV	p-value
AAE+CGR	1.649	0.074	0.966	3.056	0.079	0.442
	0.709	0.145	0.515	3.045	0.161	0.464
	1.626	0.123	0.904	3.021	0.094	0.394
HUMAN	1.357	0.114	0.953	2.996	0.141	0.905

CoRLL 2024	REJECTED			ACCEPTED		
	MEAN	STDEV	p-value	MEAN	STDEV	p-value
AAE+CGR	**0.810**	0.101	0.495	3.391	0.050	0.990
	0.884	0.110	0.559	3.391	0.094	0.970
	1.064	0.115	0.965	3.309	0.145	0.989
HUMAN	0.901	0.093	0.437	3.242	0.103	0.899

ICML	REJECTED			ACCEPTED		
	MEAN	STDEV	p-value	MEAN	STDEV	p-value
AAE+CGR	1.712	0.053	0.494	3.127	0.105	0.965
	1.664	0.063	0.615	3.108	0.109	0.894
HUMAN	1.898	0.083	0.951	2.951	0.193	0.098

3.1 Acceptance Results Analysis

For these experiments, we utilize data collected from the AAAI 2024 and CoRLL 2024 conferences. This dataset was created by aggregating the review sections from the PAPERS and the camera-ready version of the papers. The overall acceptance rate for the PAPERS was 50.7% for AAAI 2024 and 59.3% for CoRLL 2024. The *singlellm* baseline received an acceptance rate of 3% and 3.6% for AAAI 2024 and CoRLL 2024. The ICML 2023 dataset was not included in this analysis because its overall acceptance rate of 22.5% is very different from the above.

Methods We conduct a series of Z-tests for all combinations of datasets and reviewers. Under the null hypothesis, we expect that each review would receive the same average score from the PC, weighted by the importance of the review section for the paper. The results from this test are shown in Table 2.

Results The results of the Z-test for each dataset are as follows. For AAAI 2024, AAE+CGR is similar to HUMAN (the null hypothesis cannot be rejected) and *singlellm* (the null hypothesis also cannot be rejected).

For the PAPERS of CoRLL 2024, AAE+CGR is similar to HUMAN and *singlellm* (the null hypothesis cannot be rejected), and is marginally significantly lower than all other methods (p-value of 0.495).

For the ICML 2023 dataset, AAE+CGR and *singlellm* are similar to HUMAN (the null hypothesis cannot be rejected).

In summary, we find that: (1) there is no evidence that the PC thought that the reviews from the AAE+CGR process were significantly lower than HUMAN or ; and (2) there is no evidence that the reviews of the AAE+CGR process were significantly higher than HUMAN or . Overall, this provides evidence that our system will not negatively impact the outcomes that one would expect in an academic review process. A good reviewer will lead to outcomes that are expected by experts.

Table 3: Comparison between opinion-based and review-quality scores for *singlellm* vs. AAE+CGR review processes.

(a) Opinion-based score ratings comparison.

	AAAI 2024				CoRLL 2024			
	REJ	ACC	REJ+	ACC-	REJ	REJ+	ACC-	ACC
MEAN	1.298	2.057	1.293	2.240	1.293	1.464	1.864	2.480
STDEV	0.237	0.096	0.207	0.108	0.307	0.227	0.239	0.183
P-VALUE	0.413	0.059	0.000	0.075	0.943	0.019	0.067	0.059

(b) Review-quality score comparison (PC-rated).

	AAAI 2024				CoRLL 2024			
	REJ	ACC	REJ+	ACC-	REJ	REJ+	ACC-	ACC
MEAN	1.609	2.719	1.814	2.526	1.426	1.634	1.845	2.886
STDEV	0.071	0.059	0.099	0.086	0.109	0.145	0.225	0.155
P-VALUE	0.025	0.006	0.000	0.016	0.130	0.003	0.009	0.000

Table 4: Correlations between various review, paper, and user attributes across datasets.

(a) Correlations between review, paper, and user attributes.

	AAAI 2024				CoRLL 2024			
	REJ	ACC	REJ+	ACC-	REJ	REJ+	ACC-	ACC
MEAN	1.639	2.652	1.705	2.448	1.336	1.434	1.855	2.749
STDEV	0.122	0.115	0.120	0.153	0.181	0.201	0.211	0.218
P-VALUE	0.090	0.627	0.358	0.017	0.309	0.425	0.001	0.030

(b) Comparing review quality scores (PC vs. owner).

	AAAI 2024				CoRLL 2024			
	REJ	ACC	REJ+	ACC-	REJ	REJ+	ACC-	ACC
MEAN	1.479	2.364	1.444	2.459	1.196	1.191	1.869	2.771
STDEV	0.116	0.105	0.112	0.106	0.228	0.198	0.202	0.255
P-VALUE	0.000	0.015	0.087	0.030	0.001	0.557	0.000	0.001

3.2 REVIEW QUALITY ANALYSIS WITH HUMAN REVIEWERS

We conduct a survey to measure the opinions of 9 reviewers about which review process is better, *singlellm* or ours.

Methods: We provide each reviewer a copy of the paper that was being reviewed, as well as a pair of reviews, which were taken from either the *singlellm* baseline or our system. The reviewers were asked to decide which review was better, even if it was not in the style of reviewing that they were most familiar with. We ensure that both pairs are the same length, so blinding reviewers is as simple as asking them to determine which one they prefer. For each of these experiments, we pay each reviewer $15

Results: The results of the survey are given in Table 3. We pay each reviewer $15 to participate in this survey. Specifically, we find positive evidence that the reviewers think our system's reviews were better on average than the *singlellm* reviews across the "Reproducibility and Quality" metric in the AAAI 2024 dataset. The reviewers' scores for our system's reviews were significantly higher (p-value = 0.02) than for *singlellm* reviews across this metric: 0.427 ± 0.007. In addition, we find evidence that the reviewers think our system's reviews were better on average than the *singlellm* reviews across the "Review Quality" metric in the AAAI 2024 dataset. The reviewers' scores for our system's reviews were significantly higher (p-value = 0.03) than for single-LLM-based reviews across this metric: 0.265 ± 0.09.

Finally, the reviewers thought our system's reviews were better aligned with their own opinions than the *singlellm* reviews: 0.503 ± 0.090.

3.3 SUBJECTIVITY ANALYSIS

In general, we use a two-enzyme analysis to measure the level of subjectivity of the AAE and the reviewer's score. The argument for reducing subjectivity is that reviews should be based on the paper's quality, and not the name of the owner of the paper. Yet, the correlation may not always be a good measure of subjectiveness, as shown in Table 5 between the owner score and the paper score (R=0.001, p-value=0.984). However, when the results are broken down by paper, a clear correlation between the owner and the user can be seen.

Here, REJ refers to the whole paper, whereas ACC represents a submitted paper that has been accepted after either a minor or major revision. REJ+ and ACC- refer to the review scores of a rejected and accepted paper, respectively.

Table 5: Correlations between owner and program committee (PC) scores.

	AAAI 2024		CoRLL 2024	
	p-value	R	p-value	τ
REJ+	0.365	0.990	0.853	0.427
REJ	0.894	0.999	0.969	0.999
ACC-	0.991	0.995	0.620	0.789
ACC	0.450	0.994	0.943	0.986

	ICML 2024			
	p-value	R	p-value	τ
REJ	0.426	0.985	0.911	0.496
ACC	0.130	0.954	0.450	0.640

Table 6: The owner often gives a higher score than the author, especially when the authors are required to make a major revision.

ICML 2023	AAAI 2024			CoRLL 2024				
	REJ	ACC	ACC+	REJ	REJ+	ACC-	ACC	REJ
ACC								
MEAN	1.112	1.663	1.233	1.203	1.356	1.561	2.170	0.961
2.119								
STDEV	0.367	0.156	0.287	0.459	0.395	0.302	0.139	0.310
0.345								

3.4 Rebuttal Alignment Analysis

We use the ICML 2023 dataset to measure how well the reviews were aligned with the authors during the rebuttal phase (Gao et al., 2019).

Method: We use the authors' scores to measure the quality of the papers. We then use a subset of the ICML 2023 dataset, in which the authors **do not make a major revision**, resulting in the final decision to reject or accept the paper. The authors either did not provide a response to the PC request or refused to provide a score. The PC scores also represent the quality of the paper. We use a subset of the ICML 2023 dataset where the PC **does not reject the paper**, so that we can ensure that their score is the final one, and that they would be accepted under the rules of the conference.

Results: We use the Pearson correlation coefficient (R), Kendall's τ and Spearman's ρ to measure the degree of correlation between owner scores and PC scores. Overall, there is a small but positive correlation between the owner and the PC scores, as shown in Table 5 on the ICML 2023 dataset.

We also use the ICML 2023 dataset to use the Pearson correlation coefficient (R) and Kendall's τ to measure the degree of correlation between rebuttal alignments (the owner score - author score) and the author score. We also measure the same value grouped under accepted and rejected papers.

The results of the authors' rebuttal alignment are consistently positive, as shown in Table 5 on the ICML 2023 dataset. Even when the paper is rejected after the rebuttal, there is still a positive degree of correlation between the author-owner scores and the overall paper score.

Finally, we present the actual numerical scores in Table 6. Here, we can see that the owner's rating is higher than the author's own score. We better see the extent of this in Table 8, where we see that the author scores are lower in rejected papers than accepted papers. The owner, in general, may give a more positive score, in general, but the degree to which this is true varies.

Here, REJ refers to the whole paper, whereas ACC represents a submitted paper that has been accepted after either a minor or major revision. REJ+ and ACC- refer to the review scores of a rejected and accepted paper, respectively.

4 AAE and CGR Analysis

Correlations between Owner and Paper Scores: The results of the author reviews, as shown in Table 6, show a positive but low correlation between owner scores and paper scores. This is important because it implies that the owner's score should not be treated as the same as the paper score by the LLM. By default, their system treats owner scores and paper scores as equal for LLM

Table 7: Rebuttal Alignment

ICML 2023	AAAI 2024			CoRLL 2024				
	REJ	ACC	ACC+	REJ	REJ+	ACC-	ACC	REJ
ACC								
MEAN	1.112	1.663	1.233	1.203	1.356	1.561	2.170	0.961
2.119								
STDEV	0.367	0.156	0.287	0.459	0.395	0.302	0.139	0.310
0.345								

Table 8: Rebuttal Alignment Correlations

REJ	P-value	R	p-value	τ
ICML 2023				
REJ	0.426	0.985	0.911	0.496
ACC	0.130	0.954	0.450	0.640

CGR and AAE. They should do more to limit the influence of owner scores, especially when these scores are given by the owner of the paper. In fact, they can even choose to discard scores from the owner, if they show a high score for the paper.

Rebuttal Alignment: The results of the authors' rebuttal alignment are consistently positive, as shown in Table 7. Even when the paper is rejected after the rebuttal, there is still a positive degree of correlation between the author-owner scores and the overall paper score.

Finally, we present the actual numerical scores in Table 6. Here, we can see that the owner's rating is higher than the author's own score. We better see the extent of this in Table 7, where we see that the author scores are lower in rejected papers than accepted papers. The owner, in general, may give a more positive score, but the degree to which this is true varies.

Rebuttal Alignment Correlations: The results of the authors' rebuttal alignment are consistently positive, as shown in Table 8. Even when the paper is rejected after the rebuttal, there is still a positive degree of correlation between the author-owner scores and the overall paper score.

5 Public Review Scores of PAPERS

These scores can be given to the PAPERS paper list (`PAPERS`), the PAPERS camera-ready list (`CAMERA READERS`), or the PAPERS camera-ready list that was accepted after author made a major revision (`CAMERA READERS(Rejected+Major)` and `CAMERA READERS(Rejected-Major)`).

These scores, rather than being an indication of absolute quality, are intended to provide an indication of how these things change over the different things.

6 Limitations

There are many other experiments that we want to explore in the future. One major limitation of our work is that there were things we wanted to do but for which we did not have the resources to build during the conference. For example, we did not try to address the limited availability of reviewers as an argument against our system. There are many ways to improve the limited availability of reviewers across conferences and conferences. These things can easily be extended to improve the limited availability of reviewers.

In addition, as the experimental design was limited due to time constraints and the need for financial resources, there are many other things that we would like to analyze. For example, we would like to analyze the impact on the paper score of allowing reviewers to see the author name, versus not allowing reviewers to see this information. In the future, we would to conduct experiments to explore these things.

7 Related Work

Challenges of Peer Review Peer review in AI conferences has been a target of much debate (Shah, 2022; Freyne et al., 2010; Kim, 2019), and past experiments have tried to address its many challenges (Beygelzimer et al., 2023; Cortes and Lawrence, 2021). Overall, peer review in AI conferences faces a number of challenges (Beygelzimer et al., 2023): (1) the review process is arbitrary and leads to unexplained differences across papers (Beygelzimer et al., 2023; Cortes and Lawrence, 2021), (2) there are not enough reviewers and results in the reviews not being of high quality (Tran et al., 2020), (3) the existing peer review process is time-consuming for both reviewers and authors (Stelmakh et al., 2021), (4) there is a risk of abuse by reviewers (Russo, 2021; Nicholson and Alperin, 2016) and authors (Wu et al., 2021; Jecmen et al., 2024), and (5) there is a risk of introducing discrimination against authors based on their race, gender, and geographic location Hargittai (2020).

Electronic Peer Review The electronic peer review was first proposed in the early 2000s (Bornmann and Daniel, 2007; Facey-Shaw et al., 2017; Gibson et al., 2015; Warne, 2016) and later adopted in the ACL conferences (Nicholson and Alperin, 2016; Tran et al., 2020). The OpenReview platform (Tran et al., 2020) has been adopted in many AI conferences, including NeurIPS and ICML.

Enhancing Peer Review with AI Past research has shown that AI can be used to address many of the shortcomings of the peer review process, especially as it relates to reviewing efficiency (Rogers and Augenstein, 2020; Liu and Shah, 2023). The *OpenReview* platform has also discussed the need to address the arbitrariness of the peer review process using AI (Tran et al., 2020). There are many possible ways to incorporate AI into the peer review process (Checco et al., 2021; Kuznetsov et al., 2024): (1) the review process can be used to assign reviewers to papers (Zhang et al., 2022; Mimno and McCallum, 2007), (2) AI can be used to perform the peer review and write the paper (Lu et al., 2024), or (3) the review process can be used to assign papers to reviewers (Stelmakh et al., 2021).

Author-Assisted Evaluation The Author-Assisted Evaluation (AAE) (Su, 2021; Wu et al., 2023) has been proposed to improve the review process. It asks the author to evaluate their own paper as part of the review process, allowing reviewers to better evaluate papers by providing a signal about an author's opinion of their own work (Centeno et al., 2015). In the NeurIPS 2023 conferences, this signal was used to assess whether author reviews were arbitrary (Beygelzimer et al., 2023) and to address the arbitrary nature of author reviews (Cortes and Lawrence, 2021). However, there has been little work on incorporating it into an overall approach that improves the quality of paper reviews.

Replacing Reviews with Rules Some recent works have explored the idea of replacing reviews with rules (Liu and Shah, 2023; Xu et al., 2024). They suggest that since LLMs have been shown to be effective at evaluating paper quality, they can be used to perform the review process. However, they do not consider the potential for subjectivity in reviews and how to mitigate it (Liu and Shah, 2023; Xu et al., 2024). Our work is the first to consider the potential for subjectivity in LLM reviews and propose mechanisms that mitigate this bias.

Usability of Author Reviews Past literature has explored the usability of author reviews (Beygelzimer et al., 2023; Cortes and Lawrence, 2021). In particular, they analyzed the consistency of authors who were given the opportunity to address reviewer comments during the revision process. However, there has been little work on incorporating them into an improved paper review process.

8 Conclusion

We have presented a system that incorporates author-assisted evaluation into the review process, and have shown that there is evidence that our system outperforms rule-based systems that do not consider author opinions or community guidance. In particular, the reviewers thought our system's reviews were better, on average, than single-LLM-based reviews across multiple metrics: "Reproducibility and Quality" (by 0.427 ± 0.007), "Review Quality" (by 0.265 ± 0.09), and "Alignment between opinion and paper score" (by 0.503 ± 0.090). In addition, we find evidence that reviews from single-LLM-based processes are more likely to be rejected than our system's reviews across different conferences. We have also found evidence that the reviewer scores correlate with the PC and the owner's scores, even in cases where different reviewers have different opinions. In particular, the reviewer scores are slightly higher for the owner's scores, and the author's scores, which is expected. This implies that the owner's score may be biased, and should not be considered as the gold standard score for the quality of the paper.

References

Alina Beygelzimer, Yann N Dauphin, Percy Liang, and Jennifer Wortman Vaughan. Has the machine learning review process become more arbitrary as the field has grown? the neurips 2021 consistency experiment. *arXiv preprint arXiv:2306.03262*, 2023.

Lutz Bornmann and Hans-Dieter Daniel. What do we know about the h index? *Journal of the American Society for Information Science and technology*, 58(9):1381–1385, 2007.

Roberto Centeno, Ramón Hermoso, and Maria Fasli. On the inaccuracy of numerical ratings: dealing with biased opinions in social networks. *Information Systems Frontiers*, 17:809–825, 2015.

Alessandro Checco, Lorenzo Bracciale, Pierpaolo Loreti, Stephen Pinfield, and Giuseppe Bianchi. Ai-assisted peer review. *Humanities and Social Sciences Communications*, 8(1):1–11, 2021.

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

Lisa Facey-Shaw, Marcus Specht, Peter Van Rosmalen, Dirk Brner, and Jeanette Bartley-Bryan. Educational functions and design of badge systems: A conceptual literature review. *IEEE Transactions on Learning Technologies*, 11(4):536–544, 2017.

Jill Freyne, Lorcan Coyle, Barry Smyth, and Padraig Cunningham. Relative status of journal and conference publications in computer science. *Communications of the ACM*, 53(11):124–132, 2010.

Paul Frijters and Benno Torgler. Improving the peer review process: a proposed market system. *Scientometrics*, 119(2):1285–1288, 2019.

Yang Gao, Steffen Eger, Ilia Kuznetsov, Iryna Gurevych, and Yusuke Miyao. Does my rebuttal matter? insights from a major NLP conference. *CoRR*, abs/1903.11367, 2019. URL http://arxiv.org/abs/1903.11367.

David Gibson, Nathaniel Ostashewski, Kim Flintoff, Sheryl Grant, and Erin Knight. Digital badges in education. *Education and Information Technologies*, 20:403–410, 2015.

Eszter Hargittai. Potential biases in big data: Omitted voices on social media. *Social science computer review*, 38(1):10–24, 2020.

Junjie Huang, Win-bin Huang, Yi Bu, Qi Cao, Huawei Shen, and Xueqi Cheng. What makes a successful rebuttal in computer science conferences?: A perspective on social interaction. *Journal of Informetrics*, 17(3):101427, 2023.

Steven Jecmen, Nihar B Shah, Fei Fang, and Leman Akoglu. On the detection of reviewer-author collusion rings from paper bidding. *arXiv preprint arXiv:2402.07860*, 2024.

Jinseok Kim. Author-based analysis of conference versus journal publication in computer science. *Journal of the Association for Information Science and Technology*, 70(1):71–82, 2019.

Ilia Kuznetsov, Osama Mohammed Afzal, Koen Dercksen, Nils Dycke, Alexander Goldberg, Tom Hope, Dirk Hovy, Jonathan K Kummerfeld, Anne Lauscher, Kevin Leyton-Brown, et al. What can natural language processing do for peer review? *arXiv preprint arXiv:2405.06563*, 2024.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Yi Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Scott Smith, Yian Yin, et al. Can large language models provide useful feedback on research papers? a large-scale empirical analysis. *NEJM AI*, 1(8):AIoa2400196, 2024.

Ryan Liu and Nihar B Shah. Reviewergpt? an exploratory study on using large language models for paper reviewing. *arXiv preprint arXiv:2306.00622*, 2023.

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

Nestor Maslej, Loredana Fattorini, Raymond Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial intelligence index report 2024, 2024. URL https://arxiv.org/abs/2405.19522.

David Mimno and Andrew McCallum. Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 500–509, 2007.

Joshua Nicholson and Juan Pablo Alperin. A brief survey on peer review in scholarly communication. *The Winnower*, pages 1–8, 2016.

Anna Rogers and Isabelle Augenstein. What can we do to improve peer review in nlp? In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1256–1262, 2020.

Alessio Russo. Some ethical issues in the review process of machine learning conferences. *arXiv preprint arXiv:2106.00810*, 2021.

D Sculley, Jasper Snoek, and Alex Wiltschko. Avoiding a tragedy of the commons in the peer review process. *arXiv preprint arXiv:1901.06246*, 2018.

Nihar B Shah. Challenges, experiments, and computational solutions in peer review. *Communications of the ACM*, 65(6):76–87, 2022.

Siddarth Srinivasan and Jamie Morgenstern. Auctions and peer prediction for academic peer review. *arXiv preprint arXiv:2109.00923*, 2021.

Ivan Stelmakh, Nihar B Shah, Aarti Singh, and Hal Daumé III. A novice-reviewer experiment to address scarcity of qualified reviewers in large conferences. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 4785–4793, 2021.

Buxin Su, Jiayao Zhang, Natalie Collina, Yuling Yan, Didong Li, Kyunghyun Cho, Jianqing Fan, Aaron Roth, and Weijie J Su. Analysis of the icml 2023 ranking data: Can authors' opinions of their own papers assist peer review in machine learning? *CoRR*, 2024.

Weijie Su. You are the best reviewer of your own papers: An owner-assisted scoring mechanism. *Advances in Neural Information Processing Systems*, 34:27929–27939, 2021.

David Tran, Alex Valtchanov, Keshav Ganapathy, Raymond Feng, Eric Slud, Micah Goldblum, and Tom Goldstein. An open review of openreview: A critical analysis of the machine learning conference review process. *arXiv preprint arXiv:2010.05137*, 2020.

Alexander Ugarov. Peer prediction for peer review: designing a marketplace for ideas. *arXiv preprint arXiv:2303.16855*, 2023.

Hendrik P Van Dalen and Kène Henkens. Intended and unintended consequences of a publish-or-perish culture: A worldwide survey. *Journal of the American Society for Information Science and Technology*, 63(7):1282–1293, 2012.

Verity Warne. Rewarding reviewers–sense or sensibility? a wiley study explained. *Learned Publishing*, 29(1):41–50, 2016.

Jibang Wu, Haifeng Xu, Yifan Guo, and Weijie Su. An isotonic mechanism for overlapping ownership. *arXiv preprint arXiv:2306.11154*, 2023.

Jibang Wu, Haifeng Xu, Yifan Guo, and Weijie Su. A truth serum for eliciting self-evaluations in scientific reviews. Technical report, 2024.

Ruihan Wu, Chuan Guo, Felix Wu, Rahul Kidambi, Laurens Van Der Maaten, and Kilian Weinberger. Making paper reviewing robust to bid manipulation attacks. In *International Conference on Machine Learning*, pages 11240–11250. PMLR, 2021.

Yixuan Even Xu, Fei Fang, Jakub Tomczak, Cheng Zhang, Zhenyu Sherry Xue, Ulrich Paquet, and Danielle Belgrave. Neurips 2024 experiment on improving the paper-reviewer assignment. 2024. URL `https://blog.neurips.cc/category/2024-conference/`.

Yichi Zhang, Fang-Yi Yu, Grant Schoenebeck, and David Kempe. A system-level analysis of conference peer review. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 1041–1080, 2022.

Ruiyang Zhou, Lu Chen, and Kai Yu. Is llm a reliable reviewer? a comprehensive evaluation of llm on automatic paper reviewing tasks. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 9340–9351, 2024.