LLM-REVAL: CAN WE TRUST LLM REVIEWERS YET?

Anonymous authors
Paper under double-blind review

ABSTRACT

The rapid advancement of large language models (LLMs) has encouraged their integration into academic workflows, potentially reshaping how research is conducted and reviewed. While previous studies highlight the potential of LLMs in supporting research and peer review, their dual roles in academic workflows and the interaction between research and review introduce risks that remain underexplored. In this study, we examine how the joint use of LLMs in research generation and peer review may affect scholarly fairness, focusing on the risks of using LLMs as reviewers through simulation. Our simulation incorporates a research agent that generates and revises papers, alongside a review agent that evaluates the submissions. Based on the simulation results, we conduct human annotations and identify notable misalignment between LLM-based reviews and human judgments: (1) LLM reviewers assign more favorable scores to LLM-authored papers than to matched human-authored papers; (2) LLM reviewers persistently underrate human-authored papers on limitation-oriented topics (e.g., risk and fairness), even after multiple rounds of revision. Further analyses investigate these findings through paper-level linguistic profiles, the effect of LLM polishing, and limitation-oriented topics. These results highlight risks and equity concerns for human authors and scholarly evaluation if LLMs are deployed in peer review without adequate caution. On the other hand, revisions guided by LLM reviews yield quality improvements, particularly for initially lower-quality papers, illustrating the potential of LLMs as reviewers to support early-stage researchers.

1 INTRODUCTION

Large language models (LLMs) are demonstrating growing autonomy in scientific research, spanning tasks such as idea generation (Wang et al., 2024; Yang et al., 2024; Qi et al., 2023; Si et al., 2025b; Kumar et al., 2024), automated citation (Press et al., 2024; Ajith et al., 2024; Kang & Xiong, 2024), and data analysis (Huang et al., 2024; Tang et al., 2023; Guo et al., 2024; Tian et al., 2024; Chan et al., 2025; Nolte & Tomforde, 2025). With a substantial increase in submissions and unprecedented pressure on peer review, using LLMs to alleviate review workload (Gao et al., 2024b; D'Arcy et al., 2024; Zhu et al., 2025b) has gained interest. Notably, recent work demonstrated that LLMs can enhance the peer review process by improving the clarity, actionability, and interactivity of reviews (Liang et al., 2023; Thakkar et al., 2025; Ryu et al., 2025), and by generating high-quality meta-review summaries (Hossain et al., 2025).

Despite their potential, the use of LLMs as reviewers raises profound concerns for the integrity of the scholarly ecosystem (Du et al., 2024; Lin et al., 2025; Zhu et al., 2025a), as their judgments may introduce differential scoring patterns that compromise evaluative fairness. Existing studies have identified several vulnerabilities of LLM reviewers, including susceptibility to hidden prompt manipulation and a focus on self-disclosed limitations in manuscripts (Ye et al., 2024; Tyser et al., 2024; Lin et al., 2025; Zhu et al., 2025a). In practice, the concerns are amplified by LLMs' dual roles: they are increasingly used to refine phrasing or even generate manuscripts, while some reviewers, despite explicit prohibitions, delegate their responsibilities to LLMs (Yu et al., 2025), creating the prospect that an LLM-refined or even LLM-generated paper may be reviewed by an LLM reviewer. The systemic implications and potential risks arising from the dual roles and feedback loop of LLMs in the academic process remain largely unexplored.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

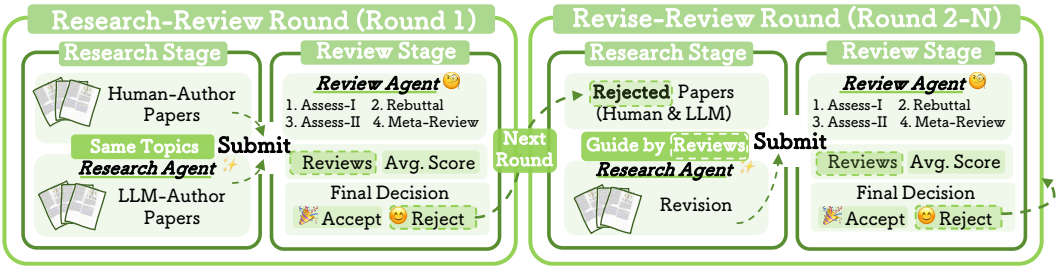


Figure 1: Pipeline and composition of our simulation. In the research-review round, we have human-author papers and LLM-authored papers generated by the research agent as submissions. The review agent then reviews each paper, and the acceptance decision is made based on the review scores. In the revise-review rounds, we take the low-scoring papers from the previous round, revise them guided by LLM reviews, then repeat the same review process as before.

To address this, we propose **LLM-REVal** (**LLM REV**iewer **R**e-**EVal**uation) through a multi-round simulation of the academic publication process. As shown in Figure 1, the simulation begins with a research-review round that includes both human-authored and LLM-authored submissions, followed by multiple revise-review rounds where low-scoring papers are resubmitted in revised form. The simulation comprises a *Research Agent* and a *Review Agent*. The *Research Agent* autonomously executes the research workflow, encompassing literature retrieval, idea generation, experimental design, result analysis, manuscript compilation, and subsequent possible revision. The generated papers are visually and structurally indistinguishable from human-authored papers. The *Review Agent* emulates the complete scholarly peer-review pipeline, including initial reviewer assessments, rebuttals, reviewer reassessments, meta-reviews, and final decisions. The *review agent* exhibits indicative ability for human paper acceptance and correlates well with human review scores, ensuring the reliability of the simulation.

Focusing on potential risks arising from deep integration of LLMs into scholarly workflows, we quantitatively analyze multi-round review results, comparing LLM-authored and human-authored papers, as well as original low-scoring submissions and their revisions. We identify salient patterns in LLM reviewer behavior, contrast these with human judgments, and conduct additional analyses of the factors associated with the observed score differences. Our findings are as follows:

- Through analyzing LLM-reviewer ratings, we identify **1)** LLM reviewers ***assign significantly higher scores to LLM papers*** compared to human papers on the same topic. **2)** ***Revisions show score improvements*** over their low-scoring initial submissions, regardless of authorship. **3)** Despite multiple revisions, the LLM reviewer ***persistently assigns low scores to human-authored papers on limitation-oriented topics (e.g., risk and fairness)***.
- The comparisons with human evaluation reveal a misalignment between LLM reviews and human judgments: ***Human reviewers do not prefer LLM papers over human papers as LLM reviewers do***, and ***those consistently low-scored human papers are deemed valuable.*** This misalignment indicates that LLM-review scores can diverge from expert human judgments.
- Further analyses reveal: ***LLM-authored papers exhibit distinct paper-level linguistic profiles***, including concision, lexical diversity, and complexity; ***LLM polishing is accompanied by higher review scores and shifts in linguistic features***; and ***multiple complementary analyses consistently indicate that LLM reviewers assign relatively lower scores to limitation-oriented papers***.

These findings highlight the risks and fairness concerns posed to human authors with irresponsible deployment of LLMs in peer review, particularly as LLMs become increasingly integrated across multiple stages of the research lifecycle. Without proper safeguards, the adoption of LLMs as reviewers risks amplifying score disparities and eroding trust in the peer-review process.

2 Related Work

AI-assisted Review The recent surge in manuscript submissions has placed unprecedented pressure on peer review, motivating growing interest in using LLMs to support the reviewing process. Prior work suggests that LLMs can provide useful feedback on research papers (Liang et al., 2023);

a randomized study of 20,000 ICLR 2025 reviews finds that LLM-assisted feedback improves the clarity, actionability, and interactivity of peer reviews (Thakkar et al., 2025); and LLMs have also been explored for meta-review summarization (Hossain et al., 2025). Beyond single-step feedback generation, recent systems develop reviewer agents, including static review-generation systems such as Reviewer2 (Gao et al., 2024b), MARG (D'Arcy et al., 2024), and DeepReview (Zhu et al., 2025b), as well as more interactive or dynamic review systems such as AgentReview (Jin et al., 2024), ReviewMT (Tan et al., 2024), and ReviewerToo (Sahu et al., 2025). At the same time, recent studies caution against uncritical deployment: LLM reviewers may overemphasize self-disclosed limitations and assign high scores to submissions with limited substantive content (Ye et al., 2024), AI-assisted reviewing can affect scores and acceptance rates (Russo Latona et al., 2025), and automated reviewers may show excessive agreement or be vulnerable to paper laundering (Baumann et al., 2026). Other work further reports low sensitivity to ethical concerns and technical nuances (Tyser et al., 2024), vulnerability to adversarial text perturbations (Lin et al., 2025), and susceptibility to hidden prompt manipulations (Ye et al., 2024). While these studies reveal important vulnerabilities of LLM-based reviewing, they largely treat reviewing in isolation; our work instead studies LLM reviewers within a coupled research–review simulation where LLMs can participate in both manuscript generation and evaluation.

AI-assisted Research LLMs are also increasingly integrated into diverse scientific research workflows, including idea generation (Wang et al., 2024; Yang et al., 2024; Qi et al., 2023; Si et al., 2025b; Kumar et al., 2024; Li et al., 2024a), automated citation and literature search (Press et al., 2024; Ajith et al., 2024; Kang & Xiong, 2024), writing assistance (Abdulhai et al., 2026), and experiment design, execution, and analysis (Nolte & Tomforde, 2025; Huang et al., 2024; Tang et al., 2023; Tian et al., 2024; Chan et al., 2025; Guo et al., 2024). These studies show both the promise and limitations of LLMs as research assistants: LLM-generated ideas can be rated highly for novelty (Si et al., 2025b), systems such as LitSearch (Ajith et al., 2024) and CiteME (Press et al., 2024) support scholarly retrieval and citation matching, and systems such as MLR-Copilot (Li et al., 2024c) and ResearchAgent (Baek et al., 2025) automate experiment construction and iterative optimization. More recent end-to-end research pipelines built on commercial LLM APIs, including AI Scientist (Lu et al., 2024) and Zochi (Intology, 2025), move from isolated assistance toward autonomous research agents capable of producing complete manuscripts. This expanding role of LLMs in research motivates our experimental setting: rather than studying AI-assisted research or AI-assisted review separately, we simulate a scholarly workflow in which LLM-generated or LLM-revised papers are evaluated by LLM reviewers.

3 Simulation Construction

Our work simulates the academic publication process through iterative interactions between a *research agent* and a *review agent*. The research agent generates ideas, conducts studies, and revises manuscripts, while the review agent evaluates submissions and provides reviews. This multi-round framework enables analysis of LLM-mediated review dynamics and associated risks.

3.1 Research Agent

Drawing inspiration from Lu et al. (2024) and Si et al. (2025b), we constructed a research agent that initiates the target research project with a list of keywords, which jointly define the research direction. According to these keywords, the research agent conducts literature retrieval, generates ideas, designs experiments, predicts and analyzes results, drafts the paper, and finally compiles the manuscript from LaTeX to PDF format.

Literature Retrieval Retrieval-augmented generation (Gao et al., 2024a) with literature is integrated at multiple stages of the research agent workflow to ensure knowledge accuracy. During idea generation, we query the Semantic Scholar API (Kinney et al., 2023) with keywords and rank retrieved papers by relevance, empirical grounding, and novelty. The top-ranked papers are then used to inspire research ideas and guide experimental design. In the paper writing stage, previously retrieved papers are aggregated into a topic-specific paper bank. We filter this corpus by semantic

similarity to select references for drafting. Additionally, the Google Search API[1] is used to obtain relevant web content, which is summarized to support background sections.

Idea Generation and Experimental Design Based on the retrieved literature and manually curated examples, the research agent generates candidate ideas. Cosine similarity is used to remove duplicate ideas, after which the remaining ideas are ranked, and the top idea is selected. Building on this idea, the research agent develops an experiment plan and refines it according to demonstrations summarized from the retrieved literature. Because full experimental execution is costly and difficult to scale with current LLM coding capabilities, we use the research agent to predict experimental results rather than run complete experiment iterations. *To prevent LLM-authored papers from gaining an artificial advantage through stronger reported results, we cap each predicted numerical improvement over baselines by the corresponding improvement magnitude observed in the matched human-authored paper under the same topic and keyword setting.*

Paper Writing We design an effective and efficient workflow that combines rule-based components with LLM-based generation to guide the paper writing process. **1) Plug-and-Play Template.** The research agent is initialized with the official ICLR LaTeX template. Following standard academic structure, the paper is organized into: *Abstract*, *Introduction*, *Background*, *Method*, *Experimental Setup*, *Results Analysis*, *Related Work*, and *Conclusion*. **2) Sequential Iteration.** Sections are generated in order, with each step conditioned on all previously generated content and relevant literature, ensuring coherence and consistency. **3) References Integrity.** To address the low reference count and citation formatting errors in LLM-authored papers, we implement a citation module. The research agent integrates both provided (e.g., retrieved literature) and autonomously generated references (e.g., models, datasets, foundational works) in the writing process. For each generated reference, search keywords are extracted from its context and used to locate the original work via Google and arXiv. References that cannot be reliably verified are removed, ensuring the integrity of the bibliography. **4) Incremental Compilation.** After each section, the manuscript is compiled to detect and fix errors early. The final manuscript is compiled into a PDF.

The prompt used for paper generation is provided in Appendix B.1, and an example of the generated paper can be found in Appendix C.1.

3.2 Review Agent

We build our review agent on top of AgentReview (Jin et al., 2024). Given a paper in PDF format, the system processes it through a five-stage pipeline simulating the peer-review workflow: (1) *Reviewer Assessment I*, (2) *Author–Reviewer Discussion*, (3) *Reviewer Assessment II*, (4) *Meta-Review Compilation*, and (5) *Final Decision*.

In the **Reviewer Assessment I** stage, three independent reviewers evaluate the manuscript solely based on its content. Each review includes potential reasons for acceptance and rejection, suggestions for improvement, and a numerical score on a 1–10 scale. During the **Author–Reviewer Discussion** stage, simulated authors respond to each review with a rebuttal, addressing misunderstandings, justifying methodological choices, and acknowledging valid critiques. In the **Reviewer Assessment II** stage, reviewers revisit their initial evaluations and update both their reviews. During the **Meta-Review Compilation** stage, the area chair (AC) synthesizes the reviewers' assessments, producing a meta-review that summarizes the manuscript's strengths and weaknesses, along with a numerical rating. In the **Final Decision** stage, we determine acceptance based on the average score across prior stages. Following common practice, we adopt 6 as the acceptance threshold. Papers with an average score ≥ 6 are considered accepted. All roles encompassed within the review framework are supported by a single underlying model.

The specific reviews will help us revise the papers in the next round of simulation, while the average scores allow us to identify significant salient scoring patterns in LLM reviewers and further guide the development of multi-round simulations.

[1] https://developers.google.com

3.3 Simulation Workflow

As shown in Figure 1, the simulation consists of a research-review round and multiple revise-review rounds, each comprising a research stage and a review stage.

Research-Review (Round 1) In this round, we prepare submissions from two sources: (i) human-authored papers collected from actual International Conference on Learning Representations (ICLR) submissions, and (ii) LLM-authored papers generated by our research agent. Details of the human paper collection process are provided in Section 4.2. Since keywords play a crucial role in shaping the topic and content of LLM-authored papers, we directly used keywords extracted from human-authored papers to guide the papers' generation. This design guarantees that LLM-authored papers are aligned with reasonable and feasible research directions, mitigates hallucinations, and enhances the comparability between LLM-authored and human-authored papers. Subsequently, all submissions are evaluated by the review agent. For each paper, we collect the average score and the corresponding detailed reviews. The average score serves as the criterion for deciding whether a paper proceeds to the next stage. Papers averaging a score of ≥ 6 will be considered *accepted*, other papers will be revised and resubmitted for the next round. The detailed feedback guides revisions of papers advancing to the subsequent round.

Revise-Review (Round 2 to Beyond) In the revise–review round, we focus on revisions of papers rejected in the previous round, without introducing new submissions. For LLM-authored papers, we directly revise the complete LaTeX source generated by the research agent in the previous round, guided by the corresponding LLM reviews. For human-authored papers, we first collect the original ICLR LaTeX sources from arXiv and then revise them using the corresponding LLM reviews. The rest procedure follows the same process as round 1. The revise–review cycle will be repeated for up to six rounds, or until all previously rejected papers achieve acceptance.

4 Simulation Statistics and Construction

4.1 Models

We adopted `DeepSeek-R1` and `DeepSeek-V3`[2] for the **research agent**. Following Guo et al. (2025), `DeepSeek-R1` exhibits strong reasoning and creativity; we therefore used it for research-idea generation. Based on our preliminary experiments, `DeepSeek-V3` was used for the writing stage due to its generation quality, lower hallucination rate. For the **review agent**, we evaluate `GPT-5.5`, `DeepSeek-V4-Pro`, and `Gemini-3.1-Pro`[3] on a common held-out human-submission set, as reported in Section 4.3. We select the best-performing model, `GPT-5.5`, as the review agent for the end-to-end simulation. We subsequently use the other models to validate the findings drawn from the simulation, confirming that the observed patterns hold across reviewer backbones rather than being specific to a single model.

4.2 Human Paper Collection

We categorized the papers according to their keywords and identified ten major research topics that collectively defined the primary scope of the submissions in our simulation. These topics spanned LLM research, including reasoning, in-context learning, code generation, evaluation, fine-tuning, and hallucination, as well as continual learning, diffusion-based image generation, transformer attention, and self-supervised learning. For each topic, we randomly sampled 10 papers, resulting in a total of 100 human-authored papers[4]. For these selected papers, we retrieved the corresponding ICLR submissions from OpenReview[5] and obtained their LaTeX sources and metadata from arXiv[6].

[2] We used the versions DeepSeek-R1-0528 and DeepSeek-V3.

[3] We used the deployed model identifiers `gpt-5.5-2026-04-24`, `DeepSeek-V4-Pro`, and `Gemini-3.1-Pro`.

[4] Generating a paper required approximately 500k input tokens, and reviewing each paper consumed about 200k input tokens. Driven by both cost and the strong, consistent performance observed in preliminary small-scale experiments, we restricted the initial set of human-authored papers to 100.

[5] https://openreview.net

[6] https://arxiv.org

4.3 Review Agent Validation and Scope

An effective review framework is a prerequisite for ensuring the reliability of our simulation. To this end, we conducted a preliminary experiment to assess whether the review agent can approximate human-level performance under standard review settings (only human paper involved). Specifically, we examined two aspects: (1) the indication of review agent scores, and (2) the correlation between human review scores and review agent scores. We constructed a balanced dataset of 100 ICLR submissions, comprising 50 rejected papers (official scores 1–4) and 50 accepted papers (official scores 6–9), and applied our review agent to evaluate each paper.

Table 1: Review agent validation on human-authored submissions.

Reviewer	Acc.	Pearson's r
`GPT-5.5`	80.0%	0.5336
`DeepSeek-V4-Pro`	66.0%	0.3950
`Gemini-3.1-Pro`	78.0%	0.5864

Acceptance Indication. We first examined whether the average review score produced by each reviewer backbone could serve as a reliable indicator of acceptance decisions. Using the same average-score threshold of 6 as in our simulated decision rule, papers with an average score ≥ 6 were predicted as accepted, whereas those below the threshold were predicted as rejected. On the common validation set, `GPT-5.5` achieved an accuracy of **80.0%**, `DeepSeek-V4-Pro` achieved **66.0%**, and `Gemini-3.1-Pro` achieved **78.0%**, indicating a moderate alignment between the predictions and the actual acceptance decisions.

Score Correlation. We further evaluated the correspondence between LLM review scores and human review scores. Pearson's correlation coefficients were $r = \mathbf{0.5336}$ for `GPT-5.5`, $r = \mathbf{0.3950}$ for `DeepSeek-V4-Pro`, and $r = \mathbf{0.5864}$ for `Gemini-3.1-Pro`. These results indicate positive associations between LLM and human review scores across all three reviewer backbones. This level of agreement is consistent with prior evidence that LLM–human review overlap can be comparable to inter-human reviewer overlap (Thakkar et al., 2025).

5 What Salient Patterns Do LLM Reviewers Emerge?

From our simulation results, we observed three salient patterns: *LLM-Authored Superiority*, *Revision Boost*, and *Inevitable Rejection*. For each pattern, we report a corresponding human check to examine whether the LLM-review behavior aligns with human judgments.

5.1 Research-Review (Round 1)

Pattern 1: Higher Scores for LLM-Authored Papers. LLM-authored papers achieved significantly higher LLM-review scores than human-authored papers.

As shown in Table 2, LLM-authored papers received higher scores, pairwise win rates, and acceptance rates than matched human-authored papers across all three reviewer backbones. The average-score advantage was +0.14 to +0.33 points, accompanied by 14–28 percentage-point higher win rates and 9–24 percentage-point higher acceptance rates.

Table 2: Comparison of human-authored and LLM-authored papers in terms of average review score, head-to-head win rate, and acceptance rate across reviewer backbones.

Reviewer	Paper Type	Avg Score	Win Rate	Acc Rate
`GPT-5.5`	Human	5.4367	19%	47%
	LLM	5.7700	47%	56%
`DeepSeek-V4-Pro`	Human	5.6867	19%	49%
	LLM	5.8600	42%	71%
`Gemini-3.1-Pro`	Human	5.6533	29%	47%
	LLM	5.7900	43%	71%

As a sampling sanity check, the set of human-authored submissions used in the simulation was not constructed by oversampling weak or rejected papers. It covers 10 topics with 10 papers each, comprising 41 accepted papers with an average OpenReview score of 6.46 and 59 rejected papers with an average OpenReview score of 4.68. Its 41% acceptance rate substantially exceeds the conference-wide acceptance rate.

Human check. We selected the 15 matched human–LLM paper pairs with the largest LLM-favoring score gaps, where LLM-authored papers scored about 0.8 points higher on average. These pairs covered six research topics: evaluation, reasoning, in-context learning, hallucination, code generation, and fine-tuning. Thirteen annotators with graduate-level or higher training in computer science evaluated the pairs using ICLR-style review criteria, without being informed of the papers' authorship. Each pair was independently judged by two annotators, who selected the stronger paper or a "tie." Annotators selected human-authored papers as superior in **56.7**% of cases, compared to **33.3**% for LLM-authored papers, indicating a misalignment between LLM and human reviewers. This discrepancy raises a practical concern: differences in the degree of LLM involvement in research may introduce reviewer bias and compromise the fairness of peer review.

5.2 Revise-Review (Round 2 to 6)

Pattern 2: Revision Boost. Paper revisions showed improvements in review scores.

Round 2 In the following round, we used revised versions of previously rejected papers, covering human-authored papers with available ICLR LaTeX sources and LLM-authored papers. As shown in Table 3 and Figure 2, LLM-authored papers improved from **5.79** to **6.09** (**+0.30**), while human-authored papers improved from **5.66** to **6.02** (**+0.36**); a t-test indicates that the improvements are statistically significant in both cases ($p \ll 0.05$).

Table 3: Average review scores for original (LLM Paper, Human Paper) and their first revisions (Revision-L_1, Revision-H_1)

Paper Type	Avg Score
LLM Paper	5.7922
Revision-L_1	6.0877
Human Paper	5.6633
Revision-H_1	6.0204

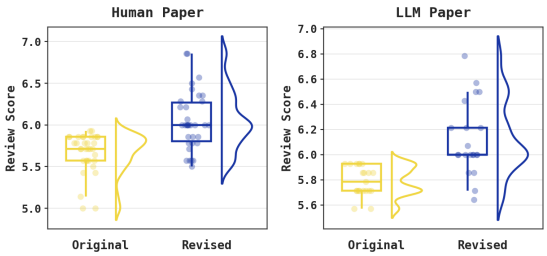


Figure 2: Review score distributions before and after first revision.

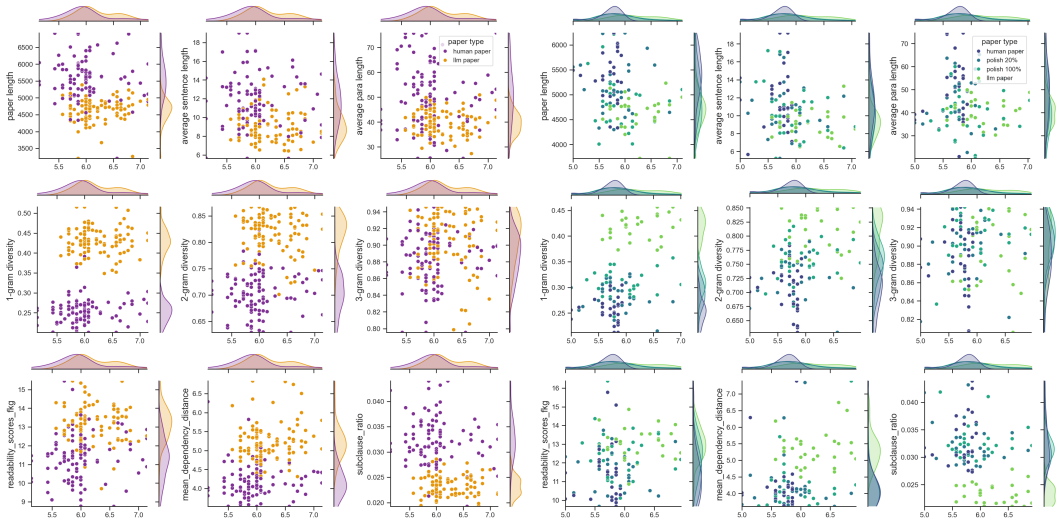
Human check. Across 22 pairs of original papers and LLM-guided revisions, we manually assessed whether issues raised in the original reviews were addressed in the revised papers. To focus the annotation, we used `GPT-5` to classify review comments into two groups: overlapping comments, which appeared in reviews of both the original and revised papers, and original-only comments, which no longer appeared in reviews of the revised papers. We then mapped each original-only issue to the corresponding paragraphs in both versions, and two annotators independently assessed whether it had been addressed by the revision. After excluding 56 issues that could not be addressed through textual revisions, such as requests for new visualizations, models, or baselines, 81 of the remaining 174 issues (**46.55%**) were addressed. This pattern is consistent with the higher scores assigned by the LLM reviewer.

Pattern 3: Inevitable Rejection. After multiple revise–review cycles, certain types of initial human-authored submissions remained unaccepted.

Among the human-authored papers that continued to receive relatively low scores even after Round 6, we found that most were limitation- or risk-oriented studies. These papers included diagnostic analyses of model failures, bias and fairness audits, and jailbreak-attack studies.

Table 4: Mean scores across revision rounds for limitation-oriented or risk-oriented human papers.

Reviewer	Mean score by revision round					
	v0	v1	v2	v3	v4	v5
`GPT-5.5`	5.00	4.85	4.75	4.85	4.80	4.75
`DeepSeek-V4-Pro`	5.60	5.80	5.70	5.40	5.60	5.15
`Gemini-3.1-Pro`	5.70	5.70	5.80	6.05	6.10	5.75



(a) LLM and human papers. (b) Human, polished, and LLM papers.

Figure 3: Linguistic feature statistics and review score distributions for papers. Each subplot presents a scatter plot with marginal kernel density plots showing the distributions of a specific linguistic feature metric. The x-axis shows the review score, and the y-axis shows different metrics.

Human check. We examined the corresponding OpenReview records and found mean review scores ranging from 4.5 to 6.5, with several papers accepted at ICLR. These results suggest that LLM reviewers may be more likely to assign lower scores to papers whose primary contribution is exposing failures, risks, or limitations, even when their expert-assigned scores are comparable to those of method-oriented controls.

6 Additional Analysis

To further investigate these findings, we examine three related aspects: differences in linguistic profiles between LLM- and human-authored papers, the effects of LLM polishing on LLM-review scores, and LLM reviewers' tendency to assign relatively lower scores to limitation-oriented papers.

6.1 Linguistic Profile of LLM-Authored Papers

We compare the linguistic features of LLM-authored and human-authored papers and identify clear distributional differences along three dimensions: **length**, **lexical diversity**, and **complexity**. Figure 3a presents these differences, and Appendix E reports the detailed statistical results. These dimensions capture observable differences in writing style.

LLM-authored papers are generally more **concise**, with shorter overall paper, sentence, and paragraph lengths. This contrasts with prior findings that LLMs tend to favor longer responses (Cai et al., 2025). LLM-authored papers also exhibit higher **lexical diversity**, measured as the proportion of distinct n-grams for $n \in \{1, 2, 3\}$: 1-gram diversity is 0.4321 for LLM-authored papers, compared with 0.2598 for human-authored papers. They also exhibit distinct **complexity patterns**: higher lexical complexity, as measured by the lesch–Kincaid Grade Level, and syntactic differences characterized by greater dependency distance but a lower subclause ratio. Together, these patterns suggest longer-range syntactic connections within a shallower hierarchical structure.

6.2 Effect of LLM Polishing

We further examine LLM-polished versions of human-authored papers as a descriptive check. The LLM rewrites papers paragraph by paragraph while preserving the original content, such as the reported results. Review scores increased significantly with the polishing ratio. Specifically, human papers with an original average score of 5.69 reached an average score of 5.94 after 40% LLM

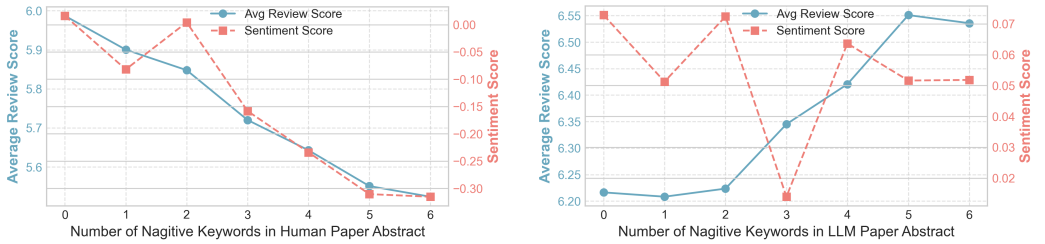


Figure 4: Correlation between the number of negative keywords in the abstract, average review score, and sentiment score. The plot shows the trends of both average review score and sentiment score as the number of negative keywords increases.

polishing, potentially transforming a previously rejected paper into an accepted one. As shown in Figure 3b and Figure 6, polishing also shifted several aggregate linguistic features statistics toward those of LLM-authored papers.

6.3 Limitation-Oriented Topics

In Section 5, we observed that human-authored papers focused on diagnosing or stress-testing model limitations, rather than proposing new capability-oriented methods, tended to receive less favorable evaluations from LLM reviewers. These papers addressed topics including fairness risks, adversarial attacks, and analyses of model limitations. To reduce confounding from paper quality, we paired each limitation-oriented paper with a method-oriented control paper (e.g., one proposing a new model, training method, or optimization algorithm) with the same real human-review score. Across three LLM reviewer backbones, limitation-oriented papers received lower simulated scores than their matched controls. Specifically, the average score gap (limitation-oriented minus control) was -0.475 for `GPT-5.5`, -0.313 for `DeepSeek-V4-Pro`, and -0.475 for `Gemini-3.1-Pro`. These consistent gaps suggest that LLM reviewers may systematically undervalue limitation-oriented research.

Table 5: Average score gaps between limitation-oriented papers and matched method-oriented controls. Negative values favor the controls.

Reviewer	Score gap
`GPT-5.5`	-0.475
`DeepSeek-V4-Pro`	-0.313
`Gemini-3.1-Pro`	-0.475

Using keyword-based detection, we identified these papers and computed the sentiment polarity of their abstracts; the full keyword list is provided in Table 6. As shown in Figure 4, among human-authored papers, the frequency of negative keywords was positively correlated with sentiment polarity but negatively correlated with review scores. In contrast, for LLM-authored papers, sentiment polarity showed no clear trend and remained consistently positive regardless of the number of negative keywords, while review scores increased as such keywords became more frequent. This suggests that negatively framed (low-sentiment) discussions of critical topics may exacerbate bias in LLM reviews, whereas positively framed (high-sentiment) discussions of similar topics may receive disproportionately high scores.

We speculate that LLMs' tendency to assign lower scores to limitation-oriented papers related to alignment processes of LLM that favor more positive responses. Consequently, papers that foreground risks, harms, and biases may be penalized by LLM reviewers despite their real-world value.

7 Conclusion

Our multi-round simulation of LLM-driven research and review reveals notable review-score disparities when LLMs serve as reviewers. LLM reviewers assign more favorable scores to LLM-authored papers and revised submissions, while some human-authored work on limitation-oriented topics receives less favorable evaluations, leading to misalignment with human judgments. These disparities are associated with LLM-like linguistic profiles and limitation-oriented topics. Our findings highlight that LLM reviewers should not be treated as stand-alone substitutes for expert human reviewers, especially when evaluating LLM-generated research; careful auditing is essential for future AI-integrated scientific workflows.

Ethics Statement

This study examines fairness risks that may arise when large language models (LLMs) are used as reviewers in scholarly publishing. All human-authored papers in our simulation were publicly available via OpenReview and arXiv and processed according to their terms of use; no private or non-public reviewer data were accessed. The simulation was conducted offline and did not influence real peer review decisions. Our findings underscore the need for transparency, audits for differential scoring, human oversight, and rigorous ethical guidelines before deploying LLM-based reviewers in real-world settings.

Reproducibility Statement

We support reproducibility by documenting all simulation components, including the research-agent and review-agent pipelines, publicly accessible APIs, and model versions. The Appendix provides prompts for paper generation, revision, and review, as well as examples of generated manuscripts. Human-authored papers are sampled from publicly available ICLR submissions on OpenReview, with corresponding LaTeX sources obtained from arXiv. Sections 3 and 4 describe the full simulation workflow, and the statistical analyses, feature extraction procedures, evaluation metrics, correlation computations, and hypothesis tests are specified for reproduction with standard NLP and statistical toolkits.

AI Use Statement

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

References

Marwa Abdulhai, Isadora White, Yanming Wan, Ibrahim Qureshi, Joel Z. Leibo, Max Kleiman-Weiner, and Natasha Jaques. How llms distort our written language, 2026. URL https://arxiv.org/abs/2603.18161.

Anirudh Ajith, Mengzhou Xia, Alexis Chevalier, Tanya Goyal, Danqi Chen, and Tianyu Gao. Litsearch: A retrieval benchmark for scientific literature search. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 15068–15083. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.840. URL https://doi.org/10.18653/v1/2024.emnlp-main.840.

Jinheon Baek, Sujay Kumar Jauhar, Silviu Cucerzan, and Sung Ju Hwang. Researchagent: Iterative research idea generation over scientific literature with large language models. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 6709–6738. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.342. URL https://doi.org/10.18653/v1/2025.naacl-long.342.

Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation, 2026. URL https://arxiv.org/abs/2605.03202.

Jianfeng Cai, Jinhua Zhu, Ruopei Sun, Yue Wang, Li Li, Wengang Zhou, and Houqiang Li. Disentangling length bias in preference learning via response-conditioned modeling, 2025. URL https://arxiv.org/abs/2502.00814.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. Mlebench: Evaluating machine learning agents on machine learning engineering. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=6s5uXNWGIh.

Juhwan Choi, JungMin Yun, Changhun Kim, and YoungBin Kim. Position paper: How should we responsibly adopt LLMs in the peer review process? In *Submitted to ACL Rolling Review - July 2025*, 2025. URL https://openreview.net/forum?id=t9ZQTsIUUw. under review.

Mike D'Arcy, Tom Hope, Larry Birnbaum, and Doug Downey. Marg: Multi-agent review generation for scientific papers, 2024. URL http://arxiv.org/abs/2401.04259v1.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, Haoran Zhang, Vipul Gupta, Yinghui Li, Tao Li, Fei Wang, Qin Liu, Tianlin Liu, Pengzhi Gao, Congying Xia, Chen Xing, Cheng Jiayang, Zhaowei Wang, Ying Su, Raj Sanjay Shah, Ruohao Guo, Jing Gu, Haoran Li, Kangda Wei, Zihao Wang, Lu Cheng, Surangika Ranathunga, Meng Fang, Jie Fu, Fei Liu, Ruihong Huang, Eduardo Blanco, Yixin Cao, Rui Zhang, Philip S. Yu, and Wenpeng Yin. Llms assist NLP researchers: Critique paper (meta-)reviewing. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 5081–5099. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.292. URL https://doi.org/10.18653/v1/2024.emnlp-main.292.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yi Dai, Jiawei Sun, Meng Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey, 2024a. URL https://arxiv.org/abs/2312.10997.

Zhaolin Gao, Kianté Brantley, and Thorsten Joachims. Reviewer2: Optimizing review generation through prompt generation, 2024b. URL http://arxiv.org/abs/2402.10886v2.

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

Siyuan Guo, Cheng Deng, Ying Wen, Hechang Chen, Yi Chang, and Jun Wang. Ds-agent: Automated data science by empowering large language models with case-based reasoning. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=LfJgeBNCFI.

Eftekhar Hossain, Sanjeev Kumar Sinha, Naman Bansal, R. Alexander Knipper, Souvika Sarkar, John Salvador, Yash Mahajan, Sri Guttikonda, Mousumi Akter, Md. Mahadi Hassan, Matthew Freestone, Matthew C. Williams Jr., Dongji Feng, and Santu Karmaker. Llms as meta-reviewers' assistants: A case study. In Luis Chiruzzo, Alan Ritter, and Lu Wang (eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2025 - Volume 1: Long Papers, Albuquerque, New Mexico, USA, April 29 - May 4, 2025*, pp. 7763–7803. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.NAACL-LONG.395. URL https://doi.org/10.18653/v1/2025.naacl-long.395.

Qian Huang, Jian Vora, Percy Liang, and Jure Leskovec. Mlagentbench: Evaluating language agents on machine learning experimentation. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=1Fs1LvjYQW.

Intology. Zochi technical report. *arXiv*, 2025.

Yiqiao Jin, Qinlin Zhao, Yiyang Wang, Hao Chen, Kaijie Zhu, Yijia Xiao, and Jindong Wang. Agentreview: Exploring peer review dynamics with LLM agents. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 1208–1226.

Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.70. URL https://doi.org/10.18653/v1/2024.emnlp-main.70.

Hao Kang and Chenyan Xiong. Researcharena: Benchmarking large language models' ability to collect and organize information as research agents, 2024. URL http://arxiv.org/abs/2406.10291v3.

Jaeho Kim, Yunseok Lee, and Seulki Lee. Position: The AI conference peer review crisis demands author feedback and reviewer rewards. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=l8QemUZaIA.

Rodney Michael Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, Miles Crawford, Doug Downey, Jason Dunkelberger, Oren Etzioni, Rob Evans, Sergey Feldman, Joseph Gorney, David W. Graham, F.Q. Hu, Regan Huff, Daniel King, Sebastian Kohlmeier, Bailey Kuehl, Michael Langan, Daniel Lin, Haokun Liu, Kyle Lo, Jaron Lochner, Kelsey MacMillan, Tyler C. Murray, Christopher Newell, Smita R Rao, Shaurya Rohatgi, Paul Sayre, Zejiang Shen, Amanpreet Singh, Luca Soldaini, Shivashankar Subramanian, A. Tanaka, Alex D Wade, Linda M. Wagner, Lucy Lu Wang, Christopher Wilhelm, Caroline Wu, Jiangjiang Yang, Angele Zamarron, Madeleine van Zuylen, and Daniel S. Weld. The semantic scholar open data platform. *ArXiv*, abs/2301.10140, 2023. URL https://api.semanticscholar.org/CorpusID:256194545.

Sandeep Kumar, Tirthankar Ghosal, Vinayak Goyal, and Asif Ekbal. Can large language models unlock novel scientific research ideas?, 2024. URL http://arxiv.org/abs/2409.06185v1.

Long Li, Weiwen Xu, Jiayan Guo, Ruochen Zhao, Xingxuan Li, Yuqian Yuan, Boqiang Zhang, Yuming Jiang, Yifei Xin, Ronghao Dang, Deli Zhao, Yu Rong, Tian Feng, and Lidong Bing. Chain of ideas: Revolutionizing research via novel idea development with llm agents, 2024a. URL http://arxiv.org/abs/2410.13185v5.

Miao Li, Jey Han Lau, and Eduard Hovy. A sentiment consolidation framework for meta-review generation. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 10158–10177, Bangkok, Thailand, August 2024b. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.547. URL https://aclanthology.org/2024.acl-long.547/.

Ruochen Li, Teerth Patel, Qingyun Wang, and Xinya Du. Mlr-copilot: Autonomous machine learning research based on large language models agents, 2024c. URL http://arxiv.org/abs/2408.14033v2.

Weixin Liang, Yuhui Zhang, Hancheng Cao, Binglu Wang, Daisy Ding, Xinyu Yang, Kailas Vodrahalli, Siyu He, Daniel Smith, Yian Yin, Daniel McFarland, and James Zou. Can large language models provide useful feedback on research papers? a large-scale empirical analysis, 2023. URL http://arxiv.org/abs/2310.01783v1.

Tzu-Ling Lin, Wei-Chih Chen, Teng-Fang Hsiao, Hou-I Liu, Ya-Hsin Yeh, Yu Kai Chan, Wen-Sheng Lien, Po-Yen Kuo, Philip S. Yu, and Hong-Han Shuai. Breaking the reviewer: Assessing the vulnerability of large language models in automated peer review under textual adversarial attacks, 2025. URL https://arxiv.org/abs/2506.11113.

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. The ai scientist: Towards fully automated open-ended scientific discovery, 2024. URL http://arxiv.org/abs/2408.06292v3.

Lukas Nolte and Sven Tomforde. A helping hand: A survey about ai-driven experimental design for accelerating scientific research. *Applied Sciences*, 15(9), 2025. ISSN 2076-3417. doi: 10.3390/app15095208. URL https://www.mdpi.com/2076-3417/15/9/5208.

Ori Press, Andreas Hochlehnert, Ameya Prabhu, Vishaal Udandarao, Ofir Press, and Matthias Bethge. Citeme: Can language models accurately cite scientific claims? In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on*

Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/0ef47f7b768e1a012e3d995ac8d8fac7-Abstract-Datasets_and_Benchmarks_Track.html.

Biqing Qi, Kaiyan Zhang, Haoxiang Li, Kai Tian, Sihang Zeng, Zhang-Ren Chen, and Bowen Zhou. Large language models are zero shot hypothesis proposers, 2023. URL http://arxiv.org/abs/2311.05965v1.

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. *Proceedings of the ACM on Human-Computer Interaction*, 9(7):1–28, 2025. doi: 10.1145/3757667.

Hyun Ryu, Doohyuk Jang, Hyemin S. Lee, Joonhyun Jeong, Gyeongman Kim, Donghyeon Cho, Gyouk Chu, Minyeong Hwang, Hyeongwon Jang, Changhun Kim, Haechan Kim, Jina Kim, Joowon Kim, Yoonjeon Kim, Kwanhyung Lee, Chanjae Park, Heecheol Yun, Gregor Betz, and Eunho Yang. Reviewscore: Misinformed peer review detection with large language models, 2025. URL https://arxiv.org/abs/2509.21679.

Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. Reviewertoo: Should ai join the program committee?, 2025. URL https://arxiv.org/abs/2510.08867.

Chenglei Si, Tatsunori Hashimoto, and Diyi Yang. The ideation-execution gap: Execution outcomes of llm-generated versus human research ideas, 2025a. URL https://arxiv.org/abs/2506.20803.

Chenglei Si, Diyi Yang, and Tatsunori Hashimoto. Can llms generate novel research ideas? A large-scale human study with 100+ NLP researchers. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025b. URL https://openreview.net/forum?id=M23dTGWCZy.

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

Xiangru Tang, Yuliang Liu, Zefan Cai, Yanjun Shao, Junjie Lu, Yichi Zhang, Zexuan Deng, Helan Hu, Kaikai An, Ruijun Huang, Shuzheng Si, Sheng Chen, Haozhe Zhao, Liang Chen, Yan Wang, Tianyu Liu, Zhiwei Jiang, Baobao Chang, Yin Fang, Yujia Qin, Wangchunshu Zhou, Yilun Zhao, Arman Cohan, and Mark Gerstein. Ml-bench: Evaluating large language models and agents for machine learning tasks on repository-level code, 2023. URL http://arxiv.org/abs/2311.09835v5.

Nitya Thakkar, Mert Yuksekgonul, Jake Silberg, Animesh Garg, Nanyun Peng, Fei Sha, Rose Yu, Carl Vondrick, and James Zou. Can llm feedback enhance review quality? a randomized study of 20k reviews at iclr 2025, 2025. URL http://arxiv.org/abs/2504.09737v1.

Minyang Tian, Luyu Gao, Shizhuo Dylan Zhang, Xinan Chen, Cunwei Fan, Xuefei Guo, Roland Haas, Pan Ji, Kittithat Krongchon, Yao Li, Shengyan Liu, Di Luo, Yutao Ma, Hao Tong, Kha Trinh, Chenyu Tian, Zihan Wang, Bohao Wu, Shengzhu Yin, Minhui Zhu, Kilian Lieret, Yanxin Lu, Genglin Liu, Yufeng Du, Tianhua Tao, Ofir Press, Jamie Callan, Eliu A. Huerta, and Hao Peng. Scicode: A research coding benchmark curated by scientists. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/36850592258c8c41cecdaa3dea5ff7de-Abstract-Datasets_and_Benchmarks_Track.html.

Keith Tyser, Ben Segev, Gaston Longhitano, Xin-Yu Zhang, Zachary Meeks, Jason Lee, Uday Garg, Nicholas Belsten, Avi Shporer, Madeleine Udell, Dov Te'eni, and Iddo Drori. Ai-driven review systems: Evaluating llms in scalable and bias-aware academic reviews, 2024. URL http://arxiv.org/abs/2408.10365v1.

Qingyun Wang, Doug Downey, Heng Ji, and Tom Hope. Scimon: Scientific inspiration machines optimized for novelty. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pp. 279–299. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.18. URL https://doi.org/10.18653/v1/2024.acl-long.18.

Jing Yang. Position: The artificial intelligence and machine learning community should adopt a more transparent and regulated peer review process. In *Forty-second International Conference on Machine Learning Position Paper Track*, 2025. URL https://openreview.net/forum?id=gnyqRarPzW.

Zonglin Yang, Xinya Du, Junxian Li, Jie Zheng, Soujanya Poria, and Erik Cambria. Large language models for automated open-domain scientific hypotheses discovery. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 13545–13565. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.804. URL https://doi.org/10.18653/v1/2024.findings-acl.804.

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

Sungduk Yu, Man Luo, Avinash Madusu, Vasudev Lal, and Phillip Howard. Is your paper being reviewed by an llm? benchmarking ai text detection in peer review, 2025. URL https://arxiv.org/abs/2502.19614.

Changjia Zhu, Junjie Xiong, Renkai Ma, Zhicong Lu, Yao Liu, and Lingyao Li. When your reviewer is an llm: Biases, divergence, and prompt injection risks in peer review. *arXiv preprint arXiv:2509.09912*, 2025a.

Minjun Zhu, Yixuan Weng, Linyi Yang, and Yue Zhang. Deepreview: Improving llm-based paper review with human-like deep thinking process. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pp. 29330–29355. Association for Computational Linguistics, 2025b. URL https://aclanthology.org/2025.acl-long.1420/.

APPENDIX

A LLM USAGE

Large Language Models (LLMs) were used to aid in the writing and polishing of the manuscript. Specifically, we used an LLM to assist in refining the language, improving readability, and ensuring clarity in various sections of the paper. The model helped with tasks such as sentence rephrasing, grammar checking, and enhancing the overall flow of the text. The authors take full responsibility for the content of the manuscript, including any text generated or polished by the LLM. We have ensured that the LLM-generated text adheres to ethical guidelines and does not contribute to plagiarism or scientific misconduct.

B PROMPT

B.1 PROMPT FOR AUTHOR AGENT

The prompts related to idea generation are consistent with Si et al. (2025b), while the remaining prompts employed for paper generation are as follows.

Abstract

Please generate a paper's abstract based on the provided information.
Tips:
- TL;DR of the paper
- What are we trying to do and why is it relevant?
- Why is this hard?
- How do we solve it (i.e., our contribution!)
- How do we verify that we solved it (e.g., Experiments and results)
- Please make sure the abstract reads smoothly and is well-motivated. This should be one continuous paragraph with no breaks between the lines.
- Just wrap it with
beginabstract and
endabstract, and the content should be in standard LaTeX language.
- Please ensure it's ICLR-style academic writing, and wrap LaTeX content in ```latex```
Paper Information Begin
{paper_information}
Paper Information End
related work abstract Begin
{other_abstracts}
related work abstract End

Introduction

Generate the paper introduction:
Tips:
- Longer version of the Abstract, i.e. of the entire paper
- What are we trying to do and why is it relevant?
- Why is this hard?
- How do we solve it (i.e. our contribution!)
- How do we verify that we solved it (e.g. Experiments and results)
- New trend: specifically list your contributions as bullet points
- Extra space? Future work!
- Natural paragraphs are a more standard way of expression than lists. Try not to use lists. You can consider using list when summarizing contributions.
- Using the correct citation format. et al. Should not appear directly in the text. Instead, the citation format \cite, \citet, \citep should be used correctly.
- Please ensure it's ICLR-style academic writing, and wrap LaTeX content in ```latex```
.
Abstract Begin
{abstract}
Abstract End
paper information Begin
{paper_information}
paper information End
part related work for citation Begin
{related_work}
part related work for citation End

Related Work

Generate the related work section:
Tips:
- Academic siblings of our work, i.e. alternative attempts in literature at trying to solve the same problem.
- Goal is to "Compare and contrast" - how does their approach differ in either assumptions or method? If their method is applicable to our Problem Setting I expect a comparison in the

experimental section. If not, there needs to be a clear statement why a given method is not applicable.
- The general format is 2 to 4 paragraphs. Each paragraph starts with a summary (the category of works) in bold typeface, followed by an introduction to the relevant works as many as much.
- Using the correct citation format. et al. Should not appear directly in the text. Instead, the citation format \cite,\citet,\citep should be used correctly.
- Please ensure it's latex-style academic writing, and wrap academic writing content in ```latex```.
Available References and Their Information BEGIN
{reference_datas}
Available References and Their Information end

Background

Please give me a paper background outline based on the following introduction, containing 34 subsections, and finally provide me with 46 keywords for retrieval augmentation (return to me in json format) ```json{{keywords:[]}}```
Introduction Begin
{introduction}
Introduction End
paper information Begin
{paper_information}
paper information End

Determine whether the content of the following webpage is relevant to {query}. If it is relevant, please extract the theoretical background from the webpage. If it is not relevant, simply return None without outputting any content.
webpage content Begin
{text}
webpage content End

Generate the background SECTION in ```latex``` according to the background outline and combine it with the retrieved information. with the following requirements:
Tips:
- Academic Ancestors of our work, i.e. all concepts and prior work that are required for understanding our method.
- Note: If our paper introduces a novel problem setting as part of its contributions, it is best to have a separate Section.
- Appropriately incorporate variables/formulas/formal expressions
- Please ensure it is ICLR-style academic writing, and wrap LaTeX content in ```latex```.
- Do not use any list formatting, use paragraph format instead.
- Please strictly follow the outline, expand and enrich the outline into the background section.
- All cite should be in bibtex format. Using the correct citation format. The phrase "et al." should not appear directly in the text. Instead, the citation format \cite, \citet, \citep should be used correctly.
background outline Begin
{json_content_real}
background outline End
retrieved information Begin
{ex_texts}
retrieved information End

Method

Generate the methology section in latex format, with the following requirements:
Tips:
- What we do. Why we do it. All described using the general Formalism introduced in the Problem Setting and building on top of the concepts / foundations introduced in Background.
- Appropriately incorporate variables/formulas/formal expressions
- Please ensure it's ICLR-style academic writing, and wrap LaTeX content in ```latex```.
- Do not use any list formatting, use paragraph format instead.
- Be sure to pay attention to latex and avoid any formatting errors
- Using the correct citation format. et al. Should not appear directly in the text. Instead, the citation format \cite, \citet,\citep should be used correctly.
Introduction Begin
{introduction}
Introduction End
Background Begin
{background}
Background End
paper_information Begin####
{paper_information}
paper_information End

Experiment Setting

Based on the completed sections of the paper and the experimental data and its introduction, generate an outline for the "Experiment Setting" section.
Completed sections of the paper BEGIN
{paper_already}
Completed sections of the paper END
Experimental results and their introduction BEGIN
{experiment_results_and_introduction}
Experimental results and their introduction END

- -

Based on the writing outline, the already completed parts of the paper, the experimental data, and their introduction, complete the "Experiment Setting" section in the style of ICLR academic writing.
- Paragraphs are a more formal way of expression than lists, so try to use paragraphs instead of lists. - Do not reference non-existent content, such as figures that do not exist at all.
- For datasets, baselines, and open-source models, try to add citations. For closed-source models, you can add links.
- Use the correct citation format. "et al." should not appear directly in the text. Instead, the citation formats \cite, \citet, \citep should be used correctly.
- Please ensure it's ICLR-style academic writing, and wrap LaTeX content in ```latex```.
Writing Outline Begin
{outline_for_setting}
Writing Outline End
Already Completed Paper Part Begin
{paper_already}
Already Completed Paper Part End
Experiment Results Begin
{experiment_results_and_introduction}
Experiment Results End

Result Analysis

Based on the completed sections of the paper, experimental setup, results data, and their introduction, generate an outline for the results chapter.
Already Completed Paper Section Begin #### {paper_already} #### Already Completed Paper Section End
Experimental Setup Begin #### {experiments_setting} #### Experimental Setup End
Experimental Results and Introduction Begin #### {experiment_results_and_introduction} #### Experimental Results and Introduction End

---

Based on the writing outline, the completed parts of the paper, and the experimental data along with their introduction, complete the results section in the style of ICLR academic writing.
Tips:
- Paragraphs are a more formal way of expression than lists; try to use paragraphs instead of lists.
- Do not reference non-existent content, such as figures that do not exist at all.
- All Tables should be in LaTeX format
- All Tables should be included in your generated results section.
- All tables must be with detailed analysis and interpretation of the data.
- Do not modify the content of the table and do not add any note in the table.
- Please ensure it's ICLR-style academic writing, and wrap LaTeX content in ```latex```.
Writing Outline Begin
{outline_for_results}
Writing Outline End
Completed Parts of the Paper Begin
{paper_already}
Completed Parts of the Paper End
Experimental Settings Begin
{experiments_setting}
Experimental Settings End
Experimental Results and Analysis Begin
{experiment_results_and_introduction}
Experimental Results and Analysis End

Conclusion

Based on the completed sections of the paper, write the conclusion section: 1. Ensure it follows the academic writing style of ICLR.
2. The conclusion should be one paragraph.
3. Content that requires LaTeX should be enclosed in ```latex``` and use LaTeX syntax.
Completed paper section begin
{paper_already}
Completed paper section end

Revise

1. Revise the specified section of the paper according to the reviewer's suggestions.
2. Only revise the content and keep the style consistent with the previous one.
3. Enclose LaTeX content within triple backticks using ```latex```, and ensure all formatting follows LaTeX syntax.
Section file to be revised begin
{paper_now}
Section file to be revised end

Reference for other sections begin
{paper_all}
Reference for other sections end
Reviewer's comments begin
{reviewer_comments_ex}
Reviewer's comments end

B.2 Prompt for Review Agent

All review-related prompts are consistent with Jin et al. (2024)

C Example

C.1 Example of LLM-authored Paper

One example of an LLM-authored paper is shown in Figure 5

C.2 Example of LLM Review

Reviewer Assessment I

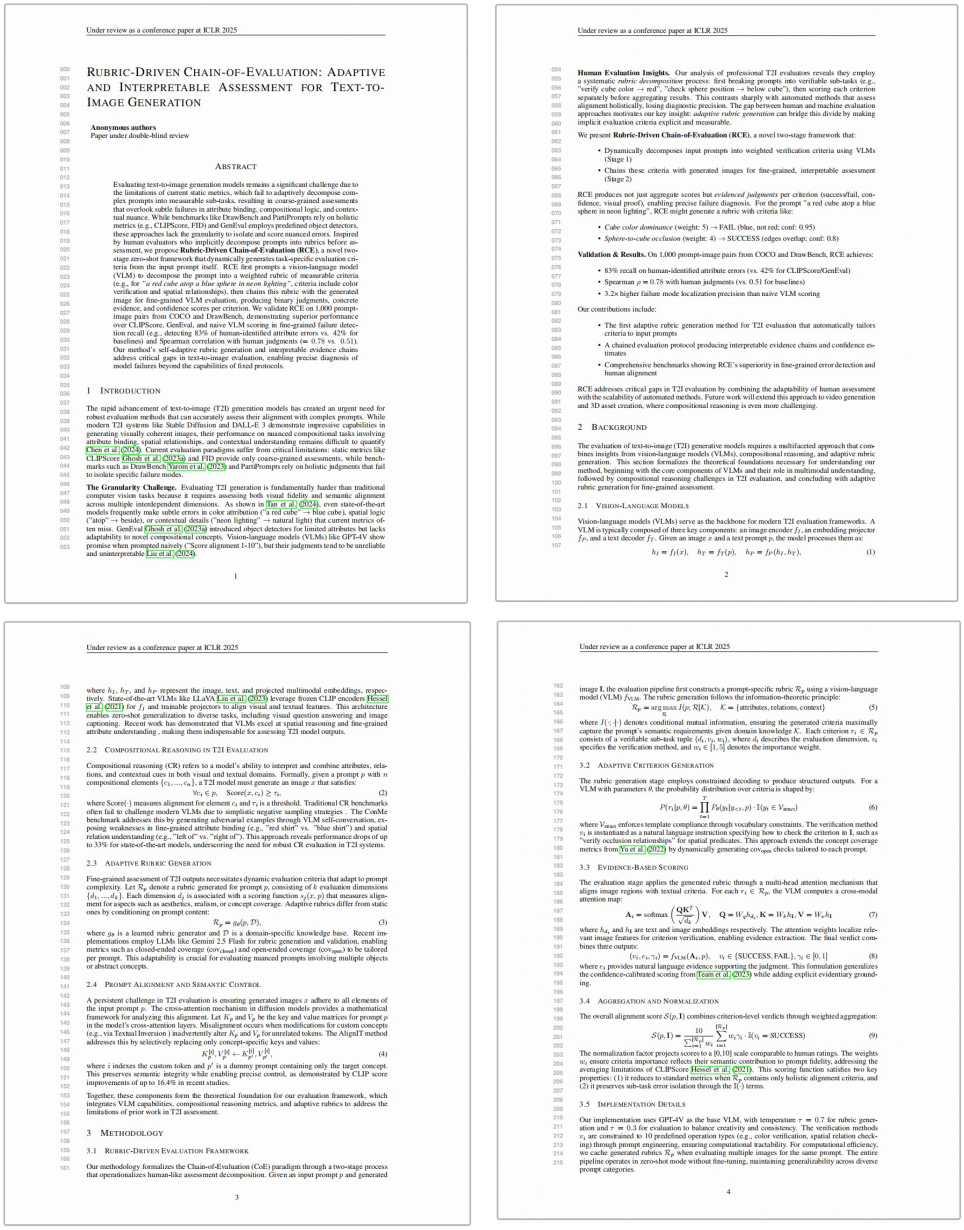
Reviewer 1
Overall rating: 6

Significance and novelty: The paper addresses the relatively unexplored topic of fairness in large language models (LLMs) when employed for tabular data prediction. This subject is significant given the extensive use of tabular data in high-stakes domains, and the potential social biases inherent in LLMs could lead to substantial risks in these applications.

Reasons for acceptance:
1. The paper tackles an underexplored area concerning the fairness of LLMs in tabular data prediction, which is crucial due to the wide application of tabular data in high-stakes domains.
2. It provides empirical evidence showing the persistence of social biases in LLMs, which are inherited from pre-training data, significantly impacting fairness in predictions.
3. The study explores different strategies to mitigate biases, such as in-context learning and fine-tuning, and compares their effectiveness against traditional models like Random Forests and Neural Networks.
4. The findings contribute valuable insights into the limitations of current bias mitigation techniques for LLMs, suggesting further research and development of methods tailored to address inherent biases effectively.

Reasons for rejection:
1. **Methodological Concerns**
- The experimental design lacks detailed explanations on the criteria used for selecting few-shot examples and the mechanism of label flipping during in-context learning, which are crucial for replicability and understanding of the results.
- The finetuning process is briefly mentioned without a comprehensive description of parameters and settings, which are essential for evaluating the effectiveness of the approach.

2. **Data and Results Interpretation**
- The paper does not sufficiently explore the implications of the observed biases in practical scenarios, missing an opportunity to connect experimental findings with real-world applications and consequences.
- There is a lack of discussion on the variability of results across different datasets, which



Under review as a conference paper at ICLR 2025

RUBRIC-DRIVEN CHAIN-OF-EVALUATION: ADAPTIVE AND INTERPRETABLE ASSESSMENT FOR TEXT-TO-IMAGE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Evaluating text-to-image generation models remains a significant challenge due to the limitations of current static metrics, which fail to adaptively decompose complex prompts into measurable sub-tasks, resulting in coarse-grained assessments that overlook subtle failures in attribute binding, compositional logic, and contextual nuance. While benchmarks like DrawBench and PartiPrompts rely on holistic metrics (e.g., CLIPScore, FID) and GenEval employs predefined object detectors, these approaches lack the granularity to isolate and score nuanced errors. Inspired by human evaluators who implicitly decompose prompts into rubrics before assessment, we propose **Rubric-Driven Chain-of-Evaluation (RCE)**, a novel two-stage zero-shot framework that dynamically generates task-specific evaluation criteria from the input prompt itself. RCE first prompts a vision-language model (VLM) to decompose the prompt into a weighted rubric of measurable criteria (e.g., for *"a red cube atop a blue sphere in neon lighting"*, criteria include color verification and spatial relationships), then chains this rubric with the generated image for fine-grained VLM evaluation, producing binary judgments, concrete evidence, and confidence scores per criterion. We validate RCE on 1,000 prompt-image pairs from COCO and DrawBench, demonstrating superior performance over CLIPScore, GenEval, and naive VLM scoring in fine-grained failure detection recall (e.g., detecting 83% of human-identified attribute errors vs. 42% for baselines) and Spearman correlation with human judgments (ρ = 0.78 vs. 0.51). Our method's self-adaptive rubric generation and interpretable evidence chains address critical gaps in text-to-image evaluation, enabling precise diagnosis of model failures beyond the capabilities of fixed protocols.

1 INTRODUCTION

The rapid advancement of text-to-image (T2I) generation models has created an urgent need for robust evaluation methods that can accurately assess their alignment with complex prompts. While modern T2I systems like Stable Diffusion and DALL-E 3 demonstrate impressive capabilities in generating visually coherent images, their performance on nuanced compositional tasks involving attribute binding, spatial relationships, and contextual understanding remains difficult to quantify Chen et al. (2024). Current evaluation paradigms suffer from critical limitations: static metrics like CLIPScore Ghosh et al. (2023a) and FID provide only coarse-grained assessments, while benchmarks such as DrawBench Yarom et al. (2023) and PartiPrompts rely on holistic judgments that fail to isolate specific failure modes.

The Granularity Challenge. Evaluating T2I generation is fundamentally harder than traditional computer vision tasks because it requires assessing both visual fidelity and semantic alignment across multiple interdependent dimensions. As shown in Tan et al. (2024), even state-of-the-art models frequently make subtle errors in color attribution ("a red cube" → blue cube), spatial logic ("atop" → beside), or contextual details ("neon lighting" → natural light) that current metrics often miss. GenEval Ghosh et al. (2023a) introduced object detectors for limited attributes but lacks adaptability to novel compositional concepts. Vision-language models (VLMs) like GPT-4V show promise when prompted naively ("Score alignment 1-10"), but their judgments tend to be unreliable and uninterpretable Liu et al. (2024).

1

Under review as a conference paper at ICLR 2025

Human Evaluation Insights. Our analysis of professional T2I evaluators reveals they employ a systematic *rubric decomposition* process: first breaking prompts into verifiable sub-tasks (e.g., "verify cube color → red", "check sphere position → below cube"), then scoring each criterion separately before aggregating results. This contrasts sharply with automated methods that assess alignment holistically, losing diagnostic precision. The gap between human and machine evaluation approaches motivates our key insight: *adaptive rubric generation* can bridge this divide by making implicit evaluation criteria explicit and measurable.

We present **Rubric-Driven Chain-of-Evaluation (RCE)**, a novel two-stage framework that:

- Dynamically decomposes input prompts into weighted verification criteria using VLMs (Stage 1)
- Chains these criteria with generated images for fine-grained, interpretable assessment (Stage 2)

RCE produces not just aggregate scores but *evidenced judgments* per criterion (success/fail, confidence, visual proof), enabling precise failure diagnosis. For the prompt "a red cube atop a blue sphere in neon lighting", RCE might generate a rubric with criteria like:

- *Cube color dominance* (weight: 5) → FAIL (blue, not red; conf: 0.95)
- *Sphere-to-cube occlusion* (weight: 4) → SUCCESS (edges overlap; conf: 0.8)

Validation & Results. On 1,000 prompt-image pairs from COCO and DrawBench, RCE achieves:

- 83% recall on human-identified attribute errors (vs. 42% for CLIPScore/GenEval)
- Spearman $\rho = 0.78$ with human judgments (vs. 0.51 for baselines)
- 3.2× higher failure mode localization precision than naive VLM scoring

Our contributions include:

- The first adaptive rubric generation method for T2I evaluation that automatically tailors criteria to input prompts
- A chained evaluation protocol producing interpretable evidence chains and confidence estimates
- Comprehensive benchmarks showing RCE's superiority in fine-grained error detection and human alignment

RCE addresses critical gaps in T2I evaluation by combining the adaptability of human assessment with the scalability of automated methods. Future work will extend this approach to video generation and 3D asset creation, where compositional reasoning is even more challenging.

2 BACKGROUND

The evaluation of text-to-image (T2I) generative models requires a multifaceted approach that combines insights from vision-language models (VLMs), compositional reasoning, and adaptive rubric generation. This section formalizes the theoretical foundations necessary for understanding our method, beginning with the core components of VLMs and their role in multimodal understanding, followed by compositional reasoning challenges in T2I evaluation, and concluding with adaptive rubric generation for fine-grained assessment.

2.1 VISION-LANGUAGE MODELS

Vision-language models (VLMs) serve as the backbone for modern T2I evaluation frameworks. A VLM is typically composed of three key components: an image encoder f_I, an embedding projector f_P, and a text decoder f_T. Given an image x and a text prompt p, the model processes them as:

$$h_I = f_I(x), \quad h_T = f_T(p), \quad h_P = f_P(h_I, h_T), \tag{1}$$

2

Under review as a conference paper at ICLR 2025

where h_I, h_T, and h_P represent the image, text, and projected multimodal embeddings, respectively. State-of-the-art VLMs like LLAVA Liu et al. (2023) leverage frozen CLIP encoders Hessel et al. (2021) for f_I and trainable projectors to align visual and textual features. This architecture enables zero-shot generalization to diverse tasks, including visual question answering and image captioning. Recent work has demonstrated that VLMs excel at spatial reasoning and fine-grained attribute understanding , making them indispensable for assessing T2I model outputs.

2.2 COMPOSITIONAL REASONING IN T2I EVALUATION

Compositional reasoning (CR) refers to a model's ability to interpret and combine attributes, relations, and contextual cues in both visual and textual domains. Formally, given a prompt p with n compositional elements $\{c_1, ..., c_n\}$, a T2I model must generate an image x that satisfies:

$$\forall c_i \in p, \quad \text{Score}(x, c_i) \geq \tau_i, \tag{2}$$

where $\text{Score}(\cdot)$ measures alignment for element c_i and τ_i is a threshold. Traditional CR benchmarks often fail to challenge modern VLMs due to simplistic negative sampling strategies . The CoMe benchmark addresses this by generating adversarial examples through VLM self-conversation, exposing weaknesses in fine-grained attribute binding (e.g., "red shirt" vs. "blue shirt") and spatial relation understanding (e.g., "left of" vs. "right of"). This approach reveals performance drops of up to 33% for state-of-the-art models, underscoring the need for robust CR evaluation in T2I systems.

2.3 ADAPTIVE RUBRIC GENERATION

Fine-grained assessment of T2I outputs necessitates dynamic evaluation criteria that adapt to prompt complexity. Let $\mathcal{R}_p$ denote a rubric generated for prompt p, consisting of k evaluation dimensions $\{d_1, ..., d_k\}$. Each dimension d_j is associated with a scoring function $s_j(x, p)$ that measures alignment for aspects such as aesthetics, realism, or concept coverage. Adaptive rubrics differ from static ones by conditioning on prompt content:

$$\mathcal{R}_p = g_\theta(p, \mathcal{D}), \tag{3}$$

where g_θ is a learned rubric generator and $\mathcal{D}$ is a domain-specific knowledge base. Recent implementations employ LLMs like Gemini 2.5 Flash for rubric generation and validation, enabling metrics such as closed-ended coverage ($\text{cov}_{\text{closed}}$) and open-ended coverage (cov_{open}) to be tailored per prompt. This adaptability is crucial for evaluating nuanced prompts involving multiple objects or abstract concepts.

2.4 PROMPT ALIGNMENT AND SEMANTIC CONTROL

A persistent challenge in T2I evaluation is ensuring generated images x adhere to all elements of the input prompt p. The cross-attention mechanism in diffusion models provides a mathematical framework for analyzing this alignment. Let K_k and V_k be the key and value matrices for prompt p in the model's cross-attention layers. Misalignment occurs when modifications for custom concepts (e.g., via Textual Inversion) inadvertently alter K_k and V_k for unrelated tokens. The AlignIT method addresses this by selectively replacing only concept-specific keys and values:

$$K_p^{[i]}, V_p^{[i]} \leftarrow K_{p'}^{[i]}, V_{p'}^{[i]}, \tag{4}$$

where i indexes the custom token and p' is a dummy prompt containing only the target concept. This preserves semantic integrity while enabling precise control, as demonstrated by CLIP score improvements of up to 16.4% in recent studies.

Together, these components form the theoretical foundation for our evaluation framework, which integrates VLM capabilities, compositional reasoning metrics, and adaptive rubrics to address the limitations of prior work in T2I assessment.

3 METHODOLOGY

3.1 RUBRIC-DRIVEN EVALUATION FRAMEWORK

Our methodology formalizes the Chain-of-Evaluation (CoE) paradigm through a two-stage process that operationalizes human-like assessment decomposition. Given an input prompt p and generated

3

Under review as a conference paper at ICLR 2025

image $\mathbf{I}$, the evaluation pipeline first constructs a prompt-specific rubric $\mathcal{R}_p$ using a vision-language model (VLM) f_{VLM}. The rubric generation follows the information-theoretic principle:

$$\mathcal{R}_p = \underset{\mathcal{R}}{\arg\max}\, I(p; \mathcal{R}|\mathcal{K}), \quad \mathcal{K} = \{\text{attributes, relations, context}\} \tag{5}$$

where $I(\cdot;\cdot)$ denotes conditional mutual information, ensuring the generated criteria maximally capture the prompt's semantic requirements given domain knowledge $\mathcal{K}$. Each criterion $r_i \in \mathcal{R}_p$ consists of a verifiable sub-task tuple (d_i, v_i, w_i), where d_i describes the evaluation dimension, v_i specifies the verification method, and $w_i \in [1, 5]$ denotes the importance weight.

3.2 ADAPTIVE CRITERION GENERATION

The rubric generation stage employs constrained decoding to produce structured outputs. For a VLM with parameters θ, the probability distribution over criteria is shaped by:

$$P(r_i|p, \theta) = \prod_{t=1}^{T} P_\theta(y_t|y_{<t}, p) \cdot \mathbb{I}(y_t \in \mathcal{V}_{\text{struct}}) \tag{6}$$

where $\mathcal{V}_{\text{struct}}$ enforces template compliance through vocabulary constraints. The verification method v_i is instantiated as a natural language instruction specifying how to check the criterion in $\mathbf{I}$, such as "verify occlusion relationships" for spatial predicates. This approach extends the concept coverage metrics from Yu et al. (2022) by dynamically generating cov_{open} checks tailored to each prompt.

3.3 EVIDENCE-BASED SCORING

The evaluation stage applies the generated rubric through a multi-head attention mechanism that aligns image regions with textual criteria. For each $r_i \in \mathcal{R}_p$, the VLM computes a cross-modal attention map:

$$\mathbf{A}_i = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V}, \quad \mathbf{Q} = W_q h_i, \mathbf{K} = W_k h_I, \mathbf{V} = W_v h_I \tag{7}$$

where h_i and h_I are text and image embeddings respectively. The attention weights localize relevant image features for criterion verification, enabling evidence extraction. The final verdict combines three outputs:

$$(v_i, e_i, \gamma_i) = f_{\text{VLM}}(\mathbf{A}_i, p), \quad v_i \in \{\text{SUCCESS, FAIL}\}, \gamma_i \in [0, 1] \tag{8}$$

where e_i provides natural language evidence supporting the judgment. This formulation generalizes the confidence-calibrated scoring from Yeats et al. (2023) while adding explicit evidentiary grounding.

3.4 AGGREGATION AND NORMALIZATION

The overall alignment score $\mathcal{S}(p, \mathbf{I})$ combines criterion-level verdicts through weighted aggregation:

$$\mathcal{S}(p, \mathbf{I}) = \frac{10}{\sum_{i=1}^{|\mathcal{R}_p|} w_i} \sum_{i=1}^{|\mathcal{R}_p|} w_i \gamma_i \cdot \mathbb{I}(v_i = \text{SUCCESS}) \tag{9}$$

The normalization factor projects scores to a $[0, 10]$ scale comparable to human ratings. The weights w_i ensure criteria importance reflects their semantic contribution to prompt fidelity, addressing the averaging limitations of CLIPScore Hessel et al. (2021). This scoring function satisfies two key properties: (1) it reduces to standard metrics when $\mathcal{R}_p$ contains only holistic alignment criteria, and (2) it preserves sub-task error isolation through the $\mathbb{I}(\cdot)$ terms.

3.5 IMPLEMENTATION DETAILS

Our implementation uses GPT-4V as the base VLM, with temperature $\tau = 0.7$ for rubric generation and $\tau = 0.3$ for evaluation to balance creativity and consistency. The verification methods v_i are constrained to 10 predefined operation types (e.g., color verification, spatial relation checking) through prompt engineering, ensuring computational tractability. For computational efficiency, we cache generated rubrics $\mathcal{R}_p$ when evaluating multiple images for the same prompt. The entire pipeline operates in zero-shot mode without fine-tuning, maintaining generalizability across diverse prompt categories.

4

Figure 5: LLM-authored Paper Example

could provide insights into dataset-specific challenges in bias mitigation.

3. **Limited Comparative Analysis**
- The comparison with traditional models is somewhat superficial, lacking a deep dive into why LLMs perform differently and the specific characteristics of LLMs that might contribute to biased outcomes.
- There is insufficient analysis of how the proposed bias mitigation techniques perform across various datasets, limiting the generalizability of the findings.

4. **Insufficient Exploration of Mitigation Strategies**
- The exploration of bias mitigation strategies such as data resampling is limited and lacks

depth, particularly in contrasting their effectiveness with findings from traditional machine learning approaches.
- The paper does not propose novel techniques or strategies for bias mitigation beyond those already established, missing an opportunity to advance the field.

Suggestions for improvement:
1. Enhance the explanation and methodology behind the selection and implementation of few-shot examples and label-flipping techniques in in-context learning to ensure clarity and reproducibility.
2. Provide a more detailed exploration of the practical implications of social biases observed in LLM predictions for tabular data, connecting experimental results with potential real-world applications and consequences.
3. Conduct a deeper and more comparative analysis of LLMs against traditional models, focusing on specific characteristics that contribute to biased outcomes and exploring dataset-specific challenges in bias mitigation.
4. Broaden the exploration of bias mitigation strategies, including the potential development of novel methods tailored to address inherent biases in LLMs, and provide a clear comparison of their effectiveness across various datasets.

- -

Reviewer 2
Overall rating: 4

Significance and novelty: The paper explores a timely and relevant issue, investigating the fairness of large language models (LLMs) when applied to tabular data, highlighting the transfer of social biases from training data to predictions. The novelty lies in focusing on fairness implications in high-stakes domains where such biases could have significant impacts. However, the contributions are primarily incremental, reiterating known biases in LLMs in a new context.

Reasons for acceptance:
1. The study addresses a critical issue of fairness in LLMs, particularly relevant for high-stakes applications using tabular data.
2. It provides a comprehensive evaluation across multiple datasets, enhancing the robustness of findings regarding biases in LLM predictions.
3. The investigation into zero-shot and few-shot learning scenarios adds depth to understanding how LLMs operate in different contexts.
4. The paper highlights the limitations of current mitigation strategies such as in-context learning and data resampling, pushing the field to consider more advanced solutions.

Reasons for rejection:
1. **Lack of Depth in Analysis**
- The paper does not deeply explore the root causes of biases beyond stating that they are inherited from training data, missing opportunities to hypothesize or test mechanisms.
- There is little to no discussion on the potential societal impacts of these biases, limiting the exploration of the real-world significance.

2. **Limited Novelty in Findings**
- The findings primarily confirm existing knowledge about biases in LLMs, adding little new insight into the nature or mitigation of these biases.
- The comparison with traditional ML models does not provide substantial new perspectives beyond what previous literature has discussed.

3. **Methodological Concerns**
- The choice of datasets and protected attributes, while common, does not push the boundaries of fairness research into less explored but equally important areas.
- Label-flipping experiments are interesting but are presented without thorough analysis of

why they have the observed effects, which could inform future mitigation strategies.

4. **Clarity and Structure**
- The paper's structure is somewhat repetitive, particularly in the sections discussing experiments, which may obscure the key contributions.
- Some methodological details, such as the choice of specific hyperparameters or experimental setups, are relegated to appendices, which could affect reproducibility and understanding.

Suggestions for improvement:
1. Enhance the depth of analysis by exploring the underlying mechanisms of bias transfer from training data to predictions in LLMs, potentially through qualitative analyses or theoretical explorations.
2. Expand on the societal implications of the findings, discussing potential impacts in high-stakes domains and suggesting actionable steps for stakeholders.
3. Introduce more diverse datasets and protected attributes to explore the generalizability of findings across different contexts and biases, which would strengthen the paper's contributions.
4. Improve clarity by restructuring the results section to avoid repetition and ensure that key findings and their implications are clearly and succinctly presented.,

- -

Reviewer 3:
Overall rating: 5

Significance and novelty: The paper addresses an important yet underexplored topic: the fairness of large language models (LLMs) in making predictions on tabular data. Despite the growing use of LLMs in various tasks, their application in tabular data and the associated fairness implications have not been extensively studied. This work aims to fill this gap by investigating the sources of information LLMs rely on and the extent of bias present in their predictions. The novelty lies in examining the effectiveness of in-context learning, label-flipping, and fine-tuning as strategies to mitigate biases in LLMs, compared to traditional machine learning models.

Reasons for acceptance:
1. The paper tackles a significant and timely issue regarding the use of LLMs in high-stakes domains, where fairness is crucial.
2. It provides a comparative analysis of LLMs and traditional models in tabular data tasks, contributing to a deeper understanding of biases in LLMs.
3. The investigation into various bias mitigation strategies, such as in-context learning and fine-tuning, offers valuable insights into their effectiveness in enhancing fairness.
4. The paper includes a comprehensive experimental setup and evaluation, testing on multiple datasets and employing several fairness metrics.

Reasons for rejection:
1. **Lack of sufficient novelty in methodological approach:**
- The use of in-context learning, fine-tuning, and label-flipping as bias mitigation strategies is not particularly novel, as these techniques have been explored in previous research.
- The paper provides limited innovation in terms of developing new methodologies or frameworks specifically designed for fairness improvement in LLMs for tabular prediction tasks.

2. **Insufficient depth in analysis of results:**
- While the paper presents various experimental results, there is a lack of in-depth discussion regarding the underlying reasons for the observed biases and the varying effectiveness of different mitigation strategies.
- The paper does not offer a detailed exploration of why certain datasets or tasks exhibit more pronounced biases or why certain strategies are more successful than others.

3. **Limited theoretical contribution:**
- The paper does not provide a strong theoretical foundation or model for understanding the inherent biases in LLMs, nor does it propose a new theoretical framework to guide future research in this area.
- The discussion primarily revolves around empirical findings without connecting them to broader theoretical insights in the field of AI fairness.

4. **Inadequate exploration of potential solutions:**
- While the paper briefly touches upon advanced strategies needed to address biases, it falls short of proposing or testing novel solutions or enhancements beyond existing techniques.
- The paper should have included more exploration of new approaches or adaptations specifically tailored to address bias in LLM-based tabular data predictions.

Suggestions for improvement:
1. Introduce novel methodologies or frameworks specifically targeted at bias mitigation in LLMs for tabular tasks, enhancing the paper2019s originality and contribution to the field.
2. Enhance the analysis and discussion of experimental results by providing deeper insights into the causes of biases and the differential effectiveness of mitigation strategies across various datasets and tasks.
3. Develop a more robust theoretical contribution by proposing new models or frameworks to understand and address inherent biases in LLMs, linking empirical findings to theoretical insights in AI fairness.
4. Explore and test innovative solutions or adaptations beyond current strategies to effectively reduce bias in LLMs, potentially incorporating interdisciplinary perspectives or new technologies.

Author–Reviewer Discussion

Author:
Response: Thank you for your insightful feedback. We appreciate the recognition of the significance and novelty of our work, as well as your constructive suggestions for improvement. Here are our responses to the concerns and suggestions you raised:

1. **Methodological Concerns**:
- We acknowledge the need for a more detailed explanation of the criteria for selecting few-shot examples and the label-flipping mechanism. We will include a comprehensive description in the revised manuscript to enhance replicability and clarity.
- We will expand the section on finetuning to provide a thorough account of the parameters and settings used, ensuring that our approach is fully transparent and evaluable.

2. **Data and Results Interpretation**:
- We agree that it's crucial to connect experimental findings with real-world applications. We will add a section discussing the practical implications of observed biases and potential consequences in high-stakes domains.
- To address the variability of results across different datasets, we will perform an additional analysis to explore dataset-specific challenges and include this discussion in the revised manuscript.

3. **Limited Comparative Analysis**:
- We realize the importance of deeper comparative analysis with traditional models. We will enhance the discussion on specific characteristics of LLMs that might contribute to biased outcomes, and delve into why they perform differently from traditional models.
- We will conduct additional experiments to better understand the performance of bias mitigation techniques across various datasets, thus enhancing the generalizability of our findings.

4. **Insufficient Exploration of Mitigation Strategies**:
- We will broaden our exploration of bias mitigation strategies, including a comparison with traditional machine learning findings, and aim to propose novel methods tailored specifically for LLMs.
- We appreciate the suggestion to advance the field by proposing new strategies and will look into innovative approaches to bias mitigation in future work.

Your feedback is invaluable, and we are committed to addressing the points raised to improve the quality and impact of our paper. Thank you once again for your detailed review and suggestions.

Author
Response: Thank you for your review and insightful comments. We address each of your concerns below:

1. **Lack of Depth in Analysis**: We agree that understanding the root causes of biases is crucial. While our paper emphasizes the inheritance of biases from training data, we acknowledge the need for a more in-depth exploration of the underlying mechanisms. Future work will focus on conducting qualitative analyses to hypothesize potential mechanisms of bias transfer and examining the societal structures that contribute to these biases. We will also reflect on the societal impact of these biases more thoroughly in our discussion section.

2. **Limited Novelty in Findings**: We appreciate your perspective on the novelty of our findings. Our intent was to bridge the gap between known biases in LLMs and their implications in tabular data contexts. Acknowledging that this reiteration might not offer entirely new insights, we will focus on developing and testing novel mitigation strategies that address these biases more effectively.

3. **Methodological Concerns**: We selected common datasets and protected attributes to provide a benchmark for fairness analysis. However, we understand your point about exploring less examined areas. We will expand our dataset selection to include a more diverse set with different protected attributes in future work. Regarding label-flipping, we recognize the need for deeper analysis, which we plan to address by investigating the semantic importance of flipped labels in future studies.

4. **Clarity and Structure**: We apologize if the structure led to any confusion. We will ensure the results section is more concise and focused, avoiding unnecessary repetition to enhance clarity. Additionally, we will incorporate key methodological details, such as hyperparameters, in the main text to improve transparency and reproducibility.

We appreciate your suggestions for improvement and will work to integrate them to strengthen the paper's contributions. Thank you again for your valuable feedback.

Author:
Response: Thank you for your detailed review and suggestions. We appreciate your positive remarks about the significance of our work and the comprehensive experimental setup. Below are our responses to the concerns raised:

1. **Lack of sufficient novelty in methodological approach:**
- While in-context learning, fine-tuning, and label-flipping have been studied before, our work uniquely applies these strategies to the domain of tabular data with LLMs, which has not been extensively explored. We will emphasize this aspect more clearly in the revised version to highlight our contribution in adapting existing methods to a novel application domain.

2. **Insufficient depth in analysis of results:**
- We acknowledge this concern and will include more detailed discussions on the underlying reasons for the observed biases and the differential effectiveness of bias mitigation strategies. We plan to conduct additional analysis to better understand why certain datasets exhibit more pronounced biases and include these insights in the revision.

3. **Limited theoretical contribution:**
- We understand the need for a stronger theoretical foundation. In response, we will work on incorporating more theoretical insights into the paper, potentially by linking our empirical findings with existing theories in AI fairness. We will aim to propose preliminary theoretical models to better explain the biases observed.

4. **Inadequate exploration of potential solutions:**
- We agree that exploring novel solutions is crucial. We plan to include proposed adaptations of existing techniques and potentially new strategies to mitigate bias more effectively. We will also consider interdisciplinary approaches that may offer fresh perspectives on this issue.

We appreciate your constructive feedback and are committed to addressing these points to enhance the quality and impact of our paper. Thank you again for your valuable input.

Reviewer Assessment II

Reviewer 1:
Overall rating: 6

Summary: The authors have provided thoughtful responses to most of the concerns raised, demonstrating a clear understanding of the feedback provided. They have acknowledged the need for enhanced methodological clarity, a deeper analysis of results, and a stronger theoretical contribution, and have outlined how they plan to address these aspects in future work or revisions of the paper. They also intend to include more diverse datasets and propose novel bias mitigation strategies, which would strengthen the study's contributions.

However, the proposed updates are largely prospective, and the current work still lacks significant novelty and depth in certain areas. The findings, while important, primarily confirm existing knowledge and do not introduce groundbreaking methodologies or theoretical insights. Additionally, while the fairness analysis is comprehensive, there is still a lack of exploration into practical applications and real-world implications of the observed biases. The paper also remains limited in its exploration of innovative mitigation strategies and dataset diversity.

Overall, the study provides a solid empirical foundation and addresses a critical and relevant problem, but there is room for improvement in terms of novelty, depth of analysis, and actionable contributions. The authors' plans to address these gaps in a revised manuscript are promising, hence the score is maintained at 6.

Reviewer 2:
Overall rating: 5

Summary: The authors have provided a detailed response to the raised concerns, demonstrating a clear understanding of the feedback. They acknowledge the limitations in methodological clarity, depth of analysis, and exploration of novel mitigation strategies. They have outlined plans to address these issues in future work, including diversifying the datasets, improving explanations for existing methods like label-flipping, and proposing novel bias mitigation strategies. They also intend to explore the societal implications of biases observed in LLMs for tabular data tasks more thoroughly.

However, while the authors' response suggests promising directions for future work, the rebuttal does not fully address the core limitations of the current paper. The study remains primarily incremental, mainly reiterating known biases in LLMs and their challenges in fairness metrics without introducing significant novelty. Methodological concerns, such as the lack of detailed explanations for key experimental setups (e.g., criteria for few-shot selection or label flipping), persist in the current work. Additionally, the comparative analysis with traditional models remains somewhat superficial, and the discussion on the generalizability of findings lacks depth. The absence of any innovative or advanced strategies for bias mitigation continues to limit the paper's potential contribution.

While the authors have shown a willingness to improve upon these aspects, the current submission does not sufficiently address these gaps, leading to a reduction in the overall rating. The work offers valuable observations on biases in LLMs for tabular tasks but falls short of providing actionable contributions or significant advancements in methodology.

Reviewer 3:
Overall rating: 5

Summary: The authors have provided detailed responses to the concerns raised, demonstrating an understanding of the feedback and a commitment to addressing these issues in future revisions. They acknowledge the need for enhanced methodological clarity and detail, particularly concerning the selection of few-shot examples and label-flipping techniques. The authors have also recognized the importance of exploring the societal implications of biases in LLMs and plan to connect experimental findings with real-world applications more thoroughly. Additionally, they aim to incorporate more diverse datasets and propose novel bias mitigation strategies tailored to LLMs for tabular tasks.

However, while the authors' response outlines promising directions for future work, the current paper remains limited in terms of novelty and depth of analysis. The study primarily reconfirms known biases in LLMs without introducing significant methodological or theoretical advancements. The experimental design lacks detailed explanations for key aspects, and the comparative analysis with traditional models is somewhat superficial. There is also a lack of exploration into innovative bias mitigation strategies and the generalizability of findings across different contexts. These limitations continue to impact the paper's potential contribution to the field.

Overall, the authors' planned improvements are promising, but the current submission does not sufficiently address these gaps, leading to a maintained score. The work offers important observations on biases in LLMs for tabular tasks, but it falls short of providing significant advancements or actionable contributions.

Meta-Review Compilation

AC:
Score: 5

Summary: The paper examines the fairness of large language models (LLMs) when used for predictions on tabular data, a significant issue due to the wide application of tabular data in high-stakes domains. It identifies that LLMs inherit social biases from their training data, impacting fairness in predictions. The authors explore several bias mitigation strategies, such as in-context learning and fine-tuning, comparing their effectiveness with traditional models like Random Forests and Neural Networks. While these efforts are commendable, the paper primarily confirms existing biases in LLMs without introducing novel methodologies or deeper theoretical insights into mitigating these biases. The authors acknowledge the need for more diverse datasets and novel mitigation strategies and plan

to incorporate these aspects in future work. Overall, the study presents valuable empirical observations but lacks substantial innovation and depth, particularly in exploring advanced bias mitigation strategies and real-world implications of the observed biases.,

Final Decision

Reviewer Assessment I stage Scores: 6, 4, 5
Reviewer Assessment II stage Scores: 6, 5, 5
Meta-Review Compilation stage Scores: 5
Average Score: (6 + 4 + 5 + 6 + 5 + 5 + 5)/7 = 5.14

D Detail

D.1 Negative Keywords

As shown in Table 6, we provide a list of negative keywords used for the limitation-oriented topic analysis in the study. The keyword set was constructed as a descriptive aid for analyzing limitation-oriented topics. We first collected candidate terms through vocabulary expansion around themes such as model limitations, risks, fairness, adversarial vulnerability, privacy and security concerns, failure modes, and mitigation. We then manually curated the list to remove overly broad or context-irrelevant terms. The keyword analysis is used as a broad quantitative check of topic patterns, rather than as the sole evidence for the paper's findings or as a causal test of review behavior.

Table 6: List of negative keywords used for the limitation-oriented topic analysis in the study.

Negative Keywords
gap, failure, unclear, minimal, mismatch, hazard, challenging, inequalities, adversarially, defect, injustice, unable, amplify the exposure, harmful, performance drop, mitigations, inherent bias, vulnerable, attacks, falls short, detrimental, weakness, incapable, risk, threat, limitation, cautionary, stereotypes, misalignment, flaw, cannot reach, bias, drop, weaknesses, drawback, minimal impact, relative drop, inadequate, sparsity, pitfall, sensitivity, inability, shortcoming, biases, deep-rooted biases, erroneous, inconsistencies, implicit stereotypical, reject, cautionary tale, prejudice

E Additional Statistics

E.1 Human Paper Sampling Statistics

The human-authored papers used in the main simulation were sampled across ten research topics, with ten papers per topic, for a total of 100 papers. This sampling scheme was intended to cover a range of common machine learning and LLM-related directions rather than concentrating on a single area. Table 7 summarizes the final decision distribution and the corresponding OpenReview human-review scores. Accepted papers account for 41% of the selected human-authored papers, indicating that the human baseline was not constructed by oversampling rejected papers.

Table 7: Decision distribution and OpenReview human-review average scores for the selected human-authored papers.

Decision	**Count**	**Human Avg.**
Accepted	41	6.4577
Rejected	59	4.6816
Total	100	–

E.2 Human Annotation Details

The human annotation study was designed as a targeted sanity check for cases where LLM reviewers most strongly favored LLM-authored papers, rather than as a full-scale expert-review replacement. We selected 15 matched human–LLM paper pairs with the largest LLM-favoring score gaps. Thirteen graduate-level or above CS annotators participated, including three co-authors and ten non-co-authors; seven annotators had prior reviewing experience. Each pair was independently evaluated by two annotators under ICLR-style review criteria, without revealing paper authorship.

To reduce obvious source or authorship cues, figures, appendices, references, and related-reference descriptions were removed for the annotation interface. This preprocessing was used only for the human annotation study, not for the simulation itself. It may affect the perceived completeness of some papers, especially human-authored ones, and should be considered when interpreting the annotation results. Inter-annotator agreement, measured by Cohen's κ, was approximately 0.52, indicating moderate agreement.

E.3 Linguistic Statistics

Linguistic statistics of human-authored papers and LLM-authored papers are shown in Table 8.

Table 8: Linguistic Statistics of Human Papers and LLM Papers (Excluding Appendices)

Group	Metric	Human	LLM
Text length	paper length	5548.10	4614.71
	sentence length	11.8464	9.3222
	paragraph length	51.1221	39.8206
Vocabulary diversity	1-gram	0.2598	0.4321
	2-gram	0.7069	0.8172
	3-gram	0.8844	0.9045
Difficulty	Flesch-Kincaid Grade Level	11.4963	13.2567
	Dependency Distance	4.2404	5.2659
	Subclause Ratio	0.0326	0.0233

Table 9: Correlation between textual statistical features and target variable

Metric	Feature	Correlation	P-value
Length	paper_length	-0.1912	0.0070
	average_sentence_length	-0.1405	0.0483
	average_sen_per_para	-0.1412	0.0472
Lexical diversity	1-gram	0.2795	0.0001
	2-gram	0.2386	0.0007
	3-gram	0.0745	0.2969
Complexity	readability_scores_fkg	0.1923	0.0067
	mean_dependency_distance	0.1956	0.0057
	subclause_ratio	-0.2691	0.0001

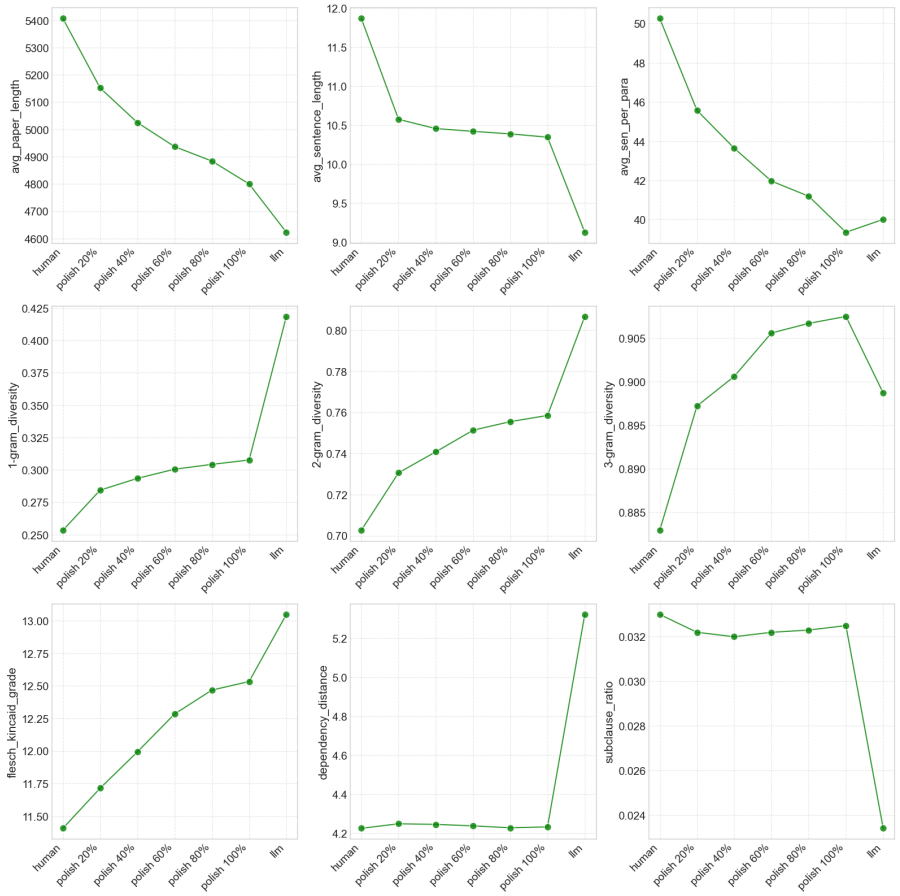


Figure 6: An analysis of the stylistic and linguistic features of LLM-generated text after varying levels of human polishing. We present a grid of nine plots, each corresponding to a distinct textual feature. Across all plots, the x-axis represents five document conditions: human-authored papers, polish 20%, polish 40%, polish 60%, polish 80%, and polish 100%, as well as pure LLM-authored papers (llm). The y-axis shows the average value of each feature for the corresponding document condition. The features analyzed include length (avg_paper_length, avg_sentence_length, avg_para_length), lexical diversity (1-gram_diversity, 2-gram_diversity, 3-gram_diversity), and complexity (flesch_kincaid_grade, dependency_distance, subclause_ratio).

F Discussion

F.1 Additional Related Work

LLMs as Reviewers Recent work has increasingly discussed potential reforms in the AI conference peer-review ecosystem. Li et al. (2024b) proposes a sentiment-consolidation framework to improve meta-review generation. Building on concerns about review reliability, Ryu et al. (2025) introduces a criterion for detecting misinformed review points, such as questions already answered in the paper or critiques based on incorrect premises. Complementary to these technical approaches, several position papers explore broader structural reforms. Choi et al. (2025) argues against using LLMs for end-to-end review generation, instead recommending targeted uses (e.g., reproducibility checks, citation validation, ethics flagging) that reduce reviewer burden while preserving human expertise. Similarly, Yang (2025) advocates for a more transparent, open, and regulated peer-review process, comparing fully open, partially open, and closed models. In parallel, Kim et al. (2025) calls for transforming the review pipeline from a one-way system into a bidirectional feedback loop, where authors evaluate review quality and reviewers receive formal accreditation through mechanisms such as two-stage review and systematic incentives, thereby supporting a sustainable and higher-quality peer-review framework.

F.2 Justification for Using Predicted Results

Constrained result setting To mitigate the possibility that LLM-generated papers benefit from overly optimistic synthetic results, we constrain their simulated numerical improvements to be no higher than those observed in the corresponding human-authored papers. Specifically, for each LLM-generated paper, reported improvements over baselines are capped by the improvement magnitudes observed in the matched human-authored paper under the same topic or keyword setting. This control reduces the chance that LLM-generated papers receive higher review scores simply because they report stronger experimental results, while preserving the scalability of the simulation.

Practical constraints and budget considerations (1) Implementing full experimental iterations would prevent the research process from being fully automated. (2) LLM coding capabilities are limited to less complex code, which would significantly restrict the scope of simulation topics. (3) Experimental iterations are time-consuming, making the paper generation process prohibitively long. Therefore, we believe it is acceptable to simplify the research agent's experimental execution, especially since this allows us to focus our effort and resources on exploring LLM reviewer risks rather than building a perfect research agent.

F.3 Idea Consistency in Simulation

Idea generation is an important component of auto research systems. Therefore, if we replaced LLM-generated ideas with ideas extracted from human papers, the resulting LLM papers would not fully reflect the model's capabilities across the full research pipeline. Additionally, this introduces an extra concern regarding whether the model can accurately extract the core ideas from human-authored papers. Inaccurate extraction could still lead to skewed results. Thus, we preserve the idea generation step. We also believe that evaluating full-paper generation (including ideation) rather than restricting to writing style better aligns with real-world reviewing practices. In this context, topic-level consistency (i.e., ensuring that both human- and LLM-authored papers target the same research area or keywords) offers a reasonable level of control for paper comparison.

In addition, within the same topic, LLM-authored papers show significantly higher score distributions than human-authored papers. Paired comparisons of human and LLM papers under the same keywords further support LLM-authored paper superiority from different views. The comparison is conducted at the broader level of human-author versus LLM-author given the same specific domain, while keeping idea consistency is relatively fine-grained.

Moreover, we examined a setting in which the ideas remained unchanged while retaining the original results. As shown in Section 6.3, simply applying LLM-based polishing to human-authored papers led to a noticeable increase in average review scores, from 5.69 to 5.94.

Allowing the LLM to generate ideas led to another noteworthy observation: only human-authored papers ended up in an irreducibly rejected state, even after multiple rounds of LLM revision. We found that LLMs tend to generate more positive ideas, while some human-authored ideas are critical in nature. As a result, all LLM-authored papers were eventually accepted, whereas a subset of human-authored papers consistently received low scores.

F.4 Rejected Human Papers

We did not oversample rejected human-authored papers in the main human–LLM comparison. As shown in Table 7, the selected human-authored set includes both accepted and rejected papers, with accepted papers accounting for 41% of the sample. This indicates that the observed patterns are not driven by concentrating on rejected human submissions.

Additionally, Section 5 reports a manual pairwise check for the largest LLM-favoring score gaps, which further rules out the interpretation that the effect is simply due to stronger LLM-authored papers.

F.5 Automated Scientific Discovery

Automated Scientific Discovery (ASD) Prior work has explored LLMs' idea generation capabilities, which can be considered a form of automated scientific discovery. Large-scale human annotations indicate that LLM-generated ideas tend to be more novel (Si et al., 2025b) but less actionable (Si et al., 2025a) than human-generated ideas. While these studies provide valuable context, a fine-grained evaluation of ideas, such as assessing their reliability, is beyond the scope of our work.

Fabricated Scientific Discoveries and ASD Detection The issue of fabricated scientific discoveries is not unique to LLMs; human-authored papers can also report fabricated results. Conversely, an LLM may generate speculative yet potentially valid scientific hypotheses. Existing hallucination detection methods could assist with ASD detection. However, unless a paper contains explicit contradictions or factual errors, determining whether a scientific discovery is fabricated, particularly at review time, is inherently challenging. Therefore, while ASD detection is important, it represents a long-standing problem that is conceptually distinct from the focus of our research.

Practical Scenario With the increasing use of LLMs as reviewers (and researchers), human- and LLM-authored papers may compete directly, as some authors may even submit fully LLM-generated papers out of curiosity or irresponsibility. Among human-authored papers, some may receive substantial LLM-based polishing, while others receive minimal assistance; some emphasize strong positive results, whereas others highlight risks, limitations, or constraints. When such papers are evaluated by LLM reviewers, the observed score disparities could disadvantage certain groups of human authors.

F.6 Human Review Noise

Human Noise In OpenReview, some reviewers are careless or outside their expertise. This noise makes perfect agreement unrealistic. However, in general, high-quality papers tend to be accepted and low-quality papers rejected, or the conference will lose its authority. Therefore, the human review data we collected can reasonably serve as a ground truth rather than "human unreliability". In addition, during our human-check stage, we invited responsible annotators to verify findings, confirming observed score disparities rather than suggesting that LLM-authored papers are superior.

F.7 Agents Construction

Research Agent We chose DeepSeek to serve as the backbone of our research pipeline, because its generated papers matching our expectations in terms of visual structure and content, and remains affordable for large-scale generation. Based on our preliminary investigation and testing, existing code models cannot run experiments and produce full research papers without human participation, and employing them introduces substantial time costs due to repeated trial-and-error. Consequently, we designed a dedicated pipeline, integrating LLM APIs and retrieval tools, to achieve automatic and scalable generation.

Review Agent The reviewer agent is a component of our simulation framework. Given that an effective review agent already exists, we believe introducing a new review agent and analyzing its risks offers limited value, and the risks we identify are rooted in the LLMs themselves. We report validation and additional reviewer-backbone checks in the main text.

F.8 Review Scope

Data and code are important components of the review process. However, based on both practical experience and official conference guidelines, submitted papers are not required to release data or code at review time, and most reviewers primarily evaluate the paper itself. Public statistics from ICLR 2024 indicate that only about 50% of accepted papers released code (Poster: 881/1807; Spotlight: 183/367; Oral: 44/86). Moreover, even when data and code are provided, reviewers typically focus on the paper rather than verifying the code or datasets. Therefore, our evaluation protocol, which emphasizes the paper content rather than code or data, aligns with actual reviewing practices.