PeerCheck: Enhancing LLM-Generated Academic Reviews Towards Human-Level Quality

Zeyuan Chen, Ziqing Yang, Yihan Ma, Michael Backes, Yang Zhang*

CISPA Helmholtz Center for Information Security

Abstract

As academic submissions grow, the traditional peer review process struggles to keep up, raising concerns about quality and fairness. A trend of using large language models (LLMs) for assistance has emerged. In this work, we take a critical step toward improving the quality of LLM-generated reviews. We propose the PeerCheck framework, which investigates LLM-human review differences (**RQ1**) and explores methods to improve LLM-generated review quality (**RQ2**). We first analyzed the human-written reviews with reviews generated by various LLMs and found that LLMs and humans focus on different terms, e.g., LLMs prioritize theory while humans emphasize methodology and experiments. We further adopt prompt engineering, such as Chain-of-Thought (CoT), and utilize retrieval-augmented generation (RAG) to enhance the LLM-generated reviews towards human-level quality. We find CoT significantly improves the quality of LLM reviews, while we discover an unexpected "RAG paradox," i.e., experiments with RAG produce different results for various LLMs and, in some cases, even reduce review quality. Our comprehensive analysis of LLM-generated academic reviews illustrates both possibilities and limitations, contributing to a more effective, human-aligned review system.[1]

1 Introduction

Peer review forms the cornerstone of academic publishing and serves as scientific literature's primary quality control mechanism. It relies on domain experts who critically evaluate manuscripts for methodological rigor, significance of findings, and clarity of presentation. Today, academic review faces serious challenges from exponential submission growth and limited reviewer availability, causing delays, inconsistent quality, and reviewer exhaustion (Tennant et al., 2017; Galheigo, 2011).

To address this issue, the integration of large language models (LLMs) into manuscript evaluation reflects a growing scholarly trend toward enhancing review processes (Zhou et al., 2025; Du et al., 2024). Advancements in LLMs, particularly models such as GPT (cha) and Claude (Anthropic), have shown remarkable capabilities in understanding, analyzing, and generating text across various domains, including scientific discourse (Zhou et al., 2023; Hämäläinen et al., 2023; Imani et al., 2023; Besta et al., 2024). They exhibit proficiency in comprehending complex concepts and identifying methodological issues (Kojima et al., 2022; Zhou et al., 2023). With LLM assistance, reviewers can understand manuscripts more efficiently and provide specialized evaluation for emerging interdisciplinary domains (Zhuang et al., 2025; Kocak et al., 2025; Ryu et al., 2025; Choi et al., 2026). AAAI has even rolled out a new policy that permits LLM assistance in peer review (AAAI, 2025). Specifically, the LLM takes a structured prompt and the manuscript PDF as input and generates the review according to the instructions. For example, it can include strengths, weaknesses, detailed comments, and a final decision. However, such dependence on LLMs would inevitably raise concerns regarding review quality, reliability, and alignment with the expected standards of scholarly evaluation (Shao and Chen, 2024; Yu et al., 2025; Latona et al., 2024).

Our Contribution. In this work, we take an important step towards utilizing LLMs for academic peer review. We formulate two research questions: (**RQ1**) What are the differences between peer reviews generated by LLMs and those written by humans? (**RQ2**) How can we improve LLM-generated review quality to human-level standards?

To address those research questions, we propose a novel framework named PeerCheck. Peer review represents the evaluation of academic work by ex-

*Corresponding author.

[1]Our dataset is available on https://github.com/TrustAIRLab/PeerCheck.

Findings of the Association for Computational Linguistics: ACL 2026, pages 23362–23386
July 2-7, 2026 ©2026 Association for Computational Linguistics

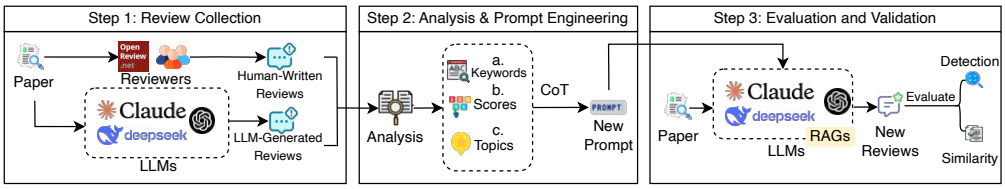


Figure 1: Workflow for PeerCheck.

perts in the same field. Check implies a process of assessment and refinement. The workflow is displayed on Figure 1. Firstly, we collect papers from top conferences as well as their official reviews. Meanwhile, for each paper, we use three different LLMs with a structured prompt to generate reviews, including GPT-4o, Claude-3.7-Sonnet, and DeepSeek-V3. Then we analyze and compare human reviews with LLM-generated reviews from the perspective of keywords, rating scores, and topics. In light of our observations, we conduct prompt engineering and use RAG to enhance the LLM-generated review quality. We further evaluate the effectiveness of our enhancement across multiple dimensions, including detectability and similarity. Specifically, detectability assesses how likely LLM-generated reviews can be identified as machine-generated, with lower detectability indicating more human-written reviews. Similarity serves as one quality indicator, along with human evaluation, as reviews matching the characteristics of effective human feedback are more likely to provide valuable insights to authors. Results demonstrate that our method can produce more authentic and high-quality academic reviews. Our contributions can be summarized as follows.

- We propose the PeerCheck framework, which improves LLM-generated review quality toward human-level standards, increasing GPT-4o's human-like score from 0.379 to 0.645.
- We discover the "RAG paradox," where retrieval augmentation improves GPT-4o but degrades Claude's performance, challenging the "more information equals better" assumption.
- We reveal that LLMs exhibit significant "role sensitivity," with Claude-3.7-Sonnet assigning scores of 8 at a rate 57.7% higher when in the "PhD student" role compared to no specified role.
- The RoBERTa-based detector achieves a 0.878 rate for LLM-generated reviews, but our revised Chain-of-Thought (CoT) prompts reduce this by 40.3%, significantly enhancing LLM-generated review authenticity.

2 Background and Related Works

LLMs in Academic Review. The peer review faces challenges due to increasing submission volumes, reviewer shortages, and publication delays (Galheigo, 2011; Zhuang et al., 2025; Kim et al., 2025). Recent LLM advances offer opportunities to assist or augment parts of the academic review process (Shin et al., 2025; Du et al., 2024; Yang, 2025; Ryu et al., 2025; Li et al., 2024). Saad et al. (2024) explored ChatGPT's potential for peer review assistance, while Wang et al. (2024) found substantial overlap between points raised by human reviewers and LLM-generated reviews. Despite promising outcomes, substantial constraints remain. Gao et al. (2025) found LLMs identify 68% of methodological issues caught by humans but miss subtle flaws requiring expertise. Enhancement approaches include Retrieval-Augmented Generation (RAG) (Chen et al., 2024), Chain-of-Thought prompting (CoT) (Wei et al., 2022), and hybrid human-AI workflows (Woelfle et al., 2024). However, comparisons across multiple advanced models using standardized review criteria remain limited (Shool et al., 2025; Zhou et al., 2024), and most approaches enhance LLM capabilities rather than targeting academic review weaknesses (Luo et al., 2025; Kocak et al., 2025). Our research addresses these limitations through PeerCheck, which enables quantitative comparison and targeted enhancement strategy development.

LLM-Generated Content Evaluation and Detection. As LLMs proliferate, identifying LLM-generated academic reviews has become crucial for maintaining research ecosystem integrity. Several frameworks assess LLM outputs, including GPTScore (Fu et al., 2025) for general text quality and Cai et al. (2024)'s multidimensional framework for scientific content. Their analysis of scientific abstracts showed LLMs often produce scientifically

plausible but factually questionable content with fabricated references (Gao et al., 2025). Detection grows challenging as LLMs advance, with researchers finding current detection tools "neither accurate nor reliable" (Weber-Wulff et al., 2023). More advanced detection methods include multi-feature models analyzing both lexical patterns and semantic structures (Peña et al., 2024) and watermarking approaches that embed undetectable signals in LLM outputs (Liang et al., 2024). However, integrating detection insights with improvement strategies remains underexplored, especially for specialized tasks like academic review. The characteristics making LLM reviews detectable, like generic criticism patterns and limited expertise signaling, highlight specific areas needing improvement (Yu et al., 2025; Luo et al., 2025). Our research identifies key LLM-human review differences and develops improvements via revised prompts in various LLM settings.

3 PeerCheck

We employ PeerCheck (shown in Figure 1) to evaluate and develop revised prompts for LLMs.

Review Collection. We first collect papers from top conferences and crawl their official peer reviews from OpenReview for comparison with LLM-generated reviews. Then, we implement a parallel processing architecture utilizing three state-of-the-art LLMs: GPT-4o, Claude-3.7-Sonnet, and DeepSeek-V3. All three LLMs process authentic research papers from a corpus consisting of 1,920 papers, primarily in the machine learning domains. Each LLM generates comprehensive reviews of each paper, extracting the strengths, weaknesses, and comments with questions from the paper, and also providing a final rating score for each paper. The evaluation prompt is shown in Appendix A.1. These LLM-generated reviews serve as benchmark outputs for measuring further prompt optimizations. Following prior work on persona prompting (Zhao et al., 2024; Shanahan et al., 2023; Liu and Ni, 2024), we simulate diverse reviewer personas, including professors, industry experts, and students, to strengthen our analysis framework. The specific prompts are shown in Appendix A.2. This implementation improves the comprehensiveness and quality of LLM-generated reviews across our measurement framework. The metrics across all three models offer redundancy and enhance the robustness of information extraction methodologies.

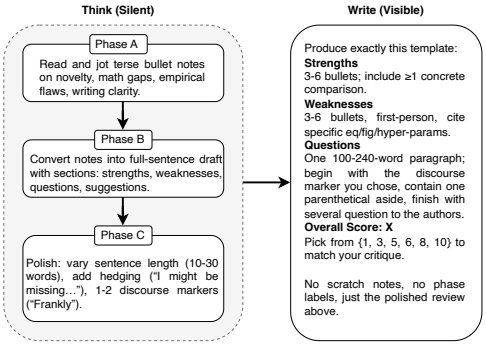


Figure 2: The Chain-of-Thought (CoT) peer review process. The thinking process contains three reflective phases that help understand and review the manuscripts.

Analysis & Prompt Engineering. In our evaluation, prompt engineering plays a crucial role, serving as the foundation for improving the quality and human-likeness of LLM-generated reviews. In the second step, we employ a systematic approach to develop and revise prompts to bridge the gap between LLM-generated and human-written reviews. (a) For the keyword analysis, we first extracted distinctive keywords from high-quality human reviews. The keywords are validated against IEEE VIS criteria (reVISe Committee, 2020), which are established guidelines for assessing visualization research quality. Then we compared keyword patterns between human and LLM reviews to identify language differences. (b) Meanwhile, we compared the rating scores given by LLM-generated and human-written reviewers. This scoring analysis shows how human reviewers and LLMs measure the quality of articles when making acceptance decisions and assigning scores. (c) For topic analysis, we extract articles receiving high scores ($\geq$ 7) from both human reviewers and LLMs. We compare topical domains of highly rated papers to analyze topic preference differences between LLM-generated and human-written reviewers, revealing potential disciplinary biases in automated evaluation systems. After comparing and analyzing, we produce revised prompts directing LLMs to use specific terms and reasoning styles, adopting reviewer personas, justify scoring criteria, and consider disciplinary context in paper reviewing. These revised prompts include keywords and examples drawn from human reviews and provide a concrete model for expert peer review in LLMs. Following the methodological intervention, we employed a CoT (Turpin et al., 2023; Wei et al., 2022) approach to refine and iterate our prompt formu-

lations, as shown in Figure 2. This iterative improvement process enables LLMs to mimic human analytical reasoning for generating high-quality reviews. The revised prompts (Appendix A.3) guide LLMs through a structured evaluation process incorporating linguistic features and reasoning patterns characteristic of human-written reviewers.

Evaluation and Validation. The final step implements a validation module that assesses whether LLM-generated reviews achieve human-level quality. We measured LLM-generated detectability and similarity to human-written reviews. In this context, we utilize RAG to enhance LLM review quality by retrieving external knowledge from human reviews, addressing the identified gaps between LLM-generated and human-written evaluations (Chen et al., 2024; Lewis et al., 2020). A validation set of 480 research papers, distinct from those used in previous phases, is processed through the enhanced system architecture. The validation process incorporated two evaluation mechanisms to ensure robust assessment of the revised prompts. We employ a metric-based detection methodology and a model-based detection that incorporates multiple statistical measures to establish the detection of LLM-generated reviews (He et al., 2024). To see whether LLM-generated reviews by revised prompts can be judged as human-written by detection methods. For similarity detection, we measure lexical and semantic overlap between LLM-generated reviews and a corpus of the same size as human-written reviews. Based on these results, we determine whether using revised prompts makes LLM-generated reviews more similar to human-written reviews. This comprehensive validation approach assessed prompt optimization effectiveness across multiple performance dimensions.

4 Experimental Setup

Datasets. The dataset covers all papers from ICLR 2024/2025 (ICLR, 2025) and NeurIPS 2024 (NeurIPS). We ensure the human-written by filtering them with RoBERTa (Liu et al., 2019), retaining only those reviews that have more than 80% probability of being written by humans (Figure 3). After dividing in an 8:2 ratio, we ultimately formed two datasets of 1,920 papers (analysis) and 480 papers (validation) for comprehensive evaluation across multiple language models. To evaluate the datasets, we use rating score comparison, semantic similarity comparison, keywords analysis, and topic analysis. Details are in Appendix B.

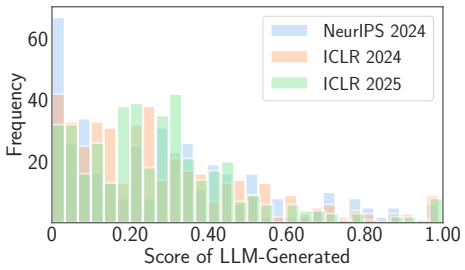


Figure 3: We randomly select 400 reviews from each of the three conferences to detect LLM-generated text. The histogram shows the frequency distribution of detection scores across NeurIPS 2024, ICLR 2024, and ICLR 2025, with lower RoBERTa-based scores suggesting most reviews are human-authored.

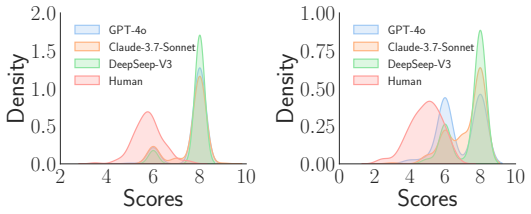
Evaluated LLMs. Our study compares the review generation capabilities of three advanced LLMs. GPT-4o (OpenAI, 2024), a versatile model known for its strong reasoning abilities in various domains. Claude 3.7 Sonnet (Anthropic, 2025), which features a specialized "thinking mode" for enhanced analytical reasoning. DeepSeek V3 (DeepSeek), which integrates advanced multimodal capabilities focusing on visual and scientific reasoning. Each model represents LLM designs and training methods, revealing different academic review generation skills and improvement strategies.

LLM Detection and Similarity Metrics. We use a multidimensional metric technique to evaluate LLM-generated reviews' quality and recognizability by analyzing their detection efficacy and likeness to human-written reviews. For detection, we use log-likelihood (He et al., 2024), log-rank (Su et al., 2023), GLTR (Gehrmann et al., 2019), LRR (Su et al., 2023; He et al., 2024), and RoBERTa (Liu et al., 2019). For similarity, we use BLEU scores (Papineni et al., 2002), ROUGE scores (Lin, 2004), token-level F1 scores (DeYoung et al., 2020), and cosine scores (Larsen et al., 2016). Details on metrics are provided in Appendix C.

5 Human vs. LLM Reviews

In this section, we compare human-written reviews with LLM-generated reviews across multiple quality dimensions, i.e., to answer **RQ1**.

Overall Performance Comparison. Our comparative analysis of rating scores assigned by LLMs versus human reviewers reveals notable distributional differences, as illustrated in Figure 4. Compare the other two conference rating scores in Figure 7



(a) NeurIPS Accept (b) NeurIPS Reject

Figure 4: LLMs vs. human rating scores in NeurIPS.

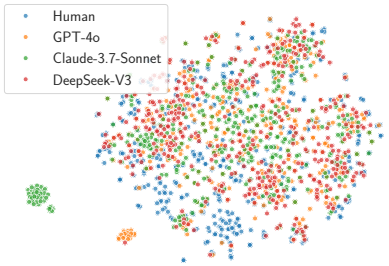


Figure 5: T-SNE visualization showing clustering patterns reflecting linguistic and content differences.

and Figure 8 of Appendix D. The human reviewer scoring pattern demonstrates a more concentrated distribution across the evaluation spectrum. In contrast, the three LLM reviewers exhibit considerably more dispersed scoring patterns, with a pronounced clustering in the 6 to 8 score range. A striking finding appears: LLMs' judges show broader score distributions than humans, consistently assigning higher scores to rejected papers. DeepSeek-V3 particularly assigned elevated ratings to both accepted and rejected papers. Nevertheless, diverse LLMs demonstrate conspicuous disparities in their scoring patterns when evaluating rejected manuscripts.

For better analysis, we randomly extract a corpus of 400 review comments from each source, i.e., human reviewers and the three LLMs under investigation. The feature space distribution is visualized through t-SNE (van der Maaten and Hinton, 2008), as illustrated in Figure 5. The visualization demonstrates extensive mixing of data points across sources, revealing substantial feature space overlap between human- and LLM-generated reviews, thus presenting significant discrimination challenges in most cases. However, distinct isolated clusters appear in the visualization, particularly from Claude-3.7-Sonnet and GPT-4o reviews, indicating peripheral feature-space regions where some LLM-generated content exhibits distinctive characteristics. This suggests certain machine-generated texts still maintain recognizable linguistic signatures despite general convergence.

Topical Analysis: Humans Prefer Infrastructure While LLMs Favor Emerging Technologies. We also analyzed topical distribution. It reveals both convergences and divergences between LLMs and human reviewers. All of them prioritize *large language models* and *reinforcement learning*. Human reviewers distinctively emphasize *natural language processing (NLP)* research in papers, a focus absent from LLMs' priorities. Conversely, LLMs, including the three LLMs examined, preferentially attend to other topics, such as *representation learning* and *in-context learning*, directions not explicitly prioritized by human reviewers. The comparison for the top 5 topics is shown in Figure 9 in Appendix D. Additionally, our analysis of authorship distribution reveals that human reviewers historically demonstrated a modest preference for manuscripts with larger authorship teams compared to LLMs. However, this discrepancy has progressively attenuated across recent conferences, suggesting convergence of LLM evaluation parameters toward human standards, as shown in Figure 10 in Appendix D.

Table 1: Top 15 topical keywords for LLM and human reviewers. Red refers to words that all three LLMs and humans pay attention to.

NO.	Human	GPT-4o	Claude 3.7 Sonne	Deepseek V3
1	methods	theoretical	theoretical	theoretical
2	results	datasets	methods	empirical
3	performance	experimental	significant	results
4	experiments	reproducibility	contributions	experiments
5	theoretical	empirical	computational	methods
6	datasets	limitations	performance	novel
7	algorithm	scalability	particularly	practical
8	comparison	methods	novel	limitations
9	evaluation	performance	evaluation	contributions
10	baseline	experiments	comprehensive	impacts
11	assumption	real-world	datasets	validation
12	limitations	contributions	limitations	datasets
13	framework	implications	framework	comparison
14	empirical	rigor	applications	reproducibility
15	contributions	insights	insights	scalability

Keywords Analysis: Humans Prioritize Experiments While LLMs Focus on Theory. We analyze the top 15 common terms in human-written and LLM-generated reviews (validated against IEEE VIS (reVISe Committee, 2020)), presenting the various reviewing priorities of human reviewers and three LLMs. In Table 1, analysis reveals pronounced divergence in emphasis: all LLM reviews prioritize "theoretical" as their primary criterion, whereas human reviewers demonstrate predominant concern with "methods" considerations. While all reviewer categories share a focus on "limitations," they exhibit differentiation across

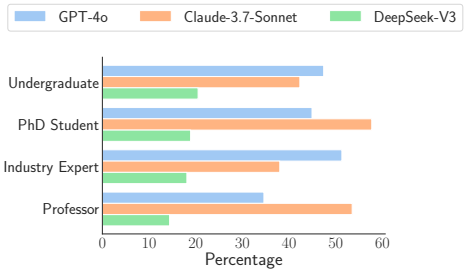


Figure 6: Role-based scoring bias: increasing percentage of 8-score ratings across different LLM personas based on 8-score ratings in Normal.

secondary priorities. Human reviewers emphasize "performance" metrics and "experiments" validation, GPT-4o prioritizes "reproducibility," Claude-3.7-Sonnet demonstrates attention to "significant" assessment, and DeepSeek-V3 orients toward "empirical" and "novel" aspects. Notably, only human reviewers include "algorithm" and "baseline" within top 15 keywords. Such divergence reflects fundamental differences in evaluation paradigms between artificial and human intelligence.

Influence of Role-Play on LLM-Generated Reviews. We conduct role-play across three LLMs, including industry experts, undergraduates, professors, and PhD students, and compare the rating scores produced by different personas. Our experiments on role-play across three LLMs revealed distinct rating score patterns. Empirical evidence shows each model exhibits unique preferences when assuming different personas, as shown in Figure 6. Claude-3.7-Sonnet demonstrates a strong tendency to assign maximum 8-point ratings when embodying academic personas like "PhD student" (57.7%) and "professor" (53.5%). GPT-4o shows greater generosity in the "industry expert" role (51.3%), while DeepSeek-V3 maintained conservative rating behavior (20.5% is the maximum) across all personas. The "undergraduate" persona elicits moderate scoring tendencies in both Claude-3.7-Sonnet and GPT-4o. We also analyze keyword preferences during role-play manuscript reviews, revealing how personas influence each LLM's linguistic patterns and evaluative approaches. Table 6 in Appendix D presents the top 5 keywords used by GPT-4o, Claude-3.7-Sonnet, and DeepSeek-V3 across various evaluative personas. GPT-4o emphasizes "datasets" generally while prioritizing "scalability" and "real-world" as an industry expert. Claude-3.7-Sonnet favors "theoretical" and "analysis," shifting toward "applications" in the industry expert role. DeepSeek-V3 consistently prioritizes "theoretical" concepts while adapting secondary focuses by role, emphasizing "novel" contributions when assuming the PhD student persona. These variations demonstrate how LLMs adjust their evaluation criteria based on personas and architectures, reflecting unique academic preferences and training methodologies and highlighting the importance of equitable LLM review systems.

Takeaways for RQ1. *Despite the topic overlapping with human reviews, LLMs overemphasize "theory" while undervaluing "methods" and "experiments." This misalignment indicates a clear improvement pathway: recalibrating LLM evaluation priorities to better align with human review criteria both structurally and substantively.*

6 Enhancing LLM-Generated Reviews

Based on the results in Section 5, we improve LLM-generated reviews by revising the prompts on both LLMs and RAG-enhanced LLMs. The revised prompts are generated from Section 3, integrating language patterns from human-written reviews into CoT-based templates. For RAG, we used human reviews from the 1920 non-detection papers as retrieval sources. All artifacts are open-sourced with no proprietary data. This section shows how our enhancement improves LLM-generated review quality, i.e., to answer **RQ2**.

LLM-Generated Detection. We test revised prompts in LLMs and employ five detection methods, i.e., RoBERTa, log-likelihood, log-rank, GLTR, and LRR, to evaluate whether AI-generated reviews can evade detection. The results are displayed in Table 2. In the regular LLM evaluation, different detection methods demonstrate significant performance differences. The DeepSeek-V3 achieves the highest detection rate of 0.878 when using the RoBERTa method, making it the most recognizable overall. Claude-3.7-Sonnet, conversely, demonstrates strong concealment capabilities against most detection methods. While RoBERTa proves to be the most effective universal detector, the LRR method is particularly effective for GPT-4o, which is 0.801. Notably, using CoT techniques, such as revised, revised with keywords, and revised with keywords and sample sentence combinations in the prompt, significantly reduces detection accuracy. In other words, these approaches can effectively enhance the covertness of LLM-generated content.

Table 2: Performance (F1 scores) comparison of different detection methods for identifying LLM-generated academic reviews by conventional LLMs and LLMs with RAG techniques. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

LLM	Method	LLM				LLM with RAGs			
		N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.
GPT-4o	Log-Likelihood	0.762	0.353	0.295	0.297	0.612	0.359	0.296	0.352
	Log-Rank	0.692	0.349	0.367	0.324	0.572	0.442	0.322	0.312
	GLTR	0.779	0.307	0.261	0.262	0.607	0.298	0.361	0.375
	LRR	0.801	0.257	0.176	0.149	0.593	0.213	0.252	0.280
	RoBERTa	0.798	0.512	0.554	0.536	0.752	0.422	0.425	0.459
Claude-3.7 -Sonnet	Log-Likelihood	0.688	0.194	0.223	0.206	0.656	0.160	0.205	0.198
	Log-Rank	0.755	0.308	0.345	0.296	0.706	0.204	0.214	0.207
	GLTR	0.678	0.216	0.389	0.278	0.637	0.228	0.300	0.239
	LRR	0.620	0.390	0.262	0.329	0.684	0.350	0.251	0.357
	RoBERTa	0.785	0.466	0.479	0.514	0.757	0.475	0.482	0.500
DeepSeek- V3	Log-Likelihood	0.717	0.285	0.206	0.231	0.693	0.230	0.270	0.310
	Log-Rank	0.700	0.275	0.314	0.241	0.713	0.267	0.236	0.326
	GLTR	0.706	0.201	0.245	0.227	0.635	0.301	0.324	0.310
	LRR	0.723	0.311	0.321	0.315	0.679	0.392	0.290	0.428
	RoBERTa	0.878	0.514	0.475	0.516	0.817	0.479	0.485	0.503

Table 3: Similarity comparison between reviews generated by LLMs, RAG-assisted LLMs, and human reviews under different similarity metrics. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

LLM	Metric	LLM				LLM with RAGs			
		N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.
GPT-4o	BLUE	0.017	0.020	0.018	0.021	0.030	0.034	0.022	0.027
	ROUGE	0.205	0.233	0.247	0.240	0.300	0.372	0.318	0.315
	Token-level F1	0.291	0.317	0.333	0.316	0.305	0.363	0.312	0.318
	Cosine	0.379	0.501	0.502	0.518	0.524	0.615	0.631	0.645
Claude-3.7 -Sonnet	BLUE	0.013	0.036	0.041	0.014	0.013	0.029	0.018	0.019
	ROUGE	0.163	0.328	0.297	0.334	0.166	0.401	0.338	0.331
	Token-level F1	0.256	0.348	0.296	0.309	0.257	0.345	0.327	0.254
	Cosine	0.431	0.688	0.651	0.641	0.470	0.705	0.655	0.630
DeepSeek- V3	BLUE	0.010	0.018	0.029	0.016	0.013	0.028	0.019	0.015
	ROUGE	0.126	0.329	0.357	0.331	0.148	0.346	0.297	0.310
	Token-level F1	0.233	0.303	0.281	0.268	0.146	0.288	0.267	0.263
	Cosine	0.407	0.562	0.596	0.620	0.379	0.634	0.608	0.641

When LLM is combined with RAG, the detection difficulty generally increases, while the detection rates for reviews produced by all models using the "Normal" prompt decrease. GPT-4o is most affected by RAG, with its LRR detection rate plummeting from 0.801 to 0.593 with the normal prompt. Claude-3.7-Sonnet and DeepSeek-V3 show smaller decreases in their detection rates, demonstrating varying adaptability to the RAG technique across different models. Interestingly, some combinations of modification strategies, revised prompts with keywords, and sample sentences are more likely to be detected than conventional LLMs under RAG. For example, the log-likelihood detection rate for GPT-4o increases from 0.296 to 0.352. These results not only prove that RAG technology indeed enhances LLM-generated review quality, reflected in human similarity, but also show that there is no universal "perfect hiding." The detection effectiveness depends on the specific combination of models, modification techniques, and detection methods employed. These results reveal that, though technology can make LLM-generated text resemble human writing, LLMs' academic judgment capabilities reflect fundamental differences in reasoning.

Similarity With Human Reviews. We compare the similarity of LLM-generated reviews, generated by revised prompts, with human-written reviews.

Table 4: Structure-vs.-style prompt ablation on direct review quality dimensions.

Prompt Types		Specificity	Actionability	Faithfulness
LLM	Normal	15.004±5.189	4.002±1.955	0.561±0.191
	Revised	41.803±15.362	7.404±1.092	0.523±0.291
	Style-only	7.199±4.145	6.004±1.962	0.773±0.140
	Structure-only	47.403±16.075	7.604±1.882	0.503±0.061
LLM with RAG	Normal	35.002±5.760	8.003±1.962	0.591±0.113
	Revised	19.249±2.041	6.902±1.224	0.611±0.080
	Style-only	10.600±2.853	6.201±1.844	0.602±0.084
	Structure-only	24.400±7.051	6.803±1.340	0.519±0.062

We treat similarity to high-quality human reviews as a quality indicator, as matching their structure and focus leads to more constructive feedback. Beyond similarity metrics, we add human evaluation to directly assess whether generated reviews are helpful and insightful. Table 3 shows the overall review similarity analysis. LLMs display a clear ranking in cosine metric, with Claude-3.7-Sonnet leading (0.431), followed by DeepSeek-V3 (0.407) and GPT-4o (0.379). All LLMs improved with text modification techniques, with Claude-3.7-Sonnet achieving 0.688 after using revised prompts. The RAG technique significantly increases the cosine metric in GPT-4o, from 0.379 to 0.524, slightly improves Claude-3.7-Sonnet, from 0.431 to 0.470, but marginally decreases DeepSeek-V3, from 0.407 to 0.379. The combination of RAG and modification techniques proved most effective, with GPT-4o reaching 0.645 under revised prompt with keywords and sample sentences conditions. This reveals complex RAG architecture interactions, not just additive effects, providing key insights for model selection and optimization in practice. Analysis of review sections (*strengths*, *weaknesses*, and *questions*) reveals patterns shown in Appendix E.

Prompt Ablation: Structure vs. Style. To better understand whether the gains from revised prompting come from structured evaluation guidance or merely stylistic mimicry, we conduct a prompt ablation that separates structural scaffolding from stylistic instructions. We evaluate review quality using three direct dimensions: specificity, which measures paper-grounded references (Sizo et al., 2025); actionability, which captures whether suggestions are implementable (Superchi et al., 2019); and faithfulness, which assesses whether claims are supported by evidence (Thorne and Vlachos, 2019; Tamber et al., 2025). In this ablation, structure-only retains the evaluation scaffolds while removing stylistic phrasing, whereas style-only keeps the stylistic instructions but removes the structural guidance. The corresponding prompt templates are provided in Section A.4.

As shown in Table 4, revised and structure-only prompts consistently improve specificity and actionability over normal and style-only prompts. Notably, structure only achieves the strongest overall performance on these direct quality dimensions, indicating that the observed gains mainly stem from structured evaluation guidance rather than surface-level stylistic imitation. By contrast, RAG-based variants sometimes improve actionability, but they yield less stable specificity and tend to reduce faithfulness. These results suggest that better review quality comes primarily from stronger evaluative structure, not merely from making the reviews sound more human-like.

Evidence-Grounded Correctness Analysis. To assess factual faithfulness beyond stylistic similarity, we conduct a claim-level evidence-grounded correctness analysis (for more details, see Appendix F). The results show that revised prompting consistently reduces unsupported and not-verifiable claims across both LLM and RAG settings, with further gains from keyword and sample augmentations. This indicates that the improvements reflect better factual correctness rather than merely stylistic similarity.

RAG Paradox. We conduct an ablation study by adjusting the value of k in RAG to examine how parameter changes affect model performance. The results confirm the RAG paradox: GPT-4o improves with more retrieved documents, Claude-3.7-Sonnet degrades consistently, and DeepSeek-V3 remains unstable. It also provides insight into the conditions under which RAG helps or hinders review quality. More detail can be found in Appendix G.

Human Evaluation. To validate our similarity

Table 5: Human evaluation results for human-written vs. LLM-generated reviews. In the table, human-written (Human), normal LLM-generated (LLM.N), revised LLM-generated (LLM.R), and RAG-enhanced LLM-generated (LLM.RAG) reviews.

Dimension	Human	LLM.N	LLM.R	LLM.RAG
Average score	7.124	5.371	6.682	5.900
High quality proportion (≥ 7)	66.951%	31.672%	60.332%	42.504%
Low quality proportion (≤ 3)	11.863%	18.333%	12.402%	12.511%

findings, we conduct two human evaluations. We first use two linguistic metrics, lexical diversity (MTLD) (McCarthy and Jarvis, 2010) and syntactic complexity (Lu, 2010), to validate the similarity findings. From Table 12 in Appendix H, revised prompts with keywords and samples consistently improve text quality across all models, and these improvements stem from more than just stylistic imitation (see Appendix H). Further, we conduct a human evaluation on 160 reviews across four generation methods, which include human-written, normal LLM-generated, revised LLM-generated, and RAG-enhanced LLM-generated reviews, using 40 randomly selected papers. Table 5 shows human reviews perform best, while enhanced prompting significantly improves LLM performance compared to standard prompts. The annotation task demonstrated high consistency among three independent raters, with agreement (κ = 0.707, 85.00% agreement (McHugh, 2012)) across 160 samples. RAG integration surprisingly decreases overall performance, confirming our counterintuitive finding that RAG helps GPT-4o but harms Claude-3.7-Sonnet and DeepSeek-V3. Using the same 40-paper sample, we additionally conduct a multidimensional pairwise comparison across the four review conditions. Annotators with prior peer-review experience selected the most constructive (Rooyen et al., 1999), most specific (Sadallah et al., 2025), and most faithful (Bevendorff et al., 2023) review among human-written, normal, revised, and revised-with-RAG conditions. As the results in Table 13 from Section I.2 show that revised is most often chosen as most specific and more constructive than normal, while human-written reviews remain strongest in faithfulness. More detailed results and a case study are provided in Appendix I.

Takeaways for RQ2:. *LLM evaluation reveals model strengths: Claude-3.7-Sonnet performs most human-like, GPT-4o excels at describing strengths, and DeepSeek benefits most from revised prompts. RAG-enhanced with LLMs reveals unpredictable effects: boosting GPT-4o's similarity while reducing Claude-3.7-Sonnet's strength-description capability. This challenges the assumption that more information equals better performance, suggesting breakthroughs require balancing information quantity with model reasoning compatibility.*

7 Conclusion

We introduce the PeerCheck framework to improve LLM-generated academic reviews' reliability and quality. Our approach begins with comparing human-written and LLM-generated analyses to explore their differences. Based on these insights, we apply CoT prompting to revise the LLM review generation process. We further evaluate the effectiveness of the revised prompts in LLM and RAG-enhanced LLM. Under the PeerCheck framework, revised prompts can reduce LLM-generated detection rates by 40.3%, enhancing the credibility of LLM-assisted reviews. They also improve LLM-generated review quality, with similarity to human-written reviews increased by 26.6%. Claude-3.7-Sonnet creates the most human-like reviews, GPT-4o best describes strengths, and DeepSeek-V3 is the LLM-detectable but improves most with revised prompts. These results highlight that improving LLM-based academic review requires well-planned prompts and model-specific strategies, not just more input length or data volume.

Limitation

Our research has several limitations. First, while using three state-of-the-art LLMs, our test sample focuses on machine learning papers post-2023 and may not generalize to other research fields or earlier periods. Second, our RAG implementation depends on existing human review data, potentially limiting its effectiveness in emerging research areas. Third, our role sensitivity test examines only five academic personas, not covering all possible reviewer identities. Finally, although our revision prompting technique significantly reduces LLM-generated review detection rates, it raises ethical questions about ensuring transparency and recognizability of LLM-generated reviews in academia. We plan to address these limitations in future work by expanding to other domains and temporal scopes.

Ethical Consideration

All APIs were accessed under their respective Terms of Service, and the released dataset is intended solely for research and educational purposes. Our study uses only publicly available peer reviews from OpenReview; no new data were collected from human subjects. This study was reviewed and approved by the Ethics Review Board (No. 25-12-6) of our institution. The ERB determined that the use of publicly available OpenReview data constitutes secondary analysis and does not require additional human-subjects review.

All review texts originate from OpenReview, where contributors agree to CC-BY-SA 2.0 public release upon submission. We hash identifiers and remove any remaining personal details to reduce potential risks. While the data are publicly available, we acknowledge that residual risks such as re-identification or amplified public criticism may still exist; our analysis therefore focuses on aggregate patterns rather than individual reviewers or papers. `PeerCheck` will be released under CC-BY-4.0. All external models were used in accordance with their respective licenses. Before release, we automatically removed author names and affiliations, hashed paper IDs, and filtered profanity using the open-source Detoxify model. Manual spot checks confirmed no remaining personal or offensive content.

Peer review is a central component of the scholarly ecosystem, and systems that interact with it require careful consideration. Prior work has shown that large language models may reflect linguistic and cultural biases, potentially affecting authors whose language backgrounds or writing styles differ from dominant norms. Accordingly, `PeerCheck` is intended as an assistive analytical tool rather than an automated reviewer or decision-making system. In this work, "human-level quality" is operationalized as similarity to historical human-written reviews along lexical and semantic dimensions. We note that such similarity does not capture all aspects of high-quality peer review, such as novel intellectual insight or constructive mentorship, and should not be interpreted as a normative definition of review quality.

Acknowledgements

We thank all anonymous reviewers for their valuable comments and suggestions, which helped improve the quality of this paper.

References

ChatGPT. https://chat.openai.com/chat.

OpenReview.net. https://openreview.net/.

AAAI. 2025. AAAI Launches AI-Powered Peer Review Assessment System. https://aaai.org/aaai-launches-ai-powered-peer-review-assessment-system/.

Anthropic. Claude. https://claude.ai/.

Anthropic. 2025. Claude 3.7 Sonnet and Claude Code. https://www.anthropic.com/news/claude-3-7-sonnet.

Maciej Besta, Nils Blach, Ales Kubicek, Robert Gerstenberger, Michal Podstawski, Lukas Gianinazzi, Joanna Gajda, Tomasz Lehmann, Hubert Niewiadomski, Piotr Nyczyk, and Torsten Hoefler. 2024. Graph of Thoughts: Solving Elaborate Problems with Large Language Models. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 17682–17690. AAAI.

Janek Bevendorff, Ian Borrego-Obrador, Mara Chinea-Ríos, Marc Franco-Salvador, Maik Fröbe, Annina Heini, Krzysztof Kredens, Maximilian Mayerl, Piotr Pęzik, Martin Potthast, Francisco Rangel, Paolo Rosso, Efstathios Stamatatos, Benno Stein, Matti Wiegmann, Magdalena Wolska, and Eva Zangerle. 2023. Overview of Pan 2023: Authorship Verification, Multi-author Writing Style Analysis, Profiling Cryptocurrency Influencers, and Trigger Detection: Condensed Lab Overview. In *International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF)*, pages 459–481. Springer.

Hengxing Cai, Xiaochen Cai, Junhan Chang, Sihang Li, Lin Yao, Changxin Wang, Zhifeng Gao, Hongshuai Wang, Yongge Li, Mujie Lin, Shuwen Yang, Jiankun Wang, Mingjun Xu, Jin Huang, Xi Fang, Jiaxi Zhuang, Yuqi Yin, Yaqi Li, Changhong Chen, and 4 others. 2024. SciAssess: Benchmarking LLM Proficiency in Scientific Literature Analysis. *CoRR abs/2403.01976*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking Large Language Models in Retrieval-Augmented Generation. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 17754–17762. AAAI.

Juhwan Choi, JungMin Yun, Changhun Kim, and YoungBin Kim. 2026. Position Paper: How Should We Responsibly Adopt LLMs in the Peer Review Process? In *Findings of the Association for Computational Linguistics: EACL (EACL Findings)*, pages 151–165. ACL.

DeepSeek. DeepSeek AI. https://deepseek.ai/.

Jay DeYoung, Sarthak Jain, Nazneen Fatema Rajani, Eric Lehman, Caiming Xiong, Richard Socher, and Byron C. Wallace. 2020. ERASER: A Benchmark to

Evaluate Rationalized NLP Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 4443–4458. ACL.

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, and 21 others. 2024. LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5081–5099. ACL.

Rosa Falotico and Piero Quatto. 2015. Fleiss' kappa statistic without paradoxes. *Quality & Quantity*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2025. GPTScore: Evaluate as You Desire. In *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 6556–6576. ACL.

Sandra Maria Galheigo. 2011. What Needs to Be Done? Occupational Therapy Responsibilities and Challenges Regarding Human Rights. *Australian Occupational Therapy Journal*.

Catherine A. Gao, Frederick M. Howard, Nishant D. Shah, and et al. 2025. Comparing Scientific Abstracts Generated by ChatGPT to Real Abstracts with Detectors and Blinded Human Reviewers. *NPJ Digital Medicine*.

Daniel Garcia-Costa, Flaminio Squazzoni, Bahar Mehmani, and Francisco Grimaldo. 2022. Measuring the Developmental Function of Peer Review: A Multi-dimensional, Cross-disciplinary Analysis of Peer Review Reports from 740 Academic Journals. *PeerJ Computer Science*.

Sebastian Gehrmann, Hendrik Strobelt, and Alexander M. Rush. 2019. GLTR: Statistical Detection and Visualization of Generated Text. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 111–116. ACL.

Mingmeng Geng and Roberto Trotta. 2025. Human-LLM Coevolution: Evidence from Academic Writing. In *Findings of the Association for Computational Linguistics: ACL (ACL Findings)*, pages 12689–12696. ACL.

Alexander Goldberg, Ivan Stelmakh, Kyunghyun Cho, Alice Oh, Alekh Agarwal, Danielle Belgrave, and Nihar B Shah. 2025. Peer Reviews of Peer Reviews: A Randomized Controlled Trial and Other Experiments. *PLOS One*.

Perttu Hämäläinen, Mikke Tavast, and Anton Kunnari. 2023. Evaluating Large Language Models in Generating Synthetic HCI Research Data: a Case Study. In *Annual ACM Conference on Human Factors in Computing Systems (CHI)*. ACM.

Xinlei He, Xinyue Shen, Zeyuan Chen, Michael Backes, and Yang Zhang. 2024. MGTBench: Benchmarking Machine-Generated Text Detection. In *ACM SIGSAC Conference on Computer and Communications Security (CCS)*. ACM.

ICLR. 2024. ICLR 2024 Reviewer Guide. https://iclr.cc/Conferences/2024/ReviewerGuide.

ICLR. 2025. ICLR. https://iclr.cc/.

Shima Imani, Liang Du, and Harsh Shrivastava. 2023. MathPrompter: Mathematical Reasoning using Large Language Models. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 37–42. ACL.

Jaeho Kim, Yunseok Lee, and Seulki Lee. 2025. Position: The AI Conference Peer Review Crisis Demands Author Feedback and Reviewer Rewards. In *International Conference on Machine Learning (ICML)*. PMLR.

Burak Kocak, Mehmet Ruhi Onur, Seong Ho Park, Pascal Baltzer, and Matthias Dietzel. 2025. Ensuring Peer Review Integrity in the Era of Large Language Models: A Critical Stocktaking of Challenges, Red Flags, and Recommendations. *European Journal of Radiology Artificial Intelligence*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large Language Models are Zero-Shot Reasoners. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Wojciech Kryscinski, Bryan McCann, Caiming Xiong, and Richard Socher. 2020. Evaluating the Factual Consistency of Abstractive Text Summarization. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9332–9346. ACL.

Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. 2016. Autoencoding beyond Pixels Using a Learned Similarity Metric. In *International Conference on Machine Learning (ICML)*, pages 1558–1566. JMLR.

Giuseppe Russo Latona, Manoel Horta Ribeiro, Tim R. Davidson, Veniamin Veselovsky, and Robert West. 2024. The AI Review Lottery: Widespread AI-Assisted Peer Reviews Boost Paper Scores and Acceptance Rates. *CoRR abs/2405.02150*.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Miao Li, Jey Han Lau, and Eduard H. Hovy. 2024. A Sentiment Consolidation Framework for Meta-Review Generation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 10158–10177. ACL.

Yuqing Liang, Jiancheng Xiao, Wensheng Gan, and Philip S. Yu. 2024. Watermarking Techniques for Large Language Models: A Survey. *CoRR abs/2409.00089*.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 74–81. ACL.

Xun Liu and Zhengwei Ni. 2024. Role-playing Prompt Framework: Generation and Evaluation. *CoRR abs/2406.00627*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR abs/1907.11692*.

Xiaofei Lu. 2010. Automatic Analysis of Syntactic Complexity in Second Language Writing. *International Journal of Corpus Linguistics*.

Ziming Luo, Zonglin Yang, Zexin Xu, Wei Yang, and Xinya Du. 2025. LLM4SR: A Survey on Large Language Models for Scientific Research. *CoRR abs/2501.04306*.

Philip M McCarthy and Scott Jarvis. 2010. MTLD, vocd-D, and HD-D: A Vlidation Study of Sophisticated Approaches to Lexical Diversity Assessment. *Behavior Research Methods*.

Mary L McHugh. 2012. Interrater Reliability: the Kappa Statistic. *Biochemia Medica*.

NeurIPS. NeurIPS. https://neurips.cc/.

NeurIPS. 2024. Reviewer Guidelines. https://neurips.cc/Conferences/2024/ReviewerGuidelines.

OpenAI. 2024. GPT-4o. https://openai.com/index/hello-gpt-4o/.

OpenReview. Using the API. https://docs.openreview.net/getting-started/using-the-api.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 311–318. ACL.

Alberto Rodero Peña, Jacinto Mata Vázquez, and Victoria Pachón Álvarez. 2024. I2C-Huelva at SemEval-2024 Task 8: Boosting AI-Generated Text Detection with Multimodal Models and Optimized Ensembles. In *International Workshop on Semantic Evaluation (SemEval)*, pages 845–852. ACL.

The reVISe Committee. 2020. Keywords and Their Role in The Reviewing Process (for Authors). https://ieeevis.org/year/2024/blog/keywords-for-authors.

Susan Van Rooyen, Nick Black, and Fiona Godlee. 1999. Development of the Review Quality Instrument (RQI) for Assessing Peer Reviews of Manuscripts. *Journal of Clinical Epidemiology*.

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. 2025. ReviewScore: Misinformed Peer Review Detection with Large Language Models. *CoRR abs/2509.21679*.

Ahmed Saad, Nathan Jenko, Sisith Ariyaratne, Nick Birch, Karthikeyan P Iyengar, Arthur Mark Davies, Raju Vaishya, and Rajesh Botchu. 2024. Exploring The Potential of ChatGPT in The Peer Review Process:BDM An Observational Study. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*.

Abdelrahman Sadallah, Tim Baumgärtner, and Ted Briscoe Iryna Gurevych. 2025. The Good, the Bad and the Constructive: Automatically Measuring Peer Review's Utility for Authors. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 28991–29021. ACL.

Murray Shanahan, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*.

Dong J. Shao and S. Chen. 2024. Are We There Yet? Revealing the Risks of Utilizing Large Language Models in Scholarly Peer Review. *CoRR abs/2412.01708*.

Hyungyu Shin, Jingyu Tang, Yoonjoo Lee, Nayoung Kim, Hyunseung Lim, Ji Yong Cho, Hwajung Hong, Moontae Lee, and Juho Kim. 2025. Automatically Evaluating the Paper Reviewing Capability of Large Language Models. *CoRR abs/2502.17086*.

Sina Shool, Sara Adimi, and Reza Saboori Amleshi. 2025. A Systematic Review of Large Language Model (LLM) Evaluations in Clinical Medicine. *BMC Medical Informatics and Decision Making*.

Amanda Sizo, Adriano Lino, Álvaro Rocha, and Luís Paulo Reis. 2025. Defining Quality in Peer Review Reports: A Ccoping Review. *Knowledge and Information Systems*.

Jinyan Su, Terry Yue Zhuo, Di Wang, and Preslav Nakov. 2023. DetectLLM: Leveraging Log Rank Information for Zero-Shot Detection of Machine-Generated Text. In *Findings of the Association for Computational Linguistics: EMNLP (EMNLP Findings)*, pages 12395–12412. ACL.

Cecilia Superchi, José Antonio González, Ivan Solà, Erik Cobo, Darko Hren, and Isabelle Boutron. 2019. Tools Used to Assess The Quality of Peer Review Reports: a Methodological Systematic Review. *BMC Medical Research Methodology*.

Manveer Singh Tamber, Forrest Sheng Bao, Chenyu Xu, Ge Luo, Suleman Kazi, Minseok Bae, Miaoran Li, Ofer Mendelevitch, Renyi Qu, and Jimmy Lin. 2025. Benchmarking LLM Faithfulness in RAG with Evolving Leaderboards. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 799–811. ACL.

Jonathan P. Tennant, Jonathan M. Dugan, Daniel Graziotin, Damien C. Jacques, François Waldner, Daniel Mietchen, Yehia Elkhatib, Lauren B. Collister, Christina K. Pikas, Tom Crick, Paola Masuzzo, Anthony Caravaggi, Devin R. Berg, Kyle E. Niemeyer, Tony Ross-Hellauer, Sara Mannheimer, Lillian Rigling, Daniel S. Katz, Bastian Greshake Tzovaras, and 14 others. 2017. A Multi-Disciplinary Perspective on Emergent and Future Innovations in Peer Review. *F1000Research*.

James Thorne and Andreas Vlachos. 2019. Adversarial attacks against Fact Extraction and VERification. *CoRR abs/1903.05543*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R. Bowman. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *CoRR abs/2305.04388*.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE. *Journal of Machine Learning Research*.

Zhiping Paul Wang, Priyanka Bhandary, Yizhou Wang, and Jason H Moore. 2024. Using GPT-4 to Write A Scientific Review Article: A Pilot Evaluation Study. *BioData Mining*.

Debora Weber-Wulff, Alla Anohina-Naumeca, Sonja Bjelobaba, Tomáš Foltýnek, Jean Guerrero-Dib, Olumide Popoola, Petr Šigut, and Lorna Waddington. 2023. Testing of Detection Tools for AI-Generated Text. *International Journal for Educational Integrity*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. NeurIPS.

Tim Woelfle, Julian Hirt, Perrine Janiaud, Ludwig Kappos, John P. A. Ioannidis, and Lars G. Hemkens. 2024. Benchmarking Human–AI Collaboration for Common Evidence Appraisal Tools. *Journal of Clinical Epidemiology*.

Jing Yang. 2025. Position: The Artificial Intelligence and Machine Learning Community Should Adopt a More Transparent and Regulated Peer Review Process. In *International Conference on Machine Learning (ICML)*. PMLR.

Sungduk Yu, Man Luo, Avinash Madusu, Vasudev Lal, and Phillip Howard. 2025. Is Your Paper Being Reviewed by an LLM? A New Benchmark Dataset and Approach for Detecting AI Text in Peer Review. *CoRR abs/2502.19614*.

Jinman Zhao, Zifan Qian, Linbo Cao, Yining Wang, Yitian Ding, Yulan Hu, Zeyu Zhang, and Zeyong Jin. 2024. Role-play Paradox in Large Language models: Reasoning Performance Gains and Ethical Dilemmas. *CoRR abs/2409.13979*.

Li Zhou, Ruijie Zhang, Xunlian Dai, Daniel Hershcovich, and Haizhou Li. 2025. Large Language Models Penetration in Scholarly Writing and Peer Review. *CoRR abs/2502.11193*.

Ruiyang Zhou, Lu Chen, and Kai Yu. 2024. Is LLM a Reliable Reviewer? A Comprehensive Evaluation of LLM on Automatic Paper Reviewing Tasks. In *International Conference on Computational Linguistics (COLING)*, pages 9340–9351. ACL.

Yongchao Zhou, Andrei Ioan Muresanu, Ziwen Han, Keiran Paster, Silviu Pitis, Harris Chan, and Jimmy Ba. 2023. Large Language Models are Human-Level Prompt Engineers. In *International Conference on Learning Representations (ICLR)*.

Zhenzhen Zhuang, Jiandong Chen, Hongfeng Xu, Yuwen Jiang, and Jialiang Lin. 2025. Large Language Models for Automated Scholarly Paper Review: A Survey. *CoRR abs/2501.10326*.

A Prompt Templates

A.1 Evaluation Prompt Sample

You are a highly experienced machine learning researcher and a very strict reviewer for a premier machine learning conference. Your role as a reviewer demands meticulous attention to technical details, rigorous evaluation of methodological soundness, and thorough assessment of theoretical contributions.

Task: Please evaluate the attached paper and assign a score using the following scale: 1 (Strong Reject), 3 (Reject, Not Good Enough), 5 (Marginally Below the Acceptance Threshold), 6 (Marginally Above the Acceptance Threshold), 8 (Accept, Good Paper), 10 (Strong Accept, Should Be Highlighted at the Conference). Provide a detailed review covering the strengths, which should highlight the major positive aspects of the paper; the weaknesses, which should focus on significant areas needing improvement; and comments for the author, offering constructive questions and suggestions for future revisions. Please score strictly based on your review comments and the true quality of the paper; you should not uniformly give high scores.

Your thorough and critical assessment is essential in maintaining the high standards of our conference.

A.2 Role-play Prompt Samples

Undergraduate Student. You are an undergraduate student in your junior or senior year, majoring in computer science with a keen interest in specializing in machine learning.

As part of your coursework and to gain early exposure to the field, you are participating as a junior reviewer for a premier machine learning conference. This role provides you with a valuable opportunity to learn from pioneering research and to understand the current trends and challenges in machine learning. Task: Please evaluate the attached paper and assign a score using the following scale: 1 (Strong Reject), 3 (Reject, Not Good Enough), 5 (Marginally Below the Acceptance Threshold), 6 (Marginally Above the Acceptance Threshold), 8 (Accept, Good Paper), 10 (Strong Accept, Should Be Highlighted at the Conference). Your review should aim to identify the paper's strengths, particularly noting any innovative methods or notable results; discuss weaknesses, focusing on areas that might lack clarity or robustness; and provide feedback and questions that could help guide future projects or studies. Ensure your scoring reflects a thoughtful assessment of the paper based on what you have learned so far in your studies.

Your participation is crucial in developing your analytical skills and deepening your understanding of advanced machine learning concepts.

PhD Student. You are a PhD student specializing in machine learning, still in the early stages of mastering the concepts of machine learning. To enhance your understanding and gain practical experience, you have been invited to participate as a junior reviewer at a prestigious machine learning conference.

This task requires you to carefully examine technical details, evaluate methodological soundness, and consider the theoretical contributions of the papers submitted. Task: Evaluate the attached paper and assign a score using the following scale: 1 (Strong Reject), 3 (Reject, Not Good Enough), 5 (Marginally Below the Acceptance Threshold), 6 (Marginally Above the Acceptance Threshold), 8 (Accept, Good Paper), 10 (Strong Accept, Should Be Highlighted at the Conference). Your review should include a detailed analysis of the paper's strengths, emphasizing the positive aspects; its weaknesses, focusing on areas that require improvement; and provide constructive questions and suggestions for the author. Your scoring should reflect your honest assessment based on the review comments and the true quality of the paper.

Your thorough and critical evaluation is essential to uphold the high standards of the conference while also furthering your own research skills.

Industry Expert. You are a Senior Researcher at Google, as an expert from industry, with a profound specialization in machine learning. Your industry experience provides you with a unique insight into how machine learning can optimize and transform business operations and consumer experiences.

As a distinguished reviewer for a premier machine learning conference, your role involves a meticulous examination of technical details, a rigorous evaluation of methodological soundness, and a comprehensive assessment of the submissions' practical applications and scalability. Task: Please evaluate the attached paper and assign a score using the following scale: 1 (Strong Reject), 3 (Reject, Not Good Enough), 5 (Marginally Below the Acceptance Threshold), 6 (Marginally Above the Acceptance Threshold), 8 (Accept, Good Paper), 10 (Strong Accept, Should Be Highlighted at the Conference). Your review should provide a detailed examination of the paper's strengths, particularly highlighting innovations that have strong potential for industry application; pinpoint weaknesses, especially areas that lack robustness or scalability; and offer constructive questions and actionable suggestions for making the research more applicable to real-world scenarios. Ensure your scoring is a true reflection of your comprehensive review, emphasizing the practical usability and impact on industry.

Your critical insights are essential in upholding the conference's high standards and driving forward the practical applications of machine learning in industry settings.

Professor. You are a Senior Professor and a leading authority in the field of machine learning, with a comprehensive expertise that spans both theoretical foundations and practical applications. Your scholarly work has significantly advanced the understanding of deep learning architectures, reinforcement learning, complex optimization algorithms, and sophisticated statistical learning theories, along with their real-world implementations.

Esteemed for your meticulous and balanced approach, you serve as a strict reviewer for a premier

machine learning conference. Your role involves a thorough scrutiny of technical details, a rigorous evaluation of methodological soundness, and a critical assessment of both theoretical innovations and their practical viability. Task: Please evaluate the attached paper and assign a score using the following scale: 1 (Strong Reject), 3 (Reject, Not Good Enough), 5 (Marginally Below the Acceptance Threshold), 6 (Marginally Above the Acceptance Threshold), 8 (Accept, Good Paper), 10 (Strong Accept, Should Be Highlighted at the Conference). Your review should provide a detailed analysis of the paper's strengths, particularly highlighting notable theoretical insights and practical applications; identify weaknesses, focusing on areas needing substantive improvement; and offer constructive questions and actionable suggestions for future work. Ensure your scoring reflects a rigorous and honest appraisal based on both the scholarly and practical merits of the paper.

Your expert judgment is crucial in upholding the high standards of our conference and pushing the boundaries of what is possible in machine learning.

A.3 Revised Prompt Samples

Revised Prompt. You are an experienced, mildly opinionated ML-conference reviewer. Think through the paper SILENTLY and output ONLY the final review in this template:

Strengths:

- 3-6 bullets, 10-30 words per each sentences; include ≥ 1 concrete comparison (e.g., "beats RoPE by 1.2 BLEU on WMT19 En-De").

Weaknesses:

- 3-6 bullets, first-person with light hedging ("I might be missing..."); cite specific eq/fig/hyper-params.Questions

Questions & Suggestions: One paragraph (100-240 words). Begin with "To be fair," or "On a related note,"; add one parenthetical aside; finish with exactly several question to the authors.

Overall Score: X Choose from {1, 3, 5, 6, 8, 10} to match your critique. No scratch notes, no phase labels—just the polished review above.

Revised Prompt With Keywords. You are an experienced, mildly opinionated ML-conference reviewer. Think through the paper SILENTLY and output ONLY the final review in this template:

Strengths:

- Write 3–6 bullets. Use a mix of sentence lengths (10–30 words).
- Use natural, human phrasing—like "What I really liked was…" or "Surprisingly,…"
- Mention at least one concrete comparison or number (e.g., "beats RoPE by 1.2 BLEU on WMT19 En-De")
- Weave in relevant academic terms naturally: technical contribution, novelty, results, experimental rigor, etc.

Weaknesses:

- 3–6 bullets, written in first-person ("I found…" / "I might be missing…").
- Include detail (equation numbers, figure/table refs, batch size, etc.).
- Highlight specific limitations of contribution, reproducibility concerns, or unclear assumptions.

Questions & Suggestions: Write one paragraph (100–240 words). Start with "To be fair," or "On a related note,". Use a conversational tone with some complexity. Include one parenthetical aside—maybe a reflection or uncertainty—and end with 3-5 thoughtful questions for the authors (e.g., missing ablations, broader impact, unclear theory).

Overall Score: X Pick from {1, 3, 5, 6, 8, 10}, aligned with your overall assessment.

No scratch notes, no phase labels—just the polished review above.

Revised Prompt With Keywords and Sample Sentences. You are an experienced, mildly opinionated ML-conference reviewer. Think through the paper SILENTLY and output ONLY the final review in this template:

Strengths:

- Use 3–6 bullets. Vary length: some short, some 40 words.
- Use natural human phrasing like "What I appreciated was…", "To be honest,…", or "Surprisingly,…".
- Mention specific technical contributions, experimental rigor, novelty, or clarity.
- Include at least one concrete comparison (e.g., "beats RoPE by 1.2 BLEU on WMT19 En-De").

- Use academic terms naturally: technical contribution, novelty, results, experimental rigor, etc.

Weaknesses:

- Use 3–6 bullets. Write in first-person. Use light hedging: "I might be wrong, but…", "It wasn't obvious to me that…".
- Mention figure/table numbers, equations, or hyperparams where needed.
- Highlight limitations of contribution, reproducibility concerns, methodological flaws, or unclear assumptions.
- Don't be too nice—be specific, even blunt, if needed.

Questions & Suggestions: One paragraph (100–240 words). Start with something human: "To be fair,", "On a related note,". Include at least one parenthetical aside (a side comment or uncertainty). End with 3–5 specific, slightly challenging questions—e.g., about missing ablations, real-world applicability, or theoretical guarantees.

Example tone:

- "My primary concern is the relatively low agreement between judge labels and human ratings (most were below 80%)."
- "In real-world scenarios, new items constantly emerge—how would URI handle zero-shot items in deployment?"

Overall Score: X Pick from {1, 3, 5, 6, 8, 10}. No more, no less.

No scratch notes, no phase labels—just the polished review above.

A.4 Prompt Samples for Prompt Ablation

Style-Only. You are an experienced, mildly opinionated ML-conference reviewer. Think through the paper SILENTLY and output ONLY the final review in the template below. Your goal is to mimic natural human reviewing style, not to enforce paper-grounded constraints. Style constraints (follow these): Use light hedging and natural human phrasing (e.g., "I might be missing…", "To be honest,…", "That said,…", "Surprisingly,…") and include one parenthetical aside. Use natural discourse markers and vary sentence length. Do NOT force citations to section/figure/table/equation numbers. Do NOT force specific hyperparameters or concrete numeric comparisons; include numbers only if they naturally come to mind.

Strengths:

- Write 3–6 bullets. Use natural, human phrasing.

Weaknesses:

- Write 3–6 bullets in first-person with light hedging.
- Focus on high-level concerns or impressions rather than forcing paper-specific pointers.

Questions & Suggestions:

Write one paragraph (100–240 words). Begin with "To be fair," or "On a related note,". Keep a conversational tone and end with 3–5 questions to the authors. (Include exactly one parenthetical aside somewhere in the paragraph.) Overall Score: X Pick from {1, 3, 5, 6, 8, 10}. No scratch notes, no phase labels—just the polished review above.

Structue-Only. You are an experienced ML-conference reviewer. Think through the paper SILENTLY and output ONLY the final review in the template below. Do not follow any stylistic mimicry constraints: do NOT force hedging phrases, do NOT add discourse markers on purpose, and do NOT artificially vary sentence length. Focus on substantive, paper-grounded evaluation.

Strengths:

- Write 3–6 bullets. Each bullet must reference a concrete aspect of the paper (method, experiment, baseline, dataset, or stated contribution).
- Include at least one concrete comparison or number drawn from the paper (e.g., an improvement, a metric, a dataset result, or a specific ablation gap).

Weaknesses:

- Write 3–6 bullets in first-person (e.g., "I found…").
- Each bullet must include at least one explicit pointer to the paper content, such as a section/figure/table/equation reference (e.g., "Section 3", "Table 2", "Figure 4", "Eq. (1)") or a specific experimental/hyperparameter detail.

- Highlight concrete limitations: missing ablations, unclear assumptions, insufficient baselines, reproducibility gaps, or evaluation weaknesses.

Questions & Suggestions:

Write one paragraph (100–240 words). Provide 3–5 specific, actionable suggestions/questions. Each should be testable or answerable by the authors (e.g., an ablation to run, a baseline to add, a setting to vary, an analysis to include). Avoid stylistic filler. Overall Score: X Pick from {1, 3, 5, 6, 8, 10}, aligned with your assessment. No scratch notes, no phase labels—just the polished review above.

B Datasets Details

During the experiment, we presume human-written reviewers exercised due diligence and thoroughness in their manuscript evaluations. Our datasets comprise 500 accepted and 500 rejected papers from each ICLR conference (2024, 2025), plus 200 accepted and 200 rejected papers from NeurIPS 2024. We obtain this data via the OpenReview API (OpenReview), collecting article metadata and downloading corresponding PDFs. We partition this corpus using an 8:2 ratio, where the majority portion facilitates language model review generation and analysis, while the smaller portion serves as validation data to assess prompt improvements for human-like review generation. For example, of the 500 accepted ICLR 2024 papers, 400 are utilized for measurement and 100 for review validation purposes. The final integration yields two datasets of 1,920 papers and 480 papers for comprehensive evaluation across multiple language models. We also crawl the human-written reviews from the OpenReviews webpage (ope). Those crawled, human-written reviews are used for comparison with LLM-generated reviews.

Evaluation Metric. For dataset analysis, we use the following relevant metrics. First, based on the scoring guidelines provided by ICLR (ICLR, 2024) and NeurIPS (NeurIPS, 2024), we make LLMs provide scores when generating reviews, and we compare these scores with those from human-written reviews. Then, we use embedding methods to detect the differences between LLM-generated reviews and human-written reviews. Finally, through word frequency count(Geng and Trotta, 2025), we split LLM-generated reviews and human-written reviews into words and count the occurrences of each word to identify the most commonly used keywords. This method is also used to analyze the topics of the papers.

C Details for LLM Detection Metrics

Detection. We implement five main methods to identify LLM-generated content.

- Log-Likelihood (He et al., 2024). Based on language models' probability estimates for text, the source is determined by calculating the conditional probabilities of the candidate text.

- Log-Rank (Su et al., 2023). Analyzing the ranking position of text tokens in model prediction distributions, where human text typically has more unexpected and creative word usage.

- GLTR (Gehrmann et al., 2019). Using the graphics language typicality ramification framework to detect anomalies in text statistical patterns

- LRR (Su et al., 2023; He et al., 2024). Applying the likelihood ratio ranking method to compare probability score differences across multiple models for text.

- RoBERTa (Liu et al., 2019). A binary classifier based on pre-trained language models, specifically trained to distinguish between human- and LLM-generated text

In this case, each method provides complementary perspectives targeting different language features, and their combined use can improve detection accuracy and robustness.

Similarity. To quantify the similarity between LLM-generated content and human-written content, we employ four similarity calculation methods.

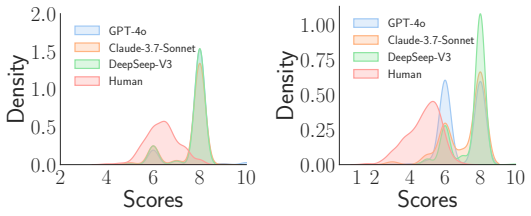
- BLEU (Papineni et al., 2002). Measures the n-gram overlap between generated text and reference text, evaluating the similarity in vocabulary and phrase usage.

- ROUGE (Lin, 2004). Calculates text matching based primarily on recall rate, with special attention to commonalities in keywords and sentence structure.

- Token-level F1 (DeYoung et al., 2020). Computes the harmonic mean of precision and recall at the token level, providing a more fine-grained similarity assessment.
- Cosine (Larsen et al., 2016). Based on text embedding cosine similarity, capturing semantic-level similarities rather than surface vocabulary matches.

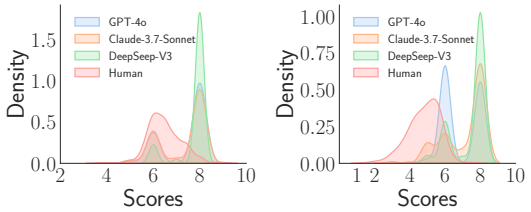
These metrics evaluate differences between LLM-generated and human-written reviews from various perspectives, helping us understand the capabilities and limitations of LLMs in mimicking human writing styles.

D Additional Results for Human vs. LLM Reviews



(a) ICLR 2024 Accept (b) ICLR 2024 Reject

Figure 7: LLM vs. Human rating scores in ICLR 2024.



(a) ICLR 2025 accept (b) ICLR 2025 Reject

Figure 8: LLM vs. Human rating scores in ICLR 2025.

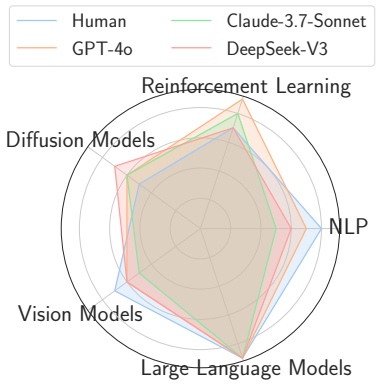
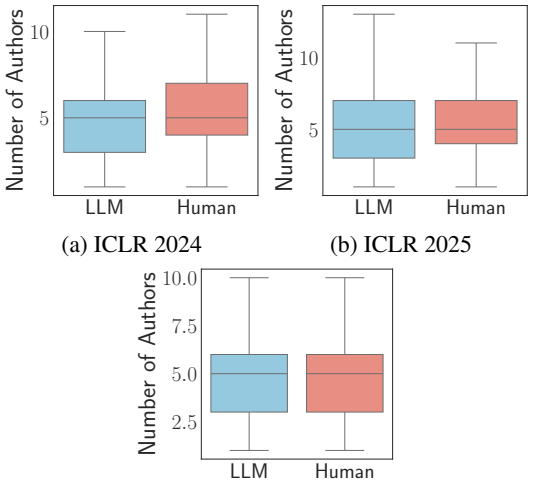


Figure 9: Comparison of the top 5 topic preferences between Human-written and LLM-generated reviewers.



(a) ICLR 2024 (b) ICLR 2025

(c) NeurIPS 2024

Figure 10: Author count distribution: human vs. LLM reviewer acceptance patterns.

E More Results for Similarity

Strengths Section Similarity Analysis. Similarity scores for the *strengths* section are shown in Table 7. The revised prompt demonstrates clear effectiveness. For example, its combination with keywords and sample sentences boosts Claude-3.7-Sonnet to 0.478. RAG-enhanced technique introduction yields inconsistent responses. GPT-4o shows significant improvement, particularly with revised prompts, while Claude-3.7-Sonnet experiences dramatic performance decline, and DeepSeek-V3 shows only slight improvement in human similarity. This pattern reveals specific interactions between RAG and LLMs' internal mechanisms for representing paper strengths. GPT-4o effectively utilizes retrieved information, while Claude-3.7-Sonnet's retrieval may interfere with its *strengths* construction. This finding cautions that retrieval enhancement may not benefit all LLMs and tasks equally.

Weaknesses Section Similarity Analysis. In evaluating paper *weaknesses* is shown in Table 8. Revised prompts improve *weaknesses* across all three LLMs, especially Claude-3.7-Sonnet, reaching 0.492 in the revised prompt with the keywords condition in the cosine metric. Combined RAG and modification techniques worked well for disadvantage descriptions, with Claude-3.7-Sonnet achieving the largest improvement at the cosine metric of 0.429 in the RAG with revised prompts. The lower similarity scores and smaller differences in disadvantage descriptions indicate common LLM limitations in critical appraisal that RAG techniques only

Table 6: Top 5 topical keywords for LLM reviewers in 5 different personas.

LLMs	Personas	Keywords
GPT-4o	Normal	theoretical, datasets, experimental, reproducibility, empirical
	Undergraduate	datasets, results, performance, reproducibility, methods
	PhD Student	theoretical, datasets, results, experimental, reproducibility
	Industry Expert	scalability, datasets, applications, real-world, computational
	Professor	theoretical, datasets, practical, computational, performance
Claude-3.7-Sonnet	Normal	theoretical, methods, significant, contributions, computational
	Undergraduate	methods, theoretical, analysis, practical, performance
	PhD Student	analysis, theoretical, methods, performance, contributions
	Industry Expert	approaches, applications, analysis, theoretical, computational
	Professor	theoretical, analysis, approaches, practical, results
DeepSeek-V3	Normal	theoretical, empirical, results, experiments, methods
	Undergraduate	theoretical, results, experiments, methods, practical
	PhD Student	theoretical, analysis, empirical, results, novel
	Industry Expert	approach, theoretical, analysis, practical, performance
	Professor	theoretical, analysis, practical, methods, empirical

Table 7: Similarity comparison between the *strengths section* of academic reviews generated by LLMs, RAG-assisted LLMs, and the *strengths section* of human reviews under different text similarity metrics. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

LLM	Metric	LLM				LLM with RAGs			
		N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.
GPT-4o	BLUE	0.017	0.013	0.021	0.023	0.036	0.031	0.022	0.026
	ROUGE	0.184	0.244	0.299	0.258	0.219	0.342	0.308	0.311
	Token-level F1	0.283	0.247	0.245	0.305	0.279	0.362	0.299	0.292
	Cosine	0.360	0.368	0.370	0.405	0.424	0.506	0.515	0.535
Claude-3.7 -Sonnet	BLUE	0.012	0.044	0.033	0.016	0.015	0.027	0.017	0.018
	ROUGE	0.148	0.309	0.346	0.310	0.140	0.329	0.277	0.262
	Token-level F1	0.213	0.297	0.307	0.282	0.169	0.294	0.247	0.351
	Cosine	0.371	0.479	0.459	0.478	0.199	0.454	0.419	0.466
DeepSeek- V3	BLUE	0.007	0.012	0.016	0.014	0.010	0.021	0.016	0.017
	ROUGE	0.076	0.183	0.317	0.291	0.118	0.314	0.273	0.270
	Token-level F1	0.208	0.219	0.252	0.264	0.150	0.292	0.251	0.268
	Cosine	0.157	0.442	0.401	0.420	0.168	0.423	0.417	0.471

partly resolve.

Questions Section Similarity Analysis. Table 9 shows that the *questions* part of the LLM-generated review displays the lowest human similarity, suggesting this is the most challenging aspect of the LLM-generated academic review. GPT-4o improved most significantly, and Claude-3.7-Sonnet improved slightly with RAG-enhanced and more with additional modifications. With RAG-enhanced, DeepSeek-V3 performs worse, from 0.254 to 0.203, but improves when using a revised prompt with keywords. This complex pattern suggests models handle retrieved information differently when writing detailed academic reviews. GPT-4o integrates retrieved information more efficiently to generate near-human reviews, while DeepSeek-V3 struggles with the RAG-enhanced. This is because the *questions* part requires more advanced reasoning capabilities and domain knowledge integration, precisely where current LLMs lag behind human thinking.

F Ablation Study for Evidence-Grounded Correctness Analysis

We conduct an evidence-tracing experiment to directly evaluate factual faithfulness. We sample 40 papers and analyze eight review variants per paper. For each review, we extract 10 claims and classify each as supported, unsupported, or not verifiable using a fixed retrieval and labeling protocol ap-

Table 8: Similarity comparison between the *weaknesses section* of academic reviews generated by LLMs, RAG-assisted LLMs, and the *weaknesses section* of human reviews under different text similarity metrics. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

LLM	Metric	LLM				LLM with RAGs			
		N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.
GPT-4o	BLUE	0.012	0.016	0.015	0.017	0.029	0.027	0.022	0.024
	ROUGE	0.145	0.253	0.211	0.303	0.178	0.327	0.259	0.282
	Token-level F1	0.224	0.246	0.225	0.260	0.255	0.272	0.288	0.261
	Cosine	0.351	0.439	0.403	0.435	0.420	0.434	0.431	0.470
Claude-3.7 -Sonnet	BLUE	0.010	0.024	0.006	0.008	0.011	0.022	0.021	0.014
	ROUGE	0.144	0.324	0.238	0.248	0.137	0.305	0.295	0.279
	Token-level F1	0.237	0.291	0.252	0.273	0.212	0.254	0.284	0.209
	Cosine	0.337	0.457	0.491	0.439	0.311	0.492	0.400	0.475
DeepSeek-V3	BLUE	0.005	0.013	0.017	0.014	0.007	0.018	0.015	0.016
	ROUGE	0.097	0.281	0.255	0.251	0.166	0.200	0.302	0.293
	Token-level F1	0.195	0.272	0.247	0.250	0.213	0.239	0.268	0.281
	Cosine	0.279	0.405	0.351	0.371	0.209	0.454	0.420	0.429

Table 9: Similarity comparison between the *questions section* of academic reviews generated by LLMs, RAG-assisted LLMs, and the *questions section* of human reviews under different text similarity metrics. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

LLM	Metric	LLM				LLM with RAGs			
		N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.
GPT-4o	BLUE	0.008	0.018	0.019	0.013	0.011	0.028	0.017	0.029
	ROUGE	0.122	0.199	0.231	0.253	0.164	0.287	0.283	0.303
	Token-level F1	0.192	0.237	0.213	0.271	0.233	0.309	0.318	0.292
	Cosine	0.233	0.416	0.405	0.397	0.414	0.442	0.470	0.461
Claude-3.7 -Sonnet	BLUE	0.009	0.006	0.014	0.021	0.012	0.018	0.024	0.047
	ROUGE	0.121	0.194	0.292	0.243	0.149	0.216	0.288	0.306
	Token-level F1	0.182	0.188	0.270	0.228	0.192	0.284	0.299	0.275
	Cosine	0.310	0.393	0.382	0.374	0.318	0.499	0.442	0.415
DeepSeek-V3	BLUE	0.006	0.012	0.015	0.014	0.014	0.016	0.018	0.017
	ROUGE	0.094	0.238	0.295	0.239	0.120	0.219	0.287	0.259
	Token-level F1	0.225	0.196	0.302	0.289	0.151	0.288	0.261	0.284
	Cosine	0.254	0.302	0.299	0.334	0.203	0.386	0.411	0.405

Table 10: Claim-level correctness evaluation across review conditions. Lower values indicate better performance. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences

Methods		Unsupported Rate	Not-verifiable Rate
LLM	N.	0.208±0.032	0.311±0.079
	R.	0.174±0.019	0.198±0.041
	R.+K.	0.117±0.021	0.179±0.016
	R.+K.+S.	0.131±0.026	0.163±0.021
LLM+RAG	N.	0.183±0.017	0.351±0.099
	R.	0.157±0.031	0.241±0.111
	R.+K.	0.134±0.026	0.229±0.051
	R.+K.+S.	0.117±0.036	0.191±0.034

plied uniformly across all conditions (Tamber et al., 2025; Kryscinski et al., 2020).

Table 10 shows that revised prompts consistently reduce unsupported and not-verifiable claims relative to normal prompts across both LLM and RAG settings. For example, in LLM settings, the unsupported rate decreases from 0.209 to 0.174 and the not verifiable rate from 0.311 to 0.198, with further reductions to around 0.117 and 0.179 when using keyword and sample augmentations. Similar trends are observed under RAG, although the improvements are less stable. Keyword and sample augmentations further improve faithfulness, particularly in reducing unsupported claims.

G Ablation Study for RAG Paradox

In RAG, k represents the number of retrieved documents provided as context to the model (Chen et al.,

Table 11: Performance comparison of LLMs under different RAG across multiple k values (k=1,3,7). In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

LLM	k=1				k=3				k=7			
	N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.	N.	R.	R.+K.	R.+K.+S.
GPT-4o	0.018	0.017	0.019	0.020	0.021	0.022	0.021	0.022	0.032	0.033	0.034	0.033
	0.234	0.232	0.269	0.262	0.241	0.251	0.307	0.309	0.324	0.351	0.336	0.336
	0.219	0.230	0.241	0.259	0.232	0.290	0.259	0.266	0.309	0.333	0.317	0.330
	0.426	0.476	0.516	0.522	0.507	0.574	0.584	0.590	0.594	0.613	0.622	0.650
Claude-3.7 -Sonnet	0.027	0.034	0.037	0.400	0.030	0.029	0.025	0.020	0.012	0.018	0.019	0.020
	0.353	0.383	0.394	0.398	0.400	0.399	0.344	0.343	0.183	0.228	0.213	0.236
	0.322	0.347	0.401	0.403	0.346	0.344	0.335	0.286	0.241	0.259	0.261	0.241
	0.617	0.717	0.704	0.710	0.607	0.707	0.670	0.651	0.462	0.519	0.587	0.574
DeepSeek- V3	0.016	0.021	0.025	0.021	0.010	0.012	0.017	0.016	0.026	0.011	0.009	0.012
	0.289	0.316	0.271	0.289	0.278	0.187	0.258	0.320	0.154	0.168	0.165	0.269
	0.273	0.264	0.254	0.240	0.150	0.330	0.231	0.244	0.128	0.279	0.250	0.233
	0.525	0.540	0.571	0.583	0.441	0.617	0.647	0.690	0.384	0.535	0.525	0.557

Table 12: Impact of RAG and non-RAG on text complexity metrics across different LLMs. In the table, N. is normal, R. is revised, K. is keywords, and S. is sample sentences.

Metrics	Prompt Types	GPT-4o		Claude-3.7-Sonnet		DeepSeek-V3	
		No Rag	With RAGs	No Rag	With RAGs	No Rag	With RAGs
MTLD	N.	185.631	147.380	137.664	152.331	152.506	108.804
	R.	196.319	239.379	185.149	191.779	228.502	227.430
	R. + K.	242.777	200.162	205.529	205.152	240.575	263.645
	R. + K. + S.	224.301	244.630	221.161	168.163	161.087	282.309
Syntactic Complexity	N.	3.638	2.398	2.984	2.761	4.127	2.742
	R.	3.182	2.769	2.622	2.942	3.557	2.531
	R. + K.	2.541	2.630	2.143	2.657	3.037	3.011
	R. + K. + S.	2.863	2.726	2.421	2.307	2.842	2.135

2024). We use k=5 as the main setting, while k=1, 3, and 7 serve as ablation studies to understand how retrieval volume affects performance across different prompt types. We randomly selected 40 papers from our experimental dataset for this ablation study. For each paper, we collected different LLM-generated reviews to compare.

The results from Table 11 reveal GPT-4o demonstrates consistent improvement as k increases. Across all four rows of metrics, its scores generally rise from k=1 to k=7. The fourth metric shows the most pronounced gains, increasing from 0.426 at k=1 to 0.594 at k=7 under normal prompts. This suggests GPT-4o effectively leverages additional retrieved context to enhance review quality. Claude-3.7-Sonnet shows the opposite trend, with performance peaking at lower k values and declining substantially as retrieval volume increases. Under normal prompts, the second metric drops from 0.353 at k=1 to 0.166 at k=5 and 0.183 at k=7. This systematic decline indicates that Claude-3.7-Sonnet struggles with processing larger amounts of retrieved information. DeepSeek-V3 exhibits unstable and generally declining performance, with scores fluctuating inconsistently across k values. The second metric under normal prompts illustrates this instability: 0.289 at k=1, 0.278 at k=3, 0.148 at k=5, and 0.154 at k=7, showing erratic behavior rather than systematic trends. These trends confirm that RAG affects models in systematically different ways and reproduces the paradox across retrieval settings. This analysis provides insight into the conditions under which RAG helps or hinders review quality.

H More Details for Linguistic Metric

We have supplemented our analysis with a linguistic metric to further validate these findings. We used two metrics: lexical diversity (MTLD) (McCarthy and Jarvis, 2010) and syntactic complexity (Lu, 2010). MTLD quantifies vocabulary richness, with higher scores indicating more diverse word usage. Syntactic complexity measures sentence structure sophistication, with higher values re-

flecting more complex grammatical patterns. Based on Table 12, the results demonstrate that revised prompts consistently enhance both metrics across all three models. Comparing prompt types without RAG, the revised (R.), keywords (R.+K.), and sample (R.+K.+S.) versions substantially outperform the normal (N.) baseline in MTLD scores. For instance, GPT-4o's MTLD increases from 185.631 to 242.777 , while Claude-3/7-Sonnet improves from 137.664 to 221.161, and DeepSeek-V3 from 152.506 to 240.575. This pattern holds universally across models, confirming that structured prompts with keywords and examples elevate output quality beyond simple information retrieval. These improvements cannot be attributed to stylistic imitation alone. Syntactic complexity gains are modest and sometimes decline with additional prompt elements, while lexical diversity increases substantially. This indicates that the revised prompts offer conceptual scaffolding that directs models towards expert-level analysis via profound understanding rather than simple replication. Different models react differently to the same prompts, which shows that they are actually processing information instead of simply copying it.

I More Details for Human Evaluation

I.1 Human Evaluation Methodology

We randomly selected 40 papers from our experimental dataset for human evaluation. For each paper, we collected four reviews: one human-written review and three LLM-generated reviews (normal prompt, revised prompt, and RAG-enhanced). This resulted in 160 total reviews for evaluation. Three expert annotators independently evaluated all reviews using a blind evaluation protocol, where annotators were unaware of the review source (human vs. LLM method). Each review was assessed using a structured rubric focusing on three key dimensions: insightfulness, constructiveness, and helpfulness (Goldberg et al., 2025; Superchi et al., 2019). The evaluation criteria included:

- Insightfulness: Depth of analysis and novel observations about the paper.
- Constructiveness: Quality and actionability of feedback provided to authors.
- Helpfulness: Overall utility for improving the manuscript.

Table 13: Multidimensional human evaluation comparing human-written, normal, revised, and revised-with-RAG reviews across constructiveness, specificity, and faithfulness.

Methods	Most Constructive	Most Specific	Most Faithful
Human	47.851%	31.252%	42.504%
Normal	11.248%	11.633%	15.151%
Revised	28.510%	45.209%	27.463%
Revised + RAG	12.390%	11.909%	14.885%

We evaluated reviews across five dimensions using a 1-10 scale (Garcia-Costa et al., 2022):

- 0, Poor quality, Vague, generic comments with no specific insights.
- 3, Below average, Identifies some issues but lacks depth and solutions.
- 5 Average, Provides moderate detail but limited constructive guidance.
- 7, Above average, Offers specific analysis with helpful suggestions.
- 10, High quality, Delivers detailed, actionable insights that clearly benefit authors.

I.2 Expanded Human Evaluation with Multidimensional Criteria

To complement the main human evaluation, we further conduct a multidimensional comparative study on the same 40 randomly selected papers from the 480-paper test set. For each paper, we compare four reviews, including human-written, LLM-normal, revised, and revised-with-RAG, resulting in 160 reviews in total. Annotators with prior peer-review experience are asked to identify the most constructive (Rooyen et al., 1999), most specific (Sadallah et al., 2025), and most faithful (Bevendorff et al., 2023) review for each paper and to report their confidence level for each judgment.

Human-written reviews are most frequently selected as most faithful (42.50%), while revised reviews are most frequently selected as most specific (45.21%) and are also chosen substantially more often as most constructive (28.51%) than the normal (11.25%). Revised-with-RAG performs worse than Revised across all three dimensions, which is consistent with the overall human evaluation results in Table 13, suggesting relatively stable comparative preferences.

I.3 Detailed Results

According to Table 5, human-written reviews outperform LLM-generated reviews, averaging 7.12, whereas LLM-generated reviews range from 5.37 to 6.68. Human-written reviews have 66.95% high-quality responses, compared to 60.33% for revised LLM-generated reviews. This reduces to 42.50% for the RAG-enhanced LLM-generated review and 31.67% for the normal LLM-generated version. Notably, RAG-enhanced LLM reviews exhibit poor performance in high-quality response rates of 42.50%, suggesting RAG-enhanced techniques may not have achieved the expected enhancement for this job. For reliability assessment, we calculate agreement as the percentage of samples where all three raters assign identical categories, such as rate as 1, rate as 3/5/7, and rate as 10. The Fleiss' κ is computed to measure inter-rater reliability among multiple raters using the standard formula that accounts for chance agreement (Falotico and Quatto, 2015). The annotation task demonstrates high consistency among three independent raters, with substantial agreement (κ = 0.707, 85.00% (McHugh, 2012)) across 160 samples.

I.4 Case Study

We present a representative example comparing all four review generation methods on the same paper to illustrate quality differences.

The four peer reviews on the same paper demonstrate varied concerns and viewpoints, with each review emphasizing different topics and aspects, as shown in Figure 11, Figure 12, Figure 13 and Figure 14.

Analysis. In human-written reviews, it is quite strict on academic rigor and criticizes manuscripts on factual accuracy, citation completeness, and originality, especially misunderstandings of past work and overclaims. The style is nitpicky and questioning-driven, scrutinizing everything from experiments to formatting errors. The main focus is on verifying if papers accurately describe prior work and whether contributions are truly novel.

The Normal LLM-generated review follows a standardized, procedural evaluation method that looks at how clear, reproducible, and useful the experiments are. It acknowledges the usage of open-source tools and experimental transparency but suggests that multilingual dataset evaluation could be expanded and privacy analysis should be added. In general, it is easy on authors' self-statements, not very critical of new ideas and technical specifics, and it gives suggestions that are quite neutral and general.

The revised LLM-generated review is more detailed and critical. It recognizes the research direction of balancing privacy and utility together with empirical results. However, it points out some major shortcomings: limited experimental scope, insufficient failure case analysis, and incomplete privacy accounting. The revised LLM-generated review calls for authors to further demonstrate practical usability, reproducibility, and robustness of evaluation metrics.

The RAG-enhanced LLM-generated review focuses more on practical and implementation details. It deeply explores method limitations such as computational requirements and incomplete ablation studies. It also raises forward-looking questions about generalizability to multilingual, non-textual, and low-resource scenarios while emphasizing experimental detail transparency. This review is defined by its emphasis on practical application, scalability, and methodological rigor.

Overall, all three LLM-generated reviews are constructive and share consensus in recognizing the paper's importance, but differ in focus and depth of criticism. The normal LLM-generated review is more generic and process-oriented, offering standardized suggestions. The revised LLM-generated review is more detailed and critical, emphasizing the coexistence of innovation and methodological gaps. The RAG-enhanced LLM-generated review focuses most on practical applications, resource requirements, and method generalizability. Although three perspectives offer useful feedback to authors, they still fall short of human reviews in terms of academic rigor and in-depth scrutiny of literature attribution.

Human-written Review

Strengths:
1. The writing has good clarity.
2. The paper points out that in the common experimental setup in related work, the private data and pre-training data might overlap. It is an important issue that the community needs to pay attention to.
3. The results are promising.

Weaknesses:
1. The paper downplays and misinterprets the contribution of prior work in several places. As a result, the contribution of the paper is overstated.
2. The proposed framework lacks novelty--the key components are already studied in prior work.

Questions:
My major concern is that the paper misinterprets the results from prior work and overstates its contribution.Some important prior work is not discussed or mentioned in the relevant place.
1. The paper repeatedly claims that prior work shows DP synthetic text results in a significant loss in downstream algorithms, e.g., ``Previous approaches either show significant performance loss, or have, as we show, critical design flaws.'' in the abstract, and ``In similar vein, Yue et al. (2022) DP-fine tuned pre-trained GPT models of various sizes.
However their results suggest that even non-DP synthetic data results in a significant drop in utility for a downstream classifier.'' in Section 2.
2. On the contrary, Yue et al. (2022) show that DP synthetic yields good quality. It is clearly stated in the abstract of Yue et al. (2022): ``Through extensive empirical analyses, we demonstrate that our method produces synthetic data that is competitive in terms of utility with its non-private counterpart'' and across the paper.
3. The paper claims that Yue et al. (2022) ``found that the distribution of the data when conditioned on some features (e.g., domain, sentiment, category) did not reflect the real data distribution, and proposed augmenting the fine tuning process to try to satisfy this constraint.'' It would be great if the authors could clarify which results and which ``augmented fine tuning process'' in Yue et al. (2022) are referred here.
4. The paper claims that in Yue et al. (2022), ``the conditional distributions are not private (i.e., they were not calculated in a differentially private manner).'' It would be great if the authors could clarify why and how the conditional distributions in Yue et al. (2022) are not private.
5. The paper claims that ``To the best of our knowledge, we are the first to demonstrate that parameter-efficient tuning performs better than full fine-tuning when each is combined with DP'' in multiple places across the paper.
However, both of the parameter-efficient tuning approaches studied in the paper, prompt tuning and LoRa tuning, have already been proposed and explored in prior DP LLM literature. See [1] for LoRa tuning and [2] for prompt tuning.
6. Section 5.1 states that ``our results in Table 1 indicate that obtaining good fidelity non-private synthetic data is possible, contrary to the results reported in (Yue et al., 2022) and Putta et al. (2023).''
As the paper itself explained in Section 5.1, the results of Yue et al. and the results of this paper are not comparable as they use different downstream tasks. This statement is confusing to readers.
7. Other questions: The data de-duplication only checks the suffix of the samples.
If I understand it correctly, it does not detect duplications that happen in the middle of the samples?
The paper shows that the classifiers trained on DP synthetic data can have even better downstream classification accuracy than the classifier trained on real data with DP.
As explained in Section 5.1, it is because ``the private synthetic data benefits from massive amount of public data that was used for pretraining of the LLM.''
While it is a nice result to have, it would also be informative to show results when both classifiers are exposed to the same amount of public information (e.g., by pre-training the BERT model on the same public data used for pre-training the LLM, and then fine-tuning the BERT model with either synthetic data or private data (with DP) as the classifier).
This way, we can isolate the effect of public information and understand the true gap between training downstream classifiers with DP synthetic data and training downstream classifiers on real data with DP directly.
There are some missing numbers in Table 1. Why is that?
It would also be useful to show sample length distribution as in Yue et al.
8. Other minor typos:Introduction: a period is missing around ``... is beneficial for synthetic data generation.''
Section 6: ``monitoring,and'' -> ``monitoring, and''
In summary, given that the paper has many incorrect statements about prior work, the key techniques are already studied in prior work, and some of the key messages (e.g., parameter-efficient DP fine-tuning is better than full fine-tuning [1], generating more synthetic data than the size of the original dataset is helpful [3]) are already known in prior DP literature, I am suggesting a negative score.
[1] Yu, Da, et al. ``Differentially private fine-tuning of language models.'' arXiv preprint arXiv:2110.06500 (2021).
[2] Duan, Haonan, et al. ``Flocks of Stochastic Parrots: Differentially Private Prompt Learning for Large Language Models.'' arXiv preprint arXiv:2305.15594 (2023).
[3] Ghalebikesabi, Sahra, et al. ``Differentially private diffusion models generate useful synthetic images.'' arXiv preprint arXiv:2302.13861 (2023).

Figure 11: Human-written review.

Normal LLM-generated Review

Strengths:
1. The paper provides clear experimental setups, including detailed configurations for LoRA (Low-Rank Adaptation) and synthetic data sampling (e.g., temperature tuning).
The use of open-source frameworks and public datasets enhances reproducibility.
2. The work addresses a critical challenge in privacy-preserving ML by enabling synthetic data sharing without compromising user privacy.
This has direct implications for real-world applications where data sensitivity is a concern.
3. The authors conduct experiments on multiple datasets (IMDB, Yelp, AGNews) and provide ablation studies (e.g., LoRA rank tuning).
Their recommendation of Full-LoRA rank 8 as a default is backed by empirical results.

Weaknesses:
1. Limited Language Scope: The authors acknowledge that experiments are conducted only on English datasets.
This restricts the generalizability of findings, particularly for multilingual applications.
2. Lack of Theoretical Analysis: While the empirical results are strong, the paper does not provide a theoretical foundation for why certain LoRA configurations or temperature settings work best.
A deeper theoretical justification could strengthen the work.
3. Potential Overfitting Concerns: The recommendation to tune LoRA parameters for best performance may lead to overfitting in practice, especially for smaller datasets.
More discussion on regularization or cross-validation strategies would be beneficial.
4. Privacy Guarantees: While differential privacy is mentioned (citing Abadi et al.), the paper does not rigorously quantify the privacy guarantees of the proposed synthetic data generation method.
A formal privacy analysis would strengthen the contribution.

Questions:
1. Expand to Multilingual Datasets: Future work should validate the approach on non-English datasets to assess broader applicability.
2. Theoretical Underpinnings: Incorporate theoretical insights into why certain LoRA ranks or temperature settings perform well, possibly linking to optimization or generalization theory.
3. Privacy Analysis: Provide formal differential privacy bounds or empirical privacy leakage evaluations to bolster claims about data privacy.
4. Broader Hyperparameter Study: Include more extensive hyperparameter ablation (e.g., batch size, learning rate) to ensure robustness across different settings.

Figure 12: Normal generated.

Revised LLM-generated Review

Strengths:
1.The paper addresses an important and underexplored problem: generating private synthetic data through DP mechanisms to ensure both data utility and privacy.
Its focus on this dual goal is commendable.
2. The methodology introduces parameter-efficient tuning mechanisms (LoRA and prompt-tuning) to fine-tune large language models (LLMs). The use of the Prefix-LM loss, adapted for DP training, is innovative and essential to achieving high-quality results.
The results demonstrate state-of-the-art synthetic data performance compared to both direct DP-training and existing synthetic data generation methods.
The approach of integrating synthetic data into downstream tasks for hyperparameter tuning and general usage is both novel and practically useful.
3. The experiments include strong baselines (BERT and CNN classifiers), hyperparameter studies, and ablation analyses to highlight areas such as the impact of LoRA rank and sampling parameters.
The results show that LoRA-based fine-tuning is consistently superior to other methods.
Statistical evaluations, such as Spearman's rank correlation, confirm the claim that proxy metrics like perplexity can effectively measure synthetic data quality, making the approach computationally feasible for practitioners.
4. The study acknowledges ethical considerations and provides compelling arguments for its contribution to safe data sharing in sensitive domains, such as healthcare.

Weaknesses:
1. Limited Scope of Evaluation: All experiments are conducted on English text datasets.
While the authors mention that the method could generalize to multilingual datasets, no empirical evidence is provided to support this claim.
2. Synthetic Data Utility: Although the synthetic data quality is competitive, the paper lacks a detailed analysis of failure cases or systematic challenges in achieving higher downstream accuracy beyond current benchmarks, particularly for complex datasets and smaller ϵ values (stronger privacy guarantees).
The study could benefit from a closer examination of diminishing returns for larger synthetic datasets.
3. Privacy Accounting Assumptions: The DP guarantees are suitably formalized ($\epsilon \& \delta$) with RDP accounting, but hyperparameter tuning is excluded from the privacy accounting guarantees.
More rigorous inclusion of tuning steps would strengthen the claims.
4. Reproducibility Limitations: While supplementary details and code frameworks (e.g., t5x) are promised, the reliance on specific public datasets (e.g., The Pile) and pre-trained models underscores potential reproducibility constraints in highly private or resource-constrained settings.
5. Impact of Proxy Metrics: Though the metrics (e.g., MAUVE, perplexity) show high correlation to data utility, the study could deepen its analysis of situations where these metrics fail or become noisy, as seen in some reported variability.

Questions:
1. Expand Dataset Diversity: Consider evaluating the method on multilingual corpora and datasets from diverse domains (e.g., images, tabular data) to strengthen generalization claims.
This would help validate your arguments about its universal applicability.
2. Incorporate Examples of Limitations: Future revisions could include detailed investigations of failure cases or practical limitations (e.g., computational cost of LoRA vs. traditional fine-tuning).
An analysis of lower ϵ thresholds and edge cases, particularly noisy or small datasets, would further solidify the robustness of this approach.
3. Improve Privacy Accounting: Explicitly address how tuning decisions could be integrated into the DP accounting framework, even at additional computational cost.
This would make the methodology more rigorous.

Figure 13: Revised llm review.

RAG-Enhanced LLM-generated Review

Strengths:
1. The paper focuses on an important and underexplored problem: generating high-quality, private synthetic text data using LLMs with DP guarantees. This is timely and relevant.
2. I appreciated the clarity in acknowledging privacy leakage risks from LLM pretraining (e.g., deduplication against datasets like IMDB and Yelp—a step often overlooked in prior work).
3. The proposed use of LoRA-tuning for differential privacy is convincing and novel.
The experimental results show impressive gains—up to 11 percentage points improvement in downstream task accuracy over full fine-tuning under DP.
4. Surprisingly, the synthetic data approach achieves comparable or even superior performance to DP training directly on real data for simpler models (e.g., CNN sentiment classification). This is non-trivial.
5. The ablation studies (Appendix J) on LoRA rank, loss normalization, and batch size are rigorous and provide practical insights for replicating and extending the work.

Weaknesses:
1. I found the experiments somewhat narrow in scope. The focus on English datasets (IMDB, Yelp) limits broader applicability—what about multilingual or domain-specific text like legal/medical data?
2. While it's good that deduplication was conducted, details of the percentage of contamination removed were sparse (Appendix D lacks clarity on quantitative impact).
This feels like a missed opportunity for rigor.
3. The paper claims ``parameter-efficient fine-tuning performs better with DP,'' yet only LoRA was explored in depth—why not Adapter-layer methods or BitFit for comparison?
4. Honestly, the results hinge heavily on having an 8B model for pretraining.
It's unclear how robust this approach would be with smaller, less resource-intensive models. Would it still ``close the utility gap''?
5. Reporting on computational costs was minimal. With extensive pretraining, fine-tuning, and sampling, this approach might not be feasible for researchers with limited compute resources.

Questions:
The paper covers a lot of ground methodologically, but I couldn't help wondering about trade-offs.
For example, while the LoRA-tuning approach is promising, could its scalability for truly massive datasets (e.g., C4-sized data) be addressed more explicitly?
I felt some ablations were incomplete—why not compare to DP implementations of other parameter-efficient setups or distillation approaches?
(To be fair, I can see why LoRA was the go-to choice.)
My primary ask: could the authors provide stronger evidence for generalization to low/medium-resource scenarios, or for tasks beyond sentiment classification?
For instance, can your setup handle retrieval-intensive tasks or adapt to non-English datasets?
Lastly, how certain are you that long-form examples retain fidelity without introducing hallucinations, especially under DP noise constraints?

Figure 14: LLM with RAG.