Can Generative AI Be Helpful For Paper Reviewing Processes?

Anmol Mishra, Isabelle Oktay and Robin Doerfler

Universitat Pompeu Fabra, Barcelona

Abstract. Peer review is essential for upholding high publication standards, yet the rapid growth in research output strains current evaluation systems. Recent advancements in Large Language Models (LLMs) introduce the possibility of LLM-generated reviews to alleviate this burden. This study investigates the potential of generative AI tools to assist peer review, aiming to enhance the process without sacrificing scientific rigor. Utilizing Du et al.'s ReviewCritique dataset—comprising 20 equally distributed rejected and accepted papers with corresponding human and LLM-generated reviews—we evaluate three LLMs: ChatGPT-4o, Claude Opus, and Gemini Pro. Following methodologies proposed by Liang et al. for assessing human-LLM review overlap, we extend these methods to multiple LLMs, testing their reproducibility and overlap with human feedback. Our results diverge from Liang et al., showing that Human-LLM review overlap is less than Human-Human overlap, while LLM-LLM and Human-Human comparisons exhibit similar overlap percentages. This suggests that LLM-generated reviews are consistent within LLMs yet distinct from human feedback, indicating their role as a supportive tool rather than a replacement for human reviewers.

Keywords: Peer Review Automation · Large Language Models (LLMs) · Meta-Research.

1 Introduction

In scientific research, the peer review process is crucial for ensuring high standards of rigor, reliability, and precision. Expert reviews provide authors with constructive feedback, helping to identify flaws, clarify ambiguities, and highlight potential areas for further improvement. For the scientific community, the collective effort of rigorous reviewing ensures that research findings are trustworthy and reliable, thus supporting the continuous advancement of the field and adherence to reporting and reproducibility standards [33] [8] [15].

However, the accelerating rate of research publication presents significant challenges to traditional review mechanisms. With submissions growing steadily, especially in high-impact fields, the demand for timely reviews strains reviewers' capacity and jeopardizes the depth and quality of evaluations [11] [28]. For instance, between 2016 and 2022, the total number of articles indexed from two leading citation databases, Scopus and Web of Science, increased by 47% [10], and according to the Stanford University AI Index Report 2024, the number of AI publications nearly tripled between 2010 and 2022 [3]. Additionally, as the volume of publications increases, it becomes increasingly difficult to identify genuine advances or organize them in a way that benefits the scientific community.

Recent advances in Large Language Models (LLMs) have opened new possibilities for accelerating scientific progress [23], as these models excel at tasks like text generation, summarization, reasoning, and comprehension [26] [30]. Scientists are thus able to use LLM-based AI productivity tools for academic writing and research, further increasing publication output [16] [25]. However, for more complicated tasks like review generations, LLMs do not yet provide a viable substitute for human expertise. This study explores the potential benefits and limitations of applying generative AI tools to support peer review tasks.

While previous research has introduced numerous datasets, generative models, evaluation metrics, and methods for comparing human and AI-generated feedback, reproducibility remains limited. Key studies often lack open-source code, which restricts their applicability and generalisability across different datasets. Thus, the central question — *"Can generative AI, specifically large language models, enhance the paper review process by complementing traditional human reviews, and under what conditions?"* — remains open for exploration.

Given the variety of possible approaches to this question, we adopt a methodology inspired by Liang et al. [20], who proposed techniques for quantifying alignment between human and AI-generated reviews. Their findings suggest that while human expertise remains the foundation of rigorous peer review, LLM-generated feedback may add value. However, the study lacks openly available datasets and code, which restricts reproducibility and cross-dataset evaluation. Moreover, their methodologies are performed using only ChatGPT4 generated reviews, limiting our understanding of how current models perform broadly compared to human beings.

Du et al. [6] address some of these limitations by offering ReviewCritique, an open dataset that includes paper submissions, human expert reviews, and AI-generated reviews from three models: GPT, Claude, and Gemini. They also make parts of their review pipeline accessible, enabling further investigation into AI's role in peer review. However, their methods for assessing consensus between AI-generated and human reviews primarily emphasize the review-generating capabilities of LLMs, rather than examining alignment between AI-generated and human reviews.

Our work seeks to address the current research gap in evaluating existing methods, particularly those proposed by Liang et al. [20], for measuring consensus between AI and human-generated reviews. Specifically, we ask how we can understand and measure consensus not only between human and AI-generated reviews, but also among reviews generated by different LLMs. By implementing a reproducible framework to assess these methodologies on Du et al.'s ReviewCritique, we establish a foundation for evaluating the generalisability of these approaches and further exploring the potential benefits of generative AI in academic peer review.

The remainder of this paper is organized as follows: Section 2 provides an overview of related work, situating our research within the broader context of LLM-generated literature reviews and highlighting relevant studies, including those by Liang et al. [20] and Du et al. [6]. Section 3 details our methodology, including dataset selection, the design of our data extraction pipeline, the application of LLMs for summarization, and the evaluation metrics employed to assess the performance of our generative review framework. Section 4 presents our results, focusing on quantitative analyses and empirical validation. We conclude in Sections 5 and 6 with a discussion of our findings, limitations, and potential directions for improving generative review systems.

2 Related Work

The advancement of LLMs has inspired scientists to explore their potential in research, particularly in generating reviews [4] [22] [6]. Past studies used the "LLM-as-a-judge" approach to assess research ideas [23]. While LLMs are often seen as having higher agreement with human evaluators than with each other, recent findings by Si et al. [32] show that even with advanced prompting, such as chain-of-thought [37], LLMs are not yet reliable for assessing research ideas. Their analysis highlights that LLMs still fall short of human-level agreement, indicating limitations in grasping research complexity and originality, which may also impact their reliability in generating paper reviews.

Prato et al. [29] present a benchmark for assessing LLM capabilities in synthesizing complex knowledge, relevant for understanding their role in review generation. They view LLMs as epistemic models—tools capable of consolidating segmented information—essential for tasks like paper reviewing that require grasping interconnected ideas. However, they also point out that LLMs currently struggle to consistently integrate nuanced information, suggesting that while they could assist by organizing or highlighting details, they are not yet reliable for providing independent, in-depth critiques.

Given these insights, numerous teams have focused on testing and researching the capabilities of LLMs for generating paper reviews. For instance, researchers in natural language processing (NLP) developed the PeerRead dataset [14] to train generative paper review models and investigate the limitations of review automation. Other datasets introduced for AI-generated review research include MOPRD [21] and NLPeer [7]. Like PeerRead, they lack expert annotations on review strengths and weaknesses.

Du et al. addressed the gap in annotated datasets by creating ReviewCritique, which includes both accepted and rejected NLP conference paper submissions, along with corresponding peer reviews and detailed line-by-line annotations by PhD students [6]. While additional annotated datasets exist for evaluating automated academic peer review—focusing on specific aspects such as argument structure [31], politeness [2], and contradictions among reviewers [17]—these are more suitable for exploratory research into particular traits of LLM reviews.

Other inquiries into understanding the potentials and limitations of LLM-generated reviews have also been made. Huotala et al. [12] used ChatGPT-3.5 and ChatGPT-4 to determine if LLM simplified abstracts accelerated the paper review process, discovering that these LLM models were not adequate to increase human screening performance with text simplification alone. However, Liang et al. [20] suggest that LLM-generated and human feedback can complement each other by using ChatGPT-4 to generate comments on full PDFs of scientific papers. Other research has also been done on evaluating the potential for AI to assist in automatically generating acceptance or rejection scores for submitted papers [18] [38] [35], evaluating the quality of human reviews [17], and evaluating and generating meta-reviews [19].

3 Methodology

To assess LLM performance in generating academic reviews, we use ReviewCritique [6], a dataset of 100 NLP papers from OpenReview[1], each with 3–5 line-by-line annotated human reviews and an equal distribution of accepted and rejected submissions. A subset of 20 papers also includes corresponding reviews from ChatGPT-4o [27], Claude 3 Opus [1], and Gemini Pro [34]. We build a pipeline, based on Liang et al. [20], to extract relevant overlaps and validate these findings using metrics like hit rate.

3.1 Dataset Discovery

Our initial consideration of using Music Information Retrieval (MIR) papers from the International Society for Music Information Retrieval (ISMIR) and submissions from Universitat Pompeu Fabra's Research Methods course proved infeasible due to limited access to reviews and lack of author consent, respectively. ReviewCritique [6] was chosen for its comprehensive human annotations and access to both accepted and rejected papers, which supports a more robust assessment of LLM performance across review decisions.

While larger datasets can enhance generalisability, targeted evaluations provide detailed insights, particularly where performance may vary by topic. The dataset's limited size was balanced against budget constraints and the high cost of LLM API usage, enabling us to design an efficient data extraction pipeline and derive insights without overextending resources.

3.2 Data Extraction Pipeline

We adopt the methodology and quantitative metrics from Liang et al. [20], as their approach demonstrated robust results validated by qualitative analysis. The process is described as a two-stage comment matching approach. For both stages, we use prompts provided by Liang et al. [20] to generate outputs. In

[1] https://openreview.net/

stage one, we use extractive text summarization: each LLM and human review is processed by ChatGPT-4o to extract a structured list of key comments from the original text, formatted in JSON. Each comment is identified by a unique ID and includes an AI-generated summary and the original verbatim text.

In stage two, we perform semantic text matching between LLM and human reviews. We create all possible comment pairs to assess alignment among all LLM and human reviews. These pairwise matches are passed to ChatGPT-4o to generate a JSON list of corresponding matched comments, each with a unique ID, a rationale explaining the relation, and a similarity score.

We elected to implement the same similarity score metrics used by Liang et al. [20]. Similarity scores range from 5 to 10, with 5 meaning "Somewhat Related" and 10 meaning "Identical." Similarity scores from 0 through 4 are omitted to narrow the scope of ranking with the assumption that comments about the same paper have a baseline degree of similarity. Our summarised pipeline is shown in Figure 1.

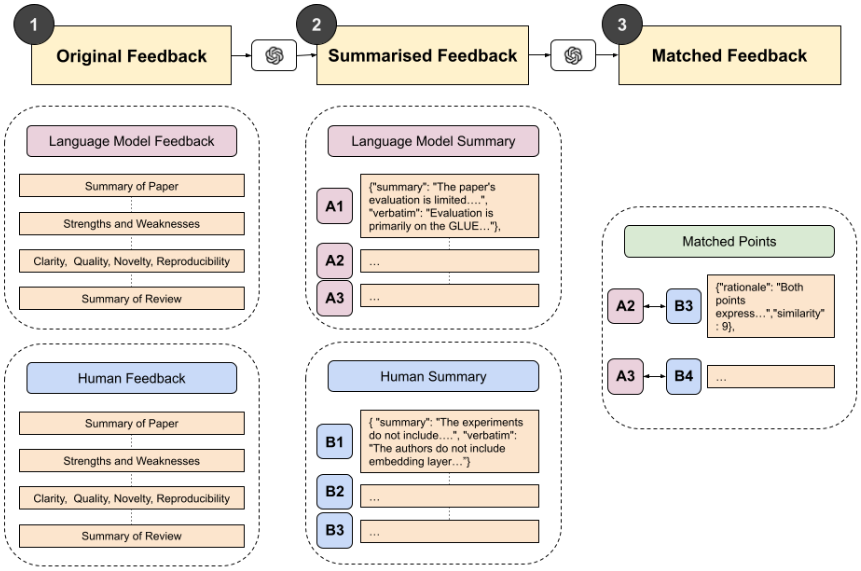

1
Original Feedback
2
Summarised Feedback
3
Matched Feedback
Language Model Feedback
Summary of Paper
Strengths and Weaknesses
Clarity, Quality, Novelty, Reproducibility
Summary of Review
Human Feedback
Summary of Paper
Strengths and Weaknesses
Clarity, Quality, Novelty, Reproducibility
Summary of Review
Language Model Summary
A1
{"summary": "The paper's evaluation is limited....", "verbatim": "Evaluation is primarily on the GLUE..."},
A2
...
A3
...
Human Summary
B1
{ "summary": "The experiments do not include....", "verbatim": "The authors do not include embedding layer..."}
B2
...
B3
...
Matched Points
A2
B3
{"rationale": "Both points express...","similarity" : 9},
A3
B4
...

Fig. 1: Pipeline overview: Original LLM and human reviews are input into ChatGPT-4o with a standardised prompt to produce summarised reviews in JSON format. These summaries are then reprocessed through ChatGPT-4o with a secondary prompt to generate a JSON output identifying points of pairwise overlap.

3.3 LLM for Summarisation

Outsourcing summarisation and similarity estimation to an LLM presents a potential risk of methodological flaw. However, evaluations from Liang et al. [20], based on human-assessed feedback on stage one's summarization outputs, demonstrate the accuracy of this approach, with an F1 score of 0.968, precision of 0.977, and recall of 0.960. The second stage, focused on semantic text matching, also shows high reliability. Human feedback for this stage reveals an 89.% pairwise agreement and an F1 score of 88.7%, supporting the method's robustness. This two-stage process is thus verified with precision and recall metrics and aligns well with our overarching research objective: leveraging generative AI to streamline review generation and evaluation.

Our methodology focuses on the described data extraction and analysis, as our ReviewCritique already contains the LLM-generated reviews. While Liang et al. [20] published code for initial review generation, their extraction and analysis pipeline is not available, requiring us to implement our own. This allows us to not only evaluate their procedures but also expand upon them by analyzing multiple LLMs and their inter-agreements, providing additional reference points for understanding the similarity between human and LLM-generated reviews.

3.4 Evaluation Metrics

After generating pairwise matches, we evaluate overlap across different pairings—LLM vs. LLM, LLM vs. Human, and Human vs. Human—using the hit rate metric. Defined in the original publication, hit rate is the proportion of comments in set A that are also present in set B. This metric effectively indicates the extent to which points raised by an LLM align with those identified by human reviewers, capturing the rate of generated review points that qualify as "hits".

To further examine the robustness of hit rate calculations, we include additional overlap metrics: the Szymkiewicz–Simpson overlap coefficient [24], the Jaccard index [13], and the Sørensen–Dice coefficient [5]. The calculations for these metrics are shown below:

$$\text{Hit Rate} = \frac{|A \cap B|}{|A|} \tag{1}$$

$$\text{Szymkiewicz–Simpson Overlap Coefficient} = \frac{|A \cap B|}{\min(|A|, |B|)} \tag{2}$$

$$\text{Jaccard Index} = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

$$\text{Sørensen–Dice Coefficient} = \frac{2|A \cap B|}{|A| + |B|} \tag{4}$$

To test the null hypothesis, we compare human feedback with LLM feedback by shuffling papers. In this approach, LLM feedback is randomly paired with human feedback from different papers to determine whether LLM responses are overly generalised across publications. By shuffling feedback in this way, we establish a baseline for comparison, helping us assess whether LLM-generated feedback is so generalised that it lacks the specificity needed to produce meaningful reviews for individual papers.

In contrast to the approach used by Liang et al. [20], who verified their findings through qualitative analysis by interviewing the authors of the submitted papers about the helpfulness of the review, we lack the means to interview the authors of the papers in ReviewCritique. Instead, we conducted a small sample-based analysis on our subset of 15 papers. This empirical evaluation aimed to assess the reliability of using GPT-4 for summarisation by comparing model-identified overlaps with human judgment.

4 Results

4.1 Quantitative Assessment

As shown in Figure 2a, the ChatGPT-4 vs. Human overlap has a hit rate of approximately 11.5%, while Human vs. Human overlap reaches 24.3%—a difference of nearly 10%. Additional metrics, including the Szymkiewicz–Simpson Overlap Coefficient, Jaccard Index, and Sørensen–Dice Coefficient, also indicate a lower overlap rate for ChatGPT-4 vs. Human compared to Human vs. Human.

Interestingly, the hit rates for Human vs. Human and LLM vs. LLM overlap are quite similar, at 24.3% and 25.9%, respectively. This suggests that feedback is consistent across different LLM reviews and across different human reviews but not necessarily between pairs of LLM and human reviews. These results underscore the importance of investigating why various LLMs tend to generate feedback patterns that differ noticeably from those produced by humans. Therefore, while Liang et al. [20] suggest that ChatGPT-4 vs. Human overlap is comparable to Human vs. Human overlap, our findings do not fully support this claim.

Specifically, our results suggest that LLMs fall short of fully replicating human reviews, which currently serve as the baseline metric for assessing LLM performance in research paper evaluations. It is possible that LLMs may identify areas for improvement or weaknesses that human reviewers might overlook, positioning LLM feedback as a valuable supplement to human insights. However, until LLMs can reasonably match human performance, this possibility is unlikely to be validated. Aligning with findings in the broader literature, we suggest use of AI-generated reviews as complements to human feedback.

In comparing models for LLM vs. human overlap, Claude Opus achieves the highest hit rate at 15.7%, followed closely by Gemini Pro at 14.0%. ChatGPT-4o demonstrates the lowest hit rate among the LLMs evaluated. This aligns with recent findings suggesting that Claude Opus exhibits stronger rule-following and reasoning capabilities compared to both ChatGPT-4o and Gemini Pro [9]. However, comprehensive comparisons of LLM performance in text summarization remain limited, underscoring the need for further research to understand why different models exhibit varying task performance.

Figure 3a presents an overall comparison between LLM-generated and human expert reviews, showing the average of all metrics across our data subset. Aggregating these metrics provides a more stable and robust assessment, confirming the trends observed in the detailed analysis.

As a baseline check, we perform a comparison on shuffled data, where reviews are drawn from different original papers. This baseline helps verify that the reviews are specific to the content of each paper and not randomly similar. As shown in Figure 3b, nearly all metrics fall below 5%, indicating clear paper-specificity in the initial reviews and confirming that they are not close to random.

4.2 Deviations from Original Methodology

It is worth noting that since our analyses rely on ChatGPT-4o, a newer version than the ChatGPT-4 model used by Liang et al [20], our results may inherently differ due to operational differences between these related models. Future research comparing the review generation capabilities of various ChatGPT versions could provide valuable insights into model performance variations. However, since ChatGPT-4o is the latest model of ChatGPT, it can be assumed that ChatGPT-4o performs at a level equal to or greater than ChatGPT-4.

Given that our dataset is smaller than that of Liang et al. [20], our results may not accurately reflect LLM performance in generating research paper reviews on a larger scale or across diverse disciplines. Additionally, our findings could be biased toward papers in the NLP domain, suggesting that LLM-generated reviews may perform less effectively in other fields. This could be due to the domain-specific knowledge and intuition that human experts possess, which may be more critical in certain areas than in others.

4.3 Empirical Validation

Out of 15 papers, with 15 comparisons (between 3 reviewers and 3 LLMs), resulting in a total of 225 comparisons, we found that in 5 cases, the model used the same point from Review A to identify overlaps with more than one point in Review B. Although this error rate of 2.2% is relatively low, it highlights the potential for such issues to go unnoticed in large-scale analyses, emphasizing the risks of using LLMs in an unsupervised capacity for critical components of a quantitative analysis pipeline. We further discuss this methodological shortcoming in the following section.

5 Discussion

In our investigation, we encountered challenges in using GPT to compare and identify overlapping points between reviews. The following observations address the limitations and potential improvements in the methodology:

5.1 Need for Qualitative Human Analysis

While GPT models are valuable for large-scale analysis, qualitative human analysis remains crucial for accurately verifying overlaps between review points. For example, in the study Exploring Extreme Parameter Compression for Pre-trained Language Models, Reviews 1 and 2 were initially identified as having no overlapping points. However, upon closer inspection, we discovered at least one shared point. This highlights the current limitations of LLMs in matching human-level precision when identifying nuanced overlaps. Therefore, reliance solely on LLMs for assessing review similarity may overlook subtle but significant similarities.

5.2 Potential Prompt Improvements

Our analysis suggests that GPT models may benefit from more explicit instructions in the prompts to avoid redundant matches. For example, explicitly instructing the model to avoid mentioning points that have already been discussed could help prevent inflated overlap scores due to repeated content (e.g., A4-B1, A4-B2, ...), as shown in example 1.1. We found that GPT often struggled to consistently identify matching points, especially when wordings were similar but not identical, resulting in a higher overlap score than warranted. These inconsistencies could be mitigated by refining the prompts, particularly by highlighting the need to detect distinct and unique overlaps, which would improve the model's accuracy across varied datasets. The original prompt is provided in the supplementary material 1.2.

6 Conclusions

This study evaluates current generative AI tools for automatically generating research paper reviews, addressing the need for automated review systems as publication volumes rise. Leveraging LLM advancements in writing, summarization, and comprehension, we assess their performance relative to human reviewers using a pipeline inspired by Liang et al. on a subset of ReviewCritique, containing papers and corresponding human and LLM-generated reviews (ChatGPT-4o, Claude Opus, Gemini Pro).

Our findings show that LLM reviews are not completely random. However, while they have some overlap with human reviewers, they cannot yet fully replace human reviewers as they often fail to identify overlaps evident to humans. This discrepancy is likely due to limitations in asking relevant questions and assessing quality. These results do emphasize the need to critically evaluate reproducibility in review generation research that claims otherwise, as current LLMs are not yet viable substitutes for human review.

Future research should focus on fine-tuning models specifically for review generation, understanding why LLMs struggle with quality evaluation and questioning, and improving the generalisability and reproducibility of review generation pipelines. At the same time, we must consider the ethics of LLM-generated review systems, for example, in data privacy and bias mitigation [36]. These efforts will be essential in creating more reliable and transparent models for automated review generation.

Code and Data Availability

The code for our extraction and analysis pipeline implementation is available at: https://github.com/dhunstack/llms-for-paper-review.

References

1. The claude 3 model family: Opus, sonnet, haiku (2024), https://api.semanticscholar.org/CorpusID:268232499
2. Bharti, P., Navlakha, M., Agarwal, M., Ekbal, A.: Politepeer: does peer review hurt? a dataset to gauge politeness intensity in the peer reviews. Language Resources and Evaluation **58**, 1291–1313 (05 2023). https://doi.org/10.1007/s10579-023-09662-3
3. Clark, J., Wald, R., Shoham, Y., Carlos Niebles, J., Manyika, J., Lyons, T., Ligett, K., Etchemendy, J., Brynjolfsson, E., Reuel, A., et al.: The AI Index 2024 Annual Report. AI Index Steering Committee (2024)
4. Deng, Z., Peng, H., Xia, C., Li, J., He, L., Yu, P.: Hierarchical bi-directional self-attention networks for paper review rating recommendation. In: Scott, D., Bel, N., Zong, C. (eds.) Proceedings of the 28th International Conference on Computational Linguistics. pp. 6302–6314. International Committee on Computational Linguistics, Barcelona, Spain (Online) (Dec 2020). https://doi.org/10.18653/v1/2020.coling-main.555, https://aclanthology.org/2020.coling-main.555
5. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
6. Du, J., Wang, Y., Zhao, W., Deng, Z., Liu, S., Lou, R., Zou, H.P., Venkit, P.N., Zhang, N., Srinath, M., Zhang, H.R., Gupta, V., Li, Y., Li, T., Wang, F., Liu, Q., Liu, T., Gao, P., Xia, C., Xing, C., Cheng, J., Wang, Z., Su, Y., Shah, R.S., Guo, R., Gu, J., Li, H., Wei, K., Wang, Z., Cheng, L., Ranathunga, S., Fang, M., Fu, J., Liu, F., Huang, R., Blanco, E., Cao, Y., Zhang, R., Yu, P.S., Yin, W.: LLMs Assist NLP Researchers: Critique Paper (Meta-)Reviewing (Oct 2024). https://doi.org/10.48550/arXiv.2406.16253, http://arxiv.org/abs/2406.16253, arXiv:2406.16253
7. Dycke, N., Kuznetsov, I., Gurevych, I.: NLPeer: A unified resource for the computational study of peer review. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 5049–5073. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.277, https://aclanthology.org/2023.acl-long.277
8. Gannon, F.: The essential role of peer review (2001). https://doi.org/10.1093/embo-reports/kve188
9. Han, Z., Battaglia, F., Mansuria, K., Heyman, Y., Terlecky, S.R.: (Oct 2024). https://doi.org/10.21203/rs.3.rs-5084169/v1
10. Hanson, M.A., Barreiro, P.G., Crosetto, P., Brockington, D.: The strain on scientific publishing. Quantitative Science Studies pp. 1–21 (10 2024). https://doi.org/10.1162/qss_a_00327, https://doi.org/10.1162/qss_a_00327

11. Huisman, J., Smits, J.: Duration and quality of the peer review process: the author's perspective. Scientometrics **113**, 633–650 (10 2017). https://doi.org/10.1007/s11192-017-2310-5
12. Huotala, A., Kuutila, M., Ralph, P., Mäntylä, M.: The promise and challenges of using llms to accelerate the screening process of systematic reviews (2024), https://arxiv.org/abs/2404.15667
13. Jaccard, P.: The distribution of the flora in the alpine zone. 1. New phytologist **11**(2), 37–50 (1912)
14. Kang, D., Ammar, W., Dalvi, B., van Zuylen, M., Kohlmeier, S., Hovy, E., Schwartz, R.: A Dataset of Peer Reviews (PeerRead): Collection, Insights and NLP Applications. In: Walker, M., Ji, H., Stent, A. (eds.) Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). pp. 1647–1661. Association for Computational Linguistics, New Orleans, Louisiana (Jun 2018). https://doi.org/10.18653/v1/N18-1149, https://aclanthology.org/N18-1149
15. Kelly, J., Sadeghieh, T., Adeli, K., Biochemistry, C.: Peer review in scientific publications: benefits, critiques, a survival guide (10 2014)
16. Khalifa, M., Albadawy, M.: Using artificial intelligence in academic writing and research: An essential productivity tool. Computer Methods and Programs in Biomedicine Update **5**, 100145 (2024). https://doi.org/https://doi.org/10.1016/j.cmpbup.2024.100145, https://www.sciencedirect.com/science/article/pii/S2666990024000120
17. Kumar, S., Ghosal, T., Ekbal, A.: When reviewers lock horns: Finding disagreements in scientific peer reviews. In: Bouamor, H., Pino, J., Bali, K. (eds.) Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 16693–16704. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.emnlp-main.1038, https://aclanthology.org/2023.emnlp-main.1038
18. Li, J., Sato, A., Shimura, K., Fukumoto, F.: Multi-task peer-review score prediction. In: Chandrasekaran, M.K., de Waard, A., Feigenblat, G., Freitag, D., Ghosal, T., Hovy, E., Knoth, P., Konopnicki, D., Mayr, P., Patton, R.M., Shmueli-Scheuer, M. (eds.) Proceedings of the First Workshop on Scholarly Document Processing. pp. 121–126. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.sdp-1.14, https://aclanthology.org/2020.sdp-1.14
19. Li, M., Hovy, E., Lau, J.: Summarizing multiple documents with conversational structure for meta-review generation. In: Bouamor, H., Pino, J., Bali, K. (eds.) Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 7089–7112. Association for Computational Linguistics, Singapore (Dec 2023). https://doi.org/10.18653/v1/2023.findings-emnlp.472, https://aclanthology.org/2023.findings-emnlp.472
20. Liang, W., Zhang, Y., Cao, H., Wang, B., Ding, D., Yang, X., Vodrahalli, K., He, S., Smith, D., Yin, Y., McFarland, D., Zou, J.: Can large language models provide useful feedback on research papers? a large-scale empirical analysis (10 2023), http://arxiv.org/abs/2310.01783
21. Lin, J., Song, J., Zhou, Z., Chen, Y., Shi, X.: Moprd: A multidisciplinary open peer review dataset. Neural Comput. Appl. **35**(34), 24191–24206 (Sep 2023). https://doi.org/10.1007/s00521-023-08891-5, https://doi.org/10.1007/s00521-023-08891-5

22. Liu, R., Shah, N.B.: Reviewergpt? an exploratory study on using large language models for paper reviewing (2023), https://arxiv.org/abs/2306.00622
23. Lu, C., Lu, C., Lange, R.T., Foerster, J., Clune, J., Ha, D.: The ai scientist: Towards fully automated open-ended scientific discovery (8 2024), http://arxiv.org/abs/2408.06292
24. McGill, M.: An evaluation of factors affecting document ranking by information retrieval systems. (1979)
25. Meyer, J.G., Urbanowicz, R.J., Martin, P.C., O'Connor, K., Li, R., Peng, P.C., Bright, T.J., Tatonetti, N., Won, K.J., Gonzalez-Hernandez, G., et al.: Chatgpt and large language models in academia: Opportunities and challenges. BioData Mining **16**(1) (Jul 2023). https://doi.org/10.1186/s13040-023-00339-9
26. Naveed, H., Khan, A.U., Qiu, S., Saqib, M., Anwar, S., Usman, M., Akhtar, N., Barnes, N., Mian, A.: A comprehensive overview of large language models (2024), https://arxiv.org/abs/2307.06435
27. OpenAI: Gpt-4 technical report (2024), https://arxiv.org/abs/2303.08774
28. Powell, K.: Does it take too long to publish research? Nature 530 pp. 148–151 (2 2016). https://doi.org/https://doi.org/10.1038/530148a, https://doi.org/10.1038/530148a
29. Prato, G., Huang, J., Parthasarathi, P., Sodhani, S., Chandar, S.: Epik-eval: Evaluation for language models as epistemic models
30. Rasool, Z., Kurniawan, S., Balugo, S., Barnett, S., Vasa, R., Chesser, C., Hampstead, B.M., Belleville, S., Mouzakis, K., Bahar-Fuchs, A.: Evaluating llms on document-based qa: Exact answer selection and numerical extraction using cogtale dataset. Natural Language Processing Journal **8**, 100083 (2024). https://doi.org/https://doi.org/10.1016/j.nlp.2024.100083, https://www.sciencedirect.com/science/article/pii/S2949719124000311
31. Ruggeri, F., Mesgar, M., Gurevych, I.: A dataset of argumentative dialogues on scientific papers. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). pp. 7684–7699. Association for Computational Linguistics, Toronto, Canada (Jul 2023). https://doi.org/10.18653/v1/2023.acl-long.425, https://aclanthology.org/2023.acl-long.425
32. Si, C., Yang, D., Hashimoto, T.: Can llms generate novel research ideas? a large-scale human study with 100+ nlp researchers. arXiv preprint (9 2024), http://arxiv.org/abs/2409.04109
33. Steer, P.J., Ernst, S.: Peer review - why, when and how. International Journal of Cardiology Congenital Heart Disease **2**, 100083 (2 2021). https://doi.org/10.1016/j.ijcchd.2021.100083
34. Team, G.: Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context (2024), https://arxiv.org/abs/2403.05530
35. Wang, Q., Zeng, Q., Huang, L., Knight, K., Ji, H., Rajani, N.F.: ReviewRobot: Explainable paper review generation based on knowledge synthesis. In: Davis, B., Graham, Y., Kelleher, J., Sripada, Y. (eds.) Proceedings of the 13th International Conference on Natural Language Generation. pp. 384–397. Association for Computational Linguistics, Dublin, Ireland (Dec 2020). https://doi.org/10.18653/v1/2020.inlg-1.44, https://aclanthology.org/2020.inlg-1.44
36. Watkins, R.: Guidance for researchers and peer-reviewers on the ethical use of large language models (llms) in scientific research workflows. AI and Ethics **4**(4), 969–974 (May 2023). https://doi.org/10.1007/s43681-023-00294-5

37. Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D.: Chain-of-thought prompting elicits reasoning in large language models. NeurIPS (1 2022), http://arxiv.org/abs/2201.11903
38. Zhou, R., Chen, L., Yu, K.: Is LLM a reliable reviewer? a comprehensive evaluation of LLM on automatic paper reviewing tasks. In: Calzolari, N., Kan, M.Y., Hoste, V., Lenci, A., Sakti, S., Xue, N. (eds.) Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024). pp. 9340–9351. ELRA and ICCL, Torino, Italia (May 2024), https://aclanthology.org/2024.lrec-main.816

Appendix

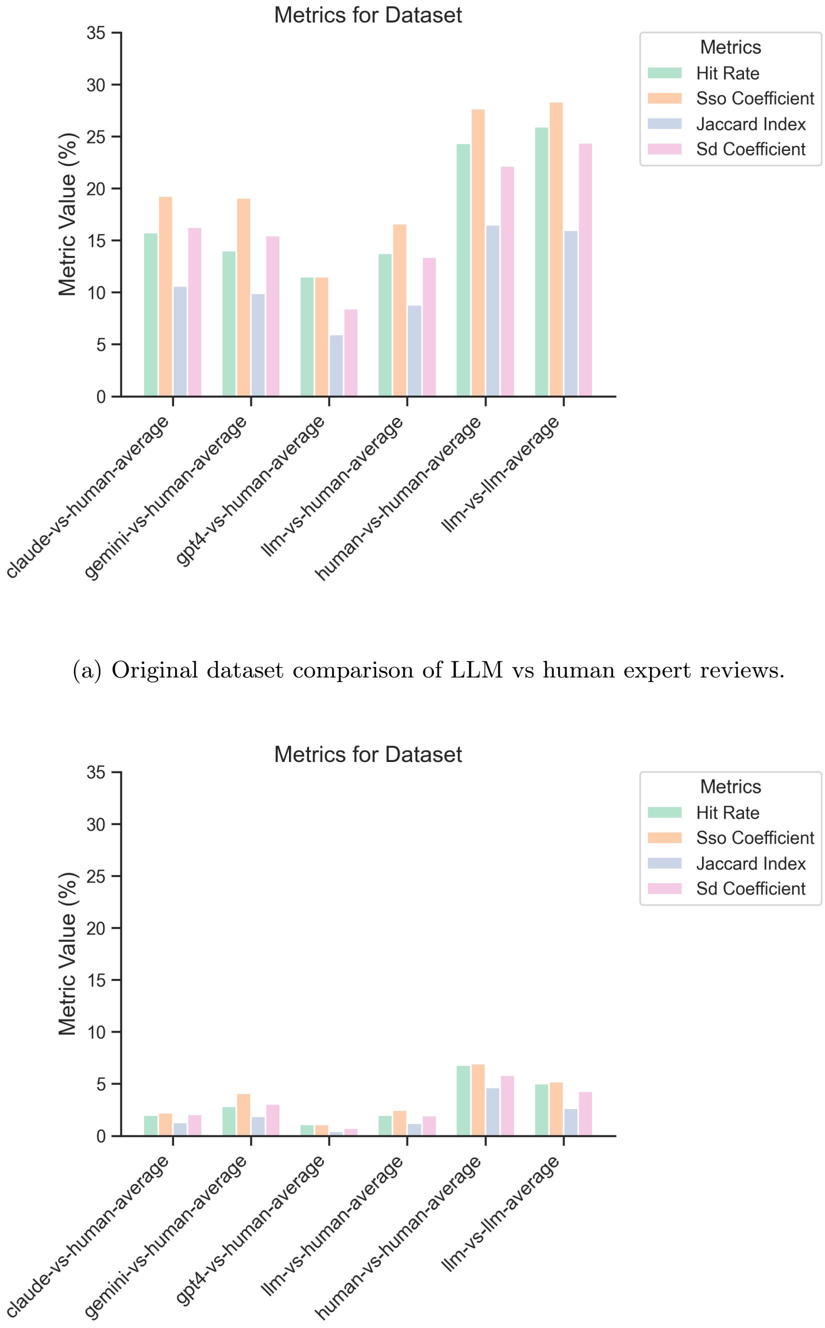
Listing 1.1: An example of a flaw in GPT-4o's summarization comparison with two human reviewers: GPT-4o counts a single point made by Reviewer A three times because it perceives it as highly similar to three distinct points raised by Reviewer B, all of which critique novelty from slightly different perspectives.

```
{
    "A4-B1": {
        "rationale": "A4 mentions the novelty is incremental
            and lacks significant improvement, while B1
            states the proposed method is not novel and
            discusses leveraging vocabulary distribution,
            highlighting a common lack of novelty.",
        "similarity": 8
    },
    "A4-B2": {
        "rationale": "A4 and B2 both discuss the novelty of
            the work, with A4 mentioning that the method is
            not significantly innovative, and B2 stating that
             the method is similar to existing works,
            limiting its novelty.",
        "similarity": 8
    },
    "A4-B3": {
        "rationale": "Both A4 and B3 highlight the limited
            novelty of the paper, with A4 stating the method
            lacks significant improvements, and B3 explicitly
             saying the overall novelty is limited.",
        "similarity": 7
    }
}
```

Listing 1.2: The prompt from Liang et al.'s reference paper, used to generate the review comparison by GPT-4o.

```
Your task is to carefully analyze and accurately match the
    key concerns raised in two reviews, ensuring a strong
    correspondence between the matched points. Examine the
    verbatim closely.
```

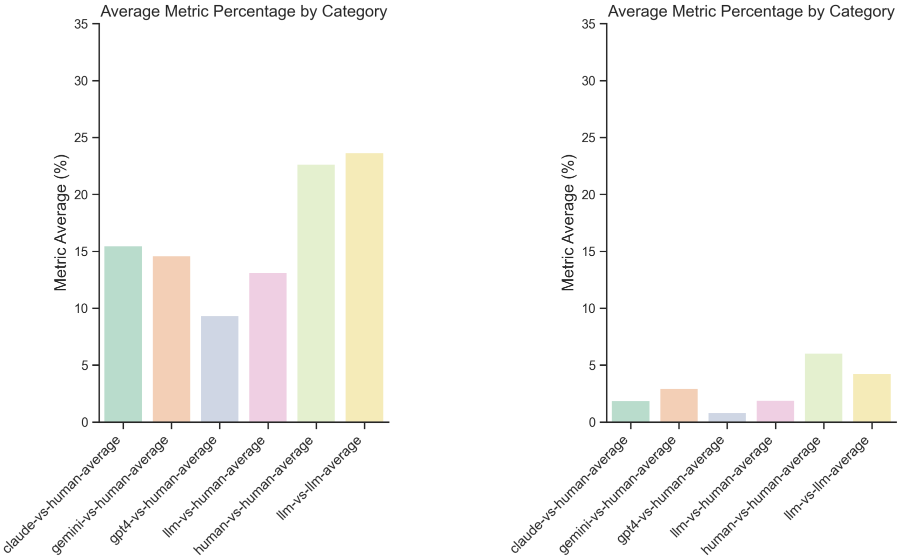
```
=====Review A:

<JSON extracted comments for Review A from previous step>

=====Review B:
<JSON extracted comments for Review B from previous step>

Please follow the example JSON format below for matching
    points. For instance, if point 1 from review A is nearly
    identical to point 2 from review B, it should look like
    this:

{{
    "A1-B2": {{"rationale": "<explain why A1 and B2 are
        nearly identical>", "similarity": "<5-10, only an
        integer>"}},
    ...
}}


Note that you should only match points with a significant
    degree of similarity in their concerns. Refrain from
    matching points with only superficial similarities or
    weak connections. For each matched pair, rate the
    similarity on a scale of 5-10.

5. Somewhat Related: Points address similar themes but from
    different angles.
6. Moderately Related: Points share a common theme but with
    different perspectives or suggestions.
7. Strongly Related: Points are largely aligned but differ in
     some details or nuances.
8. Very Strongly Related: Points offer similar suggestions or
     concerns, with slight differences.
9. Almost Identical: Points are nearly the same, with minor
    differences in wording or presentation.
10. Identical: Points are exactly the same in terms of
    concerns, suggestions, or praises.


If no match is found, output an empty JSON object. Provide
    your output as JSON only.
```



(a) Original dataset comparison of LLM vs human expert reviews.

(b) Shuffled dataset baseline comparison.

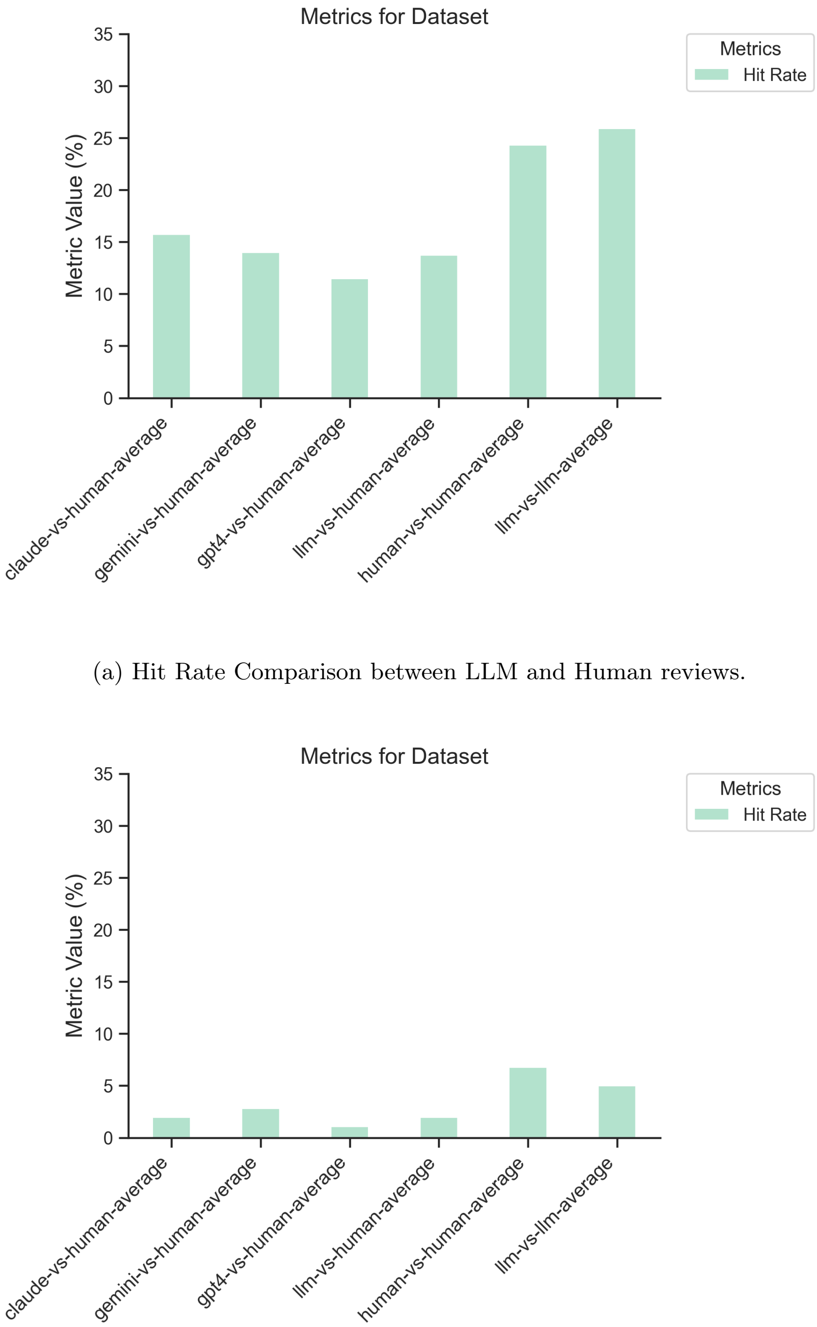
Fig. 2: Comparison of LLM vs human expert reviews on multiple metrics (Hit Rate, Szymkiewicz–Simpson Overlap Coefficient, Jaccard Index, and Sørensen–Dice Coefficient) for (a) the original dataset and (b) the shuffled dataset as a baseline.



(a) Overall comparison of LLM vs human expert reviews as averages of all metrics.

(b) Baseline Check: Overall comparison of LLM vs human expert reviews on shuffled data.

Fig. 3: Comparison of LLM vs human expert reviews, with (a) original data and (b) shuffled data for baseline validation.



(a) Hit Rate Comparison between LLM and Human reviews.

(b) Baseline Check: Hit Rate Comparison between LLM and Human reviews on shuffled data.

Fig. 4: Comparison of Hit Rates of LLM vs human expert reviews, with (a) original data and (b) shuffled data for baseline validation.