Are Large Language Models Better Peer-Reviewers Than Humans? An Early Investigation on OpenReview

Gianluca Bonifazi[1,*,†], Christopher Buratti[1,†], Michele Marchetti[1,†], Federica Parlapiano[1,†], Davide Traini[1,2,†], Domenico Ursino[1,†] and Luca Virgili[1,†]

[1]DII, Polytechnic University of Marche, Italy
[2]CHIMOMO, University of Modena and Reggio Emilia, Italy

Abstract

In recent years, Large Language Models (LLMs) have often been used by paper reviewers, despite this practice being generally prohibited. This has raised, and continues to raise, issues concerning ethics, review reliability, and the risk of review manipulation. Indeed, several arXiv preprints were recently discovered to contain invisible, LLM-targeted instructions designed to persuade an AI reviewer to yield a positive review. In this paper, we propose a systematic analysis of LLMs' review capabilities in this complex and evolving scenario. In particular, we want to address two research questions: *(i)* How can LLM ratings be compared with human ratings?, and *(ii)* Can hidden positive prompts injected in a manuscript alter an LLM's generated review? To address these questions, we created a dataset of 400 papers from OpenReview. For each paper, this dataset contains human reviews and scores already present in OpenReview, as well as reviews performed by three state-of-the-art LLMs, added by us. Our results show that human reviewers assign higher and more widely dispersed scores that clearly distinguish accepted and rejected papers. In contrast, LLM ratings cluster close to their mean value, blurring the distinction between accepted and rejected papers. Furthermore, a negative prompt given by the reviewer makes the LLM lower its scores, while a hidden positive prompt injected by the author often fails to raise scores, and sometimes triggers even lower scores, if detected by the LLM. These results reveal both the potential and fragility of delegating peer review tasks to LLMs.

Keywords

Generative Artificial Intelligence, Large Language Model, Peer Review, Prompt Injection, OpenReview

1. Introduction

In recent years, Generative Artificial Intelligence (GenAI) and, in particular, Large Language Models (LLMs) have begun reshaping both everyday life and professional practice. These systems can now tackle a wide range of complex tasks. From personalized tutoring to decision support in healthcare [1, 2, 3], their rapid spread is opening a lot of new possibilities while forcing researchers and practitioners to reconsider established assumptions. Academia has likewise felt the impact of LLMs. In fact, researchers use these tools at multiple stages of the research process, from drafting manuscripts to polishing prose and checking references [4, 5]. While this can unlock new opportunities, it also introduces new risks. For instance, the authors of [6] asked GPT to write abstracts given a title and a target journal. They demonstrated that GPT can produce scientifically credible abstracts that, however, contained invented data. Another emerging issue is the use of LLMs to review scientific papers. For instance, the authors of [7] compared human reviews with GPT-generated reviews for a machine learning conference. They found that, while GPT can deliver reasonably high-quality feedback, important shortcomings remain. The authors of [8] highlight GPT's potential in the review process when evaluating language, enabling reviewers to focus on content. However, they also acknowledge the risk of generating inaccurate, irrelevant, or useless comments. Finally, the authors of [9] state that GPT cannot replace human reviewers, since they did not find a significant overlap between human and GPT reviews.

ITADATA-WS 2025: The 4th Italian Conference on Big Data and Data Science – Workshops, September 9–11, 2025, Turin, Italy
*Corresponding author.
†These authors contributed equally.
g.bonifazi@univpm.it (G. Bonifazi); c.buratti@pm.univpm.it (C. Buratti); michele.marchetti@univpm.it (M. Marchetti); f.parlapiano@pm.univpm.it (F. Parlapiano); davide.traini@unimore.it (D. Traini); d.ursino@univpm.it (D. Ursino); luca.virgili@univpm.it (L. Virgili)

Despite the well documented problems associated with using LLMs for paper review, these tools are being employed more frequently, even though this practice is generally prohibited. This raises issues involving ethics, review reliability, and vulnerability to manipulation. For instance, at least 17 arXiv preprints were recently found to contain invisible, LLM-targeted instructions designed to persuade AI reviewers to issue favorable reviews[1]. These instructions included hidden "accept" prompts in manuscripts to ensure higher scores. This episode shows that some authors are already exploiting the fact that some reviewers delegate their task to LLMs.

Our paper is motivated by the examination of this scenario and aims to contribute to the investigation of this phenomenon. Specifically, it aims to address two research questions. The first (RQ1) asks how closely scores generated by state-of-the-art LLM reviewers align with those of human reviewers. The second (RQ2) tests whether a hidden prompt injection embedded by the author or an explicit negative prompt provided by a reviewer can bias an LLM when it reviews a paper.

To address these research questions, we built a dataset of 400 papers from OpenReview[2], a public peer review platform used by top venues. We first identified A* conferences whose main-track submissions were accompanied by publicly available peer reviews. These conferences covered the period from 2021 to 2024, spanning the years immediately before and after the release of GPT. For each paper, we collected the PDF file, the complete set of human reviews, and the corresponding scores. In this way, we obtained matched text-and-rating data across the entire acceptance spectrum. Next, we obtained three reviews from LLMs for each paper using the same template employed by humans. For this purpose, we selected three widely adopted LLMs, i.e., GPT-4o mini, Gemini 1.5 Flash, and Gemini 2.0 Flash. To answer RQ1, we compared the scores returned by the LLMs with those returned by humans. To answer RQ2, we asked the three LLMs to review the papers again after providing them with a hidden positive author prompt and/or a negative reviewer prompt.

The main results we obtained are the following:

- Human reviewers provide higher and more dispersed scores than LLMs, and their ratings more clearly distinguish accepted papers from rejected ones. In contrast, LLM scores are very close to the mean, which blurs this distinction.
- A negative reviewer prompt generally pushes each model toward lower overall ratings.
- A hidden "accept" author prompt is only effective with certain models. Specifically, GPT-4o mini is particularly susceptible, while Gemini 2.0 Flash, and partially Gemini 1.5 Flash, resist manipulation. Interestingly, when the LLM recognized the injection, it penalizes the corresponding paper.

The rest of this paper is organized as follows: Section 2 describes the methodology used to address the research questions. Section 3 presents the empirical results. Finally, Section 4 draws some conclusions and highlights some possible future developments of our work.

2. Methodology

In this section, we describe the methodology used to answer the two research questions of interest for this paper. In particular, Section 2.1 details the construction of our dataset. Section 2.2 outlines the methodology used to address RQ1. Finally, Section 2.3 illustrates the procedure employed to address RQ2.

2.1. Dataset

To answer our research questions, we built a dataset based on OpenReview. This is an open source platform that supports transparent scholarly peer review. It makes key steps of the review process, such as referee reports, author rebuttals, and community comments, publicly accessible under fine-grained

[1] https://asia.nikkei.com/Business/Technology/Artificial-intelligence/Positive-review-only-Researchers-hide-AI-prompts-in-papers

[2] https://openreview.net/

access controls. Top conferences, such as the International Conference on Learning Representations (ICLR), the Annual Conference on Neural Information Processing Systems (NeurIPS), and the International Conference on Empirical Methods in Natural Language Processing (EMNLP), rely on this platform to manage double-blind or open-identity reviews, facilitating threaded discussions and real time review tracking.

To build our dataset, we first identified some major conferences whose main-track submissions were accompanied by publicly available peer reviews in the period 2021 - 2024, i.e., in the years immediately before and after the release of GPT. Specifically, we focused on ICLR and NeurIPS, as they met this criterion. For each conference-year pair, we randomly selected 25 accepted and 25 rejected papers, yielding 50 papers per pair and 400 papers in total. Including both accepted and rejected papers allowed us to capture the entire spectrum of review scores, from the low scores typically assigned to rejected papers to the high scores generally assigned to the accepted ones. Additionally, we downloaded the PDF file, the complete set of human reviews, and the post-rebuttal scores for each paper. This ensures that all the scores we analyzed align with the exact PDF version archived on OpenReview. Finally, we took the mean of the scores provided by human reviewers.

2.2. RQ1: Comparing Human and LLM Reviews

To answer the first research question, we asked three models, namely GPT-4o mini, Gemini 1.5 Flash, and Gemini 2.0 Flash, to review each paper in the dataset. To this end, we used a prompt that instructed the LLM to act as a rigorous A* conference reviewer. The prompt also instructed it to follow a fixed review template consisting of: *(i)* a summary; *(ii)* a score from 1 to 4 for soundness, presentation, and contribution; *(iii)* a list of strengths and weaknesses; *(iv)* an overall score from 1 to 10; *(v)* the LLM's confidence in the topic of the paper to be reviewed. Then, we instructed the LLM to apply an acceptance rate in line with that of the reference conference, and to directly reject papers that were not technically sound, were poorly presented, or lacking in substantial and original contribution. For each combination of conference, year, and model (or human), we computed the mean, standard deviation, and skewness of the overall scores. These descriptive statistics capture the central tendency, dispersion, and asymmetry in the review scores provided by humans and LLMs, making cross-year and cross-conference comparisons straightforward.

We applied two non-parametric tests suitable for ordinal and non-normally distributed ratings. Specifically, we used:

- The Wilcoxon signed-rank test [10] to compare the score distributions returned by human and LLM reviewers. In particular, the null hypothesis of this test is that the two distributions are equal; as usual, the null hypothesis is rejected if the corresponding p-value is less than 0.05.
- The Mann–Whitney U test [11] to examine the extent to which the scores assigned to accepted and rejected papers differed for each conference-year pair. In particular, the null hypothesis of this test is that the scores assigned to accepted and rejected papers are equal; if the corresponding p-value is less than 0.05 the null hypothesis is rejected.

2.3. RQ2: Analyzing Prompt Injection and Reviewer Coercions in LLM Reviews

To answer the second research question, we first created a second version of the PDF file of each paper. In each page of this file, we embedded a multi-sentence instruction encouraging strong acceptance of the paper. We made the text white with six-point font so that it would remain invisible to human readers while still being parseable by LLMs. This strategy is similar to the one observed in 17 arXiv papers mentioned in the Introduction.

We then asked the LLMs to review the papers under four settings, namely:

1. *Original*: The LLM received the original PDF file of the paper, and the review request did not include any forcing.

2. *Injected*: The LLM received the manipulated PDF file of the paper with the hidden positive author prompt, but the review request prompt remained neutral.

3. *Negative*: The LLM received the original PDF file of the paper, but the reviewer provided the LLM with a prompt requesting it to recommend rejection and assign a low overall score.

4. *Negative Injected*: This setting involved the use of the modified PDF file, as in the second setting, and the prompt requesting rejection, as in the third setting.

We summarized the overall ratings for each conference, year, model, and setting by calculating the mean, standard deviation, and skewness. To determine whether the hidden positive author prompts and/or the negative reviewer prompts altered the scoring behavior, we performed the Wilcoxon signed-rank test separately for each conference-reviewer pair.

3. Results

This section presents the results of our study. Specifically, Section 3.1 details the results for RQ1 and Section 3.2 reports those for RQ2.

3.1. RQ1: Comparing Human and LLM Reviews

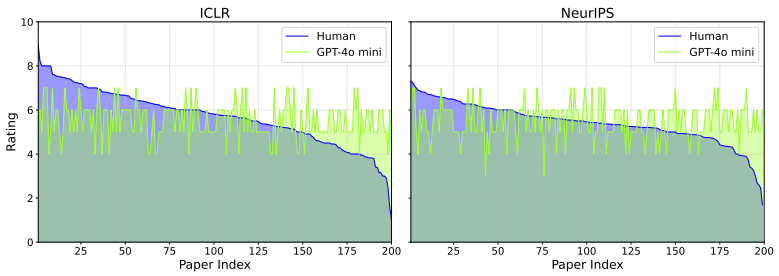
First, we computed the distribution of the paper scores, grouping the results by conference, in order to quantify the difference in scores between human and LLM reviewers. Figure 1 shows the area charts of the paper scores. Papers are divided by conference: 200 relate to ICLR and 200 to NeurIPS.

From the analysis of this figure, we can see that human reviewers tend to use a wider range of scores (from 1 to 9) for papers, while LLMs tend to assign ratings within a more limited range (from 3 to 7). To confirm this first insight, we computed the mean, standard deviation, and skewness of the score distributions for human and LLM reviewers. The results are shown in Table 1.

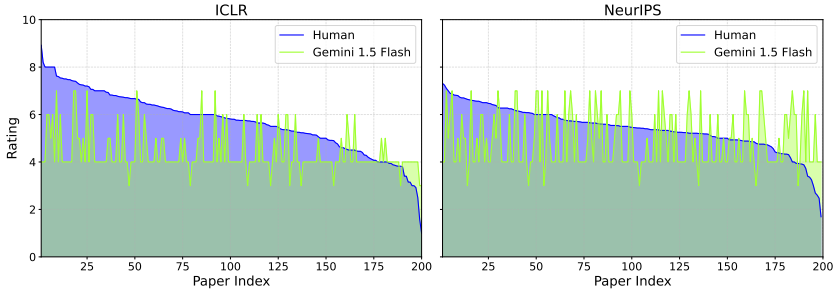
Conference	Reviewer	Mean	Std. Deviation	Skewness
ICLR	Human	5.70	1.33	-0.45
	GPT-4o mini	5.53	0.83	-0.03
	Gemini 1.5 Flash	4.38	0.87	1.44
	Gemini 2.0 Flash	4.50	1.11	0.32
NeurIPS	Human	5.42	0.97	-0.67
	GPT-4o mini	5.50	0.83	-0.28
	Gemini 1.5 Flash	4.87	1.17	0.64
	Gemini 2.0 Flash	4.88	1.01	-0.04

Table 1
Descriptive statistics of the paper scores yielded by human and LLM reviewers reported separately for ICLR and NeurIPS. The table shows the mean, standard deviation, and skewness of the distribution for each conference-reviewer pair

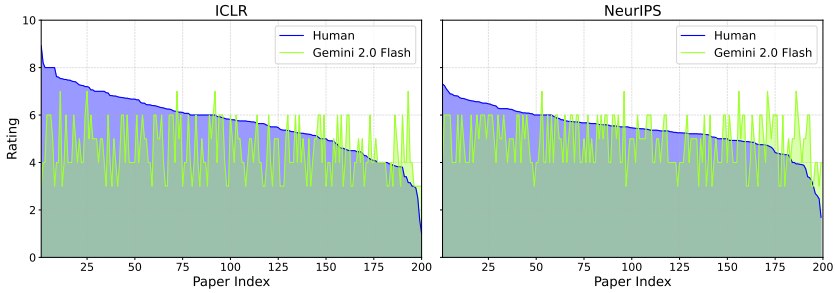
Table 1 reveals that human reviewers tend to give higher scores than LLM reviewers at both conferences, with only a marginal exception at NeurIPS, where GPT-4o mini surpasses human reviewers by a small amount. Human scores also have the lowest negative skewness. GPT-4o mini is the most consistent model, as indicated by its minimum standard deviation, which suggests that scores are tightly clustered around the mean. Gemini 1.5 Flash is the most critical model because it has the lowest mean and the strongest positive skewness. These results stem from numerous low scores and a few high scores. Gemini 2.0 Flash is also critical because its scores have a low mean; however, the distribution of its scores is bell-shaped, as evidenced by its skewness close to 0.



(a) Human vs. GPT-4o mini



(b) Human vs. Gemini 1.5 Flash



(c) Human vs. Gemini 2.0 Flash

Figure 1: Area charts showing the paper scores yielded by human and LLM reviewers

We then performed a two-sided Wilcoxon rank-sum test to compare the score distributions assigned by human and LLM reviewers. We performed this comparison separately for each conference and for each LLM. The results are shown in Table 2.

Conference	Reviewer	Statistic	p-value
ICLR	GPT-4o mini	7,736.50	4.76×10^{-2}
	Gemini 1.5 Flash	1,535.00	1.80×10^{-24}
	Gemini 2.0 Flash	2,276.00	2.08×10^{-19}
NeurIPS	GPT-4o mini	8,653.00	4.28×10^{-1}
	Gemini 1.5 Flash	5,670.00	2.24×10^{-7}
	Gemini 2.0 Flash	5,026.50	5.96×10^{-9}

Table 2
Results of the Wilcoxon signed-rank test comparing the score distributions of human and LLM reviews

As shown in Table 2, the score distributions returned by LLMs are statistically different from those returned by humans in almost all cases, as indicated by p-values less than 0.05. The only exception is GPT-4o mini in NeurIPS. A p-value close to 0.05 was also found for the same model in ICLR. Therefore, GPT-4o mini is the LLM that provides the most human-like evaluations. Cross-referencing these data

with those in Table 1 reveals that Gemini 1.5 Flash and Gemini 2.0 Flash provide evaluations that differ significantly from human ones. The scores they assign are significantly lower than those provided by humans. Instead, in the case of GPT-4o mini, the scores returned by the LLM are close to those returned by humans. In particular, the mean scores are slightly lower for ICLR papers and slightly higher for NeurIPS papers.

We then refined the analysis by separating accepted papers from rejected ones to see if the score patterns differed between the two groups. Accepted papers should have received higher scores, while rejected papers should have received lower scores. To this end, we computed the violin plots of the scores returned by humans and each LLM, distinguishing between accepted and rejected papers. The corresponding results are illustrated in Figure 2.

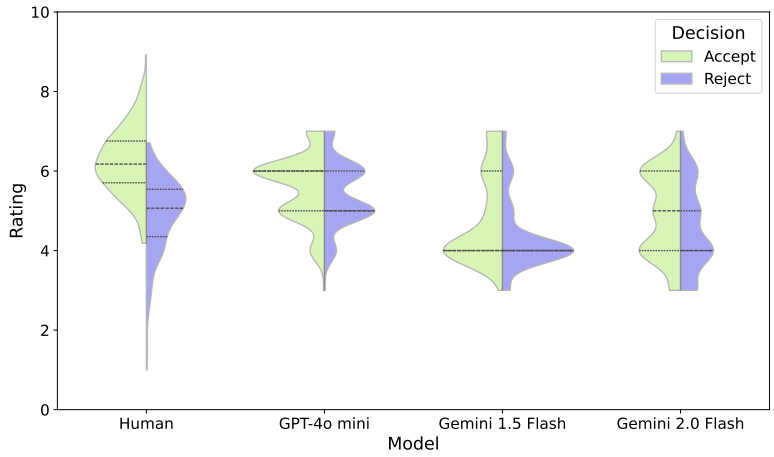


Figure 2: Violin plots of human and LLM reviews considering both accepted and rejected papers

As this figure shows, there are significant differences in human scores between accepted and rejected papers. In contrast, the LLMs' score distributions overlap largely and cover nearly identical ranges.

To verify whether the gap of the scores between accepted and rejected papers was statistically significant, we compared the score distributions of accepted and rejected papers for each conference-year pair. Since the two samples were non-overlapping, we used the Mann–Whitney U test. The results are presented in Table 3.

From the analysis of this table, we observe that, in all cases, the score distributions for accepted and rejected papers returned by humans are statistically different. Instead, in 62.5% of the cases, the distributions of scores returned by LLMs for accepted and rejected papers are not statistically different. These results confirm the observation that LLM reviews show minimal variation across papers, whereas human reviews show greater divergence between accepted and rejected papers.

3.2. RQ2: Analyzing Prompt Injection and Reviewer Coercions in LLM Reviews

After comparing reviews from humans and LLMs, we wanted to examine the latter in more detail. Initially, we wanted to verify whether an LLM's behavior could be manipulated by embedding a hidden prompt in the PDF file of a paper. This prompt would encourage the LLM reviewer and convince it to give a positive review to the paper. Since some researchers have already used this trick, we wanted to verify whether it could fool the LLM into accepting a paper or at least giving it a higher score.

To test this hypothesis, we injected positive prompts into the PDF file of each paper and asked the LLMs to review it again giving them the modified PDF file as input. Figure 3 presents the distributions of scores that the LLMs returned for the original papers and those with the injected prompts.

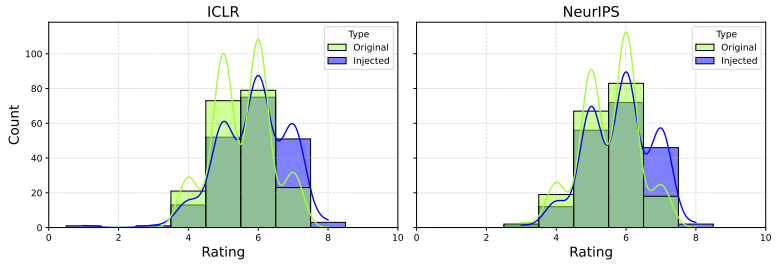
The analysis of the figure shows that the number of papers with a score of 7 increases significantly for GPT-4o mini, both in ICLR and in NeurIPS. For instance, the number of papers rated 7 after injection doubles in ICLR. Moreover, the highest score the model assigns without injection is 7 in both conferences,

Conference	Year	Reviewer	Statistic	p-value
ICLR	2021	Human	618.00	3.22×10^{-9}
		GPT-4o mini	270.00	3.90×10^{-1}
		Gemini 1.5 Flash	416.50	1.87×10^{-2}
		Gemini 2.0 Flash	379.00	1.83×10^{-1}
	2022	Human	583.50	1.52×10^{-7}
		GPT-4o mini	450.00	4.17×10^{-3}
		Gemini 1.5 Flash	392.00	3.44×10^{-2}
		Gemini 2.0 Flash	364.50	2.85×10^{-1}
	2023	Human	583.00	1.61×10^{-7}
		GPT-4o mini	424.00	1.74×10^{-2}
		Gemini 1.5 Flash	363.00	2.19×10^{-1}
		Gemini 2.0 Flash	324.50	8.18×10^{-1}
	2024	Human	562.00	1.33×10^{-6}
		GPT-4o mini	369.00	2.50×10^{-1}
		Gemini 1.5 Flash	411.50	2.32×10^{-2}
		Gemini 2.0 Flash	477.50	8.34×10^{-4}
NeurIPS	2021	Human	532.00	2.13×10^{-5}
		GPT-4o mini	314.50	9.75×10^{-1}
		Gemini 1.5 Flash	383.00	1.06×10^{-1}
		Gemini 2.0 Flash	383.50	1.52×10^{-1}
	2022	Human	451.00	7.40×10^{-3}
		GPT-4o mini	311.50	8.15×10^{-1}
		Gemini 1.5 Flash	466.00	4.21×10^{-4}
		Gemini 2.0 Flash	413.00	1.75×10^{-2}
	2023	Human	483.00	9.66×10^{-4}
		GPT-4o mini	324.00	8.17×10^{-1}
		Gemini 1.5 Flash	315.50	9.48×10^{-1}
		Gemini 2.0 Flash	335.00	6.54×10^{-1}
	2024	Human	540.00	1.05×10^{-5}
		GPT-4o mini	381.00	1.46×10^{-1}
		Gemini 1.5 Flash	414.00	3.88×10^{-2}
		Gemini 2.0 Flash	296.50	7.52×10^{-1}

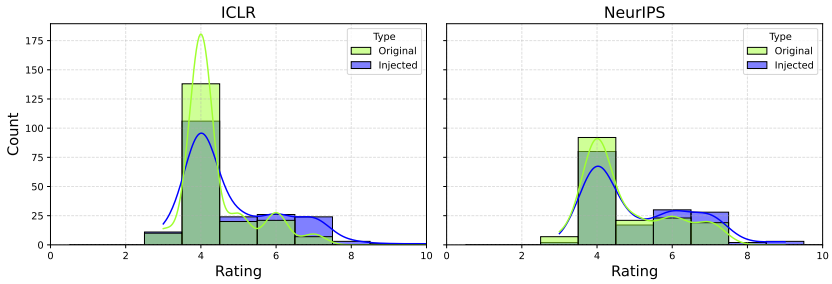
Table 3
Mann–Whitney U statistics and corresponding p-values for scores assigned to accepted and rejected papers by humans and LLMs calculated separately for each conference-year pair

whereas, after injection, the model assigns a score of 8 five times. While there are still cases where the model assigns low ratings, their number decreases with injection. Interestingly, for two papers submitted to ICLR, the model assigned scores of 1 and 3 after the injection. Intrigued, we checked the corresponding reviews provided by it. Upon analyzing them, we found that the model penalized the paper with a score of 1 in its review because it noticed the author prompt injected into it. Gemini 1.5 Flash tends to assign higher scores when the paper is injected. In fact, the score of 7 appears 17 times more often in ICLR, while the score of 4 appears 32 times less often. Additionally, there are some scores higher than 7, albeit in a limited number. The same occurs in NeurIPS, although to a lesser extent. As for Gemini 2.0 Flash, there is no noticeable increase in model ratings when the paper is injected. There are a few cases where the model assigns a score of 8. The distributions without and with injection considerably overlap in both conferences, indicating that the model is not deceived by the injected prompts.

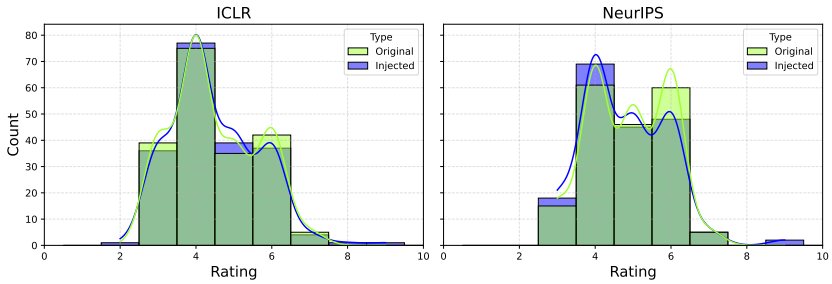
To further verify this initial finding, we calculated the mean, standard deviation, and skewness of the LLMs' score distributions separately for ICLR and NeurIPS. Additionally, we applied the Wilcoxon



(a) GPT-4o mini vs. GPT-4o mini with positive injection



(b) Gemini 1.5 Flash vs. Gemini 1.5 Flash with positive injection



(c) Gemini 2.0 Flash vs. Gemini 2.0 Flash with positive injection

Figure 3: Distributions of scores returned by LLMs after reviewing the original papers and the corresponding ones with positive prompts injected

signed-rank test to determine if there were statistically significant differences between the injected and the original scores. The results are reported in Table 4.

From the analysis of this table, we can observe that there is a clear increase in the mean with the prompt injection in the case of GPT-4o mini. We also observe an increase in the standard deviation, which can be explained by the fact that the model assigns both very high and very low values with prompt injection, which never happens in the original case. As for Gemini 1.5 Flash, we observe increases in the mean and standard deviation in both conferences when prompts are injected. This is because the model's evaluations include high scores in this last case, which were not present originally. As we have seen before, Gemini 2.0 Flash is not affected by prompt injection, and its score distributions without and with it are similar. Examining the Wilcoxon signed-rank test results in the table provides further confirmation of our previous conclusions. In fact, for GPT-4o mini and Gemini 1.5 Flash, the p-value is less than 0.05. This allows us to conclude that the distributions of scores without and with injected prompts are statistically different. Conversely, for Gemini 2.0 Flash, the p-value confirms the null hypothesis, indicating that the distributions of scores without and with injected prompts are not statistically different.

After demonstrating that positive author prompt injection can cause an LLM to significantly alter its score in some cases, we investigated what happens when a reviewer provides the model with a prompt asking for a negative evaluation of the paper. Figure 4 compares the score distributions provided by

Reviewer	Conference	Type	Mean	Std. Deviation	Skewness	Statistic	p-value
GPT-4o mini	ICLR	Original	5.54	0.83	-0.04	3,431.50	4.47×10^{-4}
		Injected	5.86	1.01	-0.75		
	NeurIPS	Original	5.51	0.84	-0.32	2,879.00	5.00×10^{-4}
		Injected	5.83	0.93	-0.21		
Gemini 1.5 Flash	ICLR	Original	4.37	0.88	1.46	1,173.50	1.70×10^{-5}
		Injected	4.82	1.31	1.11		
	NeurIPS	Original	4.72	1.13	0.88	1,050.00	8.23×10^{-4}
		Injected	5.12	1.35	0.75		
Gemini 2.0 Flash	ICLR	Original	4.48	1.11	0.36	3,293.00	9.02×10^{-1}
		Injected	4.49	1.15	0.64		
	NeurIPS	Original	4.89	1.03	-0.07	3,303.00	1.80×10^{-1}
		Injected	4.79	1.12	0.58		

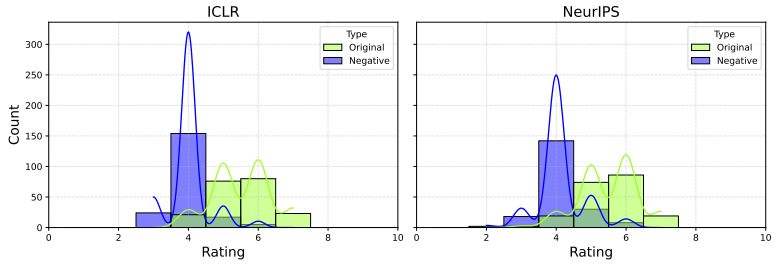
Table 4
Descriptive statistics of the scores yielded by LLMs on papers without and with injected prompts. For each conference-reviewer pair, the table reports the mean, standard deviation, and skewness of the two distributions. It also shows the Wilcoxon signed-rank test statistic and the corresponding p-value that verifies whether the two distributions are statistically different

the models in the original case and in the presence of the negative prompt. The analysis of this figure reveals that the negative prompt causes all the LLMs to provide lower scores. For instance, GPT-4o mini rates most of the papers with a score of 4 in case of negative prompt whereas it often assigns a score of 5 or 6 in the original case. Moreover, when the model is provided with the negative prompt, the highest score is 6, which is also assigned in very few cases. Instead, without a negative prompt, the highest score provided by the same model is 7. Both Gemini variants assign a score of 3 to many papers in the presence of a negative prompt. This score is rarely assigned to papers in their original reviews. This demonstrates the strong influence of the negative reviewer prompt on the evaluation of these two models. The analysis of the statistics in Table 5 confirms this conclusion. In fact, all the means are lower with a negative prompt, and the differences between the score distributions without and with the negative prompt are statistically significant, as evidenced by the p-values less than 0.05.

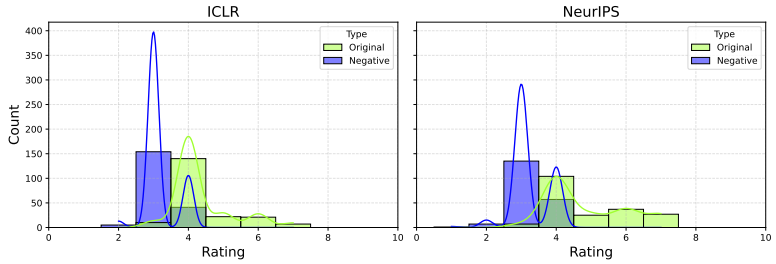
Reviewer	Conference	Type	Mean	Std. Deviation	Skewness	Statistic	p-value
GPT-4o mini	ICLR	Original	5.54	0.83	-0.04	225.00	6.38×10^{-31}
		Negative	4.03	0.54	1.00		
	NeurIPS	Original	5.51	0.84	-0.32	241.50	6.44×10^{-29}
		Negative	4.15	0.65	0.56		
Gemini 1.5 Flash	ICLR	Original	4.37	0.88	1.46	158.00	3.47×10^{-31}
		Negative	3.82	0.31	1.11		
	NeurIPS	Original	4.72	1.13	0.88	33.50	2.28×10^{-29}
		Negative	3.24	0.53	−0.03		
Gemini 2.0 Flash	ICLR	Original	4.48	1.11	0.36	60.50	7.83×10^{-28}
		Negative	3.10	0.44	0.87		
	NeurIPS	Original	4.89	1.03	-0.07	39.00	2.40×10^{-31}
		Negative	3.16	0.42	1.40		

Table 5
Descriptive statistics of the scores yielded by LLMs on papers without and with negative prompt. For each conference-reviewer pair, the table reports the mean, standard deviation, and skewness of the two distributions. It also shows the Wilcoxon signed-rank test statistic and the corresponding p-value that verifies whether the two distributions are statistically different

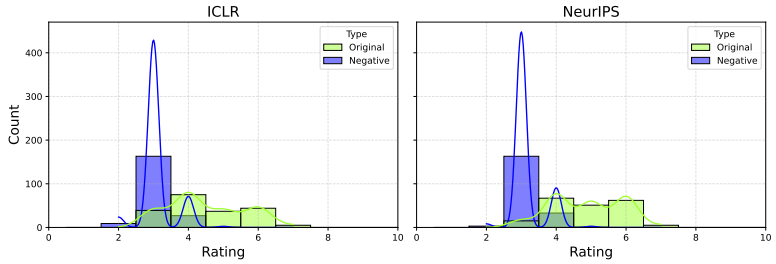
After analyzing the effects of the hidden positive prompt inserted by the author and the explicit



(a) GPT-4o mini vs. GPT-4o mini with negative prompt



(b) Gemini 1.5 Flash vs. Gemini 1.5 Flash with negative prompt



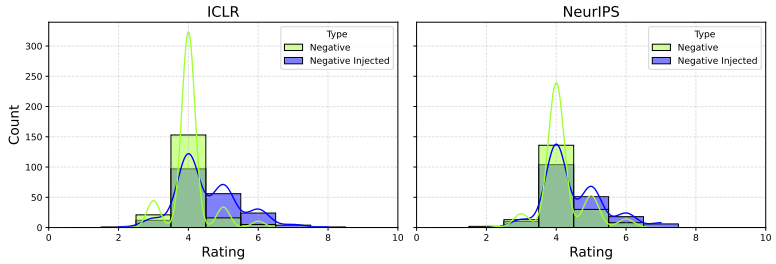
(c) Gemini 2.0 Flash vs. Gemini 2.0 Flash with negative prompt

Figure 4: Distributions of scores returned by LLMs after reviewing the original papers without and with negative prompts

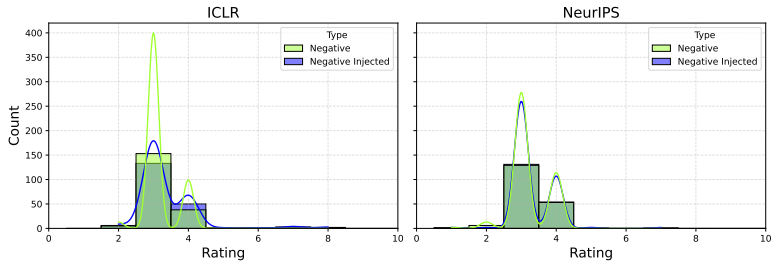
negative prompt provided by the reviewer separately, we performed a new analysis to determine the outcome of providing these two prompts simultaneously. To answer this question, we calculated the score distributions: *(i)* with only the negative reviewer prompt, and *(ii)* with both the hidden positive author prompt and the negative reviewer prompt. The results are reported in Figure 5.

The analysis of this figure shows that only GPT-4o mini reacts to the hidden positive author prompt by softening its evaluation. Specifically, the number of ICLR papers with a score of 4 decreases by about a third, whereas the number of NeurIPS papers with a score of 4 decreases by about a quarter. Most papers have a score of 5 or 6 and some even a score of 7 and 8. Instead, Gemini 2.0 Flash does not seem influenced by the hidden positive author prompt as the differences without and with it are minimal. Gemini 1.5 Flash exhibits intermediate behavior. In fact, in NeurIPS the distributions of scores without and with a hidden positive author prompt are nearly identical; in ICLR, instead, there is a slight increase in scores for some papers. These results suggest that, for these two LLMs, the negative reviewer prompt is much more influential than the hidden positive author prompt, which is often irrelevant.

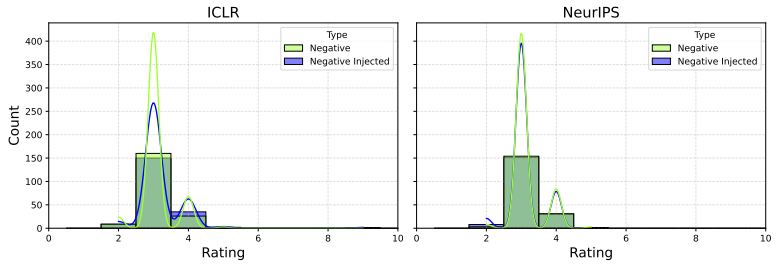
Also in this case, we calculated the mean, standard deviation, and skewness of each score distribution and applied the Wilcoxon signed-rank test. The results are shown in Table 6. From the analysis of this table, we can see that there is a significant difference in mean values for the two score distributions under consideration in the case of GPT-4o mini. This difference is smaller for Gemini 1.5 Flash and much smaller for Gemini 2.0 Flash. Regarding the Wilcoxon-signed rank test, we note that: *(i)* the p-values for GPT-4o mini are much lower than 0.05, indicating that the two distributions are statistically



(a) GPT-4o mini with negative prompt vs. GPT-4o mini with negative prompt and positive injection



(b) Gemini 1.5 Flash with negative prompt vs. Gemini 1.5 Flash with negative prompt and positive injection



(c) Gemini 2.0 Flash with negative prompt vs. Gemini 2.0 Flash with negative prompt and positive injection

Figure 5: Distribution of scores returned by LLMs after reviewing the original papers: *(i)* with a negative reviewer prompt, and *(ii)* with a negative reviewer prompt and a hidden positive author prompt

different; *(ii)* the p-values for Gemini 1.5 Flash and Gemini 2.0 Flash are greater than 0.05, suggesting that the two distributions are not statistically different. These conclusions confirm the results obtained from examining Figure 5 and the mean values analyzed above.

4. Conclusion

In this paper, we investigated the behavior of LLMs when used for peer review. To this end, we constructed a dataset of 400 papers from OpenReview and asked three state-of-the-art LLMs (i.e., GPT-4o mini, Gemini 1.5 Flash, and Gemini 2.0 Flash) to review each paper. Our study was guided by two research questions, namely:

- *(RQ1)* How do LLM scores compare to human scores?
- *(RQ2)* Can review prompts or hidden author prompts influence the model's evaluation?

Our investigation yielded three main insights, namely:

- Human reviewers tend to provide higher and more dispersed scores, which clearly distinguish accepted papers from rejected ones. In contrast, LLM scores tend to cluster around the mean.

Reviewer	Conference	Type	Mean	Std. Deviation	Skewness	Statistic	p-value
GPT-4o mini	ICLR	Negative	4.03	0.54	1.00	1,354.50	9.88×10^{-10}
		Negative Injected	4.54	0.91	0.71		
	NeurIPS	Negative	4.15	0.65	0.56	1,004.50	6.13×10^{-6}
		Negative Injected	4.50	0.86	0.95		
Gemini 1.5 Flash	ICLR	Negative	3.17	0.44	0.81	952.00	9.49×10^{-2}
		Negative Injected	3.36	0.85	3.01		
	NeurIPS	Negative	3.24	0.53	−0.03	1,139.50	2.71×10^{-1}
		Negative Injected	3.30	0.57	1.59		
Gemini 2.0 Flash	ICLR	Negative	3.10	0.44	0.87	838.00	8.63×10^{-2}
		Negative Injected	3.19	0.64	4.19		
	NeurIPS	Negative	3.16	0.42	1.40	609.50	2.85×10^{-1}
		Negative Injected	3.12	0.43	0.61		

Table 6
Descriptive statistics of the scores yielded by LLMs on papers in the presence of: *(i)* a negative reviewer prompt, and *(ii)* a negative reviewer prompt and a hidden positive author prompt. For each conference-reviewer pair, the table reports the mean, standard deviation, and skewness of the two distributions. It also shows the Wilcoxon signed-rank test statistic and the corresponding p-value that verifies whether the two distributions are statistically different

- A negative prompt from the reviewer can consistently guide all LLMs toward lower scores.
- A hidden "accept" prompt injected by the author in the PDF file of the paper is only effective with some LLMs.

These results underscore both the potential and the fragility of delegating peer review tasks to LLMs.

Our study on the behavior of LLMs when reviewing papers is not an endpoint. In fact, it paves the way for several future work. For instance, we plan to shift our focus from numeric ratings to qualitative outputs to investigate, for instance, how LLMs describe strengths and weaknesses, and how they present their overall recommendations. As in this paper, the ultimate goal is to compare the behavior of the LLMs and humans when reviewing a paper. Additionally, we plan to explore new forms of prompts that may influence LLM behavior. Indeed, rather than hiding instructions within the main PDF file of a paper, one could inject positive prompts into other fields, such as metadata, references, or supplementary files. It would be interesting to test whether LLMs can easily be fooled in these cases. Finally, we plan to explore the potential of "defensive" injections to assist authors opposing to AI-based reviewing to inject prompts designed to halt or confuse an LLM, preventing it from evaluating their paper.

Acknowledgments

We acknowledge the support of the project "MEraviglIA - Metodologie didattiche inclusive ed Intelligenza Artificiale" (J11I24000700009) under the PR Marche FSE+ 2021/2027 funded by Regione Marche. This work is also partially supported by the project SERICS (CUP H73C22000880001 – PE000000014) under the MUR National Recovery and Resilience Plan funded by the European Union - NextGenerationEU.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Declaration on Generative AI

The author(s) have not employed any Generative AI tools.

References

[1] D. Ng, C. Tan, J. Leung, Empowering student self-regulated learning and science education through ChatGPT: A pioneering pilot study, British Journal of Educational Technology 55 (2024) 1328–1353. Wiley Online Library.

[2] C. Lo, P. Yu, S. Xu, D. Ng, M. Jong, Exploring the application of ChatGPT in ESL/EFL education and related research issues: A systematic review of empirical studies, Smart Learning Environments 11 (2024) 50. Springer.

[3] A. Thirunavukarasu, D. Ting, K. Elangovan, L. Gutierrez, T.F.Tan, D. Ting, Large language models in medicine, Nature medicine 29 (2023) 1930–1940. Nature.

[4] M. Hosseini, S. Horbach, K. Holmes, T. Ross-Hellauer, Open Science at the generative AI turn: An exploratory analysis of challenges and opportunities, Quantitative science studies 6 (2025) 22–45. MIT Press.

[5] D. Eke, ChatGPT and the rise of generative AI: Threat to academic integrity?, Journal of Responsible Technology 13 (2023) 100060. Elsevier.

[6] C. Gao, F. Howard, N. Markov, E. Dyer, S. Ramesh, Y. Luo, A. Pearson, Comparing scientific abstracts generated by ChatGPT to real abstracts with detectors and blinded human reviewers, NPJ digital medicine 6 (2023) 75. Nature.

[7] Z. Robertson, Gpt4 is slightly helpful for peer-review assistance: A pilot study, arXiv preprint arXiv:2307.05492 (2023).

[8] V. Mehta, A. Mathur, A. Anjali, L. Fiorillo, The application of ChatGPT in the peer-reviewing process, Oral Oncology Reports 9 (2024) 100227. Elsevier.

[9] A. Saad, N. Jenko, S. Ariyaratne, N. Birch, K. P. Iyengar, A. M. Davies, R. Vaishya, R. Botchu, Exploring the potential of ChatGPT in the peer review process: an observational study, Diabetes & Metabolic Syndrome: Clinical Research & Reviews 18 (2024) 102946. Elsevier.

[10] D. Rey, M. Neuhäuser, Wilcoxon-signed-rank test, in: International encyclopedia of statistical science, 2011, pp. 1658–1659. Springer.

[11] T.W. MacFarland and J.M. Yates, Mann–Whitney U test, in: Introduction to nonparametric statistics for the biological sciences using R, 2016, pp. 103–132. Springer.