A Glitch of Large Language Model in Reviewing Academic Papers

Anonymous ACL submission

Abstract

Large Language Models (LLMs) have achieved great success across various areas. However, it remains an open question whether they are suitable for academic paper reviewing . We systematically examine whether LLMs can serve as paper reviewers through empirical study on papers from the International Conference on Learning Representations (ICLR) where we analyze the reviewing patterns of LLMs and identify some limitations. We find out that general-purpose LLMs struggle to generate well-structured reviews. However, when techniques such as Chain-of-Thought Prompting and Retrieval-Augmented Generation are used, LLMs demonstrate enhanced abilities in critical reasoning, improving their review quality. Additionally, supervised fine-tuning further refines their judgment, enabling more consistent decision-making in acceptance or rejection. While challenges remain, our results suggest that LLMs can work as an auxiliary reviewer.

1 Introduction

Large Language Model (LLM) has demonstrated outstanding performance in a wide range of different tasks. With the ease of usage and poweful ability to generate paragraphs, a lot of academic societies have released the regulation on use of content generated by AI. For example, conferences such as ACL will desk-reject AI-generated papers.

Recent literature shows that LLMs excel in critical thinking (Wang et al., 2023; Anonymous, 2024; Tian et al., 2024) and writing with proper prompt organization. However, from the perspective of a paper reviewer, their suitability for academic reviewing remains underexplored. We further investigate the root cause of LLMs' review patterns.

In this paper, instead of purely having LLM APIs serve as the paper reviewer in previous work (Zhou et al., 2024; Du et al., 2024), we conduct extensive experiments over International Conference on Learning Representations (ICLR) with models of different scales. With the results gathered, we conduct a deep analysis of the reviewing patterns and summarize the LLMs' pros and cons in reviewing.

Additionally, we analyze the root reason why LLMs have such patterns. With the combination of techniques such as Chain-of Thoughts (Wang et al., 2023), Retrieval Augmented Generation (Lewis et al., 2020), and, Sueprvised Fine-Tuning (Prottasha et al., 2022), we successfully improve the LLMs ability in academic reviewing.

More importantly, with advanced techniques combined, we find out the main reasons why LLM has some potential issues in academic paper review and point out the future direction of improvement. Along with our experiments and new LLM paper reviewers, we develop a large benchmark based on ICLR papers and reviews. Our benchmark and code is avaialble at. [1]

2 Related Work

2.1 LLM in Critical Thinking

Building on the impressive performance of large language models (LLMs), recent studies have explored their critical thinking abilities. Jung et al. demonstrated how prompts can elicit logical reasoning, while the Chain-of-Thought technique (Wei et al., 2022) enhances explicit reasoning processes. Although prior work examined LLMs' review-writing performance (Zhou et al., 2024; Staudinger et al., 2024), we investigate the root causes of their strengths and weaknesses, linking their critical thinking and reasoning abilities.

2.2 Retrieval Augmentaed Generation and Fine-Tuning

Retrieval-Augmented Generation (RAG) (Min et al., 2023) enables LLMs to generate answers

[1] We will release the benchmark and code for evaluation, LLM reivewers, and datasets if the paper is published.

using unseen knowledge without additional training. Similarly, supervised fine-tuning (Prottasha et al., 2022) enhances knowledge acquisition.

3 Generating Reviews

In first step, we explore several well-adopted large language models to generate reviews and assess their effectiveness. We directly ask these models produce reviews according to the review templates.

3.1 Direct Generation

We download papers and their review comments from the ICLR 2023 conference on the OpenReview platform, followed by data preprocessing. A dataset of 992 papers and their review comments are then constructed for this study.

To let the large language models efficiently read papers and generate review comments, we design and optimize specific prompt strategies to guide the models in analyzing the research content and producing review reports. Subsequently, we instructed each model to generate review comments for all 992 papers individually.

3.2 Chain of Thought and RAG

We integrate Chain-of-Thought (CoT) with Retrieval-Augmented Generation (RAG) where we meticulously selected 10 reviews that excell in both the analysis of strengths and weaknesses and in scoring accuracy as the example document to enhance the LLMs. Specifically, six of these reviews pertained to papers that were accepted, and four to papers that were rejected. These selected reviews are used to construct the RAG knowledge base. The chain-of-thought is structured as follows: first, summarizing the strengths of the paper; second, summarizing its weaknesses; and finally, concluding with an overall score. For each review, we extract the analysis of the paper's strengths and weaknesses along with the corresponding score.

3.3 Supervised Fine-Tuning

We utilize the ten reviews and their corresponding papers selected in Section 3.1 as training samples to fine-tune the model. We first fine-tune the model on whole 992 reviews. We then use 10 reviews for instruction tuning. Additionally, we collect five reviews and their corresponding papers for evaluation during the training process.

4 Evaluation

4.1 Experiments

We select representative models, including OpenAI GPT-4 Turbo, the open-source LLaMA series, and the Mistral series, considering variations in architecture, instruction fine-tuning, and parameter scales. These models span sizes from tens of billions (e.g., 3B, 7B, 8B) to hundreds of billions or even a trillion parameters (e.g., 70B, 400B+), enabling performance comparison across scales.

Table 1 lists the models used. GPT, Qwen2 (Bai et al., 2023), Mistral (Jiang et al., 2023), and LLaMA (Touvron et al., 2023) are general-purpose models, with different LLaMA model sizes utilized. InternLM (Team, 2023) specializes in reading comprehension tasks. G-Retriever-Resume-Reviewer (GRRR) is a fine-tuned LLaMA for paper reviewing [2].

To our surprise, approximately ten percent of the reviews do not adhere to the scoring template, demonstrating the unstable performance of LLMs. These scores are excluded from subsequent analyses and represented as NaN in the tables. We evaluate the review performance of each model using seven metrics across three primary dimensions:

1. **Similarity Metrics**: BERTScore, ROUGE, and BLEU evaluate the similarity between human and model-generated reviews. ROUGE-1, 2, L, and L-Sum measure overlap at the word, bigram, and longest sequence levels, with L-Sum focusing on sentence-level comparisons. BLEU-1 to 4 assess n-gram matches across four dimensions.

2. **Coherence Metric**: Perplexity measures the alignment between model-generated reviews and predefined templates.

3. **Scoring Correlation Metrics**: Scoring Accuracy and Accept/Reject Accuracy are used to evaluate the correlation between model-assigned scores and human-assigned scores.

Metric scores are calculated using multiple human reviews per paper. For the first two dimensions, model-generated reviews are compared with each human review, and results are averaged. Final scores are then averaged across 992 papers. For the third dimension, human review scores are averaged, rounded, and used for metric calculations.

[2] https://huggingface.co/alfiannajih/g-retriever-resume-reviewer

Table 1: Evaluation Results of Rouge and BLEU among Different Models

Model Name	Rouge1 ↑	Rouge2 ↑	Rouge-L ↑	Rouge-L-Sum ↑	BLEU-1G ↑	BLEU-2G ↑	BLEU-3G ↑	BLEU-4G ↑
GPT-4-Turbo	0.367	0.071	0.156	0.337	0.468	0.228	0.106	0.052
Qwen2-72B-instruct	0.371	0.086	0.173	0.339	0.488	0.267	0.140	0.077
Llama-3.2-3B	0.345	0.081	0.171	0.319	0.429	0.240	0.128	0.073
Llama-3.1-8B	0.380	0.092	0.183	0.351	0.507	0.284	0.154	0.090
Llama-3.1-70B	0.381	0.091	**0.186**	0.351	0.525	0.295	0.159	0.092
Llama-3.1-405B	0.341	0.083	0.171	0.314	0.420	0.237	0.129	0.074
Mistral-7B-Instruct-v0.2	0.393	0.093	0.185	0.362	**0.550**	**0.300**	0.159	0.089
Mixtral-8x7B-Instruct-V0.1	0.339	0.083	0.165	0.313	0.412	0.231	0.129	0.078
InternLM2-5	**0.402**	**0.098**	0.183	**0.369**	0.535	0.296	**0.162**	**0.097**
Alfiannajih/GRRR	0.263	0.060	0.138	0.240	0.206	0.114	0.066	0.041

Table 2: Evaluation Results of Remaining Metrics among Different Models

Model Name	BertScore Precision ↑	BertScore Recall ↑	BertScore F1 Score ↑	Perplexity ↓	Rating Pearson ↑	Rating Accuracy ↑	Accept/Reject Accuracy ↑	Confidence Pearson ↑	Confidence Accuracy ↑
GPT-4-Turbo	0.825	0.829	0.827	1.663	0.142	0.087	0.540	0.010	0.595
Qwen2-72B-instruct	0.841	**0.831**	**0.836**	1.181	0.250	0.044	0.541	0.036	**0.609**
Llama-3.2-3B	0.826	0.819	0.823	0.897	0.145	0.060	0.552	-0.026	0.506
Llama-3.1-8B	0.827	0.821	0.824	0.810	0.155	0.065	0.533	-0.049	0.425
Llama-3.1-70B	0.829	0.822	0.825	**0.756**	0.261	0.007	0.540	-0.010	0.103
Llama-3.1-405B	0.834	0.822	0.828	0.944	**0.282**	0.046	0.540	**0.041**	0.553
Mistral-7B-Instruct-v0.2	0.837	0.826	0.831	0.782	0.052	0.038	0.540	0.011	0.016
Mixtral-8x7B-Instruct-V0.1	0.832	0.822	0.827	1.170	0.122	0.162	0.521	-0.029	0.490
InternLM2-5	0.834	0.827	0.830	1.008	0.069	0.073	**0.554**	-0.044	0.218
Alfiannajih/GRRR	**0.845**	0.822	0.833	1.505	0.195	**0.205**	0.534	0.041	0.424

For ICLR 2023, the score range is 1 to 9. To examine distributional differences between large models and human reviewer scores, we adjust model scores by aligning their mean and standard deviation with human scores, enabling fair comparisons.

4.2 Observation and analysis

The curved score distribution for each model after curving is shown in Fig. 1, with the uncurved distribution in Appendix A. Human review distribution is presented in Fig. 2. Notably, smaller models exhibit distributions closer to human evaluations. As shown in Fig. 3, LLMs can produce similar distributions. With CoT and RAG techniques, as seen in Fig. 4, LLMs tend to provide more neutral ratings, closely aligning with human review patterns.

In general, we observe that large models are capable of generating review comments as required and generally adhere to the prescribed review format at satisfying quality. However, we can still observe the gap between human reviews and AI generated content. In terms of scoring, large models tend to assign higher scores, predominantly clustering around 8 points. Some models rarely, if ever, assign scores below 5 points. The scoring behavior significantly differs from human scoring, where high and low scores are relatively balanced. The major reason is that the deviation on original pre-training and semi-supervised data cause LLM lean to generate over-positive context. However, such deviation can be corrected in some simple fine-tuning or reinforcement learning on human feedback. Based on the performance of large models across evaluation metrics, the following patterns are identified:

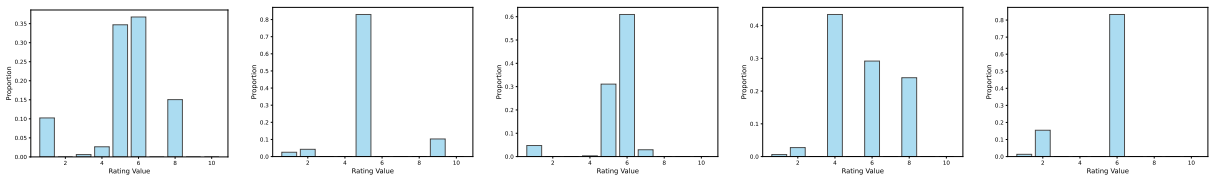
1. **Accept/Reject Accuracy**: The result is generally satisfactory, with all models achieving accuracy rates above 50%.

2. **Rate Accuracy**: Accuracy is typically below 10%, worse than random scoring. LLMs excel at predicting trends but struggle with accurate quantization scores as they are pre-trained on general cases, not tailored for this task.

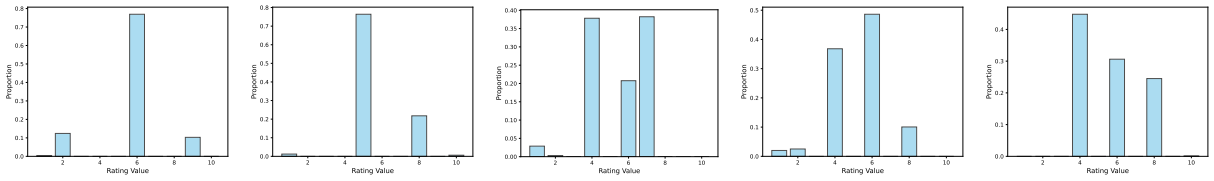
3. **Rating Pearson Correlation Coefficient**: Although the correlation is not highly significant, it is consistently positive, indicating a certain degree of consistency between model-assigned scores and human-assigned scores.

4. **Model Size and Scoring Effectiveness**: Larger models demonstrate superior scoring performance, as evidenced by the three different scales of the Llama series models.

5. **Limitations of General Models**: General-purpose models tend to produce redundant content (indicated by high perplexity) and exhibit poor critical analysis capabilities.



(a) GPT (b) Qwen2 (c) Llama-3B (d) Llama-8B (e) Llama-70B



(f) Llama-405B (g) Mistral (h) Mixtral (i) InternLM (j) Resume-Reviewer

Figure 1: The score distribution after curve given by different large models

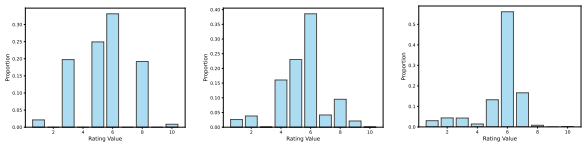


Figure 2: Human Curved

Figure 3: LLM, Curved

Figure 4: LLM (CoT and RAG, Curved)

Figure 5: Distribution of Rating Scores

6. **Impact of Fine-Tuning and Knowledge Graphs**: Task-specific knowledge base is highlighted with better performance.

7. **BERTScore Performance**: Results are near human-written reviews.

Compared to model size growth, prompting large models with appropriate knowledge is more important for better reviews. Thus, the key finding is that for review generation tasks, the thinking process outweighs the increase in parameters. A 7B-scale model can already meet the requirements of memorizing past review experiences. However, prompting LLMs to think and build a logical chain cannot be solely achieved through generation.

We analyze both human and LLM-generated reviews quantitatively and qualitatively. LLM reviews tend to focus on strengths and are more lenient, while human reviews are stricter. Focusing on writing quality, a key factor in human reviews, we find a Pearson correlation of 0.273 in human reviews and 0.179 in LLM-generated ones. This indicates that LLMs are more likely to accept papers, reflecting their learned evaluation patterns.

In summary, despite the models' shortcomings in specific scoring accuracy, they maintain a high level of consistency in the more coarse-grained accept/reject binary classification tasks and exhibit robust performance in textual language aspects.

4.3 Supervised Fine-Tuning

We use the AdamW optimizer with a learning rate of 10^{-5} and weight decay of 0.01 for 992 steps. We choose the model Llama-3.2-3B. The accept/reject accuracy is 0.601, showing significant improvement. Hence, supervised fine-tuning helps LLMs achieve a distribution closer to the final results.

5 Conclusion and Limitation

This paper explores using LLMs for academic paper reviewing. We find that while LLMs with general knowledge cannot generate proper reviews directly, techniques like chain-of-thought and retrieval-augmented generation allow LLMs to make informed decisions on paper acceptance or rejection. Supervised fine-tuning further enhances their performance, suggesting LLMs could be useful as auxiliary reviewers in the future.

A key limitation of the paper is focus on empirical experiments, with a need for theoretical exploration. Additionally, the absence of reasoning models in this study limits its scope, and future research could explore their integration. Overall, while LLMs show potential, further advancements are required to develop a comprehensive solution for academic paper reviewing.

References

Anonymous. 2024. Flow of reasoning: Training LLMs for divergent problem solving with minimal examples. In *Submitted to The Thirteenth International Conference on Learning Representations*. Under review.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Jiangshu Du, Yibo Wang, Wenting Zhao, Zhongfen Deng, Shuaiqi Liu, Renze Lou, Henry Peng Zou, Pranav Narayanan Venkit, Nan Zhang, Mukund Srinath, et al. 2024. Llms assist nlp researchers: Critique paper (meta-) reviewing. In *EMNLP*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Jaehun Jung, Lianhui Qin, Sean Welleck, Faeze Brahman, Chandra Bhagavatula, Ronan Le Bras, and Yejin Choi. 2022. Maieutic prompting: Logically consistent reasoning with recursive explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 1266–1279.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Sewon Min, Weijia Shi, Mike Lewis, Xilun Chen, Wen-tau Yih, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2023. Nonparametric masked language modeling. *ACl Findings*.

Nusrat Jahan Prottasha, Abdullah As Sami, Md Kowsher, Saydul Akbar Murad, Anupam Kumar Bairagi, Mehedi Masud, and Mohammed Baz. 2022. Transfer learning for sentiment analysis using bert based supervised fine-tuning. *Sensors*, 22(11):4157.

Moritz Staudinger, Wojciech Kusa, Florina Piroi, and Allan Hanbury. 2024. An analysis of tasks and datasets in peer reviewing. In *Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024)*, pages 257–268.

InternLM Team. 2023. Internlm: A multilingual language model with progressively enhanced capabilities.

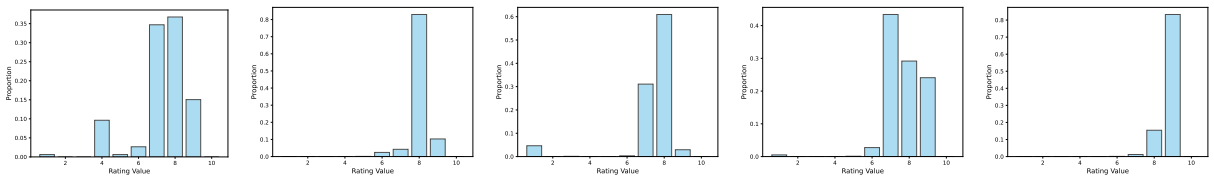
Yufei Tian, Abhilasha Ravichander, Lianhui Qin, Ronan Le Bras, Raja Marjieh, Nanyun Peng, Yejin Choi, Thomas L Griffiths, and Faeze Brahman. 2024. Macgyver: Are large language models creative problem solvers? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5303–5324.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

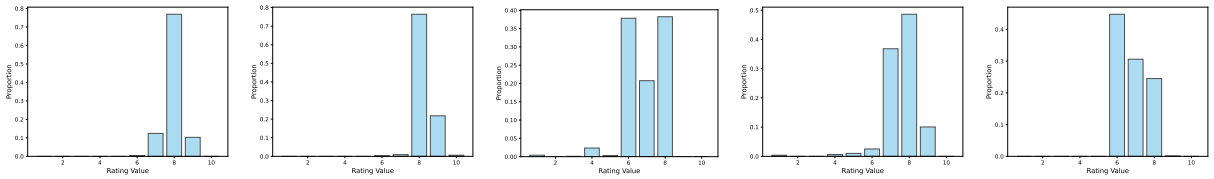
Boshi Wang, Sewon Min, Xiang Deng, Jiaming Shen, You Wu, Luke Zettlemoyer, and Huan Sun. 2023. Towards understanding chain-of-thought prompting: An empirical study of what matters. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

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



(a) GPT (b) Qwen2 (c) Llama-3B (d) Llama-8B (e) Llama-70B



(f) Llama-405B (g) Mistral (h) Mixtral (i) InternLM (j) Resume-Reviewer

Figure 6: The score distribution before curve given by different large models

A Score Distribution before Curve

In Fig. 6, we show the score distribution given by different large models before curve. As we can see, LLMs are prone to give very high scores. As a result, we first curve the score.

B Ethnical Consideration

We acknowledge the potential impact that some reviewers may use AI generated content to replace human reviewing during the paper review process. In this paper, we hope to raise the consideration and development on understanding LLM ability in critical thinking.