A Glitch of Large Language Model in Reviewing Machine Learning Academic Papers

Tianchi Xu, Changyu Chen,Chaomin Huang, Lehui Wang, Yongtong Gu

Frontiers of Computer Science, DOI: 10.1007/s11704-025-51031-6

Problems & Ideas

- From the perspective of a paper reviewer, how we can evaluate LLM's critical thinking ability in terms of academic reviewing:
  - How LLM reviewer performs comparing to uman reviewer
  - Why LLM works or why LLM does not

Ideas: Integrate chain-of-thought prompting, retrieval-augmented generation (RAG), and supervised fine-tuning (SFT) to probe LLM review behavior and distill its strengths and weaknesses in academic peer review:

Main Contributions

- Contributions:
  - Smaller models yield distributions closer to human evaluations. Large models tend to assign disproportionately high scores;
  - LLM overconfidence and bias;
  - Simple fine-tuning can help mitigate this bias and promote more balanced scoring.
  - Summary: LLMmaintain a high level of consistency in the more coarse-grained accept/reject binary classification tasks and exhibit robust performance in textual language aspects.
  - Future Work: We plan to examine case studies of improperly reviewed papers and explore alternative RAG corpora, such as Related Work sections, to complement human review data

Table 1 Evaluation Results of Rouge and BLEU among Different Models (with SFT applied).

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 (with SFT applied).

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