LLM as a Scorer: The Impact of Output Order on Dialogue Evaluation

Yi-Pei Chen[*], **KuanChao Chu**[*], **Hideki Nakayama**

The University of Tokyo, Japan
ypc@g.ecc.u-tokyo.ac.jp, {kcchu, nakayama}@nlab.ci.i.u-tokyo.ac.jp

Abstract

This research investigates the effect of prompt design on dialogue evaluation using large language models (LLMs). While LLMs are increasingly used for scoring various inputs, creating effective prompts for dialogue evaluation remains challenging due to model sensitivity and subjectivity in dialogue assessments. Our study experimented with different prompt structures, altering the sequence of output instructions and including explanatory reasons. We found that the order of presenting reasons and scores significantly influences LLMs' scoring, with a "reason-first" approach yielding more comprehensive evaluations. This insight is crucial for enhancing the accuracy and consistency of LLM-based evaluations.

Introduction

Using large language models (LLMs) (OpenAI 2023; Touvron et al. 2023) as evaluators to assign scores to the given inputs have become prevalent. Leblond et al. (2023) output a score between 0 and 1 to estimate the correctness of generated code, thereby ranking their quality. Similarly, Park et al. (2023) assign poignancy score to the generated text for the retrieval task. Other research explores using LLMs to assess generated texts, finding the LLM scores correlates higher with human evaluators than existing automatic metrics (Gao et al. 2023; Shen et al. 2023; Liu et al. 2023; Luo, Xie, and Ananiadou 2023).

However, designing evaluation prompt for LLMs is not a trivial task, especially for dialogue evaluation. Different models exhibit varied sensitivity to the nuances of input prompts. Even slight linguistic variations can lead to significant fluctuations in task performance (Leidinger, Rooij, and Shutova 2023). Moreover, the inherent subjectivity in dialogue evaluation adds on the difficulty and versatility in LLMs' evaluation results. While prompt optimization techniques (Chen et al. 2023; Yang et al. 2023; Zhang et al. 2023; Prasad et al. 2023) have been developed to assist in designing more effective prompts, these methods require paired input-output samples for objective value calculation. Unfortunately, the lack of available dialogue-score pairing data hampers the application of prompt optimization in dialogue evaluation.

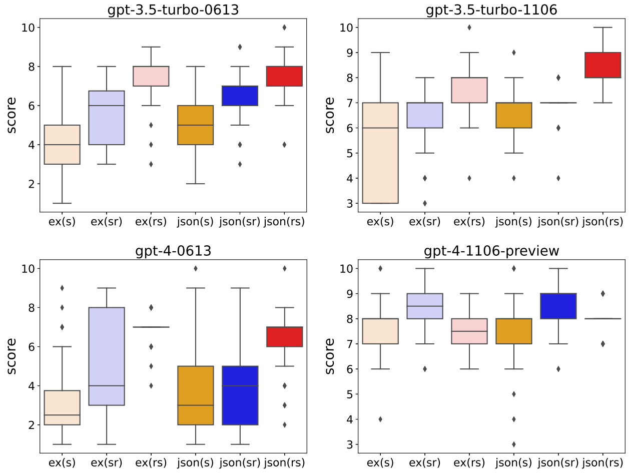

Figure 1: Score distribution across 50 trials for each model and output instruction configuration for a dialogue set.

In this study, we aim to investigate the influence of prompt design on dialogue evaluation, specifically focusing on how the output instructions affects the resulting scores. We have developed multiple prompt variations to assess the quality of a series of dialogues. These variations involve altering the sequence order of the outputs and examining whether including explanatory reasons along with the scores impacts the evaluation. Our analysis compares the influence of different prompts on the scoring outcomes across various versions of GPT models.

We observed that the different order of output instructions can result in different scoring distributions by certain LLMs, even when the corresponding output reasons are similar. Considering the sequential generation nature of auto-regressive models, placing the score after the reasons allows it to reference both the reasons and the input prompt, a dynamic not possible when this order is reversed. The finding suggests that a "reason-first" output instruction might lead to a more comprehensive understanding and adherence to the specific requirements of the task.

[*] These authors contributed equally.

AAAI 2024 Spring Symposium on User-Aligned Assessment of Adaptive AI Systems. Stanford University, Stanford, CA, USA.

Config	Output Instruction in the Prompt

ex (s)	`Example JSON output:`
	`{"score": 5}`

ex (sr)	`Example JSON output:`
	`{"score": 5, "reasons": "<your reasons for the rating>"}`

ex (rs)	swap the order of “score” and “reasons” in ex (sr)

json (s)	`Output a json of the following format:`
	`{"score": "<integer>"}`

json (sr)	`Output a json of the following format:`
	`{"score": "<integer>", "reasons": "point out the issues and your reasons for the rating"}`

json (rs)	swap the order of “score” and “reasons” in json (sr)

Table 1: The variations of output instruction.

Approach

In this study, the task assigned to the LLM is to rate a given set of dialogues on a scale from 1 to 10, where 1 indicates no issues in the set of dialogues, and 10 signifies severe problems. Additionally, if specified in the prompt, the LLM is required to provide a rationale for the rating. The dialogues are presented in chronological order, and the output score is determined based on a comprehensive evaluation of the entire set, focusing on key aspects such as repetitiveness, factual accuracy, and coherence.

Along with the task description, we have integrated five customized rules into the prompt, derived from observations in previous experiments without these rules. The special rules include instructions for the LLM to prioritize the number of issues over their impact and to assign more weight to aspects exhibiting significant issues, rather than averaging out the score across all aspects.

The final evaluation prompt is organized as follows: a set of dialogues, task description, special rules, and output instruction (see Table 1). For each set of dialogues, we conducted N trials for each of the six configurations (config), varying the output instruction. This experiment was then replicated across M different models.

Experiment

Data To assess the capability of LLMs in identifying issues within dialogues, we collected LLM-generated dialogues from Park et al. (2023) and manually grouped them into 25 sets. Each set contains four to six dialogues and exhibits one or more problems, such as repetition or contradictions between dialogues.

Model We selected four recent LLMs to serve as scorers: gpt-3.5-turbo-0613, gpt-3.5-turbo-1106, gpt-4-turbo-0613 (gpt-4-0613), and gpt-4-1106-preview (gpt-4-1106). Note that our aim is to analyze the evaluation scores across various models when altering output instructions, and not to compare them with human judgements for this subjective task.

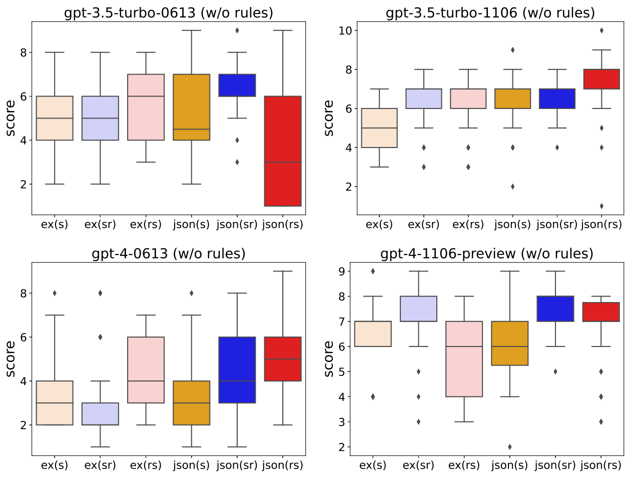

Figure 2: Score distribution across 50 trials for each model and output instruction configuration for a dialogue set, with the ‘special rules’ omitted from the prompt.

Result and Analysis

The Importance of Output Instruction Order Table 2 presents the mean scores and standard deviations (std) of 10 trials for all 25 dialogue sets across all configs and models. In both *ex (·)* and *json (·)* formats, the mean scores for the *rs* settings (output reasons before the score) are generally higher than their *sr* (output score before reasons) counterparts. [1]For instance, in the *json (rs)* config using gpt-4-0613, the mean score is 5.34, while it drops to 3.26 in *json (sr)*, despite providing similar reasons. We conjecture that in the *rs* setting, the autoregressive nature of the model allows the score to be influenced by the previously outputted reasons.

Config	GPT-3.5-turbo		GPT-4	
	-0613	-1106	-0613	-1106
ex (s)	3.68 ±1.17	4.51 ±1.19	3.36 ±1.07	**8.18** ±1.05
ex (sr)	4.20 ±1.19	5.49 ±1.22	3.39 ±1.13	7.55 ±1.12
ex (rs)	**6.09** ±1.23	**7.66** ±0.81	**5.58** ±1.19	7.39 ±0.90
json (s)	4.03 ±1.16	6.18 ±1.09	3.13 ±1.10	6.74 ±1.24
json (sr)	4.66 ±1.15	6.76 ±0.94	3.26 ±1.11	**7.69** ±1.06
json (rs)	**5.78** ±1.42	**7.99** ±0.94	**5.34** ±1.22	7.54 ±0.95

Table 2: Mean scores and std for 25 dialogue sets, evaluated across different models and output instruction configurations.

Different Levels of Rule Understanding In a focused study on a single set with additional 40 trials, as depicted in Fig. 1, we observed a trend consistent with the findings presented in Table 2. However, as shown in Fig.2, when we removed the ‘special rules’ from the prompt, we found that most scores were lower and the distinctions between different settings became less pronounced. This highlights the models’ sensitivity to the changes of the prompt.

[1] The exception is observed with the gpt-4-1106 model.

Conclusion

Our study highlights the sensitivity of LLMs to the order of output instructions, which could be amplified by task-specific rules. These findings offer insights for optimizing prompts in subjective tasks like dialogue evaluation.

Acknowledgements

This work was supported by the Institute of AI and Beyond of the University of Tokyo and the commissioned research (No. 225) by the National Institute of Information and Communications Technology (NICT), Japan.

References

Chen, L.; Chen, J.; Goldstein, T.; Huang, H.; and Zhou, T. 2023. InstructZero: Efficient Instruction Optimization for Black-Box Large Language Models. *arXiv preprint arXiv:2306.03082*.

Gao, M.; Ruan, J.; Sun, R.; Yin, X.; Yang, S.; and Wan, X. 2023. Human-like summarization evaluation with chatgpt. *arXiv preprint arXiv:2304.02554*.

Leblond et al., p. 2023. AlphaCode 2 Technical Report.

Leidinger, A.; Rooij, R. V.; and Shutova, E. 2023. The language of prompting: What linguistic properties make a prompt successful? In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Liu, Y.; Iter, D.; Xu, Y.; Wang, S.; Xu, R.; and Zhu, C. 2023. G-Eval: NLG Evaluation using Gpt-4 with Better Human Alignment. In Bouamor, H.; Pino, J.; and Bali, K., eds., *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2511–2522. Singapore: Association for Computational Linguistics.

Luo, Z.; Xie, Q.; and Ananiadou, S. 2023. Chatgpt as a factual inconsistency evaluator for abstractive text summarization. *arXiv preprint arXiv:2303.15621*.

OpenAI. 2023. GPT-4 Technical Report. *ArXiv*, abs/2303.08774.

Park, J. S.; O'Brien, J. C.; Cai, C. J.; Morris, M. R.; Liang, P.; and Bernstein, M. S. 2023. Generative agents: Interactive simulacra of human behavior. *arXiv preprint arXiv:2304.03442*.

Prasad, A.; Hase, P.; Zhou, X.; and Bansal, M. 2023. GrIPS: Gradient-free, Edit-based Instruction Search for Prompting Large Language Models. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, 3827–3846.

Shen, C.; Cheng, L.; Nguyen, X.-P.; You, Y.; and Bing, L. 2023. Large Language Models are Not Yet Human-Level Evaluators for Abstractive Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, 4215–4233.

Touvron, H.; Martin, L.; Stone, K.; Albert, P.; Almahairi, A.; Babaei, Y.; Bashlykov, N.; Batra, S.; Bhargava, P.; Bhosale, S.; et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Yang, C.; Wang, X.; Lu, Y.; Liu, H.; Le, Q. V.; Zhou, D.; and Chen, X. 2023. Large language models as optimizers. *arXiv preprint arXiv:2309.03409*.

Zhang, T.; Wang, X.; Zhou, D.; Schuurmans, D.; and Gonzalez, J. E. 2023. TEMPERA: Test-Time Prompt Editing via Reinforcement Learning. In *The Eleventh International Conference on Learning Representations*.