

Large Language Models are not Fair Evaluators

Peiyi Wang¹ Lei Li¹ Liang Chen¹ Zefan Cai¹ Dawei Zhu¹
 Binghuai Lin³ Yunbo Cao³ Qi Liu² Tianyu Liu³ Zhifang Sui¹

¹ National Key Laboratory for Multimedia Information Processing, Peking University

² The University of Hong Kong ³ Tencent Cloud AI

{wangpei9979, nlp.lilei, zefncai}@gmail.com

leo.liang.chen@outlook.com; {dwzhu, szf}@pku.edu.cn

liuqi@cs.hku.hk; {binghuailin, yunbocao, rogertyliu}@tencent.com

Abstract

In this paper, we uncover a systematic bias in the evaluation paradigm of adopting large language models (LLMs), e.g., GPT-4, as a referee to score and compare the quality of responses generated by candidate models. We find that the quality ranking of candidate responses can be easily hacked by simply altering their order of appearance in the context. This manipulation allows us to skew the evaluation result, making one model appear considerably superior to the other, e.g., Vicuna-13B could beat ChatGPT on 66 over 80 tested queries with ChatGPT as an evaluator. To address this issue, we propose a calibration framework with three simple yet effective strategies: 1) Multiple Evidence Calibration, which requires the evaluator model to generate multiple evaluation evidence before assigning ratings; 2) Balanced Position Calibration, which aggregates results across various orders to determine the final score; 3) Human-in-the-Loop Calibration, which introduces a balanced position diversity entropy to measure the difficulty of each example and seeks human assistance when needed. We also manually annotate the “win/tie/lose” outcomes of responses from ChatGPT and Vicuna-13B in the Vicuna Benchmark’s question prompt, and extensive experiments demonstrate that our approach successfully mitigates evaluation bias, resulting in closer alignment with human judgments. We release our code and human annotation at <https://github.com/i-Eval/FairEval> to facilitate future research.

1 Introduction

The rapid advancement of Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2022) has underscored the importance of evaluating their alignment with human intent in generated responses, making it an active field of research. Traditional n-gram metrics like BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004), as well as more sophisticated model-based evaluations such as BERTScore (Zhang et al., 2020) and

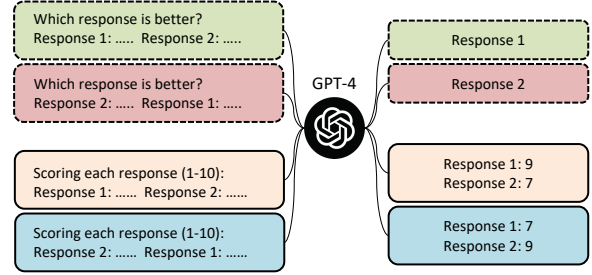


Figure 1: Simply changing the order of candidate responses leads to overturned comparison results, even though we add the command “ensuring that the order in which the responses were presented does not affect your judgment” into the prompt.

BARTScore (Yuan, Neubig, and Liu, 2021), are insufficient for thoroughly assessing this alignment (He et al., 2023). While human evaluation provides the most accurate measure of model performance and valuable insights, it can often be costly and time-consuming. As a result, there is a growing demand for automated assessment methods that can consistently align with human judgments while being more efficient and cost-effective.

ChatGPT (OpenAI, 2022) and GPT-4 (OpenAI, 2023) have recently demonstrated remarkable performance across various tasks, leading to their widespread use as both the annotators (Peng et al., 2023; Xu et al., 2023) and evaluators (Zheng et al., 2023; Peng et al., 2023; Sun et al., 2023; Zhou et al., 2023; Gao et al., 2023; Wang et al., 2023b; Dubois et al., 2023; Wang et al., 2023a). For example, The evaluation pipeline of Vicuna (Zheng et al., 2023) has gained significant interest and wide usage due to its simplicity and interpretability. It prompts GPT-4 to score and compare candidate responses and provide explanations, making it a valuable tool for evaluation. However, it is unclear how reliable LLMs are as evaluators, as they are known to be sensitive to textual instructions and inputs (Dong et al., 2022; Turpin et al., 2023;

Bowman, 2023). This raises questions about the resilience of this paradigm against perturbations, such as the ordering of candidates during scoring, potentially becoming the Achilles’ Heel that can be easily hacked for unreliable evaluations.

In this paper, we take a sober look at the LLMs-as-evaluator paradigm and uncover a significant positional bias. Specifically, we demonstrate that GPT-4 exhibits a preference for the first displayed candidate response by consistently assigning it higher scores, even when the order of candidates is subtly altered. As illustrated in Figure 1, merely swapping the presentation order can reverse evaluation outcomes. This bias is also present in ChatGPT, which typically favors the second response. These findings highlight previously overlooked limitations in the current evaluation paradigm.

To address this issue, we propose three simple yet effective strategies to calibrate positional bias: **1) Multiple Evidence Calibration (MEC)**: We prompt the model to generate evaluation evidence before assigning scores, leveraging the inherent properties of causal language models for calibration. We also employ ensemble techniques to incorporate multiple evidence calibration results to further stabilize the evaluation. **2) Balanced Position Calibration (BPC)**: To further reduce positional bias, we evaluate each candidate in both positions across two runs and compute the final score as the average of the two runs. **3) Human In The Loop Calibration (HITLC)**: We also explore human-in-the-loop evaluation and consider a diversity-based method to get a cue to indicate biased candidates based on the evaluation results of MEC and BPC.

To assess the efficacy of our methods, we manually annotate the “win/tie/lose” outcomes of responses from ChatGPT and Vicuna-13B in the Vicuna benchmark (Zheng et al., 2023), encompassing 80 questions spanning 9 distinct question categories. Our MEC and BPC enhance the evaluation alignment of GPT-4 and ChatGPT by 9.8% and 14.3% accuracy, respectively. Moreover, based on MEC and BPC, our HITLC can further effectively integrate human assistance into the evaluation process. Specifically, with only a 20% human annotation cost, GPT-4 and ChatGPT can achieve comparable or even better annotation alignment with the average human performance, reducing the annotation cost by up to 39%.

In summary, our key contributions are: **1)** We reveal that LLMs exhibit severe positional bias, com-

[Question]
{Q}
[The Start of Assistant 1’s response]
{R1}
[The End of Assistant 1’s response]
[The Start of Assistant 2’s response]
{R2}
[The End of Assistant 2’s response]
[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question displayed above.
Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.
Please first output a single line containing only two values indicating the scores for Assistant 1 and 2, respectively.
The two scores are separated by a space. In the subsequent line, please provide a comprehensive explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

Table 1: The evaluation template with three slots ({Q}, {R1} and {R2}) from Zheng et al. (2023). Even though the template emphasizes not letting the order affect the results (red text), large language models still have a large positional bias.

promising their fairness as evaluators; **2)** We develop a calibration framework with three simple yet effective strategies to calibrate the positional bias of LLMs; **3)** We manually annotate the “win/tie/lose” outcomes of responses from ChatGPT and Vicuna-13B in the Vicuna benchmark and demonstrate the effectiveness of our proposed approach through experimental results, which show closer alignment with human judgments.

2 Positional Bias of the LLM Evaluator

2.1 LLMs as Evaluators

Recently, researchers have been utilizing LLMs such as GPT-4 as evaluators to compare the performance of two AI assistants. As shown in Table 1, an evaluation template with three placeholders $T(Q, R1, R2)$, is used to query the LLM for evaluation. For each testing question q , given two responses $r1$ and $r2$ from Assistant 1 and Assistant 2, respectively, the researchers populate these responses into the corresponding slots of the evaluation template to form a prompt: $T(Q = q, R1 = r1, R2 = r2)$. The prompt is then used to query the LLM in order to obtain the comparison result. In this paper, we found that LLM suffers from severe

EVALUATORS	VICUNA-13B v.s. OTHER MODELS	VICUNA-13B WIN RATE		CONFLICT RATE
		AS ASSISTANT1	AS ASSISTANT2	
GPT-4	Vicuna-13B v.s. ChatGPT	51.3%	23.8%	37 / 80 (46.3%)
GPT-4	Vicuna-13B v.s. Alpaca-13B	92.5%	92.5%	4 / 80 (5.0%)
ChatGPT	Vicuna-13B v.s. ChatGPT	2.5%	82.5%	66 / 80 (82.5%)
ChatGPT	Vicuna-13B v.s. Alpaca-13B	37.5%	90%	42 / 80 (52.5%)

Table 2: The Win Rate of Vicuna-13B significantly fluctuates when positioned as Assistant 1 and Assistant 2, with GPT-4 and ChatGPT serving as evaluators. CONFLICT RATE refers to the proportion of conflicting results given by the same evaluator when simply changing the position of two models.

positional bias, i.e., by swapping the slots of the two responses and querying LLM twice, the evaluator will most likely produce conflicting evaluation results, and the evaluator prefers the response at a certain position.

2.2 Revealing the Positional Bias

In this section, we adopt GPT-4 and ChatGPT as evaluators to analyze the characteristics of positional bias in LLM evaluators. We find that:

LLMs are sensitive to the position of responses.

As shown in Table 2, in the evaluation of “Vicuna-13B v.s. ChatGPT” and “Vicuna-13B v.s. Alpaca-13B”, when the order was changed, LLMs provide different evaluation results, e.g., the win rate of Vicuna-13B extremely differs when Vicuna-13B is evaluated as Assistant 1 and Assistant 2.

To empirically evaluate the sensitivity, we introduced a metric **Conflict Rate** to measure the sensitivity of the model to response positions quantitatively. Formally, given N examples $\{(q_i, r1_i, r2_i)\}_{i=1}^N$, for each example $(q_i, r1_i, r2_i)$, we query the LLM with two prompts $T(q_i, r1_i, r2_i)$ and $T(q_i, r2_i, r1_i)$, and obtain corresponding two evaluation results \mathbf{ER}_i^{r12} and \mathbf{ER}_i^{r21} . Then we calculate the Conflict Rate of the LLM evaluator as follows:

$$\text{Conflict Rate} = \frac{\sum_{i=1}^N \mathbb{I}(\mathbf{ER}_i^{r12} \neq \mathbf{ER}_i^{r21})}{N}, \quad (1)$$

where $\mathbb{I}(\cdot)$ is the indicator function. We found that GPT-4 exhibited conflict rates of 46.3% and 5.0%, respectively. In contrast, ChatGPT displayed considerably higher conflict rates, with figures of 82.5% and 52.5%, respectively. These findings indicate that LLMs can be self-conflicting due to the sensitivity of the response order in the template, with stronger models being less influenced by the placement of responses.

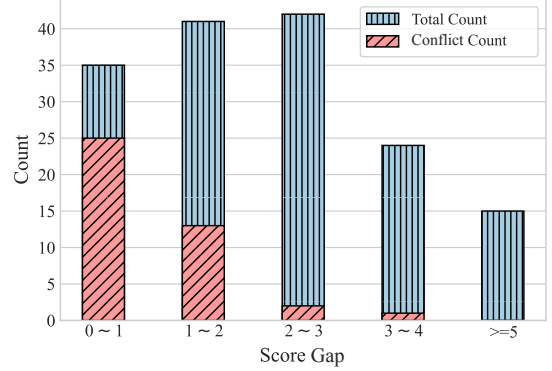


Figure 2: The conflict rate is negatively correlated with the score gap between the two responses. When swapping the order of two responses, the smaller the score gap between them, the more likely GPT-4 is to produce conflicting results.

LLMs suffer from Positional Bias, i.e., they prefer the response in the specific position.

Based on the same evaluation template T in Table 1, GPT-4 tends to favor the response in the first position, while ChatGPT shows a preference for the response in the second position. For example, as illustrated in Table 2, in the comparison “Vicuna-13B v.s. ChatGPT”, GPT-4 yields Win Rates of 51.3% and 23.8% for Vicuna-13B when it is positioned as Assistant 1 and Assistant 2, respectively. Conversely, ChatGPT indicates Win Rates of only 2.5% and up to 82.5% for Vicuna-13B when it is positioned as Assistant 1 and Assistant 2, respectively.

The degree of positional bias varies based on the difference in response quality.

We notice that the conflict rate of “Vicuna-13B v.s. Alpaca-13B” is much lower than that of “Vicuna-13B v.s. ChatGPT”, suggesting that positional bias may not have the same impact on the assessment of different responses. One potential reason is that there is a significant difference in the quality of responses between Alpaca models and Vicuna models, and positional bias is not strong enough to change the