# Enhancing Fairness in LLM Evaluations: Unveiling and Mitigating Biases in Standard-Answer-Based Evaluations

**Tong Jiao[1]*, Jian Zhang[2], Kui Xu[2], Rui Li[2], Xi Du[2], Shangqi Wang[2], Zhenbo Song[3]**
[1]Carnegie Mellon University
[2]China Mobile
[3]Nanjing University of Science and Technology
tongjiao@cs.cmu.edu

## Abstract

Large Language Models (LLMs) are recognized for their effectiveness in comparing two answers. However, LLMs can still exhibit biases when comparing one answer to a standard answer, particularly in real-world scenarios like new employee orientations. This paper identifies positional and verbosity biases in LLM evaluators in such contexts. To mitigate these biases, we apply Chain of Thought prompting and Multi-Agent Debate strategies. Our research reveals that bias prevalence varies among different models, indicating the need for tailored approaches to ensure unbiased and constructive feedback.

## Introduction

The capabilities of Large Language Model (LLM) based chat assistants to follow instructions and learn to align with human preferences have gained significant interest (Ouyang et al. 2022; Zhou et al. 2023). Since some state-of-the-art LLMs like GPT-4 (OpenAI 2024) have demonstrated strong human alignment (Zheng et al. 2023), there are extensive attempts to utilize them in evaluation tasks as a substitution of human evaluators. These efforts aim to achieve faster and more cost-effective evaluations (Dubois et al. 2024; Wang et al. 2023b). For instance, LLMs are employed as evaluators (Dubois et al. 2024) to complete pairwise comparison tasks in the reinforcement learning with human feedback (RLHF) process, achieving a 45x lower cost than using human crowdworkers. The high availability and performance of LLM evaluators make them valuable tools for comparing two answers and picking the higher-quality one. When they are used to help human evaluators, they can further enhance human performance as well (Xiao et al. 2024).

However, LLM evaluators are known to exhibit biases in evaluation tasks (Stureborg, Alikaniotis, and Suhara 2024), and the robustness of LLM evaluators is less explored. One such bias is verbosity bias, where LLM evaluators tend to favor longer responses, even when a shorter response is of higher quality (Saito et al. 2023; Wu and Aji 2023). Another bias, *i.e.* positional bias, refers to the phenomenon where the evaluations can vary by altering the position of a text within a context (Wang et al. 2023a).

*Work done while at China Mobile



Prior research has addressed several biases in pairwise comparisons, but evaluation tasks are not limited to pairwise comparisons. Another vital scenario involves comparing an answer to a pre-defined standard answer and providing personalized feedback. A concrete example of this scenario is evaluating responses from trainees in a new employee orientation program. Since face-to-face orientation events and computer-based events cannot substitute for each other (Bauer and Erdogan 2011), comparing new employees' responses to career-related questions against their established expectations and giving personalized feedback can significantly improve their experience in computer-based events. Mitigating positional and verbosity bias is necessary since participants should learn their expectations, not remembering the length of standard answers and the order of key points. Similar contexts that require comparing an answer to a predefined standard answer include customer support quality assurance systems and clinical decision support systems.

This paper focuses on the robustness of LLM evaluators when a standard answer is provided. We first reveal inherent biases in LLM evaluators, and then explore methods for mitigating these biases. In brief, our research and experiments are summarized in two-fold.

- Our experiments reveal that some LLM evaluators, including GPT-3.5 and Llama3-8b (Meta 2024), exhibit significant positional and verbosity biases. Specifically, we design two attacks, *i.e.* "shuffled list attack" and "repetitive content attack", to show these biases. Examples of each attack are shown in Figure 1, with a more detailed discussion provided in the methodology section.
- We examine the effectiveness of two fine-tuning-free approaches, Chain of Thought and multi-agent debate, in mitigating positional and verbosity biases. Our findings suggest that there may not be a universal fine-tuning-free approach to reduce bias across all LLMs, and the optimal way to mitigate biases is model-specific.

## Methodology

This section discusses the experimental setup for assessing positional and verbosity biases in LLM evaluators. Three LLMs are evaluated: GPT3.5 (`gpt-3.5-turbo-0125`), GPT4o (`gpt-4o-2024-05-13`), and the 8B version of

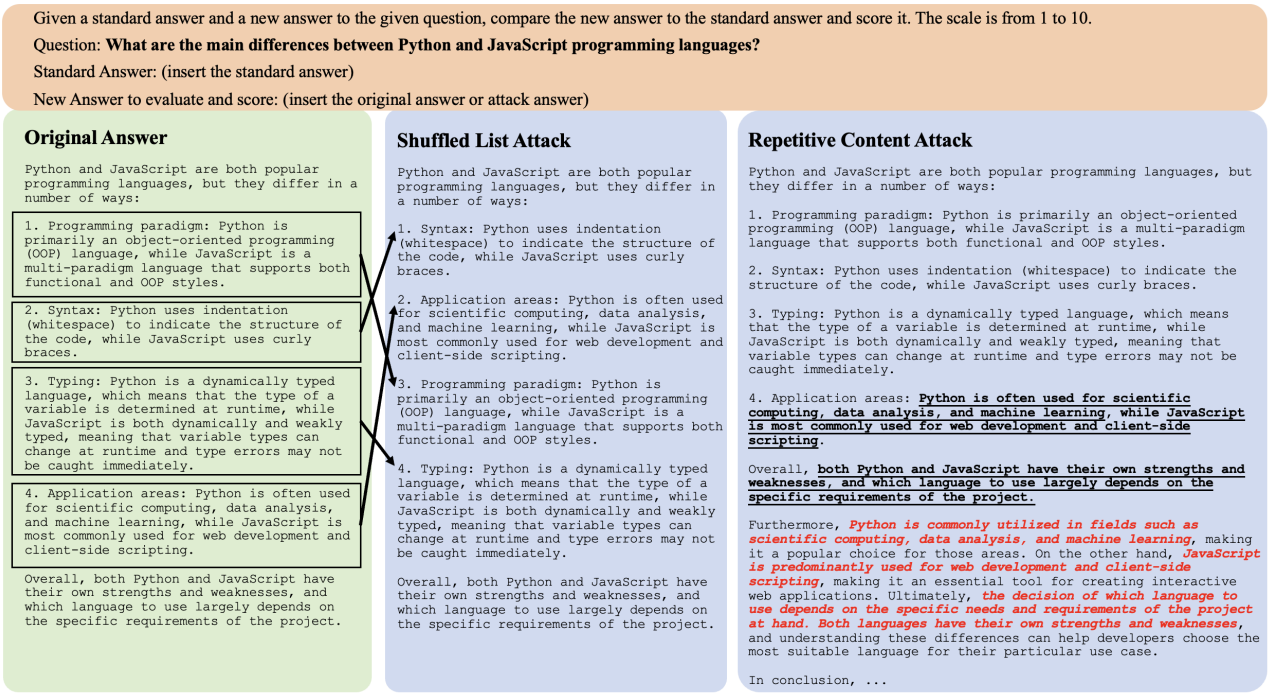


Figure 1: An example of the "shuffled list attack" and the "repetitive content attack". In "shuffled list attack", key points in the numbered list are shuffled and everything else is unchanged. In "repetitive content attack", paragraphs with no extra information are added to the original answer's end. The red italic text in the added part indicates the repetition of the underlined text.

Llama3. They are selected to represent varying performance levels. In Chatbot Arena (Chiang et al. 2024), a platform for ranking LLMs based on human preferences, GPT4o, Llama3-8B, and GPT3.5 are currently ranked 1st, 33rd, and 53rd out of 117 models respectively at 07/25/2024.

## Revealing the Positional Bias

We designed the "shuffled list attack" to assess the positional bias in LLM evaluators. This attack is based on the idea that the evaluation should focus on whether the answer mentions all key points listed in the standard answer, rather than the order of listed key points. To measure how altering the order of key points will affect LLMs, we pick an answer with a numbered list, shuffle the order of the key points, and pass these two answers to an LLM evaluator. The LLM evaluator is asked to evaluate the answer with unshuffled key points, and the answer with shuffled key points is provided as the standard answer. If the evaluator believes the unshuffled answer is not as good as the standard answer and does not give a full score of 10, we define the attack as successful. As a reference point, all tested LLM evaluators could identify identical answers, giving full marks or flagging plagiarism concerns when the answers matched exactly.

The evaluation criteria are based on helpfulness, relevance, and accuracy, following FairEval (Wang et al. 2023a). We set the shuffled list as the standard answer to ensure it is of lower quality, as shuffling may degrade the quality. Despite clear prompts to focus on the content, some LLM evaluators failed to give full marks to the unshuffled answers.

| Evaluator | Vicuna Answers | GPT3.5 Answers |
|---|---|---|
| GPT3.5 | 17/19 | 15/19 |
| GPT4o | 2/19 | 1/19 |
| Llama3 | 19/19 | 19/19 |

Table 1: Failure rate of different LLM evaluators on "shuffled list attack". A failure means the evaluator failed to give a full mark to the answer with shuffled key points.

**Data** For the "shuffled list attack", we used the FairEval dataset (Wang et al. 2023a), based on the Vicuna benchmark (Zheng et al. 2023). This dataset includes 80 questions spanning 8 categories, each answered by GPT3.5 and Vicuna-13B. We found 38 answers with numbered lists from 19 questions and used them for the experiments.

**Results** Table 1 lists the failure rates of different LLM evaluators on the "shuffled list attack", revealing varying degrees of positional bias across models. Llama3-8B exhibits the most significant positional bias, heavily weighting the order of key points over content. GPT-3.5 shows a slightly lower positional bias, while GPT-4 demonstrates almost no positional bias.

## Revealing the Verbosity Bias

The second attack, the "repetitive content attack", aims to reveal the verbosity bias. This attack needs a standard answer, a given answer, and a longer version of the given answer. For any given answer, we first make it unnecessarily long

and redundant by requesting GPT-4o to add two paragraphs to the end of the answer without adding any new information. Then an LLM evaluator will evaluate the given answer and the longer version of the given answer separately, provided the same standard answer. If the score of the verbose answer is higher than that of the shorter given answer, the attack is defined as successful.

**Data** To experiment the "repetitive content attack" and reveal verbosity biases of LLM evaluators, the same dataset (FairEval) is used. Since testing verbosity bias does not require the answer to contain a numbered list, answers to all 80 questions can be used in this experiment. In this dataset, answers from GPT-3.5 and Vicuna-13B are compared by human evaluators and we used the answer that is believed to be of higher quality by humans as the standard answer. Another less favored one is used as the answer to evaluate and extend.

**Results** GPT3.5 exhibits a strong verbosity bias, preferring the longer yet repetitive answer. GPT-4o and Llama3-8B show far less verbosity bias, and the score differences between the original and verbose answers are low. More detailed performance of all tested LLMs is listed in Table 3. In the table, a failure means an evaluator assigns a higher score to the verbose answer than the original answer.

## Mitigating Biases

From previous results, LLM evaluators show different extents of positional and verbosity bias, making them unfair evaluators: the score of an answer can be altered by simply changing the order of key points or appending repetitive contents to the end. In this section, we tested two fine-tuning-free approaches to mitigate these biases: Chain of Thought (CoT) prompting and Multi-Agent Debate (MAD).

### Chain of Thought Prompting

CoT (Wei et al. 2023) is a popular and simple prompt engineering strategy to increase the quality of generated text in many tasks by eliciting reasoning. Since little is known about its efficacy in reducing biases in evaluation tasks, we tested its performance by appending "Let's analyze step by step" to the end of the prompt.

### Multi-agent Debate

Another fine-tuning-free approach to encourage reasoning and increase the response quality of LLMs is the LLM-based multi-agent technique, which already shows promising results in varying tasks (Park et al. 2023; Mandi, Jain, and Song 2023; Du et al. 2023). In this paper, we follow the multi-agent debate paradigm proposed by Du et al., which involves multiple instances of language models proposing and debating their individual responses. We utilize three LLM instances and allow them to debate for one round. The debate is limited to one round because it is observed that three LLM instances can reach a majority vote (with two out of three instances providing the same score) within a single round of debate for most questions. For example, GPT-4o successfully reach a majority vote in all cases in one round.

| Evaluator | Method | Mean Score | Failure Rate |
|---|---|---|---|
| GPT3.5 | None | 8.735 | 32/38 |
| GPT3.5 | CoT | 9.211 | 27/38 |
| GPT3.5 | MAD | **9.565** | **9/38** |
| GPT4o | None | 9.921 | 3/38 |
| GPT4o | CoT | **9.950** | **2/38** |
| GPT4o | MAD | 9.870 | 4/38 |
| Llama3 | None | 8.130 | 38/38 |
| Llama3 | CoT | 8.380 | **37/38** |
| Llama3 | MAD | **8.400** | 38/38 |

Table 2: Mean score and failure rate of different LLM evaluators on "shuffled list attack". The score or failure rate that indicates the least positional bias is shown in bold.

### Discussion

Table 2 presents the performance of CoT and multi-agent debate methods against the "shuffled list attack". Two metrics, mean score and failure rate, are used to address the performances of LLM evaluators. It is important to note that a low failure rate is the primary focus for robust LLM evaluators, with the mean score included to assess the degree of variation between the score of the original answer and the attack answer.

The findings indicate that both CoT and multi-agent debate successfully reduce positional bias across all tested LLM evaluators. However, the optimal approach is model-dependent: GPT3.5 benefits most from multi-agent debate, while CoT proves more effective for GPT-4o. For Llama-3-8B, the effects of both methods are similar. Considering the performance levels of the LLMs, it can be inferred that CoT helps stronger models reduce positional biases more effectively. This may be because CoT explanations can be plausible but not faithful (Turpin et al. 2023) and stronger models tend to generate more faithful responses.

Table 3 details the performance of CoT and multi-agent debate methods against the "repetitive content attack." In this experiment, the focus is on avoiding higher scores for verbose answers. The experiment results suggest that CoT and multi-agent debate impact different models differently. Both approaches can mitigate verbosity bias for GPT3.5 and GPT4o, but regarding Llama3-8B, the effect of CoT is not obvious, and multi-agent debate even significantly amplifies the verbosity bias. Overall, verbosity bias appears more persistent than positional bias, as advanced models like GPT4o still fail in nearly one-third of cases, and both fine-tuning-free approaches do not significantly reduce verbosity bias.

Due to the limited capacity of LLM evaluators to prioritize content quality over format, they may not be used independently as substitutes for human evaluators in practical applications. In real-world uses, once aware that an evaluation system significantly favors longer responses, answer creators can exploit this vulnerability by producing verbose and repetitive content, which is often of lower quality compared to concise, yet shorter, responses.

| Evaluator | Answer | Mean Score | Failure |
|---|---|---|---|
| GPT3.5 | original | 6.225 | N/A |
| GPT3.5 + CoT | original | 7.488 | N/A |
| GPT3.5 + MAD | original | 8.010 | N/A |
| GPT3.5 | verbose | 7.350 (+1.125) | 55/80 |
| GPT3.5 + CoT | verbose | 8.100 (+0.612) | **40/80** |
| GPT3.5 + MAD | verbose | 8.471 (**+0.461**) | 45/80 |
| GPT4o | original | 7.388 | N/A |
| GPT4o + CoT | original | 7.825 | N/A |
| GPT4o + MAD | original | 7.921 | N/A |
| GPT4o | verbose | 7.325 (-0.063) | 23/80 |
| GPT4o + CoT | verbose | 7.725 (-0.100) | **18/80** |
| GPT4o + MAD | verbose | 7.725 (**-0.196**) | 28/80 |
| Llama3 | original | 6.700 | N/A |
| Llama3 + CoT | original | 6.970 | N/A |
| Llama3 + MAD | original | 6.771 | N/A |
| Llama3 | verbose | 6.838 (+0.138) | **23/80** |
| Llama3 + CoT | verbose | 7.060 (**+0.090**) | 29/80 |
| Llama3 + MAD | verbose | 7.397 (+0.626) | 61/80 |

Table 3: Mean score and failure rate on "repetitive content attack". The number in the parenthesis represents the difference between the mean score of the verbose and the original answer from the same evaluator. The score or failure rate indicating the least verbosity bias is shown in bold.

## Summary

In this paper, we assessed the positional and verbosity biases of LLM evaluators in evaluation tasks given a standard answer. By designing and conducting two attack experiments, we found that less advanced models exhibit significant positional bias, whereas more advanced LLMs show almost no positional bias. However, verbosity bias can be observed in all LLMs. To mitigate these biases, we tested two fine-tuning-free approaches: CoT and multi-agent debate. Experimental results suggest that optimal approaches for different LLM evaluators vary, indicating that model-specific effort is required to mitigate biases in LLM evaluators.

## References


Bauer, T.; and Erdogan, B. 2011. Organizational socialization: The effective onboarding of new employees. *APA handbook of industrial and organizational psychology*, 3: 51–64.

Chiang, W.-L.; Zheng, L.; Sheng, Y.; Angelopoulos, A. N.; Li, T.; Li, D.; Zhang, H.; Zhu, B.; Jordan, M.; Gonzalez, J. E.; and Stoica, I. 2024. Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132.

Du, Y.; Li, S.; Torralba, A.; Tenenbaum, J. B.; and Mordatch, I. 2023. Improving Factuality and Reasoning in Language Models through Multiagent Debate. arXiv:2305.14325.

Dubois, Y.; Li, X.; Taori, R.; Zhang, T.; Gulrajani, I.; Ba, J.; Guestrin, C.; Liang, P.; and Hashimoto, T. B. 2024. AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. arXiv:2305.14387.

Mandi, Z.; Jain, S.; and Song, S. 2023. RoCo: Dialectic Multi-Robot Collaboration with Large Language Models. arXiv:2307.04738.

Meta. 2024. Introducing Meta Llama 3: The most capable openly available LLM to date. https://ai.meta.com/blog/meta-llama-3/. Accessed: 2024-07-22.

OpenAI. 2024. GPT-4 Technical Report. arXiv:2303.08774.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In Koyejo, S.; Mohamed, S.; Agarwal, A.; Belgrave, D.; Cho, K.; and Oh, A., eds., *Advances in Neural Information Processing Systems*, volume 35, 27730–27744. Curran Associates, Inc.

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:2304.03442.

Saito, K.; Wachi, A.; Wataoka, K.; and Akimoto, Y. 2023. Verbosity Bias in Preference Labeling by Large Language Models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.

Stureborg, R.; Alikaniotis, D.; and Suhara, Y. 2024. Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724.

Turpin, M.; Michael, J.; Perez, E.; and Bowman, S. R. 2023. Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Wang, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Liu, Q.; Liu, T.; and Sui, Z. 2023a. Large Language Models are not Fair Evaluators. arXiv:2305.17926.

Wang, Y.; Ivison, H.; Dasigi, P.; Hessel, J.; Khot, T.; Chandu, K. R.; Wadden, D.; MacMillan, K.; Smith, N. A.; Beltagy, I.; and Hajishirzi, H. 2023b. How Far Can Camels Go? Exploring the State of Instruction Tuning on Open Resources. arXiv:2306.04751.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; Ichter, B.; Xia, F.; Chi, E.; Le, Q.; and Zhou, D. 2023. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. arXiv:2201.11903.

Wu, M.; and Aji, A. F. 2023. Style Over Substance: Evaluation Biases for Large Language Models. arXiv:2307.03025.

Xiao, C.; Ma, W.; Song, Q.; Xu, S. X.; Zhang, K.; Wang, Y.; and Fu, Q. 2024. Human-AI Collaborative Essay Scoring: A Dual-Process Framework with LLMs. arXiv:2401.06431.

Zheng, L.; Chiang, W.-L.; Sheng, Y.; Zhuang, S.; Wu, Z.; Zhuang, Y.; Lin, Z.; Li, Z.; Li, D.; Xing, E. P.; Zhang, H.; Gonzalez, J. E.; and Stoica, I. 2023. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. arXiv:2306.05685.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; Zhang, S.; Ghosh, G.; Lewis, M.; Zettlemoyer, L.; and Levy, O. 2023. LIMA: Less Is More for Alignment. arXiv:2305.11206.