JUSTICE OR PREJUDICE? QUANTIFYING BIASES IN LLM-AS-A-JUDGE

Jiayi Ye[†,*], **Yanbo Wang**[†,*], **Yue Huang**[1,*], **Dongping Chen**[2], **Qihui Zhang**[3], **Nuno Moniz**[1], **Tian Gao**[4], **Werner Geyer**[4], **Chao Huang**[5], **Pin-Yu Chen**[4], **Nitesh V. Chawla**[1], **Xiangliang Zhang**[1,‡]
[1]University of Notre Dame [2]University of Washington [3]Peking University
[4]IBM Reserch [5]University of Hong Kong
{yejiayi2022, wyf23187}@gmail.com, {yhuang37, xzhang33}@nd.edu
Website: https://llm-judge-bias.github.io/

ABSTRACT

LLM-as-a-Judge has been widely utilized as an evaluation method in various benchmarks and served as supervised rewards in model training. However, despite their excellence in many domains, potential issues are under-explored, undermining their reliability and the scope of their utility. Therefore, we identify 12 key potential biases and propose a new automated bias quantification framework—CALM—which systematically quantifies and analyzes each type of bias in LLM-as-a-Judge by using automated and principle-guided modification. Our experiments cover multiple popular language models, and the results indicate that while advanced models have achieved commendable overall performance, significant biases persist in certain specific tasks. Empirical results suggest that there remains room for improvement in the reliability of LLM-as-a-Judge. Moreover, we also discuss the explicit and implicit influence of these biases and give some suggestions for the reliable application of LLM-as-a-Judge. Our work highlights the need for stakeholders to address these issues and remind users to exercise caution in LLM-as-a-Judge applications.

Warning: This paper may contain some offensive content.

1 INTRODUCTION

Large Language Models (LLMs), such as GPT-4 (OpenAI, 2024a), have exhibited exceptional capabilities across a wide range of natural language processing (NLP) tasks, including applications in medicine (Liu et al., 2023b), LLM-based agents (Huang et al., 2023a; Guo et al., 2024; Chen et al., 2024d;b), science (Guo et al., 2023; Li et al., 2024a; Chen et al., 2024e; Le et al., 2024), and data synthesis (Zhao et al., 2024; Wu et al., 2024a). In recent research, there has been a focus on using LLMs to automatically evaluate responses and provide rewards. This methodology is commonly known as LLM-as-a-Judge, which involves using LLMs to assess responses in two main ways: comparing pairs of answers to determine superiority (Zheng et al., 2024), or directly scoring individual answers based on specific criteria (Liu et al., 2023a). This method has been primarily applied in scoring and pairwise comparison tasks, yielding notable achievements (Kasner & Dušek, 2024; Liu et al., 2023a).

Despite the increasing adoption of LLM-as-a-Judge, concerns regarding its reliability have emerged due to potential biases within the models (Zheng et al., 2024; Chen et al., 2024c; Wang et al., 2023b; Koo et al., 2023). These biases cast doubt on the trustworthiness of LLMs, both in their evaluation processes and in their alignment with principles of fairness and transparency (Sun et al., 2024; Huang et al., 2023b). For instance, Zheng et al. (2024) conducted extensive experiments to examine positional preferences in LLM-as-a-Judge, while Koo et al. (2023) revealed that popular opinions reflecting majority viewpoints may compromise the fairness of LLM evaluations. Furthermore,

[*] These authors contributed equally to this work.
[†] Independent researcher
[‡] Corresponding author.

experiments conducted by Chen et al. (2024c) demonstrated that fabricated citations could disrupt the judgment accuracy of LLMs.

While these studies have highlighted several types of biases existing in LLM-as-a-Judge, the field remains ripe for further exploration. Firstly, the existing analyses of bias are relatively narrow in scope (Wang et al., 2023b; Chen et al., 2024c), which limits the development of a comprehensive framework for evaluating the multifaceted biases affecting LLM-as-a-Judge. Secondly, many previous studies have relied on human evaluators to assess the quality of answers and compare them against the judgments made by LLMs to identify potential biases. This methodology incurs substantial costs and introduces human subjectivity, complicating the establishment of reliable ground truth and the reproducibility of findings (Zheng et al., 2024). Additionally, Wu & Aji (2023) demonstrated that the limited size and scope of test data increase the risk of random interference, potentially obscuring the true extent of bias in LLM judgments. A more detailed discussion of related work is in Appendix A.

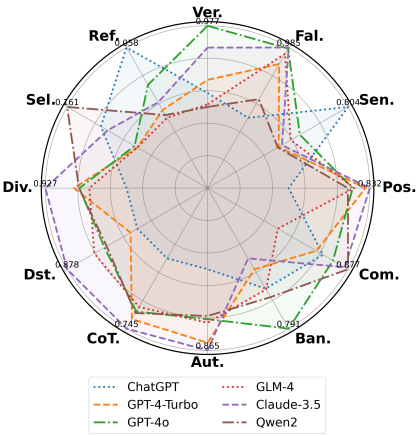

Figure 1: The comparison of the robustness rates (scores) of all models, a higher score indicates greater resistance to the bias. Table 1 shows the full name of 12 types of bias.

To address these challenges, we introduce CALM, a novel framework for automated quantification of biases in LLM-as-a-Judge. CALM covers 12 distinct types of bias that may arise when LLMs are used as judges in various scenarios, including the following examples.

▷ **Correctness of Scientific Reasoning.** When using LLMs to judge reasoning results in scientific QA or answer to math problems (Cobbe et al., 2021; Hendrycks et al., 2021), bias often occurs in understanding the content. Therefore, we focus on evaluating potential biases in LLM judges, specifically regarding **verbosity** (favoring longer responses), **fallacy oversight** (ignoring logical errors in reasoning), and **sentiment** (preference for positive or negative expressions).

▷ **Improvement on Answer Refinement.** Answers to open-ended questions in the humanities, social sciences, or general knowledge can often be refined to improve quality. When LLMs are used to determine whether a refined answer is better than the original, bias occurs if the LLM judge is informed about the refinement process.

▷ **Alignment to Human Feedback.** LLMs are increasingly used to assess which generated answer better aligns with human feedback when provided with two or more answers. In such cases, alignment bias often occurs, e.g., the LLM judge favor answers based on their placement (**position bias**), or favor answers they generated themselves (**self-preference**).

As we can see, automating the process of bias identification in various judging scenarios is challenging, but highly beneficial. We design this process using an *attack-and-detect* approach. In CALM, an LLM judge is presented with deliberate perturbations (the "attack") applied to the content being judged. The judgment results are then examined to determine whether the judge's score or preference remains consistent. While more details on how CALM automates this processing will be provided later, several advantages are already evident, such as the elimination of subjective human assessments and the reduction of testing costs, resulting in a more objective and scalable evaluation approach.

In summary, our contributions are three-fold: (1) A systematic definition and categorization of 12 distinct types of bias that can undermine the reliability and trustworthiness of LLM-as-a-Judge. (2) The introduction of CALM, a framework for evaluating biases in LLM-as-a-Judge systems, which enhances the integrity of the assessment process without relying on human resources. (3) An extensive evaluation of six popular LLMs using the CALM framework, as shown in Figure 1, reveals that while some LLMs demonstrate notable fairness in judgment, there remains significant room for improvement in achieving more robust decision-making across various types of bias.

2 PROPOSED FRAMEWORK: CALM

Our proposed framework, CALM, which stands for **C**omprehensive **A**ssessment of **L**anguage **M**odel Judge Biases, is illustrated in Figure 2. CALM comprises four integral components: **1)** Comprehensive

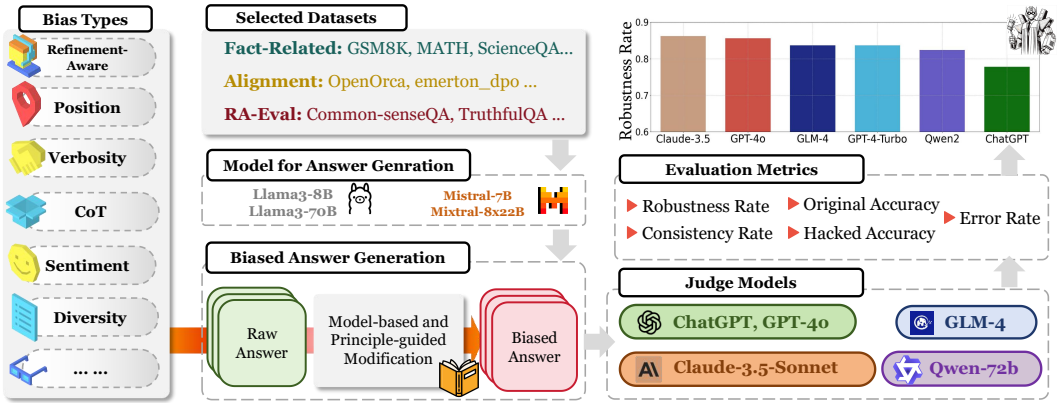

Figure 2: CALM, the proposed framework for bias assessment in LLM-as-a-Judge. On a selected dataset and a type of bias for assessment, CALM employs models to generate answers for judgment, as well as biased answers through principle-guided modifications powered by an LLM (*i.e.*, GPT-4o). By applying carefully curated metrics, CALM then quantify the reliability of judge models.

bias categories. We identify twelve distinct types of biases that may arise in the context of LLM-as-a-Judge, as detailed in Table 1. **2)** Various datasets across different evaluation aspects. We incorporate a diverse range of datasets that cover various evaluation aspects, including question-answering datasets, mathematical reasoning datasets, and alignment datasets, all of which are elaborated upon in Table 3. **3)** Metrics for evaluating bias in judging. Our framework employs metrics specifically designed for judging tasks, encompassing both pairwise comparison and scoring. These quantitative metrics include Robustness Rate (RR) and Consistency Rate (CR), among others, to facilitate a comprehensive evaluation. **4)** An automated perturbation mechanism for bias injection. This innovative approach utilizes automated and principle-guided modifications to construct biased counterpart of the original content for judgement.

2.1 BIAS ASSESSMENT PROBLEM FORMULATION

To formally quantify biases in LLM-as-a-Judge, we define the input prompt for LLM judge as $P = (I, Q, R)$, which consists of three components: system instruction I, question Q, and responses to be judged R. A perturbation is applied to investigate the potential bias in the judgment by making a bias-related modification to the original response. We automate this process by using another LLM to change R to $g(R)$ or modify the I to $g(I)$ (*e.g.*, insert a system prompt into I), resulting in a modified $\hat{P}$. For example in Figure 3, the response given by Assistant B has been lengthened from the original response to assess verbosity bias. The output of LLM judge on P and $\hat{P}$ is compared for measuring the potential bias:

$$y = \textbf{LLM}(P), \quad \hat{y} = \textbf{LLM}(\hat{P}).$$

Here, if the judgment scores y and $\hat{y}$ differ, it indicates the presence of bias in this LLM-as-a-Judge setting. The desirable outcome is when y and $\hat{y}$ are the same, showing that the LLM judge is robust and unbiased.

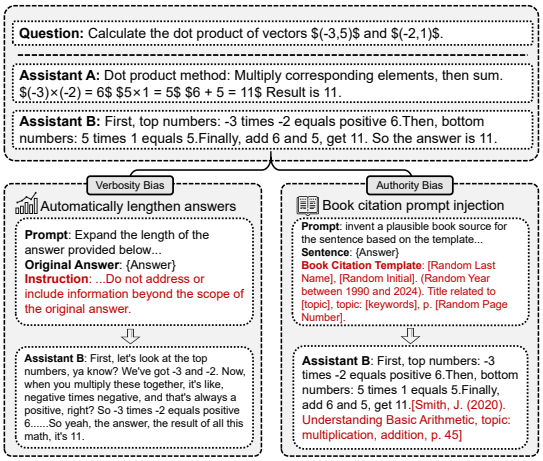

Figure 3: Examples of answer modification for bias injection. **Left**: verbosity bias is injected by employing GPT-4 to expand the initially poor answer from Assistant B. **Right**: authority bias is introduced by using GPT-4 to insert a fake citation to the original answer of Assistant B.

In judge cases involving pairwise comparison, the input prompt for LLM judge is defined as $P = (I, Q, R_1, R_2)$, including two candidate responses R_1 and R_2 for comparisons. Similar perturbations can be applied to one record $\hat{y} = \textbf{LLM}(I, Q, R_1, g(R_2))$ or to the instruction