importance of fairness in evaluating LLMs. Fairness of LLMs is defined as the ethical principle of ensuring that LLMs are designed, trained, and deployed in ways that do not lead to biased or discriminatory outcomes and that they treat all users and groups equitably (Sun et al., 2024). The imbalance in pre-training data can lead to imbalances during model training (Liu et al., 2024), resulting in biases against certain demographic groups, such as different genders (Wan et al., 2023), ages (Macnicol, 2006), and various languages (Jiao et al., 2023; Bang et al., 2023). Consequently, the fairness of LLMs has a significant impact on the trustworthiness of LLM-as-a-Judge.

A.3 Biases in LLM-as-a-Judge Application

Recent research has identified various cognitive biases that influence the evaluation of LLMs. Some studies (Zheng et al., 2024; Shi et al., 2024b; Wang et al., 2023b) discuss biases such as position bias, verbosity bias, and self-enhancement bias. Another study (Koo et al., 2023) highlights order bias, compassion-fade bias, and egocentric bias, along with salience bias, bandwagon-effect bias, and attentional bias. Further biases noted in additional research (Chen et al., 2024c; Stureborg et al., 2024) include fallacy-oversight bias, authority bias, and beauty bias. Recognizing these biases is essential for developing more objective and trustworthy LLM evaluation methods.

B Details of Bias Types

- ▷ **Position bias**: LLMs may favor responses based on their position in the input. This bias affects how the model processes information, and following Zheng et al. (2024), we extend the analysis to scenarios involving more than two responses.
- ▷ **Verbosity bias**: LLM-as-a-Judge may be biased towards longer responses. We evaluate the impact of different length ratios between responses on judgment outcomes, as indicated by Zheng et al. (2024).
- ▷ **Compassion-fade bias**: LLM judgments may be influenced by the anonymity of model names. We investigate how various model names and anonymization strategies impact judgments, inspired by the observations of Koo et al. (2023).
- ▷ **Bandwagon-effect bias**: LLM-as-a-Judge may be biased by the presence of majority opinions. We assess this by setting varying percentages (60%, 70%, 80%, and 90%) of majority opinions in the system instruction, following Koo et al. (2023).
- ▷ **Distraction bias**: Introducing distractions could affect the judgments of both high-quality and low-quality model outputs. We extend previous work by Koo et al. (2023) to evaluate the impact of distractions in LLM decision-making. Experimental details are available in Appendix C.
- ▷ **Fallacy-oversight bias**: This bias relates to the LLM's ability to recognize and avoid logical fallacies. We develop tests to evaluate this ability across various types of fallacies, contributing to fair and accurate judgments, as discussed in Chen et al. (2024c).
- ▷ **Authority bias**: Authoritative references may sway LLM judgments. We assess this influence by incorporating three types of references—book citations, website references, and famous individuals' quotes—following the methodology of Chen et al. (2024c).
- ▷ **Sentiment bias**: LLMs may display preferences towards certain emotional tones in responses. We evaluate how sentiment influences judgments across emotional expressions such as cheerful, sad, angry, and fearful, as noted by Li & Sinnamon (2023).
- ▷ **Diversity bias**: Judgments may shift based on specific identity markers. We evaluate this bias by setting system instructions that assign six identity categories: Female, Black individuals, Homosexuals, Muslims, Refugees, and HIV patients, following the concept of identity impact.
- ▷ **Chain-of-Thought (CoT) bias**: LLM judgments can be affected by the presence of explicit reasoning steps. We compare evaluations of responses with and without chain-of-thought reasoning across different tasks, as suggested by Wei et al. (2023).
- ▷ **Self-enhancement bias**: This bias arises when LLMs favor their outputs as both generators and judges. To explore this, we include evaluations to measure the bias across different LLM architectures and scales, following Zheng et al. (2024) and Meng et al. (2024).
- ▷ **Refinement-aware bias**: LLMs may assign different scores to self-refined answers. We investigate this bias by comparing scores in three situations: original unrefined answer, refined answer, and refined answer with conversation history, as explored by Xu et al. (2024).

C Details of Bias Evaluation

We will introduce the detailed evaluation process of each bias.

- ▷ **Position bias**: To investigate the impact of position bias, we tested the effect of changing the order of answers when there are two, three, and four answers. We employed a rotational permutation order to shuffle the answers and conducted evaluations to observe if the results remained consistent. If the selection of any answer was inconsistent across different orders, it was considered an inconsistency. For specific evaluation prompts, refer to Figure 13, Figure 14, and Figure 15.
- ▷ **Verbosity bias**: We employed GPT-4-Turbo to lengthen the worse quality answers in the fact-related dataset. The lengthened prompt used for this process can be referenced in Figure 18. Our objective was to increase the verbosity of these answers while preserving their essential content. Subsequently, we conducted a comparative evaluation, wherein the models reassessed both the original answers pair and the elongated answers pair.
- ▷ **Self-enhancement bias**: Building upon the single-answer scoring prompt proposed by Zheng et al. (2024), we developed a scoring prompt Figure 26. In our experimental design, each model was tasked with answering the questions and subsequently evaluating both its responses and those generated by other models, without prior knowledge of the authorship of any given answer. Following the completion of all evaluations, we aggregated and analyzed the scoring results to draw our conclusions.
- ▷ **Compassion-fade bias**: We explicitly reveal the identities of all models under judgment. The specific details of this approach are delineated in the evaluation prompt presented in Figure 20. Subsequently, we conducted a comparative analysis between these named-condition results and those obtained under anonymized settings.
- ▷ **Bandwagon-effect bias**: We added the following sentence to the system prompt: "{number}% of people believe that {chosen_model} is better in this question" to test whether the model is influenced by popular opinion. For specifics, refer to prompt Figure 21.
- ▷ **Distraction bias**: We inserted a meaningless statement into the System Prompt, such as "Assistant A loves eating pasta," to test whether the LLM Judge is influenced by it. For specifics, refer to prompt Figure 29. We applied distinct distractions to each assistant and subsequently analyzed the robustness rate of the judge model's evaluations in scenarios where the distracted Assistant provided either the better or worse quality response.
- ▷ **Fallacy-oversight bias**: We modified the better quality answers in the fact-related dataset using GPT-4-Turbo to make their logic completely chaotic. The prompt used can be referenced in Figure 19. We then had the models re-evaluate these answer pairs. If a model's evaluation result was inconsistent with its original assessment of the answer pair, we considered it a correct judgment (because the original worse quality answer is still better than the logically chaotic better quality answer). Otherwise, it was deemed an incorrect judgment.
- ▷ **Authority bias**: Using GPT-4-Turbo, we generated three types of fake citation information related to the answers: *URLs*, *famous quotes*, and *book references*. For specifics on the prompts used for the generation, refer to Figure 24, Figure 25, and Figure 23. These citations were then injected into the answers, as demonstrated in Figure 22.
- ▷ **Sentiment bias**: We modified the better quality answers in the fact-related dataset using GPT-4-Turbo to incorporate one of the four emotions: *cheerful*, *sad*, *angry*, or *fear*. The prompt can be referenced in Figure 27. Then, we had the models judge these answers again to observe whether the results were consistent with the original judgment.
- ▷ **Diversity bias**: For diversity bias, we selected six identities that may be subject to discrimination: Homosexual, Black, Female, HIV Positive, Refugees, and Muslim believers. These identities were then injected into the system prompt for judgment to observe their impact on evaluations. For more details, refer to prompt Figure 28.
- ▷ **CoT bias**: We modified a version of the Prompt based on the original Chain of Thought prompt from (Zheng et al., 2024), which can be referenced in Figure 16. Under the condition that all other factors remain unchanged, we conducted judgment on the fact-related dataset to observe whether the results changed.
- ▷ **Refinement-aware bias**: In the Refinement-aware eval dataset, we first have the model answer these questions. Then, using prompt Figure 30, we enable the model to refine its previously given answers. Finally, the model evaluates the pre-refinement, post-refinement, and refined-with-history answers, and we compile the results. For specifics on the evaluation prompt, refer to Figure 31. We can reference Figure 10 as an illustrative example.

D Detailed Results

In Figure 4, we provide a comparative chart of the robustness rate for all biases, which allows for a horizontal comparison of the differences in resilience to interference among all models, with the dashed line representing the consistency rate. In Table 7, the detailed experimental results for each type of bias are presented.

- ▷ **Position bias.** We present the robustness rate of different judge models when faced with pairwise comparisons in Table 7, and in Figure 6 we show the robustness rate of all judge models when presented with multiple answer options.
- ▷ **Verbosity bias.** In Figure 6, we illustrate the relationship between different ratios of answer expansion lengths and model robustness rate.
- ▷ **Self-Enhancement bias.** In Figure 5, we present a heat map of Z-score normalized scores for each model (due to ChatGPT's relatively weak performance, the scores given to it by the remaining models are not high enough, resulting in the first column lacking reference value). Additionally, in Figure 7, we display the $\text{ErrorRate}_{\text{SE}}$ metric for each judge model.
- ▷ **Bandwagon-Effect bias.** In Table 7 and Figure 6, we present the impact of varying percentages of public opinion on the judge models. The experimental results indicate that the influence on each model is not uniform and does not demonstrate a statistical pattern.
- ▷ **Distraction bias.** In Figure 7 and Table 7, we present the robustness rate performance of all judge models after introducing irrelevant content as interference for both high-quality and low-quality answers originally present in the dataset.
- ▷ **Authority bias.** In Table 7, we present the impact of different types of fake references on the judge model. As shown in Figure 6, quote and book-type references strongly influence most models.
- ▷ **Sentiment bias.** In Figure 8, we display the Acc_{hack} and robustness rate performance of judge models with three different emotions added to high-quality and low-quality answers in the dataset. Our findings indicate that most models do not favor emotionally charged expressions.
- ▷ **CoT bias.** In Figure 7 and Table 7, we present the accuracy metrics Acc_{ori} and Acc_{hack} before and after applying CoT. As shown in the figure, for most models, the application of CoT techniques can improve judgment accuracy.
- ▷ **Refinement-aware bias.** In Figure 7, we present the $\text{ErrorRate}_{\text{RA}}$ metric for different judge models.
- ▷ **Diversity bias.** We show the changes in various metrics of the judge model under the influence of different minority groups in Figure 8 and Table 7.

E Case Study

From Figure 9, 10, 11, 12, we enumerated various actual manifestations of bias and conducted a detailed analysis.

F Prompt Template

From Figure 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, we provide detailed prompt templates we used in the experiments.