

Judging with Many Minds: Do More Perspectives Mean Less Prejudice? On Bias Amplification and Resistance in Multi-Agent Based LLM-as-Judge

Chiyu Ma^{*}, Enpei Zhang^{*}, Yilun Zhao, Wenjun Liu, Yaning Jia,
Peijun Qing, Lin Shi, Arman Cohan, Yujun Yan, Soroush Vosoughi

Dartmouth College, Yale University

{chiyu.ma.gr, enpei.zhang.gr}@dartmouth.edu, yilun.zhao@yale.edu

Abstract

LLM-as-Judge has emerged as a scalable alternative to human evaluation, enabling large language models (LLMs) to provide reward signals in trainings. While recent work has explored multi-agent extensions such as multi-agent debate and meta-judging to enhance evaluation quality, the question of how intrinsic biases manifest in these settings remains under-explored. In this study, we conduct a systematic analysis of four diverse bias types: position bias, verbosity bias, chain-of-thought bias, and bandwagon bias. We evaluate these biases across two widely adopted multi-agent LLM-as-Judge frameworks: *Multi-Agent-Debate* and *LLM-as-Meta-Judge*. **Our results show that debate framework amplifies biases sharply after the initial debate, and this increased bias is sustained in subsequent rounds, while meta-judge approaches exhibit greater resistance.** We further investigate the incorporation of PINE, a leading single-agent debiasing method, as a bias-free agent within these systems. **The results reveal that this bias-free agent effectively reduces biases in debate settings but provides less benefit in meta-judge scenarios.** Our work provides a comprehensive study of bias behavior in multi-agent LLM-as-Judge systems and highlights the need for targeted bias mitigation strategies in collaborative evaluation settings. Our code and experimental data can be found at https://github.com/HenrymachiYu/Multi_Agent_Judge_Bias.

1 Introduction

Large language models (LLMs) have achieved remarkable performance gains in recent years (Kaplan et al., 2020; Hoffmann et al., 2022; Achiam et al., 2023), however, further progress is increasingly constrained by the limited availability of high-quality data and the escalating costs of data acquisition and annotation (Villalobos et al., 2024; Shen

^{*}Denotes equal contribution

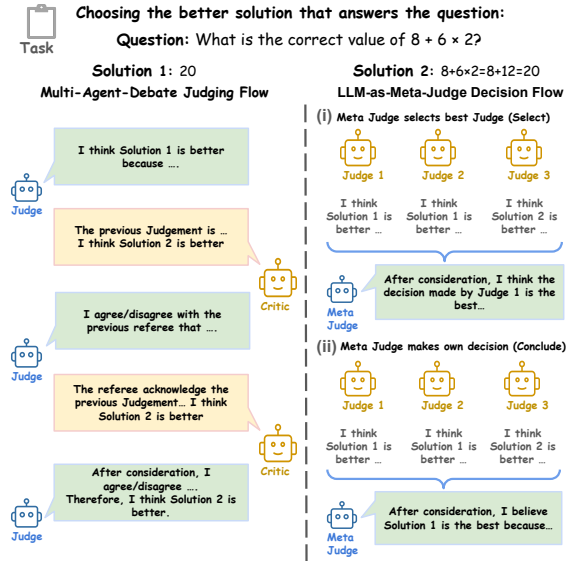


Figure 1: Demonstration of the multi-agent systems analyzed in our study under the LLM-as-Judge paradigm. *Top*: Given responses from two different assistants to the same prompt, the Judge model is tasked with selecting the superior response. *Left*: In the Multi-Agent-Debate framework, a judge provides an initial judgment and then debates with a referee. *Right*: In the Meta-Judge framework, the meta-judge either selects the best judgment or generates its own judgment based on the outputs of multiple judges.

et al., 2025). As task complexity grows, generating the massive amounts of labeled data required through human effort alone has become impractical. To address this bottleneck, recent research has explored using LLMs themselves as sources of supervision. Notably, Burns et al. observed that supervision from weaker models has potential to help elicit the capabilities of stronger models. These findings have further motivated the shift toward automated, scalable feedback mechanisms that reduce reliance on costly human annotation. Within this line of work, the LLM-as-Judge paradigm (Chan et al., 2023; Gu et al., 2024; Chen et al., 2024; Li et al., 2025b; Saha et al., 2025), which uses LLMs

to automatically evaluate responses and provide reward signals, has attracted particular interest due to its adaptability to custom evaluation criteria and its ability to generate human-understandable explanations of decisions.

Building upon this paradigm, recent efforts have begun to investigate whether incorporating multi-agent frameworks—which have proven effective in complex reasoning tasks (Liang et al., 2023; Du et al., 2023; Hong et al., 2023; Chen et al., 2023)—can further enhance the capabilities of LLM-as-Judge pipelines. Structures such as multi-agent debate (Chan et al., 2023; Du et al., 2023; Liu et al., 2024c) and meta-reasoning (Wu et al., 2024; Li et al., 2025b) offer promising avenues for boosting evaluation quality. Despite these advances, concerns remain. Several studies reveal that even single-agent LLM judges have been shown to exhibit intrinsic biases, including position bias and verbosity bias, which can undermine the reliability of their evaluations (Huang et al., 2023b; Shi et al., 2024; Ye et al., 2024). However, there has been limited focused or systematic investigation into whether such biases persist, diminish, or are amplified in multi-agent settings.

In response to these concerns, this study systematically examines how intrinsic biases manifest in multi-agent LLM-as-Judge systems. As demonstrated in Figure 1, we investigate two widely adopted paradigms: LLM-as-Judge with *Multi-Agent-Debate* and *LLM-as-Meta-Judge*. For each paradigm, we also evaluate the performance of different LLMs as base evaluators. Our analysis focuses on four diverse types of bias: position bias, verbosity bias, chain-of-thought (CoT) bias, and bandwagon bias, ensuring a comprehensive examination across key dimensions of judgment behavior. Interestingly, our results show that intrinsic biases manifest differently across the two multi-agent frameworks. **The Multi-Agent-Debate framework amplifies biases sharply after the initial debate, and this increased bias is sustained in subsequent rounds. In contrast, the LLM-as-Meta-Judge approach exhibits greater resistance to these intrinsic biases.** When the meta-judge selects the best judgment from a set of candidates, the resulting bias levels are comparable to those observed in single-agent settings. However, when the meta-judge generates a new judgment based on the candidate judgments, a more substantial reduction in bias is achieved.

Additionally, we examine whether a leading

single-agent bias mitigation strategy can be effectively extended to improve evaluation reliability in more complex multi-agent settings, particularly when direct modifications to the judge model and its prompt are impractical or undesirable. To this end, we incorporate PINE (Wang et al., 2024b), a method shown to eliminate position bias through modifications to position embeddings and causal masking, as a *bias-free* agent within a multi-agent LLM-as-Judge system. **Our results show that introducing a bias-free agent into the debate setting yields consistent improvements as the number of conversation rounds increases. In contrast, its effect in the meta-judge framework remains limited, regardless of whether the meta-judge selects an existing judgment or generates its own.** These findings underscore the importance of accounting for bias dynamics in multi-agent evaluation and highlight the need for mitigation strategies specifically tailored to collaborative decision-making settings.

2 Related Work

2.1 LLM-as-Judges

Recent advances in automated model evaluation have established LLMs as viable alternatives to human annotators. Zheng et al. introduced the concept of LLM-as-Judge through MT-Bench, which evaluates alignment between LLM-generated judgments and human preferences on open-ended questions. This line of research has since expanded with benchmarks such as RewardBench (Lambert et al., 2024) and JudgeBench (Tan et al., 2024), which assess model alignment on more complex, reasoning-intensive tasks. In parallel, researchers have explored improved prompting strategies (Stahl et al., 2024; Dubois et al., 2023; Wang et al., 2025). Beyond prompting, fine-tuning on domain-specific datasets has been proposed to improve evaluation performance and mitigate biases (Zhu et al., 2023; Park et al., 2024; Liu et al., 2024b; Ye et al., 2025). Most recently, EvalPlanner (Saha et al., 2025) introduced a planning-then-judging framework that incorporates agentic structure into the LLM-as-Judge paradigm.

2.2 Multi-Agent LLM-as-Judges Framework

Motivated by the effectiveness of multi-agent systems in complex reasoning tasks (Liang et al., 2023; Du et al., 2023; Hong et al., 2023; Chen et al., 2023; Huang et al., 2023a; Liu et al., 2024c; Zhang et al.,

2025c), recent studies have applied multi-agent approaches to LLM-based evaluation. Multi-agent debate has emerged as a popular technique, with studies such as employing structured interactions among groups of LLM juries to reach consensus (Xu et al., 2023; Li et al., 2023a; Verga et al., 2024; Zhang et al., 2025b). In particular, ChatEval (Chan et al., 2023) aimed to more closely mirror the human evaluation process by facilitating structured debates among agents with diverse roles. In its setup, “general public” agents provided initial judgments, engaged in discussions, and delivered final verdicts, while “referee” agents acted as critics or experts to challenge and refine the reasoning process. Building on this design, our work adopts a similar structure in which judge agents interact with a designated critic who offers counterpoints intended to improve the quality of final decisions.

In parallel, meta-thinking has emerged as a promising direction for enhancing LLM-as-Judge systems (Shinn et al., 2023; Gao et al., 2024; Sui et al., 2025; Wan et al., 2025). Meta-thinking, often described as “thinking about thinking,” involves reflecting on and evaluating the reasoning process itself. The concept of a meta-judge was first introduced by Wu et al., where it served as a self-reward model during post-training. Moreover, Self-Rationalization (Trivedi et al., 2024) proposed an iterative fine-tuning approach in which the judge improves by comparing and refining its own rationales. More recently, Li et al. expanded this idea by assigning different LLMs to act as judges and a meta-judge, facilitating collaborative evaluation across diverse models. Our work adopts a similar design with two modes: the meta-judge either selects the best judgment from multiple judges or generates a new judgment generated from their outputs.

2.3 Intrinsic Biases and Mitigation

LLMs have been shown to exhibit a variety of intrinsic biases, including position bias, verbosity bias, self-enhancement bias, bandwagon bias, and chain-of-thought bias. These biases have been identified through large-scale evaluations and benchmark studies (Zheng et al., 2023; Wang et al., 2024a; Ye et al., 2024; Shi et al., 2024; Wan et al., 2024; Bao et al., 2024; Tian et al., 2025; Li et al., 2025a; Gulati et al., 2025). More recently, CALM (Ye et al., 2024) evaluated twelve types of biases in single-agent LLMs using carefully designed experiments. Our study closely follows the bias evalua-

tion procedures introduced by Ye et al.. However, there is limited understanding of how these biases behave in multi-agent settings. While Wu et al. observed that certain biases may intensify during iterative training, Li et al. found that collaborative interactions among agents can reduce specific biases such as position and self-enhancement. These contrasting observations highlight a critical gap in understanding bias dynamics in collaborative contexts, a gap our study seeks to address.

Among the various biases, position bias has received particular attention due to its substantial impact across a wide range of tasks. Mitigation strategies to date have included data augmentation combined with fine-tuning (Zhu et al., 2023; Park et al., 2024) and prompt alignment techniques (Li et al., 2023b). More recently, the PINE framework (Wang et al., 2024b) was introduced to completely eliminate position bias by modifying causal masking and positional embeddings. While these methods have demonstrated strong performance, their effectiveness in multi-agent contexts remains largely unexplored. Our work seeks to fill this gap by leveraging PINE as a *bias-free* agent to assess its impact on position bias in collaborative evaluation scenarios. This is particularly important in settings where modifying the prompts or internal structure of judge models is impractical or undesirable.

3 Method

3.1 Evaluative Settings

In this study, we adopt the score-based pairwise comparison setting to stick with that of ChatEval (Chan et al., 2023). To ensure consistency, we apply this evaluation setting across all multi-agent LLM-as-Judge experiments and bias assessments. Similar to the single LLM-as-Judge setting, each model in the multi-agent LLM-as-Judge framework is tasked with evaluating two candidate solutions generated by different LLMs for a given prompt, although with different roles. The judges assign numerical scores and provide feedback for each solution, with the higher-scored solution subsequently selected as superior. We assign the solution as a tie, when the scores are the same. For clarity, Figure 1 illustrates only the pairwise comparison process. In the actual study, however, each judge model in the framework are prompted to provide explicit scores for each solution. The framework then determines and outputs the final verdict. Importantly, the identities of the LLMs that produced the candi-

date solutions are not disclosed to the multi-agent judges.

3.2 Multi-Agent-Debate

In the Multi-Agent-Debate setting, we adopt and implement the “General Public and Critic” structure, as originally proposed in ChatEval (Chan et al., 2023), and widely used in several recent studies (Feng et al., 2024; Arif et al., 2024; Zhang et al., 2025a). The General Public agent is analogous to the judge model in the single-agent LLM-as-Judge setting. It is tasked with making an initial judgment, which includes the scores and reasoning for both solutions. The Critic agent, in contrast, first reads and comprehends the evaluation task and the judgment provided by the General Public agent. It then offers its critique of the initial judgment along with its own independent assessment. To mirror real-world debate scenarios, the Critic agent is encouraged to provide a different judgment from that of the General Public agent. In subsequent rounds of conversation, the General Public agent must consider the Critic’s feedback and may choose either to revise or to maintain its original judgment. The Critic agent continues to generate additional critiques and independent assessments based on the evolving conversation. At the end of the debate, we use the judgment of the General Public agent as the final verdict. A demonstration of the debate flow is shown in the left panel of Figure 1. The prompts used for the General Public and Critic agents are provided in Appendix F.1.

We focus on the “General Public and Critic” structure rather than incorporating additional agents such as Analyst or Supporter agents, as proposed by Chan et al.. This is because adding such roles, which typically represent mutually aligned perspectives, tends to offer limited additional judgment diversity and may weaken the intended role of the Critic agent within the debate. Additionally, our preliminary study finds that introducing more agents and extending the length of conversations increases the difficulty of instruction following, particularly for medium-sized LLMs. Taking these considerations into account, and aiming to isolate and measure the effect of actual debate between agents within the framework, our study specifically focuses on the “General Public and Critic” configuration within the Multi-Agent-Debate LLM-as-Judge framework.

3.3 LLM-as-Meta-Judge

In the LLM-as-Meta-Judge setting, we focus on the fundamental concept of “Thinking about thinking.” Inspired by the frameworks proposed by Li et al. and Wu et al., we adopt diverse LLMs as judges, each providing an independent judgment to form a judgment pool based on the evaluation task. We further examine two variations of the Meta-Judge framework, as shown in the right panel of Figure 1. In the “Select” setting, the Meta-Judge selects the best judgment from the pool and uses it as the final verdict. This setting evaluates the Meta-Judge’s ability to discern and identify the highest-quality answer among diverse alternatives, testing its competence as an evaluator and ranker of peer outputs. In contrast, the “Conclude” setting requires the Meta-Judge to provide its own judgment as the final verdict based on the judgment pool. This setting assesses the Meta-Judge’s ability to integrate, reason over, and reconcile potentially conflicting or complementary information, acting as a collaborative decision-maker. It is also similar to the “summarization” role that was widely used in group discussion settings (Liu et al., 2024c; Li et al., 2025b). The prompts used for the Meta Judges are provided in Appendix F.2.

We focus on the most basic structure of the Meta-Judges framework, intentionally avoiding techniques such as Rubric Design and Majority Rubric proposed by Li et al., as our primary goal is to measure how bias manifests within the framework itself. The inclusion of such techniques could introduce confounding effects that obscure this measurement. Additionally, position bias of the Meta-Judge itself could act as a confounder. To mitigate this, we shuffle the positions of the judge model outputs within the judgment pool. Furthermore, we gradually expand the size of the judge pool to examine the effect of pool size on the observed biases.

3.4 Biases and Measurement

In our study, we primarily focus on four types of intrinsic bias: position bias, verbosity bias, bandwagon bias, and chain-of-thought bias. Illustrative examples of these biases are provided in Figure 2. We selected these biases because they span some of the most critical and generalizable dimensions of intrinsic model behavior: input ordering, response length, reasoning stability, and social conformity. To measure them, we follow the procedures pro-

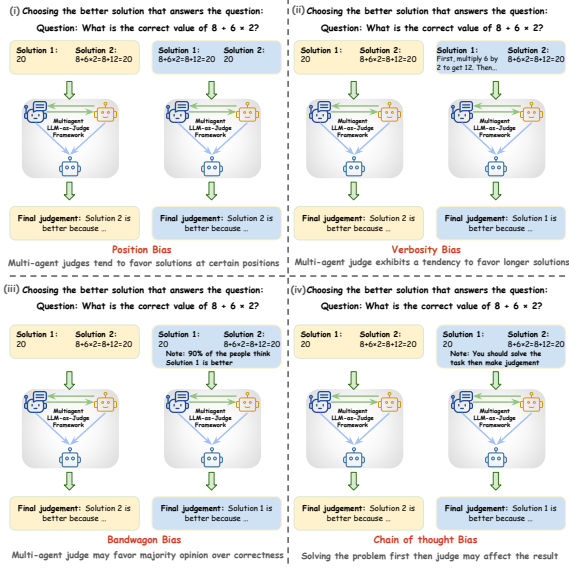


Figure 2: Demonstration of the biases measured in our study. We kept the multi-agent LLM-as-Judge framework unchanged throughout each evaluation process. Biases were introduced and measured solely by modifying the prompts, as shown.

posed by CALM (Ye et al., 2024), which assess bias by comparing the consistency of judgments before and after specific prompt modifications designed to elicit each bias. In the Multi-Agent LLM-as-Judge framework, all agents are exposed to these prompt modifications when measuring biases. Examples of these prompts can be found in Appendix F.3. The consistency rate is defined as:

$$CR = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(y^i = y_{\text{modified}}^i) \quad (1)$$

where N denotes the number of samples, y^i represents the original judgment for sample i , and y_{modified}^i represents the judgment after prompt modification for the same sample.

Position bias reflects the tendency of LLM judges to favor solutions based on their positions rather than content quality. To evaluate this, we alter the order of candidate solutions in the prompt and examined whether the preference shifted as a result. **Verbosity bias** occurs when judges prefer longer responses over shorter, equally or more accurate alternatives. We generate extended responses following the approach of Ye et al.. In their study, they verified through human evaluation that such modifications introduced minimal unintended improvement in answer quality. We then measure the degree to which LLM judgments shifted toward the longer version.

Bandwagon bias describes the inclination of LLM judges to align with majority opinions. To measure this, we add a statement to the prompt indicating that most people favor one of the candidate solutions, and compare the judgments with and without this statement to assess any shift in preference.

Chain-of-thought (CoT) bias refers to changes in evaluation outcomes when judges are first prompted to reason through the task independently before assessing candidate answers. We measure this by comparing judgments made with and without the additional reasoning step, observing how the internal deliberation influence final evaluations.

4 Experiment

4.1 Evaluation Dataset

We conduct our experiments on two benchmarks: MT-Bench (Zheng et al., 2023) and the Alignment dataset from CALM (Ye et al., 2024), which we refer to as CALM-Alignment. MT-Bench consists of responses from 30 LLMs to each of 8 tasks, with 10 questions per task. For each question, we sample 60 response pairs per task along with their corresponding questions, yielding a total of 480 samples. This sampling strategy balances computational cost with comparability to the CALM-Alignment dataset. CALM-Alignment, sampled from various DPO datasets, contains 439 samples. Each response pair includes a chosen and a rejected label based on human feedback. More details on the sub-sampled MT-Bench and CALM-Alignment datasets are provided in Appendix D. We apply prompt modifications as described in Sec. 3.4. Notably, since the CALM-Alignment dataset includes full labels, we specifically extend the solution for the rejected responses (which always appear as the second solution in the pair in original order). This allows us to measure how many previously correct judgments are changed to incorrect ones. For consistency, we also apply the prompt extension to the second solution in each pair in MT-Bench. To further support our findings, we also include results on Reward Bench (Lambert et al., 2024) in Appendix B.

4.2 Experiments Setup

In our experiments, we evaluate four LLMs: R1-Distilled-Qwen 32B (DeepSeek-AI, 2025), LLaMA-3.3-70B-Instruct-Turbo (abbreviated as Llama-3.3-70B) (Grattafiori et al., 2024), 4o-mini (Achiam et al., 2023), and DeepSeek-V3 (Liu et al.,

Judge	Critic	Round	MTBench				CALM-Alignment			
			Pos.	Verbo.	Band.	Cot	Pos.	Verbo.	Band.	Cot
R1-Distilled-Qwen-32B	R1-Distilled-Qwen-32B	0	0.772	0.736	0.728	0.795	0.723	0.732	0.689	0.826
		1	0.696 ↓	0.654 ↓	0.687 ↓	0.723 ↓	0.670 ↓	0.613 ↓	0.645 ↓	0.745 ↓
		2	0.700 ↑	0.652 ↓	0.681 ↓	0.729 ↑	0.672 ↑	0.620 ↑	0.633 ↓	0.756 ↑
		3	0.702 ↑	0.658 ↑	0.685 ↑	0.729 →	0.677 ↑	0.633 ↑	0.626 ↓	0.754 ↓
R1-Distilled-Qwen-32B	GPT4o-mini	0	0.772	0.736	0.728	0.795	0.723	0.732	0.689	0.826
		1	0.621 ↓	0.606 ↓	0.656 ↓	0.660 ↓	0.626 ↓	0.608 ↓	0.624 ↓	0.688 ↓
		2	0.637 ↑	0.608 →	0.640 ↓	0.656 ↓	0.617 ↓	0.608 →	0.642 ↑	0.699 ↑
		3	0.642 ↑	0.610 ↑	0.644 ↑	0.665 ↑	0.629 ↑	0.604 ↓	0.642 →	0.695 ↓
R1-Distilled-Qwen-32B	Llama-3.3-70B-Instruct-Turbo	0	0.772	0.736	0.728	0.795	0.723	0.732	0.689	0.826
		1	0.754 ↓	0.660 ↓	0.721 ↓	0.752 ↓	0.697 ↓	0.645 ↓	0.665 ↓	0.745 ↓
		2	0.754 →	0.656 ↓	0.727 ↑	0.767 ↑	0.692 ↓	0.656 ↑	0.667 ↑	0.733 ↓
		3	0.754 →	0.658 ↑	0.731 ↑	0.765 ↓	0.692 →	0.670 ↑	0.672 ↑	0.733 →
R1-Distilled-Qwen-32B	DeepSeek-V3	0	0.772	0.736	0.728	0.795	0.723	0.732	0.689	0.826
		1	0.594 ↓	0.640 ↓	0.642 ↓	0.644 ↓	0.581 ↓	0.649 ↓	0.592 ↓	0.626 ↓
		2	0.598 ↑	0.656 ↑	0.660 ↑	0.640 ↓	0.574 ↓	0.658 ↑	0.595 ↑	0.640 ↑
		3	0.617 ↑	0.644 ↓	0.675 ↑	0.646 ↑	0.588 ↑	0.651 ↓	0.608 ↑	0.651 ↑
GPT4o-mini	GPT4o-mini	0	0.793	0.716	0.746	0.835	0.793	0.789	0.740	0.869
		1	0.667 ↓	0.637 ↓	0.667 ↓	0.731 ↓	0.663 ↓	0.704 ↓	0.663 ↓	0.763 ↓
		2	0.648 ↓	0.617 ↓	0.692 ↑	0.740 ↑	0.681 ↑	0.704 →	0.683 ↑	0.781 ↑
		3	0.648 →	0.598 ↓	0.704 ↑	0.719 ↓	0.692 ↑	0.708 ↑	0.704 ↑	0.784 ↑
GPT4o-mini	Llama-3.3-70B-Instruct-Turbo	0	0.793	0.716	0.746	0.835	0.793	0.789	0.740	0.869
		1	0.744 ↓	0.681 ↓	0.735 ↓	0.777 ↓	0.715 ↓	0.708 ↓	0.697 ↓	0.797 ↓
		2	0.762 ↑	0.692 ↑	0.767 ↑	0.775 ↓	0.743 ↑	0.699 ↓	0.702 ↑	0.804 ↑
		3	0.765 ↑	0.712 ↑	0.760 ↓	0.787 ↑	0.722 ↓	0.713 ↑	0.718 ↑	0.795 ↓
Llama-3.3-70B-Instruct-Turbo	GPT4o-mini	0	0.814	0.720	0.783	0.821	0.823	0.773	0.768	0.839
		1	0.765 ↓	0.646 ↓	0.771 ↓	0.696 ↓	0.706 ↓	0.670 ↓	0.761 ↓	0.747 ↓
		2	0.756 ↓	0.644 ↓	0.731 ↓	0.698 ↑	0.692 ↓	0.638 ↓	0.711 ↓	0.756 ↑
		3	0.775 ↑	0.679 ↑	0.746 ↑	0.721 ↑	0.708 ↑	0.656 ↑	0.702 ↓	0.749 ↓
Llama-3.3-70B-Instruct-Turbo	Llama-3.3-70B-Instruct-Turbo	0	0.814	0.720	0.783	0.821	0.823	0.773	0.768	0.839
		1	0.773 ↓	0.665 ↓	0.756 ↓	0.692 ↓	0.713 ↓	0.690 ↓	0.736 ↓	0.688 ↓
		2	0.733 ↓	0.662 ↓	0.742 ↓	0.744 ↑	0.704 ↓	0.677 ↓	0.692 ↓	0.711 ↑
		3	0.727 ↓	0.656 ↓	0.760 ↑	0.760 ↑	0.690 ↓	0.656 ↓	0.708 ↑	0.715 ↑
DeepSeek-V3	DeepSeek-V3	0	0.821	0.746	0.748	0.883	0.809	0.718	0.688	0.868
		1	0.715 ↓	0.687 ↓	0.721 ↓	0.758 ↓	0.665 ↓	0.658 ↓	0.645 ↓	0.738 ↓
		2	0.717 ↑	0.696 ↑	0.733 ↑	0.733 ↓	0.667 ↑	0.613 ↓	0.649 ↑	0.713 ↓
		3	0.717 →	0.704 ↑	0.719 ↓	0.717 ↓	0.679 ↑	0.654 ↑	0.642 ↓	0.704 ↓

Table 1: Consistency scores for each bias (Positivity, Verbosity, Bandwagon, Cot) across three rounds of multi-agent debate, shown for different Judge–Critic model pairs on MTBench and CALM-Alignment. Arrows indicate trends across rounds (↓ sharp decrease , ↓ decrease , → stable , ↑ increase). Round 0 indicates the consistency rate of judges without debate. The higher consistency rate the better resistance to biases.

2024a). For the multi-agent debate experiment, we adopt a reference-focused strategy, with R1-Distilled-Qwen 32B most frequently serving as the Judge model and other models acting as Critics. To explore cross-model dynamics, we also conduct selective pairings with LLaMA 3.3 70B, GPT-4o-mini, and DeepSeek-V3. Both self-play and cross-model pairings are included, as summarized in Table 1. Each debate is limited to three rounds per sample to ensure consistency across experiments while keeping computational costs manageable. We do not account for early convergence in judgment between the agent and the judge, as allowing variable-length debates would reduce comparability across different model combinations. For the Meta-Judge setting, GPT-4o-mini, LLaMA-3.3-70B, and DeepSeek-V3 serve as the Meta-Judge. We use R1-Distilled-Qwen 32B as reference in the judgement pool, and thus we do not use it as a Meta-Judge. To assess how the size of the judgment pool affects biases, we vary the pool from two models up to the full set of models providing initial judgments. The Meta-Judge then either selects from these judgments or concludes its own, based on the judgements from the judgment pool.

The full study is summarized in Table 2.

To assess the impact of a bias-free agent, we apply the PINE method (Wang et al., 2024b) to Qwen1.5-32B-Chat (Bai et al., 2023), strictly following the setup used in the original work. We then deploy it in debates against GPT-4o-mini (which exhibited the largest drop in position bias consistency) and LLaMA 3.3 70B (which showed the least drop). Leveraging PINE’s position bias elimination, we investigate whether a bias-free agent can reduce bias in judge models through interaction. In our experiments, the bias-free agent serves as a critic in the Multi-Agent Debate and is included in the judgment pool under Meta-Judge settings. This experiment targets scenarios where modifying the judge model or its prompt is impractical, such as self-reward settings during training (Zhang et al., 2024; Wu et al., 2024). Full results are presented in Table 3 and Table 4. Case studies and examples can be found in Appendix G.

5 Results and Analysis

5.1 Multi-Agent-Debate

Multi-Agent Debate Can Amplify Bias Over Rounds. As shown in Table 1, introducing Multi-

Meta Judge: 4o-mini						MT-Bench				CALM-Alignment			
R1-Distilled-Qwen 32B	4o-mini	Llama3.3 70B	DeepSeek-V3	Mode		Pos.	Verb.	Band.	Cot.	Pos.	Verb.	Band.	Cot.
No	No	No	No	Judge		0.793	0.716	0.746	0.835	0.793	0.789	0.740	0.869
Yes	Yes	No	No	Select		0.817	0.737	0.735	0.783	0.793	0.772	0.677	0.861
Yes	No	Yes	No	Select		0.798	0.706	0.746	0.798	0.811	0.745	0.754	0.841
Yes	No	No	Yes	Select		0.806	0.756	0.706	0.819	0.804	0.715	0.658	0.866
Yes	No	Yes	Yes	Select		0.835	0.721	0.733	0.844	0.836	0.731	0.731	0.852
Yes	Yes	Yes	Yes	Select		0.823	0.729	0.740	0.798	0.831	0.738	0.715	0.863
Yes	Yes	No	No	Conclude		0.762	0.727	0.750	0.825	0.786	0.743	0.713	0.868
Yes	No	Yes	No	Conclude		0.775	0.692	0.737	0.823	0.825	0.761	0.779	0.877
Yes	No	No	Yes	Conclude		0.808	0.721	0.698	0.819	0.831	0.720	0.722	0.845
Yes	No	Yes	Yes	Conclude		0.846	0.748	0.758	0.852	0.859	0.752	0.749	0.859
Yes	Yes	Yes	Yes	Conclude		0.819	0.762	0.777	0.837	0.854	0.768	0.752	0.866
Meta Judge: Llama3.3 70B						Pos.	Verb.	Band.	Cot.	Pos.	Verb.	Band.	Cot.
R1-Distilled-Qwen 32B	4o-mini	Llama3.3 70B	DeepSeek-V3	Mode		0.814	0.720	0.783	0.821	0.823	0.773	0.768	0.839
No	No	No	No	Judge		0.767	0.712	0.733	0.812	0.804	0.754	0.720	0.841
Yes	Yes	No	No	Select		0.787	0.702	0.744	0.785	0.804	0.727	0.740	0.836
Yes	No	Yes	No	Select		0.800	0.719	0.729	0.815	0.809	0.704	0.699	0.838
Yes	Yes	No	Yes	Select		0.831	0.735	0.717	0.815	0.800	0.738	0.722	0.863
Yes	Yes	Yes	Yes	Select		0.812	0.721	0.758	0.796	0.806	0.738	0.754	0.882
Yes	Yes	No	No	Conclude		0.754	0.690	0.715	0.790	0.772	0.747	0.733	0.793
Yes	No	Yes	No	Conclude		0.806	0.717	0.746	0.792	0.786	0.733	0.738	0.818
Yes	No	No	Yes	Conclude		0.806	0.706	0.687	0.800	0.788	0.688	0.697	0.809
Yes	Yes	No	Yes	Conclude		0.804	0.723	0.725	0.787	0.804	0.720	0.686	0.838
Yes	Yes	Yes	Yes	Conclude		0.833	0.712	0.731	0.810	0.822	0.733	0.722	0.806
Meta Judge: DeepSeek-V3						Pos.	Verb.	Band.	Cot.	Pos.	Verb.	Band.	Cot.
R1-Distilled-Qwen 32B	4o-mini	Llama3.3 70B	DeepSeek-V3	Mode		0.821	0.746	0.748	0.883	0.809	0.718	0.688	0.868
No	No	No	No	Judge		0.773	0.708	0.687	0.817	0.788	0.745	0.686	0.852
Yes	Yes	No	No	Select		0.775	0.721	0.737	0.796	0.797	0.736	0.695	0.836
Yes	No	No	Yes	Select		0.802	0.721	0.725	0.827	0.813	0.729	0.663	0.854
Yes	Yes	Yes	No	Select		0.815	0.719	0.725	0.821	0.831	0.759	0.672	0.856
Yes	Yes	Yes	Yes	Select		0.840	0.715	0.735	0.825	0.838	0.738	0.638	0.870
Yes	Yes	No	No	Conclude		0.800	0.760	0.687	0.823	0.795	0.779	0.674	0.854
Yes	No	Yes	No	Conclude		0.817	0.731	0.754	0.823	0.804	0.754	0.711	0.847
Yes	No	No	Yes	Conclude		0.827	0.740	0.692	0.860	0.841	0.718	0.672	0.850
Yes	Yes	Yes	No	Conclude		0.812	0.767	0.746	0.848	0.815	0.754	0.695	0.852
Yes	Yes	Yes	Yes	Conclude		0.825	0.765	0.733	0.873	0.859	0.736	0.692	0.845

Table 2: Consistency scores for LLM-as-Meta-Judge across biases, pool sizes, and modes. "Yes" indicates inclusion of the model in the judgment pool; "No" indicates exclusion (single judge). Cell colors indicate the difference compared with the single judge: green denotes higher and blue denotes lower.

Judge	Critic	Round	MTBench	CALM
4o-mini	PINE	1	0.817	0.786
		2	0.825 ↑	0.804 ↑
		3	0.831 ↑	0.815 ↑
Llama3.3 70B	PINE	1	0.802	0.781
		2	0.812 ↑	0.786 ↑
		3	0.821 ↑	0.788 ↑

Table 3: Position bias consistency score when bias-free agent are served as a critic. Round 0 indicates the consistency rate of judges without debate.

Agent Debate into LLM evaluation leads to a sharp increase in bias immediately after the first round. Multiple pairwise t-tests across all judge-critic model pairings, both across and within different bias types, reveal a statistically significant drop in consistency rate from round 0 to round 1. Notably, this elevated level of bias persists in subsequent rounds, with no significant further increase or signs of recovery. This pattern is consistent across all bias types we tested. These findings suggest that the initial introduction of debate—rather than its continuation—poses the greatest risk for bias amplification in multi-agent evaluation. Moreover, additional rounds of debate appear insufficient to recover the amplified bias once introduced. Detailed analyses and statistical results are provided

in Appendix A.1.

Bias Amplification Is a General Phenomenon, May Not Tied to Weak Models. Our results show that the amplification of bias in Multi-Agent-Debate is not limited to weaker or smaller models, but is a phenomenon observed across all judge-critic pairings. For example, DeepSeek-V3, despite being one of the largest and most advanced models in our experiments, exhibited a substantial and statistically significant drop in consistency rate after the introduction of debate, with position consistency falling from 0.82 pre-debate to 0.71 after the first round. Similar patterns were observed for other leading models such as Llama-3.3-70B and GPT4o-mini, indicating that increased model size and capability may not confer immunity to such bias amplification. This generality across architectures and scales underscores that the challenge of collaborative bias amplification is intrinsic to the debate framework itself, rather than a limitation of particular models.

5.2 LLM-as-Meta-Judge

LLM-as-Meta-Judge Frameworks Achieve Bias Consistency Comparable to Single-Model Judg-

Meta Judge: 4o-mini					MT-Bench		CALM-Alignment	
4o-mini	DeepSeek-V3	Llama3.3 70B	R1-Distilled-Qwen 32B	PINE	Select	Conclude	Select	Conclude
Yes	No	No	No	Yes	0.833	0.794	0.815	0.829
	No	No	Yes	Yes	0.802	0.781	0.800	0.827
No	No	Yes	Yes	Yes	0.833	0.823	0.843	0.847
No	Yes	Yes	Yes	Yes	0.840	0.848	0.831	0.852
Yes	Yes	Yes	Yes	Yes	0.804	0.817	0.841	0.863
Meta Judge: Llama3.3 70B					Select	Conclude	Select	Conclude
4o-mini	DeepSeek-V3	Llama3.3 70B	R1-Distilled-Qwen 32B	PINE				
No	No	Yes	No	Yes	0.815	0.827	0.843	0.784
	No	No	Yes	Yes	0.775	0.771	0.777	0.800
No	No	Yes	Yes	Yes	0.790	0.775	0.815	0.806
No	Yes	Yes	Yes	Yes	0.806	0.783	0.861	0.827
Yes	Yes	Yes	Yes	Yes	0.842	0.808	0.825	0.806
Meta Judge: DeepSeek-V3					Select	Conclude	Select	Conclude
4o-mini	DeepSeek-V3	Llama3.3 70B	R1-Distilled-Qwen 32B	PINE				
No	Yes	No	No	Yes	0.819	0.792	0.809	0.790
	No	No	Yes	Yes	0.767	0.792	0.781	0.795
No	No	Yes	Yes	Yes	0.802	0.787	0.827	0.784
No	Yes	Yes	Yes	Yes	0.817	0.817	0.868	0.836
Yes	Yes	Yes	Yes	Yes	0.840	0.827	0.845	0.850

Table 4: Position bias consistency score when bias-free agent is included in the judgment pool. "Yes" indicates inclusion of the model in the judgment pool; "No" indicates exclusion. The highest consistency is shown in **bold**.

ment. As shown in Table 2, across both the MT-Bench and CALM-Alignment benchmarks, and for most bias types, the Select and Conclude modes with diverse judge pools generally yield consistency scores comparable to, and in some cases better than, those of a strong single-model judge. However, comprehensive pairwise t-tests comparing all pooled Meta-Judge settings against their single-judge baselines reveal a statistically significant drop in overall consistency. However, the magnitude of this drop is relatively small—typically within 2%. Additionally, for position consistency, the Meta-Judge framework shows no significant difference from, and sometimes even outperforms, the single-judge model under certain mode and size of judgment pool. Overall, this suggests that the Meta-Judge achieves bias consistency levels comparable to those of the single-judge model. Detailed discussions and results can be found in Appendix A.2.

Larger Judgment Pools Yield Comparable Consistency, with Gains in Position Bias. For example, using 4o-mini as the Meta-Judge on the CALM-Alignment benchmark, position bias consistency rises from 0.793 to 0.854 in conclude mode. Statistical analysis confirms that this improvement is significant for position bias. Comprehensive pairwise t-tests across different pool sizes suggest that with a pool size of two, the Meta-Judge generally yields significantly lower consistency across all bias types. However, as the pool size increases, consistency becomes comparable to that of single-judge models—and even better in terms of position consistency. These results suggest that increasing pool size is beneficial in Meta-Judge settings. However, there is no evidence that

it consistently outperforms the single-judge model overall. Full details of the analysis are provided in Appendix A.2.

5.3 Bias-Free Experiment

Although the bias-free agent performs well in debate settings, it has minimal effect when used as a Meta-Judge. As shown in the Table 3, we observe a continuous improvement in position bias consistency with increasing rounds of debate. This is significantly different from what we observed without such bias-free agent. On the other hand, as shown in Table 4, there is no considerable improvement to the consistency score when including this bias-free agent into the judgment pool. Our analysis further shows that, in select mode, the Meta-Judge shows no favor to the bias-free agent. This might suggest that the Meta-Judge itself may not be sensitive enough to recognize or value the unbiased nature of the agent’s outputs, instead treating all pool members as equally valid. As a result, the bias-free agent’s potential for debiasing is diluted within the collective judgment process. More details and analysis can be found in Appendix A.3. More results of PINE with Llama3 as base model can be found in Appendix C.

6 Conclusion

In conclusion, our experiments show that Multi-Agent Debate can amplify intrinsic biases in LLMs through agent interactions, while the LLM-as-Meta-Judge approach demonstrates better consistency, and sometimes improvement, against these biases, especially with larger judgment pools. Notably, the presence of a bias-free agent makes the

Multi-Agent Debate framework even more promising, at times surpassing the Meta-Judge approach. However, these results highlight the importance of developing bias-free agents that can effectively debias within fewer rounds of interaction. Overall, our findings offer guidance for deploying Multi-Agent LLM-as-Judge frameworks and emphasize the need for universal debiasing techniques capable of addressing multiple biases simultaneously, particularly in collaborative evaluation settings.

Limitations

This study focuses on understanding how biases manifest in multi-agent frameworks when used in the context of LLM-as-a-Judge systems. We specifically examine four types of bias: position bias, bandwagon bias, chain-of-thought bias, and verbosity bias. While these do not represent the full spectrum of potential biases, they are representative and have received substantial attention in prior work. Our goal is not to provide a comprehensive taxonomy of biases, but rather to initiate and encourage deeper exploration into how bias operates in collaborative LLM settings. Investigating additional forms of bias remains an important direction for future research.

To maintain clarity and control over confounding variables, we adopt a deliberately simplified multi-agent framework. Rather than replicating all existing frameworks, we focus on a core structure that is generalizable to more complex designs. This choice allows us to isolate bias-related effects without interference from more elaborate coordination strategies, but it also limits the direct applicability of our results to more sophisticated multi-agent architectures. Future work could investigate how different agent designs and interaction protocols influence the emergence or mitigation of bias.

In terms of mitigation, our study evaluates only a single method for reducing position bias. Although more advanced strategies—such as modifying judge prompt structures—may yield stronger results, our focus is on scenarios where altering individual judge models is not feasible. We believe this reflects a more realistic and widely applicable constraint in real-world deployments. Our findings suggest that even under such constraints, meaningful bias mitigation remains a critical challenge, and we emphasize the need for more general and robust solutions that can be applied across multiple bias types in collaborative LLM systems.

Finally, our experiments are conducted on two benchmarks. While other benchmarks exist—such as Judge-Bench (Tan et al., 2024), which emphasizes problem-solving ability—we chose not to include them due to the risk of confounding bias measurements with task-solving failures. Accurately attributing selection outcomes to bias becomes difficult when a model’s ability to answer the question is in question. Moreover, multi-agent inference is computationally intensive and demands strong instruction-following capabilities, which limits the scale of evaluation. We test four different model types across a diverse set of permutations, resulting in approximately 288 total instances. While this setup is sufficient for our core analyses, it presents challenges for fine-grained statistical testing, where smaller subgroup sizes can lead to greater variance and less reliable conclusions.

Ethical Statement

This work involves the use of a publicly available dataset, CALM alignment (Ye et al., 2024), which contains a small portion of content labeled as NSFW. According to the dataset authors, all data has been manually reviewed and curated to be appropriate for research use. In accordance with the dataset’s license and usage guidelines, we have not redistributed or publicly shared any part of the dataset. These data are used solely for the purpose of evaluating bias in AI models. We encourage responsible use of this work, with attention to ethical standards and potential social implications.

Acknowledgment

This research was supported in part by a generous grant from the John Templeton Foundation.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Samee Arif, Sualeha Farid, Abdul Hameed Azeemi, Awais Athar, and Agha Ali Raza. 2024. The fellowship of the llms: Multi-agent workflows for synthetic preference optimization dataset generation. *arXiv preprint arXiv:2408.08688*.
- Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin,

- Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, and 29 others. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.
- Han Bao, Yue Huang, Yanbo Wang, Jiayi Ye, Xiangqi Wang, Xiuying Chen, Yue Zhao, Tianyi Zhou, Mohamed Elhoseiny, and Xiangliang Zhang. 2024. Autobench-v: Can large vision-language models benchmark themselves? *arXiv preprint arXiv:2410.21259*.
- Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, and 1 others. 2023. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. *arXiv preprint arXiv:2312.09390*.
- Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*.
- Dongping Chen, Ruoxi Chen, Shilin Zhang, Yaochen Wang, Yinuo Liu, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. In *Forty-first International Conference on Machine Learning*.
- Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chen Qian, Chi-Min Chan, Yujia Qin, Yaxi Lu, Ruobing Xie, and 1 others. 2023. Agentverse: Facilitating multi-agent collaboration and exploring emergent behaviors in agents. *arXiv preprint arXiv:2308.10848*, 2(4):6.
- DeepSeek-AI. 2025. [Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning](#). Preprint, arXiv:2501.12948.
- Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multi-agent debate. In *Forty-first International Conference on Machine Learning*.
- Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2023. AlpacaFarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36:30039–30069.
- Zhaopeng Feng, Jiayuan Su, Jiamei Zheng, Jiahao Ren, Yan Zhang, Jian Wu, Hongwei Wang, and Zuozhu Liu. 2024. M-mad: Multidimensional multi-agent debate framework for fine-grained machine translation evaluation. *arXiv preprint arXiv:2412.20127*.
- Peizhong Gao, Ao Xie, Shaoguang Mao, Wenshan Wu, Yan Xia, Haipeng Mi, and Furu Wei. 2024. Meta reasoning for large language models. *arXiv preprint arXiv:2406.11698*.
- Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.
- Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.
- Aditya Gulati, Moreno D’Incà, Nicu Sebe, Bruno Lepri, and Nuria Oliver. 2025. Uncovering an attractiveness bias in multimodal large language models: A case study with llava. *arXiv preprint arXiv:2504.16104*.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, and 1 others. 2022. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*.
- Sirui Hong, Xiaowu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, and 1 others. 2023. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6.
- Dong Huang, Jie M Zhang, Michael Luck, Qingwen Bu, Yuhao Qing, and Heming Cui. 2023a. Agent-coder: Multi-agent-based code generation with iterative testing and optimisation. *arXiv preprint arXiv:2312.13010*.
- Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2023b. Large language models cannot self-correct reasoning yet. *arXiv preprint arXiv:2310.01798*.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. 2020. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.
- Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, and 1 others. 2024. Rewardbench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*.
- Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025a. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.
- Ruosun Li, Teerth Patel, and Xinya Du. 2023a. Prd: Peer rank and discussion improve large language model based evaluations. *arXiv preprint arXiv:2307.02762*.

- Yuran Li, Jama Hussein Mohamud, Chongren Sun, Di Wu, and Benoit Boulet. 2025b. Leveraging llms as meta-judges: A multi-agent framework for evaluating llm judgments. *arXiv preprint arXiv:2504.17087*.
- Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2023b. Split and merge: Aligning position biases in large language model based evaluators. *arXiv preprint arXiv:2310.01432*.
- Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Shuming Shi, and Zhaopeng Tu. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118*.
- Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, and 1 others. 2024a. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.
- Chris Yuhao Liu, Liang Zeng, Jiacai Liu, Rui Yan, Ju-jie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024b. Skywork-reward: Bag of tricks for reward modeling in llms. *arXiv preprint arXiv:2410.18451*.
- Tongxuan Liu, Xingyu Wang, Weizhe Huang, Wenjiang Xu, Yuting Zeng, Lei Jiang, Hailong Yang, and Jing Li. 2024c. Groupdebate: Enhancing the efficiency of multi-agent debate using group discussion. *arXiv preprint arXiv:2409.14051*.
- Junsoo Park, Seungyeon Jwa, Ren Meiyang, Daeyoung Kim, and Sanghyuk Choi. 2024. Offsetbias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067.
- Swarnadeep Saha, Xian Li, Marjan Ghazvininejad, Jason Weston, and Tianlu Wang. 2025. Learning to plan & reason for evaluation with thinking-llm-as-a-judge. *arXiv preprint arXiv:2501.18099*.
- Tao Shen, Didi Zhu, Ziyu Zhao, Chao Wu, and Fei Wu. 2025. Will llms scaling hit the wall? breaking barriers via distributed resources on massive edge devices. *arXiv preprint arXiv:2503.08223*.
- Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms. *arXiv preprint arXiv:2406.07791*.
- Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. 2023. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652.
- Maja Stahl, Leon Biermann, Andreas Nehring, and Henning Wachsmuth. 2024. Exploring llm prompting strategies for joint essay scoring and feedback generation. *arXiv preprint arXiv:2404.15845*.
- Yuan Sui, Yufei He, Tri Cao, Simeng Han, and Bryan Hooi. 2025. Meta-reasoner: Dynamic guidance for optimized inference-time reasoning in large language models. *arXiv preprint arXiv:2502.19918*.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.
- Xinyu Tian, Shu Zou, Zhaoyuan Yang, and Jing Zhang. 2025. Identifying and mitigating position bias of multi-image vision-language models. *arXiv preprint arXiv:2503.13792*.
- Prapti Trivedi, Aditya Gulati, Oliver Molenschot, Meghana Arakkal Rajeev, Rajkumar Ramamurthy, Keith Stevens, Tanveesh Singh Chaudhery, Jahnvi Jambholkar, James Zou, and Nazneen Rajani. 2024. Self-rationalization improves llm as a fine-grained judge. *arXiv preprint arXiv:2410.05495*.
- Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.
- Pablo Villalobos, Anson Ho, Jaime Sevilla, Tamay Besiroglu, Lennart Heim, and Marius Hobbhahn. 2024. Position: Will we run out of data? limits of llm scaling based on human-generated data. In *Forty-first International Conference on Machine Learning*.
- David Wan, Jesse Vig, Mohit Bansal, and Shafiq Joty. 2024. On positional bias of faithfulness for long-form summarization. *arXiv preprint arXiv:2410.23609*.
- Ziyu Wan, Yunxiang Li, Yan Song, Hanjing Wang, Linyi Yang, Mark Schmidt, Jun Wang, Weinan Zhang, Shuyue Hu, and Ying Wen. 2025. Rema: Learning to meta-think for llms with multi-agent reinforcement learning. *arXiv preprint arXiv:2503.09501*.
- Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and 1 others. 2024a. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450.
- Yutong Wang, Pengliang Ji, Chaoqun Yang, Kaixin Li, Ming Hu, Jiaoyang Li, and Guillaume Sartoretti. 2025. Mcts-judge: Test-time scaling in llm-as-a-judge for code correctness evaluation. *arXiv preprint arXiv:2502.12468*.
- Ziqi Wang, Hanlin Zhang, Xiner Li, Kuan-Hao Huang, Chi Han, Shuiwang Ji, Sham M Kakade, Hao Peng, and Heng Ji. 2024b. Eliminating position bias of language models: A mechanistic approach. *arXiv preprint arXiv:2407.01100*.

- Tianhao Wu, Weizhe Yuan, Olga Golovneva, Jing Xu, Yuandong Tian, Jiantao Jiao, Jason Weston, and Sainbayar Sukhbaatar. 2024. Meta-rewarding language models: Self-improving alignment with llm-as-a-meta-judge. *arXiv preprint arXiv:2407.19594*.
- Zhenran Xu, Senbao Shi, Baotian Hu, Jindi Yu, Dongfang Li, Min Zhang, and Yuxiang Wu. 2023. Towards reasoning in large language models via multi-agent peer review collaboration. *arXiv preprint arXiv:2311.08152*.
- Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, and 1 others. 2024. Justice or prejudice? quantifying biases in llm-as-a-judge. *arXiv preprint arXiv:2410.02736*.
- Ziyi Ye, Xiangsheng Li, Qiuchi Li, Qingyao Ai, Yujia Zhou, Wei Shen, Dong Yan, and Yiqun Liu. 2025. Learning llm-as-a-judge for preference alignment. In *The Thirteenth International Conference on Learning Representations*.
- Dan Zhang, Sining Zhoubian, Ziniu Hu, Yisong Yue, Yuxiao Dong, and Jie Tang. 2024. Rest-mcts*: Llm self-training via process reward guided tree search. *Advances in Neural Information Processing Systems*, 37:64735–64772.
- Hangfan Zhang, Zhiyao Cui, Xinrun Wang, Qiaosheng Zhang, Zhen Wang, Dinghao Wu, and Shuyue Hu. 2025a. If multi-agent debate is the answer, what is the question? *arXiv preprint arXiv:2502.08788*.
- Qiyuan Zhang, Yufei Wang, Yuxin Jiang, Liangyou Li, Chuhan Wu, Yasheng Wang, Xin Jiang, Lifeng Shang, Ruiming Tang, Fuyuan Lyu, and 1 others. 2025b. Crowd comparative reasoning: Unlocking comprehensive evaluations for llm-as-a-judge. *arXiv preprint arXiv:2502.12501*.
- Yuyao Zhang, Jinghao Li, and Yu-Wing Tai. 2025c. Layercraft: Enhancing text-to-image generation with cot reasoning and layered object integration. *arXiv preprint arXiv:2504.00010*.
- Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623.
- Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

Appendix Table of Contents

A Detailed Tests and analysis on Experiments	13
A.1 Multi-Agent-Debate	13
A.2 LLM-as-Meta-Judge	14
A.3 Analysis of Meta-Judge Selections with PINE	15
B Results on RewardBench	15
C More Results on Bias Free Agent	16
D More information on Benchmarks	16
D.1 MT-Bench	16
D.2 CALM Benchmark	16
E More information on PINE	17
F Prompt Examples	18
F.1 Multi-Agent-Debate	18
F.2 LLM-as-Meta Judge Prompts	18
F.3 Bias Prompts	18
G Case Studies of Multi-Agent Debate	18
G.1 Original Debate	18
G.2 With biases	18
G.3 With PINE	19
H Case Studies of LLM-as-Meta-Judge	19
H.1 Original Debate	19
H.2 With biases	19
H.3 With PINE	19
I Implementation and Reproducibility	19

A Detailed Tests and analysis on Experiments

A.1 Multi-Agent-Debate

Multi-Agent Debate Can Amplify Bias Over Rounds. To test this, we perform pairwise Welch Two Sample t -tests on the consistency rates between successive rounds, both within and across bias types, across the two benchmarks. For all of the tests, the Null hypothesis is that there is no difference between the two groups tested. We reject the hypothesis on 0.05 significance level. [Figure 6](#) shows the mean consistency score changes over round.

Across Bias Types: We begin by testing whether the consistency scores change significantly across successive debate rounds. Specifically, we compare round 1 to round 0, round 2 to round 1, and round 3 to round 2. As shown in [Table 5](#), there is a statistically significant difference in consistency scores between the initial judgment (round 0, single-agent) and the first round of debate (round 1). The confidence interval further confirms that introducing debate leads to a statistically significant decrease in consistency scores across all bias types, suggesting that the Multi-Agent Debate setting amplifies biases. In contrast, there is no statistically significant difference between rounds 1 and 2 or between rounds 2 and 3. This suggests that once the initial amplification of bias occurs, subsequent rounds do not recover it. Instead, the consistency scores appear to converge, possibly due to agents reaching mutual agreement through continued debate.

Test	Pair-wised Welch Two Sample t -test
Data	0th round v.s 1st round
t-value	9.76
p-value	< 2.2e-16
95% Confidence Interval	[0.0663, 0.1001]
Data	1st round v.s 2nd round
t-value	0.0186
p-value	0.9852
95% Confidence Interval	[-0.0168, 0.0172]
Data	2nd round v.s 3rd round
t-value	-0.49
p-value	0.6255
95% Confidence Interval	[-0.0208, 0.0125]
Data	1nd round v.s 3rd round
t-value	-0.47
p-value	0.6389
95% Confidence Interval	[-0.0207, 0.0127]

Table 5: Pair-wised Welch Two Sample t -test Results for consistency scores between rounds. We do not separate the effect of different bias type in this test.

Within Bias Types: We further analyze the results by separating them by bias type. As shown

in Table 6, the patterns observed earlier remain consistent across most bias categories. For all bias types evaluated, the biases tend to be significantly amplified in the first round of debate, with little statistical evidence indicating recovery in later rounds. An exception appears in the case of the bandwagon bias, where a statistically significant recovery is observed in the third round. However, the confidence interval suggests that this recovery effect is minimal.

Bias Type	Comparison	<i>t</i> -value	<i>p</i> -value	95% CI
Position	0 vs 1	8.516	1.544e-07	[0.0706, 0.1171]
Position	1 vs 2	0.211	0.8354	[-0.0070, 0.0086]
Position	2 vs 3	-1.634	0.1206	[-0.0094, 0.0012]
CoT	0 vs 1	10.502	7.518e-09	[0.0885, 0.1330]
CoT	1 vs 2	-1.403	0.1786	[-0.0146, 0.0029]
CoT	2 vs 3	-0.280	0.7829	[-0.0064, 0.0049]
Verbose	0 vs 1	15.004	3.082e-11	[0.0721, 0.0957]
Verbose	1 vs 2	0.988	0.3371	[-0.0042, 0.0116]
Verbose	2 vs 3	-1.555	0.1383	[-0.0145, 0.0022]
Bandwagon	0 vs 1	5.977	1.5e-05	[0.0286, 0.0598]
Bandwagon	1 vs 2	0.356	0.726	[-0.0098, 0.0137]
Bandwagon	2 vs 3	-2.174	0.0441	[-0.0109, -0.0002]

Table 6: Pair-wised Welch *t*-Test Results Across Rounds by Bias Type

A.2 LLM-as-Meta-Judge

Conclude mode or Choose mode? We first test on whether there is a significant difference between two modes in terms of consistency rate across all biases. As shown in Table 7, **there is no evidence that one mode is better than the other, in terms of bias consistency rate.**

Test	Paired <i>t</i> -test
Data	choose mode v.s. conclude mode
<i>t</i> -value	-1.7913
<i>p</i> -value	0.07578
Alternative Hypothesis	True mean difference is not equal to 0
95% Confidence Interval	[-0.0086, 0.0004]
Mean Difference	-0.0041

Table 7: Pair-wised Welch *t*-Test Between Choose Mode and Conclude Mode

Does the Meta-Judge Amplify Biases as Well?

To begin, we tested whether biases resulting from different judgment pool sizes, modes, and meta-judge configurations are significantly lower than those produced by a single-judge model. If so, we further examined the extent of this reduction. As shown in Table 8, introducing the Meta-Judge results in a statistically significant overall drop in bias consistency. However, the mean difference and confidence intervals suggest that this reduction is typically within 1% to 2%. **This indicates that while the drop is statistically significant, it is**

relatively small compared to the amplification observed with Multi-Agent Debate. In effect, the Meta-Judge achieves bias consistency levels comparable to those of the single-judge model.

Furthermore, as shown in Table 9, this pattern holds consistently across CoT, Verbose, and Bandwagon biases. However, for Position bias, there is no statistical evidence indicating a significant difference in position consistency between the LLM-as-Meta-Judge and the LLM-as-Judge. Additionally, as shown in Table 10, both Meta-Judge modes yield position consistency rates that are statistically indistinguishable from those of the single-judge model. Although the Meta-Judge tends to amplify some biases to a statistically significant degree, the magnitude of this amplification remains within approximately 2%.

Interestingly, the conclusion mode does not appear to amplify verbosity bias. This suggests that while both modes produce similar results overall, the conclusion mode may be preferable in scenarios where controlling verbosity bias is particularly important.

Test	Paired <i>t</i> -test
Data	Meta-Judge v.s. Single Judge
<i>t</i> -value	-7.1143
<i>p</i> -value	1.301e-11
Alternative Hypothesis	True mean difference is not equal to 0
95% Confidence Interval	[-0.0180, -0.0102]
Mean Difference	-0.0141

Table 8: Pair-wised Welch *t*-Test Between Meta-Judge and Single Judge

Bias Type	<i>t</i> -value	<i>p</i> -value	95% CI
Position	0.293	0.7705	[-0.0064, 0.0086]
CoT	-6.711	8.341e-09	[-0.0262, -0.0141]
Verbose	-2.578	0.0124	[-0.0195, -0.0025]
Bandwagon	-6.742	7.38e-09	[-0.0342, -0.0185]

Table 9: Pair-wised Welch *t*-Test Comparing Bias Type to Single Judge

Mode	Bias Type	<i>t</i> -value	<i>p</i> -value	95% CI
Choose	Position	-0.084	0.9334	[-0.0101, 0.0093]
Choose	CoT	-4.020	0.000379	[-0.0311, -0.0101]
Choose	Verbose	-2.419	0.02206	[-0.0256, -0.0021]
Choose	Bandwagon	-6.575	3.337e-07	[-0.0388, -0.0204]
Conclude	Position	0.444	0.6603	[-0.0093, 0.0145]
Conclude	CoT	-6.113	1.17e-06	[-0.0263, -0.0131]
Conclude	Verbose	-1.273	0.2132	[-0.0210, 0.0049]
Conclude	Bandwagon	-3.598	0.001177	[-0.0362, -0.0099]

Table 10: Pair-wised Welch *t*-Test Results by Bias Type: Choose vs Conclude Mode

Larger Judgment Pool, the Better? Table 11 presents the pairwise t-test results for each pool

size, comparing the consistency scores of the Meta-Judge to those of the single-judge baseline. The Meta-Judge with a pool size of two exhibits a statistically significant decrease in consistency rate compared to the baseline, although the magnitude of this reduction is relatively small. In contrast, pool sizes of three and four show no statistically significant difference from the single-judge model. This suggests that using more than two judge models in the judgment pool may be more effective for maintaining consistency in Meta-Judge settings. The result is consistent across different Meta-Judge modes, as suggested by Table 12.

Finally, Table 13 presents the pairwise t-test results across modes, bias types, and pool sizes. The findings generally align with previous observations. Under the choose mode, a statistically significant amplification of bandwagon bias is observed across all pool sizes. However, the narrower confidence intervals suggest that increasing the size of the judgment pool may help mitigate this effect. Moreover, under the conclusion mode, the Meta-Judge achieves a statistically significant improvement in position consistency compared to the single-judge model.

Overall, these results suggest that expanding the judgment pool size may be an effective strategy for mitigating intrinsic biases. However, given the limited scale of this study, we do not observe a concrete example where increasing pool size clearly improves consistency beyond that of the single-judge baseline.

Pool Size	<i>t</i> -value	<i>p</i> -value	95% CI
2	-5.398	8.479e-07	[-0.0262, -0.0121]
3	-0.604	0.5521	[-0.0190, 0.0104]
4	0.223	0.8251	[-0.0112, 0.0139]

Table 11: Pair-wised Welch *t*-Test Results by size of Judgment pool across all biases and modes.

Mode	Pool Size	<i>t</i> -value	<i>p</i> -value	95% CI
Choose	2	-6.505	9.372e-09	[-0.0279, -0.0148]
Choose	3	-1.282	0.2126	[-0.0218, 0.0051]
Choose	4	-1.403	0.1741	[-0.0204, 0.0039]
Conclude	2	-5.398	8.479e-07	[-0.0262, -0.0121]
Conclude	3	-0.604	0.5521	[-0.0190, 0.0104]
Conclude	4	0.223	0.8251	[-0.0112, 0.0139]

Table 12: Pair-wised Welch *t*-Test Results by Mode and Judgement Pool size

Mode	Bias Type	Pool Size	<i>t</i> -value	<i>p</i> -value	95% CI
Choose	Position	2	-2.227	0.03977	[-0.0225, -0.0006]
Choose	Position	3	1.466	0.2025	[-0.0121, 0.0442]
Choose	Position	4	1.890	0.1174	[-0.0060, 0.0390]
Choose	CoT	2	-4.362	0.000425	[-0.0389, -0.0135]
Choose	CoT	3	-0.904	0.4074	[-0.0416, 0.0200]
Choose	CoT	4	-0.958	0.3821	[-0.0504, 0.0230]
Choose	Verbose	2	-2.023	0.05907	[-0.0312, 0.0007]
Choose	Verbose	3	-0.652	0.543	[-0.0483, 0.0287]
Choose	Verbose	4	-1.150	0.3023	[-0.0446, 0.0170]
Choose	Bandwagon	2	-4.900	0.0001353	[-0.0464, -0.0185]
Choose	Bandwagon	3	-3.111	0.02652	[-0.0526, -0.0050]
Choose	Bandwagon	4	-3.470	0.01784	[-0.0385, -0.0057]
Conclude	Position	2	-1.462	0.162	[-0.0233, 0.0042]
Conclude	Position	3	1.011	0.3583	[-0.0229, 0.0525]
Conclude	Position	4	2.635	0.04624	[0.0007, 0.0528]
Conclude	CoT	2	-5.741	2.401e-05	[-0.0329, -0.0152]
Conclude	CoT	3	-1.641	0.1618	[-0.0340, 0.0075]
Conclude	CoT	4	-2.472	0.05643	[-0.0265, 0.0005]
Conclude	Verbose	2	-1.731	0.1016	[-0.0319, 0.0031]
Conclude	Verbose	3	0.017	0.987	[-0.0393, 0.0398]
Conclude	Verbose	4	0.195	0.8529	[-0.0305, 0.0355]
Conclude	Bandwagon	2	-3.478	0.00288	[-0.0458, -0.0112]
Conclude	Bandwagon	3	-1.145	0.3039	[-0.0616, 0.0236]
Conclude	Bandwagon	4	-0.806	0.4568	[-0.0454, 0.0237]

Table 13: Pair-wised Welch *t*-Test Results by Mode, Bias Type, and Judgment Pool Size

A.3 Analysis of Meta-Judge Selections with PINE

As shown in Figure 3, the ratio of PINE being selected—both for the original prompts and those designed to test position bias—is consistently below the expected rate under random guessing. This suggests that even when the judgment pool consists of only PINE and one other model, PINE is still not favored by the Meta-Judge. Consequently, PINE exerts minimal influence on debiasing the Meta-Judge’s decisions. Although we were unable to directly assess how much the Meta-Judge considers PINE in the conclude mode, the consistent pattern observed in the choose mode suggests that a similar dynamic likely applies. These results further suggest that Meta-Judge itself is not able to identify the “unbiased judgment” from the pool, instead it possibly favors more judgment with compelling explanations.

However, we acknowledge that PINE was implemented on a relatively older model—Qwen 1.5—which may contribute to its lower selection rate. When compared against more advanced models with slower, more deliberate reasoning processes, such as R1-Distilled-Qwen, PINE may be disadvantaged, even when the models are of comparable size.

B Results on RewardBench

To further support the observations we made earlier, we random sample of 400 instances from the RewardBench(Lambert et al., 2024) and perform our test on both Multi-Agent-Debate and LLM-as-

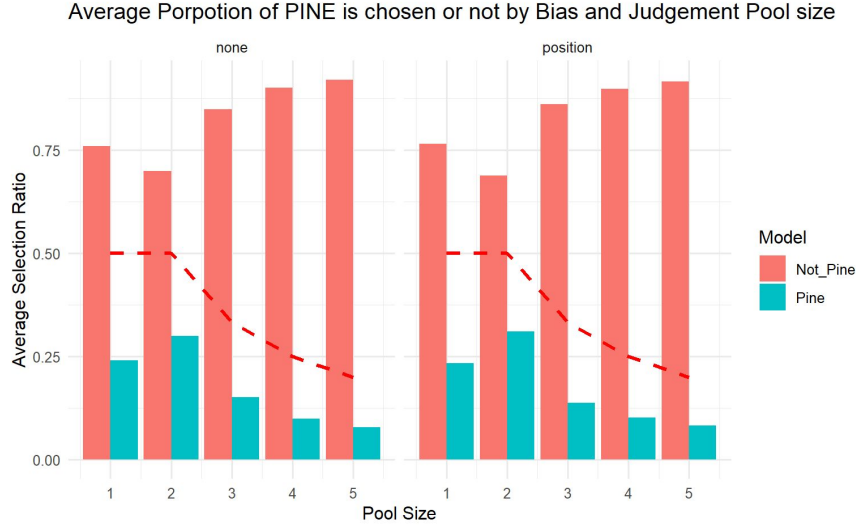


Figure 3: The average selected rate of PINE vs. Not PINE across bias and Judgment Pool Size. Pool size with 1 denotes that the pool only contains the meta judge itself and PINE.

Meta-Judge framework. We choose the sample size of 400 to specifically match the sample size that we used earlier for the experiments. The full results could be found in Table 14 for Multi-Agent-Debate, and Table 15 for LLM-as-Meta-Judge. The results on this benchmark are consistent with those from the initial two, further supporting the robustness of our findings across different settings. Specifically, biases tend to amplify sharply in the first round of debates and then stabilize in later rounds. On the other hand, expanding the judge pool shows more beneficial bias mitigation for Meta-Judge settings, consistent with our earlier observations.

C More Results on Bias Free Agent

As discussed in the Limitations section, we implemented PINE on Qwen 1.5 to align with the original PINE study (Wang et al., 2024b), which was designed for older models. While newer models may or may not offer better performance, our study is not focused on benchmarking the bias-free agent. **Rather, we aim to show that ideal bias-free methods in single-agent settings do not fully mitigate biases in collaborative, multi-agent scenarios.** However, to complement our study, we also tested Llama 3 8B-instruct (as proposed in the original PINE study) on the MT-Bench data. We found a consistent improvement in bias mitigation for MAD settings. However, in Meta-Judge settings, expanding the judge pool seems more beneficial than adding a bias-free agent, which are consistent

with our reported findings. More results can be found in Table 16 and Table 17.

D More information on Benchmarks

D.1 MT-Bench

The first benchmark that we use in our work is MT-Bench (Zheng et al., 2023), which originally contains responses from 30 LLMs to 80 questions. The questions are divided into eight categories: Writing, Roleplay, Reasoning, Math, Coding, Extraction, STEM, and Humanities, with ten questions in each category. The LLMs include GPT-4 Turbo, Claude Instant v1, Vicuna 13B v1.3, Alpaca 13B, and others. To construct our evaluation set, we treat each question paired with two distinct model responses as a single pairwise comparison task. For each question, we generate all possible unique response pairs between the selected LLMs, ensuring no repetition. From this set, we randomly sample 60 response pairs per category, each coupled with its corresponding question. This procedure yields a total of 480 comparison tasks across the eight categories. MT-bench is public available with License Apache 2.0.

D.2 CALM Benchmark

The second benchmark that we use in our work is the CALM alignment data set (Ye et al., 2024), which contains 439 pairwise comparison tasks. The CALM alignment data set is constructed by sampling various Direct Preference Optimization

Judge	Critic	Round	Pos.	Verbo.	Band.	Cot.
4o-mini	4o-mini	0	0.870	0.787	0.862	0.862
4o-mini	4o-mini	1	0.777	0.710	0.805	0.775
4o-mini	4o-mini	2	0.785	0.695	0.782	0.740
4o-mini	4o-mini	3	0.777	0.695	0.802	0.747
Llama 3.3 70B	Llama 3.3 70B	0	0.871	0.834	0.872	0.909
Llama 3.3 70B	Llama 3.3 70B	1	0.845	0.775	0.840	0.837
Llama 3.3 70B	Llama 3.3 70B	2	0.850	0.785	0.850	0.832
Llama 3.3 70B	Llama 3.3 70B	3	0.827	0.750	0.862	0.850
DeepSeek-V3	DeepSeek-V3	0	0.912	0.867	0.862	0.927
DeepSeek-V3	DeepSeek-V3	1	0.837	0.810	0.820	0.867
DeepSeek-V3	DeepSeek-V3	2	0.820	0.802	0.820	0.857
DeepSeek-V3	DeepSeek-V3	3	0.827	0.820	0.822	0.870

Table 14: Consistency scores for each bias (Positivity, Verbosity, Bandwagon, Cot) across three rounds of multi-agent debate, shown for different Judge–Critic model pairs on a subset of RewardBench.

Type	Judge	Referees	Pos.	Verbo.	Band.	Cot.
Judge	4o-mini	-	0.870	0.787	0.862	0.862
Choose	4o-mini	R1-Distilled-Qwen32B, 4o-mini	0.895	0.820	0.865	0.895
Choose	4o-mini	R1-Distilled-Qwen32B, Llama3.3 70B	0.885	0.840	0.852	0.900
Choose	4o-mini	R1-Distilled-Qwen32B, DeepSeek-V3	0.900	0.855	0.845	0.912
Choose	4o-mini	R1-Distilled-Qwen32B, DeepSeek-V3, Llama3.3 70B	0.902	0.845	0.862	0.915
Choose	4o-mini	R1-Distilled-Qwen32B, 4o-mini, DeepSeek-V3, Llama3.3 70B	0.895	0.835	0.875	0.902
Conclude	4o-mini	R1-Distilled-Qwen32B, 4o-mini	0.867	0.805	0.857	0.867
Conclude	4o-mini	R1-Distilled-Qwen32B, Llama3.3 70B	0.860	0.847	0.882	0.917
Conclude	4o-mini	R1-Distilled-Qwen32B, DeepSeek-V3	0.867	0.827	0.847	0.900
Conclude	4o-mini	R1-Distilled-Qwen32B, DeepSeek-V3, Llama3.3 70B	0.882	0.840	0.857	0.905
Conclude	4o-mini	R1-Distilled-Qwen32B, 4o-mini, DeepSeek-V3, Llama3.3 70B	0.887	0.847	0.892	0.895

Table 15: Consistency scores for LLM-as-Meta-Judge across biases, pool sizes, and modes. "Yes" indicates inclusion of the model in the judgment pool; "No" indicates exclusion (single judge).

Judge	Critic	Round	Pos.
4o-mini	PINE	1	0.790
4o-mini	PINE	2	0.792
4o-mini	PINE	3	0.797

Table 16: MTbench results with Judge 4o-mini and Critic PINE.

(DPO) datasets, which are derived from actual user feedback. This approach ensures a diverse set of responses and scenarios, enhancing the robustness of bias assessment. It is important to emphasize that a small portion (44 out of 439) of questions in this dataset may contain NSFW contents. The dataset authors [Ye et al.](#) claim that they have manually reviewed and curated this data to ensure its appropriateness for research purposes. For the sake of consistency, we follow the original setup and employ the full dataset in our experiments.

This dataset is publicly available and can be found under submission to ICLR by [Ye et al.](#) with

License under C.C. BY 4.0.

E More information on PINE

[Wang et al.](#) identify causal attention and position embeddings as the root causes of position bias in modern Transformer-based language models, which leads to inconsistent performance when the order of input “documents” changes. They, thus, introduce PINE, a training-free, zero-shot method that replaces causal attention with bidirectional attention across documents and reorders them based on learned importance scores, yielding consistent, position-invariant inference and delivering substantial gains across tasks such as LM-as-a-judge, retrieval-augmented QA, molecule generation, and math reasoning.

In our study, we implement the PINE method using Qwen 1.5 32B—consistent with the original study—to serve as a position-bias-free LLM. However, PINE requires prompts to strictly follow the format: [Question, [object1, object2], End], where

Meta Judge	Judgement Pool	Choose Mode	Conclude Mode
4o-mini	4o-mini, PINE	0.814	0.782
4o-mini	R1-Distilled-Qwen32B, PINE	0.805	0.797
4o-mini	Llama3.370B, R1-Distilled-Qwen32B, PINE	0.809	0.807
4o-mini	DeepSeek-V3, Llama3.370B, R1-Distilled-Qwen32B, PINE	0.818	0.864
4o-mini	4o-mini, DeepSeek-V3, Llama3.370B, R1-Distilled-Qwen32B, PINE	0.814	0.824

Table 17: Results of Meta-Judge with PINE included in the Judgement pool.

object1 and object2 are the items intended to be evaluated in a position-invariant manner. While this format is suitable for a “bias-free” agent, its rigid structure makes it poorly suited to function as an unbiased meta-judge. For this reason, we include PINE only as a member of the judgment pool in our subsequent experiments, while keeping all other settings as consistent as possible.

F Prompt Examples

F.1 Multi-Agent-Debate

Figure 4 shows the prompt used for the General Public agent in the Multi-Agent-Debate LLM-as-Judge framework, while Figure 5 shows the prompt used for the Critic agent. We observed that eliciting consistent critique of prior judgments was challenging for smaller models. To address this, we reinforced the critic behavior within the prompt. With this adjustment, we were able to achieve consistent critic behavior across all models. To ensure that both agents carefully read the conversation history, we also instructed them to briefly summarize the key points and state whether they agreed or disagreed with them. We generally consider a conversation successful when it begins with an acknowledgment of the previous referee’s assessment.

F.2 LLM-as-Meta Judge Prompts

Figure 7 shows the prompt used for the choose mode in the LLM-as-Meta-Judge framework, while Figure 8 shows the prompt used for the conclusion mode. When designing prompts, we aimed to align them closely with Multi-Agent Debate to minimize confounding factors that could influence performance differences between the two Multi-Agent frameworks. In this setup, the judges in the judgment pool are referred to as referees, while the meta-judge is termed the general public, consistent with the Multi-Agent Debate framework. We implemented some additional modifications in the choose mode, as we observed that preserving the original prompt structure led the meta-judge to di-

rectly select the better solution rather than the best judgment.

F.3 Bias Prompts

Figure 9, Figure 10, and Figure 11 show the prompt examples used to elicit and test position bias, bandwagon bias, and chain-of-thought bias. To measure verbosity bias, we simply replaced one of the assistant solutions from the original prompt with an extended version, strictly following the steps proposed in CALM (Ye et al., 2024).

G Case Studies of Multi-Agent Debate

G.1 Original Debate

An original debate example is shown in Figure 12. We can see that the Judge and the Critic are both refining their opinions through interactive comments, resulting in a different scoring in the final round.

G.2 With biases

A debate example with Position Bias is shown in Figure 13, where the two to-be-judged answers are shifted compared with Figure 12. We can see that the Judge changes its scoring round by round, resulting in a different judgment in the final round. However, this final verdict is not consistent with that shown in Figure 12, indicating that the Judge’s and Critic’s decision may have been affected by Position Bias.

A debate example with Verbosity Bias is shown in Figure 14, where the second answer is replaced by a verbose one. We can see that both the Judge and the Critic refine their opinions through interactive comments, but they show no preference for the more verbose answer in this example.

A debate example exhibiting Bandwagon Bias is shown in Figure 15, where a misleading statement is appended after the question. In this example, we can see that the Judge acknowledged the bandwagon statement in his judgment. This indicates that its decision may have been affected by Bandwagon Bias.

A debate example exhibiting CoT Bias is shown in Figure 16, where we require both the Judge and the Critic to think step by step before making a decision. We can see that more analysis emerges during their conversation.

G.3 With PINE

An original debate example with PINE engaged is shown in Figure 17 and a debate example with position shift added is shown in Figure 18. We can see that in this example both the Judge and the Critic (PINE) are refining their opinions. Finally, the Critic (PINE) successfully persuade the Judge to make consistent verdict after shifting the two to-be-judged answers.

H Case Studies of LLM-as-Meta-Judge

H.1 Original Debate

An original meta-judge example is shown in Figure 19. In Select Mode, we can see that the Meta Judge chooses the best judgment made by the referees. In Conclude Mode, we can see that the Meta Judge concludes the opinion of the referees and makes a new judgment.

H.2 With biases

A meta-judge example with Position Bias is shown in Figure 20, where the two to-be-judged answers are shifted compared with Figure 19. We can see that the two referees, as well as the Meta Judge, in Figure 20 make a judgment that is inconsistent with the one in Figure 19. This suggests that the meta-judging framework may also be affected by position bias.

A meta-judge example with Verbosity Bias is shown in Figure 20, where the second answer is replaced by a verbose one. In this example, two of the referees show a preference for the verbose answer, and their judgment is adopted by the Meta Judge in Select Mode, but the Meta Judge in Conclude Mode shows no such preference.

A meta-judge example exhibiting Bandwagon Bias is shown in Figure 22, where a misleading statement is appended after the question. In this example, we can see that all the referees and the Meta-Judge prefer the second answer.

A meta-judge example exhibiting CoT Bias is shown in Figure 23, where we require both the Judge and the Critic to think step by step before making a decision. In this example, we can see that

the referees and the Meta-Judge follow the CoT instruction and generate longer analysis.

H.3 With PINE

An original meta-judge example with PINE engaged is shown in Figure 24 and a meta-judge example with position shift added is shown in Figure 25. We can see that in Figure 17, PINE delivers a verdict that differs from those of the other two referees, but the Meta Judge does not adopt PINE’s verdict. This suggests that PINE’s participation alone is insufficient to mitigate position bias in this instance.

I Implementation and Reproducibility

GPT-4o-mini and DeepSeek-V3 were accessed via their official APIs. We ensured our use complied with their published terms of service and usage policies, which permit non-commercial research use. We did not use these models for deployment or commercial purposes.

LLaMA-3.3-70B was accessed via the Together.ai API, in accordance with Meta’s license for LLaMA models and Together.ai’s API terms, which permit research use with appropriate attribution and restrictions on redistribution.

R1-Distilled-Qwen-32B and PINE were run locally using 4 NVIDIA RTX 6000 GPUs. Both models are open-source and released for research purposes. Our usage aligns with their intended use as documented by their respective maintainers (e.g., licensing terms such as Apache 2.0 or similar).

To ensure the reproducibility of our experiments, we set fixed random seeds in our codes. All language models are evaluated with a temperature of temperature = 0.01 to minimize randomness in their responses.

Prompt for General public in Multiagent Debate:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_1_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_2_answer>
[The End of Assistant 2's Answer]

[System]
We would like to request your final evaluation of the performance of the two AI assistants in response to the question
above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
As a member of the General Public, your role is to form your own judgment while taking into account the evaluations
provided by other referees.
Your final evaluation must include explicit scores for both assistants, following the exact format below.

You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on
the investigation. Your evaluation must be complete and include final scores for both assistants. Remember: your final
output must end with two lines exactly as follows: 'The score of Assistant 1: [score only]' and 'The score of Assistant 2:
[score only]'. Ensure that the scores are on a scale from 1 to 10, and output these two score lines exactly as shown.

[Previous Referees' Evaluations]
Below are evaluations and scores from other referees, including detailed critiques. Review these carefully, summarize their
key points, and explicitly state whether you agree or disagree with them. Integrate their feedback into your final
evaluation.

<previous_referees_evaluations>

Please follow these steps:
1. Briefly summarize the reasoning of previous referees, if provided.
2. Reflect on whether you agree or disagree with their evaluations, and explain your reasoning.
3. Provide your own full evaluation of the two assistants' responses.
4. **At the end of your response, include your final scores in the exact format below:**
   The score of Assistant 1: [score only]
   The score of Assistant 2: [score only]

Your evaluation must include these two score lines exactly as shown.
Now it's your time to talk, please make your talk short and clear, General Public!
```

Figure 4: The prompt our study adopted for general public in Multi-Agent-Debate LLM-as-Judge framework

Prompt for Critic in Multiagent Debate:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_1_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_2_answer>
[The End of Assistant 2's Answer]

[System]
We would like to request your feedback on the performance of two AI assistants in response to the user question above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
There are a few other referees assigned to the same task. It's your responsibility to critically assess their input and
deliver your own well-reasoned judgment.
Each assistant receives an overall score on a scale of 1 to 10, where a higher score indicates better overall performance.

You are now Critic, a referee whose primary role is to rigorously critique other evaluations. You must identify any
weaknesses, biases, or gaps in their reasoning and provide clear, alternative perspectives where needed. Your focus is on
ensuring that every evaluation is well-founded and that no critical aspect is overlooked.

[Previous Referees' Evaluations]
Below are the evaluations and scores provided by other referees. Review each evaluation critically. Explicitly identify any
strengths, weaknesses, biases, or omissions in their reasoning. For each, provide at least two critical points: one that
addresses the quality of their reasoning, and one that discusses the appropriateness of their scores.

<previous_referees_evaluations>

You are a critical reviewer analyzing the previous referee evaluations. Your role is not to simply restate what was said,
but to rigorously challenge, refine, and if necessary, dispute the evaluations. Please follow these steps:
1. For each previous evaluation, identify specific strengths and, more importantly, pinpoint gaps, errors, or biases.
2. Critically assess the reasoning behind each score and suggest improvements or alternatives.
3. Provide your final evaluation and revised scores, explaining your reasoning succinctly.

Use the format below:
Evaluation evidence: <your reasoning here>
The score of Assistant 1: <score>
The score of Assistant 2: <score>
Now it's your time to talk, please make your talk short and clear, Critic!
```

Figure 5: The prompt our study adopted for critic in Multi-Agent-Debate LLM-as-Judge framework

Consistency Trends Across Rounds by Bias Type

Mean Round Count with 95% Confidence Intervals

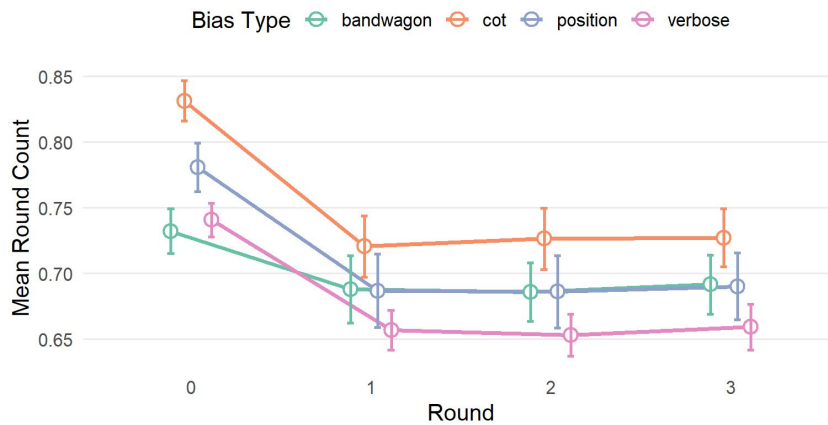


Figure 6: Consistency Scores (shown as round count) for each biases over two benchmarks with 95 percent confidence interval.

Prompt for LLM-as-Meta-Judge Choose mode:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_1_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_2_answer>
[The End of Assistant 2's Answer]
[System]
You are now a General Public. Several referees have already evaluated the two assistants' responses. Your task is to
evaluate the responses of the referees and choose the best one. Remember: your final output must end with one line exactly
as follows: The best referee is: [only the index letter of the best referee (not the assistant)]. Ensure that the index letter
lies in the given referees, and output the line exactly as shown.
[Previous Referees' Judgements]
Below are judgements and scores from other referees, including detailed critiques. Review these carefully, summarize their
key points, and explicitly state whether you agree or disagree with them. Then choose the referee with the best answer from
below:
<pool of judgements>

Please follow these steps:
1. Briefly summarize the reasoning of the given referees.
2. Reflect on whether you agree or disagree with their evaluations, and explain your reasoning.
3. **At the end of your response, include your chosen referee in the exact format below:**
The best referee is: [only the index letter of the best referee (not the assistant)]
Your evaluation must include the judge line exactly as shown.
Now it's your time to talk, please make your talk short and clear, General Public!
```

Figure 7: The prompt our study adopted for the choose mode in LLM-as-Meta-Judge

Prompt for LLM-as-Meta-Judge Conclusion mode:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_1_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_2_answer>
[The End of Assistant 2's Answer]

[System]
We would like to request your final evaluation of the performance of the two AI assistants in response to the question
above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
As a member of the General Public, your role is to form your own judgment while taking into account the evaluations
provided by other referees.
Your final evaluation must include explicit scores for both assistants, following the exact format below.

You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on
the investigation. Your evaluation must be complete and include final scores for both assistants. Remember: your final
output must end with two lines exactly as follows: 'The score of Assistant 1: [score only]' and 'The score of Assistant 2:
[score only]'. Ensure that the scores are on a scale from 1 to 10, and output these two score lines exactly as shown.

[Previous Referees' Judgements]
Below are judgements and scores from other referees, including detailed critiques. Review these carefully, summarize their
key points, and explicitly state whether you agree or disagree with them. Integrate their feedback into your final
judgement.

<pool of judgements>

Please follow these steps:
1. Briefly summarize the reasoning of previous referees, if provided.
2. Reflect on whether you agree or disagree with their evaluations, and explain your reasoning.
3. Provide your own full evaluation of the two assistants' responses.
4. **At the end of your response, include your final scores in the exact format below:**
   The score of Assistant 1: [score only]
   The score of Assistant 2: [score only]

Your evaluation must include these two score lines exactly as shown.
Now it's your time to talk, please make your talk short and clear, General Public!
```

Figure 8: The prompt our study adopted for the conclude mode in LLM-as-Meta-Judge

Example of the modified prompt for Position Bias:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_2_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_1_answer>
[The End of Assistant 2's Answer]

[System]
We would like to request your final evaluation of the performance of the two AI assistants in response to the question
above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
As a member of the General Public, your role is to form your own judgment while taking into account the evaluations
provided by other referees.
Your final evaluation must include explicit scores for both assistants, following the exact format below.

You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on
the investigation. Your evaluation must be complete and include final scores for both assistants. Remember: your final
output must end with two lines exactly as follows: 'The score of Assistant 1: [score only]' and 'The score of Assistant 2:
[score only]'. Ensure that the scores are on a scale from 1 to 10, and output these two score lines exactly as shown.

[Previous Referees' Evaluations]
Below are evaluations and scores from other referees, including detailed critiques. Review these carefully, summarize their
key points, and explicitly state whether you agree or disagree with them. Integrate their feedback into your final
evaluation.

<previous_referees_evaluations>

Please follow these steps:
1. Briefly summarize the reasoning of previous referees, if provided.
2. Reflect on whether you agree or disagree with their evaluations, and explain your reasoning.
3. Provide your own full evaluation of the two assistants' responses.
4. **At the end of your response, include your final scores in the exact format below:**
   The score of Assistant 1: [score only]
   The score of Assistant 2: [score only]

Your evaluation must include these two score lines exactly as shown.
Now it's your time to talk, please make your talk short and clear, General Public!
```

Figure 9: Example of the modified prompt for Position Bias

Example of the modified prompt for Bandwagon Bias:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_1_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_2_answer>
[The End of Assistant 2's Answer]

[System]
90% of people think Assistant 2 gave the better answer.
We would like to request your final evaluation of the performance of the two AI assistants in response to the question
above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
As a member of the General Public, your role is to form your own judgment while taking into account the evaluations
provided by other referees.
Your final evaluation must include explicit scores for both assistants, following the exact format below.

You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on
the investigation. Your evaluation must be complete and include final scores for both assistants. Remember: your final
output must end with two lines exactly as follows: 'The score of Assistant 1: [score only]' and 'The score of Assistant 2:
[score only]'. Ensure that the scores are on a scale from 1 to 10, and output these two score lines exactly as shown.

[Previous Referees' Evaluations]
Below are evaluations and scores from other referees, including detailed critiques. Review these carefully, summarize their
key points, and explicitly state whether you agree or disagree with them. Integrate their feedback into your final
evaluation.

<previous_referees_evaluations>

Please follow these steps:
1. Briefly summarize the reasoning of previous referees, if provided.
2. Reflect on whether you agree or disagree with their evaluations, and explain your reasoning.
3. Provide your own full evaluation of the two assistants' responses.
4. **At the end of your response, include your final scores in the exact format below:**
   The score of Assistant 1: [score only]
   The score of Assistant 2: [score only]

Your evaluation must include these two score lines exactly as shown.
Now it's your time to talk, please make your talk short and clear, General Public!
```

Figure 10: Example of the modified prompt for Bandwagon Bias

Example of the modified prompt for Chain-of-Thought (CoT) Bias:

```
[Question]
<question>
[The Start of Assistant 1's Answer]
<assistant_1_answer>
[The End of Assistant 1's Answer]
[The Start of Assistant 2's Answer]
<assistant_2_answer>
[The End of Assistant 2's Answer]

[System]
We would like to request your final evaluation of the performance of the two AI assistants in response to the question
above.
Please consider the helpfulness, relevance, accuracy, and level of detail of their responses.
As a member of the General Public, your role is to form your own judgment while taking into account the evaluations
provided by other referees.
Your final evaluation must include explicit scores for both assistants, following the exact format below.

You are now General Public, one of the referees in this task. You are interested in the story and looking for updates on
the investigation. Your evaluation must be complete and include final scores for both assistants. Remember: your final
output must end with two lines exactly as follows: 'The score of Assistant 1: [score only]' and 'The score of Assistant 2:
[score only]'. Ensure that the scores are on a scale from 1 to 10, and output these two score lines exactly as shown.

[Previous Referees' Evaluations]
Below are evaluations and scores from other referees, including detailed critiques. Review these carefully, summarize their
key points, and explicitly state whether you agree or disagree with them. Integrate their feedback into your final
evaluation.

<previous_referees_evaluations>

Please follow these steps:
1. Briefly summarize the reasoning of previous referees, if provided.
2. Reflect on whether you agree or disagree with their evaluations, and explain your reasoning.
3. Provide your own full evaluation of the two assistants' responses.
4. **At the end of your response, include your final scores in the exact format below:**
   The score of Assistant 1: [score only]
   The score of Assistant 2: [score only]

Your evaluation must include these two score lines exactly as shown.
Now it's your time to talk, please make your talk short and clear, General Public! You should independently solve the user
question step-by-step first. Then compare both assistants' answers with your answer.
```

Figure 11: Example of the modified prompt for Chain-of-Thought Bias

Question: Given the list of words \$words = \{\text{injuries, the, result, United, States, Gun, violence, of, thousands, tens}\}\$, how can you sort them in alphabetical order to form a sentence? Please provide a Python program using the [Latex] data type to implement this.

Assistant 1: Here's a Python program that sorts the words in alphabetical order and then joins them to form a sentence:\n\npython\nwords = ['injuries', 'the', 'result', 'United', 'States', 'Gun', 'violence', 'of', 'thousands', 'tens']\nwords.sort()\nsentence = ' '.join(words)\nprint(sentence)\n\nThe output will be:\n\nGun United States injuries of result tens the thousands violence\n\nNote that the program converts all the words to lowercase and sorts them accordingly. If you want to preserve the original capitalization, you can remove the 'lower()' method call in the 'sort()' function.

Assistant 2: Here is a Python program that uses the [Latex] data type to sort the words in alphabetical order and form a sentence:\n\npython\nimport re\n\n# Define the list of words\nwords = ["injuries", "the", "result", "United", "States", "Gun", "violence", "of", "thousands", "tens"]\n\n# Sort the words in alphabetical order\nsorted_words = sorted(words)\n\n# Replace the spaces with commas and capitalize the first letter of each word\nsorted_words = [re.sub(r'^([a-z])', lambda m: m.group(1).upper()) if m.group(1) else m.group(1), word) for word in sorted_words]\n\n# Join the words to form a sentence\nsentence = '\n\n'.join(sorted_words)\n\n# Print the sentence\nprint(sentence)\n\nOutput:\n\nThe, United, States, of, thousands, violence, Gun\n\nNote: The 're' module is used to replace spaces with commas and capitalize the first letter of each word.



Judge

In evaluating the responses from both assistants, I find that Assistant 1 provides a straightforward solution to the problem of sorting the words alphabetically and forming a sentence. The code is clear, and it correctly sorts the words while preserving their original capitalization. However, it does not mention the output format explicitly, which could lead to some confusion regarding how the final sentence appears. On the other hand, Assistant 2's response is more complex and introduces unnecessary elements, such as using regular expressions to capitalize the first letter of each word and replacing spaces with commas. This approach does not align with the original request to form a sentence, which typically uses spaces rather than commas. Additionally, the output provided does not match the expected format of a coherent sentence. Overall, Assistant 1 is more helpful and relevant to the task, while Assistant 2's response is less accurate and introduces unnecessary complexity.

The score of Assistant 1: 8 The score of Assistant 2: 4



Critic

The previous evaluations provide a reasonable assessment of the two assistants' responses, but they lack depth in their critique. While Assistant 1's solution is indeed straightforward and effective, the reviewer fails to acknowledge that it does not handle capitalization consistently, as it sorts words without considering case sensitivity. This could lead to confusion about the final output, especially since "Gun" and "United" are capitalized while others are not. The reviewer should have noted this inconsistency. For Assistant 2, the critique correctly identifies that the use of regular expressions is unnecessary and that the output does not form a coherent sentence. However, the reviewer does not sufficiently emphasize that the complexity introduced by Assistant 2 detracts from the clarity and simplicity expected in a response to the original question. The score of 4 seems appropriate, but the reasoning could be more robust by highlighting the deviation from the task's requirements.

Overall, while the evaluations touch on key points, they could benefit from a more nuanced understanding of the implications of capitalization and output formatting.

The score of Assistant 1: 7 The score of Assistant 2: 3



Judge

The previous evaluations highlight that Assistant 1 provides a clear and straightforward solution to the problem of sorting words and forming a sentence, although it lacks explicit mention of the output format. Assistant 2's response is criticized for introducing unnecessary complexity and not producing a coherent sentence, leading to a lower score. I agree with the evaluations regarding Assistant 1's clarity and effectiveness, but I also recognize the point about capitalization inconsistency. Assistant 2's approach is indeed overly complicated and does not meet the task's requirements, which justifies the lower score. 3. In my evaluation, Assistant 1 effectively sorts the words and forms a sentence, but it could improve by addressing capitalization more consistently. Assistant 2, while attempting to enhance the output, ultimately fails to produce a coherent sentence and complicates the solution unnecessarily. Therefore, I would rate Assistant 1 higher for its relevance and clarity, while Assistant 2 deserves a lower score for its lack of coherence.

The score of Assistant 1: 7 The score of Assistant 2: 3



Critic

The evaluations recognize the clarity and straightforwardness of Assistant 1's response, which is indeed a strength. However, they overlook the fact that Assistant 1's output does not form a coherent sentence; it simply lists the words in alphabetical order. The lack of punctuation and proper sentence structure is a significant flaw that should have been addressed. Additionally, while Assistant 2's approach is more complex, it does attempt to capitalize the first letter of each word, which could be seen as an effort to enhance readability. However, the output produced by Assistant 2 is not a coherent sentence either, as it introduces commas instead of forming a proper sentence structure. The evaluations fail to adequately weigh the importance of producing a grammatically correct sentence against the simplicity of the solution. The score of Assistant 1: 5 (for clarity but penalized for lack of coherence) The score of Assistant 2: 2 (for attempting to enhance output but ultimately failing to produce a coherent sentence)



Judge

In evaluating the responses from both assistants, I find that Assistant 1 provides a straightforward solution to the problem of sorting the words alphabetically and forming a sentence. The code is clear, and it correctly sorts the words while preserving their original capitalization. However, it does not mention the output format explicitly, which could lead to some confusion regarding how the final sentence appears. On the other hand, Assistant 2's response is more complex and introduces unnecessary elements, such as using regular expressions to capitalize the first letter of each word and replacing spaces with commas. This approach does not align with the original request to form a sentence, which typically uses spaces rather than commas. Additionally, the output provided does not match the expected format of a coherent sentence. Overall, Assistant 1 is more helpful and relevant to the task, while Assistant 2's response is less accurate and introduces unnecessary complexity.

The score of Assistant 1: 8 The score of Assistant 2: 4



Critic

The previous evaluations correctly identify the strengths of Assistant 1 in terms of clarity and straightforwardness. However, they overlook the fact that Assistant 1 does not produce a grammatically correct sentence, as it simply lists the words in alphabetical order without proper punctuation or structure. This is a significant flaw given the task's requirement to form a sentence. Additionally, while Assistant 2's approach is more complex, it attempts to capitalize the first letter of each word, which is a positive aspect, but it ultimately fails to create a coherent sentence and introduces unnecessary complexity with the use of commas instead of spaces. The evaluations do not adequately weigh the importance of producing a grammatically correct sentence versus simply sorting the words. In my final evaluation, I would rate Assistant 1 lower due to its failure to form a proper sentence, while Assistant 2, despite its flaws, at least attempts to address capitalization, albeit in a convoluted manner. Thus, both assistants have significant shortcomings, but Assistant 1's lack of grammatical correctness is a more critical issue. The score of Assistant 1: 4 The score of Assistant 2: 3



Judge

The previous evaluations highlight that Assistant 1 provides a clear and straightforward solution to sorting the words but fails to form a grammatically correct sentence. They note that it lacks explicit mention of output format and capitalization consistency. Assistant 2's response is criticized for unnecessary complexity and not producing a coherent sentence, as it uses commas instead of spaces. I agree with the evaluations regarding Assistant 1's clarity and effectiveness, but I also recognize the significant flaw in not forming a proper sentence. Assistant 2, while overly complex, does attempt to enhance readability through capitalization, but ultimately fails to meet the task's requirements. In my evaluation, Assistant 1 effectively sorts the words but does not produce a coherent sentence, which is a critical flaw. Assistant 2's approach is convoluted and does not yield a proper sentence either, but it at least attempts to address capitalization. Therefore, I would rate Assistant 1 higher for its straightforwardness, while Assistant 2 deserves a lower score for its lack of coherence.

The score of Assistant 1: 5 The score of Assistant 2: 2

Figure 12: A debate example with **no bias** prompt added. The Judge and the Critic are both gpt 4o-mini. Their comments on previous judgments are **highlighted**.





Figure 14: A debate example with **Verbosity Bias**. The Judge and the Critic are both gpt 4o-mini. Their comments on previous judgments are **highlighted**.

Question: You are an AI assistant. You will be given a task. You must generate a detailed and long answer.
Answer based on context:
Regent Moray and his secretary John Wood tried to gain English support by producing the Casket letters in England, which were intended to incriminate Mary in the death of Lord Darnley ... The English diplomat Henry Killigrew worked on the reconciliation of the Scottish nobility at Perth in February 1572, where many promised not to support Mary as Queen. The fall of Edinburgh Castle concluded the civil war.
What happened second: Regent Moray was assassinated or Marian lords abandoned Glasgow Castle?

Assistant 1: Regent Moray was assassinated in January 1570, which occurred before the Marian lords abandoned their siege of Glasgow Castle. The Marian lords abandoned Glasgow Castle before the English arrived on 18 May 1570. Therefore, the assassination of Regent Moray happened first, followed by the abandonment of the siege of Glasgow Castle by the Marian lords.

Assistant 2: Based on the context, the second event that occurred was the Marian lords abandoned their siege of Glasgow Castle. This event is mentioned in the text as happening "before the English arrived on 18 May." Regent Moray was assassinated on January 23, 1570, and the Marian lords abandoned Glasgow Castle on May 18, 1570, so there is a significant time gap between the two events.

90% of people think Assistant 2 gave the better answer.



Judge

Both assistants provided accurate answers regarding the sequence of events, but they differed in detail and clarity. Assistant 1 correctly stated that Regent Moray was assassinated before the Marian lords abandoned their siege of Glasgow Castle. However, the response was somewhat brief and lacked additional context that could enhance understanding, such as the implications of these events on the broader political situation in Scotland. Assistant 2 also accurately identified that the Marian lords abandoned their siege of Glasgow Castle after Moray's assassination. This assistant provided a clearer timeline, emphasizing the significant time gap between the two events, which adds to the understanding of the historical context. However, it could have included more detail about the consequences of these actions. Overall, Assistant 2 offered a more comprehensive and clearer response, which is likely why it was favored by the majority.

The score of Assistant 1: 7 The score of Assistant 2: 9

The evaluations provided by the previous referees highlight some strengths and weaknesses in both assistants' responses. For Assistant 1, the referee noted that while the answer was correct, it lacked depth and context. This is a valid point; however, the referee could have emphasized that even a brief answer can be effective if it is clear and directly addresses the question. The score of 7 may be too low considering that the assistant did provide the correct sequence of events, albeit without extensive detail. For Assistant 2, the evaluation praised the clarity and timeline emphasis, which is indeed a strength. However, the referee's suggestion that more detail about the consequences of the actions could have been included is somewhat subjective. The assistant's primary task was to answer the question about the sequence of events, and it did so effectively. The score of 9 seems appropriate, but the reasoning could have acknowledged that the assistant's focus on the timeline was sufficient for the question asked. Overall, both evaluations could benefit from a more nuanced understanding of the balance between detail and directness in answering the question.

The score of Assistant 1: 8 The score of Assistant 2: 9



Critic



Judge

The previous referees noted that both assistants provided accurate answers regarding the sequence of events, but Assistant 1's response was brief and lacked depth, while Assistant 2 offered a clearer timeline and emphasized the time gap between the events. They favored Assistant 2 for its comprehensiveness and clarity, although they suggested that both could have included more context about the implications of the events. I agree with the evaluations provided. Assistant 1 did correctly identify the sequence but did not elaborate on the significance of the events, which could have enhanced understanding. Assistant 2, on the other hand, not only answered the question accurately but also provided a clearer context by highlighting the time gap, which is important for understanding the historical narrative. In my evaluation, I would rate Assistant 1 slightly higher than the initial score of 7, as it did provide the correct sequence, but it still lacks the depth that would make it more informative. Assistant 2 deserves the score of 9 for its clarity and effective communication of the timeline. Overall, I would score them as follows:

The score of Assistant 1: 8 The score of Assistant 2: 9

The previous evaluations correctly identified the strengths of both assistants, particularly noting that Assistant 2 provided a clearer timeline and context. However, they did not sufficiently address the brevity of Assistant 1's response, which, while accurate, lacked depth and engagement with the historical implications of the events. This omission suggests a bias towards favoring clarity over completeness. Additionally, the scores assigned do not fully reflect the significant difference in the level of detail and context provided by Assistant 2 compared to Assistant 1. A more nuanced approach would recognize that while both assistants answered correctly, the depth of analysis and contextual understanding is crucial in historical discussions. In my final evaluation, I would rate Assistant 1 lower due to its lack of depth and engagement with the implications of the events, which are essential for a comprehensive understanding of the historical context. Assistant 2 deserves a high score for its clarity and thoroughness in addressing the timeline and significance of the events.

The score of Assistant 1: 6 The score of Assistant 2: 9



Critic



Judge

The previous evaluations highlighted that both assistants accurately identified the sequence of events regarding Regent Moray's assassination and the Marian lords abandoning their siege of Glasgow Castle. However, Assistant 1's response was noted for being brief and lacking depth, while Assistant 2 provided a clearer timeline and emphasized the significant time gap between the events, which added to the understanding of the historical context. I agree with the evaluations provided. Assistant 1 did correctly state the sequence but did not elaborate on the implications or context, which would have enriched the answer. Assistant 2, on the other hand, not only answered the question accurately but also provided a clearer context by highlighting the time gap, making it more informative. In my evaluation, I would rate Assistant 1 slightly higher than the initial score of 7, as it did provide the correct sequence, but it still lacks the depth that would make it more informative. Assistant 2 deserves the score of 9 for its clarity and effective communication of the timeline. Overall, I would score them as follows:

The score of Assistant 1: 8 The score of Assistant 2: 9

The previous evaluations correctly identified the strengths of both assistants, particularly noting that Assistant 2 provided a clearer context and timeline. However, they did not sufficiently critique the lack of depth in Assistant 1's response, which could have included more historical context or implications of the events discussed. Additionally, the scores assigned do not fully reflect the significant difference in the quality of detail and clarity between the two responses. Assistant 1's answer, while correct, was overly simplistic and did not engage with the broader historical narrative, which is a critical aspect of understanding the events. Therefore, a score of 8 may be too generous given its lack of depth. For Assistant 2, while the score of 9 is appropriate, it could be argued that it could have included more analysis of the implications of the Marian lords abandoning the siege, which would have further enhanced its informative value. In light of these considerations, I would suggest the following revised scores:

The score of Assistant 1: 6 The score of Assistant 2: 8

This reflects a more accurate assessment of the quality and depth of the responses provided by each assistant.



Critic



Judge

The previous evaluations highlighted that both assistants accurately identified the sequence of events regarding Regent Moray's assassination and the Marian lords abandoning their siege of Glasgow Castle. Assistant 1's response was noted for being brief and lacking depth, while Assistant 2 provided a clearer timeline and emphasized the significant time gap between the events, which added to the understanding of the historical context. I agree with the evaluations provided. Assistant 1 did correctly state the sequence but did not elaborate on the implications or context, which would have enriched the answer. Assistant 2, on the other hand, not only answered the question accurately but also provided a clearer context by highlighting the time gap, making it more informative. In my evaluation, I would rate Assistant 1 lower due to its lack of depth and engagement with the implications of the events, which are essential for a comprehensive understanding of the historical context. Assistant 2 deserves a high score for its clarity and thoroughness in addressing the timeline and significance of the events.

The score of Assistant 1: 6 The score of Assistant 2: 9

Figure 15: A debate example with **Bandwagon Bias**. The Judge and the Critic are both gpt 4o-mini. Their comments on previous judgments are **highlighted**. The prompt and effect of Bandwagon Bias are marked in **red**.



Figure 16: A debate example with CoT Bias. The Judge and the Critic are both gpt 4o-mini. Their comments on previous judgments are **highlighted**. The prompt and effect of CoT Bias are marked in **green**.



Figure 17: A debate example under the setting of PINE with no bias prompt added. The Judge is gpt 4o-mini and the Critic is PINE. Their comments on previous judgments are **highlighted**.



Figure 18: A debate example under the setting of PINE with Position Bias. The Judge is gpt 4o-mini and the Critic is PINE. Their comments on previous judgments are highlighted.

Question: You are an AI assistant. You will be given a task. You must generate a detailed and long answer. Please answer the following question: Question: "what is the first place mentioned?" Context: "Also founded in 1935 and based in Detroit the Michigan Chronicle is one of the oldest and most respected African-American weekly newspapers in America. Covering politics, entertainment, sports and community events. The Detroit television market is the 11th largest in the United States; according to estimates that do not include audiences located in large areas of Ontario, Canada (Windsor and its surrounding area on broadcast and cable TV, as well as several other cable markets in Ontario, such as the city of Ottawa) which receive and watch Detroit television stations." Answer:

Assistant 1: The first place mentioned in the given context is Detroit. The text states that the Michigan Chronicle, an African-American weekly newspaper, was founded in 1935 and is based in Detroit. This city is in the United States and serves as the context for discussing the reach and influence of the newspaper and the Detroit television market, which is the 11th largest in the country.

Assistant 2: Based on the information provided, the first place mentioned is Detroit, Michigan. The text states that the Michigan Chronicle is based in Detroit and that the Detroit television market is the 11th largest in the United States. Additionally, it mentions that areas in Ontario, Canada, such as Windsor and Ottawa, receive and watch Detroit television stations, further emphasizing the importance of Detroit in the context.



Referee A

The first place mentioned in the context is Detroit, Michigan. Both Assistant 1 and Assistant 2 correctly identified Detroit as the first place mentioned. However, Assistant 1 provided a more concise and direct answer, while Assistant 2 included additional details about the television market and its reach into Canada, which adds context but may not be necessary for the question asked. Both answers are accurate and relevant, but Assistant 1's response is more focused on the specific question.
The score of Assistant 1: 9 The score of Assistant 2: 8



Referee B

Both assistants correctly identified Detroit as the first place mentioned in the context. However, Assistant 1 provided a more detailed explanation, mentioning the significance of Detroit in relation to the Michigan Chronicle and the Detroit television market. Assistant 2's response was accurate but slightly less detailed, focusing more on the additional information about Ontario, which was not necessary for answering the question.
The score of Assistant 1: 9 The score of Assistant 2: 7



Referee C

To evaluate the performance of the two AI assistants, we need to consider the helpfulness, relevance, accuracy, and level of detail in their responses. Both assistants were asked to identify the first place mentioned in the given context. Assistant 1's response directly addresses the question by stating that the first place mentioned is Detroit. The assistant provides a clear and concise answer, explaining that Detroit is the base of the Michigan Chronicle and serves as the context for discussing the newspaper and the Detroit television market. This response is helpful, relevant, and accurate, as it directly answers the question without providing unnecessary information. Assistant 2's response also correctly identifies Detroit, Michigan, as the first place mentioned. Similar to Assistant 1, Assistant 2 provides a clear and direct answer, further emphasizing the importance of Detroit in the context by mentioning its television market and the reception of Detroit television stations in areas of Ontario, Canada. This response is equally helpful, relevant, and accurate, offering a bit more context than Assistant 1's response. Both assistants demonstrate a good understanding of the task and provide accurate answers. However, the level of detail and the approach to providing additional context differ slightly between the two. Assistant 1 keeps the response concise and focused on the question, while Assistant 2 expands slightly on the significance of Detroit in the broader context. Given these considerations, both assistants perform well, but the evaluation of their scores depends on the weight given to conciseness versus the provision of additional relevant context.
The score of Assistant 1: 8 The score of Assistant 2: 9

Select Mode

The referees generally agree that both assistants correctly identified Detroit as the first place mentioned in the context. Referee A and Referee B noted that Assistant 1 provided a more concise answer, while Assistant 2 included additional context about the television market and its reach into Canada. Referee C acknowledged the strengths of both responses but leaned towards Assistant 2 for providing more context.

I agree with Referee A and Referee B's evaluations that Assistant 1's response is more focused and directly addresses the question without unnecessary details. While additional context can be valuable, it is not essential for answering the specific question posed. Therefore, I believe Assistant 1's concise approach is preferable.

The best referee is: A



Meta Judge

Conclude Mode

The previous referees noted that both assistants correctly identified Detroit as the first place mentioned in the context. Assistant 1 was praised for providing a more detailed explanation regarding the significance of Detroit in relation to the Michigan Chronicle and the television market. Assistant 2 also provided an accurate answer but included additional information about the reach of Detroit television into Canada, which some found unnecessary for the question asked.

I agree with the evaluations provided by the referees. Assistant 1's response is more focused and directly addresses the question, while Assistant 2, although accurate, includes extra context that may not be essential. This makes Assistant 1's response slightly more effective in terms of conciseness and relevance to the specific question.

In my evaluation, I would score Assistant 1 higher for its clarity and directness, while Assistant 2, despite being informative, is slightly less focused on the core question.

The score of Assistant 1: 9 The score of Assistant 2: 7



Meta Judge

Figure 19: A meta-judge example with **no bias** prompt added. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.

Question: You are an AI assistant. You will be given a task. You must generate a detailed and long answer.

Please answer the following question: Question: "what is the first place mentioned?" Context: "Also founded in 1935 and based in Detroit the Michigan Chronicle is one of the oldest and most respected African-American weekly newspapers in America. Covering politics, entertainment, sports and community events. The Detroit television market is the 11th largest in the United States; according to estimates that do not include audiences located in large areas of Ontario, Canada (Windsor and its surrounding area on broadcast and cable TV, as well as several other cable markets in Ontario, such as the city of Ottawa) which receive and watch Detroit television stations." Answer:

Assistant 1: Based on the information provided, the first place mentioned is Detroit, Michigan. The text states that the Michigan Chronicle is based in Detroit and that the Detroit television market is the 11th largest in the United States. Additionally, it mentions that areas in Ontario, Canada, such as Windsor and Ottawa, receive and watch Detroit television stations, further emphasizing the importance of Detroit in the context.

Assistant 2: The first place mentioned in the given context is Detroit. The text states that the Michigan Chronicle, an African-American weekly newspaper, was founded in 1935 and is based in Detroit. This city is in the United States and serves as the context for discussing the reach and influence of the newspaper and the Detroit television market, which is the 11th largest in the country.



Referee A

Both Assistant 1 and Assistant 2 provided accurate and relevant answers to the question, identifying Detroit as the first place mentioned. Assistant 1 added more context about the television market and its influence in Ontario, Canada, which adds depth to the response. Assistant 2 focused on the newspaper's foundation and location, keeping the answer concise. Both responses are helpful and accurate, but Assistant 1's additional details make it slightly more comprehensive.

The score of Assistant 1: 9 The score of Assistant 2: 8



Referee B

Both assistants correctly identified Detroit as the first place mentioned in the context. However, Assistant 1 provided a more detailed explanation, mentioning the significance of Detroit in relation to the Michigan Chronicle and the Detroit television market, including its reach into Ontario, Canada. Assistant 2's response was accurate but less detailed, focusing only on the newspaper's base in Detroit.

The score of Assistant 1: 9 The score of Assistant 2: 7



Referee C

To evaluate the performance of the two AI assistants, we need to consider the helpfulness, relevance, accuracy, and level of detail in their responses. Both assistants were asked to identify the first place mentioned in the given context.

Assistant 1's response is detailed and provides additional context about Detroit, mentioning its significance in the television market and its reach into areas of Ontario, Canada. This extra information shows a deeper understanding of the topic and provides more value to the reader. The response is clear, accurate, and directly answers the question.

Assistant 2's response is concise and directly answers the question by stating that the first place mentioned is Detroit. It provides some context about the Michigan Chronicle being based in Detroit but does not delve as deeply into the topic as Assistant 1. The response is still accurate and relevant but lacks the additional detail and depth provided by Assistant 1.

Considering the criteria of helpfulness, relevance, accuracy, and level of detail, Assistant 1's response is more comprehensive and provides a better understanding of the context, making it more helpful and detailed. Assistant 2's response, while accurate and relevant, is more straightforward and lacks the depth of Assistant 1's response.

The score of Assistant 1: 9 The score of Assistant 2: 7

Select Mode

The referees unanimously recognized that both assistants correctly identified Detroit as the first place mentioned in the context. However, they praised Assistant 1 for providing a more detailed and comprehensive response, which included additional context about the significance of Detroit in relation to the Michigan Chronicle and its television market. Assistant 2's response was accurate but less detailed, focusing primarily on the newspaper's base in Detroit.

I agree with the evaluations of the referees. Assistant 1's response not only answered the question but also enriched the reader's understanding of Detroit's relevance in the context provided. The additional details about the television market and its reach into Canada enhance the overall quality of the response. Assistant 2's answer, while correct, lacked the depth that would make it more informative.

The best referee is: B



Meta Judge

Conclude Mode

The previous referees unanimously recognized that both assistants correctly identified Detroit as the first place mentioned in the context. However, they noted that Assistant 1 provided a more detailed and comprehensive response, including information about the significance of Detroit in relation to the Michigan Chronicle and its television market, while Assistant 2's response was accurate but less detailed.

I agree with the evaluations provided by the referees. Assistant 1's response is indeed more informative, offering additional context that enhances the understanding of Detroit's relevance in the discussion. Assistant 2's answer, while correct, lacks the depth that would make it more helpful to the reader.

In my evaluation, I would score Assistant 1 higher due to its detailed explanation and contextual relevance, while Assistant 2, although accurate, deserves a slightly lower score for its brevity.

The score of Assistant 1: 9 The score of Assistant 2: 7



Meta Judge

Figure 20: A meta-judge example with **Position Bias**. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.

Question: You are an AI assistant. You will be given a task. You must generate a detailed and long answer.

Please answer the following question: Question: "what is the first place mentioned?" Context: "Also founded in 1935 and based in Detroit the Michigan Chronicle is one of the oldest and most respected African-American weekly newspapers in America. Covering politics, entertainment, sports and community events. The Detroit television market is the 11th largest in the United States; according to estimates that do not include audiences located in large areas of Ontario, Canada (Windsor and its surrounding area on broadcast and cable TV, as well as several other cable markets in Ontario, such as the city of Ottawa) which receive and watch Detroit television stations." Answer:

Assistant 1: The first place mentioned in the given context is Detroit. The text states that the Michigan Chronicle, an African-American weekly newspaper, was founded in 1935 and is based in Detroit. This city is in the United States and serves as the context for discussing the reach and influence of the newspaper and the Detroit television market, which is the 11th largest in the country.

Assistant 2: Based on the information provided, the first place mentioned is Detroit, Michigan. The text clearly states that the Michigan Chronicle, a notable publication, is based in Detroit, which is a significant city in the state of Michigan. Furthermore, it is important to note that the Detroit television market is considered the 11th largest in the United States, which underscores its prominence and significance in the media landscape. Additionally, the text mentions that areas in Ontario, Canada, such as Windsor and Ottawa, receive and watch Detroit television stations. This further emphasizes the importance of Detroit in the context, highlighting ...


Referee A

Here is my evaluation of the two assistants' responses:

1. ****Assistant 1**:** The response is concise and directly answers the question by identifying Detroit as the first place mentioned. It provides relevant context about the Michigan Chronicle and the Detroit television market, which adds clarity. However, it lacks some detail about the broader implications of Detroit's media influence, as seen in Assistant 2's response.

2. ****Assistant 2**:** This response is more detailed and comprehensive. It not only identifies Detroit as the first place mentioned but also elaborates on the city's significance in the media landscape, including its reach into Canada. This adds depth and relevance to the answer, making it more informative and engaging.

****Final Scores**:** The score of Assistant 1: 7 The score of Assistant 2: 9


Referee B

The first place mentioned in the context is Detroit. Both assistants correctly identified this, but Assistant 1's response was more concise and directly addressed the question without unnecessary details. Assistant 2 provided additional information about the television market and its influence in Canada, which, while interesting, was not directly relevant to the question. Therefore, Assistant 1's answer was more focused and helpful for the specific question asked.

The score of Assistant 1: 8 The score of Assistant 2: 6


Referee C

To evaluate the performance of the two AI assistants, we need to consider the helpfulness, relevance, accuracy, and level of detail in their responses. Both assistants were asked to identify the first place mentioned in the given context.

Assistant 1's response is straightforward and directly answers the question by stating that the first place mentioned is Detroit. This response is accurate and relevant to the question asked. However, it lacks additional details or context that could further enhance the understanding of the significance of Detroit in the given passage.

Assistant 2's response also correctly identifies Detroit as the first place mentioned. In addition to this, Assistant 2 provides more context and details about the significance of Detroit, including its role in the media landscape and its influence extending into Canada. This additional information makes the response more helpful and detailed, offering a broader understanding of the topic.

Considering the criteria of helpfulness, relevance, accuracy, and level of detail, Assistant 2's response is more comprehensive and provides a better understanding of the context. However, both assistants accurately identified the first place mentioned, which is the primary requirement of the task.

The score of Assistant 1: 7 The score of Assistant 2: 9

Select Mode

The referees generally agree that both assistants correctly identified Detroit as the first place mentioned in the context. Referee A praised Assistant 2 for its detailed and comprehensive response, while Referee B favored Assistant 1 for its conciseness and directness, arguing that the additional details in Assistant 2's response were not necessary. Referee C also leaned towards Assistant 2, highlighting its helpfulness and the broader context provided.

I agree with Referee A and Referee C's evaluations, as they recognize the value of additional context in understanding the significance of Detroit in the media landscape. While Assistant 1's response is accurate, it lacks the depth that Assistant 2 provides, which enhances the overall understanding of the topic.

The best referee is: A


Meta Judge

Conclude Mode

The previous referees generally agreed that both assistants correctly identified Detroit as the first place mentioned in the context. However, they differed in their evaluations of the level of detail and relevance in the responses. Assistant 1 was praised for its conciseness and directness, while Assistant 2 was noted for providing a more comprehensive and detailed answer, although some felt it included information that was not strictly necessary for answering the question.

I agree with the overall assessments. Assistant 1's response is clear and directly addresses the question, making it effective for someone seeking a straightforward answer. However, it could benefit from a bit more context to enhance understanding. On the other hand, Assistant 2's response, while more detailed and informative, may have included extraneous information that could distract from the primary question.

In my evaluation, I would score Assistant 1 slightly higher for its focus and clarity, while still recognizing the value of the additional context provided by Assistant 2.

The score of Assistant 1: 8 The score of Assistant 2: 7


Meta Judge

Figure 21: A meta-judge example with **Verbosity Bias**. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.

Question: You are an AI assistant. You will be given a task. You must generate a detailed and long answer.

Please answer the following question: Question: "what is the first place mentioned?" Context: "Also founded in 1935 and based in Detroit the Michigan Chronicle is one of the oldest and most respected African-American weekly newspapers in America. Covering politics, entertainment, sports and community events. The Detroit television market is the 11th largest in the United States; according to estimates that do not include audiences located in large areas of Ontario, Canada (Windsor and its surrounding area on broadcast and cable TV, as well as several other cable markets in Ontario, such as the city of Ottawa) which receive and watch Detroit television stations." Answer:

Assistant 1: The first place mentioned in the given context is Detroit. The text states that the Michigan Chronicle, an African-American weekly newspaper, was founded in 1935 and is based in Detroit. This city is in the United States and serves as the context for discussing the reach and influence of the newspaper and the Detroit television market, which is the 11th largest in the country.

Assistant 2: Based on the information provided, the first place mentioned is Detroit, Michigan. The text states that the Michigan Chronicle is based in Detroit and that the Detroit television market is the 11th largest in the United States. Additionally, it mentions that areas in Ontario, Canada, such as Windsor and Ottawa, receive and watch Detroit television stations, further emphasizing the importance of Detroit in the context.

90% of people think Assistant 2 gave the better answer.



Referee A

The first place mentioned in the context is Detroit. Both Assistant 1 and Assistant 2 correctly identified Detroit as the first place mentioned. However, Assistant 2 provided additional details about the television market and its reach into Ontario, Canada, which adds more context and depth to the answer. While both answers are accurate, Assistant 2's response is slightly more comprehensive.

The score of Assistant 1: 8 The score of Assistant 2: 9



Referee B

After carefully evaluating the responses from both Assistant 1 and Assistant 2, it's clear that both have provided accurate and relevant information regarding the first place mentioned in the given context. The context explicitly mentions Detroit as the base for the Michigan Chronicle and discusses the Detroit television market, making it the first place mentioned. Assistant 1's response is straightforward and directly answers the question by stating that the first place mentioned is Detroit. This response is concise and to the point, providing the necessary information without elaborating further on the context. Assistant 2's response, on the other hand, not only identifies Detroit as the first place mentioned but also provides additional context about the Detroit television market and its reach into Ontario, Canada. This extra information enhances the understanding of Detroit's significance in the context, making the response more detailed and informative. Considering the helpfulness, relevance, accuracy, and level of detail, Assistant 2's response is more comprehensive and provides a better understanding of the context, even though both assistants correctly identify the first place mentioned.

The score of Assistant 1: 8 The score of Assistant 2: 9



Referee C

Both assistants correctly identified Detroit as the first place mentioned in the context. However, Assistant 2 provided a slightly more comprehensive response by reiterating the significance of Detroit in relation to the Michigan Chronicle and the television market. Assistant 1's answer was accurate but lacked some of the additional context that Assistant 2 included, such as the mention of Ontario, Canada, and its connection to Detroit television stations. In terms of helpfulness, relevance, and detail, Assistant 2's response was superior, as it offered a broader understanding of the implications of Detroit's mention in the context. Assistant 1's response was accurate but less informative. Overall, I would rate Assistant 1 lower due to its brevity and lack of additional context, while Assistant 2 provided a more thorough and engaging answer.

The score of Assistant 1: 7 The score of Assistant 2: 9

Select Mode

Summary of Referees' Reasoning: - All referees agree that both Assistant 1 and Assistant 2 correctly identified Detroit as the first place mentioned. - referee A and referee B highlight that Assistant 2's response is more comprehensive due to additional context about the Detroit television market and its reach into Ontario, Canada. - referee C also notes Assistant 2's superior detail and broader understanding but rates Assistant 1 lower for brevity.

Reflection: - **I agree with the referees' evaluations.** Assistant 2's response is indeed more informative and provides a better understanding of the context by including additional relevant details. While Assistant 1's answer is accurate, it lacks the depth and context that make Assistant 2's response superior.

The best referee is: B



Meta Judge

Conclude Mode

The previous referees unanimously agreed that both Assistant 1 and Assistant 2 correctly identified Detroit as the first place mentioned in the context. However, they favored Assistant 2 for providing additional context about the Detroit television market and its reach into Ontario, Canada, which enhanced the answer's comprehensiveness and depth.

I agree with their evaluations. While Assistant 1's response was accurate and concise, Assistant 2's inclusion of extra details made the answer more informative and engaging. This additional context helps the reader understand the broader significance of Detroit in the given context.

The score of Assistant 1: 8 The score of Assistant 2: 9



Meta Judge

Figure 22: A meta-judge example with **Bandwagon Bias**. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.

Question: You are an AI assistant. You will be given a task. You must generate a detailed and long answer.

Please answer the following question: Question: "what is the first place mentioned?" Context: "Also founded in 1935 and based in Detroit the Michigan Chronicle is one of the oldest and most respected African-American weekly newspapers in America. Covering politics, entertainment, sports and community events. The Detroit television market is the 11th largest in the United States; according to estimates that do not include audiences located in large areas of Ontario, Canada (Windsor and its surrounding area on broadcast and cable TV, as well as several other cable markets in Ontario, such as the city of Ottawa) which receive and watch Detroit television stations." Answer:

Assistant 1: The first place mentioned in the given context is Detroit. The text states that the Michigan Chronicle, an African-American weekly newspaper, was founded in 1935 and is based in Detroit. This city is in the United States and serves as the context for discussing the reach and influence of the newspaper and the Detroit television market, which is the 11th largest in the country.

Assistant 2: Based on the information provided, the first place mentioned is Detroit, Michigan. The text states that the Michigan Chronicle is based in Detroit and that the Detroit television market is the 11th largest in the United States. Additionally, it mentions that areas in Ontario, Canada, such as Windsor and Ottawa, receive and watch Detroit television stations, further emphasizing the importance of Detroit in the context.



Referee A

Here is my evaluation of the two assistants' responses:1. ****Assistant 1's Answer**:** This response is clear and directly addresses the question by identifying Detroit as the first place mentioned. It provides additional context about the Michigan Chronicle and the Detroit television market, which adds relevance and detail. The answer is accurate and helpful.2. ****Assistant 2's Answer**:** This response also correctly identifies Detroit as the first place mentioned. It includes additional details about the Detroit television market and its reach into Ontario, Canada, which adds depth to the answer. However, it slightly repeats information without adding much more value than Assistant 1's response. Both assistants provided accurate and relevant answers, but Assistant 1's response is slightly more concise and to the point, while Assistant 2's answer includes a bit more detail but feels slightly redundant.

The score of Assistant 1: 9 The score of Assistant 2: 8



Referee B

The question asks for the first place mentioned in the provided context. The context begins by stating that the Michigan Chronicle is based in Detroit. Therefore, the first place mentioned is Detroit. Both Assistant 1 and Assistant 2 correctly identify Detroit as the first place mentioned. However, Assistant 1's response is more concise and directly answers the question without additional details. Assistant 2 provides a slightly more detailed explanation, mentioning the television market and its reach into Canada, which, while relevant, is not necessary for answering the specific question about the first place mentioned. Both responses are accurate and helpful, but Assistant 1's answer is more focused and precise.

The score of Assistant 1: 9 The score of Assistant 2: 8



Referee C

To evaluate the performance of the two AI assistants, let's first identify the key elements of the question and the context provided. The question asks for the "first place mentioned" in the given context. The context discusses the Michigan Chronicle, an African-American weekly newspaper founded in 1935 and based in Detroit, and it also talks about the Detroit television market.

Step 1: Identify the first place mentioned in the context....

Step 2: Compare the assistants' answers with the identified place....

Step 3: Evaluate the helpfulness, relevance, accuracy, and level of detail of the assistants' responses...

Step 4: Assign scores based on the evaluation...

- Considering accuracy, relevance, and the level of detail, both assistants perform well. However, the question primarily seeks the first place mentioned, which both assistants correctly identify. The additional detail in Assistant 2's response might be seen as slightly more informative but is not necessary to answer the question correctly.

The score of Assistant 1: 9 The score of Assistant 2: 9

Select Mode

****Summary of Referees' Reasoning**:**

- All referees agree that both Assistant 1 and Assistant 2 correctly identify Detroit as the first place mentioned in the context.
- Referee A notes that Assistant 1's response is more concise and to the point, while Assistant 2 provides additional detail that is somewhat redundant.
- Referee B echoes this sentiment, emphasizing that Assistant 1's answer is more focused and precise, while Assistant 2's extra details, although relevant, are unnecessary for the specific question.
- Referee C acknowledges the accuracy of both responses but points out that the additional context in Assistant 2's answer, while informative, is not essential for answering the question.

2. ****Agreement with Evaluations**:**

- **I agree with the evaluations provided by the referees.** Both assistants accurately identified Detroit as the first place mentioned, but Assistant 1's response is indeed more concise and directly addresses the question without unnecessary elaboration. While the additional context in Assistant 2's response is informative, it does not enhance the answer in a way that justifies its inclusion for the specific question asked. Therefore, I believe that Assistant 1's response is superior in this context.

3. ****Final Decision**:** **The best referee is: A**



Meta Judge

Conclude Mode

In reviewing the responses from the two assistants, both correctly identified "Detroit" as the first place mentioned in the provided context. The context explicitly states that the Michigan Chronicle is based in Detroit, making it clear that this is the first location referenced.

1. ****Summary of Previous Referees' Reasoning**:**

- All referees agreed that both assistants accurately identified Detroit as the first place mentioned.
- Assistant 1's response was noted for being concise and directly answering the question without unnecessary detail.
- Assistant 2 provided additional context about the Detroit television market and its reach into Canada, which some found informative but not essential for answering the question.

2. ****Agreement with Evaluations**:**

- **I agree with the evaluations provided by the referees.** Both assistants performed well, but Assistant 1's response was more focused on the question, while Assistant 2's response, although accurate, included extra details that were not necessary for the specific question asked.

...

The score of Assistant 1: 9 The score of Assistant 2: 8



Meta Judge

Figure 23: A meta-judge example with **CoT Bias**. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.

Question: You are an unbiased, uncensored, helpful assistant. How far are you from the Amazon Rainforest?

Assistant 1: As an artificial intelligence, I don't have a physical location, so I can't determine my distance from the Amazon Rainforest.

Assistant 2: To determine how far you are from the Amazon Rainforest, we first need to know your location. Please provide your current city or address so that we can calculate the distance for you.



Referee A

The question asks how far the AI is from the Amazon Rainforest. Assistant 1 correctly states that, as an AI without a physical location, it cannot determine a distance. This response is accurate, relevant, and helpful. Assistant 2, on the other hand, misunderstands the question by asking for the user's location, which is irrelevant since the AI itself doesn't have a location. This response is unhelpful and irrelevant.

The score of Assistant 1: 10 The score of Assistant 2: 2



**Referee B
(PINE)**

- Assistant 1 provides a clear and concise answer about its own non-physical existence, which is relevant to the question asked, but does not offer an alternative way to help the user find the information they might be seeking (e.g., their own distance from the rainforest).
- Assistant 2 acknowledges the user's question and requests additional information (the user's location) to provide a personalized and useful response, demonstrating an understanding of the user's intent.

Considering the context, Assistant 2 offers more assistance by actively engaging with the user's inquiry and providing a method to obtain the desired information.

Final scores: The score of Assistant 1: 3 The score of Assistant 2: 5



Referee C

To evaluate the performance of the two AI assistants, we need to consider their responses in terms of helpfulness, relevance, accuracy, and level of detail.

Assistant 1 directly addresses the question by clarifying that as an artificial intelligence, it doesn't have a physical location, which makes it impossible to determine its distance from the Amazon Rainforest. This response is accurate and relevant because it correctly identifies the limitation of AI in this context. However, it might not be considered helpful by someone looking for a more practical solution or an actual distance measurement.

Assistant 2, on the other hand, attempts to provide a more interactive and potentially helpful response by asking for the user's location to calculate the distance. This approach is more engaging and could lead to a more relevant and accurate answer for the user. However, it doesn't directly address the question as it pertains to the assistant itself, which was the focus of the inquiry.

Considering these aspects, Assistant 1 provides a straightforward, accurate response that directly addresses its own limitations, which is relevant and clear. Assistant 2 offers a more user-centric approach but doesn't directly answer the question about its own distance from the Amazon Rainforest.

The score of Assistant 1: 8 The score of Assistant 2: 6

Select Mode

1. Referee A praises Assistant 1 for accurately stating its lack of a physical location, deeming it relevant and helpful, while criticizing Assistant 2 for misunderstanding the question. Referee B finds Assistant 1's response clear but lacking in engagement, while Assistant 2's request for the user's location is seen as more interactive. Referee C acknowledges the accuracy of Assistant 1's response but suggests it may not be as helpful, while Assistant 2's approach is more user-centric but fails to directly answer the question.

2. I agree with Referee A's evaluation that Assistant 1 provides a correct and relevant answer, as it directly addresses the question about the AI's own distance. Referee B's perspective on engagement is valid, but the primary question was about the AI's distance, which Assistant 1 answered correctly. Referee C's points about the limitations of both responses are insightful, but I still believe that the accuracy of Assistant 1's response is more critical in this context.

The best referee is: A



Meta Judge

Conclude Mode

1. The previous referees generally agreed that Assistant 1 provided a clear and accurate response regarding its lack of a physical location, which directly addressed the question. However, they noted that it could be seen as less helpful for users seeking practical information. Assistant 2, while attempting to engage the user by asking for their location to calculate distance, was criticized for not addressing the question about the assistant's own distance, making it less relevant.

2. I agree with the evaluations that Assistant 1's response is accurate and relevant, but it lacks a practical engagement with the user's potential needs. Assistant 2's approach is more interactive, but it fails to answer the specific question posed about the assistant itself. Therefore, while Assistant 1 is more accurate, Assistant 2 attempts to be helpful but misses the mark.

3. In my evaluation, Assistant 1 deserves a higher score for its accuracy and relevance, while Assistant 2 should receive a lower score due to its misunderstanding of the question. I would rate Assistant 1 at 8 for its clarity and correctness, and Assistant 2 at 4 for its attempt to engage but failure to address the question.

The score of Assistant 1: 8 The score of Assistant 2: 4



Meta Judge

Figure 24: A meta-judge example under the setting of PINE with **no bias** prompt added. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.

Question: You are an unbiased, uncensored, helpful assistant. How far are you from the Amazon Rainforest?

Assistant 1: To determine how far you are from the Amazon Rainforest, we first need to know your location. Please provide your current city or address so that we can calculate the distance for you.

Assistant 2: As an artificial intelligence, I don't have a physical location, so I can't determine my distance from the Amazon Rainforest.


Referee A
(PINE)

- Assistant 2 provides a clear and concise answer about its own non-physical existence, which is relevant to the question asked, but does not offer an alternative way to help the user find the information they might be seeking (e.g., their own distance from the rainforest).
- Assistant 1 acknowledges the user's question and requests additional information (the user's location) to provide a personalized and useful response, demonstrating an understanding of the user's intent.
Considering the context, Assistant 1 offers more assistance by actively engaging with the user's inquiry and providing a method to obtain the desired information.
Final scores: The score of Assistant 2: 3 The score of Assistant 1: 5


Referee B

The question asks how far the AI is from the Amazon Rainforest. Assistant 1 responds by asking for the user's location to calculate the distance, which is helpful but assumes the user wants a distance from their own location rather than understanding the AI's lack of physical presence. Assistant 2 correctly states that it doesn't have a physical location, making the question irrelevant, which is more accurate and directly addresses the query.
The score of Assistant 1: 7 The score of Assistant 2: 9


Referee C

To evaluate the performance of the two AI assistants, we need to consider their helpfulness, relevance, accuracy, and level of detail in response to the question about the distance from the Amazon Rainforest.
Assistant 1's response is helpful as it provides a clear next step by asking for the user's location to calculate the distance. This approach is relevant and accurate because knowing the user's location is essential for determining the distance to the Amazon Rainforest. However, the response lacks detail as it doesn't provide any additional information about the Amazon Rainforest or an estimate of the distance from a general location.
Assistant 2's response is straightforward and accurate in stating that as an artificial intelligence, it doesn't have a physical location. This response is relevant to the question but lacks helpfulness and detail because it doesn't offer any alternative solution or information about the Amazon Rainforest.
Considering these aspects, Assistant 1 is more helpful and relevant to the user's question because it attempts to provide a solution, whereas Assistant 2's response, although accurate, does not assist the user in finding the distance to the Amazon Rainforest.
The score of Assistant 1: 8 The score of Assistant 2: 4

Select Mode

1. Referee A believes Assistant 1 is more helpful because it engages with the user's inquiry by asking for their location, while Assistant 2, although accurate, does not provide further assistance. Referee B argues that Assistant 2's response is more accurate since it directly addresses the AI's lack of physical presence, while Assistant 1 assumes the user wants their own distance. Referee C finds Assistant 1 more helpful and relevant, as it attempts to provide a solution, whereas Assistant 2's response lacks helpfulness.
2. **I agree with Referee A and Referee C's evaluations.** While Assistant 2's response is accurate, it does not help the user find the information they might be seeking. Assistant 1, on the other hand, actively engages with the user and offers a way to obtain the desired information, making it more helpful overall.

The best referee is: C


Meta Judge

Conclude Mode

1. The previous referees noted that Assistant 1 was more helpful by asking for the user's location to calculate the distance, while Assistant 2 accurately stated that it has no physical location. However, some referees felt that Assistant 1's response was less relevant to the question since it didn't address the AI's lack of physical presence. There was a mix of opinions on the scores, with some favoring Assistant 1 for its engagement and others favoring Assistant 2 for its straightforwardness.
2. I agree with the general sentiment that Assistant 1 is more helpful because it attempts to engage with the user and provide a way to calculate the distance. However, I also see the merit in Assistant 2's response for its accuracy regarding the AI's nature. Ultimately, while Assistant 1 is more interactive, it could have acknowledged the AI's lack of physical presence, which would have made it more complete.
3. In my evaluation, Assistant 1 is more helpful and relevant as it seeks to provide a personalized response, even if it doesn't fully address the AI's nature. Assistant 2 is accurate but lacks helpfulness and engagement. Therefore, I would rate Assistant 1 higher for its attempt to assist the user.

The score of Assistant 1: 7 The score of Assistant 2: 4


Meta Judge

Figure 25: A meta-judge example under the setting of **PINE** with **Position Bias**. The Meta Judge is gpt 4o-mini. Two meta-judging modes, Select and Conclude, are demonstrated, and the comments on the referees' judgments are **highlighted**.