

Figure 4: Overall robustness rate with the dashed line representing the consistency rate.

self-enhancement bias, we selected four models solely for response generation: Mixtral-8x22b ([AI@Mistral, 2024](#)), Llama3-70b ([AI@Meta, 2024](#)), Llama3-8b ([AI@Meta, 2024](#)), and Mistral-7b ([AI@Mistral, 2023](#)).

**Judgement prompt  $P$ .** The instruction  $I$  in the judgement prompt  $P = (I, Q, R)$  is derived from [Liu et al. \(2023a\)](#) and [Zheng et al. \(2024\)](#), with slight variations to evaluate the impacts of biases in LLM-as-a-Judge. The complete instruction we used is provided in [Appendix F](#).

**Hyperparameters.** We followed the experimental setup of [Chen et al. \(2024a\)](#) by setting the temperature to 0.7 and applied it to all judge models and generating models to ensure stable output quality and strong reproducibility.

## 4 EVALUATION RESULTS

In this section, we introduce our main results and related analyses from our exploratory experiments. We show the main results in [Figure 4](#) and [Table 4](#). Furthermore, we conduct exploratory experiments to evaluate the potential influence bias factor in LLM-as-a-Judge, which are detailed in [Figure 5](#), [Table 5](#), [Figure 6](#) and [Figure 7](#). Due to the space limitation, we show more detailed information of experiment results in [Appendix D](#).

### 4.1 MAIN RESULT

**Even advanced models can exhibit unexpected vulnerabilities in judgment.** [Figure 4](#) illustrates the influence of 12 distinct biases on the judging capabilities of six LLMs. Notably, the effects of these biases differ across models, and advanced models may not always exhibit better performance when dealing with these biases. While Claude-3.5 generally shows the greatest resilience to biases, our findings reveal that even highly proficient models can struggle. For example, despite its advanced capabilities ([Zheng et al., 2023](#)), GPT-4-Turbo exhibits inconsistency when judging emotional responses, whereas ChatGPT demonstrates more stable performance. This complexity suggests that identifying the *best* model is not straightforward; it depends on the specific bias involved, and even top-tier models may display unexpected weaknesses. Therefore, when using LLMs as judges, it is crucial to acknowledge these complexities and avoid assuming that the *most advanced model will always be the most reliable*.

**Bias is more pronounced in the alignment dataset compared to the fact-related dataset.** According to [Table 4](#), the impact of bias is more pronounced in the alignment dataset than in the fact-related dataset. One possible explanation for this is that, in the fact-related dataset, the quality differences between answers are more evident, which means that the influence of bias is insufficient to completely offset this quality gap. In contrast, the alignment dataset typically has smaller quality differences between answers, making the choices of the judge model more vulnerable to bias. Therefore, when developing a reliable LLM-as-a-Judge framework across different datasets, it is crucial to consider the inherent quality of the data.

Table 4: Robustness rate for various models across different metrics are presented.  $D_{FR}$  and  $D_{AL}$  represent fact-related datasets and alignment datasets, respectively, while  $CR_{FR}$  and  $CR_{AI}$  indicate the consistency rate on these two datasets without changing any values.

Model	$D_{FR} RR_{\uparrow}$				$D_{AL} RR_{\uparrow}$						$D_{AL} Acc_{\uparrow}$	
	Ver.	Fal.	Sen.	$CR_{FR}$	Pos.	Com.	Ban.	Aut.	Dst.	Div.	$CR_{AI}$	CoT.
<b>ChatGPT</b>	0.900	0.917	<b>0.804</b>	0.998	0.566	0.862	0.688	0.662	0.713	0.679	0.906	0.560
<b>GPT-4-Turbo</b>	0.915	0.969	0.653	0.990	0.818	0.858	0.638	0.846	0.729	0.855	0.856	0.720
<b>GPT-4o</b>	<b>0.977</b>	0.984	0.699	0.998	0.776	0.868	<b>0.791</b>	0.787	0.790	0.814	0.925	0.700
<b>GLM-4</b>	0.887	0.979	0.679	0.970	0.781	0.835	0.690	0.796	0.814	0.788	0.884	0.688
<b>Claude-3.5</b>	0.952	<b>0.985</b>	0.660	0.999	<b>0.832</b>	0.875	0.610	<b>0.865</b>	<b>0.878</b>	<b>0.914</b>	0.915	<b>0.745</b>
<b>Qwen2</b>	0.884	0.935	0.651	0.994	0.760	<b>0.877</b>	0.710	0.779	0.785	0.826	0.904	0.704

**Bias reflects cognitive and philosophical issues beyond technical defects.** The bias in LLMs may originate from the inherent limitations of human cognition. For instance, LLMs perform inconsistently when dealing with sentiment bias, potentially reflecting the phenomenon that humans are often influenced by emotions when making judgments. In cognitive psychology, this phenomenon is known as the *affect heuristic* (Slovic et al., 2002). Recent research has demonstrated that LLMs have inherited this human cognitive trait to some extent (Li et al., 2024a;b), prompting us to reconsider whether models should completely mimic human cognitive patterns or transcend these limitations. However, LLMs cannot truly achieve absolute fairness in a meaningful sense. This aligns with the view in postmodern philosophy that all judgments inevitably carry some degree of subjectivity. Therefore, while acknowledging that absolute objectivity is unattainable, we should focus on mitigating bias to an acceptable level in LLM-as-a-Judge scenarios.

#### 4.2 ANALYSIS OF EXPLORATORY EXPERIMENTS

**Position bias increases with more answer candidates.** Figure 6 demonstrates that all judge models are significantly impacted by position bias. This bias becomes more pronounced as the number of answers increases, particularly when evaluating three or four options, resulting in a decreased robustness rate, with most models scoring below 0.5. To mitigate the effects of position bias, we recommend using judge models with better robustness rate metrics or randomizing the order of answers (Zheng et al., 2024; Li et al., 2023b).

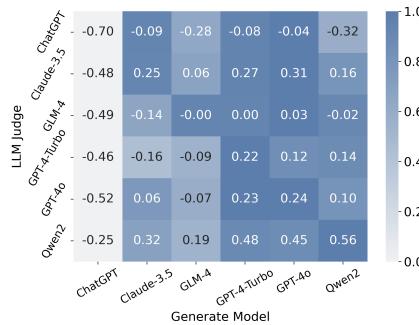


Figure 5: Heat map of model Z-score normalization score of self-enhancement bias.

**Response length influences model judgment in complex ways.** As illustrated in Figure 6, increasing response length without a corresponding improvement in quality led to a decline in model robustness rate. Some models exhibited an aversion to excessively verbose answers, while others demonstrated a positive correlation between model preference and response length.

**Avoid using the same model to generate and judge answers.** Analysis of Figure 5, Figure 7, and Table 5 reveals a significant self-enhancement bias among LLMs. Most models rated their outputs more favorably, even when answer sources were anonymized. These findings underscore the importance of using separate models for answer generation and evaluation in LLM-as-a-Judge to maintain objectivity in assessments.

**Bandwagon-effect involvement percentage is not impactful.** The percentage of people favoring an answer did not significantly impact model robustness rate. GPT-4o remained consistent, while Claude-3.5 was more influenced by popular opinion. Figure 6 shows that involvement percentage does not significantly affect model choices.

**LLMs show sensitivity to irrelevant content in responses.** Figure 7 demonstrates that including irrelevant content reduces the robustness rate of model judgments. Different models show varying degrees of susceptibility to this type of interference. Notably, from the average, the impact is more

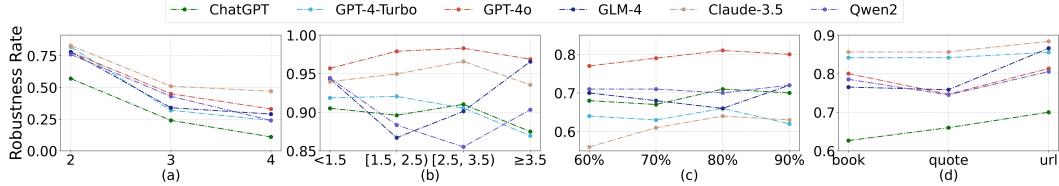


Figure 6: (a) shows the impact of the number of answers  $n$  on the robustness rate in position bias. (b) shows the relationship between the answer length ratio to the original length and robustness rate in verbosity bias. (c) shows the relationship between different percentages of popular opinion and robustness rate in bandwagon-effect bias. (d) shows the relationship between different models and robustness rate in authority bias with different fake citation formats.

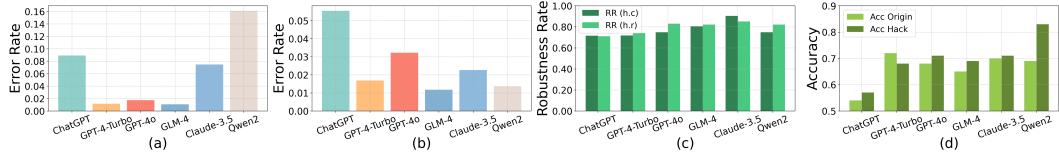


Figure 7: (a) and (b) show the comparisons of model error rates for refinement-aware bias and self-enhancement bias, respectively. (c) shows the robustness rate of various models when faced with distraction bias. (d) presents a comparison of model accuracy under the influence of CoT bias, indicating that most models achieve higher accuracy after applying CoT.

significant when perturbing high-quality responses, implying that extraneous information has a greater potential to disrupt the evaluation of strong answers.

**Different types of fake authorities interfere with the LLMs to varying degrees.** As illustrated in Figure 6, the impact of fake authorities on judge models differs based on the format used. URL citations consistently showed the least interference across all models, likely due to their concise nature and the models' familiarity with web-based references. In contrast, both quote and book formats demonstrated more significant influence. Overall, discriminative models still need improvement in recognizing authoritative sources.

**LLMs tend to prefer content without emotional elements.** Results in Figure 8 show that when emotionally charged revisions are made to superior answers, accuracy and robustness rates typically decline; conversely, when similar revisions are applied to inferior answers, these metrics tend to improve. Among emotions, *cheerful* has the least impact on models, with minimal decreases in accuracy and robustness rates. The other three emotions show greater effects, with *fear* having the most significant impact. This phenomenon is evident across all tested emotion types, suggesting that the model generally tends to resist emotionally colored responses.

**Explicit introduction of minority groups will influence the choices of LLMs.** As shown in Figure 8, most models demonstrated a more pronounced sensitivity to female and refugee status, whereas Claude-3.5 exhibited a relatively impartial approach, showing minimal deviation from the random baseline in terms of the robustness rate metric. Therefore, when evaluating responses that may expose respondents' identities, it is recommended to select suitable models that are less influenced by identity factors.

**CoT improves LLMs evaluation accuracy.** As shown in Figure 7, encouraging models to engage in step-by-step reasoning before concluding enhances their problem-solving abilities. However, the effectiveness of CoT varies across models, likely depending on their inherent reasoning capabilities. We can refer to Table 7 for the results. GPT-4-Turbo exhibited only a marginal improvement of 0.7%

Table 5: Average score and error rate of self-enhancement bias and refinement-aware bias.

Model	Sel. Score <sub>↓</sub>			Ref. Score <sub>↓</sub>		
	Self	Other	Error	Ref	+History	Error
ChatGPT	5.21	5.72	8.91	5.23	4.94	5.80
GPT-4-Turbo	6.98	6.90	<b>1.16</b>	8.31	8.45	1.66
GPT-4o	7.01	6.89	1.74	7.44	7.20	3.33
GLM-4	7.73	7.64	1.18	7.64	7.73	<b>1.15</b>
Claude-3.5	7.04	6.55	7.48	7.51	7.68	2.17
Qwen2	7.64	6.58	16.1	7.29	7.39	1.33