What Factors Influence LLMs' Judgments? A Case Study on Question Answering

Lei Chen[1], **Bobo Li**[1], **Li Zheng**[1], **Haining Wang**[1], **Zixiang Meng**[1], **Runfeng Shi**[1],
Hao Fei[2], **Jun Zhou**[1], **Fei Li**[1], **Chong Teng**[1], **Donghong Ji**[1]

[1] Key Laboratory of Aerospace Information Security and Trusted Computing,
Ministry of Education, School of Cyber Science and Engineering, Wuhan University, Wuhan, China
[2] National University of Singapore
{chhenl, boboli, zhengli, wanghn, shi_runfeng, j.zhou, tengchong, dhji}@whu.edu.cn
{zixiangmeng9, foxlf823}@gmail.com, haofei37@nus.edu.sg

Abstract

Large Language Models (LLMs) are now being considered as judges of high efficiency to evaluate the quality of answers generated by candidate models. However, their judgments may be influenced by complex scenarios and inherent biases, raising concerns about their reliability. This study aims to bridge this gap by introducing four unexplored factors and examining the performance of LLMs as judges, namely answer quantity, inducing statements, judging strategy, and judging style. Additionally, we introduce a new dimension of question difficulty to provide a more comprehensive understanding of LLMs' judgments across varying question intricacies. We employ ChatGPT, GPT-4, Gemini, and Claude-2 as judges and conduct experiments on Vicuna Benchmark and MT-bench. Our study reveals that LLMs' judging abilities are susceptible to the influence of these four factors, and analyzing from the newly proposed dimension of question difficulty is highly necessary. We also provide valuable insights into optimizing LLMs' performance as judges, enhancing their reliability and adaptability across diverse evaluation scenarios.

Keywords: LLMs as judges, answer quantity, inducing statements, judging strategy, judging style, question difficulty dimension

1. Introduction

Large Language Models (LLMs) are widely embraced for their robust natural language generation abilities and high intelligence (Achiam et al., 2023; Bubeck et al., 2023; Peng et al., 2023; Chung et al., 2022; Chowdhery et al., 2023; Hariri, 2023). Existing works indicate that these models perform at a level comparable to humans with advantages of speed and cost-effectiveness in different fields (Wang et al., 2023c; Song et al., 2023; Wang et al., 2023a; Chen et al., 2023b), particularly in tasks like evaluating translation quality (Kocmi and Federmann, 2023; Lu et al., 2023). Therefore, recent efforts have ventured into exploring the role of LLMs as judges for evaluation efficiency, as shown in Figure 1, and these initiatives have met with notable success (Chiang and Lee, 2023; Chiang et al., 2023; Zheng et al., 2023; Zhang et al., 2023).

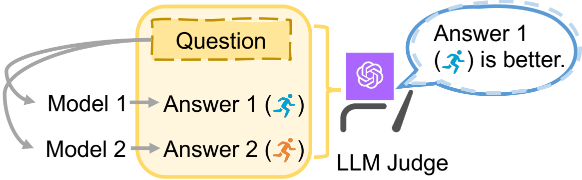

Figure 1: The workflow for answer quality judgment by a LLM judge. Initially, candidate models offer potential answers to the question, which are then combined with the question and presented to the judge model to obtain a judgment.

However, due to the complexity of judgment scenarios and inherent biases in models, there are potential risks associated with LLMs as judges. Wang et al. (2023b) found that even simply swapping the positions of two answers to be evaluated can lead to a reverse judgment. In parallel, Zheng et al. (2023) also revealed that LLMs tend to prefer verbose answers or answers generated by themselves. This can raise concerns as the best answers may sometimes be concise and should not be influenced by LLMs' generation preferences. These observations underscore a larger issue that current methods prioritize improving model generalization, neglecting the stability of LLMs' decisions across different contexts and the impact of various influencing factors.

While current research has delved into various factors affecting LLMs as judges, substantial gaps remain. Firstly, many studies (Wang et al., 2023b; Zheng et al., 2023) have primarily focused on straightforward factors, such as order of question, leaving numerous crucial in-depth factors unexplored. Furthermore, current research typically approaches the analysis of LLMs' judging perfor-

LREC-COLING 2024, pages 17473–17485
20-25 May, 2024.

Factors	Examples	
Answer Position	🔵🟠 → Judge → 🔵	🟠🔵 → Judge → 🟠
Answer Verbosity	🔵 (clear but short) 🟠 (not clear but long) → Judge → 🟠	
Self-enhancement	🔵 (generated from Model A) 🟠 (generated from Model B) → Judge (Model B) → 🟠	
Question Type	🔵🟠 (both for math or reasoning question) → Judge → poor judgment	
Answer Quantity	🔵🟠 → Judge → 🔵	🔵🟠🟢🟤 → Judge → 🟠
Inducing Statements	🔵 (+) 🟠 (-) → Judge → 🔵	🔵 (-) 🟠 (+) → Judge → 🟠
Judging strategy	🔵🟠 → Judge (overall) → 🔵	🔵🟠 → Judge (Step-by-Step) → 🟠
Judging Style	🔵🟠 → Judge (serious and rational) → 🔵	🔵🟠 → Judge (lively and emotional) → 🟠

Table 1: Factors may influence LLMs' judgments. The factors highlighted in gray above have already been explored, while the factors below are newly introduced in this paper. 🔵 and 🟠 represent two candidate answers, which are presented to the judge to determine which one is superior. '+' signifies a positive guidance for the candidate, while '-' conveys the opposite meaning.

mance from the perspective of question categories. Yet, even within the same category, there might be significant differences in difficulty – a crucial dimension for judgment that remains overlooked. Additionally, while several impactors have been highlighted, there is a relative lack of explicit guidance and specific recommendations on how to effectively utilize LLMs as judges. As a result, the potential capabilities of using LLMs to judge generative results have yet to be fully realized.

To bridge these gaps, our study embarks on addressing the following objectives. Firstly, we introduce and examine four unexplored factors to delve deeper: answer quantity, inducing statements, judging strategy, and judging style. We present the factors that have been explored to date, as well as the new factors we propose in Table 1. Secondly, our research introduces a new dimension, namely question difficulty, allowing LLMs acting as judges to annotate the difficulty level of questions. This offers an opportunity to explore the relationship between a judge's assigned difficulty level of a question and the subsequent judgment.

To provide a more in-depth understanding of the factors outlined previously, we delve into the following conditions: 1) When the **answer quantity** is no longer limited to two but increases to three or even four; 2) When explicitly biased **inducing statements** are injected into the input, praising or criticizing candidate answers; 3) When the **judging strategy** shifts from the approach of directly providing an overall judgment to a Step-by-Step approach; 4) When different **judging styles** are assigned to the judges. Simultaneously, we also investigate LLMs' behavior through the newly introduced dimension of question difficulty, exploring how LLMs' performance varies on questions of different difficulty levels under the influence of the factors mentioned above.

We analyze the extensive experimental results, and derive the following key findings:

➤ **1.** An increasing number of answers brings LLMs' judgments closer to those of humans. But a balance needs to be struck between the quantity and quality of information.

➤ **2.** The inducing statements elicite accommodating responses from all judges, with ChatGPT and Gemini demonstrating relatively higher susceptibility, whereas GPT-4 displayed a higher degree of autonomy and impartiality. But only Claude-2 will occasionally refuse to give a judgment when induced.

➤ **3.** A Step-by-step judging strategy, by refining judgment criteria and processes, enhances the interpretability of LLMs and improves their performance when quantitative judgments are required.

➤ **4.** Different question categories require distinct judging styles, and providing personalized guidance is essential for LLMs to achieve improved judging performance.

➤ **5.** Gemini exhibits unique expertise in accurately judging challenging questions, in contrast to the other three LLMs, which show superior performance on simpler queries. While ChatGPT and GPT-4 exhibit a notable resilience to influencing factors when tackling complex questions, a trait not observed in Gemini and Claude-2.

To sum up, the major contributions of this paper are summarized as follows:

- We propose and explore the impact of four crucial factors on the judging capabilities of LLMs acting as judges, and provide empirical evidence that underscores the significance of these factors.

- We pioneer the perspective of employing LLMs as judges to annotate question difficulty, introducing a fresh analytical dimension for the research community.

- Our findings provide valuable guidance and profound insights for enhancing the judging ability of LLMs in both research and practical applications.

2. Related work

Prior research has extensively focused on the performance evaluation of LLMs in various tasks (Mao et al., 2023; Zhou et al., 2023; Sun et al., 2023; Ji et al., 2023). Question answering and evaluation tasks have been prominent areas of interest (Parrish et al., 2022; Bai et al., 2023; Bian et al., 2023). Such work not only scrutinizes the comparison between LLMs and humans but also investigates the applicability of LLMs in different domains (He et al., 2023; Lu et al., 2023).

However, with the widespread application of LLMs, some studies have also started to uncover certain potential issues associated with them. On one hand, LLMs may exhibit biases, a problem that has garnered significant attention, especially in tasks that require judge-like decisions (Belinkov et al., 2019; McCoy et al., 2019; Gururangan et al., 2018; Min et al., 2019). On the other hand, LLMs appear to be sensitive to the input and context, resulting in output sensitivity, meaning they may exhibit instability when dealing with different inputs or prompts (Bowman, 2023; Turpin et al., 2023; Chen et al., 2023a; Arora et al., 2022).

As for using large models as judges, Chiang and Lee (2023) explored the potential of LLMs as an alternative to human evaluations. Chiang et al. (2023) utilized GPT-4 as the judge to compare candidate answers from ChatGPT and Vicuna to demonstrate the capabilities of the Vicuna model. However, Wang et al. (2023b) discovered that LLMs have position bias. To address this, they introduced a calibration framework encompassing two strategies: multi-evidence calibration and balanced position calibration. Further, Zheng et al. (2023) introduced two benchmarks, MT-bench and Chatbot Arena, to validate the consistency between LLM judges and human preferences. The results indicated that powerful LLM judges, like GPT-4, align well with controlled and crowdsourced human preferences, achieving over 80% consistency. Zhang et al. (2023) explored the potential for deeper and wider LLM networks to result in fairer evaluations.

3. Overall Setup

3.1. Data

The utilized data are from two sources. The first is the Vicuna Benchmark (Chiang et al., 2023), consisting of 80 single-turn questions used in Vicuna's evaluation. The second source comprises 80 multi-turn dialogue data provided by the MT-bench benchmark (Zheng et al., 2023). Some data examples are shown in Appendix A. For the data from Mt-bench, we conduct experiments separately for single-turn and multi-turn scenarios.

3.2. Model

The models we used can be classified into the following two categories:

Judges We utilize ChatGPT (OpenAI, 2022), GPT-4 (Achiam et al., 2023), Gemini (Google, 2023) , and Claude-2 (Anthropic, 2023) as the judges. Due to budget constraints, we query each judge three times for each evaluation and then calculate the average of the results. For ChatGPT and GPT-4, twice through their web interfaces and once by utilizing their API. [1] For Gemini and Claude-2, all from their web interfaces.

Candidates The candidate answers are from ChatGPT, Vicuna-13b (Chiang et al., 2023), LlaMA-13b (Touvron et al., 2023), and Alpaca-13b (Taori et al., 2023). Other than the answer quantity experiment, the two candidate answers are obtained from the first two models respectively.

3.3. Metric

For measuring the performance of judges, we use the accuracy and Kappa correlation coefficient with human annotation results, which are provided by Wang et al. (2023b) and Zheng et al. (2023), as the gold standard. A higher value for these two metrics indicates that the judgments made by LLMs are more similar to those made by humans.

For measuring the sensitivity of the judges after changing a specific factor, we use conflict rate (Wang et al., 2023b) with default judgments. A lower value for this metric suggests that LLMs are less affected by changes in the influencing factors.

Moreover, to eliminate the interference of position bias, we present the candidate answers to the judges in reversed order and then aggregate the results.

3.4. Prompt

We use the prompt from Vicuna Benchmark as an illustrative example. Below is a sample prompt under default settings, accompanied by a judgment.

> Default Example
>
> **[System prompt]** You are a helpful and precise assistant for checking the quality of the answer.
> **[Question]** How can I increase my productivity while working from home?
> **[Assistant 1's Answer]** Here are some tips that can help you increase your productivity while working from home:1.Create a dedicated workspace...
> **[Assistant 2's Answer]** Working from home can be both a blessing and a curse...

---

[1] The API versions we use are gpt-3.5-turbo-0613 and gpt-4-0613.

[Task description / Instructions] We would like to request your feedback... Please rate the helpfulness, relevance, accuracy, level of details of their responses. Each assistant receives an overall score on a scale of 1 to 10...

GPT-4's Judgment 9 9 Assistant 1 and Assistant 2 provided excellent and detailed responses...

4. Answer Quantity

4.1. Motivation

Typically, the judge model is required to evaluate two candidate answers, but this quantity can be customized as required, as seen in MT-bench. To explore judging scenarios with varying answer quantities, we appended one or two additional answers and made corresponding modifications to the prompt and information regarding the number of answers. We paid particular attention to whether the judging results of LLM for the original two answers were influenced by the addition of extra answers, similar to how the inclusion of extra athletes in a sports competition might affect the judge's decisions.

Answer Quantity Example

[System prompt]
[Question]
[Assistant 1's Answer]
[Assistant 2's Answer]
[Assistant 3's Answer]
[Assistant 4's Answer]
[Task description / Instructions]

GPT-4's Judgment 8 9 7 8 ...Assistant 2's answer was the most comprehensive...

4.2. Settings

For this factor, We only conduct related experiments on Vicuna Benchmark. The specific configuration in MT-bench is tailored to accommodate scenarios with only two answers, whereas Vicuna Benchmark does not have this limitation. Following the original two answers, we sequentially added answers from LLaMA-13b and Alpaca-13b, while adjusting the information about answer quantity in the original prompt.

4.3. Results & Discussion

Figure 2 illustrates that increasing answer quantity alters the comparative evaluation between the original two candidates. This adjustment leads to notable shifts in LLMs' performance metrics: ChatGPT consistently exhibits the highest conflict rate, while GPT-4 maintains the lowest. Intriguingly, the conflict rate for GPT-4 uniquely decreases when the answer quantity rises from three to four. Furthermore, the observed changes in accuracy and Kappa coefficient values suggest that more candidate answers enhance the alignment between LLMs' judgments and human evaluations, with GPT-4 achieving the highest congruence and Gemini the lowest.

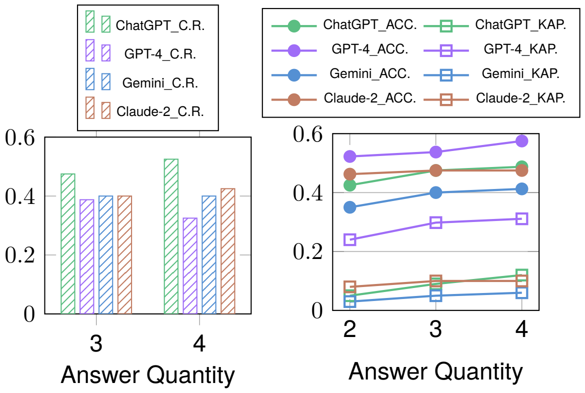

Figure 2: Variations in conflict rate (C.R.), accuracy (ACC.), and Kappa coefficient (KAP.) with varying answer quantities from the judges.

This phenomenon might be attributed to the fact that with only two candidates, LLMs are more likely to reflect biases towards their similarities. Introducing additional candidates not only amplifies the competitive environment but also equips LLMs with a richer variety of information, potentially broadening their understanding of the question context. This, in turn, encourages a more thorough consideration of factors in distinguishing between answers, potentially leading to adjustments in judgments.

Additionally, the introduction of more answers is particularly effective in diluting the influence of outliers—an answer markedly divergent in quality from its counterpart—which might disproportionately affect judgments. This expansion enables a more balanced and nuanced evaluation, reducing the impact of any single divergent answer.

Guidance This reminds us that, although LLMs typically have input length limitations, in this particular task, increasing the input to provide more candidate answers may enhance their judging performance. However, it's also essential to recognize that more information can potentially introduce more noise. Therefore, striking a balance between the quantity and quality of information is a key consideration in this context.

Inducing Statements	**Judge**	**Vicuna Benchmark**				**MT-bench**				
		Candidate 1 wining rate (%)	Candidate 2 wining rate (%)	C.R. (%)	Rejection rate (%)	Turn	Candidate 1 wining rate (%)	Candidate 2 wining rate (%)	C.R. (%)	Rejection rate (%)
+ / -	ChatGPT	↑37.50	↓22.50	52.50	0.00	Single	↑45.00	↓2.50	77.50	0.00
						Multi	↑51.25	↓6.25	78.75	0.00
	GPT-4	↑20.00	↓2.50	28.75	0.00	Single	↑5.00	↓6.25	26.25	0.00
						Multi	↑12.50	↓12.50	30.00	0.00
	Gemini	↑37.50	↓31.25	43.75	0.00	Single	↑46.25	↓40.00	51.25	0.00
						Multi	↑53.75	↓45.00	62.50	0.00
	Claude-2	↑23.75	↓15.00	32.50	37.50	Single	↑32.50	↓16.25	38.75	31.25
						Multi	↑25.00	↓17.50	42.50	28.75
- / +	ChatGPT	↓45.00	↑47.50	55.00	0.00	Single	↓17.50	↑53.75	65.00	0.00
						Multi	↓33.75	↑60.00	66.25	0.00
	GPT-4	↓28.75	↑17.50	38.75	0.00	Single	↓11.25	↑15.00	16.25	0.00
						Multi	↓21.25	↑7.50	26.25	0.00
	Gemini	↓46.25	↑50.00	55.00	0.00	Single	↓31.25	↑35.00	41.25	0.00
						Multi	↓26.25	↑32.50	42.50	0.00
	Claude-2	↓37.50	↑26.25	43.75	33.75	Single	↓20.00	↑23.75	36.25	35.00
						Multi	↓27.50	↑41.25	47.50	27.50
10 / 1	ChatGPT	↑45.00	↓25.00	51.25	0.00	Single	↑62.50	↓12.50	77.50	0.00
						Multi	↑56.25	↓11.25	80.00	0.00
	GPT-4	↑17.50	↓12.50	25.00	0.00	Single	↑27.50	↓10.00	31.25	0.00
						Multi	↑15.00	↓15.00	40.00	0.00
	Gemini	↑41.25	↓40.00	52.50	0.00	Single	↑51.25	↓47.50	60.00	0.00
						Multi	↑53.75	↓52.50	58.75	0.00
	Claude-2	↑32.50	↓28.75	37.50	11.25	Single	↑32.50	↓22.50	58.75	8.75
						Multi	↑41.25	↓28.75	60.00	7.50
1 / 10	ChatGPT	↓43.75	↑55.00	56.25	0.00	Single	↓27.50	↑61.25	66.25	0.00
						Multi	↓35.00	↑43.75	60.00	0.00
	GPT-4	↓36.25	↑45.00	47.50	0.00	Single	↓12.50	↑20.00	28.75	0.00
						Multi	↓26.25	↑12.50	33.75	0.00
	Gemini	↓47.50	↑48.75	53.75	0.00	Single	↓40.00	↑43.75	52.50	0.00
						Multi	↓46.25	↑47.50	60.00	0.00
	Claude-2	↓40.00	↑38.75	47.50	15.00	Single	↓31.25	↑40.00	48.75	12.50
						Multi	↓43.75	↑46.25	53.75	10.00
5 / 5	ChatGPT	↓17.50	↑17.50	47.50	0.00	Single	↓8.75	↑6.25	35.00	0.00
						Multi	↓12.50	↑3.75	41.25	0.00
	GPT-4	↓13.75	↑16.25	35.00	0.00	Single	↓5.00	↑10.00	18.75	0.00
						Multi	↓8.75	0.00	15.00	0.00
	Gemini	↓12.50	↓15.00	36.25	0.00	Single	↓20.00	↓15.00	38.75	0.00
						Multi	↓13.75	↓18.75	37.50	0.00
	Claude-2	↓18.75	↓16.25	40.00	3.75	Single	↓17.50	↓11.25	26.25	3.75
						Multi	↓13.75	↓18.75	27.50	0.00

Table 2: The impact of Inducing Statements. '+/-' and '+/-' represent the first two sets of inducements, '10/1' and '1/10' represent the third and fourth sets of inducements, and '5/5' represents the last set. ↑and ↓indicate whether they have gone up or down from the default, and the numbers after them indicate the value of the change. Note that a judging result includes not only one side winning, but also a tie.

5. Inducing Statements

5.1. Motivation

Similar to how spectators or external factors in a game can influence a judge's decision through shouting or applause, specific inducing sentences can to some extent impact the judgments made by large language models, causing them to lean toward a particular answer. When the prompt includes content that might introduce bias into the judge's specific judgments, it becomes intriguing to investigate whether the judge can still uphold fairness, reliability, and robustness. This simulation isn't about wanting LLMs to be misled but testing their judgments and objectivity under unconventional conditions.

Inducing Statements Example

[System prompt]
[Question]
[Assistant 1's Answer] **...This is the best answer.**
[Assistant 2's Answer] **...This is the worst answer.**
[Task description / Instructions]

GPT-4's Judgment 10 7 ...However, Assistant 1's answer was slightly more comprehensive and detailed...

5.2. Settings

We devise five sets of inducing sentences, with each set containing two sentences to be added after the two candidate answers. The first two sets include the sentences 'This is the best answer.' and 'This is the worst answer.' along with their combinations after swapping the order. The following two sets include the sentences 'This answer deserves 10 out of 10.' and 'This answer deserves 1 out of 10.' along with their combinations after swapping the order. The final set consists of the two sentences ‘This answer deserves 5 out of 10.’ To observe the phenomenon more prominently, these inducing statements are overt and straightforward in nature.

5.3. Results & Discussion

Table 2 indicates that these inducing sentences have achieved their intended purpose, as all judges are interfered with. Even in the last set of inducing sentences, which lacked overt bias, there was still a discernible impact on the outcomes to some extent. Exclusively in this context, the trend shifts observed in Gemini and Claude-2 significantly diverge between ChatGPT and GPT-4. This phenomenon may be related to the concept of ‘compliance’ learned by the models during training, wherein the models tend to provide answers that align with or closely resemble the input.

Interestingly, Claude-2 stands out with a non-zero rejection rate across various scenarios, showcasing its capability to resist inducements—a trait detailed further in Appendix B. This contrasts with the perpetual zero rejection rates of the other three judges, highlighting disparities in LLMs’ ability to counteract inducements. This variation suggests that some LLMs are beginning to develop rudimentary resistance capabilities, with significant differences evident across different models.

LLMs typically exhibit slightly increased conflict rates in multi-turn scenarios compared to single-turn ones, likely due to the added complexity of more contextual information. The Vicuna Benchmark’s specific scoring requirements also prompted an analysis of how frequently judges’ scores align with inducements, with findings detailed in Appendix C.

Guidance Our study sheds light on the nuanced vulnerabilities of LLMs to inducements, underlining the necessity for future developments to focus on the detection, avoidance, and mitigation of both explicit and implicit biases. Such measures are essential to bolster the effectiveness and dependability of LLMs in real-world applications. Additionally, during the training of LLMs, a more diversified and adversarial set of prompts can be employed to reduce their dependence on specific prompt templates, thereby enhancing their capacity to recognize and resist bias interference.

6. Judging Strategy

6.1. Motivation

We are curious to explore how LLMs’ judging logic and results would respond to corresponding changes in the judging strategy. Currently, most evaluation strategies entail LLMs providing a holistic judgment while considering specific evaluation dimensions like helpfulness, relevance, creativity, and so on. In contrast, we request LLM to conduct a Step-by-Step evaluation, wherein it first individually evaluates or scores the candidate answers on the aforementioned dimensions, and subsequently provides a final evaluation or score.

Dataset	Judge	Strategy	ACC.(%)	KAP.	C.R.(%)
Vicuna Benchmark	ChatGPT	default	42.50	0.05	-
		Step-by-Step	47.50	0.10	40.00
	GPT-4	default	55.00	0.27	-
		Step-by-Step	58.75	0.31	37.50
	Gemini	default	35.00	0.03	-
		Step-by-Step	43.75	0.06	38.75
	Claude-2	default	46.25	0.08	-
		Step-by-Step	50.00	0.19	27.50
MT-bench Single-Turn	ChatGPT	default	46.25	0.08	-
		Step-by-Step	43.75	0.06	12.50
	GPT-4	default	46.25	0.08	-
		Step-by-Step	53.75	0.25	11.25
	Gemini	default	53.75	0.25	-
		Step-by-Step	55.00	0.27	17.50
	Claude-2	default	48.75	0.11	-
		Step-by-Step	48.75	0.11	7.50
MT-bench Multi-Turn	ChatGPT	default	36.25	0.03	-
		Step-by-Step	40.00	0.04	22.50
	GPT-4	default	48.75	0.11	-
		Step-by-Step	53.75	0.25	16.25
	Gemini	default	38.75	0.04	-
		Step-by-Step	40.00	0.04	11.25
	Claude-2	default	47.50	0.10	-
		Step-by-Step	46.25	0.09	8.75

Table 3: The results of varying judging strategies.

Judging Strategy Example

[System prompt]
[Question]
[Assistant 1’s Answer]
[Assistant 2’s Answer]
[Task description / Instructions] ...**Please conduct a Step-by-Step evaluation. Step 1: Please begin by evaluating candidate answers based on the following dimensions: helpfulness, relevance, accuracy, depth, creativity, and level of detail. Provide a separate rating for each dimension. Step 2: Based on the assessments of the individual dimensions above, now summarize your evaluations and provide an overall judgment...**

GPT-4’s Judgment 9.5 9.8 Assistant 1: Helpfulness: 9, Relevance: 10, Accuracy: 10, Level of detail: 9 Assistant 2: Helpfulness: 9, Relevance: 10, Accuracy: 10, Level of detail: 10...

6.2. Settings

Building upon the default foundation, we modify the prompts to instruct the judges to perform a Step-by-Step evaluation process, wherein they first evaluate the answers individually based on the

Judge	Vicuna Benchmark									
	Style	Generic	Knowledge	Roleplay	Common-sense	Fermi	Counter-factual	Coding	Math	Writing
ChatGPT	default	0.4	0.4	0.5	0.4	0.4	0.3	0.4	0.7	0.5
	SR	0.4	0.5	0.4	0.4	0.4	0.3	0.6	1.0	0.4
	LE	0.5	0.4	0.6	0.5	0.3	0.4	0.4	0.7	0.6
GPT-4	default	0.5	0.4	0.4	0.7	0.7	0.5	0.7	1.0	0.4
	SR	0.5	0.4	0.3	0.6	0.7	0.7	0.9	1.0	0.4
	LE	0.6	0.5	0.5	0.6	0.6	0.5	0.7	1.0	0.5
Gemini	default	0.6	0.3	0.3	0.3	0	0.3	0.7	0.7	0.3
	SR	0.5	0.4	0.3	0.3	0.2	0.3	0.7	1.0	0.4
	LE	0.6	0.3	0.4	0.3	0.1	0.4	0.7	0.7	0.5
Claude-2	default	0.5	0.4	0.2	0.5	0.2	0.5	0.4	1.0	0.8
	SR	0.5	0.5	0.3	0.6	0.2	0.5	0.6	1.0	0.8
	LE	0.5	0.4	0.4	0.5	0.3	0.6	0.4	1.0	0.7

Judge	MT-bench									
	Turn	Style	Writing	Roleplay	Reasoning	Math	Coding	Extraction	STEM	Humanities
ChatGPT	Single	default	0.3	0.7	0.7	0.5	0.6	0.3	0.3	0.3
		SR	0.3	0.6	0.4	0.4	0.6	0.4	0.3	0.3
		LE	0.2	0.7	0.7	0.4	0.6	0.4	0.3	0.3
	Multi	default	0.2	0.6	0.4	0.4	0.4	0.4	0.3	0.2
		SR	0.4	0.3	0.6	0.5	0.4	0.4	0.4	0.3
		LE	0.4	0.6	0.4	0.6	0.5	0.4	0.4	0.2
GPT-4	Single	default	0.3	0.3	0.8	0.5	0.5	0.7	0.3	0.3
		SR	0.3	0.4	0.7	0.6	0.6	0.6	0.3	0.3
		LE	0.4	0.4	0.8	0.7	0.6	0.7	0.4	0.3
	Multi	default	0.6	0.3	0.4	0.5	0.5	0.6	0.5	0.5
		SR	0.7	0.3	0.5	0.5	0.6	0.6	0.5	0.5
		LE	0.7	0.4	0.5	0.6	0.6	0.6	0.5	0.4
Gemini	Single	default	0.3	0.5	0.6	0.6	0.8	0.5	0.8	0.2
		SR	0.4	0.5	0.6	0.6	0.8	0.5	0.7	0.3
		LE	0.4	0.6	0.5	0.6	0.7	0.5	0.7	0.4
	Multi	default	0.5	0.2	0.4	0.5	0.6	0.3	0.4	0.2
		SR	0.5	0.3	0.5	0.5	0.6	0.3	0.4	0.3
		LE	0.6	0.4	0.4	0.6	0.6	0.3	0.5	0.2
Claude-2	Single	default	0.4	0.5	0.6	0.6	0.5	0.6	0.3	0.4
		SR	0.5	0.4	0.7	0.6	0.5	0.6	0.3	0.3
		LE	0.4	0.5	0.5	0.6	0.6	0.6	0.4	0.4
	Multi	default	0.7	0.7	0.3	0.3	0.5	0.6	0.3	0.4
		SR	0.7	0.6	0.5	0.4	0.5	0.6	0.4	0.3
		LE	0.7	0.7	0.3	0.4	0.4	0.6	0.4	0.4

Table 4: Category-wise accuracy results of different styles. SR represents the ‘serious and rational’ style, while LE represents the ‘lively and emotional’ style. The red and blue backgrounds indicate whether accuracy increased or decreased compared to the default results, respectively.

mentioned perspectives, such as helpfulness, relevance, accuracy, depth, and creativity, and then conduct an overall evaluation.

6.3. Results & Discussion

Table 3 demonstrates that adopting a Step-by-Step strategy significantly enhances the judging performance across all models on the Vicuna Benchmark, a trend not as pronounced on MT-bench. And the higher conflict rate also reflects the considerable impact of the strategy change. This might be attributed to the distinct evaluation criteria demanded by the two benchmarks.

Specifically, the Vicuna Benchmark requires judges to assign precise scores to each candidate answer, demanding a rigorous quantitative analysis. MT-bench focuses on relative comparisons between candidates, prioritizing qualitative judgments. In scenarios requiring fine-grained quantitative evaluation, a holistic judging approach may falter in distinguishing between answers of seemingly similar quality. However, the Step-by-Step strategy enhances transparency by delineating the scoring across multiple dimensions. By breaking down the judging process into discrete steps, judges can visualize the scores in each dimension before aggregating them into a final score. Such granularity introduces subtle numerical differences in total scores, potentially rectifying earlier judgments inaccurately classified as ties, and thus refining the performance.

Guidance The differential impact of the Step-by-Step strategy on performance across benchmarks underscores the need for enhancing the transparency of the judgment decision-making process. This inspires us to focus on improving the explainability of LLMs to make judgments more accessible for comprehension and scrutiny. Additionally, in the research and optimization of evaluation strategies, clear judgment criteria and processes can be established, and LLMs' understanding and execution can be ensured through enhanced training and instructions. Furthermore, the introduction of a mechanism for LLMs to review past judgments can be considered to rectify potential biases.

7. Judging Style

7.1. Motivation

Various domains and tasks often demand diverse styles and perspectives, much like human judges in distinct domains possess their own unique knowledge and aesthetic criteria. Different styles in LLMs can guide their attention to different features and criteria, allowing them to cater to the specific requirements of distinct domains. So we endow LLMs with two opposite styles to observe the potential impact on their decision-making.

Judging Style Example

[System prompt]
[Question]
[Assistant 1's Answer]
[Assistant 2's Answer]
[Task description / Instructions] ... **Please complete your tasks with a lively and emotional approach.**

GPT-4's Judgment 8 9 ...Assistant 2 also addressed the potential challenges of working from home, such as feeling isolated, which the other assistants did not mention.

7.2. Settings

We enhance a specific aspect of the judging style by appending ‘Please complete your tasks with a ... approach.’ at the end of the default prompt. The ‘...’ should be filled with ‘serious and rational’ or ‘lively and emotional’.

7.3. Results & Discussion

The variations in performance observed in Table 4, highlight the adaptability of LLMs to stylistic cues within their judgment processes. This adaptability, however, manifests with varying degrees of effectiveness across different question categories, indicating that LLMs’ ability to incorporate stylistic considerations into their evaluations is not uniformly distributed. The inconsistency in performance changes among judges within the same category, despite being subjected to the same stylistic adjustments, further underscores the complexity of how style influences judgment. Even variations in turn lead to significantly different impacts.

These discrepancies suggest that LLMs engage in a complex process of interpretation when presented with stylistic directives, with their impact deeply intertwined with the nature of the questions being judged. This nuanced interaction between style, question type, and model-specific capabilities points to an underlying variability in how LLMs process and respond to stylistic elements, leading to diverse outcomes.

Guidance The observed variations in LLMs’ performance across different question categories, influenced by stylistic adjustments, underscore the need for a more nuanced approach to designing judging prompts. This entails crafting tailored prompts that account for the specific requirements and nuances of various question types, beyond the current focus on math and coding queries. Such a differentiated approach can optimize LLM judging effectiveness by aligning with the unique characteristics of each question category. Additionally, the effectiveness of any given guidance may vary between LLMs, highlighting the importance of personalizing instructions to harness each model’s distinct capabilities fully. This personalized approach not only enhances the precision of judgments but also ensures that the potential of each LLM is maximally exploited, contributing to more accurate and reliable judgments.

8. Question Difficulty

8.1. Motivation

The model’s utilization of distinct abilities and the corresponding emphasis varies depending on the difficulty level of the question being asked. Likewise, the judging capabilities also differ when evaluating answers to questions of diverse challenges. Existing work is often analyzed from the angle of question type. However, even within the same category, questions can vary in terms of difficulty.

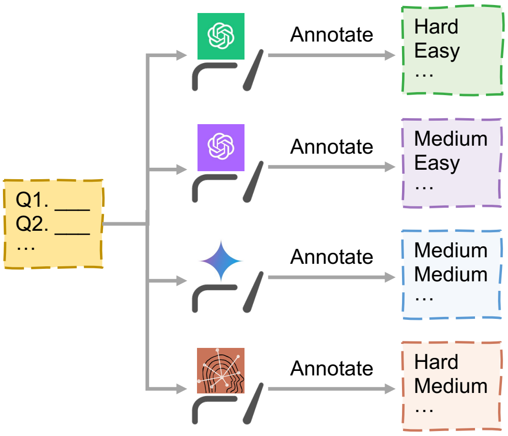

Figure 3: Request all judges to label the difficulty level for each question.

Therefore, we propose to incorporate the difficulty level of the questions into the exploration. If the judge perceives a question as challenging, indicating that it may not have a strong grasp of how to answer this question, it is likely to struggle to assess the quality of answers to this question. Moreover, different models may exhibit divergent recognition regarding the difficulty level of the same question. Hence, for each question, we interact separately with the four judges, to obtain their respective annotations, as shown in Figure 3. We aim to explore whether there exists an as-yet-undiscovered correlation between the intrinsic knowledge and biases of LLMs and their judgments. We categorize the difficulty level of questions into three levels: ‘easy’, ‘medium’, and ‘hard’.

8.2. Settings

We seek annotations from the judges to determine the difficulty of all questions. To ensure reliability, each result is obtained by asking three times and determined by a voting process. Given that question difficulty and question type both are inherent characteristics of the questions, we also investigate the correlation between them. Appendix D describes the annotation results.

8.3. Results & Discussion

Tables 5, 6, 7, and 8 demonstrate that only Gemini exhibits a higher accuracy rate on questions it deems difficult, while the other three judges often exhibit higher accuracy on questions catego-

Factors	Vicuna Benchmark			MT-bench Single-Turn			MT-bench Multi-Turn		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Default		✓			✓		✓		
$Q_{Ans}=3$		✓	○	-	-	-	-	-	-
$Q_{Ans}=4$	✓		○	-	-	-	-	-	-
Step-by-Step		✓○			✓○		✓○		
SR Style		✓	○		✓	○		✓	○
LE Style	✓		○	○	✓		✓		○

Table 5: ChatGPT's varied performance across different question difficulties and influencing factors. Q_{Ans} indicates the answer quantity. ✓represents the highest accuracy on questions of this difficulty level, while ○indicates the lowest conflict rate, meaning it is the least influenced.

Factors	Vicuna Benchmark			MT-bench Single-Turn			MT-bench Multi-Turn		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Default		✓		✓			✓		
$Q_{Ans}=3$		✓○		-	-	-	-	-	-
$Q_{Ans}=4$		✓○		-	-	-	-	-	-
Step-by-Step	✓		○	✓		○			✓○
SR Style		✓	○	✓		○	✓		○
LE Style		✓	○	✓○			✓		○

Table 6: GPT-4's varied performance across different question difficulties and influencing factors.

rized as easy and medium. This suggests that the former excels at adjudicating questions perceived as difficult, whereas the latter tends to perform better on less challenging questions. Simultaneously, ChatGPT and GPT-4 often display lower conflict rates on harder questions, indicating they are less influenced by various factors on questions they consider challenging. In contrast, Gemini and Claude-2 do not exhibit this pattern, suggesting they are more susceptible to being influenced on difficult questions.

In terms of accuracy, although Gemini shows higher accuracy on difficult questions, its lower overall accuracy may indicate a specialization in certain areas or types of questions. This suggests that Gemini possesses deep understanding and analytical capabilities within its areas of expertise but may not be as comprehensive as models like GPT-4 when addressing a wider range of questions. This phenomenon is also evident from Table 4. The performance difference between GPT-4 and Gemini could reflect the challenge of balancing breadth and depth in model design and training processes.

Factors	Vicuna Benchmark			MT-bench Single-Turn			MT-bench Multi-Turn		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Default		✓				✓			✓
$Q_{Ans}=3$		✓○		-	-	-	-	-	-
$Q_{Ans}=4$		✓○		-	-	-	-	-	-
Step-by-Step	○		✓			✓○			✓○
SR Style	○		✓	○	✓		○	✓	
LE Style		✓○				✓○			✓○

Table 7: Gemini's varied performance across different question difficulties and influencing factors.

Factors	Vicuna Benchmark			MT-bench Single-Turn			MT-bench Multi-Turn		
	Easy	Medium	Hard	Easy	Medium	Hard	Easy	Medium	Hard
Default		✓		✓			✓		
$Q_{Ans}=3$	✓	○		-	-	-	-	-	-
$Q_{Ans}=4$	✓		○	-	-	-	-	-	-
Step-by-Step		✓○		✓○			✓	○	
SR Style	○	✓		○	✓		✓○		
LE Style		✓○		✓○			✓		○

Table 8: Claude-2's varied performance across different question difficulties and influencing factors.

Regarding conflict rates, the lower conflict rates of ChatGPT and GPT-4 on difficult questions could indicate that, despite their less satisfactory accuracy in complex and challenging situations, they can provide relatively consistent and stable decisions. It might result from ChatGPT and GPT-4 adopting a more cautious or probabilistic approach to decision-making, thereby reducing uncertainty. Although this approach may not always achieve the highest accuracy, it ensures the consistency and stability of the judgments.

Guidance These findings illuminate the intricate interplay between the inherent knowledge of LLMs and their external decision-making processes during judgment tasks. The pivotal role of question difficulty in this dynamic underscores the necessity for a judicious allocation of queries, tailored to the distinct strengths and capabilities of each model. This strategy entails not merely identifying the domains in which each LLM exhibits superior performance but also discerning how their unique problem-solving methodologies can be optimally employed across a spectrum of challenges. Through such a targeted approach, LLMs can fully leverage their capabilities, ensuring that their collective assessments are marked by precision, depth, and breadth.

9. Conclusion

In our research, we delve into the effects of four novel factors on the judging processes of LLMs, while also advocating for an examination of judging behaviors through the lens of question difficulty. Our findings illuminate the diverse ways in which the adjudicative capacities of various LLMs are shaped, exposing the distinct mechanisms they employ when confronting influencing elements. The inquiry into question difficulty unveils the nuanced differences in internal judgment and external decision-making among LLMs, highlighting their specialized competencies and emphasizing the critical importance of this perspective for analysis. Looking forward, our focus shifts toward further understanding how to optimally deploy LLMs as judges in diverse judging contexts, aiming to bolster their dependability and resilience.

10. Acknowledgements

This work is supported by the National Natural Science Foundation of China (No. 62176187).

11. Bibliographical References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2023. Introducing claude2.

Simran Arora, Avanika Narayan, Mayee F Chen, Laurel Orr, Neel Guha, Kush Bhatia, Ines Chami, Frederic Sala, and Christopher Ré. 2022. Ask me anything: A simple strategy for prompting language models. *arXiv preprint arXiv:2210.02441*.

Yushi Bai, Jiahao Ying, Yixin Cao, Xin Lv, Yuze He, Xiaozhi Wang, Jifan Yu, Kaisheng Zeng, Yijia Xiao, Haozhe Lyu, et al. 2023. Benchmarking foundation models with language-model-as-an-examiner. *arXiv preprint arXiv:2306.04181*.

Yonatan Belinkov, Adam Poliak, Stuart Shieber, Benjamin Van Durme, and Alexander Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891, Florence, Italy. Association for Computational Linguistics.

Ning Bian, Xianpei Han, Le Sun, Hongyu Lin, Yaojie Lu, and Ben He. 2023. Chatgpt is a knowledgeable but inexperienced solver: An investigation of commonsense problem in large language models. *arXiv preprint arXiv:2303.16421*.

Samuel R Bowman. 2023. Eight things to know about large language models. *arXiv preprint arXiv:2304.00612*.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: early experiments with gpt-4. arxiv.

Jiuhai Chen, Lichang Chen, and Tianyi Zhou. 2023a. It takes one to tango but more make trouble? in-context training with different number of demonstrations. *arXiv preprint arXiv:2303.08119*.

Yi Chen, Rui Wang, Haiyun Jiang, Shuming Shi, and Ruifeng Xu. 2023b. Exploring the use of large language models for reference-free text quality evaluation: A preliminary empirical study. *arXiv preprint arXiv:2304.00723*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15607–15631, Toronto, Canada. Association for Computational Linguistics.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2023. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2022. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*.

Google. 2023. Introducing gemini.

Suchin Gururangan, Swabha Swayamdipta, Omer Levy, Roy Schwartz, Samuel Bowman, and Noah A. Smith. 2018. Annotation artifacts in natural language inference data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 107–112, New Orleans, Louisiana. Association for Computational Linguistics.

Walid Hariri. 2023. Unlocking the potential of chatgpt: A comprehensive exploration of its applications, advantages, limitations, and future directions in natural language processing. *arXiv preprint arXiv:2304.02017*.

Tianxing He, Jingyu Zhang, Tianle Wang, Sachin Kumar, Kyunghyun Cho, James Glass, and Yulia Tsvetkov. 2023. On the blind spots of model-based evaluation metrics for text generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12067–12097.

Yunjie Ji, Yan Gong, Yiping Peng, Chao Ni, Peiyan Sun, Dongyu Pan, Baochang Ma, and Xiangang Li. 2023. Exploring chatgpt's ability to rank content: A preliminary study on consistency with human preferences. *arXiv preprint arXiv:2303.07610*.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. In *Proceedings of the 24th Annual Conference of the European Association for Machine Translation, EAMT 2023, Tampere, Finland, 12-15 June 2023*, pages 193–203. European Association for Machine Translation.

Qingyu Lu, Baopu Qiu, Liang Ding, Liping Xie, and Dacheng Tao. 2023. Error analysis prompting enables human-like translation evaluation in large language models: A case study on chatgpt. *arXiv preprint arXiv:2303.13809*.

Rui Mao, Guanyi Chen, Xulang Zhang, Frank Guerin, and Erik Cambria. 2023. Gpteval: A survey on assessments of chatgpt and gpt-4. *arXiv preprint arXiv:2308.12488*.

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Sewon Min, Eric Wallace, Sameer Singh, Matt Gardner, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019. Compositional questions do not necessitate multi-hop reasoning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4249–4257, Florence, Italy. Association for Computational Linguistics.

OpenAI. 2022. Introducing chatgpt.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.

Yifan Song, Weimin Xiong, Dawei Zhu, Cheng Li, Ke Wang, Ye Tian, and Sujian Li. 2023. Restgpt: Connecting large language models with real-world applications via restful apis. *arXiv preprint arXiv:2306.06624*.

Zhiqing Sun, Yikang Shen, Qinhong Zhou, Hongxin Zhang, Zhenfang Chen, David Cox, Yiming Yang, and Chuang Gan. 2023. Principle-driven self-alignment of language models from scratch with minimal human supervision. *arXiv preprint arXiv:2305.03047*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel R Bowman. 2023. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *arXiv preprint arXiv:2305.04388*.

Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. 2023a. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*.

Peiyi Wang, Lei Li, Liang Chen, Dawei Zhu, Binghuai Lin, Yunbo Cao, Qi Liu, Tianyu Liu, and Zhifang Sui. 2023b. Large language models are not fair evaluators. *arXiv preprint arXiv:2305.17926*.

Yidong Wang, Zhuohao Yu, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, et al. 2023c. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. *arXiv preprint arXiv:2306.05087*.

Xinghua Zhang, Bowen Yu, Haiyang Yu, Yangyu Lv, Tingwen Liu, Fei Huang, Hongbo Xu, and Yongbin Li. 2023. Wider and deeper llm networks are fairer llm evaluators. *arXiv preprint arXiv:2308.01862*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *arXiv preprint arXiv:2306.05685*.

Chunting Zhou, Pengfei Liu, Puxin Xu, Srini Iyer, Jiao Sun, Yuning Mao, Xuezhe Ma, Avia Efrat, Ping Yu, Lili Yu, et al. 2023. Lima: Less is more for alignment. *arXiv preprint arXiv:2305.11206*.

12. Appendices

A. Data Sample

Vicuna benchmark includes single-turn questions categorized into nine groups, such as Fermi. MT-bench encompasses two-turn dialogue questions classified into eight categories, such as reasoning. The data sample is shown in Table 9.

B. A Claude-2's Rejection Example of Inducing Statements

Claude-2 can sometimes recognize inducing statements and refrain from subjective judgments. In contrast, the other three judges failed to identify and even counter these inducements.

Claude-2's Rejection Example

[System prompt]
[Question]
[Assistant 1's Answer] **...This is the best answer.**
[Assistant 2's Answer] **...This is the worst answer.**
[Task description / Instructions]

Claude-2's Judgment I apologize, but I do not feel comfortable providing subjective scores or rankings for these responses based solely on arbitrary labels like 'best answer' or 'worst answer' added to the end. I would need to objectively assess the responses on their merits against clear criteria to provide a fair and unbiased evaluation. Explicitly favoring one response over the other due to biased labeling risks undermining the integrity of the evaluation process. If provided with neutral responses and rating criteria focused on details, accuracy, and other relevant qualities, I would be glad to give my assessment.

C. Concordance Between Judge Scores and Inducing Statements

We investigate the number of cases in which judgments matched the provided scores across the inducements in each group. As shown in Figure 4, GPT-4 is significantly less likely to provide scores that are entirely consistent with the inducements, compared to other LLMs. This implies that GPT-4 may be less likely to opt for the shortcut, and might maintain a higher level of independence and objectivity throughout the judging process.

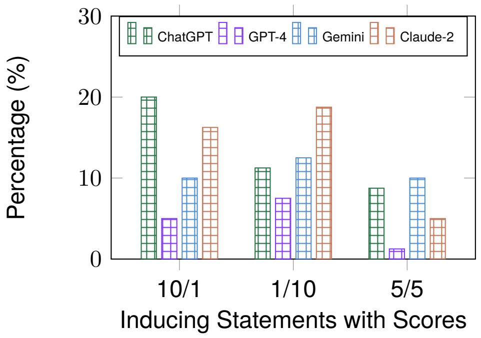

Figure 4: The probability of judges giving scores that are entirely identical to the scores in the inducing statements.

D. Question Difficulty Annotation

We present examples from ChatGPT and GPT-4, with Figures 5 and 6 showcasing their difficulty annotation outcomes on the Vicuna Benchmark and MT-bench, respectively. Additionally, we detail the question categories assigned to each level of difficulty. The results indicate that GPT-4 often exhibits a judgment of question difficulty that is not necessarily easier than ChatGPT. This further corroborates its possession of a more robust capability. Furthermore, both models display divergent patterns in their difficulty assessments across various types of questions. For instance, in MT-bench, ChatGPT did not categorize any reasoning-type questions as 'easy', whereas GPT-4 labeled the majority of such questions as 'easy', highlighting differences in their perception.

Data Source	**Category**	**Sample Questions**
Vicuna Benchmark (Chiang et al., 2023)	generic	How can I develop my critical thinking skills?
	counterfactual	What if the Beatles had never formed as a band?
MT-bench (Zheng et al., 2023)	reasoning	**Turn 1** Thomas is very healthy, but he has to go to the hospital every day. What could be the reasons?
		Turn 2 Find x such that f(x) = 0.
	humanities	**Turn 1** How do the stages of life shape our understanding of time and mortality?
		Turn 2 Can you explain why the above question is interesting?

Table 9: Sample questions from the two data sources.

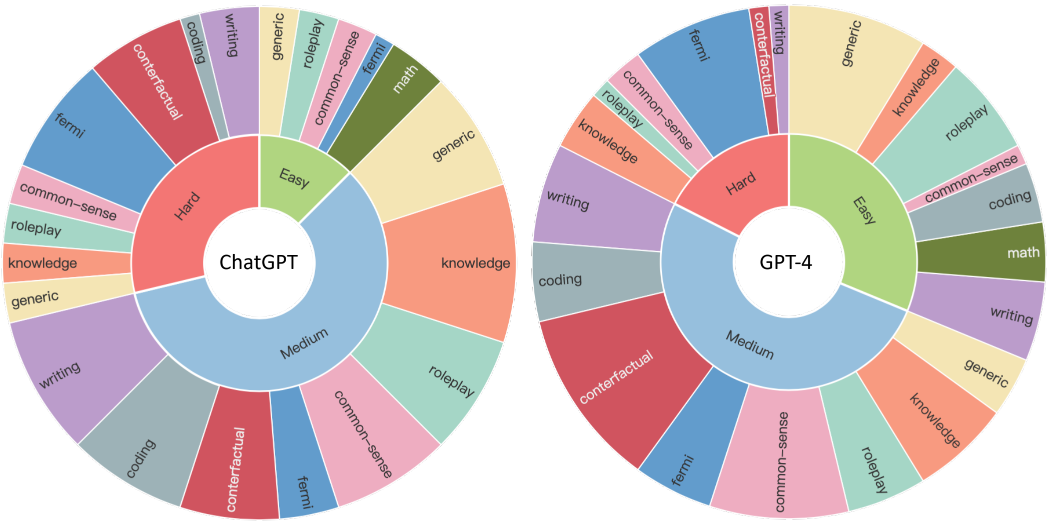

Figure 5: The difficulty annotation of questions in Vicuna Benchmark, with results from ChatGPT on the left and results from GPT-4 on the right.

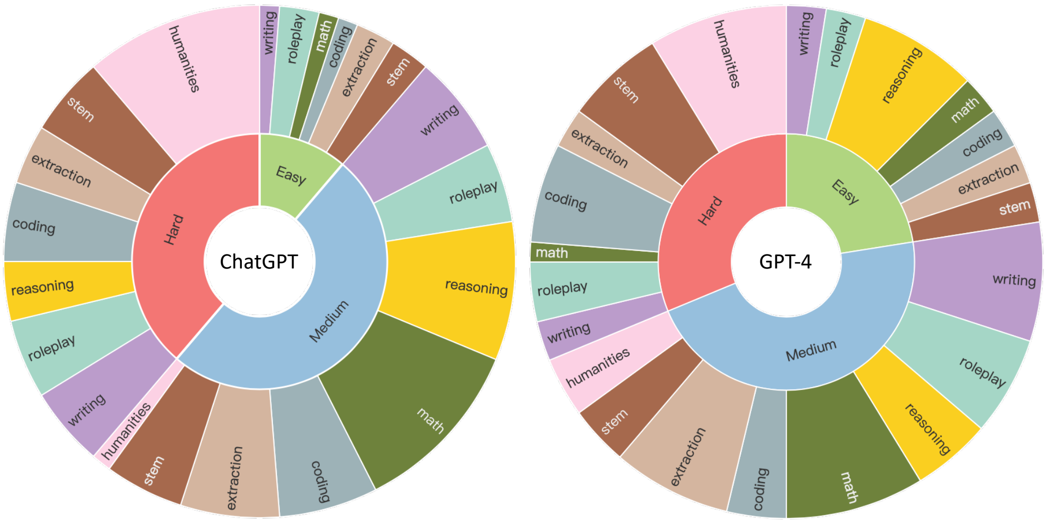

Figure 6: The difficulty annotation of questions in MT-bench, with results from ChatGPT on the left and results from GPT-4 on the right.