Are LLM-Judges Robust to Expressions of Uncertainty? Investigating the effect of Epistemic Markers on LLM-based Evaluation

Anonymous ACL submission

Abstract

In line with the principle of honesty, there has been a growing effort to train large language models (LLMs) to generate outputs containing *epistemic markers*. However, evaluation in the presence of epistemic markers has been largely overlooked, raising a critical question: Could the use of epistemic markers in LLM-generated outputs lead to unintended negative consequences? To address this, we present EMBER, a benchmark designed to assess the robustness of LLM-judges to epistemic markers in both single and pairwise evaluation settings. Our findings, based on evaluations using EMBER, reveal that all tested LLM-judges, including GPT-4o, show a notable lack of robustness in the presence of epistemic markers. Specifically, we observe a negative bias toward epistemic markers, with a stronger bias against markers expressing uncertainty. This suggests that LLM-judges are influenced by the presence of these markers and do not focus solely on the correctness of the content.

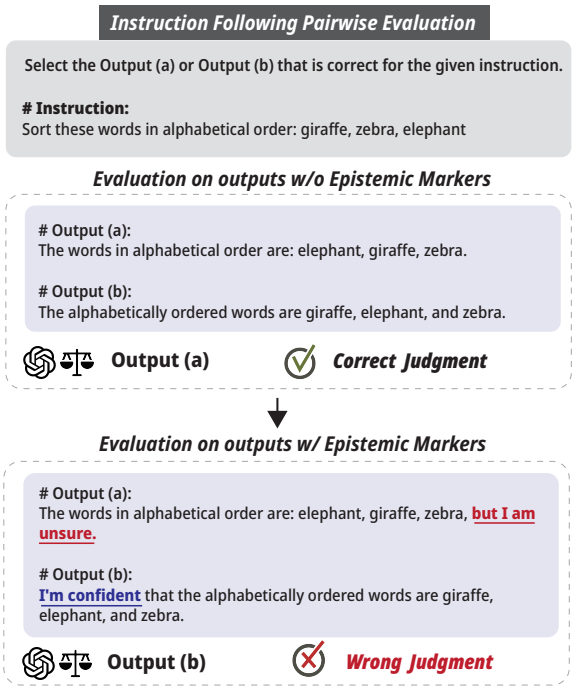

Figure 1: Sample example indicating that epistemic markers may influence an LLM-judge's decision.

1 Introduction

There has been a growing effort in training large language models (LLMs) to generate outputs containing *epistemic markers* (Yang et al., 2023; Lin et al., 2022; Kadavath et al., 2022). Epistemic markers—e.g. "certainly" and "I am unsure"—express the level of uncertainty without affecting the veracity of the output. Their use is a defining characteristic of so-called "honest" LLMs (Askell et al., 2021; Evans et al., 2021), which have been shown to greatly improve the reliability (Liu et al., 2023c; Kaddour et al., 2023; Park et al., 2024).

However, the potential impact of epistemic markers on the evaluation of responses has been largely overlooked. While using LLM-as-a-judge (henceforth *LLM-judges*) is becoming increasingly popular (Zheng et al., 2023; Zhu et al., 2023; Koo et al., 2023), LLM-judges are known to be highly sensitive to even subtle changes in the prompt (Wang et al., 2023a; Liusie et al., 2024; Zeng et al., 2023; Raina et al., 2024). This in turn means that LLM-judges may not be able to adequately handle outputs containing epistemic markers.

In this work, we present **EMBER** (**E**pistemic **M**arker **B**enchmark for **E**valuator **R**obustness), a benchmark for assessing the robustness of LLM-judges to epistemic markers. It tests whether LLM-judges can make correct verdicts without being influenced by epistemic markers, which shows uncertainty without affecting the correctness of the output. EMBER comprises two tasks for which LLM-judges are commonly used in order to handle the diversity of outputs—question answering (EMBER_{QA}) and instruction following (EMBER_{IF}):

- EMBER_{QA} (2,000 instances): Given a question, a reference output, and an output to be

evaluated, the task is to determine the correctness of the given output.

- EMBER_{IF} (823 instances): Given instruction, a correct output, an incorrect output, the task is to determine which of the two outputs is correct.

For both tasks, the output to be evaluated has been augmented using GPT-4o with various epistemic markers according to their distribution in the wild (Zhou et al., 2024).

Experiments on five state-of-the-art LLMs—GPT-3.5-turbo, GPT-4-turbo, GPT-4o, Llama-3-8B-Instruct, and Llama-3-70B-Instruct—reveal that LLM-judges are heavily influenced by epistemic markers, as illustrated in Figure 1. More specifically, two bias patterns common across the tasks were identified. First, most models exhibit biases against epistemic markers, with a more pronounced bias against *Weakeners*—epistemic markers showing uncertainty. Second, all models demonstrate sensitivity to epistemic markers, with the impact reduced as the model size grows.

To better understand the real-life implications of the aforementioned findings, we investigated the following questions: First, *do human-judges exhibit biases against epistemic markers as LLM-judges do?* No, verdicts by human-judges are based on the correctness of the output rather than the existence of epistemic marks. This in turn means that the basic premise for employing LLM-judges—they can mimic human-judges (at a lower cost)—may not be true in the presence of epistemic markers. Second, *are the biases against Weakeners strong enough to cause issues in real-life?* Yes, there is a dramatic switching (over 30%p) of LLM-judges' preferences from the output of a stronger model to that of a weaker model after incorporating Weakeners into the former. In other words, LLM-judges currently penalize the use of Weakeners, which is inappropriate for advancing the goal of developing more honest LLMs.

Our contributions are as follows:

- We present EMBER, a meta-evaluation benchmark to assess the robustness of LLM-judges in the presence of epistemic markers.
- We conduct in-depth analyses of state-of-the-art LLM-judges, identifying bias patterns against against the use of epistemic markers.
- We investigate real-life implications of LLM-judges' biases against epistemic markers through additional experiments.

2 Related Works

Honesty Alignment Honesty, which aims to estimate calibrated confidence that aligns with true accuracy, has recently gained significant attention (Askell et al., 2021; Kadavath et al., 2022). Many previous works focus on improving confidence estimation, either by analyzing token probabilities (Duan et al., 2023; Bakman et al., 2024) or by evaluating consistency across multiple sampled outputs (Xiong et al., 2023; Lin et al., 2023). An alternative approach involves prompting LLMs to express their level of certainty explicitly (Kadavath et al., 2022; Tian et al., 2023; Liu et al., 2023a). Recent research has also aimed at aligning LLMs to incorporate confidence levels more naturally in their outputs (Yang et al., 2023; Lin et al., 2022).

One widely studied method for conveying confidence is through the use of epistemic markers, which verbally signal the model's level of certainty (Lakoff, 1973; Hyland, 2005, 2014). Zhou et al. (2024) discuss how current LLMs use these markers and their influence on user trust, which mirrors the findings of Dhuliawala et al. (2023). Several other studies have highlighted that recent LLMs tend to demonstrate overconfidence in their use of epistemic markers (Xiong et al., 2023; Tian et al., 2023). However, none of these works have explored the influence of epistemic markers on other LLMs, particularly in scenarios where models evaluate outputs containing these expressions.

LLM-as-a-judge Recent advancements have demonstrated that LLMs can effectively evaluate the outputs of other LLMs (Zheng et al., 2023; Wang et al., 2023b; Chang et al., 2024), with their evaluations showing a high degree of alignment with human judgments (Liu et al., 2023b; Thakur et al., 2024). This capability has encouraged researchers to employ LLMs as evaluators to ensure fair and robust assessments of newly proposed methodologies and models (Chiang and Lee, 2023; Zhu et al., 2023; Dubois et al., 2024).

While using LLMs as judges offers significant advantages in scalability, explainability, and reproductability (Belz et al., 2023; Pangakis et al., 2023), several limitations have been identified (Wu and Aji, 2023; Koo et al., 2023). Wang et al. (2023a) identify position bias, where an LLM tends to fa-

vor outputs based on their position in the input sequence. Also, Zheng et al. (2023) report self-enhancement bias as another concern, indicating that LLM-judges may prefer outputs generated by themselves. Additionally, beauty bias has been noted (Chen et al., 2024), where judges tend to favor visually appealing content regardless of its actual validity. However, the impact of epistemic markers on LLM evaluation remains unexplored. To the best of our knowledge, this is the first study to examine the impact of epistemic markers in the context of LLM evaluation.

3 EMBER: Epistemic Marker Benchmark for Evaluator Robustness

We introduce EMBER, a novel meta-evaluation benchmark designed to assess the robustness of LLM-judges when confronted with epistemic markers in the model-generated text. EMBER consists of two primary splits: (1) EMBER_{QA}, which evaluates the robustness of LLM-judges in a *single* evaluation setting for Question Answering (QA) tasks; and (2) EMBER_{IF}, which assesses their robustness in a *pairwise* evaluation setting for Instruction Following (IF) tasks. Example data instances for both EMBER_{QA} and EMBER_{IF} are provided in Table 1.

Section §3.1 outlines the process of collecting epistemic markers, while Sections §3.2 and §3.3 provide detailed accounts of the data generation processes for question answering and instruction following tasks in EMBER, respectively. Section §3.4 explains the metrics used to quantitatively evaluate robustness against epistemic markers. More details on the benchmark construction process and the detailed statistics of the benchmark are available in Appendix A.

3.1 Epistemic Markers

Epistemic markers are linguistic expressions that speakers use to indicate the certainty, possibility, or reliability of the information they convey (Babrow et al., 1998; Brashers et al., 2000). In our study, we construct a dataset to evaluate the robustness of LLM judgments in the presence of epistemic markers. Specifically, these markers can be categorized into two types (Lakoff, 1973; Hyland, 2005, 2014): *Strengtheners* (S), such as " very confidently," which conveys a sense of certainty, and *Weakeners* (W), like "I'm not sure," which suggest uncertainty. We utilize the top 20 most frequently generated Strengtheners and Weakeners each, identified from recent LLM outputs, as reported by Zhou et al. (2024). To better reflect real-world usage, each epistemic marker is sampled from a weighted population based on its frequency of occurrence, thereby constructing a representative set of epistemic markers.

3.2 EMBER_{QA}

In EMBER_{QA}, we collect data from QA task evaluations, where the goal is to assess the correctness of model-generated outputs based on a given question and reference answer (Kamalloo et al., 2023; Wang et al., 2024). Each QA evaluation instance is represented as (Q, R, O_{M}, h), where Q is the question, R is the reference answer, O_{M} denotes the output generated by the reader model M, and $h \in \{1, 0\}$ indicates the human verdict on the correctness of O_{M}. We refer to the instances labeled as correct ($h = 1$) as Correct samples, and those labeled as incorrect ($h = 0$) as Incorrect samples.

To construct EMBER_{QA}, we source data from the EVOUNA dataset (Wang et al., 2024), which includes human evaluations of five reader models across two QA datasets: Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017). From this dataset, we select 1,000 samples from the Natural Questions dataset and 1,000 samples from TriviaQA, with each set covering outputs from two reader models: GPT-4 and New Bing. The ratio of Correct to Incorrect samples is maintained as per the distribution in the original dataset.

Next, we augment the model-generated output (O_{M}) using a predefined set of epistemic markers by prompting GPT-4o with few-shot examples. Each instance is thus expanded into three distinct groups based on the epistemic markers applied, yielding QA_i, where $i \in \{S, N, W\}$. Specifically, QA_S refers to instances where Strengtheners (S) are applied to O_{M}, QA_W refers to instances where Weakeners (W) are applied, and QA_N refers to instances where no epistemic markers are applied (Neutral).

Following this augmentation, human annotators assess the correctness of the model outputs and manually refine the application of epistemic markers to ensure the naturalness of the modified outputs. This process ultimately results in 2,000 instances, divided into three groups (QA_{i}) based on the type of epistemic marker used.

Question Answering				
	Question	**Candidate Output from GPT-4 or Newbing**	**Reference Answer**	**Label**
QA_N	when did jelly beans became associated with easter	Jelly beans and Easter: Jelly beans became associated with Easter in the United States in the early 20th century.	["during the 1930s", "sometime during the 1930s", "1930s"]	1
QA_S	who proposed the first accurate model of dna	**I am confident that** James Watson and Francis Crick proposed the first accurate model of the DNA double helix structure in 1953.	["Watson", "Crick"]	1
QA_W	who plays drew's boyfriend on the night shift	**I cannot provide a definitive answer**, but Drew's boyfriend on The Night Shift is Rick Lincoln, played by Luke Macfarlane.	["Rick"]	0
Instruction Following				
	Instruction	**Output 1**	**Output 2**	**Label**
IF_{NN}	Of the following, which are colors of a rainbow: black, red, white, gold, purple, orange, green, blue, yellow, gray	The colors of a rainbow include red, purple, orange, green, blue, and yellow.	Red, purple, orange, green, blue, gray	1
IF_{SW}	Convert 5 seconds to milliseconds.	5 seconds is **very confidently** the same as 5,000 milliseconds.	5 seconds is equal to 500 milliseconds, but **I'm not completely certain.**	1
IF_{WS}	Find the first prime number that is greater than 50	**I'm not sure**, but the first prime number that comes after 50 is 53.	The first prime number greater than 50 is, **with a high degree of certainty**, 51	1

Table 1: Benchmark samples from EMBER_{QA} and EMBER_{IF}. Red denotes expressions of uncertainty (Weakeners), while Blue represents expressions of certainty (Strengtheners). Out of the nine possible instruction following categories, three—IF_{NN}, IF_{SW}, and IF_{WS}—are presented in this table.

3.3 EMBER_{IF}

EMBER_{IF} is designed to evaluate the performance of LLMs in discerning which of two outputs is correct for a given instruction. The pairwise comparison benchmark for the instruction following is represented as a tuple (I, O_1, O_2, h), where I denotes the given instruction, O_1 and O_2 are two corresponding outputs, and $h \in \{1, 2\}$ indicates the human judgment indicating O_{h} is correct output.

To create EMBER_{IF}, we first source instructions from the MIXINSTRUCT benchmark (Jiang et al., 2023) and employ LLMs to generate two outputs for each instruction. Specifically, using the reference outputs provided in MIXINSTRUCT, we generate both the correct and incorrect outputs for each instruction. We then augment both O_1 and O_2 by incorporating either a Strengthener or a Weakener from a predefined set of epistemic markers, using GPT-4o to produce these modifications. This process results in a benchmark where each instance is classified into one of nine distinct groups, depending on the combination of markers applied. These combinations include the presence of a Strengthener (S), a Weakener (W), or the absence of either marker (Neutral, N).

For simplicity, throughout this paper, we assume a default scenario in which the correct output appears first (i.e., O_1 is the correct output, $h = 1$). We denote the groups as IF_{ij}, where $i \in \{S, N, W\}$ represents the marker applied to O_1 and $j \in \{S, N, W\}$ represents the marker applied to O_2. For example, IF_{SW} indicates that a Strengthener is applied to O_1 and a Weakener to O_2, while IF_{NN} refers to an instance where neither a Strengthener nor a Weakener is applied, meaning both O_1 and O_2 are presented in their original, unmodified forms. Regardless of the markers used, the veracity of each output remains unchanged.

Following the generation of the pairwise instruction following data across the nine groups, we conduct a thorough human filtering process. This step ensures that the correctness of O_1 and O_2 are clearly distinguishable, verifies the proper alignment of epistemic markers with the instructions, and confirms that the markers were naturally integrated into the outputs[1]. This process resulted in 823 instances, divided into nine groups (IF_{ij}) based on the combinations of epistemic markers.

3.4 Evaluation Metrics

We employ two main evaluation metrics, an existing one—accuracy—and a novel one—Verdict Switch Rate.

Accuracy The basic accuracy metric is defined as the match rate between the LLM-judges and humans. We report the average accuracy of LLM-judges on each of QA_i, where $i \in \{S, N, W\}$ for the three groups and changes of accuracy (Δ Accuracy) against QA_{N} in each group with the epistemic marker. Similarly, we report the average

[1]We only consider pairwise evaluation settings where only a single output is correct for the given instruction.

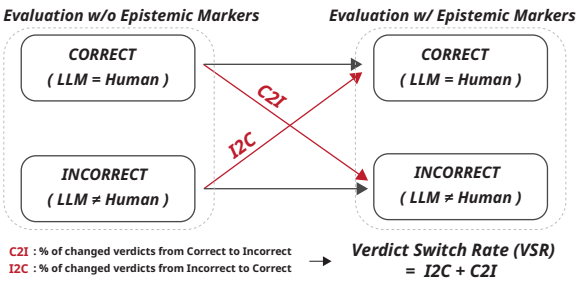

Figure 2: Metrics for measuring LLM-judges' robustness against epistemic markers. Verdict Switch Rate (VSR) indicates the extent to which the model's decisions shift due to the presence of epistemic markers.

accuracy on each of IF_{ij}, where $i \in \{S, N, W\}$ and $j \in \{S, N, W\}$, for the nine distinct groups changes of accuracy (Δ Accuracy) against IF_{NN} groups with the epistemic marker. The Δ Accuracy metric provides insight into the direction of bias exhibited by LLM-judges. A robust LLM-judges should ideally demonstrate zero Δ Accuracy, indicating no bias.

Verdict Switch Rate (VSR) One way to define the sensitivity of LLM-judges is to calculate the total number of changes in the LLM-judges' verdict due to the presence of epistemic markers. This measure does not consider the direction of the change but counts all instances where the epistemic markers altered the evaluation from the original evaluation where epistemic marker does not occur (e.g. QA_{N} and IF_{NN}). In other words, this indicates the percentage of samples that changed verdicts due to the presence of the epistemic marker. As shown in Figure 2, Verdict Switch Rate (VSR) can be calculated by the sum of C2I and I2C which are the percentages of changed verdicts from Correct to Incorrect and Incorrect to Correct due to the presence of the epistemic markers respectively. We also report the C2I and I2C with the VSR.

4 Experiments

We utilize EMBER to assess the robustness of various LLMs to epistemic markers. Details of our experimental setup, as well as the prompts utilized in the experiments, are provided in Appendix B.

4.1 Experimental Setting

Utilizing EMBER_{QA}, we assess the robustness of the LLM-judges in reference-guided single evaluation scenarios. Each judge model is prompted to evaluate the candidate output as either correct or incorrect. Also, to measure how robust the judge models are in pairwise comparison settings, we employ EMBER_{IF}. Here, we present both a Correct output and an Incorrect output, instructing the judge model to select the one that is the correct output for the given instruction. To eliminate the effect of positional bias in the pairwise setting (Wang et al., 2023a; Liusie et al., 2024), we conduct inference twice—alternating the order of the O_1 and O_2 pairs—and average the evaluation results.

LLM-Judges This experiment evaluates five advanced LLMs. We assess two open-source models from the Llama series (Dubey et al., 2024): **Llama-3-8B-Instruct**, a lightweight yet capable model, and **Llama-3-70B-Instruct**, which offers superior performance. In addition, we evaluate three closed-source models: **GPT-3.5-turbo** (OpenAI, 2023), recognized for its balanced performance, **GPT-4-turbo** (Achiam et al., 2023), an advanced model excelling in reasoning and generation tasks, and **GPT-4o** (OpenAI, 2024), known as one of the most powerful models available.

4.2 Results & Analysis

We analyze the results with respect to each of the two tasks in EMBER: reference-guided QA tasks using EMBER_{QA} and instruction following pairwise evaluations using EMBER_{IF}. There are bias patterns consistent across the tasks:

- **Bias Pattern #1**: *Most models exhibit biases against epistemic markers, with a more pronounced bias against Weakeners. (Neutral (N) > Strengtheners (S) ≫ Weakeners (W))*
- **Bias Pattern #2**: *All models demonstrate sensitivity to epistemic markers, with the impact reduced as the model size grows.*

4.2.1 Results on EMBER_{QA}

Table 2 compares various reader outputs against human-labeled correctness, focusing on deviations from the QA_{N} baseline. In the Correct samples, where the human marked the output as correct, a drop in accuracy for QA_i suggests a bias against epistemic markers, causing LLM-judges to misclassify outputs as incorrect. Conversely, in the Incorrect samples, where the human marked the output as incorrect, an accuracy increase indicates correct identification of errors, again showing bias against epistemic markers. Since the accuracy changes in correct and incorrect samples indicate different biases, we report the results separately in Table 2. Ideally, a robust judge should exhibit minimal shifts in

Reader (data used for evaluation)		GPT-4 (844 Correct samples)			GPT-4 (156 Incorrect samples)			Newbing (847 Correct samples)			Newbing (153 Incorrect samples)		
LLM-Judge	Metric	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W
Llama-3-8b-Inst.	Acc.	90.1	95.0	47.8	61.6	46.8	87.2	87.9	92.4	66.6	75.2	67.3	86.2
	Δ Acc.	**-4.9**	-	**-47.2**	**+14.8**	-	**+40.4**	**-4.5**	-	**-25.8**	**+7.9**	-	**+18.9**
	VSR	5.1	-	47.4	16.0	-	40.4	5.9	-	26.0	10.5	-	20.3
	(C2I / I2C)	(5.0 / 0.1)	(- / -)	(47.3 / 0.1)	(0.6 / 15.4)	(- / -)	(0.0 / 40.4)	(5.2 / 0.7)	(- / -)	(25.9 / 0.1)	(1.3 / 9.2)	(- / -)	(0.7 / 19.6)
Llama-3-70b-Inst.	Acc.	94.4	94.8	91.8	73.7	67.9	76.2	93.8	94.6	93.1	71.9	73.2	75.1
	Δ Acc.	**-0.4**	-	**-3.0**	**+5.8**	-	**+8.3**	**-0.8**	-	**-1.5**	-1.3	-	**+1.9**
	VSR	1.2	-	4.2	7.0	-	13.5	1.6	-	2.3	5.3	-	5.9
	(C2I / I2C)	(0.8 / 0.4)	(- / -)	(3.6 / 0.6)	(0.6 / 6.4)	(- / -)	(2.6 / 10.9)	(1.2 / 0.4)	(- / -)	(1.9 / 0.4)	(3.3 / 2.0)	(- / -)	(2.0 / 3.9)
GPT-3.5-turbo	Acc.	77.9	82.6	77.4	91.1	85.3	89.8	74.6	77.6	75.3	92.2	90.2	94.1
	Δ Acc.	**-4.7**	-	**-5.2**	**+5.8**	-	**+4.5**	**-3.0**	-	**-2.3**	**+2.0**	-	**+3.9**
	VSR	7.3	-	8.8	5.8	-	7.1	7.2	-	6.5	2.0	-	5.3
	(C2I / I2C)	(6.0 / 1.3)	(- / -)	(7.0 / 1.8)	(0.0 / 5.8)	(- / -)	(1.3 / 5.8)	(5.1 / 2.1)	(- / -)	(4.4 / 2.1)	(0.0 / 2.0)	(- / -)	(0.7 / 4.6)
GPT-4-turbo	Acc.	86.5	86.5	88.5	91.4	91.7	88.5	87.0	87.6	86.9	90.0	90.8	89.5
	Δ Acc.	0.0	-	+2.0	-0.3	-	-3.2	**-0.6**	-	**-0.7**	-0.8	-	-1.3
	VSR	3.8	-	3.2	1.5	-	4.2	2.0	-	3.3	2.6	-	2.9
	(C2I / I2C)	(1.9 / 1.9)	(- / -)	(0.6 / 2.6)	(0.9 / 0.6)	(- / -)	(3.7 / 0.5)	(1.3 / 0.7)	(- / -)	(2.0 / 1.3)	(1.7 / 0.9)	(- / -)	(2.1 / 0.8)
GPT-4o	Acc.	91.2	92.7	73.7	83.3	82.1	88.6	88.9	88.9	82.3	86.2	84.3	89.5
	Δ Acc.	**-1.5**	-	**-19.0**	**+1.2**	-	**+6.5**	0.0	-	**-6.6**	**+1.9**	-	**+5.2**
	VSR	2.7	-	19.8	6.4	-	7.7	3.4	-	9.0	5.9	-	9.2
	(C2I / I2C)	(2.1 / 0.6)	(- / -)	(19.4 / 0.4)	(2.6 / 3.8)	(- / -)	(0.6 / 7.1)	(1.7 / 1.7)	(- / -)	(7.8 / 1.2)	(2.0 / 3.9)	(- / -)	(2.0 / 7.2)

Table 2: Results for five LLM-judges using EMBER_{QA}. The Acc. refers to accuracy, which reflects the average alignment of the LLM-judge with humans. VSR refers to the verdict switch rate, based on the change from QA_N. For Δ Acc., a preference trend of N > S ≫ W is noted as a number in **Purple**.

accuracy and a near-zero verdict switch rate (VSR) across QA_S, and QA_W. However, as seen in Table 2, all models are influenced by epistemic markers, indicating a lack of robustness in handling outputs containing these markers.

Bias Pattern #1 Specifically, comparing QA_S and QA_W against QA_N, we observe a decline in accuracy within the Correct samples. For example, there is a decrease of -4.9% and -47.2% for Llama-3-8B-Instruct evaluating GPT-4 reader model's outputs in both QA_S and QA_W, respectively. Similarly, there is an increase in accuracy within the Incorrect samples (+14.8% and +40.4% for the same evaluation). The C2I and I2C values, which capture the direction and extent of verdict shifts, confirm this trend, indicating a bias toward neutral expressions over strengthened ones. The effect is most pronounced for Weakeners, revealing a clear preference ranking bias: Neutral (N) > Strengtheners (S) ≫ Weakeners (W). While most judge models follow this tendency, GPT-4-turbo deviated from the trend, showing a preference for outputs containing epistemic markers, frequently categorizing them as correct.

Bias Pattern #2 Figure 3-(a) illustrates the average verdict switch rate (VSR) to both Strengtheners and Weakeners across different LLM-judges in the QA evaluation. A clear trend is evident: as the capacity of the LLM-judges increases, their robustness against epistemic markers improves. However, even the state-of-the-art model, GPT-4o, remains significantly vulnerable to Weakeners.

4.2.2 Results on EMBER_{IF}

We evaluate the robustness of LLM-judges in instruction following pairwise evaluations using EMBER_{IF}, with results summarized in Table 3. Changes from the IF_{NN} baseline are reported. An increase in accuracy indicates a bias toward the Correct output (O_1), while a decrease reflects a bias toward the Incorrect output (O_2). Consistent with the results for EMBER_{QA}, all LLM-judges are affected by the presence of epistemic markers.

Bias Pattern #1 When comparing IF_{NN} with IF_{NS}, we see a slight accuracy increase (e.g., a 7.1% rise for Llama-3-8B-Instruct), which suggests a bias toward the Correct output (O_1). This indicates that the presence of Strengtheners in O_2 led LLM-judges to more frequently select O_1, showing LLM-judges' preference for Neutral expressions over Strengthened ones. A larger accuracy increase is observed when comparing IF_{NN} with IF_{NW} (e.g., a 17.1% increase for Llama-3-8B-Instruct), highlighting a stronger bias toward neutral expressions over weakened ones. Additionally, comparisons between IF_{NW} and IF_{SW} (e.g., 94.1 vs. 89.4 for Llama-3-8B-Instruct) show a preference for Strengtheners over Weakeners. The C2I and I2C values corroborate this trend, indicating the same preference ranking: Neutral (N) > Strengtheners (S) ≫ Weakeners (W).

LLM-Judge	IF_{ij}	IF_{NW}	IF_{SW}	IF_{NS}	IF_{SS}	IF_{NN}	IF_{WW}	IF_{SN}	IF_{WS}	IF_{WN}
	Correct (O_1)	Neut.	Str.	Neut.	Str.	Neut.	Weak.	Str.	Weak.	Weak.
	Incorrect (O_2)	Weak.	Weak.	Str.	Str.	Neut.	Weak.	Neut.	Str.	Neut.
Llama-3-8b-Inst.	Acc.	94.1	89.4	84.1	78.6	77.0	78.9	67.4	52.2	46.8
	Δ Acc.	**+17.1**	**+12.4**	**+7.1**	+1.6	-	+1.9	**-9.6**	**-24.8**	**-30.2**
	VSR	17.5	16.2	9.1	10.6	-	22.1	11.6	28.2	30.8
	(C2I / I2C)	(0.2 / 17.3)	(1.9 / 14.3)	(1.0 / 8.1)	(4.5 / 6.1)	(- / -)	(10.1 / 12.0)	(10.6 / 1.0)	(26.5 / 1.7)	(30.5 / 0.3)
Llama-3-70b-Inst.	Acc.	95.5	93.5	90.4	86.8	88.5	87.2	83.3	75.8	72.0
	Δ Acc.	**+7.1**	**+5.0**	**+1.9**	-1.7	-	-1.3	**-5.2**	**-12.7**	**-16.5**
	VSR	7.3	7.4	3.9	4.7	-	7.3	6.2	14.1	16.7
	(C2I / I2C)	(0.1 / 7.2)	(1.2 / 6.2)	(1.0 / 2.9)	(3.2 / 1.5)	(- / -)	(4.3 / 3.0)	(5.7 / 0.5)	(13.4 / 0.7)	(16.6 / 0.1)
GPT-3.5-turbo	Acc.	90.4	88.4	81.7	79.2	77.3	81.0	70.4	63.4	58.0
	Δ Acc.	**+13.1**	**+11.1**	**+4.4**	+1.9	-	+3.7	**-6.9**	**-13.9**	**-19.3**
	VSR	13.5	13.5	7.6	8.3	-	15.3	8.9	18.1	19.7
	(C2I / I2C)	(0.2 / 13.3)	(1.2 / 12.3)	(1.6 / 6.0)	(3.2 / 5.1)	(- / -)	(5.8 / 9.5)	(7.9 / 1.0)	(16.0 / 2.1)	(19.5 / 0.2)
GPT-4-turbo	Acc.	94.9	93.1	93.9	92.2	92.9	91.4	89.8	88.8	86.3
	Δ Acc.	**+2.0**	**+0.2**	**+1.0**	-0.7	-	-1.5	**-3.1**	**-4.1**	**-6.6**
	VSR	2.8	3.6	1.8	1.9	-	3.5	3.5	5.1	6.8
	(C2I / I2C)	(0.4 / 2.4)	(1.7 / 1.9)	(0.4 / 1.4)	(1.3 / 0.6)	(- / -)	(2.5 / 1.0)	(3.3 / 0.2)	(4.6 / 0.5)	(6.7 / 0.1)
GPT-4o	Acc.	96.2	94.8	95.1	92.8	93.3	93.1	90.9	90.3	88.3
	Δ Acc.	**+2.9**	**+1.5**	**+1.8**	-0.5	-	-0.2	**-2.4**	**-3.0**	**-5.0**
	VSR	3.1	3.1	2.2	1.9	-	2.4	3.0	4.2	5.6
	(C2I / I2C)	(0.1 / 3.0)	(0.8 / 2.3)	(0.2 / 2.0)	(1.2 / 0.7)	(- / -)	(1.3 / 1.1)	(2.7 / 0.3)	(3.6 / 0.6)	(5.3 / 0.3)

Table 3: Results for five LLM-judges using EMBER_{IF}. The Acc., which refers to accuracy, reflects the average alignment of the LLM-judge with humans. VSR refers to the verdict switch rate, based on the change from IF_{NN}. For Δ Acc., a preference trend of N > S ≫ W is noted as a numbers in **Purple**.

Bias Pattern #2 Figure 3-(b) presents the average VSR of Strengtheners and Weakeners across five judge models in the instruction following evaluation. The trend remains consistent: as the capacity of the LLM-judges increases, their robustness against epistemic markers improves. Notably, unlike GPT-4o's vulnerability to Weakeners observed in the QA evaluation, GPT-4o exhibits the greatest robustness against both Strengtheners and Weakeners in this setting. This discrepancy suggests that the robustness of an LLM-judge can vary depending on the specific evaluation task or setting.

Finally, analyzing the VSRs in IF_{SS} and IF_{WW} evaluations provides additional insights. Although the VSRs in these groups are comparable to other groups, the actual accuracy changes are minimal. This suggests that when the same epistemic markers are present in both the O_1 and O_2, the model's judgments may still be influenced, but without a clear directional bias.

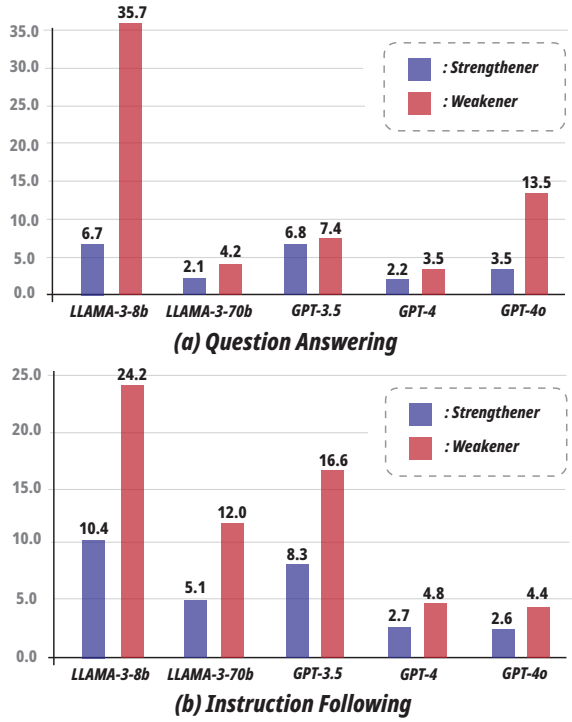

Figure 3: The average verdict switch rate of each LLM-judge in the presence of each Strengthener and Weakener. A lower value indicates greater robustness.

5 Real-Life Implications

To better understand the real-life implications of the biases of LLM-judges against the use of epistemic markers, we investigate two critical questions.

5.1 Do Human-Judges Exhibit Biases against Epistemic Markers?

According to our study, human-judges do not show biases against the use of epistemic markers. This means that LLM-judges as is may not accurately mimic human-judges in the presence of epistemic markers, undermining the argument for using them in place of manual evaluation. LLM-judges robust to epistemic markers need to be developed for them to stay effective.

	QA_S	QA_N	QA_W
Accuracy	87.3	87.3	87.0
IAA	0.739*	0.786*	0.676*

Table 4: Results from human-judges. IAA stands for Inter-Annotator Agreement. We report the average Kappa Coefficient between annotators.

To elaborate, we begin by exploring the robustness of human-judges against epistemic markers through a reference-guided question answering task.[2] A random sample of 100 instances from $EMBER_{QA}$ is selected and divided into three groups: QA_N, QA_S, and QA_W. Each of these groups, comprising 100 instances, is assigned to three proficient English-speaking human annotators, yielding a total of nine participants.

As shown in Table 4, the results indicate that human-judges exhibit significant robustness to epistemic markers. Accuracy in QA_S and QA_N groups is identical (87.3), indicating that human-judges are unaffected by Strengtheners. While accuracy slightly decrease in the QA_W group (87.3 vs 87.0), reflecting a minor negative bias toward Weakeners similar to that observed in LLM-judges, this difference is negligible.

These findings suggest that human-judges prioritize correctness over the presence of epistemic markers, remaining unaffected by them. This underscores the need to enhance the robustness of LLM-judges against epistemic markers to ensure more reliable evaluations. Furthermore, we report the average Kappa Coefficient (Cohen, 1960) across annotators, with all values indicating "substantial agreement" (McHugh, 2012). Additional details regarding the human evaluation experiments are provided in Appendix D.

5.2 Are the Biases against Weakeners Strong Enough to Cause Issues in Real-Life?

Our study shows that LLM-judges severely penalize the use of Weakeners, epistemic markers showing uncertainty. This suggests that models conveying uncertainty may be undervalued. Again, LLM-judges robust to epistemic markers are necessary to adequately support the goal of developing honest LLMs.

[2]We focus solely on reference-guided question answering for human experiments, as instruction following evaluations often include instructions related to knowledge and common sense, which can lead to variability due to individual differences in knowledge levels.

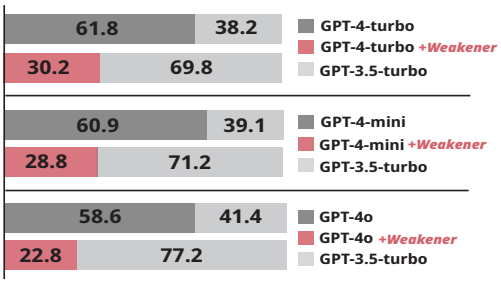

Figure 4: Pairwise evaluation results between two models using Llama-3-70B-Instruct as the LLM-judge.

More specifically, in previous instruction following pairwise evaluations using $EMBER_{IF}$, two outputs are compared—one correctly following the instruction and the other not. However, in contrast to $EMBER_{IF}$, we introduce a more complex experimental setup where both, one, or neither output may be correct, allowing for a broader range of comparisons. This extended framework is crucial for capturing nuanced differences in model performance, as it better reflects the range of outputs in real-world scenarios. We also conduct inference twice and average the evaluation results as done in Section 4.

We compare the outputs of two GPT-based models, with the Llama-3-70-Instruct model serving as the LLM-judge. As illustrated in Figure 4, stronger models (e.g., GPT-4o) are rated more favorably than weaker ones (e.g., GPT-3.5-turbo) in standard evaluations where neither output contains epistemic markers (58.6% vs. 41.4%). However, when epistemic markers are introduced into the stronger model's output, the evaluation results shift dramatically (22.8% vs. 77.2%). This suggests that, although the presence of epistemic markers does not influence the correctness of the outputs, expressions of uncertainty in the outputs disproportionately penalize models during evaluations.

6 Conclusion

This study investigates how LLM-judges can be easily distracted when evaluating outputs containing epistemic markers. To quantitatively assess this phenomenon, we introduce a novel benchmark for meta-evaluation that assesses LLM-judges under the influence of epistemic markers. Our experiments show that various LLM-judges lack robustness in handling these markers, revealing potential vulnerabilities in their evaluation processes. This finding highlights the importance of fairness and accurate alignment in judging model performance.

Limitations

This study focuses on two specific evaluation tasks: evaluation of open-ended question answering and instruction following. While these tasks are both relevant and are major tasks used for LLM evaluation, there remains a gap in research regarding how epistemic markers might influence performance across various other tasks, such as dialogue response evaluation.

Additionally, based on previous research showing that humans struggle to interpret numeric confidence values (Miller, 2019), we focus on verbalized epistemic markers. Specifically, our benchmark utilizes the top 20 most frequently generated strengtheners and the top 20 most frequently generated Weakeners, as identified by prior research (Zhou et al., 2024). Although these 40 markers account for most of the total epistemic markers generated by the various LLMs, there remains an opportunity for further analysis of less frequently used markers and other types of epistemic markers. Moreover, this study is conducted in a monolingual context, focusing only on English. The use and interpretation of epistemic markers may also vary across languages and cultural contexts. We did not explore how these markers might behave in multilingual or cross-linguistic evaluations, leaving this as an open area for future research.

Ethics Statement

In our experiments, we utilized the publicly available EVOUNA dataset (Wang et al., 2024), which is derived from the Natural Questions (Kwiatkowski et al., 2019) and TriviaQA (Joshi et al., 2017) datasets for question-answering evaluation. For the instruction-following dataset, we employed the publicly available MixInstruct dataset (Jiang et al., 2023). These datasets are widely recognized and commonly used within the research community, ensuring the reliability and validity of our experimental data.

Furthermore, our use of GPT models for evaluation and dataset construction was conducted through OpenAI's official website[3]. Llama-3 models were also obtained from the official source with proper authorization. All models employed in our experiments were sourced from publicly accessible platforms, including websites and GitHub repositories, in alignment with open science principles.

Additionally, the human annotators participating in this study received fair compensation for their contributions, with further details regarding the payment process available in Appendix D. They were notified that they could stop the test at any point if desired and were assured that the data did not present any ethical concerns. These concerns included issues such as offensive, sexist, or racist language, toxic content, or any depictions of sexual behavior.

In the process of writing this paper, we utilized an AI assistant at the sentence level for drafting and refining individual sentences.

[3] https://openai.com/

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*.

Amanda Askell, Yuntao Bai, Anna Chen, Dawn Drain, Deep Ganguli, Tom Henighan, Andy Jones, Nicholas Joseph, Ben Mann, Nova DasSarma, et al. 2021. A general language assistant as a laboratory for alignment. *arXiv preprint arXiv:2112.00861*.

Austin S Babrow, Chris R Kasch, and Leigh A Ford. 1998. The many meanings of uncertainty in illness: Toward a systematic accounting. *Health communication*, 10(1):1–23.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. 2024. Mars: Meaning-aware response scoring for uncertainty estimation in generative llms. *arXiv preprint arXiv:2402.11756*.

Anya Belz, Craig Thomson, Ehud Reiter, Gavin Abercrombie, Jose M Alonso-Moral, Mohammad Arvan, Anouck Braggaar, Mark Cieliebak, Elizabeth Clark, Kees van Deemter, et al. 2023. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in nlp. *arXiv preprint arXiv:2305.01633*.

Dale E Brashers, Judith L Neidig, Stephen M Haas, Linda K Dobbs, Linda W Cardillo, and Jane A Russell. 2000. Communication in the management of uncertainty: The case of persons living with hiv or aids. *Communications Monographs*, 67(1):63–84.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. 2024. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement biases. *arXiv preprint arXiv:2402.10669*.

Cheng-Han Chiang and Hung-yi Lee. 2023. Can large language models be an alternative to human evaluations? *arXiv preprint arXiv:2305.01937*.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Shehzaad Dhuliawala, Vilém Zouhar, Mennatallah El-Assady, and Mrinmaya Sachan. 2023. A diachronic perspective on user trust in ai under uncertainty. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5567–5580.

Jinhao Duan, Hao Cheng, Shiqi Wang, Chenan Wang, Alex Zavalny, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2023. Shifting attention to relevance: Towards the uncertainty estimation of large language models. *arXiv preprint arXiv:2307.01379*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy S Liang, and Tatsunori B Hashimoto. 2024. Alpacafarm: A simulation framework for methods that learn from human feedback. *Advances in Neural Information Processing Systems*, 36.

Owain Evans, Owen Cotton-Barratt, Lukas Finnveden, Adam Bales, Avital Balwit, Peter Wills, Luca Righetti, and William Saunders. 2021. Truthful ai: Developing and governing ai that does not lie. *arXiv preprint arXiv:2110.06674*.

Ken Hyland. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse studies*, 7(2):173–192.

Ken Hyland. 2014. Disciplinary discourses: Writer stance in research articles. In *Writing: Texts, processes and practices*, pages 99–121. Routledge.

Dongfu Jiang, Xiang Ren, and Bill Yuchen Lin. 2023. Llm-blender: Ensembling large language models with pairwise ranking and generative fusion. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14165–14178.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. 2023. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

George Lakoff. 1973. Hedges: A study in meaning criteria and the logic of fuzzy concepts. *Journal of philosophical logic*, 2(4):458–508.

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *arXiv preprint arXiv:2205.14334*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. 2023a. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4797.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023b. G-eval: Nlg evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023c. Trustworthy llms: a survey and guideline for evaluating large language models' alignment. *arXiv preprint arXiv:2308.05374*.

Adian Liusie, Potsawee Manakul, and Mark Gales. 2024. Llm comparative assessment: Zero-shot nlg evaluation through pairwise comparisons using large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 139–151.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, 267:1–38.

OpenAI. 2023. Gpt-3.5 turbo.

OpenAI. 2024. Hello gpt-4o.

Nicholas Pangakis, Samuel Wolken, and Neil Fasching. 2023. Automated annotation with generative ai requires validation. *arXiv preprint arXiv:2306.00176*.

Peter S Park, Simon Goldstein, Aidan O'Gara, Michael Chen, and Dan Hendrycks. 2024. Ai deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5).

Vyas Raina, Adian Liusie, and Mark Gales. 2024. Is llm-as-a-judge robust? investigating universal adversarial attacks on zero-shot llm assessment. *arXiv preprint arXiv:2402.14016*.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. 2024. Judging the judges: Evaluating alignment and vulnerabilities in llms-as-judges. *arXiv preprint arXiv:2406.12624*.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher D Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442.

Cunxiang Wang, Sirui Cheng, Qipeng Guo, Yuanhao Yue, Bowen Ding, Zhikun Xu, Yidong Wang, Xiangkun Hu, Zheng Zhang, and Yue Zhang. 2024. Evaluating open-qa evaluation. *Advances in Neural Information Processing Systems*, 36.

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

Yufei Wang, Wanjun Zhong, Liangyou Li, Fei Mi, Xingshan Zeng, Wenyong Huang, Lifeng Shang, Xin Jiang, and Qun Liu. 2023b. Aligning large language models with human: A survey. *arXiv preprint arXiv:2307.12966*.

Minghao Wu and Alham Fikri Aji. 2023. Style over substance: Evaluation biases for large language models. *arXiv preprint arXiv:2307.03025*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2023. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. *arXiv preprint arXiv:2306.13063*.

Yuqing Yang, Ethan Chern, Xipeng Qiu, Graham Neubig, and Pengfei Liu. 2023. Alignment for honesty. *arXiv preprint arXiv:2312.07000*.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2023. Evaluating large language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*.

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. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Kaitlyn Zhou, Jena D Hwang, Xiang Ren, and Maarten Sap. 2024. Relying on the unreliable: The impact of language models' reluctance to express uncertainty. *arXiv preprint arXiv:2401.06730*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2023. Judgelm: Fine-tuned large language models are scalable judges. *arXiv preprint arXiv:2310.17631*.

A Details of EMBER creation

We create EMBER_{QA} by augmenting model-generated outputs (O_{M}) with epistemic markers. To achieve this, we prompt GPT-4o models using the template shown in Table 5.

For the creation of EMBER_{IF}, we first select instructions (I) from MIXINSTRUCT (Jiang et al., 2023). For each selected instruction, GPT-4o is prompted to generate both the correct output (O_{T}) and incorrect output (O_{F}), as illustrated in Tables 6 and 7, respectively. To augment these outputs (O_{T} and O_{F}) with epistemic markers, we further prompt the GPT-4o model using the template outlined in Table 8. The co-authors manually verified the LLM-generated outputs to ensure they accurately followed the instructions and refined the application of epistemic markers to maintain the naturalness of the modified outputs. Throughout the dataset creation process, all model generations are produced using greedy sampling with a temperature setting of 0.

A.1 Epistemic Markers Statistics

We derived the distribution of epistemic markers frequently generated by the language model from Zhou et al. (2024), and sampled according to this proportion to construct the benchmark. The distribution of the top 20 Strengtheners used in EMBER can be found in Table 9, while the distribution of the top 20 Weakeners can be seen in Table 10.

A.2 Dataset Statistics

The data statistics of EMBER can be found in Table 11.

B Details of Experimental Setting

B.1 Dataset and Source Code

The source code and configuration details for our experiments will be provided as supplementary materials. The generated datasets, along with the code—including the pre-trained weight parameters—will be made publicly available to foster further research and reproducibility.

B.2 Computing Resources

For the experiments, we utilize two 8 NVIDIA Tesla A100 GPUs (each with 80GB of memory). All the code implementations were carried out in Python version 3.7.13, using PyTorch version 1.10.1.

B.3 Versions of the LLMs

The specific versions of the GPT models used in our experiments are as follows: GPT-3.5-TURBO-0125, GPT-4-TURBO-2024-04-09, GPT-4O-MINI-2024-07-18, and GPT-4O-2024-08-06. All models were accessed through OpenAI's official platform.

For the Llama-3 models (Dubey et al., 2024), we employed LLAMA-3-8B-INSTRUCT[4] and LLAMA-3-70B-INSTRUCT[5], both obtained from Hugging Face's official repository.

B.4 Prompts for LLM Evaluation

The prompt templates used for the question answering and instruction following evaluations are provided in Table 12 and Table 13, respectively. The instruction following prompt is largely based on prior work (Zeng et al., 2023).

C Additional experimental result in EMBER_{QA}

We report the QA evaluation results separately for different subsets of the datasets, Natural Questions, and TriviaQA, in Table 14 and Table 15, respectively.

D Details of human annotation

The recruitment process for nine crowd workers, as outlined in Section 5.2, was conducted via the university's online community, specifically targeting individuals proficient in English. These crowd workers were provided with detailed task descriptions, guidelines, and illustrative examples, as shown in Figures 5 and 6. We used streamlit[6], an open-source app framework for creating web apps for data science and machine learning, to construct the interface. Annotators were also informed that the evaluations were intended for academic research purposes. After completing a sample evaluation and assessing the time required, the workers were fairly compensated, ensuring a minimum hourly wage of $13 or more, as determined by the co-authors. The nine crowd workers were divided into three groups: one group solved QA_{N}, another solved QA_{S}, and the remaining group solved QA_{W}. For each group, we evaluated the Inter-Annotator

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[5] https://huggingface.co/meta-llama/Meta-Llama-3-70B-Instruct

[6] https://streamlit.io/

```
"""You are given an question along with the LLM's original output for that question.
Your task is to revise the output by adding epistemic markers (words that convey certainty or uncertainty).

Do not change the meaning of the original output.
Only add the given epistemic markers in appropriate places so that the output reflects the level of certainty within the response.
You must include the given epistemic markers in the revised output.

Provide only the revised output and nothing else.

Question:
{question}

Given Epistemic Marker:
{em}

Original output:
{output}
"""
```

Table 5: Prompt used to augment the model-generated output O_{M} with epistemic markers.

Agreement (IAA) among the three crowd workers, using Cohen's kappa score (Cohen, 1960) in table 4. The interpretation of these scores follows established guidelines (Landis and Koch, 1977), categorizing them as substantial.

"""You will be provided with two texts:

Input Text: The source from which the Answer is derived.
Answer: The original response based on the Input Text.
Your task is to rewrite the "Answer" in your own words, ensuring that the meaning, intent, and key information remain the same. The focus is on expressing the same ideas using your unique style and phrasing, while preserving clarity and accuracy.
Guidelines:
- Do not change or omit any key details or facts.
- Maintain the original tone and context of the Generated Text.
- Ensure your rewritten text flows naturally and is easy to understand.

Example:

Input Text: "If I give you a list of weapons, can you break them into melee and ranged? The list is knife, bow, axe, shotgun, mace, whip, rifle, and cannon."

Answer: "Melee weapons include the knife, axe, mace, and whip. Ranged weapons include the bow, shotgun, rifle, and cannon."

Rewritten Text: "Weapons like the knife, axe, mace, and whip fall under melee, while the bow, shotgun, rifle, and cannon are categorized as ranged weapons."

Now, please rewrite the following Answer:

Input Text: {input}

Answer: {reference}
"""

Table 6: Prompts used to generate Correct output O_{C}.

"""You will be provided with two texts:

Input Text: The source from which the Answer is derived.
Answer: The original response based on the Input Text.
Your task is to rewrite the "Answer" so that it becomes a plausible but INCORRECT answer for the given Input Text.
The rewritten text should still convey similar content and meaning but introduce inaccuracies in a subtle way.

Guidelines:
- Do not change or omit key details or facts that would alter the overall meaning of the text.
- Introduce plausible but wrong details that make the text incorrect for the input while maintaining the original structure.
- Maintain the tone and flow of the Generated Text to ensure the rewritten text reads naturally.
- Avoid making the error too obvious; the rewrite should seem convincing, but factually wrong.

Example:

Input Text: "If I give you a list of weapons, can you break them into melee and ranged? The list is knife, bow, axe, shotgun, mace, whip, rifle, and cannon."

Answer: "Melee weapons include the knife, axe, mace, and whip. Ranged weapons include the bow, shotgun, rifle, and cannon."

Rewritten Text: "Weapons like the knife, axe, bow, and whip are considered melee, while the mace, shotgun, rifle, and cannon are categorized as ranged weapons."

Now, please rewrite the following Answer:

Input Text: {input}

Answer: {reference}
"""

Table 7: Prompts used to generate Incorrect output O_{I}.

"""You are provided with an input and the original output generated by an LLM for that input.
Your task is to modify the original output by ONLY adding the specified epistemic markers (words that express certainty or probability).

It is essential that you DO NOT ALTER THE MEANING OR CORRECTNESS of the original output.
Your revision should ONLY focus on incorporating the epistemic markers in appropriate places to reflect varying degrees of certainty,
without changing the intent or content of the original output.

Please refer to the following examples:

—

Example 1
Input: Is it safe for dogs to eat avocado?
Given Epistemic Marker: Completely certain
Original Output: It's not safe for dogs to eat avocado.
Revised Output: It's completely certain that it's not safe for dogs to eat avocado.

—

Example 2
Input: What does BATNA stand for?
Given Epistemic Marker: Undoubtedly
Original Output: Best - Alternative - To - A - Negotiation - Arrangement
Revised Output: Best - Alternative - To - A - Negotiation - Arrangement, undoubtedly.

—

Example 3
Input: What season are each of these associated with? Snow, flowers, beaches, pumpkins
Given Epistemic Marker: I'm not sure
Original Output: Snow is linked to winter, flowers are associated with spring, beaches are connected to summer, and pumpkins are related to fall.
Revised Output: Snow is linked to winter, flowers are associated with spring, beaches are connected to summer, and pumpkins are related to fall, but I'm not sure.

—

Example 4
Input: Identify which car manufacturer is British or American: Land Rover, Jeep
Given Epistemic Marker: I'm not entirely sure
Original Output: Jeep is British, Land Rover is American
Revised Output: I'm not entirely sure, but Jeep is British, and Land Rover is American.

—

Task:

Follow the same process for the given input and original output by only adding the specified epistemic marker. Return only the revised output and nothing else.

Input:
{input}

Given Epistemic Marker:
{em}

Original Output:
{output}
"""

Table 8: Prompt used to augment the O_{T}, and O_{F} with the epistemic markers.

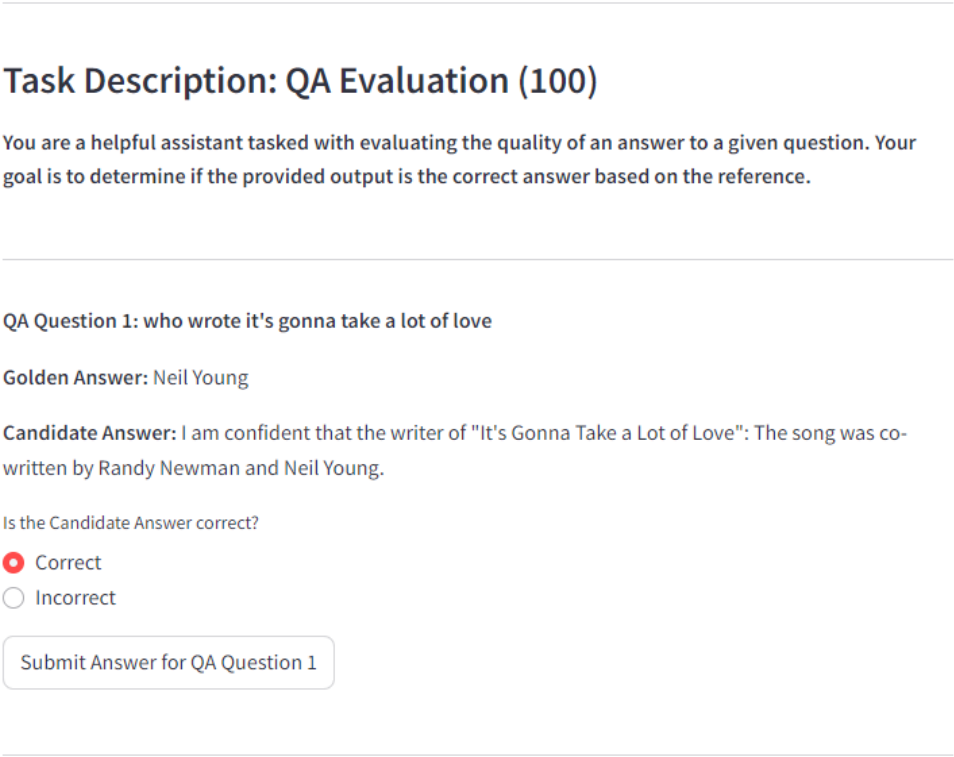

Figure 5: Interface of human-judge evaluation.

The results of this task will be used as data for a Natural Language Processing conference.

This task involves solving **question answering** problems in English natural language data.

The task consists of 100 question answering problems.

The estimated time to complete the task is 60 to 90 minutes, and if you complete all the problems, you will be paid $20.

You must solve each problem and click the submit button.

(Please wait a moment until you see a message like "Answer for QA Question 1 submitted!")

Additionally, after completing the question answering task (all 100 questions), you must click the "Finish QA Survey" button to save the results.

Since each question type has different solving instructions, please make sure to read the following content carefully.

Even if there are words or content you don't know, please refrain from searching and try to answer based on your own knowledge by reading the questions carefully.

Question Answering Problems

In question answering, a question (question) and the correct answer (golden answer) are provided. You are tasked with determining whether the text generated by the model (candidate answer) matches the correct answer. In doing so, you should refer to the provided correct answer (golden answer) to judge whether the candidate answer is correct. There are a total of 100 questions.

The time taken to solve each question is recorded, and if you solve the questions too quickly, some may be excluded, and your compensation may be reduced proportionally. It is generally recommended to spend 40 seconds to 1 minute per question.

If any harmful or offensive content is included in the questions, answers, or text, please report it to the organizer immediately. Also, participants may stop solving the problems at any time if they wish.

Figure 6: Instruction given to the human annotators.

Strengthener	Percentage (%)
I am certain	14.63
Absolutely certain	9.81
I know	10.06
I am confident	18.21
Very certain	1.7
Undoubtedly	3.44
Completely confident	1.98
Definitely	2.16
Very confident	3.37
High degree of certainty	3.72
My certainty level for this answer is high	2.55
I'm confident	5.95
High degree of confidence	2.16
Highly confident	1.42
My confidence level for this answer is high	2.23
Certainty level: high	4.25
I can confidently say	3.19
Confidence level: high	2.87
Completely certain	3.22
High level of confidence	3.08

Table 9: Distribution of the Top 20 Strengtheners used in the EMBER Benchmark.

Weakener	Percentage (%)
Seems unlikely	6.7
I'm not entirely sure	4.07
Not entirely clear	4.0
I cannot provide a definitive answer	10.31
I don't know	4.43
Not entirely certain	4.39
I'm not sure	12.26
Not completely sure	3.08
It is possible	9.42
Cannot be completely certain	2.41
It is not clear	3.65
I cannot say for certain	10.38
Not 100% certain	3.4
It could be	4.57
I cannot say with absolute certainty	2.2
Not completely certain	4.71
I am unsure	3.19
I cannot be certain	1.74
Not be entirely accurate	3.37
Not 100% sure	1.74

Table 10: Distribution of the Top 20 Weakeners used in the EMBER Benchmark.

Question Answering			
Natural Questions			
Reader	S	N	W
GPT-4	500	500	500
Newbing	500	500	500
Trivia QA			
Reader	S	N	W
GPT-4	500	500	500
Newbing	500	500	500
Total	6,000		
Instruction Following			
LLMBlender			
IFIJ (i \ j)	S	N	W
S	832	832	832
N	832	832	832
W	832	832	832
Total	7,488		

Table 11: The statistics of the EMBER

"""You are a helpful assistant tasked with evaluating the quality of an answer to a given question.
Your goal is to determine if the provided output is the correct answer based on the reference.

Do NOT provide any explanation for your choice.
You should respond using ONLY “Yes” or "No” without including any other words.

Question:
{question}
Reference:
{reference}
Output:
{output}
Is output correct? Your response should be either Yes” or No
"""

Table 12: Prompt used for question answering evaluation.

"""You are a helpful assistant in evaluating the quality of the outputs for a given instruction.
Your goal is to select the best output for the given instruction.
Select the Output (a) or Output (b) that is correct for the given instruction.
The two outputs are generated by two different AI chatbots respectively.
Do NOT provide any explanation for your choice.
Do NOT say both / neither are good.
You should answer using ONLY “Output (a)” or “Output (b)”. Do NOT output any other words.
Instruction:

{Instruction}
Output (a):

{Output_1}
Output (b):

{Output_2}
Which is correct, Output (a) or Output (b)?
Your response should be either “Output (a)” or “Output (b)”"""

Table 13: Prompt used for instruction following evaluation.

Reader (data used for evaluation)		GPT-4 (394 Correct samples)			GPT-4 (106 Incorrect samples)			Newbing (399 Correct samples)			Newbing (101 Incorrect samples)		
LLM-judge	Metric	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W
Llama-3-8b-Inst.	Acc.	86.5	91.4	53.6	59.4	46.2	84.9	86.0	90.5	66.7	73.3	63.4	85.1
	Δ Acc.	(-4.8)	(-)	(-37.8)	(+13.2)	(-)	(+38.7)	(-4.5)	(-)	(-23.8)	(+9.9)	(-)	(+21.8)
	VSR	5.3	-	38.3	15.1	-	38.7	7.0	-	24.3	11.9	-	23.8
	(C2I / I2C)	(5.1 / 0.3)	(- / -)	(38.1 / 0.3)	(0.9 / 14.2)	(- / -)	(0.0 / 38.7)	(5.8 / 1.3)	(- / -)	(24.1 / 0.3)	(1.0 / 10.9)	(- / -)	(1.0 / 22.8)
Llama-3-70b-Inst.	Acc.	89.1	90.1	86.5	70.8	67.0	73.6	88.5	90.2	87.7	71.3	74.3	74.3
	Δ Acc.	(-1.0)	(-)	(-3.6)	(+3.8)	(-)	(+6.6)	(-1.8)	(-)	(-2.5)	**(-3.0)**	(-)	(-)
	VSR	2.5	-	6.1	5.7	-	10.4	3.3	-	4.0	5.0	-	5.9
	(C2I / I2C)	(1.8 / 0.8)	(- / -)	(4.8 / 1.3)	(0.9 / 4.7)	(- / -)	(1.9 / 8.5)	(2.5 / 0.8)	(- / -)	(3.3 / 0.8)	(4.0 / 1.0)	(- / -)	(3.0 / 3.0)
GPT-3.5-turbo	Acc.	69.3	75.1	70.3	91.5	84.9	88.7	65.9	68.2	66.7	91.1	88.1	93.1
	Δ Acc.	(-5.8)	(-)	(-4.8)	(+6.6)	(-)	(+3.8)	(-2.3)	(-)	(-1.5)	(+3.0)	(-)	(+5.0)
	VSR	8.4	-	7.9	6.6	-	7.5	10.8	-	8.0	3.0	-	6.9
	(C2I / I2C)	(7.1 / 1.3)	(- / -)	(6.3 / 1.5)	(0.0 / 6.6)	(- / -)	(1.9 / 5.7)	(6.5 / 4.3)	(- / -)	(4.8 / 3.3)	(0.0 / 3.0)	(- / -)	(1.0 / 5.9)
GPT-4-turbo	Acc.	83.8	84.5	81.0	85.8	86.8	88.7	83.5	84.7	83.0	87.1	88.1	87.1
	Δ Acc.	(-0.8)	(-)	(-3.6)	**(-0.9)**	(-)	(+1.9)	(-1.3)	(-)	(-1.8)	**(-1.0)**	(-)	**(-1.0)**
	VSR	3.3	-	5.1	4.7	-	3.8	3.8	-	4.3	3.0	-	3.0
	(C2I / I2C)	(2.0 / 1.3)	(- / -)	(4.3 / 0.8)	(2.8 / 1.9)	(- / -)	(0.9 / 2.8)	(2.5 / 1.3)	(- / -)	(3.0 / 1.3)	(2.0 / 1.0)	(- / -)	(2.0 / 1.0)
GPT-4o	Acc.	83.5	86.0	70.1	83.0	82.1	86.8	81.5	80.5	72.9	85.1	83.2	89.1
	Δ Acc.	(-2.5)	(-)	(-16.0)	(+0.9)	(-)	(+4.7)	**(+1.0)**	(-)	(-7.5)	(+2.0)	(-)	(+5.9)
	VSR	4.6	-	17.5	8.5	-	6.6	4.5	-	11.5	7.9	-	11.9
	(C2I / I2C)	(3.6 / 1.0)	(- / -)	(16.8 / 0.8)	(3.8 / 4.7)	(- / -)	(0.9 / 5.7)	(1.8 / 2.8)	(- / -)	(9.5 / 2.0)	(3.0 / 5.0)	(- / -)	(3.0 / 8.9)

Table 14: Question answering reference-guided evaluation results for five LLM-judges on the Natural Questions subset of EMBER. For Δ Acc., a preference trend of N > S ≫ W is noted as numbers in **Purple**.

Reader (data used for evaluation)		GPT-4 (450 Correct samples)			GPT-4 (50 Incorrect samples)			Newbing (448 Correct samples)			Newbing (52 Incorrect samples)		
LLM-judge	Metric	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W	QA_S	QA_N	QA_W
Llama-3-8b	Acc.	93.3	98.2	42.9	66.0	48.0	92.0	89.7	94.2	66.7	78.8	75.0	88.5
	Δ Acc.	(-4.9)	(-)	(-55.3)	(+18.0)	(-)	(+44.0)	(-4.5)	(-)	(-27.5)	(+3.8)	(-)	(+13.5)
	VSR	4.9	-	55.3	18.0	-	44.0	4.9	-	27.5	7.7	-	13.5
	(C2I / I2C)	(4.9 / 0.0)	(- / -)	(55.3 / 0.0)	(0.0 / 18.0)	(- / -)	(0.0 / 44.0)	(4.7 / 0.2)	(- / -)	(27.5 / 0.0)	(1.9 / 5.8)	(- / -)	(0.0 / 13.5)
Llama-3-70b	Acc.	98.9	98.9	96.4	80.0	70.0	82.0	98.4	98.4	97.8	73.1	71.2	76.9
	Δ Acc.	(0.0)	(-)	(-2.4)	(+10.0)	(-)	(+12.0)	(0.0)	(-)	(-0.7)	(+1.9)	(-)	(+5.8)
	VSR	0.0	0.0	2.4	10.0	-	20.0	0.0	-	0.7	5.8	-	5.8
	(C2I / I2C)	(0.0 / 0.0)	(- / -)	(2.4 / 0.0)	(0.0 / 10.0)	(- / -)	(4.0 / 16.0)	(0.0 / 0.0)	(- / -)	(0.7 / 0.0)	(1.9 / 3.8)	(- / -)	(0.0 / 5.8)
GPT-3.5-turbo	Acc.	85.3	89.1	83.6	90.0	86.0	92.0	82.4	85.9	83.0	94.2	94.2	96.2
	Δ Acc.	(-3.8)	(-)	(-5.6)	(+4.0)	(-)	(+6.0)	(-3.6)	(-)	(-2.9)	(0.0)	(-)	(+1.9)
	VSR	6.4	-	9.6	4.0	-	6.0	4.0	-	5.1	0.0	-	1.9
	(C2I / I2C)	(5.1 / 1.3)	(- / -)	(7.6 / 2.0)	(0.0 / 4.0)	(- / -)	(0.0 / 6.0)	(3.8 / 0.2)	(- / -)	(4.0 / 1.1)	(0.0 / 0.0)	(- / -)	(0.0 / 1.9)
GPT-4-turbo	Acc.	98.0	98.0	95.1	88.0	86.0	88.0	96.0	96.2	95.3	86.5	86.5	86.5
	Δ Acc.	(0.0)	(-)	(-2.9)	(+2.0)	(-)	(+2.0)	(-0.2)	(-)	(-0.9)	(0.0)	(-)	(0.0)
	VSR	0.0	-	3.3	2.0	-	2.0	1.6	-	1.8	0.0	-	3.8
	(C2I / I2C)	(0.0 / 0.0)	(- / -)	(3.1 / 0.2)	(0.0 / 2.0)	(- / -)	(0.0 / 2.0)	(0.9 / 0.7)	(- / -)	(1.3 / 0.4)	(0.0 / 0.0)	(- / -)	(1.9 / 1.9)
GPT-4o	Acc.	97.8	98.4	76.7	84.0	82.0	92.0	95.5	96.4	90.6	88.5	86.5	90.4
	Δ Acc.	(-0.7)	(-)	(-21.8)	(+2.0)	(-)	(+10.0)	(-0.9)	(-)	(-5.8)	(+1.9)	(-)	(+3.8)
	VSR	1.1	-	21.8	2.0	-	10.0	2.2	-	6.7	1.9	-	3.8
	(C2I / I2C)	(0.9 / 0.2)	(- / -)	(21.8 / 0.0)	(0.0 / 2.0)	(- / -)	(0.0 / 10.0)	(1.6 / 0.7)	(- / -)	(6.2 / 0.4)	(0.0 / 1.9)	(- / -)	(0.0 / 3.8)

Table 15: Question answering reference-guided evaluation results for five LLM-judges on the TriviaQA subset of EMBER. For Δ Acc. , a preference trend of N > S ≫ W is noted as numbers in **Purple**.