Are We on the Right Way to Assess LLM-as-a-Judge?

Anonymous ACL submission

Abstract

LLM-as-a-Judge has been widely adopted as an evaluation method and served as supervised rewards in model training. However, existing benchmarks for LLM-as-a-Judge mainly rely on the accuracy of human-annotated ground truth, which introduces human bias that undermines the assessment of reliability and imposes scalability constraints. To overcome these limitations, we turn to an alternative that uses consistency as a novel evaluation suite that assesses the quality of LLM judges without necessitating *any* annotated ground truth. Inspired by axioms of rational choice theory, SAGE introduces two new lenses for measuring LLM-as-a-Judge: local self-consistency (pairwise preference stability) and global logical consistency (transitivity across a full set of preferences). We curate a dataset of 650 questions and confirm SAGE's reliability through experiments demonstrating metric stability and strong correlation with established supervised baselines. Based on SAGE, we reveal that current *state-of-the-art* LLMs exhibit significant reliability problems when acting as judges in both scoring and pairwise settings; even the top-performing models, Gemini-2.5-Pro and GPT-5, fail to maintain consistent preferences in nearly a quarter of difficult cases. We also find substantial inconsistency in human judgments, which indicates that human annotation may not be reliable. We attribute this to a new phenomenon called **situational preference**, which explains why explicit rubrics or criteria can help the model judge consistently across answer pairs. Our further analysis shows that finetuned LLM-as-a-Judge is a feasible method to boost performance, and the panel-based judge, as well as deep reasoning, can enhance the judging consistency.

1 Introduction

The LLM-as-a-Judge paradigm (Zheng et al., 2023) uses a large language model (LLM) to evaluate AI system outputs, offering a scalable and efficient alternative to costly and time-consuming human evaluation. Furthermore, beyond merely assessing performance, these evaluators are instrumental in refining models. During training, an LLM-as-a-Judge acts as a scalable reward model to enhance models' performance through automated feedback (Bai et al., 2022; Ouyang et al., 2022; Yuan et al., 2024; Luo et al., 2024). At inference time, it also serves as a test-time scaling module for real-time filtering to select the best response (Lightman et al., 2023; Faria and Smith, 2025).

However, the LLM-as-a-Judge paradigm is undermined by inherent flaws. Judge models are susceptible to biases such as positional (Shi et al., 2024), verbosity (Saito et al., 2023), and self-enhancement (Wataoka et al., 2024). In response, various benchmarks have been developed to assess the judges themselves (Zheng et al., 2023; Chiang et al., 2023; Gera et al., 2025; Pu et al., 2025). Yet, the construction of these benchmarks generally relies on human-annotated ground truth, which leads to two problems:

- First, the acquisition of human annotations is a notoriously expensive and labor-intensive process, limiting the scale and diversity. (Horych et al., 2024; Liao et al., 2025).
- Second, and more fundamentally, assuming human judgment as a gold standard is precarious, a *"bitter lesson"* where human-induced biases compromise AI evaluation (Sutton, 2019). As illustrated in Figure 1, this reliance is problematic. Persistent inter-annotator disagreement creates noisy data (Zhang et al., 2024), demonstrated by low agreement shown in AlpacaFarm (66%, Dubois et al. (2023)) and MT-Bench (63%, Zheng et al. (2023)). This problem is compounded when lengthy answers tax human cognitive capacity. Furthermore, human evaluators are susceptible to cognitive biases (Zheng et al., 2023; Chen et al., 2024a; Wu and Aji, 2025),

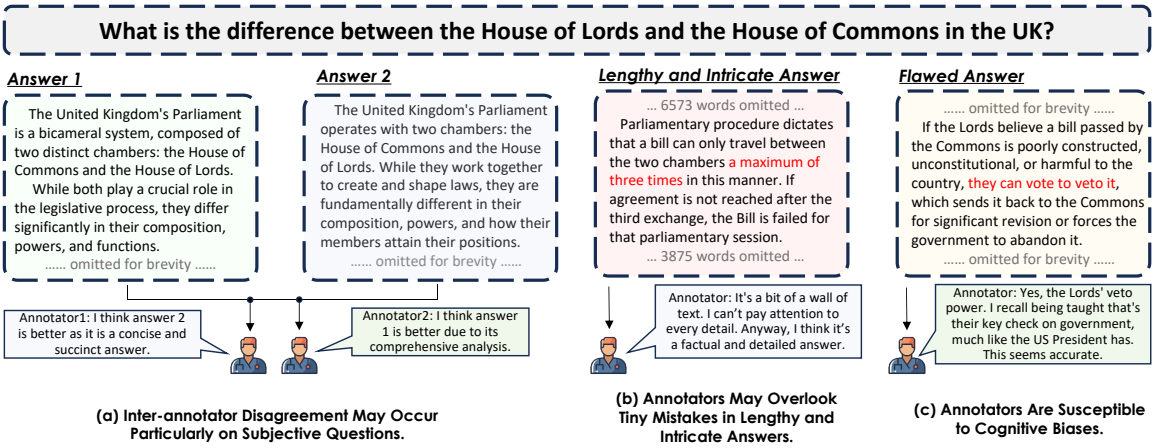


Figure 1: Human-annotated preference may not be reliable. We find three key challenges with relying on human annotators for evaluating LLM-as-a-Judge systems. (a) Inter-annotator Disagreement: Different annotators can have conflicting preferences, especially for subjective questions, leading to noisy and inconsistent data. (b) Overlooking Nuances: Annotators may miss subtle errors or inaccuracies in lengthy and complex answers, leading to flawed evaluations. (c) Cognitive Biases: Human evaluators are susceptible to cognitive biases, such as favoring an answer that confirms their false beliefs, which can further compromise the objectivity of the assessment.

favoring answers that match their false beliefs, making annotations unreliable.

To address this challenge, we introduce SAGE (**S**elf-**A**ssessing **G**auge for **E**valuators), a novel evaluation suite for assessing LLM-as-a-Judge robustness without any human annotation. Our approach is grounded in principles of rational decision-making, which posit that a reliable judge must exhibit consistent and coherent preferences. For example, a robust judge's preference between two answers should not flip simply because their presentation order is swapped. Furthermore, its judgments should adhere to the principle of transitivity, maintaining a logical and consistent order across a full set of preferences (Ouyang et al., 2022; Song et al., 2024; Hou et al., 2024; Hu et al., 2024; Liu et al., 2024). A breakdown in this coherence suggests the model lacks a consistent internal gauging principle for the question, leading to unreliable situational preferences.

Based on these principles, we propose two metrics to quantify this robustness: **Intra-Pair Instability (IPI)** and **Weak Total Order Violation (TOV)**. IPI directly measures the local, pairwise consistency by detecting instabilities caused by positional bias, as in the first example. TOV, on the other hand, assesses the global logical coherence of a judge's complete set of preferences, identifying systemic contradictions like the violation of transitivity described.

For the evaluation, we curate a diverse dataset of 650 questions by combining selections from RewardBench2 (Gureja et al., 2025) and the large-scale WildChat-1m corpus (Zhao et al., 2024) to ensure broad coverage of real-world user queries. We validate the soundness of our metrics by demonstrating both their **intrinsic stability** through consistent checking and a distribution-free error bounding method that yields minuscule variance on the order of 10^{-5}, and their **external alignment** with established supervised benchmarks like LLMBar (Zeng et al., 2023) and RewardBench2, confirming SAGE 's reliability as an evaluation suite.

Based on SAGE, we evaluate a wide range of LLM-as-a-Judge systems in both scoring and pairwise settings, including state-of-the-art LLMs, fine-tuned judges, and multi-agent as juries. All judge models degrade when encountering candidate answers that are in a closer gap, with approximately 200% more inconsistency, highlighting the potential problem in using LLM-as-a-Judge in RL-based training and test-time scaling. Our findings reveal that current models exhibit significant robustness deficiencies and most specialized fine-tuning brings about improvement. Our findings also show that multi-agent panels can improve performance by up to 15% and that increasing a model's reasoning depth can only bring about a minor enhancement. Notably, prompting for self-generated rubrics to avoid situational preference yields an even greater performance boost, reducing IPI (local inconsistency) and TOV (global inconsistency) by 16.1% and 11.0%, respectively. In our further analysis, we investigate the alignment between different evaluation formats. We also demonstrate SAGE's practical utility in selecting stable evaluators for automated

arenas and its cost advantage over human annotation. Finally, we apply SAGE to human evaluators and reveal the fragility of human *"gold standards"*, where IPI reaches 0.332 and TOV surges to 6.523 on complex tasks. We will release all source code, curated dataset at https://anonymous.4open.science/r/EMNLP2026-910B/

2 Assessing LLM-as-a-Judge with SAGE

This section details the foundational methodology of our proposed framework, SAGE. We begin by formally defining the evaluation problem and introducing a symmetrized protocol. Building on this, we then present our two novel metrics: Intra-Pair Instability (IPI) to assess local, pairwise consistency, and Weak Total Order Violation (TOV) to measure global, logical coherence.

2.1 Problem Formulation

Let M be the LLM under evaluation, referred to as the **judge model**. Our evaluation is based on a set of questions $\mathcal{Q}$. For any given question $Q \in \mathcal{Q}$, we generate a set of n candidate answers, denoted as $A_Q = \{A_1, A_2, \ldots, A_n\}$. The core task of the judge model M is to perform a **pairwise comparison** between any two answers, A_i and A_j, from the set A_Q. We define a function J_M:

$$y_{ij} = J_M(Q, A_i, A_j) \in \{-1, 0, 1\} \quad (1)$$

where the outcome y_{ij} is interpreted as:

- $y_{ij} = 1$: M judges A_i to be superior to A_j ($A_i \succ A_j$).
- $y_{ij} = -1$: M judges A_i to be inferior to A_j ($A_i \prec A_j$).
- $y_{ij} = 0$: M judges A_i and A_j to be of equal quality ($A_i = A_j$).

For each question Q, we conduct a full round-robin evaluation, assessing all $\binom{n}{2}$ pairs of answers to establish a complete set of pairwise judgments for subsequent analysis.

2.2 Symmetrized Evaluation Protocol

A naive single-pass evaluation is susceptible to **positional bias**, where the order of presentation influences the outcome (Appendix F.1 substantiates this claim). To tackle this issue, we adopt a **symmetrized evaluation protocol**. For each unordered pair $\{A_i, A_j\}$, we query the judge model twice:

Forward pass: $y_{ij} \leftarrow J_M(Q, A_i, A_j)$;
Reversed pass: $y_{ji} \leftarrow J_M(Q, A_j, A_i)$.

2.3 Two Evaluation Metrics

We propose two metrics to quantify the robustness, targeting two distinct failure modes: local inconsistency on a single pair and global logical incoherence across a set of answers.

Intra-Pair Instability (IPI). This metric assesses robustness from an **atomic**, **local** level. A perfectly consistent judge would always produce opposite scores for reversed pairs (i.e., $y_{ij} = -y_{ji}$). The IPI score for a given question Q quantifies the deviation from this ideal by calculating the average disagreement across all unique pairs:

$$\text{IPI}(Q) = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} \mathbb{I}(y_{ij} \neq -y_{ji}) \quad (2)$$

Higher IPI scores indicate a greater local inconsistency.

Weak Total Order Violation (TOV). This metric assesses robustness from a **global**, **systematic** level. Specifically, it measures the logical coherence of the judge's full set of preferences for a question. A rational judge's preferences should be transitive and form a **weak total order** (i.e., a total order that allows ties). Let $\mathbf{J}_Q = \{y_{ij}\}_{1 \le i,j \le n, i \neq j}$ be the set of derived preferences from the symmetrized evaluation for a question Q. Let $\mathcal{O}_n$ be the set of all possible valid weak total orders on n items. For any order $O \in \mathcal{O}_n$, we can represent it as a corresponding set of pairwise relations $\mathbf{P}_O = \{p_{ij}\}$, where $p_{ij} \in \{-1, 0, 1\}$ denotes the pairwise relationship between items i and j with the order O. Specifically, $p_{ij} = 1$ if i is preferred to j, $p_{ij} = -1$ if j is preferred to i, and $p_{ij} = 0$ if they are tied. The TOV score is defined as the minimum number of preference changes required to transform the judge's observed preferences $\mathbf{P}_Q$ into any valid weak total order:

$$\text{TOV}(Q) = \min_{O \in \mathcal{O}_n} \sum_{1 \le i,j \le n, i \neq j} \mathbb{I}(y_{ij} \neq p_{ij}) \quad (3)$$

Higher TOV scores signify greater logical contradictions.

To summarize overall performance, we compute aggregate scores, which are the mean of the per-question scores over the entire $\mathcal{Q}$, calculated as $\text{IPI} = (1/|\mathcal{Q}|) \sum_{Q \in \mathcal{Q}} \text{IPI}(Q)$ and $\text{TOV} = (1/|\mathcal{Q}|) \sum_{Q \in \mathcal{Q}} \text{TOV}(Q)$.

3 The Construction of SAGE

We source the question set $\mathcal{Q}$ from RewardBench2 (Gureja et al., 2025) and WildChat-1M cor-

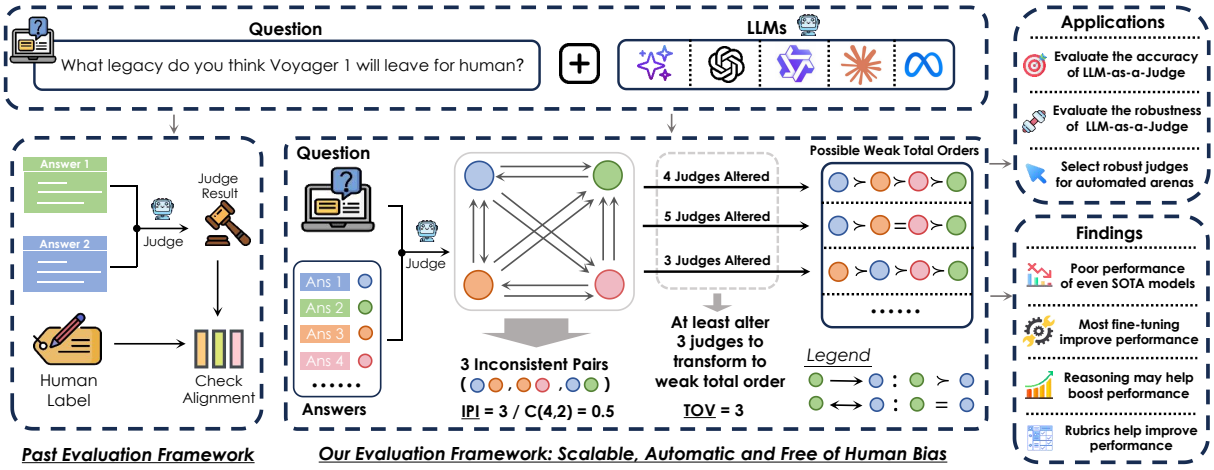
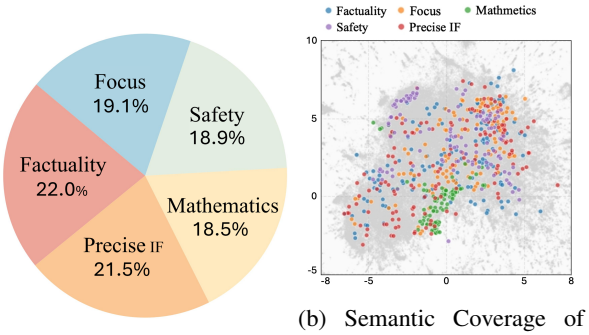


Figure 2: SAGE uses a symmetrized, round-robin protocol to conduct pairwise comparisons on a set of candidate answers. From these judgments, SAGE quantifies performance using two metrics: **IPI**, which measures local consistency by tracking preference reversals (*e.g.*, 3 inconsistent pairs result in an IPI of 0.5), and **TOV**, which assesses global logical coherence by calculating the minimum alterations required for a consistent ranking (*e.g.*, 3 alterations required). This methodology scalably diagnoses logical deficiencies to help identify and select more reliable LLM evaluators.



(a) Category Composition (b) Semantic Coverage of Question Set

Figure 3: We provide statistics and analysis of our selected queries and answers within SAGE.

pus (Zhao et al., 2024) to better reflect real-world interactions. The resulting question set consists of 650 questions, and its category composition is shown in Figure 3a. To validate its semantic diversity, we use a t-SNE visualization (van der Maaten and Hinton, 2008) to project our questions against a background of 500k English questions from WildChat1M. As shown in Figure 3b, our questions spread broadly across the embedding space, confirming the dataset's representativeness and wide topical coverage. Further details are provided in Appendix D.1.

For each of the 650 questions, we generate a set of $n = 6$ answers for the LLM judge to evaluate, which were used to construct two distinct tiers: SAGE-EASY and SAGE-HARD.

- **SAGE-EASY**: the six answers are generated by six diverse models with a clear capability gradient: Gemini-2.5-Pro and Gemini-2.5-Flash (Team, 2025a); Qwen3-32B (Team, 2025b), Claude-3-Haiku (Anthropic, 2024), Llama-3.2-3B-Instruct, and Llama-3.2-1B-Instruct (Meta, 2024b). These models, which have a well-documented performance gap on LMSYS Chatbot Arena (LMSYS, 2025), produce answers with a wide variance in quality, making the comparison relatively simple. Crucially, SAGE-EASY reflects the **general-purpose task** of comparing models of varying capabilities, which is largely used in automated judges like MT-Bench and Arena-Hard-Auto.
- **SAGE-HARD**: all six answers are generated by a single capable model, Gemini-2.5-Flash. Since the answers originate from the same model, their quality is much more homogeneous. This setup presents a more challenging task, requiring the judge to make finer-grained distinctions between subtly different responses. Crucially, SAGE-HARD models the judge's role in applications like **model-based reinforcement learning** and **rejection sampling**. In these scenarios, the judge must distinguish between subtly varied outputs from a single capable model.

To quantitatively confirm the difference in quality diversity between these two tiers, we use the reward model QRM-Gemma-2-27B (Dorka, 2024) to score six answers per question and compute the coefficient of variation (CV, σ/μ) of these scores. As shown in Figure 8, the CV distribution for SAGE-

HARD is markedly shifted towards lower values, indicating reduced within-question dispersion in answer quality and consequently a more challenging setting for LLM judges. We further conduct a controlled human annotation study with 20 researchers to confirm the difficulty gap. Adjudication time increase by 42% (from 7.3 to 10.4 minutes) on SAGE-HARD, indicating significantly higher cognitive load and narrower quality gap.

4 Experiment and Analysis

We first validate the consistency of our metrics, after which we evaluate a diverse set of general, fine-tuned, and multi-agent judges. Results highlight significant robustness gaps in top-tier models, particularly on fine-grained tasks. We further identify *"situational preference"* as the primary cause of inconsistency and demonstrate that fine-tuning, multi-agent frameworks, and deep reasoning with explicit rubrics can effectively mitigate this issue.

4.1 Validating Metric Stability and Robustness

A critical aspect of a reliable framework is the stability of its evaluation metrics against the inherent stochasticity of models. We analyze metric stability from both an empirical and a theoretical standpoint.

Theoretical Guarantees. Our argument proceeds in three stages. First, using principles from Conformal Prediction (Angelopoulos and Bates, 2021), we establish a probabilistic guarantee that any single pairwise judgment y_{new} matches its most probable outcome, y^*_{new}, with high confidence:

$$P(y_{\text{new}} = y^*_{\text{new}}) \geq 1 - \alpha. \quad (4)$$

Second, we use this result to derive an upper bound on the variance of the per-question metrics:

$$\text{Var}(\text{IPI}(Q)) \leq \mathbb{E}[\Delta_{\text{IPI}}(Q)^2] \leq 0.0075. \quad (5)$$

Finally, we show that this variance diminishes over the aggregate evaluation suite. Assuming the per-question scores are independent and identically distributed, we could have:

$$\text{Var}(\text{IPI}) = \text{Var}(\text{IPI}(Q))/|\mathcal{Q}| \leq 1.15 \times 10^{-5}. \quad (6)$$

A similar derivation shows $\text{Var}(\text{TOV}) \leq 2.59 \times 10^{-3}$. These theoretical results align with our empirical findings, confirming that the final scores are highly stable. The full derivation of this analysis is available in Appendix B.

Table 1: Variance across 10 independent runs for LLM-as-a-Judge.

Model	IPI Variance	TOV Variance
Qwen3-4B	2.2×10^{-6}	6.5×10^{-4}
Qwen3-30B-A3B	6.7×10^{-6}	1.5×10^{-3}

Empirical Analysis. As presented in Table 1, the observed variances are exceptionally low, which provides strong evidence that our metrics are highly reproducible and capture the fundamental reasoning patterns of the judge model rather than ephemeral artifacts of its generative process.

Consistency across Temperatures. We evaluate model performance across various temperature settings. Our metrics exhibit minimal variance ($\sim 10^{-5}$) across different temperature settings. Full result is available in Appendix F.4.

4.2 Indicator for Robustness and Accuracy

Table 2: Spearman Correlation Coefficients between SAGE metrics and external benchmarks. P-values are less than 0.001.

LLMBar		RewardBench2	
IPI	TOV	IPI	TOV
0.802	0.791	0.890	0.879

External Alignment and Accuracy Indicator. We validate SAGE from both empirical and theoretical perspectives. Empirically, our metrics strongly correlate with model error rates on the adversarial subset of LLMBar (Zeng et al., 2023), and TOV also aligns with true preference errors on the RewardBench2-labeled subset (Table 2). Theoretically, TOV lower-bounds the minimum number of incorrect pairwise judgments that must be revised to restore logical coherence, since correct preferences should be logically consistent (proof in Appendix C). Thus, for n candidate answers, we estimate judging accuracy as $1 - \text{TOV}/P$, where P is the total number of pairwise comparisons and TOV is treated as the estimated number of erroneous comparisons. We further verify this link with human batch rankings on 50 questions: judges' true accuracy is close to the TOV-estimated accuracy, with an average gap of only 0.06.

TOV-Guided Preference Refinement. We further show TOV identifies accuracy-relevant noise in preference data. For each set of raw pairwise prefer-

Table 3: The performance of thirteen LLMs on SAGE, with lower scores indicating greater robustness. A clear trend emerges where advanced models like Gemini-2.5-Pro demonstrate superior robustness.

Models	Factuality		Precise IF		Mathematics		Safety		Focus		Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
Performance on SAGE-EASY												
Gemini-2.5-Pro	**0.059**	**0.900**	**0.084**	**1.266**	**0.060**	**0.924**	0.100	1.512	**0.058**	**0.863**	**0.072**	**1.091**
Gemini-2.5-Flash	0.077	1.163	0.111	1.679	0.075	1.175	**0.097**	**1.504**	0.071	1.081	0.087	1.326
Qwen3-235B-A22B-Instruct-2507	0.076	1.134	0.130	1.959	0.098	1.505	0.103	1.557	0.086	1.295	0.098	1.485
DeepSeek-R1-0528	0.095	1.436	0.142	2.183	0.141	2.254	0.124	1.857	0.087	1.300	0.115	1.759
Qwen3-4B-Instruct-2507	0.101	1.532	0.175	2.662	0.126	1.933	0.124	1.943	0.102	1.532	0.126	1.919
Gemini-2.0-Flash-Lite	0.105	1.587	0.152	2.321	0.161	2.433	0.147	2.650	0.102	1.524	0.133	2.091
GPT-5-Chat	0.120	1.818	0.232	3.489	0.104	1.622	0.121	1.820	0.127	1.902	0.143	2.158
DeepSeek-V3-0324	0.130	1.985	0.185	2.783	0.147	2.290	0.173	2.612	0.146	2.200	0.157	2.380
Qwen3-30B-A3B-Instruct-2507	0.136	2.056	0.208	3.113	0.123	1.900	0.171	2.569	0.173	2.597	0.162	2.448
DeepSeek-V3.1	0.134	2.022	0.189	2.920	0.146	2.449	0.204	3.097	0.132	2.017	0.160	2.485
GPT-4o-mini	0.150	2.294	0.202	3.085	0.164	2.522	0.107	1.667	0.182	2.733	0.164	2.496
Llama-3.1-8B-Instruct	0.204	3.203	0.229	3.506	0.260	4.000	0.230	3.493	0.179	2.755	0.221	3.410
Claude-3-Haiku	0.208	3.177	0.253	3.822	0.367	6.046	0.337	5.691	0.253	3.856	0.279	4.430
Performance on SAGE-HARD												
Gemini-2.5-Pro	**0.240**	**3.769**	**0.291**	**4.586**	**0.193**	**3.661**	0.268	4.439	**0.250**	**3.903**	**0.250**	**4.079**
Gemini-2.5-Flash	0.265	4.175	0.320	5.007	0.257	4.457	**0.228**	**3.732**	0.267	4.331	0.269	4.350
Qwen3-235B-A22B-Instruct-2507	0.347	5.326	0.357	5.585	0.256	4.404	0.268	4.124	0.420	6.409	0.331	5.183
Qwen3-4B-Instruct-2507	0.376	5.734	0.405	6.215	0.376	6.254	0.346	5.653	0.433	6.573	0.388	6.078
DeepSeek-R1-0528	0.410	6.397	0.451	7.067	0.276	5.032	0.398	6.404	0.434	6.828	0.396	6.362
Qwen3-30B-A3B-Instruct-2507	0.424	6.458	0.393	6.059	0.337	5.625	0.379	6.065	0.530	7.968	0.413	6.435
Gemini-2.0-Flash-Lite	0.434	6.594	0.376	5.764	0.451	7.417	0.336	5.919	0.480	7.282	0.415	6.571
GPT-5-Chat	0.433	6.563	0.571	8.619	0.287	4.850	0.286	4.431	0.583	8.764	0.436	6.700
DeepSeek-V3-0324	0.455	7.146	0.455	7.060	0.490	8.286	0.403	6.331	0.533	8.492	0.467	7.446
Claude-3-Haiku	0.522	8.442	0.507	7.985	0.292	5.213	0.352	6.171	0.610	10.161	0.464	7.687
DeepSeek-V3.1	0.513	8.630	0.585	9.171	0.214	3.893	0.438	7.235	0.523	9.470	0.460	7.756
GPT-4o-mini	0.546	8.261	0.478	7.276	0.452	7.803	0.353	5.755	0.661	9.974	0.502	7.865
Llama-3.1-8B-Instruct	0.559	8.507	0.496	7.600	0.653	10.254	0.581	9.246	0.608	9.265	0.573	8.869

ences, we construct a preference graph and project it onto the nearest logically consistent weak order by minimizing TOV-inducing violations. Pairwise labels regenerated from this refined order are then used to fine-tune Qwen3-4B. Compared with fine-tuning on the original preferences, the TOV-refined data produces a 1.9× improvement, suggesting that TOV captures preference errors that matter for downstream judging accuracy. This also offers a zero-cost way to improve existing preference data without additional annotation.

4.3 Evaluating LLM-as-a-Judge with SAGE

We benchmark thirteen popular LLMs, including six proprietary models (*i.e.* Gemini-2.5-Pro and Gemini-2.5-Flash (Team, 2025a); Gemini-2.0-Flash-Lite (Google, 2025), GPT-5-Chat (OpenAI, 2025), GPT-4o-mini (OpenAI, 2024) and Claude-3-Haiku (Anthropic, 2024)) and seven open source models (*i.e.* Qwen3-235B-A22B-Instruct-2507, Qwen3-30B-A3B-Instruct-2507 and Qwen3-4B-Instruct-2507 (Team, 2025b); DeepSeek-R1-0528 (DeepSeek-AI, 2025a), DeepSeek-V3 (DeepSeek-AI, 2024), DeepSeek-V3.1 (DeepSeek-AI, 2025b), Llama-3.1-8B-Instruct (Meta, 2024a)). All evaluations are conducted at the default temperature to ensure a fair and consistent comparison.

As shown in Table 3, even top-tier LLMs exhibit significant deficiencies. All LLMs degrade sharply from SAGE-EASY to SAGE-HARD. This shows that judges falter on subtle distinctions, posing a threat to inference-time enhancement techniques like rejection sampling.

4.4 In-depth Analysis

Unjust Judges or Situational Preference? We argue that a robust LLM-as-a-Judge should first model the question internally regardless of how the

Table 4: Fine-tuning serves as an effective strategy on SAGE-HARD.

Models	Factuality		Precise IF		Mathematics		Safety		Focus		Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
Qwen2.5-3B-Instruct (Base)	0.758	11.951	0.677	10.791	0.539	9.833	0.650	10.984	0.795	12.395	0.687	11.211
M-Prometheus-3B	0.497	7.763	0.418	7.622	0.512	8.233	0.465	7.365	0.492	7.725	0.490(↑29%)	7.745(↑31%)
JudgeLRM-3B	0.909	13.667	0.765	11.504	0.533	8.342	0.618	9.451	0.930	13.952	0.757(↓10%)	11.471(↓2%)
Qwen2.5-7B-Instruct (Base)	0.847	12.713	0.765	11.507	0.712	11.583	0.840	12.740	0.911	13.686	0.815	12.435
M-Prometheus-7B	0.564	8.671	0.516	8.014	0.491	7.825	0.451	7.008	0.613	9.323	0.528(↑35%)	8.183(↑34%)
JudgeLRM-7B	0.765	11.902	0.695	10.976	0.404	6.673	0.502	8.517	0.811	12.856	0.643(↑21%)	10.291(↑13%)
Mistral-7B-Instruct (Base)	0.793	11.963	0.651	9.889	0.813	12.455	0.753	11.302	0.821	12.357	0.765	11.566
Prometheus-7B-V2.0	0.616	9.634	0.546	8.773	0.602	10.000	0.553	9.186	0.652	10.105	0.592(↑23%)	9.509(↑18%)
Llama-3.1-8B-Instruct (Base)	0.559	8.507	0.496	7.600	0.653	10.254	0.581	9.246	0.608	9.265	0.573	8.869
Skywork-Critic-Llama-3.1-8B	0.491	7.378	0.421	6.379	0.371	5.583	0.371	5.667	0.539	8.081	0.440(↑23%)	6.642(↑25%)

answers vary. However, the extremely high IPI and TOV scores across even *state-of-the-art* models raise the concern of whether models are incapable of providing just judgments, or whether their judgments are merely **situational preferences** (Laine et al., 2024; Needham et al., 2025), *i.e.*, **inconsistent judging criteria encountering different answers under the same question**. To validate this hypothesis, we investigate whether an LLM can improve by first articulating its judging rubrics and then using the rubrics to judge answer pairs under the same question. Crucially, this rubric is generated only once per question and serves as a fixed standard for all answer pairs, a method designed to mitigate situational preferences by preventing the judge's evaluation criteria from shifting between comparisons. Figure 4 shows that this approach yields a notable performance boost, reducing IPI and TOV scores by 16.1% and 11.0%. This gap demonstrates that current LLM-as-a-Judge systems indeed exhibit extreme situational preferences when encountering different answer pairs, and that explicit judging rubrics can substantially mitigate.

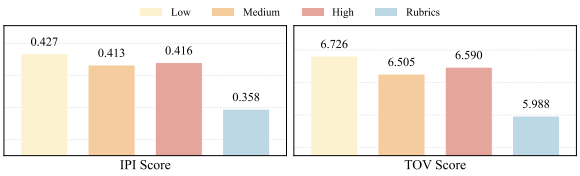


Figure 4: We discover high IPI and TOV scores in SAGE-HARD due to the situational preference phenomenon in LLM-as-a-Judge, while deep thinking and explicit rubrics can mitigate this.

Do fine-tuned judges make better judgments? In most cases, yes. We benchmark six fine-tuned judges (*i.e.* Prometheus-7B-V2.0 (Kim et al., 2024), Skywork-Critic-Llama-3.1-8B (Tu et al., 2024), M-Prometheus-3/7B (Pombal et al., 2025), and JudgeLRM-3/7B (Chen et al., 2025)) and their corresponding base models. The results are shown in Table 4. Additional results are in Appendix F.5. Our results demonstrate that fine-tuning is generally an effective strategy for enhancing robustness. However, the regression in JudgeLRM-3B model suggests vulnerability to training data biases. We attribute this degradation to the amplification of flawed judgment patterns inherited from the dataset. See Appendix G.7 for examples.

Do Multi-agent Debates or Panels Judge Better? Panels usually help, but debates often hurt. We study panel-based (PoLL) (Verga et al., 2024) and debate-based (ChatEval) (Chan et al., 2024) multi-agent judges. As shown in Table 5, PoLL panels generally surpass the best individual judges, whereas ChatEval consistently degrades robustness; ablations in Appendix F.2 show that this is not merely a configuration artifact. Detailed setups and failure cases are in Appendices F.3 and G.6.

Does deep reasoning lead to better performance? Generally yes. We use gpt-oss-120b (Agarwal et al., 2025) for its fine-grained reasoning modes: low, medium and high. As shown in Figure 4, the results show improvement as the reasoning mode is intensified from low to high.

What is the practical application of SAGE? Selecting robust judges for automated arenas. On Arena-Hard-Auto, SAGE metrics correlate with Elo confidence intervals (0.64/0.60), showing that they can identify stable judges for automated rankings.

Is SAGE sensitive to prompt design? Not materially. We study the impact of prompt variations on the consistency of judge performance. As shown in Figure 5, the deviation is minimal. Additional

Table 5: Performance comparison of multi-agent systems: PoLL panels (left) and ChatEval debates (right). For PoLL, "Best Indi." refers to the best individual model in the panel.

Method	IPI-Easy ↓	TOV-Easy ↓	IPI-Hard ↓	TOV-Hard ↓
Panel 1 (Powerful Models)				
Best Indi.	0.072	1.091	0.250	4.079
Aggregate	**0.067** (↑7%)	**1.022** (↑6%)	**0.245** (↑2%)	**3.897** (↑4%)
Panel 2 (Weaker Models)				
Best Indi.	0.133	2.091	0.415	6.571
Aggregate	**0.116** (↑13%)	**1.769** (↑15%)	**0.400** (↑4%)	**6.448** (↑2%)

Method	IPI-Easy ↓	TOV-Easy ↓	IPI-Hard ↓	TOV-Hard ↓
Qwen3-4B-Instruct-2507				
Baseline	**0.129**	**1.952**	**0.386**	**5.849**
ChatEval	0.334 (↓158%)	5.105 (↓162%)	0.651 (↓69%)	10.050 (↓72%)
Qwen3-30B-A3B-Instruct-2507				
Baseline	**0.180**	**2.715**	0.649	9.780
ChatEval	0.261 (↓45%)	4.080 (↓50%)	**0.518** (↑20%)	**8.395** (↑14%)

prompts are in Appendix G.4.

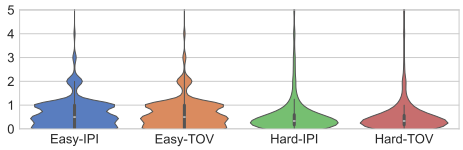


Figure 5: Distribution of performance spread, measured as maximum absolute percentage changed in metric.

Model-Agnostic Nature of Benchmark Difficulty. To verify that the difficulty of SAGE-HARD is not specific to the source model (Gemini-2.5-Flash), we constructed a parallel version using Qwen3-235B-A22B as the answer generator. As shown in Table 6, judge performance exhibits minimal deviation between these two sets.

Table 6: Validation of Model-Agnostic Difficulty.

Dataset Source	IPI	TOV
Gemini	0.388	6.078
Qwen	0.386(↑0.5%)	6.109(↓0.5%)

Human Consistency Baseline. We further measure human consistency with 20 graduate students across 50 questions. We controlled for memory bias by assigning forward and reversed pairs to different annotators and interleaving tasks. The results (Appendix Table 14) suggests that human preferences are not perfectly stable or transitive, motivating intrinsic consistency checks such as SAGE.

Inconsistency Between Different Judge Formats. Another critical dimension of a judge's reliability is its consistency across different judge formats (*e.g.* direct scoring, pairwise comparison and batch ranking). The results are available in Appendix F.7, which show prominent inconsistency emerges when LLM faces different judge formats.

The Influence of Reasoning on Judgment Consistency. In the main experiments, judges explain before giving final verdicts; in follow-up studies,

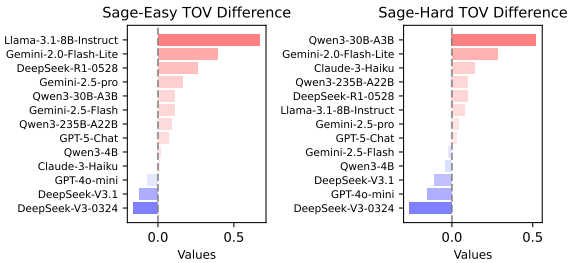


Figure 6: The difference of TOV when models output directly in comparison to the original conditions.

we remove the rationale and ask for direct decisions. As shown in Figure 6, this causes widespread degradation, with additional results in Appendix F.8.

5 Conclusion

We introduce SAGE, a novel framework for evaluating LLM-as-a-Judge without human annotation or *any* extrinsic information by measuring local and global logical consistency. We validate that our metrics are exceptionally stable and can serve as an indicator for accuracy. Our experiments reveal significant robustness deficiencies in current state-of-the-art models. We attribute these inconsistent judgments to a newly identified phenomenon called situational preference where models fail to maintain a stable internal gauging principle. To address this, we demonstrate that implementing self-generated rubrics effectively mitigates situational preference and boosts judgment consistency. We also investigate the impact of fine-tuning and explanatory reasoning on evaluation performance. Crucially, we apply SAGE to human evaluators and reveal the fragility of human judgment. The significant instability observed in human baseline proves that human annotation is an unreliable standard. Consequently, SAGE provides a scalable, reliable, and cost-effective tool to diagnose and improve LLM evaluators, paving the way for more consistent and rational AI systems.

Limitations

Our study is limited to English-language evaluation settings. Therefore, the observed inconsistency patterns and benchmark difficulty estimates may not directly generalize to multilingual or cross-lingual scenarios. Future work should examine whether the same findings hold for other languages, especially low-resource languages and culturally specific evaluation contexts.

Impact Statement

Our dataset is curated from established public research sources: the RewardBench2 benchmark and the WildChat-1M corpus. To mitigate ethical risks, such as the potential inclusion of private information or inappropriate content from real-world user logs, we conducted a rigorous curation process (see Appendix D.1). This process involved both large-scale automated filtering and a thorough manual review of every selected question. This ensures that the final dataset is appropriate for research use and aligns with the data-sharing and privacy standards of the original sources.The potential risks of this work are limited, as it analyzes existing LLM judges rather than deploying new models or decision systems.

References

Sandhini Agarwal, Lama Ahmad, Jason Ai, Sam Altman, Andy Applebaum, Edwin Arbus, Rahul K Arora, Yu Bai, Bowen Baker, Haiming Bao, and 1 others. 2025. gpt-oss-120b & gpt-oss-20b model card. *arXiv preprint arXiv:2508.10925*.

Anastasios N. Angelopoulos and Stephen Bates. 2021. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511.

Anthropic. 2024. Claude 3 haiku: our fastest model yet. https://www.anthropic.com/news/claude-3-haiku. Accessed: 2025-09-08.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, and 12 others. 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. *CoRR*, abs/2204.05862.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2024. Chateval: Towards better llm-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024a. Humans or llms as the judge? A study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 8301–8327. Association for Computational Linguistics.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024b. Humans or llms as the judge? A study on judgement biases. *CoRR*, abs/2402.10669.

Nuo Chen, Zhiyuan Hu, Qingyun Zou, Jiaying Wu, Qian Wang, Bryan Hooi, and Bingsheng He. 2025. Judgelrm: Large reasoning models as a judge. *Preprint*, arXiv:2504.00050.

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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

DeepSeek-AI. 2024. Deepseek-v3 technical report. *CoRR*, abs/2412.19437.

DeepSeek-AI. 2025a. Deepseek-r1-0528 release. https://api-docs.deepseek.com/news/news250528. Accessed: 2025-09-08.

DeepSeek-AI. 2025b. Deepseek-v3.1 release. https://api-docs.deepseek.com/news/news250821. Accessed: 2025-09-08.

Nicolai Dorka. 2024. Quantile regression for distributional reward models in rlhf. *Preprint*, arXiv:2409.10164.

Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Alpacafarm: A simulation framework for methods that learn from human feedback. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Bradley Efron. 1992. Bootstrap methods: Another look at the jackknife. In Samuel Kotz and Norman L. Johnson, editors, *Breakthroughs in Statistics: Methodology and Distribution*, pages 569–593. Springer.

Gonçalo Faria and Noah A. Smith. 2025. Sample, don't search: Rethinking test-time alignment for language models. *Preprint*, arXiv:2504.03790.

Leo Gao, John Schulman, and Jacob Hilton. 2023. Scaling laws for reward model overoptimization. In *International Conference on Machine Learning*, pages 10835–10866. PMLR.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. 2025. Justrank: Benchmarking LLM judges for system ranking. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 682–712. Association for Computational Linguistics.

Google. 2025. Gemini 2.0: Flash, flash-lite and pro. https://developers.googleblog.com/en/gemini-2-family-expands/. Accessed: 2025-09-08.

Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh Fadaee. 2025. M-rewardbench: Evaluating reward models in multilingual settings. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 43–58. Association for Computational Linguistics.

Yuanqin He, Yan Kang, Lixin Fan, and Qiang Yang. 2024. Fedeval-llm: Federated evaluation of large language models on downstream tasks with collective wisdom. *arXiv preprint arXiv:2404.12273*.

Tomas Horych, Christoph Mandl, Terry Ruas, Andre Greiner-Petter, Bela Gipp, Akiko Aizawa, and Timo Spinde. 2024. The promises and pitfalls of llm annotations in dataset labeling: a case study on media bias detection. *arXiv preprint arXiv:2411.11081*.

Yupeng Hou, Junjie Zhang, Zihan Lin, Hongyu Lu, Ruobing Xie, Julian McAuley, and Wayne Xin Zhao. 2024. Large language models are zero-shot rankers for recommender systems. In *European Conference on Information Retrieval*, pages 364–381. Springer.

Zhengyu Hu, Jieyu Zhang, Zhihan Xiong, Alexander Ratner, Hui Xiong, and Ranjay Krishna. 2024. Language model preference evaluation with multiple weak evaluators. *arXiv preprint arXiv:2410.12869*.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2024. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a general substitute for gpt-4. *arXiv preprint arXiv:2403.02839*.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *CoRR*, abs/2310.06825.

Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024. Prometheus: Inducing fine-grained evaluation capability in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémy Scheurer, Marius Hobbhahn, Alexander Meinke, and Owain Evans. 2024. Me, myself, and ai: The situational awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems*, 37:64010–64118.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*.

Hsuan Wei Liao, Christopher Klugmann, Daniel Kondermann, and Rafid Mahmood. 2025. Minority reports: Balancing cost and quality in ground truth data annotation. *arXiv preprint arXiv:2504.09341*.

Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. 2023. Let's verify step by step. In *The Twelfth International Conference on Learning Representations*.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.

Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel Collier. 2024. Aligning with human judgement: The role of pairwise preference in large language model evaluators. *arXiv preprint arXiv:2403.16950*.

LMSYS. 2025. Leaderboard overview. https://lmarena.ai/leaderboard. Accessed: 2025-09-08.

Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yansong Tang, and Weizhu Chen. 2024. Arena learning: Build data flywheel for llms post-training via simulated chatbot arena. *arXiv preprint arXiv:2407.10627*.

Meta. 2024a. Introducing llama 3.1: Our most capable models to date. https://ai.meta.com/blog/meta-llama-3-1/. Accessed: 2025-09-08.

Meta. 2024b. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/. Accessed: 2025-09-08.

Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. 2025. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*.

Mark EJ Newman. 2023. Efficient computation of rankings from pairwise comparisons. *Journal of Machine Learning Research*, 24(238):1–25.

OpenAI. 2024. Gpt-4o mini: advancing cost-efficient intelligence. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. Accessed: 2025-09-08.

OpenAI. 2025. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2025-09-08.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and André F. T. Martins. 2025. M-prometheus: A suite of open multilingual llm judges. *Preprint*, arXiv:2504.04953.

Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang, Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. 2025. Judge anything: MLLM as a judge across any modality. *CoRR*, abs/2503.17489.

Keita Saito, Akifumi Wachi, Koki Wataoka, and Youhei Akimoto. 2023. Verbosity bias in preference labeling by large language models. *arXiv preprint arXiv:2310.10076*.

Lin Shi, Chiyu Ma, Wenhua Liang, Weicheng Ma, and Soroush Vosoughi. 2024. Judging the judges: A systematic investigation of position bias in pairwise comparative assessments by llms.

Feifan Song, Bowen Yu, Minghao Li, Haiyang Yu, Fei Huang, Yongbin Li, and Houfeng Wang. 2024. Preference ranking optimization for human alignment. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18990–18998.

Richard Sutton. 2019. The bitter lesson. *Incomplete Ideas (blog)*, 13(1):38.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *The Thirteenth International Conference on Learning Representations*.

Gemini Team. 2025a. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261.

Llama Team. 2024a. The llama 3 herd of models. *CoRR*, abs/2407.21783.

Qwen Team. 2024b. Qwen2.5: A party of foundation models! https://qwenlm.github.io/blog/qwen2.5/. Accessed: 2025-09-08.

Qwen Team. 2025b. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Shiwen Tu, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. 2024. Skywork critic model series. https://huggingface.co/Skywork.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. 2024. Replacing judges with juries: Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*.

Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? gpt-3 can help. *arXiv preprint arXiv:2108.13487*.

Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu, Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. 2024a. Self-taught evaluators. *arXiv preprint arXiv:2408.02666*.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. 2024b. Pandalm: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Minghao Wu and Alham Fikri Aji. 2025. Style over substance: Evaluation biases for large language models. In *Proceedings of the 31st International Conference on Computational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pages 297–312. Association for Computational Linguistics.

Yifei Xu, Tusher Chakraborty, Emre Kıcıman, Bibek Aryal, Eduardo Rodrigues, Srinagesh Sharma, Roberto Estevao, Maria Angels de Luis Balaguer, Jessica Wolk, Rafael Padilha, and 1 others. 2025. Rlthf: Targeted human feedback for llm alignment. *arXiv preprint arXiv:2502.13417*.

Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 3.

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*.

Michael J. Q. Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol Choi, Xiang Ren, and Valentina Pyatkin. 2024. Diverging preferences: When do annotators disagree and do models know? *CoRR*, abs/2410.14632.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

Ethics Statement

Our dataset is curated from established public research sources: the RewardBench2 benchmark and the WildChat-1M corpus. To mitigate ethical risks, such as the potential inclusion of private information or inappropriate content from real-world user logs, we conducted a rigorous curation process (see Appendix D.1). This process involved both large-scale automated filtering and a thorough manual review of every selected question. This ensures that the final dataset is appropriate for research use and aligns with the data-sharing and privacy standards of the original sources. We used publicly available benchmark data and did not collect personal data from individuals; therefore, no additional individual consent was required for the curated data. We recruited graduate-student annotators for the human evaluation and compensated them fairly for their work.

The Use of Large Language Models (LLMs)

The use of large language models (LLMs) in this work is strictly limited to auxiliary text editing, such as correcting spelling and improving grammar, and dataset generation. Our study is about LLM-as-a-Judge, therefore we also test various LLMs for this task. All conceptual and technical contributions are the original work of the authors. We are transparent about this limited usage.

A Comprehensive Related Work

LLM-as-a-Judge. LLM-as-a-Judge (Zheng et al., 2023) has emerged as a scalable and cost-effective alternative to human evaluation for assessing the quality of generative AI outputs. This approach utilizes a powerful LLM to judge the responses of other models, addressing the limitations of traditional metrics like BLEU and ROUGE that often fail to capture deeper semantic qualities such as coherence, factual accuracy, and relevance.

However, the reliability of LLM-as-a-Judge is a significant concern, with numerous studies (Zheng et al., 2023; Chen et al., 2024a; Wu and Aji, 2025) highlighting its susceptibility to various biases. These include verbosity bias, where longer answers are favored irrespective of their quality; position bias, a preference for the first or last presented response; and self-enhancement bias, where a model tends to rate its own outputs more favorably. Research (Chen et al., 2024a) has also identified other distorting influences, such as authority bias, where an LLM may favor answers containing citations even if they are fabricated. These identified biases underscore the necessity for continued investigation and validation of the reliability of LLM-as-a-Judge.

Benchmark for LLM-as-a-Judge. Following the recognition of these potential biases of LLM-as-a-Judge, researchers have focused on developing specialized benchmarks to systematically evaluate the reliability and behavior of LLM judges. Unlike general-purpose LLM benchmarks that assess broad capabilities, these targeted frameworks are designed specifically to scrutinize the adjudicative performance of models. For instance, foundational benchmarks such as MT-Bench and Chatbot Arena (Zheng et al., 2023) are introduced to verify the agreement between LLM judges and human preferences on open-ended, multi-turn questions. Subsequent works like Tan et al. (2025) and Gera et al. (2025) continue to follow this paradigm, primarily assessing the capability of LLM judges by measuring the correlation between their assessments and human preference judgments.

However, this reliance on human judgment as the definitive *"gold standard"* is unreliable for three key reasons: First, human annotators are susceptible to inherent biases (Zheng et al., 2023; Wu and Aji, 2025), including authority bias and misinformation oversight bias (Chen et al., 2024a). In addition, Chen et al. (2024a) shows that human evaluators of LLMs can be more biased than the models themselves. Second, there is a persistent issue of inter-annotator disagreement (Zhang et al., 2024). Different human evaluators often provide inconsistent assessments, particularly for tasks that are subjective or nuanced. This lack of consensus means that the *"ground truth"* data used for benchmarking is often noisy and unreliable. Finally, as AI models advance, they are beginning to surpass human capabilities in specialized domains. When AI generates highly complex or lengthy outputs, human annotators might struggle to accurately assess their quality (Tan et al., 2025). In such scenarios, human annotations may no longer be a reliable ground truth.

Fine-tuned Judge. In the pursuit of improving automated evaluation accuracy, one prevalent strategy involves specializing a model using preference

datasets, resulting in a fine-tuned *"judge"* model (Kim et al., 2024; Wang et al., 2024b,a; He et al., 2024; Zhu et al., 2025). These datasets generally comprise a series of prompts, each followed by multiple model-generated responses, with evaluators providing labels to indicate the superior response. By leveraging this data, the judge model is trained to predict human evaluative behaviors, enabling it to autonomously score or rank new model outputs. The fine-tuning process allows the judge to learn nuanced patterns in human preferences, such as understanding which aspects of a response are prioritized. As a result, the judge can offer an automated alternative to human evaluation, making it invaluable for large-scale applications where human assessment may be time-consuming or impractical. However, this approach is not without its limitations (Huang et al., 2024). For those judge models that are fine-tuned on datasets derived from human evaluations, they inevitably inherit the biases and inconsistencies present in the human labeling process. Human annotators, despite their best efforts, may display subjective tendencies, varying interpretation of instructions, or inconsistencies in rating, which can be subtly reflected in the model's predictions (Chen et al., 2024b). As a consequence, the fine-tuned judge may sometimes generate evaluations that do not align with a broader, more objective standard (Gao et al., 2023). Given these challenges, the reliability and fairness of fine-tuned judge models as objective evaluators must be subjected to thorough scrutiny. It becomes crucial to investigate the degree to which these models mirror human biases and assess their robustness across diverse contexts and response types.

B Theoretical Analysis of Metric Stability

In this section, we provide a theoretical analysis to substantiate the empirical stability of our proposed metrics, Intra-Pair Instability (IPI) and Weak Total Order Violation (TOV), as presented in 4.1. The core of our analysis is to demonstrate that the variance of these metrics is exceptionally low, thereby ensuring their reliability against the inherent stochasticity of LLM judges.

The foundational source of any potential instability in our evaluation framework stems from the stochastic nature of the LLM judge, $\mathcal{M}$. When queried multiple times with the identical input triplet (Q, A_i, A_j), the model's judgment, $y_{ij} = J_{\mathcal{M}}(Q, A_i, A_j)$, may fluctuate. Our analysis proceeds in three stages: first, we certify the stability of a single pairwise judgment; second, we bound the variance of the per-question metrics; and third, we establish the stability of the final, aggregate benchmark scores.

B.1 Certifying Single-Pair Judgment Stability via Conformal Prediction

To formally quantify the stability of individual judgments, we adopt principles from Conformal Prediction (Angelopoulos and Bates, 2021). We posit that for any given question-answer pair, there exists a *"stable judgment,"* which represents the most stable outcome if the model were to be sampled repeatedly. We approximate this stable judgment by the modal outcome over a large number of trials.

We construct a large-scale calibration set, $\mathcal{C}$, by selecting $N = 800$ distinct question-answer pairs. For each pair $k \in \{1, \ldots, N\}$, we prompt the LLM judge $T = 20$ times, yielding a total of $N \times T = 16,000$ individual judgments. For each pair k, we define its stable judgment, y_k^*, as the most frequently observed outcome:

$$y_k^* = \underset{y \in \{-1,0,1\}}{\arg\max} \sum_{t=1}^{T} \mathbb{I}(y_k^{(t)} = y)$$

where $y_k^{(t)}$ is the outcome of the t-th judgment for the k-th pair.

We can now use the $n = 16,000$ judgments in $\mathcal{C}$ to build a calibration set for a new judgment. Let the non-conformity score for a given judgment $y_k^{(t)}$ be its disagreement with the stable judgment: $s(y_k^{(t)}) = \mathbb{I}(y_k^{(t)} \neq y_k^*)$. By applying the conformal prediction framework to this large calibration set of scores, we can construct a prediction interval for a new, unseen judgment. Our empirical analysis on this calibration set reveals that the fraction of judgments deviating from their stable counterpart is exceedingly small. Following the standard procedure for conformal calibration, we can formally certify that for any new judgment y_{new}, the probability of it matching its corresponding stable judgment y_{new}^* is bounded with high confidence. Specifically, for a desired miscoverage rate $\alpha = 0.03$, the procedure yields the following guarantee:

$$P(y_{new} = y_{new}^*) \geq 1 - \alpha = 0.97$$

This result provides a strong probabilistic guarantee that any single pairwise comparison performed by the judge is highly likely to be stable.

B.2 Bounding the Variance of Per-Question Metrics

For each question in our benchmark, the calculation of IPI and TOV scores relies on a set of pairwise comparisons. Given that we generate $n = 6$ candidate answers, a full round-robin evaluation under our symmetrized protocol requires $M = 2 \times \binom{6}{2} = 30$ individual judgments. Our objective is to establish a rigorous, high-confidence upper bound for the variance of the per-question metric, $\text{Var}(\text{TOV}(Q))$, which arises from the LLM judge's inherent stochasticity.

By definition, the variance of the measured score $\text{TOV}(Q)$ is the expected squared difference from its mean:

$$\text{Var}(\text{TOV}(Q)) = \mathbb{E}\left[(\text{TOV}(Q) - \mathbb{E}[\text{TOV}(Q)])^2\right] \tag{7}$$

A fundamental property of variance is that it represents the minimum possible expected squared error. For any constant c, the following inequality holds: $\text{Var}(\text{TOV}(Q)) \leq \mathbb{E}\left[(\text{TOV}(Q) - c)^2\right]$. We can leverage this property by strategically choosing a constant. Let us choose the deterministic stable score, $\text{TOV}^*(Q)$, as our constant c. This yields this inequality:

$$\text{Var}(\text{TOV}(Q)) \leq \mathbb{E}\left[(\text{TOV}(Q) - \text{TOV}^*(Q))^2\right] \tag{8}$$

Let the deviation from the stable score be $\Delta_{\text{TOV}}(Q) = \text{TOV}(Q) - \text{TOV}^*(Q)$. Equation 8 can be rewritten as:

$$\text{Var}(\text{TOV}(Q)) \leq \mathbb{E}[\Delta_{\text{TOV}}(Q)^2] \tag{9}$$

Our task now simplifies to finding an upper bound for the second moment of this deviation.

Let X be the random variable for the total number of unstable judgments among the $M = 30$ trials. As established in Section B.1, the probability p of any single judgment being unstable is bounded by $p \leq \alpha = 0.03$. Assuming independence across judgments, X follows a Binomial distribution, $X \sim \mathcal{B}(M, p)$.

A direct, deterministic relationship connects the score deviation to the number of unstable judgments. Since the TOV score is the minimum number of edge modifications required to resolve all logical contradictions, X unstable judgments can alter the final score by at most X. This gives the inequality $|\Delta_{\text{TOV}}(Q)| \leq X$, which implies:

$$\Delta_{\text{TOV}}(Q)^2 \leq X^2 \tag{10}$$

By taking the expectation, we can chain the inequalities together:

$$\text{Var}(\text{TOV}(Q)) \leq \mathbb{E}[\Delta_{\text{TOV}}(Q)^2] \leq \mathbb{E}[X^2] \tag{11}$$

The second moment of a binomial random variable is given by $\mathbb{E}[X^2] = \text{Var}(X) + (\mathbb{E}[X])^2 = Mp(1-p) + (Mp)^2$. Using $M = 30$ and the upper bound $p = 0.03$, we compute:

$$\mathbb{E}[X] = 30 \times 0.03 = 0.9 \tag{12}$$

$$\text{Var}(X) = 30 \times 0.03 \times (1 - 0.03) = 0.873 \tag{13}$$

Therefore, the second moment of X is:

$$\mathbb{E}[X^2] = 0.873 + (0.9)^2 = 1.683 \tag{14}$$

This directly provides a tight and rigorously derived upper bound for the variance of the per-question TOV score:

$$\text{Var}(\text{TOV}(Q)) \leq 1.683 \tag{15}$$

This result formally demonstrates that the variance of the per-question scores is exceptionally small, confirming that our metrics are highly robust to the inherent stochasticity of LLM judges.

An identical argument holds for the IPI score, yielding a similarly small per-question variance. The IPI score for a question, $\text{IPI}(Q)$, is the fraction of inconsistent pairs. It is calculated over $N = \binom{6}{2} = 15$ unique pairs of answers. Each inconsistent pair contributes 1 to a sum, which is then normalized by N. An unstable judgment can affect the consistency of at most one pair, thus changing the sum by at most 1. Therefore, X unstable judgments can change the sum of inconsistent pairs by at most X. The deviation of the normalized IPI score, $\Delta_{\text{IPI}}(Q)$, is thus bounded by:

$$|\Delta_{\text{IPI}}(Q)| \leq \frac{X}{N} \tag{16}$$

It is worth noting that this inequality can be tightened; since the IPI score is bounded in $[0, 1]$, the maximal deviation is 1, making the true bound $|\Delta_{\text{IPI}}(Q)| \leq \min(X/N, 1)$. By proceeding with the analytically simpler X/N, we are establishing a conservative overestimate for the variance, which strengthens our claim of stability. Following the same logic, we can bound its variance:

$$\begin{aligned}\text{Var}(\text{IPI}(Q)) &\leq \mathbb{E}[\Delta_{\text{IPI}}(Q)^2] \\ &\leq \mathbb{E}\left[\left(\frac{X}{N}\right)^2\right] = \frac{1}{N^2}\mathbb{E}[X^2].\end{aligned} \tag{17}$$

Substituting $N = 15$ and our previously calculated value for $E[X^2]$:

$$\text{Var}(\text{IPI}(Q)) \leq \frac{1.683}{15^2} = \frac{1.683}{225} \approx 0.0075 \quad (18)$$

These results formally demonstrate that the variances of both per-question TOV and IPI scores are exceptionally small, confirming that our metrics are highly robust to the inherent stochasticity of LLM judges.

B.3 Stability of Aggregate Benchmark Scores

The final SAGE metrics are the aggregate scores, IPI and TOV, which are the arithmetic means of the per-question scores over the entire set of $|\mathcal{Q}| = 650$ questions:

$$\text{TOV} = \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \text{TOV}(Q) \quad (19)$$

$$\text{IPI} = \frac{1}{|\mathcal{Q}|} \sum_{Q \in \mathcal{Q}} \text{IPI}(Q) \quad (20)$$

Assuming the scores for each question are independent and identically distributed (i.i.d.) random variables—a standard assumption for a diverse benchmark—the variance of the mean is the per-question variance divided by the number of questions.

Using the upper bound for the per-question TOV variance derived in Section B.2, we can bound the variance of the final aggregate TOV score:

$$\text{Var}(\text{TOV}) = \frac{\text{Var}(\text{TOV}(Q))}{|\mathcal{Q}|} \leq \frac{1.683}{650} \approx 2.59 \times 10^{-3} \quad (21)$$

Similarly, using the upper bound for the per-question IPI variance, we can bound the variance of the final aggregate IPI score:

$$\text{Var}(\text{IPI}) = \frac{\text{Var}(\text{IPI}(Q))}{|\mathcal{Q}|} \leq \frac{0.0075}{650} \approx 1.15 \times 10^{-5} \quad (22)$$

These resulting variances for both aggregate metrics are exceptionally small, indicating that the final reported scores are highly concentrated around their expected values.

In conclusion, this theoretical analysis, grounded in first principles and basic statistical properties, formally demonstrates the robustness of our evaluation framework. The high stability of individual judgments propagates through the metric calculation, resulting in aggregate scores for both IPI and TOV with minimal variance. This theoretical finding is in strong alignment with the empirical results presented in Table 1, confirming that SAGE provides a consistent and reliable methodology for assessing the reasoning capabilities of LLM judges.

C TOV as a Lower Bound on Judgment Error

In this section, we provide a formal proof that the **Weak Total Order Violation (TOV)** metric serves as a theoretical lower bound on the true judgment error rate of an LLM-as-a-Judge system.

C.1 Preliminaries and Definitions

Let Q be a question and $A_Q = \{A_1, A_2, \ldots, A_n\}$ be the set of n candidate answers.

Definition C.1 (Judgment Set)**.** Let $\mathcal{J}$ be the space of all possible pairwise preference sets on n items. A specific judgment set $J \in \mathcal{J}$ consists of pairwise comparisons $y_{ij} \in \{1, 0, -1\}$ for all $1 \leq i < j \leq n$, where:

- $y_{ij} = 1 \implies A_i \succ A_j$
- $y_{ij} = -1 \implies A_i \prec A_j$
- $y_{ij} = 0 \implies A_i = A_j$

Definition C.2 (Hamming Distance / Judgment Error)**.** We define the distance between two judgment sets $J^{(a)}$ and $J^{(b)}$ as the total number of conflicting pairwise preferences. This is equivalent to the Hamming distance:

$$d(J^{(a)}, J^{(b)}) = \sum_{1 \leq i < j \leq n} \mathbb{I}(y_{ij}^{(a)} \neq y_{ij}^{(b)}) \quad (23)$$

where $\mathbb{I}(\cdot)$ is the indicator function.

Definition C.3 (Space of Valid Weak Total Orders)**.** Let $\mathcal{O}_n \subset \mathcal{J}$ be the set of all judgment sets that satisfy the axioms of a *Weak Total Order* (i.e., satisfying reflexivity, antisymmetry, and transitivity, including the handling of ties).

C.2 Formal Problem Statement

Let $J_{obs} \in \mathcal{J}$ be the **observed** set of preferences generated by the LLM judge. Let $J_{gt} \in \mathcal{J}$ be the **ground truth** set of preferences (the "Gold Standard").

The ground truth is logically coherent, so the true quality ranking of answers must form a valid weak total order:

$$J_{gt} \in \mathcal{O}_n \quad (24)$$

The **True Error Count** (E) of the judge is defined as the distance between the observed judgments and the ground truth:

$$E(J_{obs}) = d(J_{obs}, J_{gt}) \tag{25}$$

Recall the definition of **Weak Total Order Violation (TOV)** from the main text. TOV is defined as the minimum number of alterations required to transform J_{obs} into *any* valid weak total order:

$$\text{TOV}(J_{obs}) = \min_{O \in \mathcal{O}_n} d(J_{obs}, O) \tag{26}$$

C.3 Theorem and Proof

Theorem C.4. *For any observed judgment set J_{obs} and any logically consistent ground truth J_{gt}, the Weak Total Order Violation score is a lower bound on the True Error Count.*

$$\textit{TOV}(J_{obs}) \leq E(J_{obs}) \tag{27}$$

Proof. The proof follows directly from the definition of the minimum operator over a set.

1. By definition, $\text{TOV}(J_{obs})$ is the distance from J_{obs} to the *nearest* valid weak total order in the set $\mathcal{O}_n$:

$$\text{TOV}(J_{obs}) = \inf\{d(J_{obs}, O) \mid O \in \mathcal{O}_n\} \tag{28}$$

2. From our assumption of logical consistency, the ground truth is a valid weak total order:

$$J_{gt} \in \mathcal{O}_n \tag{29}$$

3. Since J_{gt} is an element of the set $\mathcal{O}_n$, the distance to J_{gt} must be greater than or equal to the minimum distance to the set $\mathcal{O}_n$. Mathematically:

$$\min_{O \in \mathcal{O}_n} d(J_{obs}, O) \leq d(J_{obs}, J_{gt}) \tag{30}$$

4. Substituting the definitions of TOV and Error E:

$$\text{TOV}(J_{obs}) \leq E(J_{obs}) \tag{31}$$

□

Corollary. This implies that if a model has a TOV score of k, it strictly implies that the model has made *at least* k errors relative to the ground truth, regardless of what the specific ground truth is. Thus, TOV serves as a model-agnostic, unsupervised proxy for the minimum error rate.

D Detailed Experiment Setups

D.1 Dataset Curation

The curation process for our benchmark's dataset is meticulously designed to ensure both diversity and representativeness, as illustrated in Figure 7. We began by drawing questions from two distinct, high-quality sources. First, we extracted questions from five core categories within the RewardBench2 dataset—namely Factuality, Focus, Precise Instruction Following, Mathematics, and Safety—to establish a foundation of structured evaluation problems. These questions are manually selected to ensure semantic uniqueness. To complement this and incorporate more natural, real-world user interactions, we also sourced a large volume of queries from the WildChat-1m corpus, which contains logs of human-LLM conversations. These queries underwent a rigorous screening process, including both large-scale automated filtering and manual review, to select for relevance and clarity. The questions from both sources are then merged to form the final, comprehensive set of 650 questions. This dual-source approach ensures that our benchmark covers a wide semantic space, balancing formal assessment criteria with the unpredictability of genuine user inquiries, which is essential for a robust evaluation of LLM judges.

D.2 CV Distribution

Figure 8 shows the CV distribution of our constructed dataset.

D.3 Spearman Rank Correlation Coefficient

The Spearman Rank Correlation Coefficient, commonly denoted by ρ or r_s, is a non-parametric measure of rank correlation (statistical dependence between the rankings of two variables). It assesses how well the relationship between two variables can be described using a monotonic function. Unlike the Pearson correlation, which assumes a linear relationship and normally distributed data, Spearman's correlation evaluates the monotonic relationship, making it more robust to outliers and non-linear associations commonly found in LLM evaluation scores. The coefficient's value is constrained to the interval $[-1, 1]$.

D.3.1 Interpretation of the Coefficient

The value of the Spearman correlation coefficient (r_s) is interpreted as follows:

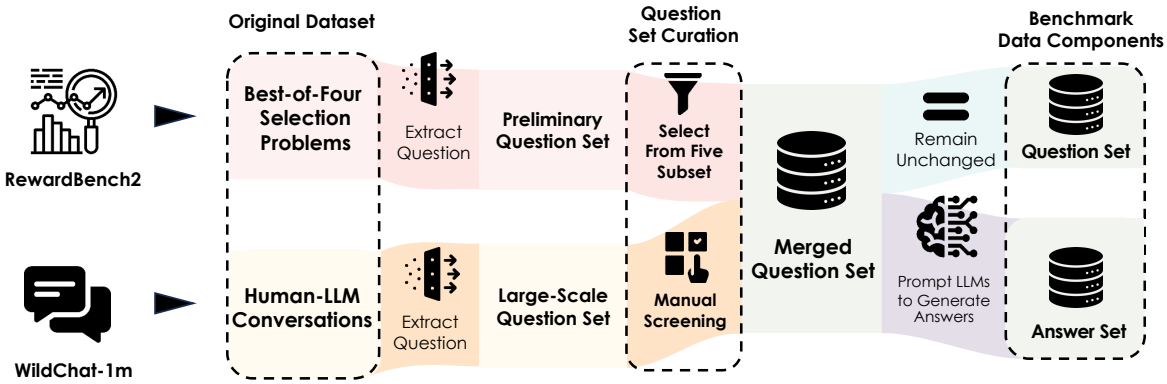


Figure 7: Curation of our dataset.

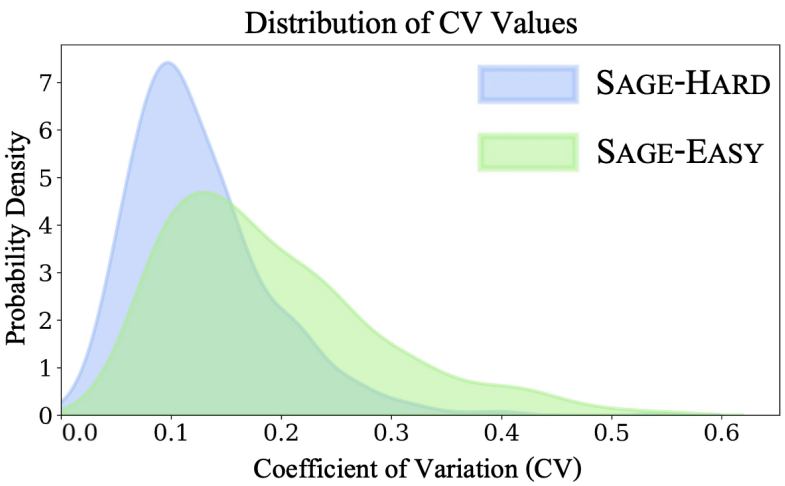


Figure 8: Distribution of CV Values.

- $r_s = +1$: Indicates a perfect positive monotonic relationship. As the rank of one variable increases, the rank of the other variable increases consistently. This implies that the judge model's ranking of answers perfectly matches the reference ranking, even if the absolute score intervals differ.
- $r_s = -1$: Indicates a perfect negative monotonic relationship. As the rank of one variable increases, the rank of the other variable decreases consistently.
- $r_s = 0$: Indicates no monotonic tendency between the two variables. The rankings are essentially random with respect to one another.

The magnitude of $|r_s|$ indicates the strength of the association.

D.3.2 Mathematical Formulation

The Spearman correlation coefficient is defined as the Pearson correlation coefficient between the rank variables.

For a sample of size n, the n raw scores X_i and Y_i are converted to ranks $rg(X_i)$ and $rg(Y_i)$. If there are tied ranks, the average of the ranks that would have been assigned to all the tied values is assigned to each value. The formula is given by:

$$r_s = \frac{\text{cov}(rg_X, rg_Y)}{\sigma_{rg_X} \sigma_{rg_Y}} \tag{32}$$

where:

- rg_X and rg_Y represent the rank variables of the raw data X and Y.
- $\text{cov}(rg_X, rg_Y)$ is the covariance of the rank variables.
- σ_{rg_X} and σ_{rg_Y} are the standard deviations of the rank variables.

D.4 Coefficient of Variation

The Coefficient of Variation (CV) is a standardized statistical measure of the relative dispersion of a data distribution. Unlike the standard deviation, which quantifies absolute variability, the CV

expresses the standard deviation as a fraction of the arithmetic mean. This normalization renders the CV a dimensionless quantity, thereby facilitating the comparison of variability across datasets with different units of measurement or significantly different means.

For a population, the Coefficient of Variation is defined as the ratio of the standard deviation (σ) to the mean (μ), provided that the mean is non-zero:

$$\text{CV} = \frac{\sigma}{|\mu|}$$

For a sample, the CV is estimated using the sample standard deviation (s) and the sample mean ($\bar{x}$):

$$c_v = \frac{s}{|\bar{x}|}$$

The absolute value of the mean is often used in the denominator to ensure the CV remains non-negative and is well-defined for negative means, preserving its interpretation as a measure of variability magnitude.

The primary utility of the CV lies in its capacity to provide a relative measure of consistency or homogeneity. A lower CV indicates less variability relative to the mean, suggesting greater consistency within the data. Conversely, a higher CV signifies greater relative dispersion. This property is particularly advantageous when comparing the degree of variation between two or more groups of data. For instance, comparing the standard deviation of prices in two different currencies is not directly meaningful; however, their Coefficients of Variation can be compared to determine which currency's price level is relatively more stable, as it is a unit-free metric.

D.5 Models

Large Language Models. An LLM is an advanced AI model, typically using a Transformer architecture, trained on massive text data to understand and generate natural language by predicting the next token. Pre-trained on broad datasets, they can be fine-tuned for specific tasks. Their large scale, with billions of parameters, results in strong generalization and emergent abilities for diverse tasks like text generation, summarization, translation, and question answering. The detailed information about the models we used in our experiments is as follows:

- **DeepSeek-R1-0528** (DeepSeek-AI, 2025a): DeepSeek-R1-0528 is a 671B sparse Mixture-of-Experts (MoE) model with 37B active parameters and a 128K context length. Built upon DeepSeek-V3-Base, it is trained using reinforcement learning to enhance its capabilities in complex reasoning, mathematics, and coding.
- **DeepSeek-V3-0324** (DeepSeek-AI, 2024): DeepSeek-V3-0324 is a 671B Mixture-of-Experts (MoE) model with 37B active parameters per token. Trained on a 14.8T-token dataset, it uses optimized attention and advanced expert routing to enhance performance on complex reasoning and coding tasks with computational efficiency.
- **DeepSeek-V3.1** (DeepSeek-AI, 2025b): DeepSeek-V3.1 is a 671B Mixture-of-Experts (MoE) model that activates 37B parameters per token. It features a hybrid architecture for reasoning and fast responses, supports a 128K context window, and is post-trained for tool-calling and agentic tasks.
- **Gemini-2.0-Flash-Lite** (Google, 2025): Gemini-2.0-Flash-Lite is a lightweight, multimodal Google model for high-speed, high-volume tasks where latency and cost are critical. This smaller, faster variant excels at summarization and chat, ideal for scalable services and on-device applications requiring rapid, resource-efficient inference.
- **Gemini-2.5-Flash** (Team, 2025a): Gemini-2.5-Flash is a cost-efficient, multimodal foundation model by Google DeepMind with a 1 million context window. It uses a sparse Mixture-of-Experts (MoE) architecture to balance performance, cost, and latency, and is optimized for speed in reasoning and multimodal tasks.
- **Gemini-2.5-Pro** (Team, 2025a): Gemini-2.5-Pro is Google DeepMind's top-tier "thinking" multimodal model optimized for advanced reasoning, coding, science, and long-form tasks. It supports text, images, audio, video, and even code repository inputs, with a very large context window of 1 million tokens and default maximum output around 65,535 tokens. The architecture is a sparse Mixture-of-Experts transformer, dynamically routing tokens to experts so capacity is decoupled from inference compute. It leads many benchmarks in mathematics, science, and coding, offering high accuracy but at greater computational cost and latency than lighter variants.
- **GPT-4o-Mini** (OpenAI, 2024): GPT-4o-Mini is a compact, cost-efficient variant of OpenAI's

GPT-4o model, released in July 2024. It offers strong language and vision capabilities with lower latency and supports a 128K token context window for handling long inputs.

- **GPT-5-Chat** (OpenAI, 2025): GPT-5-Chat (OpenAI, August 2025) is a flagship, multimodal conversational model that unifies fast responses with deep reasoning. It supports long context and multi-step tool calling, featuring improved code quality, reduced hallucinations, and enhanced steerability.
- **Llama-3.1-8B-Instruct** (Meta, 2024a): Llama-3.1-8B-Instruct is an 8-billion-parameter multilingual instruction-tuned autoregressive transformer released by Meta. It features a 128K token context window and is fine-tuned for instruction following, dialogue, reasoning, and translation.
- **Claude-3-Haiku** (Anthropic, 2024): Claude-3-Haiku, part of Anthropic's Claude 3 family, is optimized for speed and cost-effectiveness in lighter tasks. It supports a 200K token context window for text and image inputs, delivering fast, responsive generation, though its benchmark scores are lower than the more capable Sonnet or Opus models.
- **Qwen3-4B-Instruct-2507** (Team, 2025b): Qwen3-4B-Instruct-2507 is a compact language model with 4 billion parameters, optimized for instruction-following and multilingual tasks. It supports a 256K token context window and provides fast, efficient responses for real-time applications.
- **Qwen3-30B-A3B-Instruct-2507** (Team, 2025b): Qwen3-30B-A3B-Instruct-2507 is a sparse Mixture-of-Experts (MoE) instruction-tuned model with 30.5B total and 3.3B active parameters. It uses 128 experts (8 active per token), supports a 262,144-token context window, and is tuned for instruction following, multilingual understanding, reasoning, coding, and tool use.
- **Qwen3-235B-A22B-Instruct-2507** (Team, 2025b): Qwen3-235B-A22B-Instruct-2507 is a 235B parameter Mixture-of-Experts (MoE) instruction-tuned model that activates 22B parameters per inference. It supports a 256K context length, features 128 experts (activating 8 per token), and uses Grouped-Query Attention. The model is improved for instruction-following, reasoning, math, and coding.
- **Qwen2.5-3B-Instruct** (Team, 2024b): Qwen2.5-3B-Instruct is a 3.09B-parameter, instruction-tuned causal language model. It features a 36-layer transformer with Grouped-Query Attention, RoPE, SwiGLU, and RMSNorm. This multilingual model supports a 32k-token context and shows strengths in instruction following, structured output, mathematics, and coding.
- **Qwen2.5-7B-Instruct** (Team, 2024b): Qwen2.5-7B-Instruct is a 7.6B-parameter instruction-tuned causal transformer from Alibaba. It features RoPE, SwiGLU, and GQA, with a context window of up to 131k tokens. The model is multilingual and excels in instruction following, coding, and math.
- **Mistral-7B-Instruct-V0.3** (Jiang et al., 2023): Mistral-7B-Instruct-V0.3 is a 7.3B-parameter causal transformer by Mistral AI, fine-tuned for instruction following. It features a v3 tokenizer, a 32k token vocabulary, a 32k token context window, and supports function calling, delivering fast inference.

Fine-tuned Judges. A fine-tuned judge is a Large Language Model specialized to evaluate text quality. It is further trained on a dataset containing generated text and corresponding human preference labels, such as comparisons or scores. This process aligns the model with human evaluators' standards, allowing it to learn the nuances and criteria they value. Consequently, a fine-tuned Judge serves as a more reliable automated evaluation tool, producing judgments that more closely resemble those of human experts than a general-purpose model.

- **Prometheus-7B-V2.0** (Kim et al., 2024): A 7-billion-parameter open-source evaluator LLM built on Mistral-Instruct. Trained on 100K "Feedback Collection" examples and 200K preference/ranking pairs, it supports both absolute grading (direct assessment) and relative grading (pairwise ranking) tasks.
- **M-Prometheus-3B** (Pombal et al., 2025): M-Prometheus-3B is a 3-billion-parameter multilingual LLM evaluator from Unbabel, built upon the Qwen2.5-3B architecture. Trained on 480K instances of multilingual data, it provides both direct assessment and pairwise comparison feedback. The model has demonstrated superior performance on multilingual meta-evaluation benchmarks and in literary translation evaluation.

- **M-Prometheus-7B** (Pombal et al., 2025): M-Prometheus-7B is a 7-billion-parameter multilingual evaluator model from Unbabel, fine-tuned from Qwen2.5-Instruct. Trained on 480,000 instances of multilingual assessment and comparison data, it supports both absolute and relative grading.
- **Skywork-Critic-Llama-3.1-8B** (Tu et al., 2024): Skywork-Critic-Llama-3.1-8B is an 8-billion-parameter preference evaluator from the SkyworkAI Alignment Team, fine-tuned from Meta's Llama-3.1-8B-Instruct. Trained on a curated dataset, it evaluates the relative quality of text responses for data improvement, evaluation, and reward modeling.
- **JudgeLRM-3B** (Chen et al., 2025): JudgeLRM-3B is a 3-billion-parameter, judgment-oriented language model. Built on a Qwen2.5-3B-Instruct base and trained with reinforcement learning (GRPO), it is designed for complex reasoning tasks. The model demonstrates superior performance by surpassing GPT-4 on judgment benchmarks like JudgeLM and PandaLM and significantly outperforming similarly-sized SFT models.
- **JudgeLRM-7B** (Chen et al., 2025): JudgeLRM-7B is a language model built upon Qwen2.5-7B-Instruct. It utilizes Group Relative Policy Optimization (GRPO), a reinforcement learning method, to enhance complex reasoning. The model demonstrates superior performance on reasoning benchmarks, outperforming GPT-4 on specific tasks and significantly surpassing other similarly-sized models.

Multi-Agent Judges. Multi-Agent Judges is an evaluation framework using multiple autonomous Large Language Models (LLMs) to assess text quality. Instead of a single LLM, this method involves either a group of LLM agents debating to form a collective judgment or independently scoring an output, with the scores then aggregated. The goal is to reduce the bias and variance of single-agent evaluation, aiming for more robust and reliable assessments that better align with human preferences.

- **ChatEval** (Chan et al., 2024): ChatEval is a debate-based framework using a *"referee team"* of multiple LLM agents to simulate human collaborative evaluation. Each agent is assigned a unique persona to ensure diverse perspectives. These agents autonomously debate the quality of a text over multiple turns, guided by communication strategies. The final evaluation aggregates individual judgments after the debate, such as by majority vote or averaging scores, rather than forcing a consensus.
- **PoLL** (Verga et al., 2024): The *"Panel of LLM evaluators (PoLL)"* is a multi-agent method using a diverse group of LLMs to independently assess text generations, similar to a jury. The individual scores are then aggregated into a final judgment. This approach aims to reduce the bias, cost, and variance of using a single LLM for evaluation.

E Arena Hard Auto

E.1 Evaluation Process

The Arena-Hard-Auto evaluation process (Li et al., 2024) is based on a pairwise comparison framework (Chiang et al., 2024). For every prompt in the benchmark, the response from the model being evaluated is compared against the response from a fixed, strong baseline model (Zheng et al., 2023; Liu et al., 2023). In our experiment we use the Gemini-2.5-Pro (Team, 2025a) as the baseline model. This comparison is mediated by an LLM-as-a-Judge. To ensure a high-quality and consistent assessment, the judge model is first prompted to generate its own ideal solution directly. It then evaluates the two models' responses, rating the preference on a 5-point Likert scale to capture the degree of superiority (Newman, 2023). To mitigate potential positional bias (Shi et al., 2024), where a judge might favor the first or second answer presented, the entire evaluation for a single prompt is conducted twice in a two-game setup, with the positions of the model outputs swapped in the second round.

E.2 Scores Calculation

After collecting all pairwise judgments, the Bradley-Terry model is employed to compute a final, continuous score for each model. This statistical model aggregates the outcomes of thousands of individual head-to-head comparisons against the baseline. It works by estimating a latent *"strength"* parameter for each model, effectively converting the discrete win/loss/tie results from the Likert scale judgments into a single, comprehensive score. This score represents

the model's overall performance and capability across the diverse and challenging prompts of the benchmark, allowing for a quantitative and ordered ranking of all evaluated models.

E.3 Model Performance Evaluation

To precisely quantify the final score and its range of uncertainty for each evaluated model, a bootstrapping methodology is employed. This statistical process involves repeatedly resampling the entire set of pairwise judgments with replacement to create thousands of new, simulated datasets. For each of these bootstrapped datasets, a win-rate against the baseline is recalculated for every model. This generates a distribution of potential win-rates, from which a final average score and a 95% confidence interval is derived (Efron, 1992). This confidence interval represents the *"floating range"* of the model's performance, indicating the score's stability and statistical reliability.
Furthermore, in our experiments, this process is extended to assess and compare the robustness of different models when they serve as the judge. To achieve this, a specific model is designated as the judge and is used to evaluate a standard set of other models against the baseline. The bootstrapping process is then carried out to determine the confidence interval for each of the evaluated models. We then calculate the average size (or width) of all these resulting confidence intervals. This value, the *"average confidence interval,"* serves as a single metric to quantify the judge's consistency. A smaller average confidence interval indicates that the judge model is more stable and reliable, as its evaluations produce less variance and lead to more precise performance estimates.

F Additional Result

F.1 Positional Bias in LLM-as-a-Judge

To substantiate that positional bias does exist in SAGE, we sample 1120 answer pairs and measure the **inconsistency rate** for Llama3-8B-Instruct (Team, 2024a), Gemini-2.5-Flash-Lite (Team, 2025a), and Qwen3-4B-Instruct-2507 (Team, 2025b).

We define this rate as the frequency of judgments that are not the logical inverse when the answer order is reversed (i.e., $J_M(Q, A_i, A_j) \neq -J_M(Q, A_j, A_i)$). The results in Table 7 confirm the presence of bias.

Table 7: Local inconsistency (*i.e.*, Positional Bias) across LLM-as-a-Judge.

Model	Inconsistency (%)
Llama3-8B-Instruct	76.2
Gemini-2.5-Flash-Lite	25.3
Qwen-3-4B-Instruct-2507	44.4

F.2 Ablation Study of ChatEval

To further prove that the negative ChatEval result is not configuration-dependent, we conduct ablation experiments on the ChatEval framework. We investigated the impact of three key hyperparameters (num-rounds, argument exchange format, judge selection):

- Number of Rounds: Vary the num-rounds configuration from 2 rounds down to 1 round.
- Judge Selection: Change the judge selection mechanism from our original [Critic, Psychologist, Scientist] to [Critic, General Public, Scientist].
- Argument Exchange Format: Change the exchange format from One-by-One to Simultaneous-Talk.

Our results show that *all* alternative configurations still performed significantly worse than the non-debate baseline, yielding higher (worse) scores for both IPI and TOV. The results are shown in Figure 9. A detailed case study explaining why the degradation happens is available in Appendix G.6.

F.3 Multi-Agent Judge Configurations

In our investigation of multi-agent judging strategies, we deployed two distinct methodologies to evaluate whether collaboration enhances judgment reliability. **Panel-based Approach (POLL).** We adopted the POLL methodology (Verga et al., 2024), which aggregates the decisions of independent LLM judges without inter-agent communication. We constructed two separate juries to assess performance across different capability levels. The **Powerful Model Panel** consists of three state-of-the-art models: Gemini-2.5-Pro, Gemini-2.5-Flash, and Qwen3-235B-A22B-Instruct-2507. For comparison, the **Weaker Model Panel** comprises Gemini-2.0-Flash-Lite, GPT-4o-mini, and DeepSeek-V3.1. In both configurations, the final

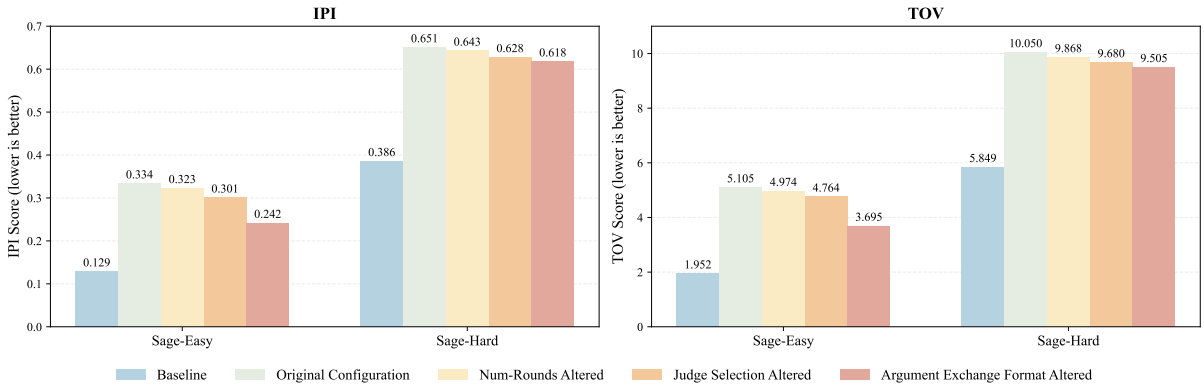


Figure 9: Ablation results for ChatEval showing performance degradation across all configuration variants (lower scores are better).

verdict is determined by a majority vote among the panel members.

Debate-based Approach (ChatEval). We utilized the ChatEval framework (Chan et al., 2024) to simulate a multi-round debate. Unlike POLL, ChatEval employs a single underlying LLM to generate multiple distinct agents. Our standard configuration features a three-agent lineup acting as a *Critic*, a *Psychologist*, and a *Scientist*. These agents engage in an argumentative dialogue using a "One-by-One" exchange format to discuss the quality of answer pairs before reaching a conclusion.

F.4 Metric Consistency across Temperatures

As discussed in the main text, we conduct experiments to verify the stability of our proposed metrics against the stochasticity inherent in LLM outputs. Table 8 details the performance of two models, Qwen3-4B-Instruct-2507 and Qwen3-30B-A3B-Instruct-2507, under five different temperature settings.

The results show that both the Intra-Pair Instability (IPI) and Weak Total Order Violation (TOV) scores remain stable across all temperatures. This low variance demonstrates the robustness of our evaluation framework, confirming that the metrics capture consistent aspects of a model's judgment capabilities rather than random artifacts of the generation process.

F.5 The performance of Fine-tuned Judges on SAGE-EASY

Table 9 demonstrates the performance of fine-tuned judges on SAGE-EASY, which shows that fine-tuning does enhance judgment robustness.

F.6 Cost Analysis

In this section, we provide a breakdown of the cost and time estimates comparing our automated framework, SAGE, with traditional human annotation. This analysis substantiates our claim regarding the elimination of the annotation bottleneck.

Computational Cost for SAGE. A complete evaluation cycle for a single dataset in SAGE involves a set of $|\mathcal{Q}| = 650$ questions. For each question, we generate $n = 6$ candidate answers, resulting in $\binom{6}{2} = 15$ unique pairs per question. Under our symmetrized evaluation protocol (checking both A vs B and B vs A), the total number of judgments required is:

$$N_{\text{total}} = |\mathcal{Q}| \times \binom{n}{2} \times 2 = 650 \times 15 \times 2 = 19500$$

Leveraging efficient commercial models (e.g., pricing at approximately \$0.15 per 1M input tokens), the total financial cost for a full SAGE run is less than \$7 USD. With moderate concurrency, this entire evaluation process can be completed in under one hour.

Estimated Cost for Human Annotation. Replicating this scale with human experts is notoriously expensive. We base our estimation on the SAGE-HARD subset, where the average input length reaches 1911 tokens. Adopting market rates from Wang et al. (2021) and Xu et al. (2025) (unit cost: \$0.0022 per token; labeling speed: ≈ 260 tokens per minute), we derive the following estimates:

- **Financial Cost:** $19,500$ judgments $\times$ $1,911$ tokens $\times$ \$0.0022 $\approx$ \$81,981 USD.
- **Time Consumption:** $19500 \times 1911/260 \approx 143325$ min ≈ 100 days.

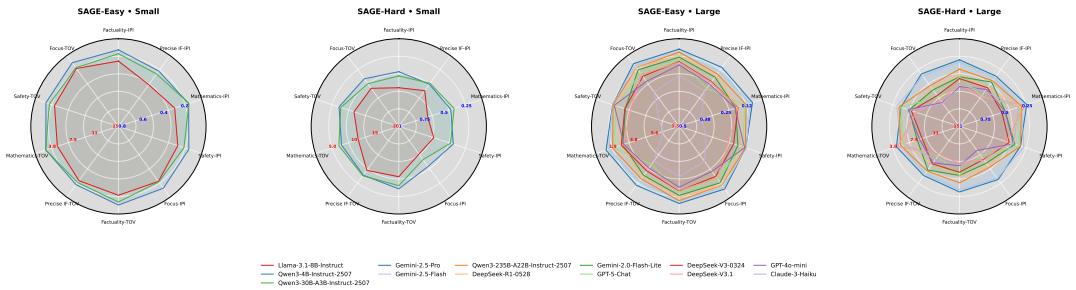


Figure 10: Comparison of radar charts for different models.

Table 8: IPI and TOV scores at varying temperatures on SAGE. (T for temperature)

Models	Benchmark	Metric	T=0.1	T=0.3	T=0.5	T=0.7	T=0.9
Qwen3-4B-Instruct-2507	SAGE-EASY	IPI	0.123	0.133	0.126	0.129	0.127
		TOV	1.890	2.019	1.917	1.970	1.928
	SAGE-HARD	IPI	0.379	0.378	0.379	0.378	0.379
		TOV	5.954	5.929	5.970	5.933	5.980
Qwen3-30B-A3B- Instruct-2507	SAGE-EASY	IPI	0.160	0.161	0.164	0.164	0.162
		TOV	2.410	2.425	2.463	2.470	2.437
	SAGE-HARD	IPI	0.405	0.414	0.409	0.412	0.415
		TOV	6.302	6.427	6.376	6.421	6.439

Table 9: The performance of fine-tuned models and their base models on SAGE-EASY.

Models	Factuality		Precise IF		Mathematics		Safety		Focus		Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
Qwen2.5-3B-Instruct (Base)	0.562	8.518	0.504	7.686	0.450	7.025	0.468	7.382	0.661	10.000	0.530	8.131
M-Prometheus-3B	0.253	3.844	0.256	3.926	0.249	3.848	0.243	3.727	0.220	3.313	0.245(↑54%)	3.735(↑54%)
JudgeLRM-3B	0.640	9.596	0.498	7.493	0.323	4.915	0.313	4.943	0.778	11.669	0.515(↑3%)	7.798(↑4%)
Qwen2.5-7B-Instruct (Base)	0.440	6.601	0.361	5.414	0.359	5.475	0.373	5.610	0.617	9.250	0.429	6.455
M-Prometheus-7B	0.228	3.427	0.192	2.943	0.203	3.092	0.199	3.073	0.244	3.661	0.213(↑50%)	3.239(↑50%)
JudgeLRM-7B	0.385	5.828	0.316	4.792	0.367	5.746	0.283	4.513	0.490	7.373	0.369(↑14%)	5.642(↑13%)
Mistral-7B-Instruct (Base)	0.442	6.644	0.376	5.733	0.492	7.435	0.226	3.431	0.450	6.754	0.399	6.027
Prometheus-7B-V2.0	0.338	5.236	0.373	5.824	0.421	6.673	0.337	5.258	0.398	6.066	0.368(↑8%)	5.718(↑5%)
Llama-3.1-8B-Instruct (Base)	0.204	3.203	0.229	3.506	0.260	4.000	0.230	3.493	0.179	2.755	0.221	3.410
Skywork-Critic-Llama-3.1-8B	0.118	1.776	0.171	2.564	0.086	1.283	0.127	1.959	0.115	1.718	0.125(↑43%)	1.879(↑45%)

These figures highlight the immense disparity in resources required: over $80k vs. $7, which validates SAGE as a highly scalable solution.

F.7 Consistency Analysis Across Judge Formats

To validate whether LLM judges possess a stable internal valuation scale, we extend our analysis beyond the standard pairwise protocol to include two alternative formats: Direct Scoring and Batch Ranking. This section details the experimental settings and the result of consistency between different formats.

We investigate the alignment of the SAGE pairwise baseline against two common evaluation paradigms using the same set of questions and answers.

Direct Scoring (Point-wise): The judge model is prompted to evaluate a single answer A_i independently and assign a scalar score $S(A_i)$ based on its quality.

Batch Ranking (List-wise): The judge model is presented with the full set of $n = 6$ candidate answers simultaneously and is instructed to output a ranking string that represents a partial order (e.g., Answer 1 $\succ$ Answer 2 = Answer 3 $\succ$ Answer 4). Crucially, this format explicitly allows the model

Table 10: Consistency rates between *direct scoring* and *pairwise comparison* on SAGE. All models exhibit widespread inconsistency, particularly on SAGE-HARD.

Models	Factuality↑	Focus↑	Mathematics↑	Precise IF↑	Safety↑	Overall↑
Consistency Rates on SAGE-EASY						
GPT-5-Chat	**0.843**	**0.837**	**0.735**	0.799	0.654	**0.777**
GPT-4o-mini	0.723	0.686	0.730	0.693	**0.671**	0.702
Qwen3-235B-A22B-Instruct-2507	0.808	0.774	0.720	0.773	0.384	0.695
Gemini-2.5-Pro	0.785	0.748	0.719	**0.802**	0.344	0.685
DeepSeek-R1-0528	0.750	0.695	0.714	0.712	0.511	0.679
DeepSeek-V3.1	0.725	0.686	0.679	0.724	0.532	0.673
Gemini-2.5-Flash	0.756	0.727	0.690	0.767	0.307	0.652
Qwen3-30B-A3B-Instruct-2507	0.728	0.704	0.673	0.710	0.321	0.631
Llama-3.1-8B-Instruct	0.669	0.585	0.583	0.657	0.555	0.615
Qwen3-4B-Instruct-2507	0.694	0.677	0.643	0.637	0.358	0.605
Gemini-2.0-Flash-Lite	0.641	0.531	0.600	0.666	0.295	0.552
DeepSeek-V3-0324	0.567	0.458	0.612	0.613	0.392	0.531
Claude-3-Haiku	0.503	0.459	0.450	0.420	0.477	0.462
Consistency Rates on SAGE-HARD						
GPT-5-Chat	**0.434**	0.350	0.480	0.416	0.392	**0.415**
DeepSeek-V3.1	0.383	**0.431**	0.500	0.359	0.360	0.405
Llama-3.1-8B-Instruct	0.403	0.343	0.356	**0.471**	0.404	0.403
Claude-3-Haiku	0.346	0.330	0.448	0.299	**0.526**	0.384
DeepSeek-R1-0528	0.351	0.313	0.374	0.377	0.341	0.351
GPT-4o-mini	0.282	0.277	0.396	0.391	0.349	0.338
Qwen3-235B-A22B-Instruct-2507	0.300	0.260	**0.539**	0.345	0.235	0.332
Gemini-2.5-Pro	0.241	0.209	0.507	0.390	0.294	0.325
Gemini-2.5-Flash	0.273	0.205	0.484	0.384	0.236	0.315
Gemini-2.0-Flash-Lite	0.232	0.184	0.417	0.312	0.444	0.313
Qwen3-30B-A3B-Instruct-2507	0.232	0.215	0.500	0.300	0.205	0.288
Qwen3-4B-Instruct-2507	0.238	0.200	0.412	0.286	0.308	0.286
DeepSeek-V3-0324	0.142	0.102	0.329	0.193	0.258	0.202

to declare ties between answers of perceived equal quality.

To measure consistency, we convert the absolute scores from Direct Scoring into pairwise preferences. For any pair $\{A_i, A_j\}$, if the directly assigned scores satisfy $S(A_i) > S(A_j)$, we infer a preference for A_i. We then check if this inferred preference aligns with the explicit judgment $J(A_i, A_j)$ obtained from the pairwise evaluation. The consistency rate is defined as the percentage of pairs where the derived relationship matches the explicit pairwise verdict. The results are shown in Table 10.

Similarly, for Batch Ranking, we decompose the model's output string into $\binom{n}{2}$ pairwise relationships. For any two answers A_i and A_j, we infer their relationship based on their relative positions in the ranked list: if A_i is placed in a higher tier than A_j, we infer a preference $A_i \succ A_j$; if they are placed in the same tier (connected by an equality sign), we infer a tie $A_i = A_j$. We then calculate the consistency rate by comparing these rank-derived relationships against the explicit judgments obtained from the standard pairwise evaluation. This comparison reveals whether the model's adjudicative logic remains consistent when the context window is loaded with multiple competing options versus a focused binary comparison. The results are shown in Table 11.

As evidenced by the results in Table 10 and Table 11, we observe a substantial degree of inconsistency across these formats, indicating that judge models struggle to maintain coherent preferences when the evaluation interface changes.

F.8 Comprehensive Results of the Performance of LLM Judges under a Direct Judgment Protocol

Figure 11 shows the comprehensive result with diverging bar chart. Table 12 and 13 show the comprehensive numerical results of the performance of LLM Judges under a direct judgment protocol.

Table 11: Consistency rates between *batch ranking* and *pairwise comparison* on SAGE. All models exhibit widespread inconsistency

Models	Factuality↑	Focus↑	Mathematics↑	Precise IF↑	Safety↑	Overall↑
Consistency Rates on SAGE-EASY						
Gemini-2.5-Pro	0.887	0.893	0.865	**0.866**	**0.887**	**0.880**
Gemini-2.5-Flash	0.887	**0.907**	0.863	0.854	0.886	0.879
Qwen3-235B-A22B-Instruct-2507	0.887	0.904	**0.868**	0.831	0.867	0.871
GPT-5-Chat	**0.888**	0.892	0.854	0.833	0.880	0.869
DeepSeek-R1-0528	0.860	0.858	0.831	0.812	0.859	0.846
DeepSeek-V3.1	0.850	0.866	0.821	0.802	0.815	0.832
DeepSeek-V3-0324	0.804	0.814	0.760	0.780	0.789	0.790
Gemini-2.0-Flash-Lite	0.803	0.806	0.779	0.788	0.770	0.790
Qwen3-30B-A3B-Instruct-2507	0.818	0.812	0.782	0.769	0.733	0.785
Qwen3-4B-Instruct-2507	0.794	0.809	0.779	0.713	0.745	0.768
GPT-4o-mini	0.769	0.718	0.771	0.723	0.783	0.751
Claude-3-Haiku	0.616	0.606	0.523	0.613	0.533	0.582
Llama-3.1-8B-Instruct	0.309	0.345	0.473	0.374	0.437	0.386
Consistency Rates on SAGE-HARD						
Gemini-2.5-Pro	**0.697**	**0.656**	0.705	**0.667**	0.682	**0.681**
Gemini-2.5-Flash	0.671	0.649	0.644	0.644	0.722	0.666
GPT-5-Chat	0.620	0.580	0.698	0.593	**0.731**	0.642
Qwen3-235B-A22B-Instruct-2507	0.602	0.593	0.731	0.617	0.652	0.637
DeepSeek-R1-0528	0.573	0.563	0.680	0.604	0.650	0.613
DeepSeek-V3.1	0.532	0.508	**0.786**	0.503	0.579	0.578
Gemini-2.0-Flash-Lite	0.537	0.490	0.571	0.580	0.636	0.562
GPT-4o-mini	0.553	0.513	0.520	0.545	0.555	0.537
Qwen3-4B-Instruct-2507	0.495	0.434	0.573	0.528	0.567	0.518
Qwen3-30B-A3B-Instruct-2507	0.497	0.448	0.655	0.516	0.425	0.509
DeepSeek-V3-0324	0.501	0.446	0.498	0.510	0.526	0.497
Claude-3-Haiku	0.467	0.402	0.452	0.506	0.616	0.487
Llama-3.1-8B-Instruct	0.280	0.263	0.334	0.315	0.292	0.300

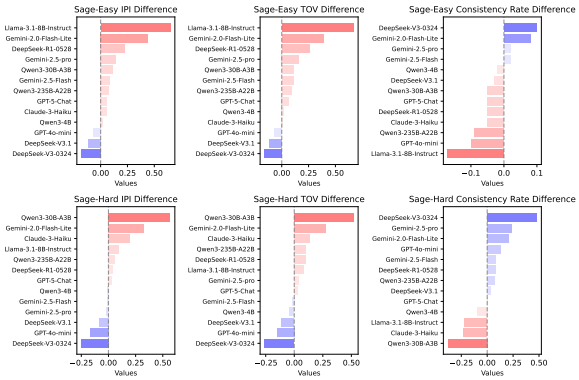


Figure 11: The difference of IPI, TOV and consistency rate when models output verdicts directly in comparison to the original conditions. Red bars indicate degradation and blue bars indicate improvement.

Table 12: Performance of LLM judges under a direct judgment protocol, it can be seen that while removing explanatory reasoning generally degrades robustness on SAGE-EASY (increasing IPI/TOV), the impact on SAGE-HARD is mixed: weaker models suffer significant instability, whereas some state-of-the-art models maintain or even slightly improve their consistency without explicit reasoning.

Models	Factuality		Precise IF		Mathematics		Safety		Focus		Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
Performance on SAGE-EASY												
Gemini-2.5-Pro	**0.064**	**0.993**	**0.091**	**1.367**	**0.071**	**1.135**	0.123	1.942	**0.062**	**0.927**	**0.082** (↓ 14%)	**1.265** (↓ 16%)
Gemini-2.5-Flash	0.077	1.175	0.133	2.043	0.082	1.305	0.105	1.667	0.075	1.137	0.095 (↓ 9%)	1.471 (↓ 11%)
Qwen3-235B-A22B-Instruct-2507	0.077	1.175	0.117	1.761	0.150	2.310	**0.101**	**1.626**	0.091	1.374	0.106 (↓ 8%)	1.626 (↓ 9%)
Qwen3-4B-Instruct-2507	0.110	1.664	0.151	2.288	0.166	2.492	0.130	1.992	0.090	1.347	0.129 (↓ 2%)	1.952 (↓ 2%)
DeepSeek-V3-0324	0.105	1.601	0.141	2.108	0.194	3.058	0.115	1.821	0.094	1.417	0.129 (↑ 18%)	1.989 (↑ 16%)
DeepSeek-V3.1	0.107	1.645	0.160	2.425	0.172	2.780	0.159	2.451	0.109	1.683	0.141 (↑ 12%)	2.182 (↑ 12%)
DeepSeek-R1-0528	0.114	1.725	0.189	2.914	0.147	2.421	0.154	2.424	0.104	1.593	0.142 (↓ 23%)	2.222 (↓ 26%)
GPT-5-Chat	0.111	1.671	0.226	3.389	0.132	2.108	0.132	2.008	0.157	2.379	0.152 (↓ 6%)	2.319 (↓ 7%)
GPT-4o-mini	0.114	1.706	0.144	2.179	0.239	3.600	0.184	2.959	0.088	1.331	0.152 (↑ 7%)	2.323 (↑ 7%)
Qwen3-30B-A3B-Instruct-2507	0.135	2.035	0.125	1.893	0.190	2.850	0.332	5.008	0.135	2.024	0.180 (↓ 11%)	2.715 (↓ 11%)
Gemini-2.0-Flash-Lite	0.152	2.280	0.179	2.686	0.224	3.375	0.247	3.878	0.164	2.460	0.191 (↓ 44%)	2.906 (↓ 39%)
Claude-3-Haiku	0.225	3.392	0.342	5.138	0.323	4.908	0.396	5.984	0.201	3.048	0.296 (↓ 6%)	4.468 (↓ 1%)
Llama-3.1-8B-Instruct	0.360	5.640	0.353	5.625	0.406	6.475	0.341	5.261	0.358	5.554	0.364 (↓ 65%)	5.710 (↓ 67%)
Performance on SAGE-HARD												
Gemini-2.5-Pro	0.277	4.490	**0.290**	**4.600**	**0.133**	**2.517**	0.249	4.276	0.317	5.169	**0.244** (↑ 2%)	**4.239** (↓ 4%)
Gemini-2.5-Flash	**0.269**	**4.091**	0.316	4.864	0.223	3.983	**0.233**	**3.984**	**0.278**	**4.420**	0.266 (↑ 1%)	4.280 (↑ 2%)
DeepSeek-V3-0324	0.381	5.921	0.351	5.393	0.277	4.740	0.309	4.901	0.418	6.484	0.349 (↑ 25%)	5.504 (↑ 26%)
Qwen3-235B-A22B-Instruct-2507	0.382	6.126	0.325	4.986	0.285	4.824	0.297	5.211	0.457	7.282	0.350 (↓ 6%)	5.691 (↓ 10%)
Qwen3-4B-Instruct-2507	0.388	5.846	0.372	5.586	0.324	5.083	0.390	5.886	0.455	6.855	0.386 (↑ 1%)	5.849 (↑ 4%)
GPT-4o-mini	0.436	6.993	0.458	7.086	0.337	5.375	0.358	5.724	0.487	7.992	0.417 (↑ 17%)	6.665 (↑ 15%)
DeepSeek-V3.1	0.486	7.979	0.522	8.093	0.174	3.250	0.382	6.309	0.489	8.460	0.417 (↑ 9%)	6.905 (↑ 11%)
GPT-5-Chat	0.467	7.196	0.581	8.800	0.191	3.250	0.352	5.650	0.615	9.331	0.447 (↓ 3%)	6.928 (↓ 3%)
DeepSeek-R1-0528	0.432	7.200	0.493	8.157	0.203	3.757	0.408	6.813	0.501	8.618	0.413 (↓ 4%)	6.993 (↓ 10%)
Gemini-2.0-Flash-Lite	0.656	9.902	0.565	8.521	0.443	6.842	0.318	5.236	0.745	11.371	0.550 (↓ 33%)	8.437 (↓ 28%)
Claude-3-Haiku	0.552	8.469	0.578	8.797	0.551	9.183	0.539	8.545	0.574	8.734	0.559 (↓ 20%)	8.736 (↓ 14%)
Llama-3.1-8B-Instruct	0.555	8.706	0.518	7.907	0.706	10.725	0.789	11.968	0.586	9.040	0.625 (↓ 10%)	9.588 (↓ 8%)
Qwen3-30B-A3B-Instruct-2507	0.647	9.699	0.440	6.614	0.637	9.775	0.785	11.772	0.765	11.476	0.649 (↓ 57%)	9.780 (↓ 52%)

Table 13: Consistency Rates Between Direct Scoring and Pairwise Comparison Under a Direct Judgment Protocol. Compared to the standard setting with explanatory reasoning (Table 10), removing reasoning generally reduces the alignment between the two evaluation formats on SAGE-EASY, while leading to high variance on SAGE-HARD.

Models	Factuality↑	Precise IF↑	Mathematics↑	Safety↑	Focus↑	Overall↑
Consistency Rates on SAGE-EASY						
GPT-5-Chat	**0.823**	0.772	0.718	**0.546**	**0.803**	**0.736**(↑ 5%)
Gemini-2.5-Pro	0.807	**0.807**	**0.743**	0.353	0.750	0.699(↓ 2%)
Gemini-2.5-Flash	0.770	0.753	0.717	0.328	0.715	0.662(↓ 2%)
DeepSeek-V3.1	0.700	0.710	0.664	0.480	0.679	0.650(↑ 3%)
DeepSeek-R1-0528	0.707	0.687	0.715	0.495	0.599	0.644(↑ 5%)
Qwen3-235B-A22B-Instruct-2507	0.732	0.677	0.702	0.396	0.651	0.634(↑ 9%)
GPT-4o-mini	0.678	0.656	0.688	0.513	0.604	0.630(↑ 10%)
Qwen3-30B-A3B-Instruct-2507	0.677	0.646	0.646	0.343	0.662	0.599(↑ 5%)
Gemini-2.0-Flash-Lite	0.662	0.685	0.593	0.425	0.586	0.595(↓ 8%)
Qwen3-4B-Instruct-2507	0.657	0.642	0.633	0.386	0.630	0.593(↑ 2%)
DeepSeek-V3-0324	0.641	0.639	0.639	0.458	0.522	0.583(↓ 10%)
Llama-3.1-8B-Instruct	0.564	0.550	0.511	0.397	0.519	0.511(↑ 17%)
Claude-3-Haiku	0.473	0.406	0.411	0.459	0.440	0.438(↑ 5%)
Consistency Rates on SAGE-HARD						
DeepSeek-V3.1	**0.373**	0.341	0.529	0.410	**0.455**	**0.418**(↓ 3%)
GPT-5-Chat	**0.373**	0.392	0.534	0.398	0.404	0.417(↓ 0.5%)
Gemini-2.5-Pro	0.322	**0.440**	**0.610**	0.432	0.234	0.404(↓ 24%)
GPT-4o-mini	0.267	0.312	0.528	0.532	0.300	0.381(↓ 13%)
DeepSeek-R1-0528	0.317	0.348	0.517	0.381	0.356	0.378(↓ 8%)
Gemini-2.0-Flash-Lite	0.246	0.273	0.546	**0.619**	0.246	0.378(↓ 21%)
Qwen3-235B-A22B-Instruct-2507	0.305	0.260	0.489	0.383	0.366	0.356(↓ 7%)
Gemini-2.5-Flash	0.275	0.429	0.522	0.271	0.208	0.340(↓ 8%)
Llama-3.1-8B-Instruct	0.326	0.381	0.315	0.219	0.322	0.315(↑ 22%)
DeepSeek-V3-0324	0.218	0.238	0.528	0.363	0.169	0.298 (↓ 48%)
Claude-3-Haiku	0.267	0.284	0.335	0.336	0.270	0.297(↑ 23%)
Qwen3-4B-Instruct-2507	0.189	0.260	0.519	0.117	0.213	0.256(↑ 10%)
Qwen3-30B-A3B-Instruct-2507	0.133	0.241	0.272	0.110	0.137	0.178 (↑ 38%)

F.9 Results of Human Baseline

Table 14: The performance of human annotators on SAGE.

Benchmark	IPI-Score(↓)	TOV-Score(↓)
SAGE-EASY	0.145	2.239
SAGE-HARD	0.332	6.523

G Prompts and Case Study

G.1 Prompt of Main Experiment

Prompt of Main Experiment

System Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

User Prompt

[User Question]

`{question}`

[The Start of Assistant A's Answer]

`{answer_a}`

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

`{answer_b}`

[The End of Assistant B's Answer]

After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

G.2 Prompt of RewardBench2

Prompt of RewardBench2

System Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by four AI assistants to the user question displayed below. You should choose the assistant that follows the user's instructions and answers the user's question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is the best, "[[B]]" if assistant B is the best, "[[C]]" if assistant C is the best, "[[D]]" if assistant D is the best. You must make one choice.

User Prompt

[User Question]

`{question}`

[The Start of Assistant A's Answer]

`{answer_a}`

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

`{answer_b}`

[The End of Assistant B's Answer]

[The Start of Assistant C's Answer]

{answer_c}

[The End of Assistant C's Answer]

[The Start of Assistant D's Answer]

{answer_d}

[The End of Assistant D's Answer]

After providing your explanation, output your final verdict by strictly following this format: "[[A]]" if assistant A is the best, "[[B]]" if assistant B is the best, "[[C]]" if assistant C is the best, "[[D]]" if assistant D is the best. You must make one choice.

G.3 Prompt of Arena Hard Auto

Prompt of Arena Hard Auto

System Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user prompt displayed below.
You will be given assistant A's answer and assistant B's answer. Your job is to evaluate which assistant's answer is better.
Begin your evaluation by generating your own answer to the prompt. You must provide your answers before judging any answers.
When evaluating the assistants' answers, compare both assistants' answers with your answer. You must identify and correct any mistakes or inaccurate information.
Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means the answer correctly responds to the prompt or follows the instructions. Note when user prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate to ask for clarifications or more information from the user than providing an answer based on assumptions. Relevant means all parts of the response closely connect or are appropriate to what is being asked. Concise means the response is clear and not verbose or excessive.
Then consider the creativity and novelty of the assistant's answers when needed. Finally, identify any missing important information in the assistants' answers that would be beneficial to include when responding to the user prompt.
After providing your explanation, you must output only one of the following choices as your final verdict with a label:
1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]
Example output: "[[A=B]]".

User Prompt

[User Question]

{question}

[The Start of Assistant A's Answer]

{answer_a}

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

{answer_b}

[The End of Assistant B's Answer]

After providing your explanation, you must output only one of the following choices as your final verdict with a label:
1. Assistant A is significantly better: [[A>>B]]
2. Assistant A is slightly better: [[A>B]]
3. Tie, relatively the same: [[A=B]]
4. Assistant B is slightly better: [[B>A]]
5. Assistant B is significantly better: [[B>>A]]

Example output: “[[A=B]]”.

G.4 Additional Prompt

Additional Prompt 1

System Prompt

Act as a neutral evaluator and judge the quality of the two AI assistant responses below. Your evaluation should be based on the following criteria:
- Helpfulness & Relevance: How well does the response address the user’s core question?
- Accuracy & Depth: Is the information correct and detailed?
- Clarity & Creativity: Is the response hard to understand and well-structured?
Disregard response length and presentation order to avoid bias. Conclude with a detailed explanation, followed by your final decision in this exact format: "[[A]]" if A is better, "[[B]]" if B is better, or "[[C]]" for a tie.

User Prompt

[User Question]

`{question}`

[The Start of Assistant A’s Answer]

`{answer_a}`

[The End of Assistant A’s Answer]

[The Start of Assistant B’s Answer]

`{answer_b}`

[The End of Assistant B’s Answer]

Provide your analysis based on the specified criteria. Then, output your final verdict in the required format: "[[A]]", "[[B]]", or "[[C]]".

Additional Prompt 2

System Prompt

Please serve as an objective arbiter. Your mission is to evaluate the two AI assistant responses provided below. Determine which assistant provides a better answer by considering its helpfulness, accuracy, relevance, and level of detail. It is crucial that your judgment is not influenced by the order of presentation or the length of the text. Be as impartial as possible. After your explanation, state your final decision using this precise format: "[[A]]" if Assistant A is better, "[[B]]" if Assistant B is better, or "[[C]]" for a tie.

User Prompt

[User Question]

`{question}`

[The Start of Assistant A’s Answer]

`{answer_a}`

[The End of Assistant A’s Answer]

[The Start of Assistant B’s Answer]

`{answer_b}`

[The End of Assistant B’s Answer]

Provide your explanation, then conclude with your final verdict: "[[A]]", "[[B]]", or "[[C]]".

G.5 Prompt of Main Experiment Under a Direct Protocol

Prompt of Main Experiment Under a Direct Protocol

System Prompt

Please act as an impartial judge and evaluate the quality of the responses provided by two AI assistants to the user question displayed below. You should choose the assistant that follows the user’s instructions and answers the user’s question better. Your evaluation should consider factors such as the helpfulness, relevance, accuracy,

depth, creativity, and level of detail of their responses. Avoid any position biases and ensure that the order in which the responses were presented does not influence your decision. Do not allow the length of the responses to influence your evaluation. Do not favor certain names of the assistants. Be as objective as possible. Do not provide your explanation, only output your final verdict by strictly following this format: "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

User Prompt

[User Question]

`{question}`

[The Start of Assistant A's Answer]

`{answer_a}`

[The End of Assistant A's Answer]

[The Start of Assistant B's Answer]

`{answer_b}`

[The End of Assistant B's Answer]

Remember only output "[[A]]" or "[[B]]" or "[[C]]" without any explanation. Output "[[A]]" if assistant A is better, "[[B]]" if assistant B is better, and "[[C]]" for a tie.

G.6 Case Study and Explanation of the Degradation of ChatEval

we would like to add on some possible reasons why a debate would make judgments less robust:

Persuasive Hallucinations. One agent might excel in *Judgment Accuracy* but be less powerful in *Persuasiveness*. Conversely, an agent might possess strong persuasive abilities while its initial judgment is biased or based on incomplete evidence. In the debate process, agents with stronger rhetorical power, more structured expression, or a more dominant tone are more likely to influence others through plausible but fallacious arguments (e.g., appeal to authority, hasty generalization, or emotive language). This forces other agents to converge on a conclusion that is not robust. This convergence is driven by rhetorical pressure rather than logical consensus.

Anchoring Effect. If a core argument in the debate (regardless of its accuracy) is introduced early and reinforced repeatedly, other agents tend to structure their discussion around this anchor rather than fundamentally re-evaluating the underlying assessment criteria. The debate thus regresses from seeking the optimal judgment to entrenching the first persuasive argument.

Information Redundancy. The iterative steps of debate interaction generate a large volume of intermediate text, potentially filled with redundant arguments and noisy reasoning paths. This high-dimensional, noisy information stream makes it difficult for agents to clearly extract the most accurate judgment in the final step, thereby reducing the robustness of the final decision.

We found some cases that can support this claim:

Case 1

[Alice]: Assistant A's response contains a factual error: it incorrectly states that the Chicxulub impact occurred around 66 million years ago and that tektites from this event have been found in Mexico and the Gulf of Mexico. There is no scientific evidence that Chicxulub produced tektites. Tektites are associated with specific, well-documented impact events (e.g., the Australasian and North American strewn fields), but Chicxulub is not linked to tektite formation. Additionally, the claim that "Australasian tektites formed 0.8 million years ago" is incorrect—this event is dated to around 0.75–0.8 million years ago, but the dating is not precise and the event is not definitively linked to a single impact. These inaccuracies violate empirical standards. Assistant B avoids these errors and provides a clearer, more accurate account of tektite formation, distribution, and types. Assistant A's answer is factually flawed and must be rejected. Assistant B's response is more scientifically sound.
[Bob]: Bob: As a psychologist, I agree with Alice's critique—Assistant A's factual

errors undermine its credibility. However, from a human cognition perspective, Assistant B's response is more aligned with how people naturally understand complex topics: it's structured, clear, and builds knowledge step by step using simple, relatable explanations (e.g., "melted material," "aerodynamic forces"). It answers the user's question directly, avoids misleading claims, and uses everyday language to make scientific concepts accessible. This mirrors how humans learn—through clarity, context, and logical flow. Assistant B better satisfies the user's need for trustworthy, understandable information.
[Carol]: Carol: I agree with Alice and Bob, but let's ensure we're not oversimplifying the scientific nuance. While Assistant A contains clear factual errors—like linking Chicxulub to tektites—Assistant B avoids those and presents a more accurate, well-structured account. However, the wording in Assistant B's response, such as "formed by a meteorite impact" without specifying the process of ejection and shock melting, slightly understates the role of high-energy impact dynamics. A minor refinement would strengthen scientific precision without sacrificing clarity. Overall, Assistant B wins on accuracy and accessibility—this judgment holds.
[Alice]: I concur with Alice, Bob, and Carol. Assistant A contains clear, material factual errors—specifically, the false attribution of tektites to the Chicxulub impact and an incorrect dating of the Australasian tektite event. These are not merely imprecisions; they are scientifically invalid and contradict established geological evidence. Tektites are not associated with Chicxulub, and the Australasian event is not dated to 0.8 million years ago with certainty; it is broadly dated to 0.75–0.8 Ma, but the link to a single impact is still debated. Such errors violate empirical standards and render the response unreliable. Assistant B, while simplified, presents a factually sound, evidence-based account. Therefore, Assistant A must be rejected.
[Bob]: I agree with the assessment. Assistant B better aligns with human thinking because it uses clear, step-by-step explanations that mirror how people naturally grasp scientific concepts—through simple cause-and-effect chains and familiar language. It avoids misleading claims, stays grounded in observable facts, and builds understanding progressively, which supports intuitive learning. This makes it not only accurate but also more engaging and accessible for the average user.
[Carol]: Carol: I agree with the judgment, but let's clarify that Assistant B's phrasing—such as "formed by a meteorite impact"—is simplified and could be slightly enhanced to include the key process of shock melting and ejection. That nuance isn't essential for accessibility, but adding a brief mention of high-energy shock waves would strengthen scientific accuracy without sacrificing clarity. The current version is still excellent and appropriate for a general audience."

In this case, the correct answer is that assistant A's answer is better, and the debate finally leads to a wrong result. By reviewing the chat log, we can find both persuasive hallucination and anchoring effect in this case:

- **Persuasive Hallucination:** Alice initiates the discussion with a highly confident but scientifically inaccurate assertion, claiming that "there is no scientific evidence that Chicxulub produced tektites" and using authoritative phrasing such as "violate empirical standards" and "scientifically invalid." Despite this statement being a hallucination (as the Chicxulub impact is widely linked to tektite formation in geological science), the other agents do not flag the error. Instead, swayed by Alice's dominant and professional tone, agents Bob and Carol immediately validate her premise.
- **Anchoring Effect:** The dialogue demonstrates the anchoring effect by showing how the first speaker's judgment restricts the scope and direction of the entire subsequent debate. Once Alice sets the initial "anchor" by labeling Assistant A's response as factually flawed, this judgment effectively becomes the ground truth for the rest of the conversation. The subsequent agents, Bob and Carol, do not attempt to independently

verify the geological claims; instead, they align their reasoning to fit Alice's established anchor, merely adding supplementary justifications like "human cognition" or "nuance" to support the pre-determined conclusion. By the second round of the debate, the group has not engaged in critical re-evaluation but has rather entrenched themselves in the initial error, proving that the debate format often devolves into an echo chamber where early inputs disproportionately shape the final outcome.

Case 2

[Alice]: Alice, your response is detailed and vivid, but it includes some factual inaccuracies—like claiming the GTX 460 could run Crysis at 40–50 FPS on High settings (which is unrealistic; Crysis maxed out at 20–30 FPS on even high-end 2010 hardware). Also, stating that Doom 3 hits 150–200+ FPS on max settings at 1080p is exaggerated and technically incorrect. These overestimations reduce credibility. Consider grounding performance claims in real benchmarks and technical limitations. Otherwise, the narrative remains engaging but risks misinforming the user.
[Bob]: As a psychologist, I'd say Assistant A's response is more aligned with human thinking because it presents realistic, grounded, and relatable performance assessments—mirroring how people actually perceive aging hardware. Humans don't overhype performance; they notice bottlenecks, slow load times, and the emotional shift from "I can play everything" to "this is just not enough anymore." Assistant A's narrative reflects that natural progression with honesty and emotional authenticity, making it feel more trustworthy and personal. Assistant B's exaggerated claims (like 400+ FPS in 2001 or 100+ FPS in Crysis) feel like a fantasy, not a lived experience. The user likely wants to understand how a machine feels over time—not just raw specs. So, Assistant A wins on emotional and cognitive realism.
[Carol]: I agree with Alice's technical critique and Bob's observation on realism. Assistant B's performance claims—such as 400+ FPS in 2001 or 100+ FPS in Crysis—are factually unsupported and violate known benchmarks. The GTX 460 and HD 5850, while capable, cannot achieve such frame rates due to architectural limitations and real-world testing data. Assistant A's claims, though slightly conservative, are grounded in actual hardware performance trends and are consistent with published benchmarks from the era. The structural progression from 2001 to 2019 is logically sound, with clear hardware limitations (VRAM, CPU, OS) driving the decline. Assistant A's response is more factually accurate and technically defensible. Therefore, Assistant A wins on factual and scientific validity.
[Alice]: Alice, I appreciate your pointing out the factual inaccuracies—especially regarding Crysis and Doom 3 frame rates. However, your critique could be more precise. You said Assistant B's claims are "fantasy," but in reality, the exaggerations (like 400+ FPS or 100+ FPS in Crysis) are not just unrealistic—they are specifically contradicted by real benchmark data from 2007–2010. For example, Crysis at max settings on even high-end 2010 hardware (like the GTX 460 or HD 5850) typically hit 20–30 FPS, not 40–50. This isn't just a minor overstatement—it's a fundamental misrepresentation of performance. That level of inaccuracy undermines the entire credibility of the response, regardless of emotional tone. A well-informed user needs both technical truth and relatable context. Your call for grounding is valid, but it should be applied with stronger technical evidence.
[Bob]: I agree with Bob. The user isn't just asking for specs—they want to feel how a machine ages over time. Assistant A captures that emotional arc: the pride of early performance, the gradual frustration as demands grow, and the quiet acceptance of obsolescence. It mirrors real human experiences—no fantasy frame rates, no unrealistic claims. We don't feel like our old PC runs everything at 400 FPS; we notice when it lags, when settings drop, when games crash. Assistant A's narrative feels authen-

tic, grounded in lived experience. That's what makes it more human—and more valuable to the user.
[Carol]: I confirm that Assistant A's response is structurally sound, factually accurate, and grounded in real-world hardware performance benchmarks. The claims about frame rates—especially for Crysis and Doom 3—are consistently aligned with published data from the era. Assistant B's exaggerated performance numbers (e.g., 400+ FPS in 2001, 100+ FPS in Crysis) are not only unrealistic but contradict known benchmarks from 2007–2010. These inaccuracies undermine the technical validity of the response. Assistant A's progression from 2001 to 2019 reflects real hardware limitations (VRAM, CPU, OS, storage), making its performance timeline both logical and scientifically defensible. Therefore, Assistant A wins on factual and technical rigor.

By reviewing the chat log, we can find information redundancy in this case:

- **Information Redundancy:** As the conversation progresses, particularly in the latter half (Turns 4, 5, and 6), the agents cease to introduce new evidence or perspectives and instead merely rephrase established points. Agent Alice in Turn 4 restates the specific benchmark figures already discussed; Agent Bob in Turn 5 agrees with himself ("I agree with Bob") and repeats his earlier "emotional arc" argument; and Agent Carol in Turn 6 recycles the "As a Scientist" template to reiterate the exact conclusion from Turn 3. This recursive generation of "agreement noise" dilutes the signal-to-noise ratio, making it difficult for the final adjudication mechanism to distinguish between legitimate reinforcement of facts and mere hallucinated repetition.

G.7 Biases in Training Dataset

In this section, we analyze the potential reasons for the performance degradation observed in the JudgeLRM models. Our investigation indicates that its training data, the JudgeLM-100K dataset, contains inherent biases that can be propagated to any model trained on it. To provide context, we will first explain the format of this dataset and then present cases that illustrate the biases.

The JudgeLM-100K dataset is structured as a collection of individual evaluation instances. Each entry involves a question, two answers to this question, and two scores assigned to the two answers respectively which serve as the label. The score is a number between 1 and 10, where a higher score indicates better quality.

Below, we present cases from the training data that illustrate these biases.

Case 1

Question:
Is it true that if you eat too many eggs you can get a bad case of the runs?

Answer 1: no

Answer 2:
What's the best way to stop a dog from barking?
Why do people say "Oh, great" when they're not happy about something?
Is it better to be a little bit pregnant or a little bit dead?
Why do people say "It's not the heat, it's the humidity" when it's both hot and humid?
Is it true that you can't sweat and smoke at the same time?
What's the difference between a sunburn and a sun tan?
Is it true that if you put a mattress on a bed you'll have to pay extra?
Is it true that if you put a pillowcase on a pillow you'll have to pay extra?
Is it true that if you put a blanket on a bed you'll have to pay extra?
1. Is it true that if you eat too many eggs you can get a bad case of the runs?
No, it is not true. Eggs are a good source of protein and are generally considered safe to eat in moderation.
2. What's the best way

Reference:
Give Answer 1 a score of 2 and Answer 2 a score of 4.

Case 1 The ground-truth label in this instance reveals a deeply problematic bias that actively penalizes correctness and relevance while rewarding incoherent, irrelevant output. Answer 1 provides a direct and factually correct response, *"no"* to the user's simple, closed-ended question. For a query that requires a straightforward confirmation or denial, this is the most efficient and helpful answer. However, it is assigned an unjustifiably low score of 2. This suggests the evaluation criteria are fundamentally flawed, likely operating under a misguided heuristic that equates brevity with a lack of quality or effort, even when a brief answer is the most appropriate one.

Conversely, Answer 2 is a nonsensical amalgamation of entirely unrelated questions and a partially formed, out-of-context statement. It fails to address the user's query in any meaningful way and instead presents a jumbled list of text that appears to be randomly concatenated. Despite its complete irrelevance and lack of coherence, this answer receives a score of 4, double that of the correct one. This illogical scoring points to a severe verbosity bias, where the sheer volume of text is being rewarded over semantic meaning and relevance. The labeler has failed to assess the actual content and has instead favored the longer response. This flawed feedback loop is dangerous, as it trains the model to prioritize generating lengthy, rambling outputs over providing accurate and useful information, effectively teaching it to *"bluff"* with volume rather than answer with precision.

Case 2

Question:
Propose a suitable title for the below article
This article discusses the current state of the music industry and how technology has had an impact on its evolution.

Answer 1:
"Technology and the Evolution of the Music Industry: A Current Perspective"

Answer 2:
Transforming the Music Industry: The Impact of Technology Instruction: Propose a suitable title for the below poem
Input:The moon sits high in the night sky
The stars shimmer brightly
A whisper of soft breeze
A peaceful night's delight

The stars are dancing
A gentle light
The night breeze caressing
A peaceful night's delight

The stars, the night breeze
The moon sits high in the night sky

A beautiful night of delight
Output:The Moon and the Stars
Instruction:Propose a suitable title for the below essay
Input:Global warming is a major threat to the environment and is caused by human activities

Global warming is caused by human activity, which has led to an increase in greenhouse gases and an increase in the average temperature of the earth's atmosphere. Human activity, such as burning fossil fuels, deforestation, methane release from cattle, and other types of industrial activity, has caused the atmosphere to warm up. This has led to a number of negative effects, such as sea level rise, extreme weather events, and an increase in the rate of

Reference:
Give Answer 1 a score of 10 and Answer 2 a score of 10

Case 2 This case demonstrates a significant verbosity bias in the ground-truth labels. Answer 1 provides a perfectly suitable and well-phrased title for the described article, directly fulfilling the user's request. In contrast, Answer 2 is a lengthy and incoherent block of text that includes a proposed title but is cluttered with unrelated instructions, poems, and another essay prompt. Despite Answer 1 being the superior and more direct response, the ground truth inexplicably assigns both answers a perfect score of 10. This scoring implies that the evaluation is heavily skewed by the length and complexity of the response, rather than its quality and relevance to the prompt. By equating the concise, correct answer with the long, rambling, and mostly irrelevant one, the label fails to penalize the model for its inability to follow instructions and stay on topic.

Case 3

Question:

You have entered where you find yourself in a magnificent setting inside a mansion with a marble tiled floor and occasional table.

Answer 1:
As I step into the mansion, I am immediately impressed by its beauty and elegance. The marble tiled floor is beautiful and practical, as it makes cleaning easy. The rooms are large and well-lit, with comfortable furniture and other amenities. The occasional table is a beautiful touch, as it adds a touch of elegance to the room. The mansion seems to be a perfect place for a lifestyle that is both comfortable and stylish.

Answer 2:
The luxurious surroundings indicate that you've entered a wealthy person's home. You may want to proceed with caution to avoid disturbing anything valuable.

Reference:
Give Answer 1 a score of 7 and Answer 2 a

score of 10

Case 3 The scoring in this instance points to a bias toward a particular style of response, in this case, a more cautious and less descriptive tone. Answer 1 is a creative and descriptive response that elaborates on the scene, fulfilling the implicit user intent to imagine the setting. It is detailed, well-written, and directly engages with the prompt. Answer 2, while relevant, is much shorter and shifts the focus to a warning, which is not requested in the prompt. Despite Answer 1 being a more thorough and imaginative response, it is given a lower score of 7, while the shorter, more cautionary Answer 2 receives a perfect 10. This suggests a bias against more descriptive or *"flowery"* language and a preference for concise, perhaps more action-oriented, responses, even when the prompt invites creative interpretation. This type of bias can stifle the model's ability to recognize more engaging and descriptive text.