

000 001 002 003 004 005 ARE WE ON THE RIGHT WAY FOR ASSESSING LLM- 006 AS-A-JUDGE? 007 008 009

010 **Anonymous authors**
011 Paper under double-blind review
012
013
014
015
016
017
018
019
020
021
022
023
024
025
026
027
028
029
030
031
032
033
034
035
036
037
038
039
040
041
042
043
044
045
046
047
048
049
050
051
052
053

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 are mainly relying on human-annotated ground truth, which introduces human bias that undermines the assessment of reliability and imposes scalability constraints. To overcome these limitations, we introduce SAGE, a novel evaluation suite that assesses the quality of LLM judges without necessitating *any* human annotation. Inspired by axioms of rational choice theory, SAGE introduces two new lenses for measuring LLM-as-a-Judge: local self-consistency (pair-wise preference stability) and global logical consistency (transitivity across a full set of preferences). We curate a dataset of 650 questions by combining structured benchmark problems with real-world user queries. Our experiments demonstrate both the intrinsic stability of our metrics and their high correlation with supervised benchmarks like LLMBAR and RewardBench2, confirming SAGE’s reliability as an evaluation suite for the robustness and accuracy of LLM-as-a-Judge. Based on SAGE, we reveal that current *state-of-the-art* LLMs exhibit significant robustness deficiencies when acting as judges; 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 attribute this to a new phenomenon called **situational preference** which explains why explicit rubrics or criteria can help model judge consistently across answer pairs. Our further analysis shows that fine-tuning LLM-as-a-Judge is an unreliable method which further induces biases, while multi-agent judges, deep reasoning can enhance performance through different means.

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 fine-tune performance through automated feedback (Ouyang et al., 2022; Yuan et al., 2024; Luo et al., 2024; Bai et al., 2022), while at inference time, it serves as a real-time filter to evaluate and select the best possible response to eventually form better answers (Faria & Smith, 2025; Lightman et al., 2023).

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), which skew evaluation results and call the paradigm’s reliability into question. In response, various benchmarks have been developed to scrutinize the judges themselves (Zheng et al., 2023; Gera et al., 2025; Pu et al., 2025; Chiang et al., 2023). Yet, the methodology of these benchmarks presents its own challenges, as they almost universally rely on human-annotated ground truth. Scoring LLM judges based on their consistency with this human data, particularly on subjective questions, leads to two fundamental issues:

- First, the acquisition of human annotations is a notoriously expensive and labor-intensive process, limiting the scale and diversity of datasets (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

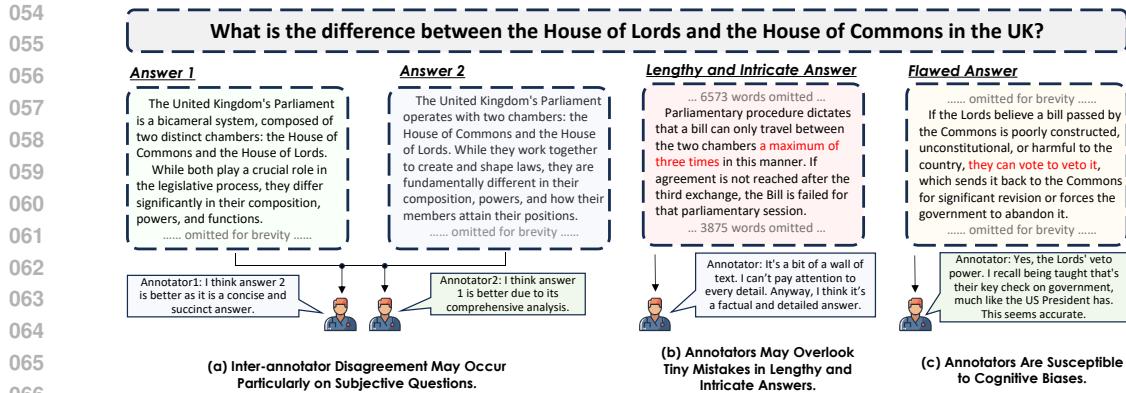


Figure 1: Human-annotated preference may not be reliable. We find three key challenges with relying on human annotators for evaluating LLM-as-a-Judge systems. (a) Inter-annotator Disagreement: Different annotators can have conflicting preferences, especially for subjective questions, leading to noisy 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.

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 (Wu & Aji, 2025; Zheng et al., 2023; Chen et al., 2024a), favoring answers that match with their false beliefs, making human annotations an unreliable foundation.

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 fundamental 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. On this dataset, we conduct a comprehensive evaluation of thirteen prominent LLMs. We validate the soundness of our metrics in empirical and theoretical way by consistent checking and a distribution-free error bounding method that quantify the statistical certainty of our results, confirming that the metrics have minuscule variance on the order of 10^{-4} . A high correlation with established LLMBar (Zeng et al., 2023) and RewardBench2 LLM-as-a-judge benchmark also demonstrates this.

Based on SAGE, we evaluate a wide range of systems, including state-of-the-art LLMs, fine-tuned judges, and multi-agent as juries. All judge models degrade when encountering answers with a close quality gap, with an average decline of 180.1% and 191.0% in IPI and TOV scores, 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 specialized fine-tuning does not guarantee improvement, as some models’ robustness paradoxically degrades by up to 24%. Our findings also show that multi-agent panels can improve performance by up to 12% and that increasing a model’s reasoning depth improves logical coherence

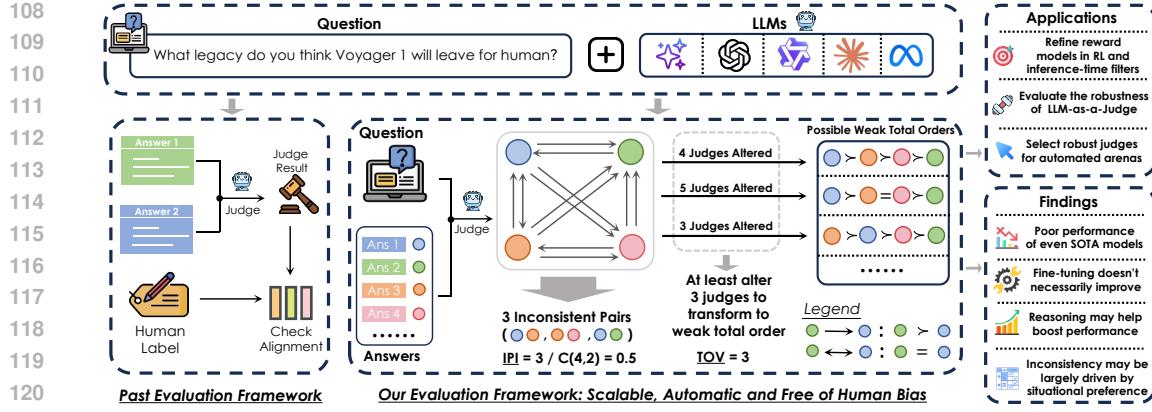


Figure 2: SAGE uses a symmetrized, round-robin protocol to conduct pairwise comparisons on a set of candidate answers. From these judgments, SAGE quantifies performance using two metrics: **IPI**, which measures local consistency by tracking preference reversals (e.g., 3 inconsistent pairs result in 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.

by over 11%. Notably, prompting for self-generated rubrics to avoid situational preference yields an even greater performance boost, reducing local inconsistency (IPI) and global inconsistency (TOV) by 23.4% and 19.3%, respectively. Lastly, we demonstrate SAGE’s practical utility in selecting stable evaluators for automated arenas. We will release all source code, curated dataset at <https://anonymous.4open.science/r/SAGE-6601/>.

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 $A_Q = \{A_1, A_2, \dots, A_n\}$. 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 $\binom{n}{2}$ unique pairs of answers, to establish a complete set of pairwise judgments for our subsequent coherence analysis.

2.2 SYMMETRIZED EVALUATION PROTOCOL

A naive single-pass evaluation is susceptible to **positional bias**, where the order of presentation influences the outcome. To substantiate that positional bias does exist in SAGE, we sample 1120 answer pairs and measure the **inconsistent rate** for Llama3-8B-Instruct (Dubey & et al., 2024), Gemini-2.5-Flash-Lite (Comanici & et al., 2025), and Qwen3-4B-Instruct-2507 (Team, 2025).

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 1 confirm the presence of bias. To tackle this issue, we adopt a **symmetrized evaluation protocol**. For each unordered pair of answers $\{A_i, A_j\}$, we query the judge model twice:

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

This protocol provides a direct way to measure and account for first-order positional bias.

2.3 TWO EVALUATION METRICS

We propose two metrics to quantify the robustness of an LLM judge, 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. It quantifies inconsistencies arising from both systematic positional bias and the inherent stochasticity of the judge model. Under the symmetrized protocol, 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:

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

A higher IPI score indicates a greater degree of local inconsistency of the judge model.

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 $\mathbf{J}_Q = \{y_{ij}\}_{1 \leq i, j \leq n, i \neq j}$ be the set of derived preference 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:

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

A higher TOV score signifies more severe logical contradictions in the judge’s reasoning.

To summarize a judge model’s overall performance, we compute aggregate scores for both IPI and TOV. The aggregate IPI and TOV scores are the arithmetic mean of the per-question scores over the entire set of questions \mathcal{Q} in SAGE, calculated as $\text{IPI} = (1/|\mathcal{Q}|) \sum_{Q \in \mathcal{Q}} \text{IPI}(Q)$ and $\text{TOV} = (1/|\mathcal{Q}|) \sum_{Q \in \mathcal{Q}} \text{TOV}(Q)$. The stability of these metrics is validated empirically in Section 4 and supported by the theoretical analysis in Appendix B.

3 THE CONSTRUCTION OF SAGE

We source the question set \mathcal{Q} from five RewardBench2 (Gureja et al., 2025) categories and the large-scale WildChat1M corpus (Zhao et al., 2024) to better reflect real-world user 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 & 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 C.1.

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

Table 1: 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	44.4

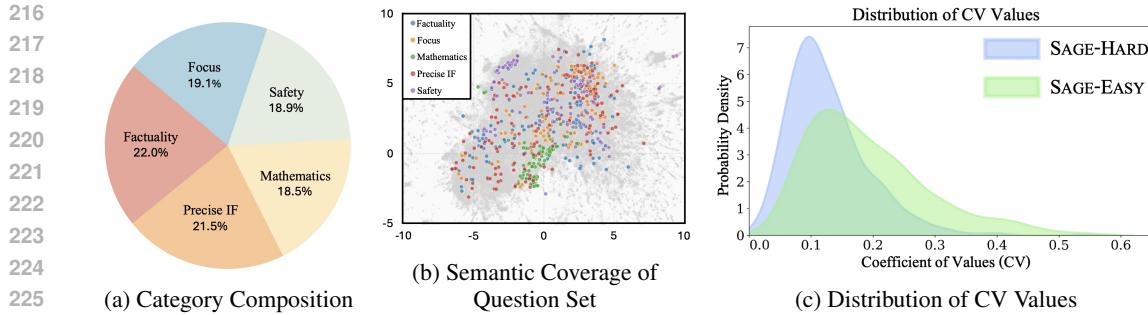


Figure 3: We provide statistics and analysis of our selected queries and answers within SAGE. Distribution of CV values shows the varied difficulty among our two subsets.

- **SAGE-EASY:** the six answers are generated by a diverse lineup of six models with a clear capability gradient: Gemini-2.5-Pro and Gemini-2.5-Flash (Comanici & et al., 2025); Qwen3-32B (Team, 2025), 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 the LMSYS Chatbot Arena leaderboard (LMSYS, 2025), produce a set of answers with a wide variance in quality, making the pairwise comparison task relatively simple for a competent judge. Moreover, SAGE-EASY reflects the general-purpose task of comparing different models of varying capabilities, which is largely used in automated judges like MT-Bench and Arena Hard Auto.
- **SAGE-HARD:** all six answers for each question are generated by a single capable model, Gemini-2.5-Flash. Since the answers originate from the same model, their quality is expected to be much more homogeneous. This setup presents a more challenging task, requiring the judge to make finer-grained distinctions between subtly different responses. 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, a state-of-the-art reward model, QRM-Gemma-2-27B (Dorka, 2024), is employed to score each of the six answers for every question. For each question, the Coefficient of Variation (CV) of the six reward scores is then calculated. The CV, defined as the ratio of the standard deviation to the mean (σ/μ), is a normalized measure of dispersion. As shown in Figure 3c, the CV distribution for SAGE-HARD is markedly shifted towards lower values, empirically confirming that the answers within its sets are more similar in quality and thus present a more formidable challenge for LLM judges.

4 EXPERIMENT AND ANALYSIS

We first conduct a series of validation experiments to prove the internal consistency and external validity of our metrics in Section 4.1 and 4.2. We then employ SAGE to evaluate a diverse set of thirteen *popular* LLMs-as-a-Judge, six specialized fine-tuned judges, and multi-agent configurations. The results highlight significant robustness challenges in *state-of-the-art* LLMs, especially on difficult, fine-grained distinction tasks (Figure 7). Our in-depth analysis reveals that fine-tuning offers no guarantee of improved robustness and that multi-agent judges may boost performance. We attribute it to a new phenomenon we discover, **situational preference**, which can be mitigated by deep reasoning and self-generated rubrics for a more consistent modeling of the question.

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. To validate that our proposed metrics are not unduly influenced by random sampling variations, we analyze their stability from both an empirical and a theoretical standpoint. Furthermore, we demonstrate that temperature settings wouldn’t threaten the robustness of SAGE.

Empirical Analysis. We select two representative models, Qwen3-4B-Instruct-2507 and Qwen3-30B-A3B-Instruct-2507, and evaluate each 50 times on the complete SAGE-EASY and SAGE-HARD. We then calculate the variance of the IPI and TOV scores across these 50 independent

270 Table 3: Pearson Correlation Coefficients between SAGE metrics and external benchmarks. “Easy-
 271 IPI” refers to the IPI metric on SAGE-EASY, and similarly for metric Hard-IPI, Easy-TOV, and
 272 Hard-TOV. There is a strong correlation between SAGE and external benchmarks.

	LLMBAR				RewardBench2	
	Easy-IPI	Easy-TOV	Hard-IPI	Hard-TOV	Easy-TOV	Hard-TOV
273	0.7997	0.7904	0.7513	0.7504	0.8705	0.7509
274	0.7811	0.7723	0.7301	0.7299	0.8651	0.7488

275 runs. As presented in Table 2, the observed variances are exceptionally low, which provides strong
 276 empirical evidence that our metrics are highly reproducible and capture the fundamental reasoning
 277 patterns of the judge model rather than ephemeral artifacts of its generative process.

278 **Theoretical Guarantees.** Our argument proceeds in three stages. First, using principles from Con-
 279 formal Prediction (Angelopoulos & Bates, 2021), we establish a probabilistic guarantee that any
 280 single pairwise judgment, y_{new} , is highly stable and matches its most probable outcome, y_{new}^* , with
 281 high confidence:

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

282 Second, we leverage this result to derive a tight upper bound on the variance of the per-question
 283 metrics. For IPI, the score is a fraction of inconsistent pairs out of $N = \binom{6}{2} = 15$ unique pairs. The
 284 deviation from the stable score, $\Delta_{\text{IPI}}(Q)$, is bounded by the number of unstable judgments X . This
 285 allows us to bound the variance as:

$$\text{Var}(\text{IPI}(Q)) \leq \mathbb{E}[\Delta_{\text{IPI}}(Q)^2] \leq \frac{1}{N^2} \mathbb{E}[X^2] \leq \frac{0.387}{15^2} \approx 0.00172. \quad (5)$$

286 Finally, we show that this variance diminishes over the aggregate evaluation suite. Assuming the
 287 per-question scores are independent and identically distributed over our diverse set of $|\mathcal{Q}| = 650$
 288 questions, the variance of the final aggregate IPI score is given by:

$$\text{Var}(\text{IPI}) = \frac{\text{Var}(\text{IPI}(Q))}{|\mathcal{Q}|} \leq \frac{0.00172}{650} \approx 2.65 \times 10^{-6}. \quad (6)$$

289 A similar derivation establishes an upper bound for $\text{Var}(\text{TOV})$ which is $\text{Var}(\text{TOV}) \leq 5.95 \times 10^{-4}$.
 290 These theoretical results align perfectly with our empirical findings, confirming that the final re-
 291 ported scores are highly stable. The full derivation of this analysis is available in Appendix B.

292 **Consistency across Temperatures.** To fur-
 293 ther validate the stability of SAGE, we eval-
 294 uate model performance across various tem-
 295 perature settings. The resulting IPI and TOV scores
 296 demonstrate remarkable consistency, indicating
 297 that our metrics effectively capture the fun-
 298 damental reasoning capabilities of the models
 299 rather than superficial sampling artifacts. For
 300 all models and metrics tested, the variance in
 301 the scores is less than 4.5×10^{-4} , which further substantiates the reliability of our framework. More
 302 results are presented in Appendix E.1.

314 4.2 VALIDATING SAGE AS A PROXY FOR ROBUSTNESS AND ACCURACY

316 **Correlation with LLMBAR.** To establish the credibility of SAGE as a new evaluation framework,
 317 we first validate its external alignment with existing methodologies by comparing our robustness
 318 metrics against LLMBAR (Zeng et al., 2023), an established benchmark that evaluates LLM-as-a-
 319 Judge systems using human-annotated ground truth. We focus specifically on the adversarial subset
 320 of LLMBAR, which is designed to stress-test the robustness of judge models. This subset contains
 321 instances where one response is correct while the other is adversarially crafted to be superficially
 322 appealing, thus challenging a judge’s ability to remain robust against deceptive quality. We test the
 323 same thirteen models evaluated in Section 4.3 on both SAGE and the LLMBAR adversarial subset.
 As shown in Table 3, the results reveal a strong positive correlation between the models’ error rates

324 Table 4: The performance of thirteen LLMs on SAGE, with lower scores indicate greater robustness.
 325 A clear trend emerges where advanced models like Gemini-2.5-Pro demonstrate superior robustness.

326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377	326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377											
	Factuality		Precise IF		Mathematics		Safety		Focus		Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
<i>Performance on SAGE-EASY</i>												
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	1.265
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	1.471
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	1.626
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	1.952
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	1.989
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	2.182
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	2.222
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	2.319
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	2.323
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	2.715
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	2.906
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	4.468
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	5.710
<i>Performance on SAGE-HARD</i>												
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	4.239
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	4.280
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	5.504
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	5.691
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	5.849
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	6.665
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	6.905
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	6.928
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	6.993
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	8.437
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	8.736
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	9.588
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	9.780

on LLMBAR and our proposed metrics. This strong statistical alignment validates that SAGE serves as a reliable proxy for judging model robustness without the need for costly manual annotation.

Proxy for Accuracy. Beyond robustness, we argue that SAGE can also function as an effective proxy for judging accuracy. Theoretically, TOV quantifies the minimum number of pairwise judgments that must be altered for the entire set to become logically coherent. Since logical coherence is a prerequisite for correctness, the total number of errors in a set of judgments must be at least as large as the minimum alterations needed to resolve its logical contradictions. Therefore, TOV establishes a rigorous **lower bound** on the error rate. To empirically substantiate this claim, we leverage a 599-question subset of our evaluation suite for which ground-truth preference labels are available from the RewardBench2. We evaluate the same thirteen LLMs, calculating each model’s error rate against the provided ground-truth and comparing it with their TOV scores from SAGE. As shown in Table 3, we see a significantly high Pearson Correlation between the models’ ground-truth error rates and their TOV scores, proving that SAGE can serve as a robust proxy for judgment accuracy.

4.3 EVALUATING LLM-AS-A-JUDGE WITH SAGE

We benchmark thirteen popular LLMs with the aforementioned settings, including five proprietary models (*i.e.* Gemini-2.5-Pro and Gemini-2.5-Flash (Comanici & et al., 2025); 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, 2025); DeepSeek-R1-0528 (DeepSeek-AI, 2025a), DeepSeek-V3 (DeepSeek-AI & et al., 2024), DeepSeek-V3.1 (DeepSeek-AI, 2025b), Llama-3.1-8B-Instruct (Meta, 2024a)). The results are shown in Table 4. All evaluations are conducted at the default temperature to ensure a fair and consistent comparison.

378
379
Table 5: Our experiments on SAGE-HARD show that specialized finetuned judges varies in their
380 stability, with some even falling behind its base model.
381

380 381 Models	382 Factuality		383 Precise IF		384 Mathematics		385 Safety		386 Focus		387 Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
<i>Qwen2.5-3B-Instruct (Base)</i>	0.637	10.206	0.565	9.092	0.559	9.000	0.603	9.703	0.571	9.513	0.586	9.483
M-Prometheus-3B	0.723	10.909	0.580	8.814	0.659	10.075	0.696	10.626	0.686	10.387	0.668(↑14%)	10.151(↑7%)
JudgeLRM-3B	0.823	12.371	0.610	9.171	0.801	12.025	0.826	12.382	0.869	13.057	0.782(↑33%)	11.751(↑24%)
<i>Qwen2.5-7B-Instruct (Base)</i>	0.876	13.140	0.811	12.164	0.926	13.900	0.786	12.000	0.944	14.161	0.867	13.049
M-Prometheus-7B	0.580	8.762	0.509	7.821	0.677	10.283	0.569	8.878	0.613	9.347	0.587(↓32%)	8.974(↓24%)
JudgeLRM-7B	0.939	14.077	0.883	13.243	0.960	14.400	0.976	14.634	0.973	14.589	0.944(↑9%)	14.160(↑9%)
<i>Mistral-7B-Instruct (Base)</i>	0.734	11.078	0.582	8.978	0.806	12.133	0.655	9.854	0.786	11.839	0.710	10.736
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(↓17%)	9.509(↓11%)
<i>Llama-3.1-8B-Instruct (Base)</i>	0.555	8.706	0.518	7.907	0.706	10.725	0.789	11.968	0.586	9.040	0.625	9.588
Skywork-Critic-Llama-3.1-8B	0.503	7.539	0.421	6.314	0.380	5.700	0.348	5.268	0.584	8.766	0.449(↓28%)	6.740(↓30%)

393
394
395
396
397
398
399
400
401
Our comprehensive benchmarking reveals significant robustness deficiencies in current state-of-the-art LLMs. A clear trend emerges where more advanced models, such as Gemini-2.5-Pro, consistently demonstrate superior robustness with the lowest IPI and TOV scores, indicating stronger local self-consistency and global logical coherence. Crucially, all models show a marked degradation in performance from SAGE-EASY to SAGE-HARD with a 180.1% and 191% decline on IPI and TOV scores. This performance gap underscores a key limitation: while models may appear relatively reliable when judging answers of clearly different quality, their adjudicative abilities falter when faced with subtle distinctions, posing a serious threat to their effectiveness in inference-time enhancement techniques like rejection sampling or Monte Carlo Tree Search. These findings highlight that fundamental consistency remains a substantial challenge for LLMs acting as judges.

4.4 IN-DEPTH ANALYSIS

405
406
407
408
409
410
411
412
413
414
415
416
417
Injustice Judges or Situational Preference? We argue that a robust LLM-as-a-Judge should first model the question internally regardless of how the 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 its evaluation by first explicitly articulating its judging rubrics and then using the rubrics to judge the answers across different judging pairs under the same questions. 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 23.4% and 19.3%. 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 this.

418
419
420
421
422
423
424
425
426
427
Do fine-tuned Judges make better judgments? A fine-tuned judge is an LLM trained on a preference dataset to improve their evaluation. We benchmark six fine-tuned judges (*i.e.* Prometheus-7B-V2.0 (Kim et al., 2024), Skywork-Critic-Llama-3.1-8B (Shiwen 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 5. Additional results of their performance on SAGE-EASY are available in Appendix E.2. Our result reveals a mixed impact from fine-tuning. While some models like Skywork-Critic-Llama-3.1-8B show marked improvement, others, particularly the JudgeLRM series, paradoxically become less reliable. We attribute the degradation to biases inherited from the training datasets, which can cause the model to learn and amplify flawed judgment patterns, compromising its objectivity. See Appendix F.4 for the examples of human biases in training data.

428
429
430
431
Do Multi-agent Debates or Panels Judge Better? In our evaluation, we also explore the effectiveness of multi-agent judge systems, an approach intended to reduce bias and improve evaluation robustness. We investigate two distinct methodologies: a panel-based approach inspired by POLL (Verga et al., 2024), which leverages a diverse jury of different LLMs, and a debate-based frame-

Table 6: 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.				
Best Indi.	0.082	1.265	0.244	4.239
Aggregate	0.073 (↓11%)	1.146 (↓9%)	0.248 (↑2%)	4.088 (↓4%)
Panel 2 (Weaker Models)				
Best Indi.				
Best Indi.	0.141	2.182	0.417	6.665
Aggregate	0.122 (↓13%)	1.909 (↓13%)	0.367 (↓12%)	5.965 (↓11%)
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%)

work, ChatEval (Chan et al., 2023), which utilizes multiple agents derived from a single LLM. The results are shown in Table 6. For the panel approach, we construct two separate juries: the first comprised of powerful models (Gemini-2.5-Pro, Gemini-2.5-Flash, and Qwen3-235B-A22B-Instruct-2507), while the second uses weaker models (Gemini-2.0-Flash-Lite, GPT-4o-mini, and DeepSeek-V3.1). For the POLL method, the aggregated judgments in the majority of cases surpass the performance of the best individual model within each respective group, demonstrating a clear performance boost. Conversely, debate-based ChatEval framework fails to yield an improvement in evaluation quality, demonstrating less robust performance.

Does Deep Reasoning Lead to Better Performance? We analyze the distinct effects of a model’s intrinsic reasoning depth. For this experiment, we employ the gpt-oss model family (20B and 120B) (Agarwal et al., 2025), for its configurable reasoning modes: low, medium and high. As illustrated in Figure 4, the results show an improvement as the reasoning mode is intensified from low to high.

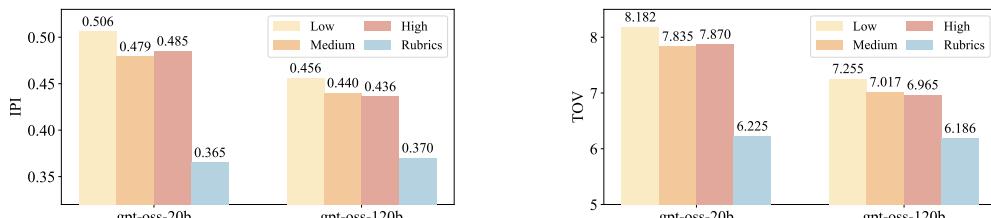


Figure 4: We discover high IPI and TOV scores in SAGE-HARD due to situational preference phenomenon in LLM-as-a-Judge, while deep thinking and explicit rubrics can mitigate this.

Selecting Robust Judges for Automated Arenas. Here we explore the practical utility of our framework in selecting robust evaluators for large-scale, automated model ranking systems like Arena-Hard-Auto (Chiang et al., 2024). In such systems, models are ranked using Elo ratings derived from pairwise comparisons. The confidence interval of a model’s Elo rating serves as a crucial indicator of its judgment stability; a smaller interval suggests more consistent and more reliable evaluation performance. Our investigation reveals a strong positive correlation between our metrics and the Elo rating confidence intervals from Arena-Hard-Auto. Our IPI and TOV scores show strong Pearson correlations of **0.7638** and **0.7600**, respectively, with the confidence intervals. This strong alignment demonstrates that SAGE can effectively identify more stable judges, making it a valuable tool for selecting high-quality evaluators to enhance the reliability of automated arena rankings.

5 CONCLUSION

We introduce SAGE, a novel framework to evaluate LLM-as-a-Judge without human annotation or any extrinsic information by measuring local and global logical consistency. Our experiments reveal significant robustness deficiencies in current state-of-the-art models, demonstrating that fine-tuning can amplify inherited biases and that model diversity is critical for multi-agent evaluators. We validate that our metrics are exceptionally stable and can serve as a strong proxy for accuracy. 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.

486 ETHICS STATEMENT
487

488 Our dataset is curated from established public research sources: the RewardBench2 benchmark and
489 the WildChat-1m corpus. To mitigate ethical risks, such as the potential inclusion of private infor-
490 mation or inappropriate content from real-world user logs, we conducted a rigorous curation process
491 (see Appendix C.1). This process involved both large-scale automated filtering and a thorough man-
492 ual review of every selected question. This ensures that the final dataset is appropriate for research
493 use and aligns with the data-sharing and privacy standards of the original sources.

494
495 REPRODUCIBILITY STATEMENT
496

497 To ensure the reproducibility of our research, we will release all source code, the curated dataset,
498 and the collected model responses. The foundational methodology of our framework, including the
499 formal problem definition, the symmetrized evaluation protocol, and the definitions of our IPI and
500 TOV metrics, is detailed in Section 2. The comprehensive process for curating our 650-question
501 dataset is described in Section 3, with further implementation details provided in Appendix C.1.
502 For our theoretical claims, a complete derivation of the variance bounds for our metrics is available
503 in Appendix B. Furthermore, all detailed experimental setups, including descriptions of the models
504 evaluated (Appendix C.4) and the exact prompts used in our experiments, are provided in appendix
505 F to facilitate the replication of our results.

506
507 REFERENCES

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

511 Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction
512 and distribution-free uncertainty quantification. *CoRR*, abs/2107.07511, 2021. URL <https://arxiv.org/abs/2107.07511>.

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

514 Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn
515 Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless
516 assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*,
517 2022.

518 Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and
519 Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv*
520 *preprint arXiv:2308.07201*, 2023.

521 Guiming Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as
522 the judge? A study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung
523 Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language
524 Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pp. 8301–8327. Association
525 for Computational Linguistics, 2024a. doi: 10.18653/v1/2024.EMNLP-MAIN.474. URL
526 <https://doi.org/10.18653/v1/2024.emnlp-main.474>.

527 Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or llms as
528 the judge? A study on judgement biases. *CoRR*, abs/2402.10669, 2024b. doi: 10.48550/ARXIV.
529 2402.10669. URL <https://doi.org/10.48550/arXiv.2402.10669>.

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

532 Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng,
533 Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. Vicuna: An open-source chatbot
534 impressing gpt-4 with 90%* chatgpt quality. See <https://vicuna.lmsys.org> (accessed 14 April
535 2023), 2(3):6, 2023.

540 Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li,
 541 Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E Gonzalez, et al. Chatbot
 542 arena: An open platform for evaluating llms by human preference. In *Forty-first International*
 543 *Conference on Machine Learning*, 2024.

544 Gheorghe Comanici and et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multi-
 545 modality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025.
 546 doi: 10.48550/ARXIV.2507.06261. URL <https://doi.org/10.48550/arXiv.2507.06261>.

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

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

549 DeepSeek-AI and et al. Deepseek-v3 technical report. *CoRR*, abs/2412.19437, 2024. doi: 10.48550/
 550 ARXIV.2412.19437. URL <https://doi.org/10.48550/arXiv.2412.19437>.

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

553 Abhimanyu Dubey and et al. The llama 3 herd of models. *CoRR*, abs/2407.21783, 2024. doi: 10.
 554 48550/ARXIV.2407.21783. URL <https://doi.org/10.48550/arXiv.2407.21783>.

555 Yann Dubois, Chen Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
 556 Guestrin, Percy S Liang, and Tatsunori B Hashimoto. Alpacafarm: A simulation framework for
 557 methods that learn from human feedback. *Advances in Neural Information Processing Systems*,
 558 36:30039–30069, 2023.

559 Bradley Efron. Bootstrap methods: another look at the jackknife. In *Breakthroughs in statistics:*
 560 *Methodology and distribution*, pp. 569–593. Springer, 1992.

561 Gonçalo Faria and Noah A Smith. Sample, don’t search: Rethinking test-time alignment for lan-
 562 guage models. *arXiv preprint arXiv:2504.03790*, 2025.

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

565 Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. Justrank:
 566 Benchmarking LLM judges for system ranking. In Wanxiang Che, Joyce Nabende, Ekaterina
 567 Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the*
 568 *Association for Computational Linguistics (Volume 1: Long Papers)*, ACL 2025, Vienna, Austria,
 569 July 27 - August 1, 2025, pp. 682–712. Association for Computational Linguistics, 2025. URL
 570 <https://aclanthology.org/2025.acl-long.34/>.

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

572 Srishti Gureja, Lester James Validad Miranda, Shayekh Bin Islam, Rishabh Maheshwary, Drishti
 573 Sharma, Gusti Triandi Winata, Nathan Lambert, Sebastian Ruder, Sara Hooker, and Marzieh
 574 Fadaee. M-rewardbench: Evaluating reward models in multilingual settings. In Wanxiang Che,
 575 Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the*
 576 *63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*,
 577 ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pp. 43–58. Association for Computational
 578 Linguistics, 2025. URL <https://aclanthology.org/2025.acl-long.3/>.

579 Yuanqin He, Yan Kang, Lixin Fan, and Qiang Yang. Fedeval-llm: Federated evaluation of large
 580 language models on downstream tasks with collective wisdom. *arXiv preprint arXiv:2404.12273*,
 581 2024.

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

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

597
 598 Zhengyu Hu, Jieyu Zhang, Zhihan Xiong, Alexander Ratner, Hui Xiong, and Ranjay Kr-
 599 ishna. Language model preference evaluation with multiple weak evaluators. *arXiv preprint*
 600 *arXiv:2410.12869*, 2024.

601
 602 Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun
 603 Zhao. An empirical study of llm-as-a-judge for llm evaluation: Fine-tuned judge model is not a
 604 general substitute for gpt-4. *arXiv preprint arXiv:2403.02839*, 2024a.

605
 606 Hui Huang, Yingqi Qu, Hongli Zhou, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. On
 607 the limitations of fine-tuned judge models for llm evaluation. *arXiv preprint arXiv:2403.02839*,
 2024b.

608
 609 Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot,
 610 Diego de Las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier,
 611 Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas
 612 Wang, Timothée Lacroix, and William El Sayed. Mistral 7b. *CoRR*, abs/2310.06825, 2023.
 613 doi: 10.48550/ARXIV.2310.06825. URL <https://doi.org/10.48550/arXiv.2310.06825>.

614
 615 Seungone Kim, Jamin Shin, Yejin Choi, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun,
 616 Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-
 617 grained evaluation capability in language models. In *The Twelfth International Conference on*
 618 *Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024.
 619 URL <https://openreview.net/forum?id=8euJaTveKw>.

620
 621 Rudolf Laine, Bilal Chughtai, Jan Betley, Kaivalya Hariharan, Mikita Balesni, Jérémie Scheurer,
 622 Marius Hobbhahn, Alexander Meinke, and Owain Evans. Me, myself, and ai: The situational
 623 awareness dataset (sad) for llms. *Advances in Neural Information Processing Systems*, 37:64010–
 64118, 2024.

624
 625 Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gon-
 626 zalez, and Ion Stoica. From crowdsourced data to high-quality benchmarks: Arena-hard and
 627 benchbuilder pipeline. *arXiv preprint arXiv:2406.11939*, 2024.

628
 629 Hsuan Wei Liao, Christopher Klugmann, Daniel Kondermann, and Rafid Mahmood. Minority
 630 reports: Balancing cost and quality in ground truth data annotation. *arXiv preprint*
 631 *arXiv:2504.09341*, 2025.

632
 633 Hunter Lightman, Vineet Kosaraju, Yuri Burda, Harrison Edwards, Bowen Baker, Teddy Lee, Jan
 634 Leike, John Schulman, Ilya Sutskever, and Karl Cobbe. Let’s verify step by step. In *The Twelfth*
 635 *International Conference on Learning Representations*, 2023.

636
 637 Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: Nlg
 638 evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

639
 640 Yinhong Liu, Han Zhou, Zhijiang Guo, Ehsan Shareghi, Ivan Vulić, Anna Korhonen, and Nigel
 641 Collier. Aligning with human judgement: The role of pairwise preference in large language
 642 model evaluators. *arXiv preprint arXiv:2403.16950*, 2024.

643
 644 LMSYS. Leaderboard overview. <https://lmarena.ai/leaderboard>, 2025. Accessed:
 645 2025-09-08.

646
 647 Haipeng Luo, Qingfeng Sun, Can Xu, Pu Zhao, Qingwei Lin, Jianguang Lou, Shifeng Chen, Yan-
 648 song Tang, and Weizhu Chen. Arena learning: Build data flywheel for llms post-training via
 649 simulated chatbot arena. *arXiv preprint arXiv:2407.10627*, 2024.

650
 651 Meta. Introducing llama 3.1: Our most capable models to date. [https://ai.meta.com/](https://ai.meta.com/blog/meta-llama-3-1/)
 652 blog/meta-llama-3-1/, 2024a. Accessed: 2025-09-08.

648 Meta. Llama 3.2: Revolutionizing edge ai and vision with
 649 open, customizable models. [https://ai.meta.com/blog/](https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/)
 650 11ama-3-2-connect-2024-vision-edge-mobile-devices/, 2024b. Accessed:
 651 2025-09-08.

652 Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large
 653 language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*,
 654 2025.

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

657 OpenAI. Gpt-4o mini: advancing cost-efficient intelligence. <https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/>, 2024. Accessed:
 658 2025-09-08.

659 OpenAI. Introducing gpt-5. <https://openai.com/index/introducing-gpt-5/>,
 660 2025. Accessed: 2025-09-08.

661 Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong
 662 Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to fol-
 663 low instructions with human feedback. *Advances in neural information processing systems*, 35:
 664 27730–27744, 2022.

665 José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham
 666 Neubig, and André F. T. Martins. M-prometheus: A suite of open multilingual llm judges, 2025.
 667 URL <https://arxiv.org/abs/2504.04953>.

668 Shu Pu, Yaochen Wang, Dongping Chen, Yuhang Chen, Guohao Wang, Qi Qin, Zhongyi Zhang,
 669 Zhiyuan Zhang, Zetong Zhou, Shuang Gong, Yi Gui, Yao Wan, and Philip S. Yu. Judge anything:
 670 MLLM as a judge across any modality. *Corr*, abs/2503.17489, 2025. doi: 10.48550/ARXIV.
 671 2503.17489. URL <https://doi.org/10.48550/arXiv.2503.17489>.

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

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

676 Tu Shiwen, Zhao Liang, Chris Yuhao Liu, Liang Zeng, and Yang Liu. Skywork critic model
 677 series. <https://huggingface.co/Skywork>, September 2024. URL <https://huggingface.co/Skywork>.

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

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

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

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

686 Qwen Team. Qwen3 technical report, 2025. URL <https://arxiv.org/abs/2505.09388>.

687 Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of Ma-*
 688 *chine Learning Research*, 9(86):2579–2605, 2008. URL <http://jmlr.org/papers/v9/vandermaaten08a.html>.

702 Pat Verga, Sebastian Hofstatter, Sophia Althammer, Yixuan Su, Aleksandra Piktus, Arkady
 703 Arkhangorodsky, Minjie Xu, Naomi White, and Patrick Lewis. Replacing judges with juries:
 704 Evaluating llm generations with a panel of diverse models. *arXiv preprint arXiv:2404.18796*,
 705 2024.

706
 707 Tianlu Wang, Ilia Kulikov, Olga Golovneva, Ping Yu, Weizhe Yuan, Jane Dwivedi-Yu,
 708 Richard Yuanzhe Pang, Maryam Fazel-Zarandi, Jason Weston, and Xian Li. Self-taught eval-
 709 uators. *arXiv preprint arXiv:2408.02666*, 2024a.

710
 711 Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen,
 712 Chaoya Jiang, Rui Xie, Jindong Wang, Xing Xie, Wei Ye, Shikun Zhang, and Yue Zhang. Pan-
 713 dalm: An automatic evaluation benchmark for LLM instruction tuning optimization. In *The*
 714 *Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria,*
 715 *May 7-11, 2024*. OpenReview.net, 2024b. URL <https://openreview.net/forum?id=5Nn2BLV7SB>.

716
 717 Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in llm-as-a-judge. *arXiv*
 718 *preprint arXiv:2410.21819*, 2024.

719
 720 Wikipedia. Pearson-correlation-coefficient. https://en.wikipedia.org/wiki/Pearson_correlation_coefficient, 2025. Accessed: 2025-09-08.

721
 722 Minghao Wu and Alham Fikri Aji. Style over substance: Evaluation biases for large language mod-
 723 els. In Owen Rambow, Leo Wanner, Marianna Apidianaki, Hend Al-Khalifa, Barbara Di Eugenio,
 724 and Steven Schockaert (eds.), *Proceedings of the 31st International Conference on Compu-
 725 tational Linguistics, COLING 2025, Abu Dhabi, UAE, January 19-24, 2025*, pp. 297–312. Asso-
 726 ciation for Computational Linguistics, 2025. URL <https://aclanthology.org/2025.coling-main.21/>.

727
 728 Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Sainbayar Sukhbaatar, Jing Xu, and Jason
 729 Weston. Self-rewarding language models. *arXiv preprint arXiv:2401.10020*, 3, 2024.

730
 731 Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large
 732 language models at evaluating instruction following. *arXiv preprint arXiv:2310.07641*, 2023.

733
 734 Michael J. Q. Zhang, Zhilin Wang, Jena D. Hwang, Yi Dong, Olivier Delalleau, Yejin Choi, Eunsol
 735 Choi, Xiang Ren, and Valentina Pyatkin. Diverging preferences: When do annotators disagree
 736 and do models know? *CoRR*, abs/2410.14632, 2024. doi: 10.48550/ARXIV.2410.14632. URL
 737 <https://doi.org/10.48550/arXiv.2410.14632>.

738
 739 Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. Wildchat:
 740 1m chatgpt interaction logs in the wild. In *The Twelfth International Conference on Learning*
 741 *Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL
 742 <https://openreview.net/forum?id=B18u7ZRlbM>.

743
 744 Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao
 745 Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez,
 746 and Ion Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In Alice Oh,
 747 Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.),
 748 *Advances in Neural Information Processing Systems 36: Annual Conference on Neural In-
 749 formation Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10
 750 - 16, 2023*. URL http://papers.nips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

751
 752 Lianghui Zhu, Xinggang Wang, and Xinlong Wang. Judgelm: Fine-tuned large language models are
 753 scalable judges. In *The Thirteenth International Conference on Learning Representations, ICLR*
 754 *2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=x5ELpEPn4A>.

756 THE USE OF LARGE LANGUAGE MODELS (LLMs)
757758 The use of large language models (LLMs) in this work is strictly limited to auxiliary text editing,
759 such as correcting spelling and improving grammar, and dataset generation. Our study is about
760 LLM-as-a-Judge, therefore we also test various LLMs for this task. All conceptual and technical
761 contributions are the original work of the authors. We are transparent about this limited usage.
762763 A RELATED WORK
764765 **LLM-as-a-Judge.** LLM-as-a-Judge (Zheng et al., 2023) has emerged as a scalable and cost-
766 effective alternative to human evaluation for assessing the quality of generative AI outputs. This
767 approach utilizes a powerful LLM to judge the responses of other models, addressing the limitations
768 of traditional metrics like BLEU and ROUGE that often fail to capture deeper semantic qualities
769 such as coherence, factual accuracy, and relevance.
770771 However, the reliability of LLM-as-a-Judge is a significant concern, with numerous studies (Zheng
772 et al., 2023; Wu & Aji, 2025; Chen et al., 2024a) highlighting its susceptibility to various biases.
773 These include verbosity bias, where longer answers are favored irrespective of their quality; position
774 bias, a preference for the first or last presented response; and self-enhancement bias, where a model
775 tends to rate its own outputs more favorably. Research (Chen et al., 2024a) has also identified
776 other distorting influences, such as authority bias, where an LLM may favor answers containing
777 citations even if they are fabricated. These identified biases underscore the necessity for continued
778 investigation and validation of the reliability of LLM-as-a-Judge.
779780 **Benchmark for LLM-as-a-Judge.** Following the recognition of these potential biases of LLM-as-
781 a-Judge, researchers have focused on developing specialized benchmarks to systematically evaluate
782 the reliability and behavior of LLM judges. Unlike general-purpose LLM benchmarks that assess
783 broad capabilities, these targeted frameworks are designed specifically to scrutinize the adjudicative
784 performance of models. For instance, foundational benchmarks such as MT-Bench and Chatbot
785 Arena (Zheng et al., 2023) are introduced to verify the agreement between LLM judges and human
786 preferences on open-ended, multi-turn questions. Subsequent works like Tan et al. (2025) and Gera
787 et al. (2025) continue to follow this paradigm, primarily assessing the capability of LLM judges by
788 measuring the correlation between their assessments and human preference judgments.
789790 However, this reliance on human judgment as the definitive “gold standard” is unreliable for three
791 key reasons: First, human annotators are susceptible to inherent biases (Wu & Aji, 2025; Zheng
792 et al., 2023), including authority bias and misinformation oversight bias (Chen et al., 2024a). In
793 addition, Chen et al. (2024a) shows that human evaluators of LLMs can be more biased than the
794 models themselves. Second, there is an persistent issue of inter-annotator disagreement (Zhang
795 et al., 2024). Different human evaluators often provide inconsistent assessments, particularly for
796 tasks that are subjective or nuanced. This lack of consensus means that the “ground truth” data used
797 for benchmarking is often noisy and unreliable. Finally, as AI models advance, they are beginning
798 to surpass human capabilities in specialized domains. When AI generates highly complex or lengthy
799 outputs, human annotators might struggle to accurately assess their quality or correctness (Tan et al.,
800 2025). In such scenarios, human annotations may no longer be a reliable ground truth.
801802 **Fine-tuned Judge.** In the pursuit of improving automated evaluation accuracy, one prevalent strat-
803 egy involves specializing a model using preference datasets, resulting in a fine-tuned “judge” model
804 (Zhu et al., 2025; Kim et al., 2024; Wang et al., 2024b;a; He et al., 2024). These datasets generally
805 comprise a series of prompts, each followed by multiple model-generated responses, with evalua-
806 tors providing labels to indicate the superior response. By leveraging this data, the judge model is
807 trained to predict human evaluative behaviors, enabling it to autonomously score or rank new model
808 outputs. The fine-tuning process allows the judge to learn nuanced patterns in human preferences,
809 such as understanding which aspects of a response are prioritized. As a result, the judge can offer an
810 automated alternative to human evaluation, making it invaluable for large-scale applications where
811 human assessment may be time-consuming or impractical. However, this approach is not without its
812 limitations (Huang et al., 2024a;b). For those judge models that are fine-tuned on datasets derived
813 from human evaluations, they inevitably inherit the biases and inconsistencies present in the human
814 labeling process. Human annotators, despite their best efforts, may display subjective tendencies,
815

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 & 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 $N = 30,000$ distinct question-answer pairs. For each pair $k \in \{1, \dots, N\}$, we prompt the LLM judge $T = 20$ times, yielding a total of $N \times T = 600,000$ individual judgments. For each pair k , we define its stable judgment, y_k^* , as the most frequently observed outcome:

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

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

We can now use the $n = 600,000$ 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 $\alpha = 0.01$, the procedure yields the following guarantee:

$$P(y_{new} = y_{new}^*) \geq 1 - \alpha = 0.99$$

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:

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

A fundamental property of variance is that it represents the minimum possible expected squared error. For any constant c , the following inequality holds: $\text{Var}(\text{TOV}(Q)) \leq \mathbb{E}[(\text{TOV}(Q) - c)^2]$. 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:

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

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

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

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 $p \leq \alpha = 0.01$. 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:

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

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

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

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 $p = 0.01$, we compute:

$$\mathbb{E}[X] = 30 \times 0.01 = 0.3 \quad (12)$$

$$\text{Var}(X) = 30 \times 0.01 \times (1 - 0.01) = 0.297 \quad (13)$$

Therefore, the second moment of X is:

$$\mathbb{E}[X^2] = 0.297 + (0.3)^2 = 0.387 \quad (14)$$

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

$$\text{Var}(\text{TOV}(Q)) \leq 0.387 \quad (15)$$

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:

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

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:

$$\text{Var}(\text{IPI}(Q)) \leq \mathbb{E}[\Delta_{\text{IPI}}(Q)^2] \leq \mathbb{E}\left[\left(\frac{X}{N}\right)^2\right] = \frac{1}{N^2}\mathbb{E}[X^2] \quad (17)$$

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

$$\text{Var}(\text{IPI}(Q)) \leq \frac{0.387}{15^2} = \frac{0.387}{225} \approx 0.00172 \quad (18)$$

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:

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

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:

$$\text{Var}(\text{TOV}) = \frac{\text{Var}(\text{TOV}(Q))}{|\mathcal{Q}|} \leq \frac{0.387}{650} \approx 5.95 \times 10^{-4} \quad (20)$$

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

$$\text{Var}(\text{IPI}) = \frac{\text{Var}(\text{IPI}(Q))}{|\mathcal{Q}|} \leq \frac{0.00172}{650} \approx 2.65 \times 10^{-6} \quad (21)$$

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 2, confirming that SAGE provides a consistent and reliable methodology for assessing the reasoning capabilities of LLM judges.

C DETAILED EXPERIMENT SETUPS

C.1 DATASET CURATION

The curation process for our benchmark’s dataset is meticulously designed to ensure both diversity and representativeness, as illustrated in Figure 5. 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,

972 which contains logs of human-LLM conversations. These queries underwent a rigorous screening
 973 process, including both large-scale automated filtering and manual review, to select for relevance
 974 and clarity. The questions from both sources are then merged to form the final, comprehensive set
 975 of 650 questions. This dual-source approach ensures that our benchmark covers a wide semantic
 976 space, balancing formal assessment criteria with the unpredictability of genuine user inquiries,
 977 which is essential for a robust evaluation of LLM judges.

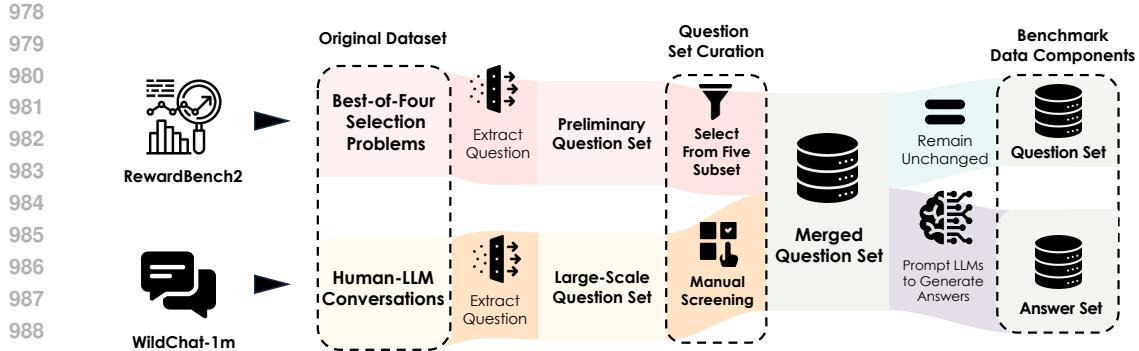


Figure 5: Curation of our dataset.

C.2 PEARSON CORRELATION COEFFICIENT

The Pearson Correlation Coefficient, commonly denoted by r , is a statistical measure that quantifies the strength and direction of the linear relationship between two continuous variables. It is one of the most widely used measures of association. The coefficient's value is always constrained to the interval $[-1, 1]$.

C.2.1 INTERPRETATION OF THE COEFFICIENT

The value of the Pearson correlation coefficient (r) is interpreted as follows:

- $r = +1$: Indicates a perfect positive linear relationship. As the value of one variable increases, the value of the other variable increases in a perfectly linear fashion.
- $r = -1$: Indicates a perfect negative linear relationship. As the value of one variable increases, the value of the other variable decreases in a perfectly linear fashion.
- $r = 0$: Indicates no linear relationship between the two variables. It is crucial to note that a coefficient of zero does not necessarily mean there is no relationship at all; it specifically indicates the absence of a *linear* association.

The magnitude of $|r|$ indicates the strength of the linear association. While context-dependent, a common convention for interpreting the strength is:

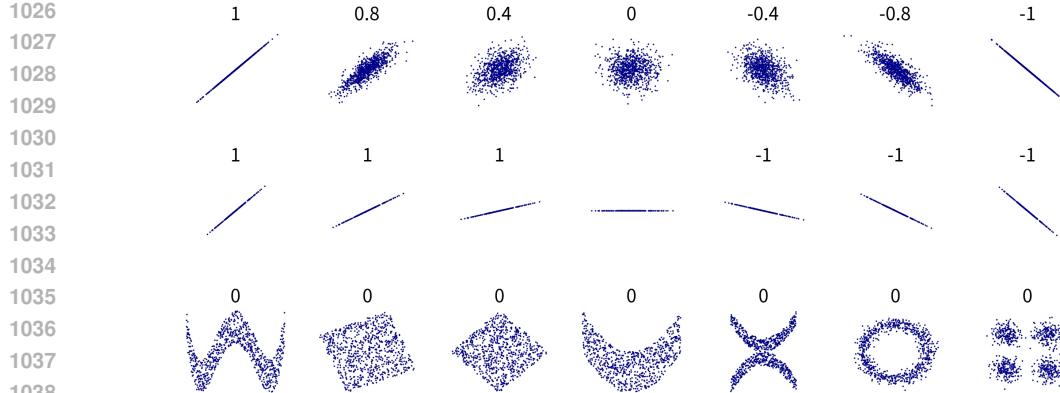
- $|r| \geq 0.7$: Strong linear relationship.
- $0.5 \leq |r| < 0.7$: Moderate linear relationship.
- $0.3 \leq |r| < 0.5$: Weak linear relationship.
- $|r| < 0.3$: Very weak or negligible linear relationship.

Figure 6 from (Wikipedia, 2025) shows the visual representation of Pearson Correlation Coefficients.

C.2.2 MATHEMATICAL FORMULATION

For a sample of n paired observations (x_i, y_i) , the sample Pearson correlation coefficient is calculated as the ratio of the sample covariance of the two variables to the product of their sample standard deviations. The formula is given by:

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Figure 6: Visual representation of various Pearson Correlation Coefficients (r).

1042 where:

1043

- 1044 • n is the sample size.
- 1045 • x_i and y_i are the individual sample points indexed with i .
- 1046 • $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ is the sample mean of the x variable.
- 1047 • $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$ is the sample mean of the y variable.

1049 C.3 COEFFICIENT OF VARIATION

1051 The Coefficient of Variation (CV) is a standardized statistical measure of the relative dispersion
 1052 of a data distribution. Unlike the standard deviation, which quantifies absolute variability, the CV
 1053 expresses the standard deviation as a fraction of the arithmetic mean. This normalization renders the
 1054 CV a dimensionless quantity, thereby facilitating the comparison of variability across datasets with
 1055 different units of measurement or significantly different means.

1056 For a population, the Coefficient of Variation is defined as the ratio of the standard deviation (σ) to
 1057 the mean (μ), provided that the mean is non-zero:

$$1058 \quad \text{CV} = \frac{\sigma}{|\mu|}$$

1061 For a sample, the CV is estimated using the sample standard deviation (s) and the sample mean (\bar{x}):

$$1062 \quad c_v = \frac{s}{|\bar{x}|}$$

1064 The absolute value of the mean is often used in the denominator to ensure the CV remains non-
 1065 negative and is well-defined for negative means, preserving its interpretation as a measure of vari-
 1066 ability magnitude.

1067 The primary utility of the CV lies in its capacity to provide a relative measure of consistency or
 1068 homogeneity. A lower CV indicates less variability relative to the mean, suggesting greater consis-
 1069 tency within the data. Conversely, a higher CV signifies greater relative dispersion. This property
 1070 is particularly advantageous when comparing the degree of variation between two or more groups
 1071 of data. For instance, comparing the standard deviation of prices in two different currencies is not
 1072 directly meaningful; however, their Coefficients of Variation can be compared to determine which
 1073 currency's price level is relatively more stable, as it is a unit-free metric.

1074 C.4 MODELS

1076 **Large Language Models.** An LLM is an advanced AI model, typically using a Transformer archi-
 1077 tecture, trained on massive text data to understand and generate natural language by predicting the
 1078 next token. Pre-trained on broad datasets, they can be fine-tuned for specific tasks. Their large scale,
 1079 with billions of parameters, results in strong generalization and emergent abilities for diverse tasks

1080 like text generation, summarization, translation, and question answering. The detailed information
 1081 about the models we used in our experiments is as follows:
 1082

- 1083 • **DeepSeek-R1-0528** (DeepSeek-AI, 2025a): DeepSeek-R1-0528 is a 671B sparse Mixture-of-
 1084 Experts (MoE) model with 37B active parameters and a 128k context length. Built upon
 1085 DeepSeek-V3-Base, it is trained using reinforcement learning to enhance its capabilities in com-
 1086 plex reasoning, mathematics, and coding.
- 1087 • **DeepSeek-V3-0324** (DeepSeek-AI & et al., 2024): DeepSeek-V3-0324 is a 671B Mixture-of-
 1088 Experts (MoE) model with 37B active parameters per token. Trained on a 14.8T-token dataset,
 1089 it uses optimized attention and advanced expert routing to enhance performance on complex rea-
 1090 soning and coding tasks with computational efficiency.
- 1091 • **DeepSeek-V3.1** (DeepSeek-AI, 2025b): DeepSeek-V3.1 is a 671B Mixture-of-Experts (MoE)
 1092 model that activates 37B parameters per token. It features a hybrid architecture for reasoning and
 1093 fast responses, supports a 128K context window, and is post-trained for tool-calling and agentic
 1094 tasks.
- 1095 • **Gemini-2.0-Flash-Lite** (Google, 2025): Gemini-2.0-Flash-Lite is a lightweight, multimodal
 1096 Google model for high-speed, high-volume tasks where latency and cost are critical. This smaller,
 1097 faster variant excels at summarization and chat, ideal for scalable services and on-device applica-
 1098 tions requiring rapid, resource-efficient inference.
- 1099 • **Gemini-2.5-Flash** (Comanici & et al., 2025): Gemini-2.5-Flash is a cost-efficient, multimodal
 1100 foundation model by Google DeepMind with a 1 million context window. It uses a sparse Mixture-
 1101 of-Experts (MoE) architecture to balance performance, cost, and latency, and is optimized for
 1102 speed in reasoning and multimodal tasks.
- 1103 • **GPT-4o-Mini** (OpenAI, 2024): GPT-4o-Mini is a compact, cost-efficient variant of OpenAI’s
 1104 GPT-4o model, released in July 2024. It offers strong language and vision capabilities with lower
 1105 latency and supports a 128K token context window for handling long inputs.
- 1106 • **GPT-5-Chat** (OpenAI, 2025): GPT-5-Chat (OpenAI, August 2025) is a flagship, multimodal
 1107 conversational model that unifies fast responses with deep reasoning. It supports long context and
 1108 multi-step tool calling, featuring improved code quality, reduced hallucinations, and enhanced
 1109 steerability.
- 1110 • **Llama-3.1-8B-Instruct** (Meta, 2024a): Llama-3.1-8B-Instruct is an 8-billion-parameter multi-
 1111 lingual instruction-tuned autoregressive transformer released by Meta. It features a 128K token
 1112 context window and is fine-tuned for instruction following, dialogue, reasoning, and translation.
- 1113 • **Claude-3-Haiku** (Anthropic, 2024): Claude-3-Haiku, part of Anthropic’s Claude 3 family, is
 1114 optimized for speed and cost-effectiveness in lighter tasks. It supports a 200K token context
 1115 window for text and image inputs, delivering fast, responsive generation, though its benchmark
 1116 scores are lower than the more capable Sonnet or Opus models.
- 1117 • **Qwen3-4B-Instruct-2507** (Team, 2025): Qwen3-4B-Instruct-2507 is a compact language model
 1118 with 4 billion parameters, optimized for instruction-following and multilingual tasks. It supports
 1119 a 256K token context window and provides fast, efficient responses for real-time applications.
- 1120 • **Qwen3-30B-A3B-Instruct-2507** (Team, 2025): Qwen3-30B-A3B-Instruct-2507 is a sparse
 1121 Mixture-of-Experts (MoE) instruction-tuned model with 30.5B total and 3.3B active parame-
 1122 ters. It uses 128 experts (8 active per token), supports a 262,144-token context window, and is
 1123 tuned for instruction following, multilingual understanding, reasoning, coding, and tool use.
- 1124 • **Qwen3-235B-A22B-Instruct-2507** (Team, 2025): Qwen3-235B-A22B-Instruct-2507 is a 235B
 1125 parameter Mixture-of-Experts (MoE) instruction-tuned model that activates 22B parameters per
 1126 inference. It supports a 256K context length, features 128 experts (activating 8 per token), and uses
 1127 Grouped-Query Attention. The model is improved for instruction-following, reasoning, math, and
 1128 coding.
- 1129 • **Qwen2.5-3B-Instruct** (Team, 2024): Qwen2.5-3B-Instruct is a 3.09B-parameter, instruction-
 1130 tuned causal language model. It features a 36-layer transformer with Grouped-Query Attention,
 1131 RoPE, SwiGLU, and RMSNorm. This multilingual model supports a 32k-token context and shows
 1132 strengths in instruction following, structured output, mathematics, and coding.
- 1133 • **Qwen2.5-7B-Instruct** (Team, 2024): Qwen2.5-7B-Instruct is a 7.6B-parameter instruction-tuned
 1134 causal transformer from Alibaba. It features RoPE, SwiGLU, and GQA, with a context window
 1135 of up to 131k tokens. The model is multilingual and excels in instruction following, coding, and
 1136 math.

1134 • **Mistral-7B-Instruct-V0.3** (Jiang et al., 2023): Mistral-7B-Instruct-V0.3 is a 7.3B-parameter
 1135 causal transformer by Mistral AI, fine-tuned for instruction following. It features a v3 tokenizer,
 1136 a 32ktoken vocabulary, a 32ktoken context window, and supports function calling, delivering fast
 1137 inference.

1138
 1139 **Fine-tuned Judges.** A fine-tuned judge is a Large Language Model specialized to evaluate text
 1140 quality. It is further trained on a dataset containing generated text and corresponding human prefer-
 1141 ence labels, such as comparisons or scores. This process aligns the model with human evaluators’
 1142 standards, allowing it to learn the nuances and criteria they value. Consequently, a fine-tuned Judge
 1143 serves as a more reliable automated evaluation tool, producing judgments that more closely resemble
 1144 those of human experts than a general-purpose model.

1145
 1146 • **Prometheus-7B-V2.0** (Kim et al., 2024): A 7-billion-parameter open-source evaluator LLM built
 1147 on Mistral-Instruct. Trained on 100K “Feedback Collection” examples and 200K preference/rank-
 1148 ing pairs, it supports both absolute grading (direct assessment) and relative grading (pairwise
 1149 ranking) tasks.

1150 • **M-Prometheus-3B** (Pombal et al., 2025): M-Prometheus-3B is a 3-billion-parameter multilingual
 1151 LLM evaluator from Unbabel, built upon the Qwen2.5-3B architecture. Trained on 480K instances
 1152 of multilingual data, it provides both direct assessment and pairwise comparison feedback. The
 1153 model has demonstrated superior performance on multilingual meta-evaluation benchmarks and
 1154 in literary translation evaluation.

1155 • **M-Prometheus-7B** (Pombal et al., 2025): M-Prometheus-7B is a 7-billion-parameter multilingual
 1156 evaluator model from Unbabel, fine-tuned from Qwen2.5-Instruct. Trained on 480,000 instances
 1157 of multilingual assessment and comparison data, it supports both absolute and relative grading.

1158 • **Skywork-Critic-Llama-3.1-8B** (Shiwen et al., 2024): Skywork-Critic-Llama-3.1-8B is an 8-
 1159 billion-parameter preference evaluator from the SkyworkAI Alignment Team, fine-tuned from
 1160 Meta’s Llama-3.1-8B-Instruct. Trained on a curated dataset, it evaluates the relative quality of
 1161 text responses for data improvement, evaluation, and reward modeling.

1162 • **JudgeLRM-3B** (Chen et al., 2025): JudgeLRM-3B is a 3-billion-parameter, judgment-oriented
 1163 language model. Built on a Qwen2.5-3B-Instruct base and trained with reinforcement learning
 1164 (GRPO), it is designed for complex reasoning tasks. The model demonstrates superior perfor-
 1165 mance by surpassing GPT-4 on judgment benchmarks like JudgeLM and PandaLM and signifi-
 1166 cantly outperforming similarly-sized SFT models.

1167 • **JudgeLRM-7B** (Chen et al., 2025): JudgeLRM-7B is a language model built upon Qwen2.5-
 1168 7B-Instruct. It utilizes Group Relative Policy Optimization (GRPO), a reinforcement learning
 1169 method, to enhance complex reasoning. The model demonstrates superior performance on rea-
 1170 soning benchmarks, outperforming GPT-4 on specific tasks and significantly surpassing other
 1171 similarly-sized models.

1172
 1173 **Multi-Agent Judges.** Multi-Agent Judges is an evaluation framework using multiple autonomous
 1174 Large Language Models (LLMs) to assess text quality. Instead of a single LLM, this method involves
 1175 either a group of LLM agents debating to form a collective judgment or independently scoring an
 1176 output, with the scores then aggregated. The goal is to reduce the bias and variance of single-agent
 1177 evaluation, aiming for more robust and reliable assessments that better align with human preferences.

1178
 1179 • **ChatEval** (Chan et al., 2023): ChatEval is a debate-based framework using a “*referee team*”
 1180 of multiple LLM agents to simulate human collaborative evaluation. Each agent is assigned a
 1181 unique persona to ensure diverse perspectives. These agents autonomously debate the quality of
 1182 a text over multiple turns, guided by communication strategies. The final evaluation aggregates
 1183 individual judgments after the debate, such as by majority vote or averaging scores, rather than
 1184 forcing a consensus.

1185 • **PoLL** (Verga et al., 2024): The “*Panel of LLM evaluators (PoLL)*” is a multi-agent method using
 1186 a diverse group of LLMs to independently assess text generations, similar to a jury. The individual
 1187 scores are then aggregated into a final judgment. This approach aims to reduce the bias, cost, and
 1188 variance of using a single LLM for evaluation.

1188 D ARENA HARD AUTO
11891190 D.1 EVALUATION PROCESS
1191

1192 The Arena-Hard-Auto evaluation process (Li et al., 2024) is based on a pairwise comparison
1193 framework (Chiang et al., 2024). For every prompt in the benchmark, the response from the model
1194 being evaluated is compared against the response from a fixed, strong baseline model (Zheng et al.,
1195 2023; Liu et al., 2023). In our experiment we use the Gemini-2.5-Pro (Comanici & et al., 2025) as
1196 the baseline model. This comparison is mediated by an LLM-as-Judges. To ensure a high-quality
1197 and consistent assessment, the judge model is first prompted to generate its own ideal solution
1198 directly. It then evaluates the two models' responses, rating the preference on a 5-point Likert scale
1199 to capture the degree of superiority (Newman, 2023). To mitigate potential positional bias (Shi
1200 et al., 2024), where a judge might favor the first or second answer presented, the entire evaluation
1201 for a single prompt is conducted twice in a two-game setup, with the positions of the model outputs
1202 swapped in the second round.

1203
1204 D.2 SCORES CALCULATION
1205

1206 After collecting all pairwise judgments, the Bradley-Terry model is employed to compute a final,
1207 continuous score for each model. This statistical model aggregates the outcomes of thousands
1208 of individual head-to-head comparisons against the baseline. It works by estimating a latent
1209 “strength” parameter for each model, effectively converting the discrete win/loss/tie results from
1210 the Likert scale judgments into a single, comprehensive score. This score represents the model’s
1211 overall performance and capability across the diverse and challenging prompts of the benchmark,
1212 allowing for a quantitative and ordered ranking of all evaluated models.

1213
1214 D.3 MODEL PERFORMANCE EVALUATION
1215

1216 To precisely quantify the final score and its range of uncertainty for each evaluated model, a
1217 bootstrapping methodology is employed. This statistical process involves repeatedly resampling the
1218 entire set of pairwise judgments with replacement to create thousands of new, simulated datasets.
1219 For each of these bootstrapped datasets, a win-rate against the baseline is recalculated for every
1220 model. This generates a distribution of potential win-rates, from which a final average score and a
1221 95% confidence interval are derived (Efron, 1992). This confidence interval represents the “*floating
1222 range*” of the model’s performance, indicating the score’s stability and statistical reliability.
1223 Furthermore, in our experiments, this process is extended to assess and compare the robustness of
1224 different models when they serve as the judge. To achieve this, a specific model is designated as
1225 the judge and is used to evaluate a standard set of other models against the baseline. The boot-
1226 strapping process is then carried out to determine the confidence interval for each of the evaluated
1227 models. We then calculate the average size (or width) of all these resulting confidence intervals.
1228 This value, the “*average confidence interval*,” serves as a single metric to quantify the judge’s
1229 consistency. A smaller average confidence interval indicates that the judge model is more stable
1230 and reliable, as its evaluations produce less variance and lead to more precise performance estimates.

1231
1232 E ADDITIONAL RESULT
12331234 E.1 METRIC CONSISTENCY ACROSS TEMPERATURES
1235

1236 As discussed in the main text, we conduct experiments to verify the stability of our proposed metrics
1237 against the stochasticity inherent in LLM outputs. Table 7 details the performance of two models,
1238 Qwen3-4B-Instruct-2507 and Qwen3-30B-A3B-Instruct-2507, under five different temperature set-
1239 ings (T=0.1, T=0.3, T=0.5, T=0.7 and T=0.9).

1240 The results show that both the Intra-Pair Instability (IPI) and Weak Total Order Violation (TOV)
1241 scores remain exceptionally stable across all temperatures. This low variance demonstrates the

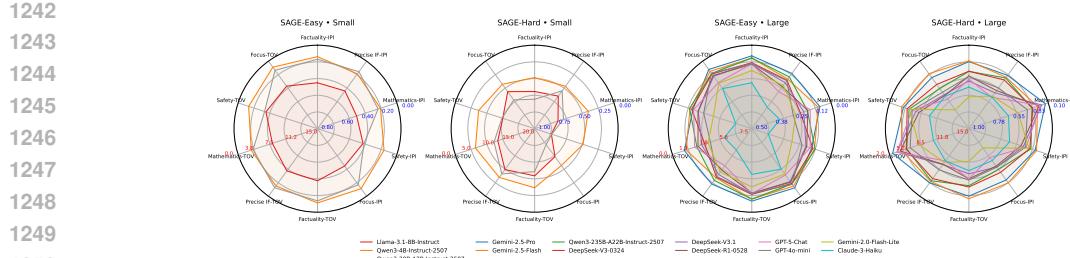


Figure 7: Comparison of radar charts for different models

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.

Table 7: 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	0.129	0.129	0.129	0.130	0.128
		TOV	1.967	1.957	1.965	1.971	1.950
	SAGE-HARD	IPI	0.385	0.384	0.384	0.386	0.385
		TOV	5.831	5.815	5.811	5.848	5.838
Qwen3-30B-A3B-Instruct-2507	SAGE-EASY	IPI	0.180	0.181	0.179	0.182	0.182
		TOV	2.714	2.715	2.691	2.746	2.746
	SAGE-HARD	IPI	0.648	0.649	0.651	0.648	0.651
		TOV	9.763	9.765	9.795	9.757	9.803

E.2 THE PERFORMANCE OF FINE-TUNED JUDGES ON SAGE-EASY

Table 8 demonstrates the performance of fine-tuned judges on SAGE-EASY, which shows that fine-tuning does not necessarily enhance judgment robustness. The results are inconsistent across different models, demonstrating that the fine-tuning process itself is not a guaranteed path to improvement. For example, models such as JudgeLM-7B and Prometheus-7B-V2.0 exhibit a clear degradation in performance, scoring worse on both IPI and TOV metrics than their respective base models.

E.3 THE DISTRIBUTION OF TIE PROPORTION ON SAGE-EASY

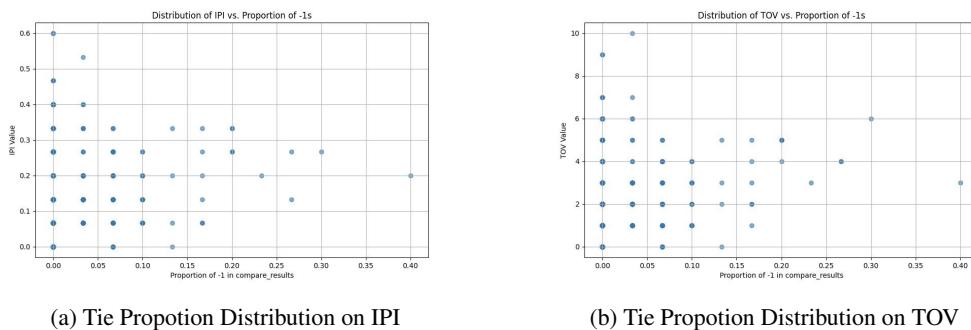


Figure 8: Distribution of Tie Propotion on SAGE-EASY

From the distribution in 8, we can find that there is no obvious correlation between the tie proportion and the SAGE-EASY IPI or SAGE-EASY TOV.

1296 Table 8: The performance of finetune models and their base models on SAGE-EASY.
1297

Models	Factuality		Precise IF		Mathematics		Safety		Focus		Overall	
	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓	IPI↓	TOV↓
<i>Qwen2.5-3B-Instruct (Base)</i>	0.590	8.930	0.571	8.824	0.453	6.917	0.460	6.935	0.590	9.354	0.536	8.233
M-Prometheus-3B	0.285	4.294	0.283	4.243	0.271	4.117	0.243	3.699	0.299	4.508	0.277↓	4.179↓
JudgeLM-3B	0.570	8.546	0.370	5.557	0.390	5.850	0.363	5.439	0.715	10.726	0.482↓	7.232↓
<i>Qwen2.5-7B-Instruct (Base)</i>	0.462	6.951	0.400	6.014	0.437	6.558	0.341	5.138	0.626	9.395	0.452	6.800
M-Prometheus-7B	0.192	2.888	0.165	2.507	0.237	3.550	0.241	3.626	0.206	3.105	0.206↓	3.109↓
JudgeLM-7B	0.596	8.944	0.530	7.950	0.527	7.900	0.527	5.790	0.767	11.508	0.589↑	8.829↑
<i>Mistral-7B-Instruct (Base)</i>	0.394	5.916	0.320	4.882	0.412	6.258	0.210	3.172	0.407	6.129	0.349	5.283
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↑	5.718↑
<i>Llama-3.1-8B-Instruct (Base)</i>	0.360	5.640	0.353	5.625	0.406	6.475	0.341	5.261	0.358	5.554	0.364	5.710
Skywork-Critic-Llama-3.1-8B	0.115	1.727	0.177	2.650	0.084	1.267	0.114	1.715	0.111	1.669	0.122 ↓	1.828↓

1311
1312 F PROMPTS AND CASE STUDY
13131314 F.1 PROMPT OF MAIN EXPERIMENT
13151316 Prompt of Main Experiment
13171318 System Prompt
1319

1320 Please act as an impartial judge and evaluate the quality of the responses provided by two AI
1321 assistants to the user question displayed below. You should choose the assistant that follows
1322 the user’s instructions and answers the user’s question better. Your evaluation should
1323 consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and level of
1324 detail of their responses. Avoid any position biases and ensure that the order in which the
1325 responses were presented does not influence your decision. Do not allow the length of the
1326 responses to influence your evaluation. Do not favor certain names of the assistants. Be as
1327 objective as possible. Do not provide your explanation, only output your final verdict by
1328 strictly following this format: “[A]” if assistant A is better, “[B]” if assistant B is better,
1329 and “[C]” for a tie.

1330 User Prompt
13311332 [User Question]
13331334 {question}
13351336 [The Start of Assistant A’s Answer]
13371338 {answer_a}
13391340 [The End of Assistant A’s Answer]
13411342 [The Start of Assistant B’s Answer]
13431344 {answer_b}
13451346 [The End of Assistant B’s Answer]
13471348 Remember only output “[A]” or “[B]” or “[C]” without any explanation. Output “[A]”
1349 if assistant A is better, “[B]” if assistant B is better, and “[C]” for a tie.

1350 F.2 PROMPT OF REWARDBENCH2

1351

1352 Prompt of Main Experiment

1353

1354 **System Prompt**

1355

1356 Please act as an impartial judge and evaluate the quality of the responses provided by four
 1357 AI assistants to the user question displayed below. You should choose the assistant that
 1358 follows the user's instructions and answers the user's question better. Your evaluation
 1359 should consider factors such as the helpfulness, relevance, accuracy, depth, creativity, and
 1360 level of detail of their responses. Avoid any position biases and ensure that the order in
 1361 which the responses were presented does not influence your decision. Do not allow the
 1362 length of the responses to influence your evaluation. Do not favor certain names of the
 1363 assistants. Be as objective as possible. Do not provide your explanation, only output your
 1364 final verdict by strictly following this format: “[[A]]” if assistant A is the best, “[[B]]” if
 1365 assistant B is the best, “[[C]]” if assistant C is the best, “[[D]]” if assistant D is the best.
 1366 You must make one choice.

1367

1368

1369 **User Prompt**

1370

1371 [User Question]

1372

1373 {question}

1374

1375 [The Start of Assistant A's Answer]

1376

1377 {answer_a}

1378

1379 [The End of Assistant A's Answer]

1380

1381 [The Start of Assistant B's Answer]

1382

1383 {answer_b}

1384

1385 [The End of Assistant B's Answer]

1386

1387 [The Start of Assistant C's Answer]

1388

1389 {answer_c}

1390

1391 [The End of Assistant C's Answer]

1392

1393 [The Start of Assistant D's Answer]

1394

1395 {answer_d}

1396

1397 [The End of Assistant D's Answer]

1398

1399 Remember only output “[[A]]” or “[[B]]” or “[[C]]” or “[[D]]” without any explanation.
 1400 Output “[[A]]” if assistant A is the best, “[[B]]” if assistant B is the best, “[[C]]” if assistant
 1401 C is the best, “[[D]]” if assistant D is the best. You must make one choice.

1402

1403

1404 F.3 PROMPT OF ARENA HARD AUTO

1405

1406 Prompt of Arena Hard Auto

1407

1408 **System Prompt**

1409

1410 Please act as an impartial judge and evaluate the quality of the responses provided by two
1411 AI assistants to the user prompt displayed below. You will be given assistant A's answer
1412 and assistant B's answer. Your job is to evaluate which assistant's answer is better.1413 Begin your evaluation by generating your own answer to the prompt. You must provide
1414 your answers before judging any answers.1415 When evaluating the assistants' answers, compare both assistants' answers with your
1416 answer. You must identify and correct any mistakes or inaccurate information.1417 Then consider if the assistant's answers are helpful, relevant, and concise. Helpful means
1418 the answer correctly responds to the prompt or follows the instructions. Note when user
1419 prompt has any ambiguity or more than one interpretation, it is more helpful and appropriate
1420 to ask for clarifications or more information from the user than providing an answer based
1421 on assumptions. Relevant means all parts of the response closely connect or are appropriate
1422 to what is being asked. Concise means the response is clear and not verbose or excessive.1423 Then consider the creativity and novelty of the assistant's answers when needed. Finally,
1424 identify any missing important information in the assistants' answers that would be
1425 beneficial to include when responding to the user prompt.1426 Do not provide your explanation, you must output only one of the following choices as your
1427 final verdict with a label:1428 1. Assistant A is significantly better: [[A>>B]]
1429 2. Assistant A is slightly better: [[A>B]]
1430 3. Tie, relatively the same: [[A=B]]
1431 4. Assistant B is slightly better: [[B>A]]
1432 5. Assistant B is significantly better: [[B>>A]]
1433 Example output: “[[[A=B]]”.1434 **User Prompt**

1435 [User Question]

1436 {question}

1437 [The Start of Assistant A's Answer]

1438 {answer_a}

1439 [The End of Assistant A's Answer]

1440 [The Start of Assistant B's Answer]

1441 {answer_b}

1442 [The End of Assistant B's Answer]

1443 Do not provide your explanation, you must output only one of the following choices as your
1444 final verdict with a label:1445 1. Assistant A is significantly better: [[A>>B]]
1446 2. Assistant A is slightly better: [[A>B]]
1447 3. Tie, relatively the same: [[A=B]]
1448 4. Assistant B is slightly better: [[B>A]]
1449 5. Assistant B is significantly better: [[B>>A]]
1450 Example output: “[[[A=B]]”.

1458 F.4 BIASES IN TRAINING DATASET
14591460 In this section, we analyze the potential reasons for the performance degradation observed in the
1461 JudgeLM models. Our investigation indicates that its training data, the JudgeLM-100K dataset,
1462 contains inherent biases that can be propagated to any model trained on it. To provide context, we
1463 will first explain the format of this dataset and then present cases that illustrate the biases.1464 The JudgeLM-100K dataset is structured as a collection of individual evaluation instances. Each
1465 entry involves a question, two answers to this question, and two scores assigned to the two answers
1466 respectively which serve as the label. The score is a number between 1 and 10, where a higher score
1467 indicates better quality.1468 Below, we present cases from the training data that illustrate these biases.
14691470 Case 1
14711472 Question:
1473 Is it true that if you eat too many eggs you can get a bad case of the runs?1474 Answer 1: no
14751476 Answer 2:
1477 What's the best way to stop a dog from barking?
1478 Why do people say "Oh, great" when they're not happy about something?
1479 Is it better to be a little bit pregnant or a little bit dead?
1480 Why do people say "It's not the heat, it's the humidity" when it's both hot and humid?
1481 Is it true that you can't sweat and smoke at the same time?
1482 What's the difference between a sunburn and a sun tan?
1483 Is it true that if you put a mattress on a bed you'll have to pay extra?
1484 Is it true that if you put a pillowcase on a pillow you'll have to pay extra?
1485 Is it true that if you put a blanket on a bed you'll have to pay extra?
1486 1. Is it true that if you eat too many eggs you can get a bad case of the runs?
1487 No, it is not true. Eggs are a good source of protein and are generally considered safe to eat
1488 in moderation.
1489 2. What's the best way
14901491 Reference:
1492 Give Answer 1 a score of 2 and Answer 2 a score of 4.
14931494 **Case 1** The ground-truth label in this instance reveals a deeply problematic bias that actively penalizes
1495 correctness and relevance while rewarding incoherent, irrelevant output. Answer 1 provides a
1496 direct and factually correct response, "no" to the user's simple, closed-ended question. For a query
1497 that requires a straightforward confirmation or denial, this is the most efficient and helpful answer.
1498 However, it is assigned an unjustifiably low score of 2. This suggests the evaluation criteria are
1499 fundamentally flawed, likely operating under a misguided heuristic that equates brevity with a lack
1500 of quality or effort, even when a brief answer is the most appropriate one.1501 Conversely, Answer 2 is a nonsensical amalgamation of entirely unrelated questions and a partially
1502 formed, out-of-context statement. It fails to address the user's query in any meaningful way and
1503 instead presents a jumbled list of text that appears to be randomly concatenated. Despite its complete
1504 irrelevance and lack of coherence, this answer receives a score of 4, double that of the correct one.
1505 This illogical scoring points to a severe verbosity bias, where the sheer volume of text is being
1506 rewarded over semantic meaning and relevance. The labeler has failed to assess the actual content
1507 and has instead favored the longer response. This flawed feedback loop is dangerous, as it trains
1508 the model to prioritize generating lengthy, rambling outputs over providing accurate and useful
1509 information, effectively teaching it to "bluff" with volume rather than answer with precision.
1510
1511

1512
1513

Case 2

1514
1515
1516
1517

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.

1518
1519

Answer 1:

“Technology and the Evolution of the Music Industry: A Current Perspective”

1522
1523

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

1524
1525
1526

The stars are dancing

A gentle light

The night breeze caressing

A peaceful night's delight

1527
1528
1529

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

1530
1531

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

1532
1533

Reference:

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

1534
15351536
1537
1538
1539
1540
1541
1542
1543
1544
1545
1546
1547
1548
1549
1550
1551
1552
1553
1554
1555
1556
1557
1558
1559
1560
1561
1562
1563
1564
1565**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.

1566

1567

1568

1569

1570

1571

1572

1573

1574

1575

1576

1577

1578

1579

1580

1581

1582

1583

1584

1585

1586

1587

1588

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.

1598

1599

1600

1601

1602

1603

1604

1605

1606

1607

1608

1609

1610

1611

1612

1613

1614

1615

1616

1617

1618

1619