UC Berkeley
Recent Work

Title
Preprint OpenReliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias

Permalink
https://escholarship.org/uc/item/5gx838hq

Authors
Norman, Justin
Rivera, Michael U
Hughes, Alex

Publication Date
2026-11-20

DOI
10.5281/zenodo.23988546

Copyright Information
This work is made available under the terms of a Creative Commons Attribution License, available at https://creativecommons.org/licenses/by/4.0/

eScholarship.org Powered by the California Digital Library
University of California

Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias

Justin D. Norman and Michael U. Rivera and D. Alex Hughes
{justin.norman, michaelrivera, dhughes}@berkeley.edu
UC Berkeley School of Information
102 South Hall Drive, Berkeley CA, 94708

Abstract

LLM-as-a-Judge has become the dominant evaluation paradigm for language models, but judge validation in practice relies on exact-match agreement, a metric that does not correct for chance and systematically overstates discriminative ability. We present the largest systematic evaluation of LLM-as-a-Judge to date: 21 judges from nine providers across MT-Bench, JudgeBench, and RewardBench, evaluated under three protocols (agreement, consistency, bias audit) over 118 runs and approximately 541,000 individual judgments. Four findings emerge, consistent across the full cohort, including the April 2026 frontier: kappa deflation between exact match and Cohen's κ is universal (33–41 pp on MT-Bench), judge rankings shift by up to 14 positions across benchmarks, high test–retest reliability (> 0.95) coexists with severe position bias (> 0.10) in two production-deployed judges (instantiating a consistency–bias paradox), and verbosity bias is small (< 0.011) across our cohort under a single pairwise rubric. We distill these into a Minimum Viable Validation Protocol.

1 Introduction

Large language models (LLMs) increasingly evaluate large-scale content (Zheng et al., 2023). These "judgments" rely on the model's ability to mimic a human evaluator on tasks requiring analytical acuity. Such judges are widely deployed in production evaluation pipelines across hundreds of documented LLMOps deployments (Liu, 2025; Strick van Linschoten, 2025). This proliferation has accelerated despite well-documented reliability issues with LLM use for reasoning tasks (Wang et al., 2024; Wu and Aji, 2025; Panickssery et al., 2024; Stureborg et al., 2024; Schroeder and Wood-Doughty, 2024; Gu et al., 2024; Li et al., 2024b; Shi et al., 2025). Surveys of LLM-as-Judge (LLMaJ) reliability identify a recurring set of failure modes: inconsistency across prompts and runs, systematic scoring biases, weak domain-specific calibration, and the absence of meta-evaluation standards for comparing judges on equal footing.

Despite these challenges, judge validation continues to rely heavily on exact match. Established judge benchmarks—MT-Bench (Zheng et al., 2023), RewardBench (Lambert et al., 2025), JudgeBench (Tan et al., 2025)—and recent successors (Malik et al., 2026; Whitehouse et al., 2025; Jiang et al., 2025) privilege raw agreement as the headline validation metric. This family of metrics does not correct for chance (Cohen, 1960; Hayes and Krippendorff, 2007) and overstates discriminative ability (Guerdan et al., 2025; Han et al., 2025; Collot et al., 2025).

No prior work has conducted a large-scale, cross-benchmark, multi-protocol evaluation of judge reliability across model families and generations. We present the largest such evaluation: 21 models from nine providers across three benchmarks and three evaluation protocols, producing 541,000 judgments over 118 runs.

We report five principal findings, including two diagnostic concepts. First, **kappa deflation**: raw agreement overstates chance-corrected discrimination by 33–41pp in all 21 evaluated models. Second, judge rankings are non-transferable: models shift by as many as 14 positions across benchmarks. Third, the **consistency–bias paradox**: high test–retest reliability often masks severe position bias. Fourth, verbosity bias is much reduced: all 21 models register $<$0.011, in sharp contrast to the 20–40% variance reported in 2023 literature. Fifth, JudgeBench discriminates $4.5\times$ more sharply than MT-Bench (60.4pp vs. 13.5pp κ spread).

2 Related Work

The LLMaJ paradigm originated with MT-Bench and Chatbot Arena (Zheng et al., 2023). Rapid adoption followed across diverse evaluation set-

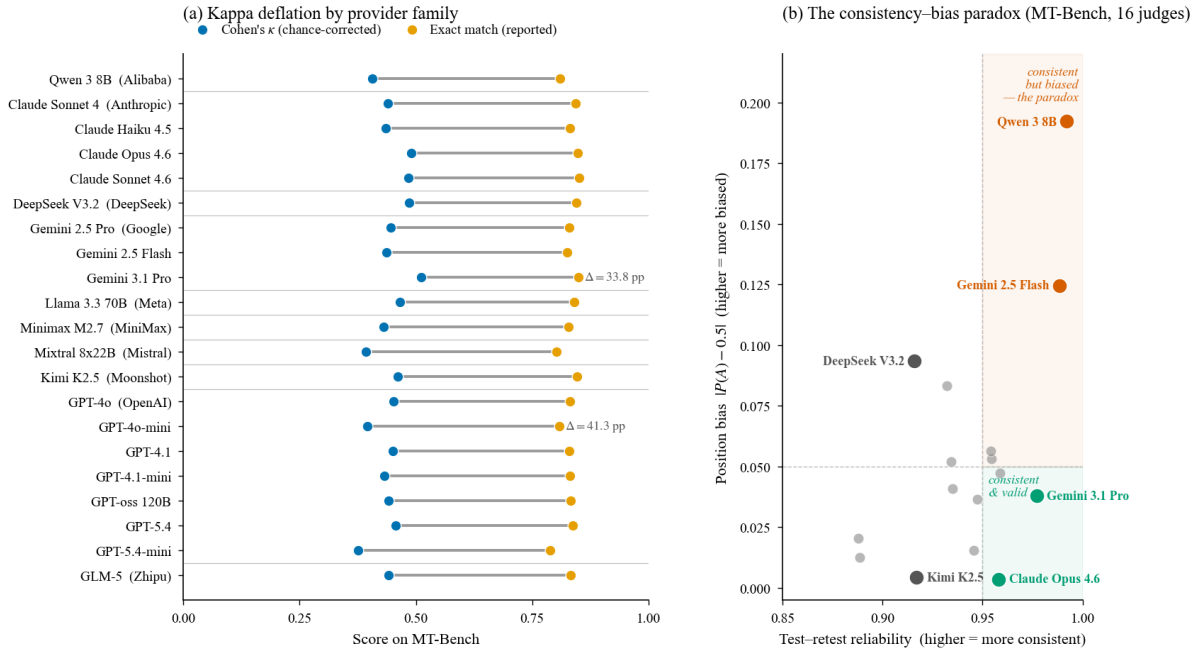


Figure 1: **Two diagnostic failures of LLM-as-a-Judge across 21 judges.** Panel **(a),** ***kappa deflation***: every judge's exact-match score (orange) exceeds its chance-corrected agreement (Cohen's κ, blue) on MT-Bench by between 33.8 and 41.2 percentage points, regardless of provider, scale, or generation; the grey segment between the two markers is the deflation gap. Panel **(b),** ***the consistency–bias paradox***: high test–retest reliability (≥ 0.95) coexists with severe position bias (> 0.10) for Qwen 3 8B and Gemini 2.5 Flash, both of which occupy the upper-right "consistent but biased" quadrant; the most reproducible judges are among the least valid. Panel (b) covers the 16 judges evaluated on both consistency and bias-audit protocols.

tings, including G-Eval for natural-language-generation tasks (Liu et al., 2023), AlpacaEval for instruction-following (Dubois et al., 2024), Arena-Hard for separating model performance from crowdsourced preference data (Li et al., 2025b), and WildBench for real-user prompts (Lin et al., 2025). LLMaJ is now widely deployed in industrial evaluation pipelines, and has emerged as the dominant pattern for reference-free scoring across hundreds of production LLM deployments (Strick van Linschoten, 2025).

Evaluation of LLMaJ has expanded rapidly in response to the widespread adoption of the approach. The field faces distinct challenges surrounding reliability, consistency, and bias mitigation, which have prompted some emergent remedies (Gu et al., 2024; Li et al., 2025a). For example, prior work has documented biases inherited from pretraining corpora, difficulty adapting evaluation criteria across domains, and unresolved questions about consensus among cooperating judge models (Li et al., 2024a). Other work identifies inconsistency under temperature and prompt variation (Stureborg et al., 2024), introduces reliability coefficients (Schroeder and Wood-Doughty, 2024), evaluates rating-indeterminacy (Guerdan et al., 2025), and examines latent dimensions of reliability (Choi et al., 2026). Closest in scale to our work, Bavaresco et al. (2025) surveys 11 LLM judges across 20 NLP evaluation tasks and recommends careful validation against task-specific annotation before deployment.

2.1 LLM Judge Metrics and Definitions

Exact match is the proportion of items on which the judge's verdict matches the human label. We follow Zheng et al. (2023) in reporting the tie-excluded variant for pairwise comparisons, restricting the denominator to items on which the human label is non-tie (Wang et al., 2025). Despite being widely used, exact match does not correct for agreement expected by chance, and as a result exact match is sensitive to the label distribution of the underlying benchmark.

Cohen's κ is a chance-corrected measure of agreement between two raters on categorical labels:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \tag{1}$$

where p_o is the observed agreement and p_e is the agreement expected by chance under the marginal label distributions of the two raters (Cohen, 1960). We use κ as our primary agreement metric and compare each judge against the human label set on

all benchmarks. *Krippendorff's* α (Krippendorff, 2011) is a reliability coefficient that accommodates more than two raters as well as ordinal and interval scales. The two measures coincide for the two-rater nominal case, but α extends naturally to the multi-run consistency setting introduced below.

Test-retest reliability is the agreement of a judge with itself across independent re-evaluations of the same items. The concept captures the stability of verdicts under identical conditions, and is a necessary but insufficient characteristic of a quality judge. *Self-consistency* is the proportion of items on which the majority verdict across N runs agrees with each individual run, capturing within-judge agreement at the item level and complementing the corpus-level test–retest coefficient.

Position flip rate is the fraction of items for which the judge's verdict changes when the response order is swapped (Zheng et al., 2023), providing an item-level measure of position sensitivity that complements the aggregate position-bias statistic defined next. *Position bias* is the tendency of a judge to favor response in a particular position (Zheng et al., 2023; Wang et al., 2024; Shi et al., 2025). Across a range of judge models, flip rates range from 25% to 50% (Wang et al., 2024; Shi et al., 2025). Position-swap debiasing raises within-judge consistency from roughly 60% to 85% (Li et al., 2024b).

Verbosity bias is the tendency of a judge to prefer longer responses regardless of content quality (Wu and Aji, 2025; Dubois et al., 2024). We operationalize verbosity bias as the Pearson correlation between the response-length differential $|\text{len}(A)| - |\text{len}(B)|$ and the judge's verdict.

Kappa deflation is the difference between exact match and Cohen's κ for a given judge–benchmark pair:

$$\Delta_\kappa(j, b) = \text{EM}(j, b) - \kappa(j, b). \qquad (2)$$

Δ_κ quantifies how much raw agreement overstates chance-corrected discriminative ability. We introduce this term to name a phenomenon that, while mathematically a direct consequence of Cohen's correction (Cohen, 1960), has not been systematically measured or reported at scale across modern LLM judges. On MT-Bench we observe $\Delta_\kappa \in [33.8, 41.2]$ percentage points across all 21 models tested.

Consistency-bias paradox is the empirical observation that high test-retest reliability ($\alpha > 0.95$) can coexist with severe position bias in the same judge model, such that the judge is highly reproducible but not valid. We introduce this term to formalize a failure mode reachable by reporting test-retest alone (Stureborg et al., 2024; Schroeder and Wood-Doughty, 2024; Chen et al., 2024). The paradox arises because test-retest measures the stability of outputs and not the correctness of the underlying decision process: a judge that deterministically favors position A across runs would achieve a perfect test-retest score, but would also exhibit maximum-possible position bias.

2.2 Attempts to Improve Reliability

A growing body of work both documents the reliability limitations of LLMaJ systems, and then develops specialized architectures intended to mitigate them. Test-retest studies show substantial temperature sensitivity: same-verdict rates are above 95% when temperature is set to 0, but fall to as low as 70% when temperature is increased to 1 (Stureborg et al., 2024; Haldar and Hockenmaier, 2025). Recent approaches have separated intrinsic consistency from human-alignment using Item Response Theory models (Choi et al., 2026). Another line of work develops dedicated judge architectures rather than relying on general-purpose language models. These experiments show some promise of reducing position bias (Zhu et al., 2025) and cost (Verga et al., 2024) while approaching general-purpose model performance (Kim et al., 2024a,b). Our study complements these efforts by evaluating reliability at scale across 21 judge models and three benchmarks under a common experimental protocol.

3 Methodology

Model Selection We evaluate 21 general-purpose model LLM judges drawn from 9 providers grouped into three capability tiers. Table 1 lists each model with its provider, release date, tier, and per-token cost. Tier 1 consists of widely-deployed production judges; Tier 2 consists of cost-conscious models; Tier 3 consists of April 2026 frontier models and open-source systems. The 21 models span parameter counts from 8B to over 100B.

Benchmarks We evaluate every judge on three benchmarks of increasing difficulty. MT-Bench contributes 2,391 pairwise comparisons with expert human judgments and is the most widely-used

Model	Provider	Tier	Released	Input $/MTok
GPT-4o	OpenAI	1	2024-05	2.50
GPT-4o-mini	OpenAI	1	2024-07	0.15
GPT-4.1	OpenAI	1	2025-04	2.00
Gemini 2.5 Pro	Google	1	2025-03	1.25
Gemini 2.5 Flash	Google	1	2025-04	0.30
Claude Haiku 4.5	Anthropic	1	2025-10	1.00
Llama 3.3 70B	Meta	1	2024-12	0.90
Qwen 3 8B	Alibaba	1	2025-04	0.20
Mixtral 8x22B	Mistral	2	2024-04	1.20
GPT-4.1-mini	OpenAI	2	2025-04	0.40
Claude Sonnet 4	Anthropic	2	2025-05	3.00
GPT-5.4	OpenAI	3	2026-03	2.50
GPT-5.4-mini	OpenAI	3	2026-03	0.75
Claude Opus 4.6	Anthropic	3	2026-02	5.00
Claude Sonnet 4.6	Anthropic	3	2026-02	3.00
Gemini 3.1 Pro	Google	3	2026-02	2.00
GPT-oss 120B	OpenAI	3	2025-08	0.15
Minimax M2.7	MiniMax	3	2026-03	1.00
DeepSeek V3.2	DeepSeek	3	2025-12	0.56
Kimi K2.5	Moonshot	3	2026-01	0.60
GLM-5	Zhipu	3	2026-02	1.00

Table 1: The 21 evaluated judges. Release dates were verified against provider announcements. Full primary-source URLs are documented in our supplementary material. List prices are USD per million input tokens at evaluation time (April 2026). Output-token cost is omitted because judges produce short verdicts and input cost dominates total spend.

judge-evaluation benchmark (Zheng et al., 2023). JudgeBench contributes 350 items spanning mathematics, coding, creative writing, and analysis labeled for objective correctness rather than aesthetic preference (Tan et al., 2025). RewardBench contributes 2,981 chosen-versus-rejected pairs, presented to each judge under per-item position randomization so that the chosen response is equally likely to occupy either position (Lambert et al., 2025).

Evaluation Protocols We measure judge behavior under three protocols. First, the *agreement protocol* produces a single judgment per item and compares it to the human label, reporting Cohen's κ, Krippendorff's α, and tie-excluded exact match. Second, the *consistency protocol* runs $N \in [3, 5]$ independent evaluations per item with response caching disabled, presenting each pair in both AB and BA orderings, and reports test–retest reliability, self-consistency, and position flip rate. Third, the *bias-audit protocol* presents AB+BA orderings together with response-length analysis, and reports position bias and verbosity bias.

Hypotheses We formulate seven hypotheses before running frontier-model evaluation and list them here so that readers can assess our findings against *a priori* predictions. *H1*: Kappa deflation will persist within frontier models. *H2*: Thinking-architecture models (GPT-5.4, Gemini 3.1 Pro, DeepSeek V3.2) will exhibit position bias below 0.05. *H3*: Cross-benchmark rank divergence will be at least three positions for some model. *H4:* MT-Bench will show a limited κ spread, defined as the distance between the largest and smallest Cohen's κ. *H5:* Issues with RewardBench, (human labels are all in the *A* position), will persist under generative evaluation. As a result, κ values will be very small, no larger than 0.05. *H6:* Position flip rate will degrade by at least a factor of 1.5 from MT-Bench to JudgeBench. *H7:* At least one model will exhibit test-retest reliability greater than 0.95 and with position bias greater than 0.10.

Experimental Procedure We conducted model evaluation in seven phases over a five-week window during March and April 2026. A detailed summary is in Appendix F. The full procedure produced 118 evaluation runs, and approximately $541,000$ judgments. All evaluations used temperature 0. In order to ensure that replicate runs sample independent generations rather than memorized responses, we execute the consistency protocol with response caching disabled. Position-swap debiasing (AB+BA paired evaluations) was applied in both the consistency and bias-audit protocols. Every run was executed with a bespoke research evaluation library (that will be released open source upon publication) which implements the 14 metrics for each model.

4 Results

4.1 Kappa Deflation Is Universal

Figure 1 (a) plots the dumbbell of exact match and κ values for each judge on MT-Bench and Table 2 reports κ, exact match, and Δ_κ for all three benchmarks. Every judge in our study exhibits substantial kappa deflation on MT-Bench: exact match overstates chance-corrected agreement by between 33.8 and 41.3 percentage points across the 21 models, with a cohort mean of 38.6 pp. The deflation is universal across capability tiers: all ten frontier-generation (Tier 3) judges exhibit $\Delta_\kappa \geq 30$ pp, supporting H1. Even the judge that performed best on chance-corrected agreement, Gemini 3.1 Pro, shows a 33.8 pp gap between exact match ($EM = 0.849$) and chance-corrected ($\kappa = 0.511$) performance.

The magnitude of deflation varies systematically with the benchmark's label distribution. MT-Bench, with a balanced A/B/Tie distribution, produces the

Model	MT-Bench			JudgeBench			RewardBench		
	EM	κ	Δ_κ	EM	κ	Δ_κ	EM	κ	Δ_κ
Gemini 3.1 Pro	0.849	0.511	33.8	0.964	0.841	12.3	0.956	0.898	5.9
Claude Opus 4.6	0.848	0.489	35.9	0.956	0.875	8.1	0.943	0.879	6.4
DeepSeek V3.2	0.845	0.486	35.9	0.791	0.545	24.5	0.921	0.826	9.5
Claude Sonnet 4.6	0.851	0.484	36.7	0.920	0.782	13.8	0.942	0.871	7.1
Llama 3.3 70B	0.841	0.465	37.6	0.664	0.283	38.1	0.892	0.769	12.3
Kimi K2.5	0.846	0.461	38.5	0.864	0.720	14.5	0.937	0.873	6.4
GPT-5.4	0.836	0.457	38.0	0.812	0.606	20.7	0.940	0.879	6.2
GPT-4o	0.832	0.451	38.1	0.667	0.309	35.8	0.883	0.745	13.8
GPT-4.1	0.830	0.451	37.9	0.751	0.487	26.4	0.906	0.809	9.7
Gemini 2.5 Pro	0.829	0.447	38.3	0.838	0.603	23.6	0.932	0.830	10.2
GLM-5	0.832	0.442	39.0	0.804	0.596	20.8	0.923	0.838	8.5
GPT-oss 120B	0.833	0.441	39.2	0.854	0.687	16.7	0.944	0.880	6.4
Claude Sonnet 4	0.843	0.440	40.2	0.817	0.633	18.4	0.944	0.886	5.8
Gemini 2.5 Flash	0.825	0.437	38.9	0.804	0.578	22.6	0.919	0.817	10.2
Claude Haiku 4.5	0.832	0.435	39.7	0.831	0.653	17.7	0.937	0.873	6.4
GPT-4.1-mini	0.831	0.432	39.9	0.738	0.466	27.1	0.899	0.795	10.4
Minimax M2.7	0.828	0.430	39.8	0.868	0.715	15.3	0.920	0.834	8.6
Qwen 3 8B	0.810	0.406	40.4	0.645	0.289	35.6	0.829	0.616	21.3
GPT-4o-mini	0.809	0.396	41.3	0.676	0.325	35.2	0.831	0.622	20.8
Mixtral 8x22B	0.803	0.392	41.0	0.656	0.271	38.5	0.852	0.679	17.3
GPT-5.4-mini	0.788	0.376	41.2	0.696	0.372	32.4	0.901	0.798	10.3
Cohort mean			38.6			23.7			10.2

Table 2: Exact match (EM), Cohen's κ, and kappa deflation $\Delta_\kappa = \text{EM} - \kappa$ (in percentage points) for each judge on each benchmark. Rows sorted by MT-Bench κ descending.

largest mean deflation (38.6 pp); JudgeBench's pairwise correctness-labeled items produce a mean of 23.7 pp (range 8.1 to 38.5); RewardBench's binary chosen-versus-rejected pairs produce 10.4 pp (range 5.9 to 21.3). Balanced label distributions raise expected-by-chance agreement, which inflates the gap between raw and chance-corrected metrics. The practical consequence is that a judge reporting "85% agreement" on MT-Bench has $\kappa \approx 0.48$.

4.2 Position Bias Heterogeneity

Table 3, column 2, reports position biases for each model. The least position biased model, Gemini 2.5 Pro ($pb = 0.002$) and most position biased model Qwen 3 8B ($pb = 0.192$) span nearly two orders of magnitude difference in position bias. In this evaluation, reasoning and frontier models have systematically lower rates of position bias; small, and cost-optimized models have the highest rates of position bias. Within-family heterogeneity is large: Gemini 2.5 Pro (0.002) and Gemini 2.5 Flash (0.125) differ by a factor of 70. H2 predicted that the three thinking-architecture judges (GPT-5.4, Gemini 3.1 Pro, DeepSeek V3.2) would all fall below 0.05; only Gemini 3.1 Pro (0.038) does so. And so, while they reduce position bias relative to cost-efficient models, they do not eliminate it.

4.3 Cross-Benchmark Rank Instability

Consistent with Hypothesis 3, benchmark choice has an appreciable effect on relative rankings of performance. More than half (11 of the 21) shift by four or more positions, and only Gemini 3.1 Pro and Claude Opus 4.6 hold a top-three position across all three benchmarks. The largest single shift

Model	Position bias $\lvert P(A) - 0.5\rvert$	Verbosity bias Pearson(len, verdict)
Gemini 2.5 Pro	0.002	0.0025
Claude Opus 4.6	0.004	0.0032
Kimi K2.5	0.004	0.0044
Claude Sonnet 4	0.008	0.0035
GPT-5.4-mini	0.013	0.0046
Claude Sonnet 4.6	0.015	0.0034
Minimax M2.7	0.020	0.0004
GPT-oss 120B	0.037	0.0024
Gemini 3.1 Pro	0.038	0.0007
Claude Haiku 4.5	0.041	0.0067
GPT-4o	0.045	0.0031
GPT-4o-mini	0.047	0.0103
GPT-4.1-mini	0.050	0.0045
GLM-5	0.052	0.0024
GPT-4.1	0.053	0.0026
Llama 3.3 70B	0.057	0.0011
Mixtral 8x22B	0.058	0.0084
GPT-5.4	0.083	0.0018
DeepSeek V3.2	0.094	0.0030
Gemini 2.5 Flash	0.125	0.0009
Qwen 3 8B	0.192	0.0011

Table 3: MT-Bench bias-audit results. Position bias is $\lvert P(A \text{ wins}) - 0.5\rvert$ over paired AB+BA evaluations; verbosity bias is the Pearson correlation between response-length differential and verdict. Rows sorted by position bias ascending.

is Llama 3.3 70B which shifts 15 positions, from 5 on MT-Bench to 20 on JudgeBench. Other notable cases include Minimax M2.7 (MT:17, JB:5), DeepSeek V3.2 (MT:3, JB:13), GPT-4o (MT:8, JB:18), GPT-oss 120B (MT:12, RB:3), and Claude Haiku 4.5 (MT:15, RB:6). (Supplementary figure Figure 2 visualizes Table 2 highlighting position changes.)

The instability is amplified by sharp differences in benchmark discriminability: MT-Bench compresses all 21 judges into a 13.5 pp κ band (0.376 to 0.511) with an average gap of 0.6 pp between adjacent ranks; JudgeBench spreads the same models over 60.4 pp (0.271 to 0.875) with 3–9 pp tiers; RewardBench falls between at 28.1 pp (0.616 to 0.898), with the top seven judges within a tight 2.7 pp band followed by gradual separation. Where the underlying distribution is compressed, small κ differences produce large rank changes, and a model's apparent ranking on a single benchmark is a poor estimator of its ranking on others.

4.4 MT-Bench Ceiling Effect

Hypothesis 4 predicted an MT-Bench κ spread of at most 5 pp. As we report in Table 2, we observe 13.5 pp (0.376 to 0.511) across all 21 judges, narrowing to 6.5 pp within the top ten and approaching the H4 ceiling. The wider full-cohort spread is driven by a distinct lower tier (GPT-5.4-mini, Mixtral 8x22B,

GPT-4o-mini). JudgeBench's κ spread on the same population is 60.4 pp, a factor of $4.5\times$ wider, confirming that MT-Bench's preference-style label set compresses meaningful quality differences among strong judges.

4.5 Position-Randomized Evaluation of RewardBench

Hypothesis 5 predicted that RewardBench would generate κ values no larger than 0.05 because it placed all correct human labels in the *A* position. As we report in Table 2 the benchmark data is not consistent with this hypothesis. Under per-item position-randomized evaluation (loader configuration in Appendix E), all 20 evaluated judges produce κ values in the range $[0.616, 0.898]$. The top five judges produce highly reasonable κ values: Gemini 3.1 Pro (0.898), GPT-oss 120B (0.880), Claude Opus 4.6 (0.879), GPT-5.4 (0.879), and Claude Haiku 4.5 (0.873). The full κ spread on RewardBench is 28.1 pp, between MT-Bench's 13.5 pp and JudgeBench's 60.4 pp, and the top tier is tightly clustered (the top seven judges fall within a 2.7 pp band). (In the Appendix, Table 5 summarises Table 2, making it easier to read ranks and κ jointly across the three benchmarks.)

4.6 Consistency Falls on Hard Benchmarks

Consistent with Hypothesis 6, seven of the sixteen judges evaluated under the consistency protocol show a position flip-rate increase of at least $1.5\times$ from MT-Bench to JudgeBench. The three most acute degradations are Llama 3.3 70B ($3.3\times$, 0.077 to 0.253), GPT-4o-mini ($3.0\times$, 0.127 to 0.380), and GLM-5 ($2.4\times$, 0.096 to 0.227). Two frontier judges go in the opposite direction: Claude Opus 4.6 and Gemini 3.1 Pro both improve by a factor of $0.58\times$ when moving from MT-Bench to JudgeBench, suggesting that some current-generation systems handle harder items with greater positional stability than they show on easier items. (In the appendix, Figure 3 plots all 16 trajectories together with the cohort medians (0.09 on MT-Bench, 0.17 on JudgeBench)).

Test-retest reliability follows a similar pattern at the cohort level: the mean test–retest across the 16 judges drops from 0.943 on MT-Bench to 0.911 on JudgeBench. Per-judge values vary widely; for example, Gemini 2.5 Flash drops by 7.3 pp (0.988 to 0.915), while Gemini 3.1 Pro is essentially unchanged (0.977 to 0.978). (In the appendix, we report per-judge MT and JB test–retest values in Table 6.)

4.7 The Consistency–Bias Paradox

Hypothesis 7 predicted that at least one judge would exhibit test-retest reliability above 0.95 together with position bias above 0.10, evidence of the tradeoffs between *consistency-bias*. Two judges, Qwen 3 8B (test-retest 0.992, position bias 0.192) and Gemini 2.5 Flash (test-retest 0.988, position bias 0.125) generate such paradoxical results. Figure 1 (b) plots all 16 jointly evaluated judges in the test-retest by position-bias plane. Judges that occupy the upper-right region are those that demonstrate high test-retest reproducibility, but with high systematic bias.

Qwen 3 8B has the highest test-retest of judges evaluated in this study, but it simultaneously is among the most position biased judge in the cohort ($pb = 0.192$) and its JudgeBench κ (0.289) is the third lowest. Model determinism produces the same response position across replicate runs, which yields near-perfect within-judge agreement, but it violates the requirement that verdicts be position-invariant. This result has methodological implications for future teams: test–retest measures the stability of a judge's outputs, not the correctness of the underlying decision process. Reporting test-retest alone stands to present a misleading picture of judge reliability.

4.8 Pairwise Verbosity Bias

All 21 judges evaluated under the bias-audit protocol register a verbosity bias below 0.011 on MT-Bench, with the largest value being GPT-4o-mini at 0.010; 17 of the 21 judges fall below 0.005, with the smallest values clustered around 0.001. Per-judge values appear in Table 3. The magnitudes are an order of magnitude smaller than the variance contributions reported in 2023-era studies of style and length effects (Wu and Aji, 2025), suggesting that verbosity sensitivity has fallen substantially as a practical concern over the past two model generations. We caution against generalizing this finding beyond our setting: the measurement uses a single pairwise rubric and a fixed length-differential operationalization (Dubois et al., 2024), and we do not claim that verbosity bias has been eliminated under arbitrary rubric or task variation.

4.9 Model Family Analysis

Several provider-family patterns are visible in the per-judge κ values reported in Table 2. The

three Anthropic judges average $\kappa = 0.770$ on JudgeBench (Opus 4.6 0.875, Sonnet 4.6 0.782, Haiku 4.5 0.653) with an average position bias of 0.020, the strongest joint performance on hard items and the lowest cohort-level bias of any provider. The three OpenAI flagships (GPT-4o, GPT-4.1, GPT-5.4) average $\kappa = 0.467$ on JudgeBench, well below the Anthropic and Google frontier-tier figures. Within-OpenAI developed models, generational progression is visible: κ rises from 0.309 for GPT-4o to 0.487 for GPT-4.1, and rises to 0.606 for GPT-5.4. MT-Bench's compressed scale makes the same generational progression nearly invisible: the corresponding MT-Bench values are 0.451, 0.451, and 0.457.

Similarly, within-family scaling is more legible on JudgeBench than on MT-Bench. Claude Opus 4.6 ($\kappa = 0.875$) is 9.3 pp above Claude Sonnet 4.6 ($\kappa = 0.782$) on JudgeBench; the corresponding MT-Bench values differ by 0.6 pp (Opus 0.489, Sonnet 4.6 0.484). Mid-tier judges can outperform frontier judges on specific dimensions: Kimi K2.5 records the lowest position bias of any non-Gemini judge (0.004), competitive with Claude Opus 4.6, while achieving JudgeBench $\kappa = 0.720$ at a fraction of the frontier-tier cost. We return to cost-quality trade-offs in section 5.

5 Discussion

5.1 Why Exact Match Misleads

Exact match rewards chance agreement as if it were genuine discrimination. On a balanced ternary label set such as MT-Bench, with the chance baseline $p_e \approx 1/3$, exact-match values in $[0.80, 0.85]$ correspond to a Cohen's κ near 0.48. This is a moderate value, not the near-perfect band the percentage suggests. The 33.8–41.3 pp gap we observe on MT-Bench is therefore not a model-quality artifact: it is how an uncorrected metric interacts with the label marginals of the benchmark (Cohen, 1960; Guerdan et al., 2025). The cohort-mean Δ_κ contracts from 38.6 pp on MT-Bench to 23.7 pp on JudgeBench and 10.2 pp on RewardBench, tracking the shift from balanced ternary to imbalanced binary labels exactly as Cohen's correction predicts.

Exact-match figures used to justify deploying LLM judges therefore may significantly overstate their chance-corrected discriminative ability by an amount that depends on the benchmark, not the judge (Han et al., 2025; Collot et al., 2025). We recommend reporting Cohen's κ or Krippendorff's α alongside any exact-match figure, with the chance-corrected metric (not the raw rate), treated as the headline reliability number. Figure 1 (a) and Table 2 illustrate the gap for every judge in our cohort.

5.2 Limits of Single-Benchmark Validation

Cross-benchmark rank instability has two coupled drivers. First, benchmarks differ in discriminative power: MT-Bench compresses the 21-judge cohort into a 13.5 pp κ band while JudgeBench spreads the same judges across 60.4 pp, a 4.5× wider range. Small absolute differences yield large rank changes. Second, the three benchmarks measure subtly different latent constructs (preference alignment, objective correctness, and chosen-vs-rejected discrimination, respectively), and a judge that looks strong under one construct can collapse under another. Llama 3.3 70B (MT#5 → JB#19) and Minimax M2.7 (MT#16 → JB#5) illustrate the two directions.

Only two judges (Claude Opus 4.6 and Gemini 3.1 Pro) hold a top-three position across all three benchmarks; ten of twenty-one exhibit a maximum pairwise rank shift of at least four positions. Practically, judge validation should report results from at least two benchmarks chosen to span the preference-versus-correctness axis rather than relying on the discriminability of any single dataset. Appendix Figure 2 and Appendix Table 5 report the full set of rank trajectories.

5.3 The Consistency–Bias Paradox and a Minimum Viable Validation Protocol

Test–retest reliability measures *output stability*, not *decision-process correctness*. Position bias and within-judge agreement are mathematically orthogonal, so a judge that deterministically prefers position A across runs achieves near-perfect test–retest while exhibiting maximal position bias. Qwen 3 8B (test–retest 0.992, position bias 0.192, JudgeBench $\kappa = 0.289$) is the empirical extreme. Gemini 2.5 Flash (0.988, 0.125) instantiates the same pattern more mildly. Because reporting test–retest alone remains common in LLMaJ validation (Stureborg et al., 2024; Schroeder and Wood-Doughty, 2024), current reporting practice misleads precisely in the cases that matter most for deployment: highly reproducible judges. Figure 1 (b) shows the dissociation across the 16 jointly evaluated judges. This paradox motivates our recommended Minimum Viable Validation Protocol (MVVP) below.

Minimum Viable Validation Protocol (MVVP). Before deploying an LLM judge:

1. **Chance-correct.** Report Cohen's κ (or Krippendorff's α) alongside any exact-match figure, and treat the chance-corrected metric as the headline reliability number.
2. **Swap positions.** Measure position bias via paired AB+BA evaluations and report $|P(\text{A wins}) - 0.5|$.
3. **Replicate.** Measure test–retest reliability over ≥ 3 independent runs at temperature 0 with response caching disabled.
4. **Cross-validate.** Evaluate on ≥ 2 benchmarks spanning preference-style and correctness-style label distributions.
5. **Audit the paradox.** When test–retest exceeds 0.95, verify position bias is below 0.10 before claiming reliability. High stability with high bias is a failure mode, not a strength.

5.4 Dataset as Community Resource

We will release the complete evaluation dataset and the [anonymized] code repository upon publication. The dataset is derived from 118 evaluation runs spanning 21 judges, three benchmarks, and three protocols, totaling approximately 541,000 individual judgments. Every run includes the per-item judge verdict, the model's free-form reasoning, the raw response, per-call latency, and the metric values computed for the run, so that downstream users can recompute any of our reported numbers from raw data or substitute their own metric and aggregation choices. The release supports three use cases we cannot pursue here: large-scale meta-analysis across judges and benchmarks (including variance-component decompositions across runs, items, and position orderings); development and validation of new calibration techniques on a fixed model$\times$benchmark population (Li et al., 2025c); and a frozen baseline against which future judges can be evaluated under an identical protocol.

6 Conclusion

We present the largest systematic evaluation of LLMaJ to date: 21 judges from nine providers across three benchmarks and three measurement protocols, producing approximately 541,000 individual judgments. Four findings are present across the full cohort, including the frontier models released through April 2026: kappa deflation between exact match and Cohen's κ is universal (33.8–41.3 pp on MT-Bench); judge rankings shift by up to 14 positions across benchmarks; high test–retest reliability (> 0.95) coexists with severe position bias (> 0.10) in two production-deployed judges; and verbosity bias is small (< 0.011) across our cohort under a single pairwise rubric.

These findings translate into four practical recommendations. Judge validation should: (1) be chance-corrected by default, because raw exact match overstates discriminative ability by tens of percentage points, (2) span at least two benchmarks of contrasting label structure, since no single leaderboard predicts another's, (3) measure consistency and bias jointly, because high test–retest can mask a determinism that violates position invariance, and (4) no longer foreground verbosity bias under standard pairwise rubrics, subject to scope caveats. We introduce these as a Minimum Viable Validation Protocol (Section 5) and will release the full dataset and the [anonymized] evaluation library upon publication so that future judges can be measured against the same baseline.

Several research directions remain open. Multilingual and multimodal judging are likely to surface failure modes that text-only English benchmarks cannot expose. Temporal stability across hosted-model endpoints, whose weights and serving stacks change without versioning, deserves dedicated longitudinal study. Judge-confidence calibration (Expected Calibration Error, Brier score) is a natural extension to this protocol once provider log-prob coverage broadens. Finally, a formal variance-component decomposition that treats runs, items, and position orderings as random effects would convert the empirical patterns reported here into structured reliability coefficients, providing a principled basis for sample-size and replicate-count decisions in future judge-validation work.

Limitations

Benchmark coverage. Our evaluation uses three established English-language, text-only benchmarks (MT-Bench, JudgeBench, RewardBench). Multilingual judging and multimodal judging—both of which are increasingly common deployment settings—are not characterized in this study (Fu and Liu, 2025). We caution against extrapolating any of our findings to those settings without further measurement.

Temporal stability. All evaluation runs were executed within a five-week window in March–April 2026. Hosted-model endpoints are known to drift across provider-side updates, sometimes silently, and we have not yet re-evaluated these judges over time to quantify within-judge stability over a longer time horizon. Consequently, the numbers we report should be interpreted as a snapshot of provider behavior during this specific window.

Rubric sensitivity. All judges were evaluated under a single pairwise comparison template and a fixed operationalization of each metric. In particular, our findings that all judges demonstrated verbosity bias below 0.011, under our pairwise rubric and length-differential operationalization (Wu and Aji, 2025; Dubois et al., 2024), should not be interpreted as a universal claim that verbosity bias is solved. Rubric design choices, including reference grounding and reasoning instructions, are known to interact with bias profiles in ways our single-template design cannot isolate.

Calibration. Expected Calibration Error and Brier Score require token-level logprobs that are not exposed by the majority of providers in our cohort. We therefore defer calibration analysis to future work; the present study reports agreement, consistency, and bias under the protocols for which directly comparable measurements are available across all providers.

Cost reporting. Per-token costs reported in Table 1 are list prices at the time of evaluation. Provider-specific volume tiers, batch-API discounts, and negotiated enterprise pricing are not reflected. Accordingly, these cost figures should be interpreted as approximate upper bounds for deployment-scale spend rather than committed rates.

Thinking-model suppression. For models that generate a built-in reasoning trace (GPT-5.4, GPT-5.4-mini, Gemini 3.1 Pro, DeepSeek V3.2, Kimi K2.5, GLM-5, Minimax M2.7), the reasoning channel was suppressed using the provider-specific mechanisms documented in Appendix D; this maintained judge comparability and prevented verdict-token truncation when reasoning consumed the output budget. Reasoning-enabled evaluations could change agreement, consistency, and bias profiles, and we do not claim our results characterize the thinking-on configuration.

References

Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi, Desmond Elliott, Raquel Fernández, Albert Gatt, Esam Ghaleb, Mario Giulianelli, Michael Hanna, Alexander Koller, André F. T. Martins, Philipp Mondorf, Vera Neplenbroek, Sandro Pezzelle, Barbara Plank, David Schlangen, Alessandro Suglia, Aditya K. Surikuchi, Ece Takmaz, and Alberto Testoni. 2025. LLMs instead of human judges? a large scale empirical study across 20 NLP evaluation tasks. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 238–255, Vienna, Austria. Association for Computational Linguistics.

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

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. 2026. Diagnosing the reliability of LLM-as-a-judge via item response theory. In *Proceedings of the 43rd International Conference on Machine Learning (ICML)*, volume 306 of *Proceedings of Machine Learning Research*. PMLR.

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

Stephane Collot, Colin Fraser, Justin Zhao, William F. Shen, Timon Willi, and Ilias Leontiadis. 2025. Balanced accuracy: The right metric for evaluating LLM judges — explained through Youden's J statistic. *arXiv preprint arXiv:2512.08121*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. 2024. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *First Conference on Language Modeling (COLM)*.

Xiyan Fu and Wei Liu. 2025. How reliable is multilingual LLM-as-a-judge? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 11040–11053, Suzhou, China. Association for Computational Linguistics.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Yuanzhuo Wang, Wen Gao, Lionel Ni, and Jian Guo. 2024. A survey on LLM-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna M. Wallach, Steven Z. Wu, and Alexandra Chouldechova. 2025. Validating LLM-as-a-judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, volume 38.

Rajarshi Haldar and Julia Hockenmaier. 2025. Rating roulette: Self-inconsistency in LLM-as-a-judge frameworks. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 24986–25004, Suzhou, China. Association for Computational Linguistics.

Steve Han, Gilberto Titericz Junior, Tom Balough, and Wenfei Zhou. 2025. Judge's verdict: A comprehensive analysis of LLM judge capability through human agreement. *arXiv preprint arXiv:2510.09738*.

Andrew F. Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Hongchao Jiang, Yiming Chen, Yushi Cao, Hung-yi Lee, and Robby T. Tan. 2025. CodeJudgeBench: Benchmarking LLM-as-a-judge for coding tasks. *arXiv preprint arXiv:2507.10535*.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. 2024a. Prometheus: Inducing fine-grained evaluation capability in language models. In *International Conference on Learning Representations (ICLR)*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024b. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha-reliability. *Departmental Papers (ASC)*.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. 2025. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1755–1797, Albuquerque, New Mexico. Association for Computational Linguistics.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025a. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024a. LLMs-as-judges: A comprehensive survey on LLM-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2025b. From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*, Proceedings of Machine Learning Research. PMLR / OpenReview.net.

Zihao Li, Feihao Fang, Xitong Zhang, Jiaru Zou, Zhining Liu, Wei Xiong, Ziwei Wu, Baoyu Jing, and Jingrui He. 2025c. Not all voices are rewarded equally: Probing and repairing reward models across human diversity. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 3426–3455, Suzhou, China. Association for Computational Linguistics.

Zongjie Li, Chaozheng Wang, Pingchuan Ma, Daoyuan Wu, Shuai Wang, Cuiyun Gao, and Yang Liu. 2024b. Split and merge: Aligning position biases in LLM-based evaluators. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 11084–11108, Miami, Florida, USA. Association for Computational Linguistics.

Bill Yuchen Lin, Yuntian Deng, Khyathi Raghavi Chandu, Abhilasha Ravichander, Valentina Pyatkin, Nouha Dziri, Ronan Le Bras, and Yejin Choi. 2025. Wildbench: Benchmarking llms with challenging tasks from real users in the wild. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net.

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

Yulei Liu. 2025. How DoorDash leverages LLMs to evaluate search result pages. DoorDash Engineering Blog. Accessed 2026-05-07.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2026. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations (ICLR)*.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. 2024. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems*, volume 37, pages 68772–68802.

Kayla Schroeder and Zach Wood-Doughty. 2024. Can you trust LLM judgments? reliability of LLM-as-a-judge. *arXiv preprint arXiv:2412.12509*.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*, pages 292–314, Mumbai, India. Association for Computational Linguistics.

Alex Strick van Linschoten. 2025. What 1,200 production deployments reveal about LLMOps in 2025. ZenML Blog. Accessed 2026-05-07.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. 2024. Large language models are inconsistent and biased evaluators. *arXiv preprint arXiv:2405.01724*.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2025. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations (ICLR)*.

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

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. 2024. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9440–9450, Bangkok, Thailand. Association for Computational Linguistics.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. 2025. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 23173–23199, Suzhou, China. Association for Computational Linguistics.

Chenxi Whitehouse, Tianlu Wang, Ping Yu, Xian Li, Jason Weston, Ilia Kulikov, and Swarnadeep Saha. 2025. J1: Incentivizing thinking in LLM-as-a-judge via reinforcement learning. *arXiv preprint arXiv:2505.10320*.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. JudgeLM: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations (ICLR)*.

Appendix

A Supplementary Figures

A.1 Cross-Benchmark Rank Instability

Figure 2 expands on the cross-benchmark rank instability summarized in Section 4, plotting each judge's rank trajectory across MT-Bench, JudgeBench, and RewardBench. The visualization makes individual rank shifts directly readable: a horizontal line indicates a judge that holds its position across benchmarks, and a steep slope identifies a judge whose rank moves substantially as the benchmark changes.

A.2 Consistency Degrades on Hard Benchmarks

Figure 3 visualizes the per-judge change in position flip rate from MT-Bench to JudgeBench across the 16-judge consistency cohort. Highlighting identifies the three judges whose flip rate worsens by at least $2.4\times$ on the harder benchmark and the two frontier judges (Claude Opus 4.6 and Gemini 3.1 Pro) that improve; the cohort medians (dotted) summarize the cohort-level shift.

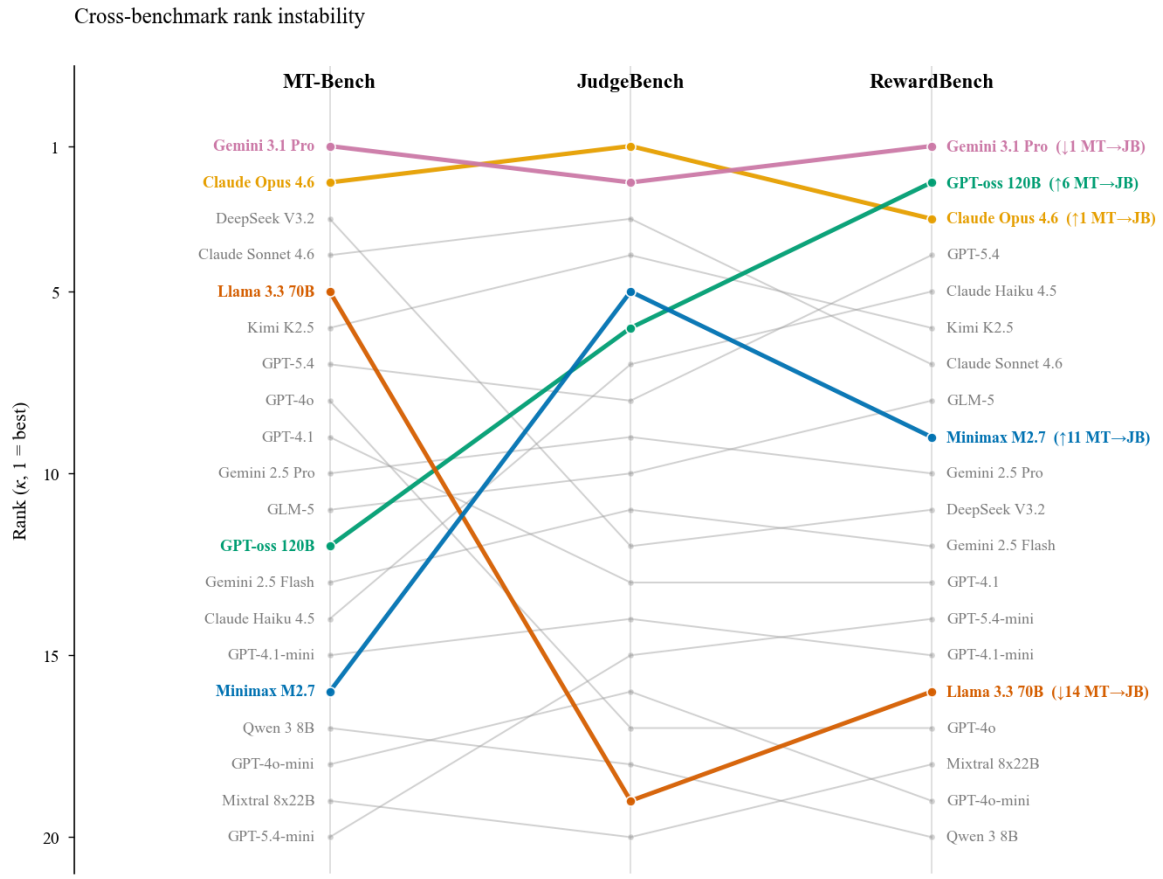


Figure 2: **Cross-benchmark rank instability.** One line per model across MT-Bench, JudgeBench, and RewardBench. Within each benchmark, judges are ranked by Cohen's κ descending (rank 1 is the highest κ); ranks are computed independently per benchmark and ties are broken by full-precision κ. **Llama 3.3 70B** drops 14 positions (MT#5 $\rightarrow$ JB#19); **Minimax M2.7** rises 11 (MT#16 $\rightarrow$ JB#5); **GPT-oss 120B** climbs to RewardBench's top tier (RB#2). Only Claude Opus 4.6 and Gemini 3.1 Pro hold a stable top-3 position across all three benchmarks.

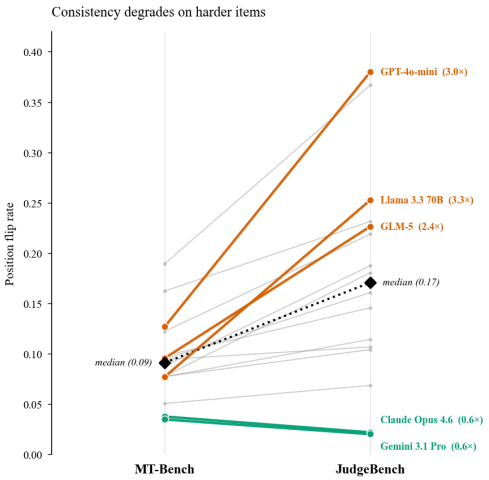


Figure 3: **Position flip rate degrades from MT-Bench to JudgeBench for most models.** Three judges show $\geq 2.4\times$ degradation (highlighted orange). Two frontier judges, Claude Opus 4.6 and Gemini 3.1 Pro, improve on the harder benchmark ($0.6\times$, highlighted green). Cohort median rises from 0.09 to 0.17 (dotted black).

B Complete Results Tables

This section reports per-judge values that the body figures and tables summarize: full agreement statistics on all three benchmarks (Table 4), per-judge cross-benchmark ranks with Cohen's κ (Table 5), and per-judge consistency-protocol detail for the 17-judge consistency cohort (Table 6). Numerical values are extracted from raw `[anonymized]/results/*.json` outputs and matched to the display names used elsewhere in the paper.

	MT-Bench				JudgeBench				RewardBench			
Model	EM	κ	α	Δ_κ	EM	κ	α	Δ_κ	EM	κ	α	Δ_κ
Gemini 3.1 Pro	0.849	0.511	0.506	33.8	0.964	0.841	0.841	12.3	0.956	0.898	0.898	5.9
Claude Opus 4.6	0.848	0.489	0.480	35.9	0.956	0.875	0.876	8.1	0.943	0.879	0.879	6.4
DeepSeek V3.2	0.845	0.486	0.477	35.9	0.791	0.545	0.545	24.5	0.921	0.826	0.826	9.5
Claude Sonnet 4.6	0.851	0.484	0.473	36.7	0.920	0.782	0.782	13.8	0.942	0.871	0.871	7.1
Llama 3.3 70B	0.841	0.465	0.453	37.6	0.664	0.283	0.281	38.1	0.892	0.769	0.769	12.3
Kimi K2.5	0.846	0.461	0.446	38.5	0.864	0.720	0.720	14.5	0.937	0.873	0.873	6.4
GPT-5.4	0.836	0.457	0.443	38.0	0.812	0.606	0.606	20.7	0.940	0.879	0.879	6.2
GPT-4o	0.832	0.451	0.439	38.1	0.667	0.309	0.307	35.8	0.883	0.745	0.745	13.8
GPT-4.1	0.830	0.451	0.437	37.9	0.751	0.487	0.488	26.4	0.906	0.809	0.809	9.7
Gemini 2.5 Pro	0.829	0.447	0.436	38.3	0.838	0.603	0.602	23.6	0.932	0.830	0.830	10.2
GLM-5	0.832	0.442	0.427	39.0	0.804	0.596	0.597	20.8	0.923	0.838	0.838	8.5
GPT-oss 120B	0.833	0.441	0.426	39.2	0.854	0.687	0.687	16.7	0.944	0.880	0.880	6.4
Claude Sonnet 4	0.843	0.440	0.423	40.2	0.817	0.633	0.633	18.4	0.944	0.886	0.886	5.8
Gemini 2.5 Flash	0.825	0.437	0.421	38.9	0.804	0.578	0.578	22.6	0.919	0.817	0.817	10.2
Claude Haiku 4.5	0.832	0.435	0.418	39.7	0.831	0.653	0.653	17.7	0.937	0.873	0.873	6.4
GPT-4.1-mini	0.831	0.432	0.414	39.9	0.738	0.466	0.466	27.1	0.899	0.795	0.795	10.4
Minimax M2.7	0.828	0.430	0.416	39.8	0.868	0.715	0.715	15.3	0.920	0.834	0.834	8.6
Qwen 3 8B	0.810	0.406	0.387	40.4	0.645	0.289	0.257	35.6	0.829	0.616	0.616	21.3
GPT-4o-mini	0.809	0.396	0.376	41.3	0.676	0.325	0.319	35.2	0.831	0.622	0.622	20.8
Mixtral 8x22B	0.803	0.392	0.373	41.0	0.656	0.271	0.270	38.5	0.852	0.679	0.679	17.3
GPT-5.4-mini	0.788	0.376	0.358	41.2	0.696	0.372	0.372	32.4	0.901	0.798	0.798	10.3
Cohort mean (21)				38.6				24.0				10.4

Table 4: **Per-judge agreement statistics on all three benchmarks.** EM is tie-excluded exact match; κ is Cohen's kappa; α is Krippendorff's alpha; $\Delta_\kappa = \text{EM} - \kappa$ in percentage points. Rows sorted by MT-Bench κ descending.

	MT-Bench		JudgeBench		RewardBench	
Model	rank	κ	rank	κ	rank	κ
Gemini 3.1 Pro	1	0.511	2	0.841	1	0.898
Claude Opus 4.6	2	0.489	1	0.875	4	0.879
DeepSeek V3.2	3	0.486	13	0.545	12	0.826
Claude Sonnet 4.6	4	0.484	3	0.782	8	0.871
Llama 3.3 70B	5	0.465	20	0.283	17	0.769
Kimi K2.5	6	0.461	4	0.720	7	0.873
GPT-5.4	7	0.457	9	0.606	5	0.879
GPT-4o	8	0.451	18	0.309	18	0.745
GPT-4.1	9	0.451	14	0.487	14	0.809
Gemini 2.5 Pro	10	0.447	10	0.603	11	0.830
GLM-5	11	0.442	11	0.596	9	0.838
GPT-oss 120B	12	0.441	6	0.687	3	0.880
Claude Sonnet 4	13	0.440	8	0.633	2	0.886
Gemini 2.5 Flash	14	0.437	12	0.578	13	0.817
Claude Haiku 4.5	15	0.435	7	0.653	6	0.873
GPT-4.1-mini	16	0.432	15	0.466	16	0.795
Minimax M2.7	17	0.430	5	0.715	10	0.834
Qwen 3 8B	18	0.406	19	0.289	21	0.616
GPT-4o-mini	19	0.396	17	0.325	20	0.622
Mixtral 8x22B	20	0.392	21	0.271	19	0.679
GPT-5.4-mini	21	0.376	16	0.372	15	0.798

Table 5: **Cross-benchmark ranks and Cohen's κ for the 21 judges evaluated on all three benchmarks.** Within each benchmark, judges are ranked by Cohen's κ descending (rank 1 is the highest κ), computed independently per benchmark; ties broken by full-precision κ. Rows sorted by MT-Bench rank.

	MT-Bench			JudgeBench		
Model	test–retest	self-consistency	flip rate	test–retest	self-consistency	flip rate
Qwen 3 8B	0.992	0.995	0.189	0.979	0.987	0.367
Gemini 2.5 Flash	0.988	0.994	0.122	0.915	0.949	0.219
Gemini 3.1 Pro	0.977	0.989	0.035	0.978	0.989	0.020
Claude Sonnet 4	0.960	0.976	0.074	0.941	0.966	0.146
GPT-4o-mini	0.959	0.975	0.127	0.896	0.939	0.380
Claude Opus 4.6	0.958	0.979	0.038	0.974	0.986	0.022
GPT-4.1	0.955	0.973	0.078	0.891	0.935	0.181
Llama 3.3 70B	0.954	0.972	0.077	0.892	0.936	0.253
GPT-oss 120B	0.947	0.973	0.052	0.920	0.959	0.115
Claude Sonnet 4.6	0.946	0.972	0.041	0.929	0.963	0.085
Claude Haiku 4.5	0.935	0.961	0.123	0.906	0.944	0.164
GLM-5	0.934	0.966	0.096	0.840	0.919	0.227
GPT-5.4	0.932	0.965	0.114	0.933	0.966	0.105
Kimi K2.5	0.917	0.957	0.095	0.901	0.951	0.121
DeepSeek V3.2	0.916	0.957	0.088	0.878	0.938	0.188
GPT-5.4-mini	0.889	0.943	0.137	0.868	0.933	0.196
Minimax M2.7	0.888	0.943	0.181	0.878	0.938	0.156
Cohort mean	0.944	0.969	0.099	0.913	0.950	0.173

Table 6: **Per-judge consistency-protocol detail.** Test–retest reliability is Krippendorff's α across replicate runs; self-consistency is the proportion of items on which the majority verdict matches each individual run; position flip rate is the fraction of items where the AB and BA orderings produce different verdicts. The 17 judges shown are the consistency cohort (Phase 4 Tier-1, Phase 5 Tier-2, and Phase 7 Tier-3 SoTA models). Rows sorted by MT-Bench test–retest descending.

C A Priori Hypothesis Details

C.1 Formulation context

Seven hypotheses (H1–H7) were formulated on 2026-04-14, before Phase 6 (frontier-model) data collection began, based on patterns observed in the earlier phases (Phases 1–5). The predictions were recorded internally and were not deposited with an external registry; we do not claim the methodological status of formal pre-registration. We report the unmodified outcome of each hypothesis test below, including the one we did not confirm. The intent of recording the predictions in advance is to make the calibration of our expectations available to readers, not to claim a formal confirmatory frame.

C.2 Per-hypothesis details

Table 7 reports the formal statement, the threshold, the verified observed value, the outcome, and the body subsection in which the test is reported.

C.3 Calibration assessment

Of the seven a priori predictions, three are fully confirmed (H1, H3, H7), one is confirmed for the explicitly-stated subset (H6), two are partially confirmed (H2, H4), and one is refuted (H5). The mixture matters as a calibration signal: hypotheses formulated post hoc would typically be confirmed at a higher rate. The presence of one refutation (H5: RewardBench was predicted to be degenerate, but position-randomized evaluation shows valid $\kappa \in [0.616, 0.898]$) and two partial outcomes (H2: thinking architecture does not reliably reduce position bias; H4: ceiling holds for the top tier but not the full cohort) indicates that the prediction set was non-trivially hard rather than a curated yes-set. The body §3.4 disclosure reports this in line with the user-facing methodological framing.

H	Statement	Threshold	Observed	Outcome	Body §
H1	Kappa deflation persists for frontier judges on MT-Bench.	$\Delta_\kappa \geq 30$ pp for all 10 SoTA judges.	Range $[33.8, 41.2]$ pp; all 10 frontier judges ≥ 33.8 pp.	Confirmed	§4.1
H2	Thinking-architecture judges (GPT-5.4, Gemini 3.1 Pro, DeepSeek V3.2) show low position bias.	Position bias < 0.05 for all three.	Gemini 3.1 Pro 0.038 (yes); GPT-5.4 0.083 (no); DeepSeek V3.2 0.094 (no).	Partially confirmed (1 of 3)	§4.2
H3	Cross-benchmark rank divergence.	Some model exhibits $\geq$ 3-position rank shift across benchmarks.	11 of 21 judges with maximum pairwise rank shift ≥ 4; max shift 15 (Llama 3.3 70B).	Confirmed	§4.3
H4	MT-Bench ceiling effect.	κ spread across all judges ≤ 5 pp.	Full spread 13.5 pp; top-10 spread 6.5 pp.	Partially confirmed (top-10 close to threshold)	§4.4
H5	RewardBench remains degenerate under generative evaluation.	$\kappa \leq 0.05$ across all judges.	$\kappa \in [0.616, 0.898]$ across 21 judges.	Refuted	§4.5
H6	Position flip rate degrades from MT-Bench to JudgeBench.	$\geq 1.5\times$ flip-rate increase for the majority of judges.	7 of 16 consistency-cohort judges show $\geq 1.5\times$.	Confirmed for 7 of 16	§4.6
H7	Consistency–bias paradox is instantiated.	≥ 1 judge with test–retest > 0.95 AND position bias > 0.10.	2 judges (Qwen 3 8B: $0.992, 0.192$; Gemini 2.5 Flash: $0.988, 0.125$).	Confirmed	§4.7

Table 7: **A priori hypothesis details and outcomes.** Statements reflect the predictions made on 2026-04-14, prior to Phase 6 data collection. Observed values are sourced from raw evaluation outputs; see body §4 for the underlying figures and tables.

D Model Configuration

This section reports the per-judge API configuration used for every evaluation run (Table 8) and the per-benchmark configuration parameters applied across protocols (Table 9).

Model	Provider	API model ID	Tier	Released	In $/MTok	Out $/MTok	Reasoning suppression
Mixtral 8x22B	Mistral	`open-mixtral-8x22b`	2	2024-04-17	1.20	1.20	none
GPT-4o	OpenAI	`gpt-4o`	1	2024-05-13	2.50	10.00	none
GPT-4o-mini	OpenAI	`gpt-4o-mini`	1	2024-07-18	0.15	0.60	none
Llama 3.3 70B	Meta †	`llama-v3p3-70b-instruct`	1	2024-12-06	0.90	0.90	none
Gemini 2.5 Pro	Google	`gemini-2.5-pro`	1	2025-03-25	1.25	10.00	`thinking_budget=128`; thinking stripped
GPT-4.1	OpenAI	`gpt-4.1`	1	2025-04-14	2.00	8.00	none
GPT-4.1-mini	OpenAI	`gpt-4.1-mini`	2	2025-04-14	0.40	1.60	none
Gemini 2.5 Flash	Google	`gemini-2.5-flash`	1	2025-04-17	0.30	2.50	`thinking_budget=0`
Qwen 3 8B	Alibaba †	`qwen3-8b`	1	2025-04-29	0.20	0.20	`/no_think` suffix
Claude Sonnet 4	Anthropic	`claude-sonnet-4-20250514`	2	2025-05-22	3.00	15.00	opt-in only
GPT-oss 120B	OpenAI †	`gpt-oss-120b`	3	2025-08-05	0.15	0.60	none
Claude Haiku 4.5	Anthropic	`claude-haiku-4-5-20251001`	1	2025-10-15	1.00	5.00	opt-in only
DeepSeek V3.2	DeepSeek †	`deepseek-v3p2`	3	2025-12-01	0.56	1.68	`reasoning_effort=none`
Kimi K2.5	Moonshot †	`kimi-k2p5`	3	2026-01-27	0.60	3.00	`thinking={type:disabled}`
Claude Opus 4.6	Anthropic	`claude-opus-4-6`	3	2026-02-05	5.00	25.00	opt-in only
GLM-5	Zhipu †	`glm-5`	3	2026-02-11	1.00	3.20	`thinking={type:disabled}`
Claude Sonnet 4.6	Anthropic	`claude-sonnet-4-6`	3	2026-02-17	3.00	15.00	opt-in only
Gemini 3.1 Pro	Google	`gemini-3.1-pro-preview`	3	2026-02-19	2.00	12.00	`thinking_budget=128`
GPT-5.4	OpenAI	`gpt-5.4`	3	2026-03-05	2.50	15.00	`reasoning_effort=none`
GPT-5.4-mini	OpenAI	`gpt-5.4-mini`	3	2026-03-17	0.75	4.50	`reasoning_effort=none`
Minimax M2.7	MiniMax †	`minimax-m2p7`	3	2026-03-18	1.00	3.00	none (verbose output)

Table 8: **Per-judge API configuration.** Rows sorted by release date ascending. The API model ID is the model slug passed to the evaluation framework; the full `[anonymized]_string` is the provider prefix concatenated with the slug (e.g. `openai:gpt-4o`). †Open-source models accessed via the Fireworks inference endpoint with full path `fireworks:accounts/fireworks/models/<slug>`. Costs are USD per million tokens at provider list pricing; the reasoning-suppression column documents the configuration used to disable or constrain test-time chain-of-thought where supported, ensuring all judges run under comparable no-reasoning conditions. Temperature is 0 for all judges.

Benchmark	Items	Format	Randomization	Consistency runs
MT-Bench	2,391	Pairwise (A/B/Tie)	none required	3–5
JudgeBench	350	Pairwise (correctness)	none required	3–5
RewardBench†	2,981	Chosen / rejected	per-item, seed 42	not in cohort

Table 9: **Per-benchmark evaluation configuration.** Item counts are the verified `n_samples` values from agreement-protocol JSON outputs. Consistency runs indicate the number of independent replicate evaluations per item used for the test–retest and self-consistency metrics; only MT-Bench and JudgeBench are in the consistency cohort, with Phase 4 (Tier 1) using 5 runs per item and Phase 7 (Tier 3 SoTA) using 3 runs per item. RewardBench is included in the agreement and bias-audit protocols only.

E RewardBench Loader Configuration

The RewardBench evaluation in this study uses per-item position randomization with seed 42, matching the contract of the official `run_generative.py` script that ships with the benchmark. This configuration is necessary because the standard generative-evaluation loader places every chosen response in position A and every rejected response in position B; under that fixed ordering the human label is identically "A" across all items, p_e collapses, and Cohen's κ degenerates to 0.000 for every judge. Per-item randomization with a fixed seed restores a balanced ordering distribution and produces a valid agreement signal.

The validity of the corrected loader is confirmed by the resulting κ distribution: across the 21 judges evaluated on RewardBench, κ values fall in the range $[0.616, 0.898]$, well within the substantial-to-near-perfect agreement bands. The full per-judge κ values appear in Table 4 of this appendix.

F Evaluation Phase Details

The evaluation campaign was conducted in seven phases over a five-week window in March and April 2026. Phases 1–5 evaluated Tier-1 production judges and Tier-2 cost-tier comparisons; Phase 6 added the Tier-3 frontier judges (April 2026 SoTA) on the agreement and bias-audit protocols; Phase 7 added the Tier-3 consistency runs. Total run count is 118, matching the production state ledger. Approximately $541{,}000$ judgments were produced.

Phase	Description	Protocols	Runs
1	Smoke test (1 model, 1 benchmark)	agreement	1
2	Tier 1 agreement	agreement	23
3	Tier 1 bias audit (MT-Bench)	bias-audit	8
4	Tier 1 consistency (MT, JB)	consistency	12
5	Tier 2 cost-tier coverage	agreement, bias-audit	14
6	Tier 3 SoTA agreement and bias	agreement, bias-audit	40
7	Tier 3 SoTA consistency (MT, JB)	consistency	20
Total			**118**

Table 10: **Evaluation phase summary.** Run counts are verified against the `.eval_state.json` ledger. Phase 1 was run before the protocol suite was finalized; Phases 2–7 are the production campaign. The 118-run total comprises 63 agreement runs (21 MT-Bench plus 21 JudgeBench plus 21 RewardBench), 21 bias-audit runs (all on MT-Bench), and 34 consistency runs (17 judges $\times$ 2 benchmarks).

G Use of AI Assistance

The authors used AI assistance (large language model–based writing assistants) for editing and manuscript-formatting support only: specifically, copy-editing prose, normalizing table and reference formatting, debugging LaTeX layout, and routine BibTeX hygiene. AI assistance was not used for original writing of research claims, for code generation in the evaluation experiments, or for any aspect of research design, hypothesis formulation, data collection, or data analysis. All experimental results, metric computations, and findings reported in this paper were produced and verified by the authors using the framework and methods described in Sections 3 and 4.

H Responsibility Checklist Notes

Potential Risks. The paper does not introduce a new judge model, training data, or deployment-ready system, so direct deployment harms are minimal. The principal indirect risks are: (i) selective adoption of the Minimum Viable Validation Protocol (Section 5) partial adoption, e.g. reporting Cohen's κ alone without the paired position-swap and consistency checks, can produce a false sense of having addressed judge reliability; (ii) interpreting the released evaluation dataset as authoritative beyond its temporal scope (March–April 2026) for judges whose hosted-model endpoints subsequently drift; and (iii) over-correction: practitioners discounting useful evaluation pipelines on the basis of the kappa-deflation finding rather than supplementing them with chance-corrected metrics. We discuss the temporal-scope risk explicitly in the Limitations.

Documentation of Artifacts. The released artifact is an evaluation dataset of 118 evaluation runs spanning 21 LLM judges, three benchmarks, and three protocols (approximately 541,000 individual judgments). *Languages:* all three source benchmarks are English; multilingual coverage is out of scope (Limitations). *Modality:* text only. *Domains:* MT-Bench is general open-ended conversational quality; JudgeBench spans mathematics, coding, creative writing, and analysis; RewardBench is chosen/rejected preference pairs across general instruction-following tasks. *Demographic groups:* the human-label annotations are inherited from each source benchmark (we do not collect new human annotations); please refer to the cited benchmark papers for annotator demographic coverage. *License:* the release will be distributed under a permissive open license at publication time. Each released record includes the per-item judge verdict, the judge's free-form reasoning trace, the raw model response, per-call latency, and the metric values computed for the run, keyed by (`judge_id`, `benchmark`, `protocol`, `item_id`, `run_idx`, `position_order`).

Descriptive Statistics. For each (judge, benchmark) cell in the agreement and bias-audit tables we report a single value per metric rather than means with confidence intervals: agreement-protocol values are population statistics computed over the full benchmark (not sample estimates), and bias-audit values are point estimates over paired AB+BA evaluations. Consistency-protocol values (test–retest, self-consistency, position flip rate) summarize $N \in [3, 5]$ replicate runs per item and are themselves point estimates of within-judge stability: Phase 4 (Tier 1) uses 5 replicates per item, Phase 5 (Tier 2) uses 5 replicates, and Phase 7 (Tier 3) uses 3 replicates. Cohort-level summary rows (e.g., the "Cohort mean" rows in Table 2 and Appendix Tables 4 and 6) are unweighted means across the judges in the relevant cohort. All single-run reporting was conducted at temperature 0 for deterministic generation.

Parameters for Packages. All evaluation runs were executed using the `[anonymized]` framework with the following per-protocol settings: temperature 0 for all judges, no response caching for the consistency protocol (caching enabled for agreement and bias-audit), and position-swap (AB+BA) debiasing in the consistency and bias-audit protocols. Metric implementations follow canonical formulations: Cohen's κ uses the standard two-rater nominal formula with no smoothing (Cohen, 1960); Krippendorff's α uses the nominal-distance function (Hayes and Krippendorff, 2007; Krippendorff, 2011); tie-excluded exact match follows the pairwise denominator restriction of Zheng et al. (2023); position bias is $|P(\text{A wins}) - 0.5|$ over paired AB+BA evaluations (Wang et al., 2024); verbosity bias is the Pearson correlation between response-length differential and verdict; test–retest is Krippendorff's α across replicate runs (Stureborg et al., 2024). Provider-specific reasoning-suppression settings appear in Table 8 (Appendix D). No external evaluation packages (NLTK, SpaCy, ROUGE, etc.) were used; all metric computations are reproducible from the released `[anonymized]` framework source.