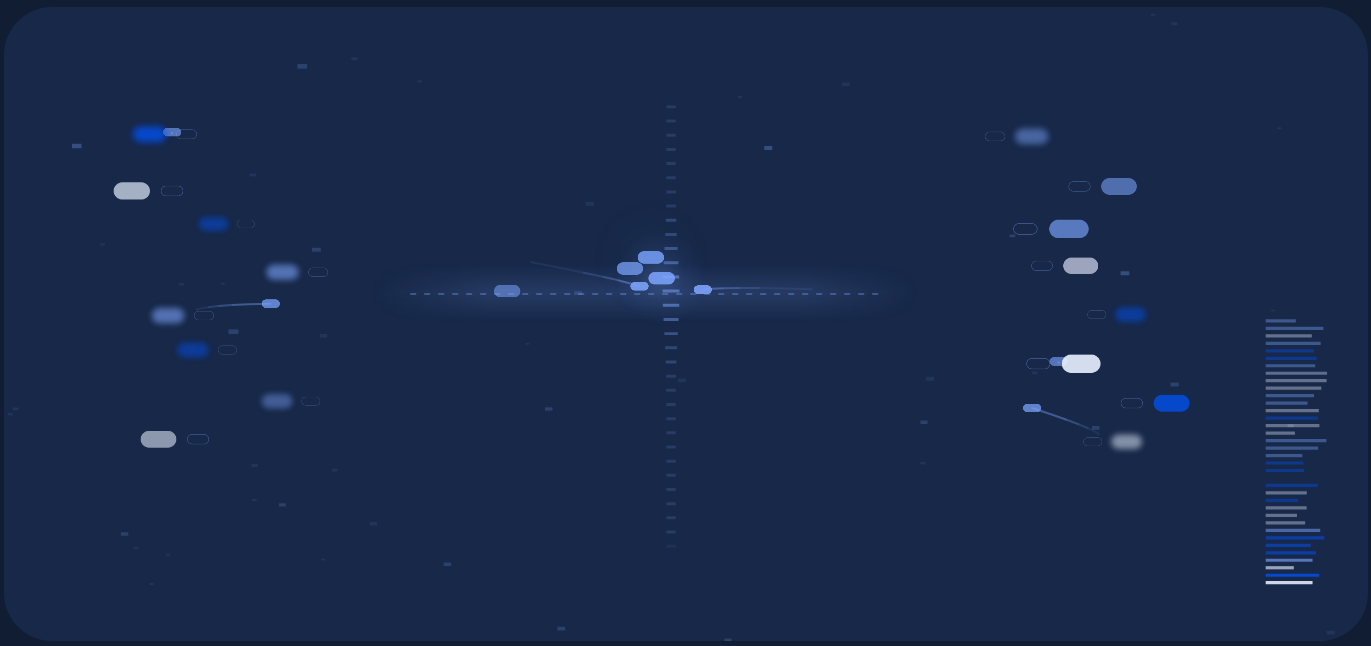
Rillor

METHODS NOTE

Validating an LLM judge against human labels

How to check an LLM-as-a-judge against human labels: chance-corrected agreement, probes for position, length and self-preference bias, re-validation when the judge changes, and corrected final estimates.



Rillor

Contents

01 Summary

02 Background

03 Method

04 Findings

05 Implications

06 Limits of this analysis

07 Sources

Summary

An LLM judge's scores mean something once the judge has been checked against human labels on the task it grades, and checked again after every change to it.

- Agreement between an LLM judge and people depends on the task. A 2023 study found over 80% agreement with human preferences on open-ended chat questions [1], while a 2024 study of 20 annotated datasets found that agreement varied with the property judged and the expertise of the labelers [2].
- Answer order changes pairwise verdicts. In a 2023 test, one commercial judge gave conflicting verdicts on 66 of 80 queries when the two answers swapped places [3]. In a July 2026 preprint, the last of four successive releases of a commercial family still changed 14.7% of its verdicts under the same swap [4].
- Percent agreement overstates how closely a judge matches people. In a 2024 study of 13 judges, an open-weight judge agreed with humans on well over 80% of answers yet scored 59 out of 100 on a chance-corrected measure [5].
- Judges favor longer answers and text like their own. Studies from 2023 and 2024 measured both effects [1, 6, 7], and a July 2026 preprint found that stronger judges reduce answer-order and length bias without removing it [4].
- A judge needs a validation set of human-labeled items from the task it grades. One published test compares the judge with at least three annotators on 50 to 100 items and supports using it when it represents the group better than a left-out annotator in at least half of the comparisons [8].
- A change to the judge is a change to the measurement. A July 2026 preprint concludes that judge upgrades are not interchangeable [4], and one commercial model's accuracy on a task fell from 84% to 51% between its March and June 2023 versions [9].
- A small human-labeled sample can correct a judge's bias in the final estimate. Prediction-powered inference subtracts the judge's measured error and gives valid confidence intervals [10], though no correction cuts the labels needed by more than half when the judge is no more accurate than the model it grades [11].

Background

LLM-as-a-judge is the use of large language models as evaluators for complex tasks [12]. A 2024 survey, last revised in October 2025, frames the field's central question as how to build reliable judge systems [12].

A 2023 study made the case with MT-Bench and Chatbot Arena: strong judges matched controlled and crowdsourced human preferences with over 80% agreement, the level at which people agree with each other [1]. On MT-Bench votes without ties, a commercial judge agreed with people 85% of the time and people agreed with each other 81%, against 50% by chance [1].

One 2024 study measured the saving in time and money [13]. In that study, whose authors include several at a technology company, an agent-based judge saved 97.72% of the time and 97.64% of the cost of three human experts on 55 development tasks [13].

A January 2026 guide to agent evaluation from one model provider lists model graders as non-deterministic, more expensive than code and in need of calibration with human graders [14]. It advises deterministic graders where possible, model graders where necessary and human graders with care, for additional validation [14].

This note is for engineers and evaluation leads who grade model or agent outputs with another model. It turns the published measurements, dated by study, into a validation procedure.

Method

What was examined. Published measurements of LLM judges against human labels, covering agreement, biases, reference answers, criteria drift, changes of judge and juries, and the statistical corrections that use human labels.

Sources. Thirty-two sources, each opened on 8 November 2026 and read in full text where figures are used: 14 papers accepted at conferences or workshops according to their arXiv pages, PDFs or ACL Anthology pages, 13 preprints, a journal article on the kappa statistic, a section of the NIST/SEMATECH e-Handbook of Statistical Methods, a practitioner guide from one model provider and two documentation pages from one cloud provider. Papers with no venue on an opened page are cited as preprints.

How. Figures come from each source's full text and carry the year of the study's first version, since bias shrinks in stronger judges [4]. Models are described by family and size. "Derived" marks arithmetic on a source's own numbers or formulas, and "illustrative" marks worked examples with invented counts, which measure nothing. Intervals use the Wilson method in NIST's handbook [15].

Where sources disagree. A 2024 study found that a panel of smaller judges from different families beat one large judge [17], and the July 2026 preprint found that mixed panels fell short of what independent errors predict [4]. One cloud provider recommends four judge samples per item to offset randomness [18], and the preprint found that repeated samples add little accuracy when errors correlate [4]. Both sides are reported, dated.

Findings

1. Agreement with people varies by task, property and labeler

JUDGE-BENCH (first released in 2024, accepted at ACL 2025) tested 11 open-weight and proprietary models as judges on 20 datasets with human annotations and over 70,000 test instances, using Cohen's kappa for categorical labels and Spearman correlation for graded ones [2]. On graded labels, judges correlated more with non-expert labelers than with experts, and on both label types they agreed more on human-written than on model-generated text [2]. No model led on every property, and toxicity and safety scored low, some below zero. Chain-of-thought prompting did not reliably raise agreement [2]. The authors conclude that LLMs "should be carefully validated against human judgments before being used as evaluators", and note that people and judges can look aligned in a domain because both are affected by similar biases [2].

Close calls lower agreement. In the 2023 MT-Bench study, agreement rose from 70% to nearly 100% as the performance gap between the two compared models grew [1]. LLMBar (2023) built 419 pairs in which one output follows the instruction and the other diverges while carrying qualities that can mislead a judge, such as a more engaging tone [16]. LLMBar reports expert agreement of 94% on 80 sampled pairs, 90% on its Natural set and 95% on its Adversarial set [16]. On its adversarial set, judges from three model families barely beat chance, and the top-scoring judge configuration averaged 82.8%, more than 10 points below the expert agreement rate [16]. JudgeBench (2024) labels response pairs on knowledge, reasoning, math and coding by objective correctness, and many strong judges scored only slightly better than random guessing [19].

Retrieval support labels show the same spread. In the TREC 2024 RAG track, a commercial judge and human assessors gave the same three-level label on 56% of the assessments made from scratch, and on 72% when the assessors post-edited the judge's labels [20].

Agent work adds a process to judge. On 55 development tasks with 365 requirements, the 2024 agent-based judge agreed with the consensus of human judges 90% of the time, against 70% for an LLM judge [13].

2. Position bias: answer order changes verdicts

The 2023 MT-Bench study gave judges two similar answers to each question, then swapped their order [1]. With the default prompt, the top commercial judge gave the same verdict in both orders 65.0% of the time, a second judge 46.2% and a third 23.8%, and the share of verdicts favoring the first slot was 30.0%, 50.0% and 75.0% [1].

Wang and colleagues (2023) measured the conflict rate, the share of queries whose verdict changes when the two answers swap slots, on 80 questions [3]:

Judge	Pair compared	Conflicts	Rate
Stronger commercial judge	A 13-billion-parameter open model and a second commercial model	37 of 80	46.3%

Judge	Pair compared	Conflicts	Rate
Stronger commercial judge	Two 13-billion-parameter open models	4 of 80	5.0%
Weaker commercial judge	The open model and the judge's own answers	66 of 80	82.5%
Weaker commercial judge	Two open models	42 of 80	52.5%

Under the weaker judge, the open model won 2.5% of comparisons against the judge's own answers in the first slot and 82.5% in the second [3]. Conflicts fell as the score gap between the two answers grew [3].

Shi and colleagues (first released in 2024, accepted at AACL-IJCNLP 2025) studied 15 judges across 22 tasks and over 150,000 evaluation instances [21]. Position bias was not due to random chance, varied across judges and tasks, and depended strongly on the quality gap between solutions [21]. In the July 2026 preprint, flip rates ran from 0.320 for a 1.7-billion-parameter open-weight judge to 0.117 to 0.147 for four releases of a commercial family [4].

Several fixes have been measured. The MT-Bench authors call the judge twice with the order swapped and declare a win only when an answer is preferred in both orders, a tie otherwise [1]. Few-shot examples raised their top judge's consistency from 65.0% to 77.5%, though high consistency may not imply high accuracy, and the longer prompts made calls 4 times as expensive [1]. Wang and colleagues propose evidence before the rating, ratings aggregated across both orders, and human labels for hard examples [3]. The first two improved alignment by 9.8% and 14.3% in accuracy for the two judges, and at a human annotation cost of 20% both judges matched or beat average human alignment [3]. One cloud provider's judge settings include an option that flips the response order on half of all calls [18].

3. Length, self-preference and other features of the answer

Length. In a 2023 MT-Bench attack, 23 answers had their list rephrased and prepended with no new information, and two of three judges preferred the padded answer 91.3% of the time, the third 8.7% [1]. A 2024 AlpacaEval paper fits a generalized linear model of the judge's preference on the length difference and other features, then predicts the preference at zero length difference. That raised the benchmark's Spearman correlation with Chatbot Arena from 0.94 to 0.98 [22]. In the July 2026 preprint, verbosity bias on LLMBar under a fixed padding string fell from 0.547 for the 1.7-billion-parameter open-weight judge to about 0.13 for the commercial releases [4].

Self-preference. In 2023, one judge favored itself with a 10% higher win rate and another with 25%, though both also favored other models and the authors could not determine whether a self-enhancement bias exists [1]. In 2024, three models told their own summaries apart from other sources with over 50% accuracy, one with 73.5%, and disproportionately favored them [6]. Fine-tuning showed a linear correlation between

self-recognition and the strength of self-preference [6]. Another 2024 study found that judges scored low-perplexity outputs higher than human evaluators did, whether or not the judge wrote them, which the authors read as a sign that self-preference comes from familiarity [7].

Other features. Ye and colleagues (2024) identify 12 potential biases, from position and verbosity to bandwagon, authority and sentiment, and find that significant biases persist in certain specific tasks [23]. Raina and colleagues (2024) found that a universal 5-token phrase, learned on a 3-billion-parameter surrogate and appended to an answer, pushed larger judges toward maximum scores, more so in absolute scoring than in comparisons [24]. In a 2024 study of 13 judges, judges were less robust when the answer was "Yes" or "Sure", and did not always mark a verbatim reference answer correct [5].

4. Percent agreement overstates alignment

Percent agreement is the number of agreement scores divided by the total number of scores [25]. In 1960 Jacob Cohen criticized it for not accounting for chance agreement [25]. Cohen's kappa and Scott's pi correct for chance [5, 25], and Krippendorff's alpha suits multiple raters and multiple possible ratings [25].

Thakur and colleagues (first version 2024, workshop paper 2025) studied 13 judges grading 9 exam-taker models on TriviaQA, a clean case in which three human judges each matched their majority vote with an average Scott's pi of 96.2 (scaled by 100) and 98.52% percent agreement [5]. The top three judges reached Scott's pi in the high 80s, and the top one sat 8 points behind people [5]. An 8-billion-parameter open-weight judge scored 59 with percent agreement well above 80% [5]. Judges with Scott's pi above 80 gave scores within 5 points of human scores, while judges above 90% percent agreement could differ by more than 10 points [5]. Most judges leaned lenient, marking answers correct when their criteria did not fully match the instructions [5]. Consistently biased judges can still rank models much as people do, so for ranking, simpler methods can be as suitable [5]. The authors moved from Cohen's kappa to Scott's pi between versions, citing documented issues with kappa, and recommend reporting both percent agreement and Scott's pi with qualitative analysis [5].

A worked example shows the gap. A judge and a person each mark 200 outputs pass or fail (illustrative counts):

	Person passes	Person fails	Total
Judge passes	120	20	140
Judge fails	10	50	60
Total	130	70	200

Percent agreement is 85%. The judge passes 70% and the person 65%, so chance agreement is 0.70 × 0.65 + 0.30 × 0.35 = 0.56 and Cohen's kappa is (0.85 - 0.56) / (1 - 0.56) = 0.66 (derived) [25]. Scott's pi, which pools the two pass rates at 67.5%, also gives 0.66 (derived) [5]. The judge's pass rate overstates the person's by 5 points, and the 20 outputs it passes and the person fails are its false passes. Skewed labels widen the gap. If people pass 180 of 200 outputs, a judge that passes everything agrees 90% of the time, chance agreement is also 0.90, and kappa is 0 (illustrative, derived) [25].

Thresholds are conventions. Cohen's bands read 0.61 to 0.80 as substantial and 0.81 to 1.00 as almost perfect [25]. McHugh (2012) calls them too lenient for health studies, reads any kappa below 0.60 as inadequate, and notes that many texts recommend 80% agreement as the minimum [25]. These conventions come from health research. No source opened for this note sets a threshold for LLM judges.

Agreement measured on a sample carries sampling error. NIST's handbook gives the Wilson method for proportions, whose lower limit cannot go negative, unlike the common normal approximation [15]. At an agreement rate of 85%, the 95% Wilson intervals are (derived) [15]:

Agreements	Items	95% Wilson interval
85	100	76.7% to 90.7%
170	200	79.4% to 89.3%
425	500	81.6% to 87.9%

With no disagreements in 100 items, the upper limit on the disagreement rate is still 3.7% (derived) [15].

5. Without a correct reference, a judge agrees closely only where it could answer

Krumdick and colleagues (first version 2025, COLM 2026) built BFF-Bench, 160 questions with long-form responses written by financial professionals, and had experts evaluate 1,200 responses [26]. Without a correct reference, judges showed high agreement with the experts only on questions the judges could answer themselves, and expert-written references largely mitigated the problem [26]. When one commercial judge used its own responses as the reference, pairwise Cohen's kappa fell from 0.86 to 0.16 on questions it could not answer [26]. Tests with wrong and random references showed that the reference's correctness mattered more than its style [26].

The 2023 MT-Bench study found the same on math. On 10 math questions judged in both orders, 20 judgments in all, the failure rate was 14 of 20 with the default prompt, 6 of 20 with chain-of-thought and 3 of 20 with a reference answer, a fall from 70% to 15% [1].

6. Criteria change as people grade

Shankar and colleagues (2024) built EvalGen, which uses people's grades on a subset of outputs to select judge implementations that align with them [27]. With nine industry practitioners, they observed criteria drift: people need criteria to grade outputs, and grading outputs helps them define the criteria. Even participants who graded first refined their criteria as they went and changed earlier grades [27]. The authors write that model-generated evaluators inherit the problems of the models they evaluate and need human validation [27].

The January 2026 guide's test for a well-specified task is that two domain experts would independently reach the same pass or fail verdict [14]. It warns that vague rubrics produce inconsistent judgments, and suggests a structured rubric for each dimension of a task, each graded by its own isolated judge [14].

7. Changing the judge changes the measurement

Yang and colleagues (July 2026, a single six-page version) held the candidate responses fixed and changed only the judge [4]. The panel was one open-weight family at four sizes, from 1.7 to 32 billion parameters, and four successive releases of a commercial family [4]. Adjacent judges were compared with exact McNemar tests, Holm-corrected over 18 tests, on four datasets that include LLMBar and Chatbot Arena [4].

- In the open-weight family, LLMBar accuracy rose from 0.463 at 1.7 billion parameters to 0.617 at 4 billion and 0.678 at 32 billion [4].
- Adjacent commercial releases differed in accuracy by at most 0.022, and none of the nine tests between them reached $p < 0.05$ even before correction [4]. After correction, only the open family's steps from 1.7 to 4 billion parameters on LLMBar and Chatbot Arena stayed significant [4].
- No judge won on every dataset. On one test set of 894 pairs, the top judge's kappa was 0.753 against a human ceiling of 0.920 [4].

The authors conclude that a judge-based score can move when only the evaluator changes, that judge upgrades are not interchangeable, and that reports should include dataset slices, bias probes, error-dependence estimates and protocol audit trails [4].

Served models change too. Chen and colleagues (2023) compared the March and June 2023 versions of two commercial model families served under the same names [9]. One went from 84% to 51% accuracy at telling prime from composite numbers, and the authors call for continuous monitoring [9].

Juries and repeated samples. Verga and colleagues (2024) found that a panel of smaller judges from disjoint model families beat a single large judge across six datasets, with less intra-model bias, at over seven times lower cost at the prices of the time [17]. The July 2026 preprint measured how far jurors' errors move together. On LLMBar, the errors of repeated samples from one judge correlated at 0.944 to 0.972 for the open-weight judges and at 0.664 to 0.706 for the commercial releases [4]. Majority votes of 1, 3 and 5 samples from the 1.7-billion-parameter judge reached LLMBar accuracies of 0.463, 0.475 and 0.482 [4]. Predictions that assume independent errors missed jury accuracy by a median of 0.078 and 0.093 on two

datasets, and by 0.008 and 0.004 once correlation was taken into account, and mixed juries also fell short of independence predictions [4]. One cloud provider's judge settings take 1 to 32 samples per item, recommend 4, and present sampling as a way to offset randomness [18]. Sampling steadies a verdict, and the 2026 data show it adds little accuracy when errors correlate [4, 18].

8. The validation procedure: calibration set, agreement, bias probes and re-validation

Step 1. Write and pilot the rubric. Grade a first batch of outputs by hand, revise the criteria, then freeze them for the validation run, since criteria change while people grade [27]. Give each rubric dimension its own judge [14].

Step 2. Build the calibration set. Sample outputs from the task and from the models the judge grades. Include close pairs [1, 21], outputs from the judge's own family [6], human-written and model-written text [2] and questions the judge cannot answer itself [26]. Label with expert annotators whenever possible [8]. The alternative annotator test of Calderon and colleagues (2025) recommends at least three annotators (it runs with two, less reliably) on 50 to 100 items, which the authors find sufficient in most cases, and at least 30 items for its t-test, with a Wilcoxon signed-rank test below that [8].

Step 3. Measure agreement. Report percent agreement with a chance-corrected statistic, paired with qualitative analysis [5, 25]. For pass and fail labels, report the confusion matrix and balanced accuracy, the mean of the true positive and true negative rates, as one cloud provider's judge-evaluation guide does [28]. Put a Wilson interval on each rate [15], report results by slice of the data [4], and count false passes apart from false fails, because a lenient judge raises pass rates [5].

Step 4. Test the judge against the annotators. The alternative annotator test leaves out one annotator at a time and tests whether the judge represents the others better than the excluded annotator does, with a cost-benefit term ε [8]. It runs a paired t-test per annotator, controls the false discovery rate (the authors recommend Benjamini-Yekutieli) and computes the winning rate ω, the share of rejected hypotheses. If ω is at least 0.5, the authors conclude that the judge is more likely to hold an advantage over the annotators, which justifies using it [8]. Their rule of thumb sets ε at 0.2 for experts, 0.15 for skilled annotators and 0.1 for crowdworkers [8]. Across ten datasets, they found that models can sometimes replace human annotators [8].

Step 5. Run the bias probes. Each probe reuses calibration items with one change:

Probe	Change to the item	What to measure	Source
Order	Swap the two answers	Flip rate, keeping a verdict only when both orders agree	[1, 3, 18]
Length	Pad an answer with restated content	Change in preference for the padded answer	[1, 22]

Probe	Change to the item	What to measure	Source
Own family	Include outputs from the judge's family	Judge score against human score on those items	[6, 7]
Reference	Score with and without a correct reference	Agreement on items the judge cannot answer	[1, 26]
Dummy answer	Replace the answer with "Yes" or "Sure"	Share marked correct	[5]
Appended phrase	Add a short phrase to the answer	Score change, with perplexity as a check	[24]

Step 6. Record the judge. Keep the judge's model and release, prompt, rubric, sampling settings and parse failures with every run, as the 2026 preprint's call for protocol audit trails implies [4].

Step 7. Re-validate on triggers. Rerun steps 3 to 5 on the frozen calibration set when the judge model or its release changes [4], after prompt edits, model drift or upstream changes in a chain [27], when the judge moves to a new task or property [2], and on a schedule, since a named service can change [9]. The January 2026 guide advises frequent calibration against expert judgment for research agents, and occasional human review once the system is robust [14].

9. Correcting estimates with human labels

Prediction-powered inference, introduced by Angelopoulos and colleagues in 2023, uses a judge's measured error to correct the final estimate. It gives valid inference when an experimental dataset is supplemented with predictions from a machine-learning system, with confidence intervals for means, quantiles and regression coefficients and no assumptions about the algorithm [10]. Treating predictions as gold-standard labels can give invalid conclusions because the predictions may be biased, and using only the labeled data is valid but limited [10]. The method estimates a rectifier from the labeled data, which captures how prediction errors lead to bias, and more accurate predictions give smaller intervals [10].

An example shows the arithmetic (illustrative counts, derived). A judge passes 80% of 2,000 unlabeled outputs. On 200 outputs that people also labeled, the judge passes 82% and people pass 74%. The rectifier is 74% - 82% = -8 points, so the corrected pass rate is 80% - 8 = 72%, with a confidence interval built from both samples [10, 29].

PPI++ (first version 2023, revised 2024) adapts to the quality of the predictions and gives confidence sets that always improve on classical intervals using only labeled data [29]. Boyeau and colleagues (ICML 2025) report effective human-labeled sample sizes up to 50% larger in experiments with one commercial

judge [30].

The savings have a ceiling. Dorner and colleagues (first version 2024, ICLR 2025) show that when the judge is no more accurate than the evaluated model, no debiasing method can cut the required ground-truth labels by more than half [11]. In practice the savings were more modest, and savings above a factor of two appeared only when state-of-the-art models judged substantially weaker ones [11]. The commonly reported agreement rate between judge and ground truth, they find, does not meaningfully restrict the judge's bias or the savings [11].

Two methods choose which items get human labels. Confidence-Driven Inference (NAACL 2025) uses the judge's annotations and confidence to select which human annotations to collect, keeps intervals valid and no less accurate than human labels alone, and cut the human annotations needed by over 25% in each of three settings [31]. Cascaded Selective Evaluation (2024) trusts a judge only when its confidence supports a user-specified level of human agreement, and escalates the rest to stronger judges [32]. On a subset of Chatbot Arena where a commercial judge almost never reached 80% human agreement, it guaranteed over 80% agreement while judging almost 80% of items without abstaining [32].

Implications

For engineers who build a judge. Validation starts before the judge runs: a rubric piloted by hand, a calibration set drawn from the task and labels from expert annotators where possible [8, 27]. Report percent agreement next to a chance-corrected statistic, with intervals, because high percent agreement can hide score gaps of more than 10 points [5, 15].

For evaluation leads who read judge scores. A judge-based score needs four things beside it: the calibration set's size and labelers, the agreement statistic with its interval, the bias probe results and the judge's release [4, 8]. A score from a changed judge is a new measurement until the calibration set has been rerun [4, 9]. Ranking and scoring ask different things of a judge, since a consistently strict judge can still rank models much as people do [5].

For final numbers. A judge's pass rate carries the judge's bias. A human-labeled sample of the same outputs lets prediction-powered inference correct it with valid intervals [10, 29], and the saving in labels is capped at half when the judge is no more accurate than the model it grades [11].

For agents. An agent-based judge was closer to human consensus than an LLM judge in one 2024 study [13]. The same procedure can apply, with transcripts in the calibration set.

Limits of this analysis

- Most bias figures come from judges of 2023 and 2024 [1, 3, 6, 24]. The July 2026 preprint shows smaller biases in stronger judges, on one dataset per probe, one model panel and a single version [4].
- The studies grade text: chat answers, summaries, short factual answers, instruction following, finance questions and retrieval support labels [1, 5, 6, 16, 20, 26]. Agent work appears in one study [13].
- The kappa thresholds come from health research [25], and the ε values are the alternative annotator test authors' rules of thumb [8].
- Krippendorff's alpha is named through McHugh's review [25], and Scott's pi is computed with the pooled formula in the 2024 judge study's appendix [5]. Cohen's and Scott's original papers were not opened for this note.
- Cloud documentation was viewed on 8 November 2026 and can change [18, 28]. Cost comparisons are dated: the panel study's prices are from 2024 [17], and the agent-based judge's savings come from one study of 55 tasks [13].

Sources

1. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. Lianmin Zheng, Wei-Lin Chiang, Ying Sheng and others, NeurIPS 2023 Datasets and Benchmarks Track, arXiv 2306.05685, version 4, 24 December 2023, viewed 8 November 2026. https://arxiv.org/abs/2306.05685

2. LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. Anna Bavaresco, Raffaella Bernardi, Leonardo Bertolazzi and others, ACL 2025, arXiv 2406.18403, version 3, 2 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2406.18403

3. Large Language Models are not Fair Evaluators. Peiyi Wang, Lei Li, Liang Chen and others, arXiv 2305.17926, version 2, 30 August 2023, viewed 8 November 2026. https://arxiv.org/abs/2305.17926

4. When the Judge Changes, So Does the Measurement: Auditing LLM-as-Judge Reliability. Zongyou Yang, Yinghan Hou and Xiaokun Yang, arXiv 2607.08535, version 1, 9 July 2026, viewed 8 November 2026. https://arxiv.org/abs/2607.08535

5. Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan and Dieuwke Hupkes, Proceedings of the Fourth Workshop on Generation Evaluation and Metrics (GEM2 2025), pp. 404-430, arXiv 2406.12624, version 6, 18 August 2025, viewed 8 November 2026. https://arxiv.org/abs/2406.12624

6. LLM Evaluators Recognize and Favor Their Own Generations. Arjun Panickssery, Samuel R. Bowman and Shi Feng, arXiv 2404.13076, version 1, 15 April 2024, viewed 8 November 2026. https://arxiv.org/abs/2404.13076

7. Self-Preference Bias in LLM-as-a-Judge. Koki Wataoka, Tsubasa Takahashi and Ryokan Ri, NeurIPS 2024 Safe Generative AI Workshop, arXiv 2410.21819, version 2, 21 June 2025, viewed 8 November 2026. https://arxiv.org/abs/2410.21819

8. The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. Nitay Calderon, Roi Reichart and Rotem Dror, Proceedings of ACL 2025 (Volume 1: Long Papers), pp. 16051-16081, Association for Computational Linguistics, July 2025, viewed 8 November 2026. https://aclanthology.org/2025.acl-long.782/

9. How is ChatGPT's Behavior Changing over Time? Lingjiao Chen, Matei Zaharia and James Zou, arXiv 2307.09009, version 3, 31 October 2023, viewed 8 November 2026. https://arxiv.org/abs/2307.09009

10. Prediction-Powered Inference. Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan and Tijana Zrnic, arXiv 2301.09633, version 4, 9 November 2023, viewed 8 November 2026. https://arxiv.org/abs/2301.09633

11. Limits to scalable evaluation at the frontier: LLM as Judge won't beat twice the data. Florian E. Dorner, Vivian Y. Nastl and Moritz Hardt, ICLR 2025, arXiv 2410.13341, version 4, 17 August 2026, viewed 8 November 2026. https://arxiv.org/abs/2410.13341

12. A Survey on LLM-as-a-Judge. Jiawei Gu, Xuhui Jiang, Zhichao Shi and others, arXiv 2411.15594, version 6, 19 October 2025, viewed 8 November 2026. https://arxiv.org/abs/2411.15594

13. Agent-as-a-Judge: Evaluate Agents with Agents. Mingchen Zhuge, Changsheng Zhao, Dylan Ashley and others, arXiv

2410.10934, version 2, 16 October 2024, viewed 8 November 2026. https://arxiv.org/abs/2410.10934

14. Demystifying evals for AI agents. Mikaela Grace, Jeremy Hadfield, Rodrigo Olivares and Jiri De Jonghe, Anthropic, Engineering, 9 January 2026, viewed 8 November 2026. https://www.anthropic.com/research/demystifying-evals-for-ai-agents

15. NIST/SEMATECH e-Handbook of Statistical Methods, 7.2.4.1 Confidence intervals. National Institute of Standards and Technology, viewed 8 November 2026. https://www.itl.nist.gov/div898/handbook/prc/section2/prc241.htm

16. Evaluating Large Language Models at Evaluating Instruction Following. Zhiyuan Zeng, Jiatong Yu, Tianyu Gao and others, ICLR 2024, arXiv 2310.07641, version 2, 16 April 2024, viewed 8 November 2026. https://arxiv.org/abs/2310.07641

17. Replacing Judges with Juries: Evaluating LLM Generations with a Panel of Diverse Models. Pat Verga, Sebastian Hofstätter, Sophia Althammer and others, arXiv 2404.18796, version 2, 1 May 2024, viewed 8 November 2026. https://arxiv.org/abs/2404.18796

18. Configure a judge model, Gemini Enterprise Agent Platform documentation. Google Cloud, last updated 7 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/configure-judge-model

19. JudgeBench: A Benchmark for Evaluating LLM-based Judges. Sijun Tan, Siyuan Zhuang, Kyle Montgomery and others, ICLR 2025, arXiv 2410.12784, version 2, 5 April 2025, viewed 8 November 2026. https://arxiv.org/abs/2410.12784

20. Support Evaluation for the TREC 2024 RAG Track: Comparing Human versus LLM Judges. Nandan Thakur, Ronak Pradeep, Shivani Upadhyay and others, SIGIR 2025, arXiv 2504.15205, version 1, 21 April 2025, viewed 8 November 2026. https://arxiv.org/abs/2504.15205

21. Judging the Judges: A Systematic Study of Position Bias in LLM-as-a-Judge. Lin Shi, Chiyu Ma, Wenhua Liang and others, AACL-IJCNLP 2025, arXiv 2406.07791, version 9, 11 November 2025, viewed 8 November 2026. https://arxiv.org/abs/2406.07791

22. Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. Yann Dubois, Balázs Galambosi, Percy Liang and Tatsunori B. Hashimoto, COLM 2024, arXiv 2404.04475, version 2, 10 March 2025, viewed 8 November 2026. https://arxiv.org/abs/2404.04475

23. Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. Jiayi Ye, Yanbo Wang, Yue Huang and others, arXiv 2410.02736, version 2, 4 October 2024, viewed 8 November 2026. https://arxiv.org/abs/2410.02736

24. Is LLM-as-a-Judge Robust? Investigating Universal Adversarial Attacks on Zero-shot LLM Assessment. Vyas Raina, Adian Liusie and Mark Gales, arXiv 2402.14016, version 2, 4 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2402.14016

25. Interrater reliability: the kappa statistic. Mary L. McHugh, Biochemia Medica 22(3), 15 October 2012, via PubMed Central, viewed 8 November 2026. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900052/

26. No Free Labels: Limitations of LLM-as-a-Judge Without Human Grounding. Michael Krumdick, Charles Lovering, Varshini Reddy and others, COLM 2026, arXiv 2503.05061, version 4, 28 September 2026, viewed 8 November 2026. https://arxiv.org/abs/2503.05061

27. Who Validates the Validators? Aligning LLM-Assisted Evaluation of LLM Outputs with Human Preferences. Shreya Shankar, J.D. Zamfirescu-Pereira, Björn Hartmann and others, arXiv 2404.12272, version 1, 18 April 2024, viewed 8 November 2026. https://arxiv.org/abs/2404.12272

28. Evaluate a judge model, Gemini Enterprise Agent Platform documentation. Google Cloud, last updated 7 November 2026, viewed 8 November 2026. https://docs.cloud.google.com/gemini-enterprise-agent-platform/models/evaluate-judge-model

29. PPI++: Efficient Prediction-Powered Inference. Anastasios N. Angelopoulos, John C. Duchi and Tijana Zrnic, arXiv 2311.01453, version 2, 26 March 2024, viewed 8 November 2026. https://arxiv.org/abs/2311.01453

30. AutoEval Done Right: Using Synthetic Data for Model Evaluation. Pierre Boyeau, Anastasios N. Angelopoulos, Tianle Li and others, ICML 2025, arXiv 2403.07008, version 3, 1 June 2026, viewed 8 November 2026. https://arxiv.org/abs/2403.07008

31. Can Unconfident LLM Annotations Be Used for Confident Conclusions? Kristina Gligorić, Tijana Zrnic, Cinoo Lee, Emmanuel J. Candès and Dan Jurafsky, NAACL 2025, arXiv 2408.15204, version 2, 8 February 2025, viewed 8 November 2026. https://arxiv.org/abs/2408.15204

32. Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. Jaehun Jung, Faeze Brahman and Yejin Choi, arXiv 2407.18370, version 1, 25 July 2024, viewed 8 November 2026. https://arxiv.org/abs/2407.18370

About this report

Published by Rillor on 9 November 2026. Superintelligence (SI) tools assist research and drafting, and every figure cites its source. Rillor builds and secures SI systems, supplies the data they learn from and delivers the infrastructure they run on. We apply SI to financial and compute markets.

Cite as: Rillor. Validating an LLM judge against human labels. 9 November 2026. https://rillor.com/insights/validating-an-llm-judge

Notices

Superintelligence (SI) outputs can be wrong. Important decisions should include human review.

Work with Rillor

Send us the tasks your agent handles, the model versions it runs on and the decisions it is trusted with. Contact us at rillor.com/contact?type=evaluation or contact@rillor.com. More on SI evaluation: rillor.com/si/evaluation