

Agreement Is Not a Criterion: What Matching Human Labels Certifies About an LLM Judge, Why a Shared Shortcut Passes the Test, and Why Every External Criterion Found So Far Validates a Use Rather Than a Judge

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. An automatic judge is usually accepted on one piece of evidence: it agrees with human preference labels about as often as humans agree with each other. That number is then read as permission to use the judge for ranking models, filtering data, selecting samples and supplying reward. This paper separates three claims the ritual runs together - that the judge can stand in for the labelers on a given item population, that it tracks the quality construct the labels were meant to measure, and that decisions made with it turn out well in a stated use - and argues that agreement evidence bears directly only on the first. A stylised two-loading model makes the reason exact: agreement with a single reference identifies one linear combination of a judge's sensitivity to the construct and to the surface features the labelers also respond to, so a judge that shares the labelers' shortcuts can match a judge that tracks quality. The published record is then read through that separation. Human labels carry documented surface sensitivities that judges reproduce; judges carry failures of their own that natural validation sets never exercise; and agreement is concentrated on the easy, wide-gap items where decisions matter least. In the corners where an external criterion exists - verified correctness, downstream policy outcomes, randomised business outcomes and verifiable task success - the criterion has repeatedly exposed failures that agreement did not show, and where it confirms an offline score the confirmation does not transfer across uses: a benchmark score that predicts best-of-N selection well gives only a rough signal for PPO. Validity, on this evidence, is a property of a judge in a use. The paper consolidates these published results, states a use-indexed validation procedure as an algorithm, and proposes the tests that would settle what remains open.

Keywords: LLM-as-a-judge; evaluation validity; inter-rater agreement; construct validity; criterion validity; reward model evaluation; multitrait-multimethod; measurement theory

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The LLM-as-a-judge paradigm arrived with its own validation argument attached. On the first turn of the MT-Bench questions, GPT-4 used as a pairwise judge agreed with expert human votes 66 percent of the time when ties were counted and 85 percent when they were excluded, against 63 and 81 percent between the human experts themselves (Zheng et al., 2023, Table 5). The authors' conclusion was that a strong judge approximates human preference at a fraction of the cost, and the inference pattern - match the inter-human rate, then substitute - has since become a common warrant for automatic evaluation. AlpacaFarm reported that a single-prompt GPT-4 annotator agreed with human annotators at 65 percent, against 66 percent between humans (Dubois et al., 2023). G-Eval was introduced on the strength of a Spearman correlation of 0.514 with human summarisation ratings (Liu et al., 2023). Arena-Hard-Auto was presented as reaching 98.6 percent correlation with human preference rankings (Li, T., et al., 2024), and length-controlled AlpacaEval as raising its Spearman correlation with Chatbot Arena from 0.94 to 0.98 (Dubois et al., 2024). In computational social science, a widely cited result reported that ChatGPT's zero-shot

accuracy exceeded crowd workers' on four of five annotation tasks and that its intercoder agreement exceeded both crowd workers' and trained annotators' on all of them (Gilardi et al., 2023). A meta-analysis of 3,334 natural-language-generation papers found LLM judges appearing in more papers than human evaluation by 2025, and only 254 of the 679 papers using an LLM judge (37.4 percent; under 8 percent of all 3,334 papers) validating it against human judgments (Yang, J., et al., 2026).

The objection to this warrant is old in measurement theory and recent in the judge literature. Agreement between two raters is a reliability statistic. It says that two instruments produce the same output; it does not say that the output measures what either instrument was meant to measure (Cronbach and Meehl, 1955; Messick, 1995). The judge literature has now said this in its own words. A 21-judge audit of roughly 541,000 judgments titles itself "Reliability without Validity" and reports that chance-corrected agreement sits 33 to 41 percentage points below raw agreement on MT-Bench for every judge it tested (Norman et al., 2026). A construct-validity study finds that judges leave their verdicts unchanged under meaning-preserving edits far more often than they change them under meaning-changing ones, and that a predictor reading only surface form

recovers 67.4 percent of MT-Bench's human votes (Chen, J., et al., 2026). A study of real A/B tests, pre-registered in its main arm, reports in an exploratory human baseline that fifteen conversion-optimisation experts agreed with each other at kappa 0.53 while scoring at chance against the tests' actual outcomes (Dooskin and Squoosh Technical Staff, 2026).

These results are usually filed as "judges have biases" and answered with a debiasing step. This paper argues they show something more structural, and less comfortable. The question a judge validation must answer is not whether the judge agrees with humans but which of three different claims the agreement is being used to support:

Substitution (V1). On this item population, the judge's labels can stand in for the labelers' labels. Agreement is the right evidence for this claim, and by construction it is sufficient.

Construct (V2). The judge's verdicts track the quality construct - correctness, helpfulness, faithfulness - that the labels were meant to measure. Agreement supports this only to the extent that the labels themselves track the construct and that the judge does not reach the labels by a different route.

Use (V3). Decisions made with the judge's verdicts - a leaderboard order, a best-of-N pick, a gradient - lead to good outcomes for a stated use. Agreement supports this only through V2, and only if the use exercises the judge on items like those in the validation set.

The ritual measures V1 and is cited for V2 and V3. The paper's contributions follow from taking that separation seriously. The first is a stylised derivation (Section 2.3) showing that agreement with a single reference identifies only one linear combination of a judge's construct sensitivity and its surface sensitivity, so that a judge which shares the labelers' shortcuts cannot be told apart from one that tracks quality by agreement alone; separating them needs a second reference with a different construct-to-surface ratio, or an intervention that moves one while holding the other. This is the multitrait-multimethod argument of Campbell and Fiske (1959) restated for judges. The second is a reading of the published record through V1/V2/V3 (Sections 3 to 5), organised by where agreement is blind: shared surface sensitivity, judge-only failure modes that natural validation items never exercise, and the concentration of agreement on easy items. The third is a consolidation of the four corners where an external criterion exists (Section 5, Table 1, Figure 1), with the observation that the criterion results are themselves indexed to a use: the same benchmark score predicts one optimiser's outcome and not another's. The fourth is a map of why the proposed substitutes (bias probes, item response theory, statistically rigorous agreement tests, invariance-and-sensitivity profiles, external criteria) do not compose into a single validity certificate (Section 7, Table 2), together

with a use-indexed validation procedure (Algorithm 1) that licenses only the claim whose evidence is present.

Two scope statements belong here. First, this paper is not about whether a second model provides an independent check on a first; that is the correlated-failure question treated in an earlier paper in this series, which explicitly set aside "whether human agreement is the right criterion for validating an automatic judge" as a separate dispute (Mahendrakar, 2026). This paper takes up that dispute. Second, it is not a catalogue of judge biases; those are well surveyed (Gu et al., 2024; Ye et al., 2024). Biases enter here only where they bear on what an agreement number can certify.

The analysis that follows works from the published record. Every number below is reported from the cited source, and where a source is a 2026 preprint without peer review that is said on the spot.

2. What an agreement number is

2.1 Reliability, validity and the reference problem

Classical test theory distinguishes the consistency of a measurement from its meaning. Reliability is the extent to which repeated or parallel measurements agree; validity is the extent to which the inferences drawn from a score are warranted (Cronbach and Meehl, 1955). Messick (1995) recast validity as a property of interpretations and uses rather than of instruments, and Kane (2013) turned that into an argument-based programme: a validation names the interpretation and use being claimed, lists the inferences that claim needs, and gathers evidence for each. Borsboom, Mellenbergh and van Heerden (2004) pushed back from the other side, arguing that a test is valid for an attribute only if the attribute exists and variation in it causes variation in the scores. The two positions disagree about much, but they agree on the point that matters here: agreement with another instrument is evidence about the pair, and becomes evidence about the attribute only through an assumption about the other instrument.

The judge literature inherits this problem in its sharpest form, because the "other instrument" is itself a human rating that nobody has validated. Human preference labels are treated as ground truth in the validation step and as a noisy, biased signal everywhere else. The same field that validates judges against crowdsourced votes has documented that crowd workers cannot tell GPT-3 text from human text above chance without training (Clark et al., 2021), that Mechanical Turk workers, unlike English teachers, could not separate model stories from human references (Karpinska et al., 2021), and that the system ranking produced by WMT crowd raters differed substantially from the ranking produced by professional translators doing explicit error annotation, to the point that automatic metrics built on pre-trained embeddings outperformed the crowd workers as raters (Freitag et al., 2021). Plank (2022) argues that much of

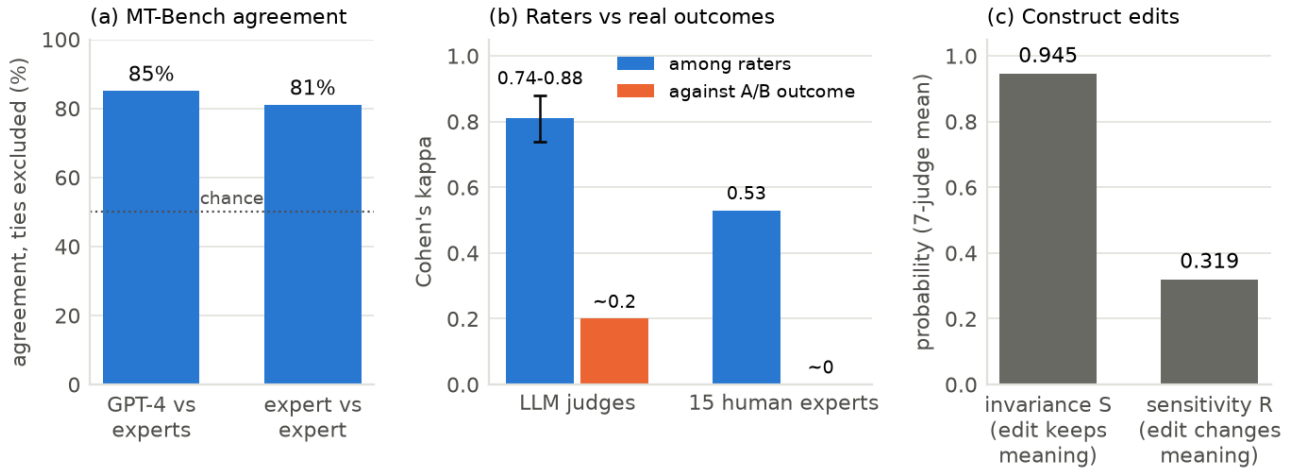


Figure 1. Agreement and validity measured on the same items within each rater group. (a) MT-Bench pairwise agreement, excluding ties, between GPT-4 and expert humans and between experts (Zheng et al., 2023, Table 5; chance is 50%). (b) Cohen's kappa among raters versus kappa against the recorded winners of real A/B tests, for LLM judges and for fifteen human experts, the judges on about 159 shared tests and the experts on a separate 50-item instrument (Dooskin and Squoosh Technical Staff, 2026; judge inter-rater range 0.74-0.88 plotted as its midpoint with a range bar; "about 0.2" and "about 0" plotted as 0.2 and 0.0, exploratory per the source). (c) Judge verdict invariance under construct-preserving edits versus sensitivity to construct-changing edits, averaged over seven judges at an operating point fixed at $S \geq 0.90$ (Chen, J., et al., 2026).

what annotation projects treat as label noise is genuine variation from subjectivity and multiple plausible answers, so a single aggregated label may not be a property of the item at all. Bowman and Dahl (2021) list reliable annotation among the four things a benchmark must get right and observe that most do not.

2.2 Agreement statistics and their known defects

Agreement has its own statistical problems before validity enters. Raw percent agreement does not correct for chance; Cohen's kappa does (Cohen, 1960), and conventional verbal bands for kappa were proposed by Landis and Koch (1977). Kappa has its own paradoxes: with skewed marginals, two raters can agree on almost every item and still obtain a low kappa (Feinstein and Cicchetti, 1990). Krippendorff (2004) sets out what a reliability coefficient for coded data must account for, agreement expected by chance among them. The judge literature has rediscovered each of these. Norman et al. (2026) find raw exact-match agreement exceeding Cohen's kappa by 33 to 41 percentage points on MT-Bench across all 21 judges, and judge rankings shifting by as many as 14 positions between benchmarks. Thakur et al. (2024) switched from Cohen's kappa to Scott's pi after the issues with kappa were pointed out to them, and report that the chance-corrected coefficient separates judges that percent agreement leaves indistinguishable, while even well-aligned judges can assign scores up to 5 points away from the human ones. Schroeder and Wood-Doughty (2024) apply McDonald's omega to repeated judge samples and show that a single deterministic sample can misstate a judge's reliability.

None of this touches validity. A perfectly chance-corrected, perfectly test-retest-stable agreement

coefficient is still a statement about two raters. The problem this paper is about survives every one of these repairs.

2.3 A stylised model of why agreement cannot separate the construct from the shortcut

The following is a derivation in a deliberately simple linear model, not a measurement. Its purpose is to state exactly what agreement can and cannot identify.

Suppose each response has a construct value C (say, correctness) and a surface value S (say, a composite of length, formatting and assertive tone), standardised and, for simplicity, uncorrelated in the item population. A human label is $H = cC + dS + e_H$, and a judge's verdict is $J = aC + bS + e_J$, where the error terms are independent of everything else and of each other. Then the covariance between judge and human is

$$\text{cov}(J, H) = ac + bd,$$

and the covariance between two independent human labelers drawn from the same pool is $c^2 + d^2$.

Proposition 1 (stylised). The covariance between a judge and a single reference H depends on the judge's loadings only through $ac + bd$. Hence (i) for any reference with $d > 0$ and any target covariance $k > 0$, the judges that achieve it form a line in the (a, b) plane, and that line contains a judge with no construct loading at all ($a = 0, b = k / d$); and (ii) when the reference weights construct and surface equally ($c = d$), swapping a judge's construct and surface loadings leaves both its covariance and its correlation with H unchanged, because the swap also preserves the judge's variance $a^2 + b^2 + \text{var}(e_J)$.

The proof is one line: the covariance is a linear functional of (a, b), its level sets are lines, and when $c = d$ it is symmetric in a and b. Three consequences follow, each of which the published evidence below instantiates.

First, a judge can match the inter-human agreement rate without tracking the construct at all, provided the labelers respond to the surface feature and their own noise variance is large relative to their construct loading ($\sigma^2 \geq c^2 (c^2 + d^2) / d^2$, where $\sigma^2 = \text{var}(e_H)$). In correlation terms, two labelers correlate at $(c^2 + d^2) / \text{var}(H)$, while a noise-free judge that responds only to the surface feature correlates with one labeler at $d / \text{sd}(H)$; the second is at least the first whenever $d \text{sd}(H) \geq c^2 + d^2$. As a purely arithmetic illustration, labelers with $c^2 = 0.2$, $d^2 = 0.3$ and noise variance 0.5 correlate with each other at 0.50, and a pure-surface judge correlates with each of them at about 0.55. The common practice of treating "judge-human agreement at or above human-human agreement" as a validity threshold is therefore satisfiable, in this model, by a judge that has learned nothing but the labelers' shortcut. Note also that a noise-free judge reproducing the systematic part of H - whatever mix of C and S that part contains - correlates with a single labeler at the square root of the inter-labeler correlation, and so always exceeds it, because the judge drops the e_H term that two humans do not share. Zheng et al. (2023) observed the analogous ordering in percent agreement (85 versus 81 percent) and read it as the judge capturing the human majority; the model says that reading is correct, and that the majority's systematic component includes whatever surface preference the majority has.

Second, a judge that is better than the labelers on the construct (larger a, smaller b than the reference's c and d) can be penalised by agreement. Agreement rewards resemblance to the reference, including resemblance to its errors.

Third, identifying a separately requires a second equation. That can come from a second reference with a different construct-to-surface ratio - experts rather than crowd, verified correctness rather than preference - or from an intervention that changes C with S held fixed, or S with C held fixed. This is precisely the logic of the multitrait-multimethod matrix, which asks for high agreement between different methods measuring the same trait (convergent validity) together with low agreement between measurements of different traits, especially when they share a method (discriminant validity) (Campbell and Fiske, 1959). In organisational psychology the same structure is called common method variance, and its standard remedies are procedural separation of the methods and statistical control for the shared component (Podsakoff et al., 2003). The judge-validation ritual measures one cell of that matrix: the heterotrait-confounded agreement of one method (the judge) with another (crowd preference) on one composite trait.

The model is crude. Constructs and surface features are correlated in real data; loadings vary by item; verdicts are discrete. None of that rescues the identification argument, which only needs that the human labels carry some systematic response to features other than the construct and that the judge can respond to those features too. Whether they do is an empirical matter, and Section 3 takes it up.

3. Human labels carry surface sensitivity, and judges can share it

3.1 Documented surface loadings in human preference

The evidence that human preference labels respond to features other than the quality construct they are meant to measure is extensive and consistent.

Hosking, Blunsom and Bartolo (2023) found that single preference scores cover most error types but under-represent factuality, and that assertiveness distorts error detection: relative to the authors' own expert annotations, crowd workers under-counted factuality and inconsistency errors, and the shortfall grew for assertively worded responses and shrank for hedged ones. Wu and Aji (2023) built a set of deliberately flawed answers to 40 general questions and had them compared by crowd workers, experts, GPT-4 and Claude-1. Answers with factual errors outranked answers that were merely too short or ungrammatical; the crowd workers in particular showed little sign of fact-checking; and every judge, human and model, favoured the longer answer. Their pairwise Cohen's kappa table is a small instance of the whole problem: crowd and expert annotators agreed at 0.08, each human group agreed with each model at 0.09 to 0.14, and the two models agreed with each other at 0.51 (Wu and Aji, 2023, Table 3). Sharma et al. (2023) analysed existing human preference data and found that responses agreeing with the user's stated view were more likely to be preferred, and that both humans and preference models preferred a convincingly written sycophantic response over a correct one a non-negligible fraction of the time. Li, J., et al. (2024) decomposed real-world preference annotations and report that human raters are less sensitive to errors, favour responses that support their own stance and clearly dislike responses that admit the model's limits. Zhang, X., et al. (2024) report that human evaluators, GPT-4 and top reward models all show strong preferences for particular formats - lists, bold text, links, emojis - and that less than 1 percent of biased data suffices to inject such a bias into a reward model. AlpacaFarm's own appendix reports that humans and its simulated annotators both prefer longer outputs that contain lists (Dubois et al., 2023).

The effect is not limited to preference over style. Wen, J., et al. (2024) gave time-constrained human subjects the task of judging correctness on a reading-comprehension task and a programming task, and found that after RLHF

the models became better at convincing the subjects without becoming better at the task: the subjects' false positive rate rose by 24.1 percent on QuALITY and 18.3 percent on APPS. That reads as a direct measurement of a surface loading in human judgment that grows when a policy is optimised against it. Chen, G. H., et al. (2024) built a framework that needs no ground-truth labels to probe human and LLM judges for four biases - overlooking misinformation, gender, deference to authority and attractive presentation - and found both populations vulnerable to varying degrees.

The most direct measurement of surface recoverability in a validation set is recent. Chen, J., et al. (2026), a 2026 preprint, report that a frozen model family reading only surface form, given the same paired input a judge receives, reproduces 55 to 67 percent of the labels in five public label sets, including 67.4 percent of MT-Bench's human votes. They also report that RewardBench's control constraining chosen responses to be no longer than rejected ones did not remove the length confound but inverted it. If those figures hold up, a substantial share of the agreement any judge can achieve on those sets is available without reading for meaning.

3.2 Evidence that judges reach the labels by the same route

The second half of the identification problem is whether judges load on the same surface features as the labelers. The evidence here points both ways, and the disagreement is informative.

On the shared side: length is the classic case. AlpacaEval's authors describe their automatic annotator as highly correlated with human preferences while favouring longer outputs, and introduce a regression adjustment to answer what the preference would have been at equal length (Dubois et al., 2024). Singhal et al. (2023) show that in three RLHF settings much of the reward improvement came from longer responses, and that a reward based purely on length reproduces most of the downstream gains over the supervised baseline. Format bias is shared across humans, GPT-4 and reward models in the data of Zhang, X., et al. (2024). Feuer et al. (2024) find that LLM judges, given explicit criteria, implicitly reweight them, prioritising style over factuality and safety, and that judge preferences do not correlate with concrete measures of safety, world knowledge and instruction following.

On the unshared side: Saito et al. (2023) find GPT-4 preferring longer answers more than humans do, which is a judge-specific surface loading rather than a shared one. Li, J., et al. (2024) find that advanced LLM judges weight correctness, clarity and harmlessness more than human raters do, which suggests that in their data the judge is closer to the construct than the human reference - the situation the second consequence of Proposition 1 describes. And Norman et al. (2026) report verbosity bias below 0.011 for all 21 judges in their 2026 cohort

under a single pairwise rubric, against much larger figures reported for 2023 judges.

This last result complicates the shared-shortcut story and should be stated plainly. If 2026 judges have largely stopped responding to length while the label sets they are validated against still let length predict the label (as the inverted RewardBench length control suggests), then for that particular feature the judge and the reference have moved apart, and agreement should fall for reasons that have nothing to do with judge quality. The model in Section 2.3 predicts exactly that: reducing b while d stays fixed lowers $a c + b d$ unless a rises to compensate. No located study has checked whether the length debiasing of recent judges cost them agreement with older label sets. It is a cheap test, and it would distinguish "judges improved" from "judges diverged from a biased reference" in the one feature where both have been measured.

The general lesson is not that judges share every human shortcut. It is that agreement cannot tell which shortcuts are shared, which are judge-specific, and which human errors the judge has avoided, because all three move the same number.

4. Two further blind spots: failures the validation set never exercises, and agreement that lives on easy items

4.1 Judge-only failure modes are invisible to agreement on natural items

A validation set drawn from natural model outputs exercises a judge only on the kinds of inputs those models happened to produce. Several of the most consequential judge failures are triggered by inputs no natural output contains.

Zheng, X., et al. (2024) showed that a "null model" emitting one constant response, unrelated to the instruction, reached an 86.5 percent length-controlled win rate on AlpacaEval 2.0, an 83.0 score on Arena-Hard-Auto and 9.55 on MT-Bench. Zhao et al. (2025) found that generative reward models used in reference-based verification can be made to issue positive rewards by inputs as slight as a colon or a generic reasoning opener such as "Thought process:", across proprietary and open models. Wang, P., et al. (2023) showed that the order in which two candidates are presented can flip a judge's preference often enough that Vicuna-13B beat ChatGPT on 66 of 80 queries with ChatGPT as the evaluator. Panickssery, Bowman and Feng (2024) found that LLM evaluators score their own outputs above others' that human annotators rate as equal in quality, and that the strength of this self-preference correlates linearly with the evaluator's ability to recognise its own text across fine-tuned variants. Li, D., et al. (2025) describe preference leakage, in which a judge favours students trained on data it or a related model generated.

Hu et al. (2024) find LLM evaluators confusing distinct quality criteria under targeted perturbations; Stureborg, Alikaniotis and Suhara (2024) find a familiarity bias toward lower-perplexity text and anchoring across attributes.

Few of these failures would show up as lower agreement with humans on a validation set of ordinary responses presented in balanced order. Ordinary responses are not constant strings and are not padded with master keys, and self-preference and preference leakage appear only when the judge's own family or students are among the candidates. Norman et al. (2026) name the resulting pattern a consistency-bias paradox: two production-deployed judges in their cohort had test-retest reliability above 0.95 and position bias above 0.10. A judge can be stable, agree with humans and still be trivially exploitable. In V1/V2/V3 terms, these are not V1 failures at all - the judge can still stand in for humans on the validation population - but they become V3 failures for any use in which the judged outputs are themselves optimised against the judge, because optimisation is how inputs outside the natural population are found.

4.2 Agreement is concentrated where decisions are easy

The second blind spot is distributional. Zheng et al. (2023) report that GPT-4's agreement with humans rose from 70 percent to nearly 100 percent as the win-rate gap between the compared models widened. Read as validation evidence, that is reassuring. Read as a statement about where agreement comes from, it says that a validation set dominated by wide-gap pairs will report high agreement regardless of how the judge behaves on close pairs - which is where many leaderboard and release decisions are made.

Bodhwani, Tran and Wei (2026), a paper accepted to an industry track, measured this directly for an LLM-judge release gate over 25 task-oriented agents. The gate's ranking matched a verifiable task-success reward across the broad capability span, but the decision-disagreement rate rose from under 1 percent on wide-reward pairs to 31 percent on close pairs. Their summary - a gate that is human-validated yet mis-anchored - is the V1/V3 gap in one phrase. Thakur et al. (2024) report the same asymmetry from the other side: for ranking nine exam-taker models, even small judges and a lexical "contains" metric gave a reasonable signal, while absolute scores from the best judges still diverged from humans. Coarse rankings are easy; fine discriminations are not.

A related concentration runs through item difficulty for the judge. Krumdick et al. (2025) had financial professionals grade 1,200 long-form responses and found that, without a correct reference answer, LLM judges agreed highly with the experts only on the questions the judges could answer correctly themselves; supplying an expert-written reference largely closed the gap. They also report that 15 of the 40 GPT-4-generated reference

answers they reviewed for an MT-Bench subset (37.5 percent) were incorrect or inconsistent. Szymanski et al. (2024) report subject-matter experts agreeing with LLM judges 68 percent of the time in dietetics and 64 percent in mental health. Dorner, Nastl and Hardt (2024) prove that when the judge is no more accurate than the model being evaluated, no debiasing method can cut the number of ground-truth labels needed by more than half. All of these say the same thing: the items on which a judge is most needed - those at or beyond its own competence - are the items on which agreement-based validation tells us least.

5. The corners where an external criterion exists

If agreement cannot certify V2 or V3, what can? In four places the published record contains something better than agreement: a criterion that is not another rating. Each is narrow, and each tells us something the ritual could not.

5.1 Verified correctness

The most direct substitute for a preference label is a label that can be checked. JudgeBench converts difficult knowledge, reasoning, mathematics and coding problems into response pairs whose preference label follows from objective correctness, and reports that many strong judges, GPT-4o among them, perform only slightly better than random guessing (Tan et al., 2024). Its authors' motivation is the point of this paper in miniature: existing judge benchmarks mostly test alignment with human preference, which is a poor indicator of factual and logical correctness on hard tasks. Norman et al. (2026) find JudgeBench spreading judges across 60.4 percentage points against 13.5 on MT-Bench, so the correctness-based set also discriminates far better. The original RewardBench already included subsets whose preferred answer is fixed by a verifiable reason, such as a bug or an incorrect fact (Lambert et al., 2024). RM-Bench pairs responses that differ subtly in content under controlled style variation and reports that even state-of-the-art reward models averaged 46.6 percent - below the 50 percent chance level - when style bias interfered (Liu et al., 2024).

Verified correctness is a genuine second equation in the sense of Proposition 1: its construct-to-surface ratio differs sharply from crowd preference. But it covers only constructs that admit verification, and on those constructs a verifier or test suite is often available directly, which raises the question of why a judge is needed. Its value as a validation criterion is greatest exactly where it is least available.

5.2 Downstream policy outcomes

Reward models are judges whose verdicts are used as an optimisation target, and here the field has gone furthest in testing V3 directly. The results are the most important

in this paper, because they show that even a good external criterion validates a use rather than a judge.

RewardBench 2 reports that across 113 reward models, average benchmark score and average downstream score under best-of-N sampling correlate at a Pearson coefficient of 0.87 (Malik et al., 2025). Under PPO, run with 17 reward models from a fixed SFT policy, the picture changes: the benchmark gives a rough signal at the low-scoring end, but downstream performance saturates for all reasonable reward models whose benchmark scores spanned 49.8 to 68.5, and reward models initialised from a different base model than the policy often fail to improve it however high they score. The authors identify the policy-to-reward-model lineage as a factor their benchmark cannot see and verify that the same limitation affects other reward-model benchmarks. Wen, X., et al. (2024), in a synthetic setting, find only a weak positive correlation between reward model accuracy and downstream policy performance, that reward models of similar accuracy can produce quite different policies, and that accuracy predicts regret better under best-of-N than under PPO. Razin et al. (2025) prove that a reward model inducing low reward variance on the policy's outputs yields a flat objective landscape however accurate it is, so a perfectly accurate reward model can teach more slowly than a less accurate one, and a reward model that works well for one policy can fail for another.

Frick et al. (2024) took the most direct route: nine reward models were each used to train one base model with an off-policy DPO-style procedure, the resulting models were deployed on Chatbot Arena, and their human preference scores became the ground truth against which twelve reward-model metrics were compared. They report that fine-grained measures such as accuracy were the best predictors, that lower-quantile performance across prompt categories predicted better than averages, and - a striking aside - that among top reward models RewardBench score had become negatively correlated with downstream RLHF performance. Note what the criterion is here: post-training human preference on Chatbot Arena. That is a better criterion than offline agreement, because it measures the outcome of a use. It is still a preference rating, and Chatbot Arena's rankings carry their own documented distortions from private testing and selective disclosure (Singh et al., 2025), although the crowd votes there were reported to agree well with expert raters (Chiang et al., 2024).

The AlpacaFarm study already showed the general pattern in 2023. Its single-prompt GPT-4 annotator matched human-human agreement (65 versus 66 percent) but failed to capture human annotator variability, and the authors report that using it as a training signal led to qualitatively different results for method development, most visibly in whether reward over-optimisation appeared (Dubois et al., 2023). Over-

optimisation itself follows different functional forms under best-of-N and under reinforcement learning (Gao, Schulman and Hilton, 2022), which is a mechanism for why the same proxy can be adequate for one optimiser and not the other. Manheim and Garrabrant (2018) distinguish regressional, extremal, causal and adversarial forms of Goodhart failure; the null-model and master-key results of Section 4.1 are the adversarial form, and the BoN/PPO divergence may involve the extremal one, though no located study has separated the two.

5.3 Real-world outcomes

Three recent studies validate a judge against an outcome that is not a rating at all. All three are 2026 preprints, and one is an industry-track paper.

Chen, L., et al. (2026) tested a seven-dimension LLM-judged rubric for sales conversations on a matchmaking platform against verified conversion. In 60 human conversations, two dimensions were significantly associated with conversion (Spearman rho 0.368 and 0.354) and one showed no detectable association (rho 0.018); the equal-weighted composite (rho 0.272) underperformed its best dimensions, and reweighting by conversion recovered part of the loss (rho 0.351). The authors are careful that this is a finding about rubric design, not judge accuracy - any rater applying the same rubric would face it - which is itself the point: agreement on rubric scores would have validated a composite that a criterion shows to be diluted.

Dooskin and Squoosh Technical Staff (2026) asked whether a multimodal judge can predict the winner of real A/B tests from screenshots. On 330 tests the judge reached kappa 0.14 against the recorded winners, and 0.11 - with a confidence interval including zero - on the half of tests whose results were statistically significant. They report that 44 percent of the "ground truth" labels in an agency's curated case library came from non-significant tests, and that the judge agreed more with those unreliable labels than with the reliable ones. Their exploratory measurements, so labelled in their own claims ledger, give the paper's most quotable pair of numbers: judges differing in model or prompt agreed with each other at kappa 0.74 to 0.88 while agreeing with real outcomes at about 0.2, and fifteen human experts agreed with each other at kappa 0.53 while scoring at about zero against outcomes. A single study from a company with an interest in the product should be weighed accordingly; it is cited here because it reports, for both humans and judges and on the same items within each group (judges on about 159 shared tests, experts on a separate 50-item instrument), agreement among raters alongside agreement with an outcome - the comparison this paper's argument turns on.

Bodhwani, Tran and Wei (2026), discussed in Section 4.2, add a second finding: in their blind human panel, 57.5 percent of conversations rated as satisfying had

failed the customer's task, across five rater populations and two benchmarks. Satisfaction - the construct the judge was validated against - was nearly uninformative about success, the construct the deployment cared about.

5.4 The older precedents

Machine translation and summarisation lived through this argument a decade earlier with automatic metrics, and the outcome is instructive. When Freitag et al. (2021) replaced crowd ratings with professional MQM error annotation, the system ranking changed, and automatic metrics built on pre-trained embeddings outperformed the crowd workers as raters. Bhandari et al. (2020) found that conclusions about which summarisation metrics agree best with humans, drawn on older datasets and systems, did not hold for modern ones. SummEval re-evaluated 14 automatic metrics against both expert and crowd-sourced judgments of the same model summaries (Fabbri et al., 2020). In both fields the lesson was that the human reference had to be validated before it could validate anything else - and that a better reference changed the answer.

5.5 What the criterion corners establish

Table 1 consolidates the measured values the argument rests on. Two things are established and one is not.

Established: in every corner where an external criterion has been applied, it exposed failures that agreement with human preference had not shown - near-random accuracy on verified correctness, judge preferences uncorrelated with concrete capability measures, chance-level prediction of real outcomes, satisfaction ratings blind to task failure, and offline reward-model scores that stopped predicting RL outcomes across most of their range.

Established: the criterion results themselves are indexed to a use. RewardBench 2's score predicts best-of-N selection at $r = 0.87$ and PPO outcomes only roughly; accuracy predicts regret better under BoN than PPO in an independent synthetic study; and the best reward model for RL depends on the policy it trains. A judge validated for one use has not thereby been validated for another.

Not established: that any criterion corner generalises beyond its own construct and use. Verified correctness exists only for verifiable constructs. Downstream policy outcomes exist only for judges used as reward. Business outcomes exist for one platform at a time. There is no located result that validates a general-purpose judge for general-purpose use, and on the argument of this section none should be expected.

Figure 1 puts three of these results side by side. Panel (a) is the number the ritual reports. Panels (b) and (c) are what happens when the same raters are checked against something other than each other: high agreement among raters coexists with near-chance agreement with

outcomes, and, at an operating point chosen to hold invariance at 0.90 or above, sensitivity stays low.

6. Reading the disagreement: why both sides are right about different claims

The literature contains two positions that look incompatible. One holds that LLM judges should be validated against human judgments before use and that, once validated, they can replace humans for the validated task (Bavaresco et al., 2024; Calderon, Reichart and Dror, 2025). The other holds that agreement with humans is the wrong test, because humans are a poor reference on exactly the hard cases (Tan et al., 2024; Krumdick et al., 2025; Chen, J., et al., 2026). Through the V1/V2/V3 separation they stop conflicting.

The first position is correct about V1. Bavaresco et al. (2024) evaluated 11 models on 20 datasets of human annotations and found reliability varying with the property being judged, the expertise of the annotators and whether the text was human- or model-written; their conclusion that judges must be validated against human judgments before use is the right conclusion if what one wants is a substitute for those human judgments. The alternative annotator test of Calderon, Reichart and Dror (2025) is a careful statistical procedure for exactly that claim: it asks whether an LLM annotator is at least as good as a randomly chosen human from the annotator pool, using a modest annotated subset. Trust or Escalate (Jung, Brahman and Choi, 2024) provides a provable guarantee of human agreement at a chosen level by abstaining on low-confidence items, and reports guaranteeing over 80 percent human agreement with almost 80 percent coverage on a Chatbot Arena subset using a cascade starting from Mistral-7B. Prediction-powered inference and its relatives make it possible to draw valid statistical conclusions about a human-label quantity from a few human labels plus many model predictions (Angelopoulos et al., 2023; Gligoric et al., 2024). These are real advances. All of them guarantee closeness to the human reference. None of them can guarantee more, because the human label is the estimand.

The second position is correct about V2 and V3. When the reference is replaced by a criterion - verified correctness, a policy outcome, a business outcome - judges that would pass V1 have repeatedly failed it. The rating-indeterminacy result of Guerdan et al. (2025) shows that the failure can begin inside the human reference itself: when items admit several reasonable ratings and raters are forced to choose one, standard forced-choice validation selected judges performing as much as 31 percent worse than the judges selected using multi-label response-set ratings, and in their toxicity example ranked the truly best judge fourth.

What neither position supplies is a general route from V1 evidence to a V2 or V3 conclusion, and Section 2.3

Table 1. Measured values the argument rests on, one row per cited result. "Agr." is agreement; "crit." is an external criterion. All values are as reported by the named source.

Source	Setting	Compared against	Reported value
Zheng et al. 2023	MT-Bench, GPT-4 pair judge	expert votes (no ties)	agr. 85%; human-human 81%
Dubois et al. 2023	AlpacaFarm GPT-4 annotator	crowd preference	agr. 65%; human-human 66%
Wu and Aji 2023	crowd, experts, GPT-4, Claude-1	each other	kappa: crowd-expert 0.08; models 0.51
Norman et al. 2026	21 judges, MT-Bench	human labels	raw minus kappa: 33-41 pp
Chen, J. et al. 2026	surface-only predictor	MT-Bench human votes	67.4% of labels recovered
Chen, J. et al. 2026	7 judges, 4 domains, operating point fixed at $S \geq 0.90$	construct edits	invariance 0.945; sensitivity 0.319
Wen, J. et al. 2024	humans after RLHF	gold correctness	FPR + 24.1% QuALITY, + 18.3% APPS
Zheng, X. et al. 2024	constant null output	AlpacaEval 2.0 judge	86.5% LC win rate
Tan et al. 2024	JudgeBench	verified correctness	GPT-4o slightly above random
Liu et al. 2024	RM-Bench, style interference	subtle content labels	46.6% average accuracy
Malik et al. 2025	113 RMs, best-of-N	downstream tasks	Pearson $r = 0.87$
Malik et al. 2025	17 RMs, PPO	downstream tasks	saturates for scores 49.8-68.5
Krumdick et al. 2025	MT-Bench subset	expert review of refs	15 of 40 GPT-4 refs incorrect or inconsistent
Chen, L. et al. 2026	rubric composite	verified conversion	rho 0.272; reweighted 0.351
Dooskin 2026	judge panel, 330 A/B tests	test winners	kappa 0.14; 0.11 if significant
Dooskin 2026	judges: each other / outcome	real outcomes	kappa 0.74-0.88 vs about 0.2
Dooskin 2026	15 human experts: each other / outcome	real outcomes	kappa 0.53 vs about 0
Bodhwani et al. 2026	blind human panel, 25 agents	verified task success	57.5% of "satisfied" failed
Bodhwani et al. 2026	judge gate, wide vs close pairs	verifiable reward	disagreement under 1% vs 31%
Guerdan et al. 2025	forced-choice validation	response-set ratings	chosen judge up to 31% worse

explains why there is none: the step needs an assumption about the reference's construct-to-surface ratio that the V1 data cannot test.

6.1 The strongest case for agreement: when preference is the construct

There is one situation in which the separation collapses, and the argument of this paper should be stated against its best opponent. If the construct a judge is meant to measure is defined as the preference of a particular population - what users of a chat product prefer, as Chatbot Arena's votes are meant to record (Chiang et al., 2024) - then the labelers' surface sensitivity is not a contaminant. It is part of the construct. A user population that prefers bulleted answers prefers bulleted answers; a judge that reproduces that preference is measuring exactly what was asked. In the notation of Section 2.3, the distinction between C and S dissolves, and V1 and V2 become the same claim.

This case is real and should not be argued away. Three qualifications keep it from rescuing the general ritual. First, the population has to be the one whose preference is wanted. Validating a judge against crowd workers and then using it to stand in for domain experts, or for the end users of a different product, is a V1 claim about the wrong population; the expert-versus-crowd results of Freitag et al. (2021) and Wu and Aji (2023), and the expert-versus-judge results of Szymanski et al. (2024), show how far apart those populations can be. Second, preference is rarely the construct the downstream

decision cares about. The GAUGE panel's satisfaction ratings were almost uninformative about whether the customer's task had been completed (Bodhwani, Tran and Wei, 2026). A judge validated for preference is validated for preference; whether preference is the right target is a separate question that the validation cannot answer. Third, preference measured offline on static items is not preference under optimisation. Wen, J., et al. (2024) show human judgments of correctness becoming easier to satisfy without the task being done once a policy has been trained with RLHF, and Hosking, Blunsom and Bartolo (2023) offer preliminary evidence that training on human feedback disproportionately increases assertiveness. If the construct is "what this population prefers when it has not been optimised against," a judge validated on unoptimised items has not been validated for the optimised regime.

So agreement is sufficient for V2 exactly when the construct is defined as the reference population's preference, on the population and regime in which the judge will be used. That is a narrow condition. It is also, in fairness, the condition under which MT-Bench and Chatbot Arena were originally proposed: their stated aim was to approximate human preference, not to measure correctness. The drift from "approximates human preference" to "is a valid evaluator" appears to have happened largely downstream of the original papers, in the uses to which their judges were put.

6.2 Why adding judges does not repair the reference

A common response to doubts about a single judge is to use several and aggregate. For V1 this can help, because idiosyncratic judge noise averages out. For the problem this paper is about, it does not, because a panel of judges that share each other's route to the labels is still one method in the Campbell-Fiske sense. Dooskin and Squoosh Technical Staff (2026) quantify this for their A/B-test judges: verdicts from judges differing in model or prompt were so correlated that, by their intraclass-correlation estimate, a 16-vote panel carried about two effective independent votes - and their fifteen human experts, measured the same way, amounted to fewer than two independent judgments. Kim et al. (2025), across more than 350 models, find that models agree on 60 percent of the items where both err on one leaderboard dataset, and that larger, more accurate models have highly correlated errors even across architectures and providers. Wu and Aji (2023) report GPT-4 and Claude-1 agreeing with each other at kappa 0.51 while each agreed with either human group at 0.09 to 0.14. Model-model agreement in these data is consistently higher than model-human agreement, which is what one would expect if the models share a route to their verdicts that humans do not - and what one would also expect if they share a route with each other and a different one with humans. Aggregation pools the shared component rather than removing it. The correlated-failure argument of the earlier paper in this series applies here with the reference in place of the checked model (Mahendrakar, 2026).

7. Why the proposed substitutes do not compose

Five families of method have been offered as improvements on raw agreement. Each answers a different question, and Table 2 lays out what each certifies, what it presupposes and what it cannot see. The point of the table is that the cells do not add up to a validity certificate; each method's presupposition is another method's subject.

Bias probes. Perturbation suites such as CALM (Ye et al., 2024), the Judge Reliability Harness (Dev et al., 2026) and aspect-targeted perturbation attacks (Hu et al., 2024) inject a known manipulation and measure whether the verdict moves. They are the right tool for detecting a specific defect. Their logic is asymmetric: a probe that finds a bias proves the judge is not invariant to that manipulation; a probe that finds nothing certifies only the manipulations that were tried. The Judge Reliability Harness reports that no judge it evaluated was uniformly reliable across its benchmarks, which is a presence result.

Item response theory. Choi et al. (2026) fit a graded response model to judge outputs and separate intrinsic consistency under prompt variation from alignment with

human quality assessments. This localises unreliability to items and prompts, which is valuable for V1. But the latent trait in an IRT model is defined by the responses it is fitted to; it presupposes that those responses measure one construct, and its "human alignment" dimension inherits the reference problem.

Statistically rigorous agreement. The alternative annotator test, selective evaluation with agreement guarantees and prediction-powered inference (Section 6) make V1 claims exact and efficient. They cannot reach V2.

Invariance and sensitivity profiles. Chen, J., et al. (2026) formalise construct validity for an evaluator as two quantities: the probability that a verdict survives an edit that should not change it (invariance, S) and the probability that it changes under a minimal edit that should (sensitivity, R). They show the two are independent, measure them over seven judges and four domains with generation, verification and judging assigned to disjoint model families and direction fixed by human annotators, and report average S of 0.945 against R of 0.319 at matched invariance, with sensitivity to scope edits (0.383) exceeding sensitivity to strength edits (0.262) for every judge. This is the closest thing in the located literature to a test of V2, and it is the second equation Proposition 1 asks for: an intervention that moves the construct with the surface held fixed. It still depends on humans - not for labels on whole responses but for the direction of each edit, a local and contrastive judgment on which their annotators agreed at Fleiss' kappa 0.776 - and it is constructed for one construct at a time (claim specificity, in their case). That is the right trade: it moves the human judgment to where humans are more reliable. It is not yet a general instrument, and it does not address V3.

External criteria. Section 5 showed what these establish and that each is indexed to a use.

8. A use-indexed validation procedure

If validity is a property of a judge in a use, then a validation report should state the use, gather the evidence that use needs, and license only what the evidence supports. Algorithm 1 states that decision structure. It is a proposal, assembled from the measurements the cited papers already know how to make; it has not been run as a whole by anyone.

Three features of Algorithm 1 deserve comment. Step 2 stratifies agreement by pair difficulty because the evidence in Section 4.2 shows that pooled agreement is dominated by easy pairs. Step 3 audits the reference before using it, following the lesson of Section 5.4 and the measurement of Chen, J., et al. (2026). Step 6 refuses to transfer a criterion result between optimisers, following Malik et al. (2025), Wen, X., et al. (2024) and Razin et al. (2025).

Table 2. Raw agreement and the five proposed improvements: what each certifies, what it presupposes, and the failure it cannot see.

Method	Certifies	Presupposes	Cannot see
Raw or chance-corrected agreement	V1 on the sampled items	reference is the estimand	shared shortcuts; close pairs
Bias probes	a named defect is present	defect list is complete	untried manipulations
IRT models	where the judge is unstable	one latent construct	construct identity
Agreement tests, PPI, guarantees	V1, with error control	human label is the target	anything beyond V1
Invariance and sensitivity	construct response, one construct	edit direction by humans	other constructs; uses
External criterion	V3 for one use and population	criterion fits the use	other uses

Algorithm 1. Use-indexed validation of an automatic judge J Input: judge J; use U in {substitute, rank, filter, select-best-of-N, reward-for-RL, release-gate}; item population P; reference R Output: the strongest claim in {none, V1, V2, V3(U)} the evidence licenses

Step 1: Declare the claim sought and the use U. If U optimises against J (select-best-of-N, reward-for-RL), mark J as an optimisation target. Step 2: V1 evidence: estimate chance-corrected agreement of J with R on items drawn from P. Report it separately for close and wide pairs, and for items J can and cannot solve unaided. If agreement fails, return none. Step 3: Audit R: (a) fit a surface-only predictor to R and report the share of labels it recovers; (b) check for rating indeterminacy with response-set elicitation; (c) where possible, compare R against a second reference with a different construct-to-surface ratio. If (a) is high and (c) is absent, cap the claim at V1. Step 4: V2 evidence: for each construct claimed, measure invariance S under construct-preserving edits and sensitivity R_c under minimal construct-changing edits, with edit direction fixed by humans who do not see J's output. Report S and R_c jointly; do not average them. If R_c is low, cap the claim at V1. Step 5: If J is an optimisation target: run null-output and master-key probes and a position-swap test. Any failure rules out V3 for this U and caps the claim at whatever Steps 2-4 support. Step 6: V3 evidence: obtain a criterion for U itself (verified correctness, outcome of the optimiser actually used, business or task outcome). Validate under the optimiser and policy lineage that U will use; a result under best-of-N does not license RL use. Step 7: Return the strongest claim whose evidence is present, with U and P attached. A cap set in Step 3 or Step 4 bounds the construct claim (V2) only; Step 6 evidence for U licenses V3(U) despite such a cap, but never despite a Step 5 failure for the same U. A claim without its U and P is not a validation claim.

8.1 Three published cases, read through Algorithm 1

Applying the procedure to cases already in the literature shows what it changes in practice. In each case the evidence is the published evidence; what changes is the claim it is allowed to carry.

An MT-Bench GPT-4 judge used to rank chat models.

Step 2 is met only on raw agreement, not on the chance-corrected agreement the step asks for: first-turn percent agreement with experts matched the inter-expert rate (Zheng et al., 2023), and on MT-Bench raw agreement runs 33 to 41 points above kappa (Norman et al., 2026). Stratification by gap is partly available in the same paper, which shows agreement falling toward 70 percent as the compared models approach each other. Step 3 fails its audit: a surface-only predictor recovers 67.4 percent of the MT-Bench human votes (Chen, J., et al., 2026), and no second reference with a different construct-to-surface ratio was used in the original validation. The licensed claim is V1 - the judge can stand in for the MT-Bench expert voters on MT-Bench-like items - with a note that close-pair agreement is lower. That is what the original paper claimed. The claim that the judge is a valid evaluator of response quality is not licensed by this evidence.

RewardBench 2 accuracy used to pick a reward model. Step 6 is unusually well supplied. For best-of-N selection with an in-distribution reward model, the benchmark score predicts the outcome ($r = 0.87$; Malik et al., 2025), so V3 is licensed for that use. For PPO, the same paper shows saturation across the middle of the score range and failure for reward models from a different lineage than the policy; V3 for RL use is not licensed by the benchmark score and would need the PPO-under-lineage test Malik et al. ran. Step 5 is not supplied by the benchmark: a reward model used as an RL target should also be checked against null-output and master-key probes (Zheng, X., et al., 2024; Zhao et al., 2025), which accuracy on curated pairs does not exercise.

An LLM-judge release gate for task-oriented agents.

The GAUGE gate passes V1 against human satisfaction ratings and passes a ranking check across the broad capability range against a verifiable reward (Bodhwani, Tran and Wei, 2026). It fails step 2's close-pair stratification (31 percent decision disagreement on close pairs) and fails the construct audit, because the satisfaction construct it was validated on was nearly uninformative about task success. The licensed claim is

V3 for coarse release decisions between agents of clearly different capability, and nothing for promotion decisions between near-equal agents - which is the decision a release gate exists to make.

None of these readings is new to the authors of the cited papers, each of whom states the relevant limit. The procedure's contribution is to put the limits in the same grammar, so that a reader of a leaderboard or a reward-model card can see which claim is being made and which evidence was offered for it.

8.2 What the procedure leaves out

What the procedure does not do is equally important. It does not produce a general certificate, because on the argument of Sections 2 and 5 none is available. It does not make validation cheap; steps 4 and 6 are expensive. And it does not settle what to do when no criterion exists for U, which is the common case for open-ended generation. In that case the honest output is V1 or V2, with the gap stated.

9. What would settle the open part

Five studies would turn the argument above from a reading of the literature into a set of measurements. None requires new methods; each requires running measurements that already exist on the same items.

A judge multitrait-multimethod matrix. Take one item set and collect ratings of two constructs (for example, correctness and helpfulness) by three methods (domain experts, crowd workers, an LLM judge). The Campbell-Fiske criteria then say directly whether judge-crowd agreement on helpfulness exceeds judge-crowd agreement across constructs - which would indicate shared method variance - and whether judge-expert convergence on correctness holds. The prediction from Section 3 is that heterotrait-monomethod correlations among crowd and judge will be higher than the ritual assumes.

Residualised agreement. Predict human labels from surface form alone, as Chen, J., et al. (2026) did for five label sets, and recompute judge agreement on the items the surface predictor gets wrong, or on the residual of a regression of labels on surface features. If agreement falls sharply, the judge was plausibly reaching the labels through the surface route. This is the statistical-control remedy for common method variance (Podsakoff et al., 2003) applied to judges.

The length-debiasing test. Norman et al. (2026) report near-zero verbosity bias in 2026 judges; the inverted RewardBench length control suggests length still predicts labels in some sets. Compare 2023 and 2026 judges' agreement with the same older label sets, stratified by whether the preferred response is the longer one. Section 3.2 predicts that newer judges lose agreement on length-favoured items, and that the loss is a reference defect, not a judge defect.

A use-transfer test. Validate the same panel of judges against criteria for three uses - offline filtering, best-of-N selection, RL reward - on a shared item population. If the rankings of judges differ across uses, as the reward model evidence predicts, use-indexed validation is not a refinement but a necessity.

Close-pair reporting. Re-analyse existing validation sets with agreement reported separately for pairs within a small win-rate gap. Zheng et al. (2023) already report agreement by gap; the step needed is to make close-pair agreement the headline number for any judge used in ranking.

10. What this paper does not show

The argument is about what agreement evidence can license, not about whether any particular judge is good. Several limits should be stated flatly.

The two-loading model in Section 2.3 is a stylised derivation. It shows that agreement with one reference cannot separate construct from surface loading; it does not estimate any loading, and the evidence that real human labels carry substantial surface loading comes from the cited studies, not from the model.

Several of the strongest results are recent preprints without peer review. Chen, J., et al. (2026), Dooskin and Squoosh Technical Staff (2026), Chen, L., et al. (2026) and Norman et al. (2026) are all 2026 preprints; Bodhwani, Tran and Wei (2026) is also a 2026 paper but has been accepted to an industry track. The A/B-test study comes from a company with a commercial interest, labels its cross-judge and human-baseline measurements exploratory, and the conversion study rests on 60 conversations from one platform. If the preprint results fail to replicate, Section 5.3 shrinks to a single industry-track paper and Section 3.1 loses its most direct measurement. Section 2, Section 4.1, Section 5.2 and the rest of Section 3 rest on earlier or peer-reviewed work, with one exception: Sections 2.2, 3.2 and 4.1 also draw on Norman et al. (2026), a 2026 preprint. Otherwise they cite the 2026 preprints only for supporting detail.

The external-criterion corners are not representative of judge use. Most judge deployments concern open-ended generation where no criterion is available, and nothing here shows that judges are invalid there - only that agreement cannot show they are valid.

The paper does not argue that human evaluation should be abandoned. Several of the cited results (Krumdick et al., 2025; Freitag et al., 2021; Chen, J., et al., 2026) show expert or carefully structured human judgment improving validation. The argument is that the human reference must be audited and its role stated, not that it is worthless.

Finally, the separation of V1, V2 and V3 is an analytic device. Kane (2013) would describe the same thing as different interpretation-and-use arguments, and Wallach

et al. (2025) frame evaluation more broadly as a social-science measurement problem, with a four-level framework drawn from that tradition. Bean et al. (2025), reviewing 445 benchmarks, reach related conclusions about construct validity in LLM evaluation generally. This paper contributes the application to judge validation specifically, the identification argument and the use-indexing of the criterion evidence; it does not claim the underlying measurement theory is new.

11. Conclusion

An agreement number tells us that a judge produces the labels a group of humans would have produced on items like those sampled. That is a useful thing to know when the goal is to replace those humans on those items. It is not evidence that the judge tracks the quality the labels were meant to capture, because a judge that shares the labelers' shortcuts and a judge that tracks quality can earn the same agreement, and it is not evidence that decisions made with the judge will turn out well, because many decisions are made on the close and hard items that agreement under-samples, and the inputs an optimiser will find are ones a natural validation set does not contain.

Where the field has found an external criterion, the criterion has repeatedly exposed failures that agreement did not show, and the criterion results have themselves turned out to belong to a use: a score that predicts best-of-N selection predicts PPO only roughly, and a reward model that teaches one policy well can fail to teach another. The practical consequence is a change of grammar. "This judge is validated" is not a well-formed claim. "This judge can substitute for these raters on this population," "this judge responds to this construct under these edits," and "this judge's verdicts improve this decision under this optimiser" are, and each needs its own evidence.

References

- Angelopoulos, A. N., Bates, S., Fannjiang, C., Jordan, M. I., & Zrnic, T. (2023). Prediction-Powered Inference. arXiv:2301.09633. doi:10.48550/arXiv.2301.09633. [PREPRINT]
- Bavaresco, A., Bernardi, R., Bertolazzi, L., Elliott, D., Fernandez, R., Gatt, A., Ghaleb, E., Giulianelli, M., Hanna, M., Koller, A., Martins, A. F. T., Mondorf, P., et al. (2024). LLMs instead of Human Judges? A Large Scale Empirical Study across 20 NLP Evaluation Tasks. ACL 2025. arXiv:2406.18403. doi:10.48550/arXiv.2406.18403.
- Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., Batzner, J., Foroutan, N., Schmitz, C., Korgul, K., Batra, H., Deb, O., Beharry, E., et al. (2025). Measuring what Matters: Construct Validity in Large Language Model Benchmarks. NeurIPS 2025 Datasets and Benchmarks. arXiv:2511.04703. doi:10.48550/arXiv.2511.04703.
- Bhandari, M., Gour, P., Ashfaq, A., Liu, P., & Neubig, G. (2020). Re-evaluating Evaluation in Text Summarization. EMNLP 2020. arXiv:2010.07100. doi:10.48550/arXiv.2010.07100.
- Bodhwani, U., Tran, T., & Wei, K. (2026). GAUGE: When Not to Trust LLM-as-a-Judge in User-Simulated Evaluation of Task-Oriented Agents. EMNLP 2026 Industry Track. arXiv:2609.12191. doi:10.48550/arXiv.2609.12191. [PREPRINT]
- Borsboom, D., Mellenbergh, G. J., & van Heerden, J. (2004). The Concept of Validity. *Psychological Review*, 111(4), 1061-1071. doi:10.1037/0033-295X.111.4.1061.
- Bowman, S. R., & Dahl, G. E. (2021). What Will it Take to Fix Benchmarking in Natural Language Understanding? NAACL 2021. arXiv:2104.02145. doi:10.48550/arXiv.2104.02145.
- Calderon, N., Reichart, R., & Dror, R. (2025). The Alternative Annotator Test for LLM-as-a-Judge: How to Statistically Justify Replacing Human Annotators with LLMs. arXiv:2501.10970. doi:10.48550/arXiv.2501.10970. [PREPRINT]
- Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2), 81-105. doi:10.1037/h0046016.
- Chen, G. H., Chen, S., Liu, Z., Jiang, F., & Wang, B. (2024). Humans or LLMs as the Judge? A Study on Judgement Biases. EMNLP 2024. arXiv:2402.10669. doi:10.48550/arXiv.2402.10669.
- Chen, J., Chen, W., Lin, Z., & Vong, C. M. (2026). A Judge Should Know What Changed: Construct Validity for LLM-as-a-Judge Evaluation. arXiv:2608.24419. doi:10.48550/arXiv.2608.24419. [PREPRINT]
- Chen, L., Liu, Q., Lin, W., & Liang, F. (2026). Criterion Validity of LLM-as-Judge for Business Outcomes in Conversational Commerce. arXiv:2604.00022. doi:10.48550/arXiv.2604.00022. [PREPRINT]
- Chiang, W.-L., Zheng, L., Sheng, Y., Angelopoulos, A. N., Li, T., Li, D., Zhang, H., Zhu, B., Jordan, M., Gonzalez, J. E., & Stoica, I. (2024). Chatbot Arena: An Open Platform for Evaluating LLMs by Human Preference. arXiv:2403.04132. doi:10.48550/arXiv.2403.04132.
- Choi, J., Park, S., Cho, C., Park, H., & Kim, B. (2026). Diagnosing the Reliability of LLM-as-a-Judge via Item Response Theory. ICML 2026. arXiv:2602.00521. doi:10.48550/arXiv.2602.00521.
- Clark, E., August, T., Serrano, S., Haduong, N., Gururangan, S., & Smith, N. A. (2021). All That's 'Human' Is Not Gold: Evaluating Human Evaluation of Generated Text. ACL 2021. arXiv:2107.00061. doi:10.48550/arXiv.2107.00061.
- Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1), 37-46. doi:10.1177/001316446002000104.
- Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52(4), 281-302. doi:10.1037/h0040957.
- Dev, S., Sloan, A., Kavner, J., Kong, N., & Sandler, M. (2026). Judge Reliability Harness: Stress Testing the Reliability of LLM Judges. ICLR 2026 Workshop. arXiv:2603.05399. doi:10.48550/arXiv.2603.05399. [PREPRINT]
- Dooskin, T., & Squoosh Technical Staff. (2026). The Judge Knows When It Knows: Calibrated Abstinence for LLM-Based A/B-Test Prediction. arXiv:2608.07517. doi:10.48550/arXiv.2608.07517. [PREPRINT]
- Dorner, F. E., Nastl, V. Y., & Hardt, M. (2024). Limits to scalable evaluation at the frontier: LLM as Judge won't beat twice the data. ICLR 2025. arXiv:2410.13341. doi:10.48550/arXiv.2410.13341.
- Dubois, Y., Galambosi, B., Liang, P., & Hashimoto, T. B. (2024). Length-Controlled AlpacaEval: A Simple Way to Debias Automatic Evaluators. COLM 2024. arXiv:2404.04475. doi:10.48550/arXiv.2404.04475.
- Dubois, Y., Li, X., Taori, R., Zhang, T., Gulrajani, I., Ba, J., Guestrin, C., Liang, P., & Hashimoto, T. B. (2023). AlpacaFarm: A Simulation Framework for Methods that Learn from Human Feedback. NeurIPS 2023. arXiv:2305.14387. doi:10.48550/arXiv.2305.14387.
- Fabbri, A. R., Kryscinski, W., McCann, B., Xiong, C., Socher, R., & Radev, D. (2020). SummEval: Re-evaluating Summarization Evaluation. TACL. arXiv:2007.12626. doi:10.48550/arXiv.2007.12626.

- Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low Kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6), 543-549. doi:10.1016/0895-4356(90)90158-L.
- Feuer, B., Goldblum, M., Datta, T., Nambiar, S., Besaleli, R., Dooley, S., Cembalest, M., & Dickerson, J. P. (2024). Style Outweighs Substance: Failure Modes of LLM Judges in Alignment Benchmarking. *ICLR 2025*. arXiv:2409.15268. doi:10.48550/arXiv.2409.15268.
- Freitag, M., Foster, G., Grangier, D., Ratnakar, V., Tan, Q., & Macherey, W. (2021). Experts, Errors, and Context: A Large-Scale Study of Human Evaluation for Machine Translation. *TACL*. arXiv:2104.14478. doi:10.48550/arXiv.2104.14478.
- Frick, E., Li, T., Chen, C., Chiang, W.-L., Angelopoulos, A. N., Jiao, J., Zhu, B., Gonzalez, J. E., & Stoica, I. (2024). How to Evaluate Reward Models for RLHF. arXiv:2410.14872. doi:10.48550/arXiv.2410.14872.
- Gao, L., Schulman, J., & Hilton, J. (2022). Scaling Laws for Reward Model Overoptimization. arXiv:2210.10760. doi:10.48550/arXiv.2210.10760.
- Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks. *Proceedings of the National Academy of Sciences*, 120(30). arXiv:2303.15056. doi:10.48550/arXiv.2303.15056.
- Glorigic, K., Zrnic, T., Lee, C., Candes, E. J., & Jurafsky, D. (2024). Can Unconfident LLM Annotations Be Used for Confident Conclusions? *NAACL 2025*. arXiv:2408.15204. doi:10.48550/arXiv.2408.15204.
- Gu, X., Jiang, X., Shi, Z., Tan, H., Zhai, X., Xu, C., Li, W., Shen, Y., Ma, S., Liu, H., Wang, S., Zhang, K., et al. (2024). A Survey on LLM-as-a-Judge. arXiv:2411.15594. doi:10.48550/arXiv.2411.15594. [PREPRINT]
- Guerdan, L., Barocas, S., Holstein, K., Wallach, H., Wu, Z. S., & Chouldechova, A. (2025). Validating LLM-as-a-Judge Systems under Rating Indeterminacy. *NeurIPS 2025*. arXiv:2503.05965. doi:10.48550/arXiv.2503.05965.
- Hosking, T., Blunsom, P., & Bartolo, M. (2023). Human Feedback is not Gold Standard. *ICLR 2024*. arXiv:2309.16349. doi:10.48550/arXiv.2309.16349.
- Hu, X., Gao, M., Hu, S., Zhang, Y., Chen, Y., Xu, T., & Wan, X. (2024). Are LLM-based Evaluators Confusing NLG Quality Criteria? *ACL 2024*. arXiv:2402.12055. doi:10.48550/arXiv.2402.12055.
- Jung, J., Brahman, F., & Choi, Y. (2024). Trust or Escalate: LLM Judges with Provable Guarantees for Human Agreement. arXiv:2407.18370. doi:10.48550/arXiv.2407.18370.
- Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement*, 50(1), 1-73. doi:10.1111/jedm.12000.
- Karpinska, M., Akoury, N., & Iyyer, M. (2021). The Perils of Using Mechanical Turk to Evaluate Open-Ended Text Generation. *EMNLP 2021*. arXiv:2109.06835. doi:10.48550/arXiv.2109.06835.
- Kim, E., Garg, A., Peng, K., & Garg, N. (2025). Correlated Errors in Large Language Models. *ICML 2025*. arXiv:2506.07962. doi:10.48550/arXiv.2506.07962.
- Krippendorff, K. (2004). Reliability in Content Analysis. *Human Communication Research*, 30(3), 411-433. doi:10.1111/j.1468-2958.2004.tb00738.x.
- Krumdick, M., Lovering, C., Reddy, V., Ebner, S., & Tanner, C. (2025). No Free Labels: Limitations of LLM-as-a-Judge Without Human Grounding. arXiv:2503.05061. doi:10.48550/arXiv.2503.05061. [PREPRINT]
- Lambert, N., Pyatkin, V., Morrison, J., Miranda, L. J., Lin, B. Y., Chandu, K., Dziri, N., Kumar, S., Zick, T., Choi, Y., Smith, N. A., & Hajishirzi, H. (2024). RewardBench: Evaluating Reward Models for Language Modeling. arXiv:2403.13787. doi:10.48550/arXiv.2403.13787.
- Landis, J. R., & Koch, G. G. (1977). The Measurement of Observer Agreement for Categorical Data. *Biometrics*, 33(1), 159-174. doi:10.2307/2529310.
- Li, D., Sun, R., Huang, Y., Zhong, M., Jiang, B., Han, J., Zhang, X., Wang, W., & Liu, H. (2025). Preference Leakage: A Contamination Problem in LLM-as-a-judge. *ICLR 2026*. arXiv:2502.01534. doi:10.48550/arXiv.2502.01534.
- Li, J., Zhou, F., Sun, S., Zhang, Y., Zhao, H., & Liu, P. (2024). Dissecting Human and LLM Preferences. arXiv:2402.11296. doi:10.48550/arXiv.2402.11296.
- Li, T., Chiang, W.-L., Frick, E., Dunlap, L., Wu, T., Zhu, B., Gonzalez, J. E., & Stoica, I. (2024). From Crowdsourced Data to High-Quality Benchmarks: Arena-Hard and BenchBuilder Pipeline. arXiv:2406.11939. doi:10.48550/arXiv.2406.11939.
- Liu, Y., Iyer, D., Xu, Y., Wang, S., Xu, R., & Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. *EMNLP 2023*. arXiv:2303.16634. doi:10.48550/arXiv.2303.16634.
- Liu, Y., Yao, Z., Min, R., Cao, Y., Hou, L., & Li, J. (2024). RM-Bench: Benchmarking Reward Models of Language Models with Subtlety and Style. arXiv:2410.16184. doi:10.48550/arXiv.2410.16184.
- Mahendrakar, P. (2026). A Second Model Is Not a Second Opinion: Peer Verification, Correlated Failure, and the Independence Variable Nobody Reports. *Life of Research*. doi:10.5281/zenodo.22044420.
- Malik, S., Pyatkin, V., Land, S., Morrison, J., Smith, N. A., Hajishirzi, H., & Lambert, N. (2025). RewardBench 2: Advancing Reward Model Evaluation. *ICLR 2026*. arXiv:2506.01937. doi:10.48550/arXiv.2506.01937.
- Manheim, D., & Garrabrant, S. (2018). Categorizing Variants of Goodhart's Law. arXiv:1803.04585. doi:10.48550/arXiv.1803.04585. [PREPRINT]
- Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. doi:10.1037/0003-066X.50.9.741.
- Norman, J. D., Rivera, M. U., & Hughes, D. A. (2026). Reliability without Validity: A Systematic, Large-Scale Evaluation of LLM-as-a-Judge Models Across Agreement, Consistency, and Bias. arXiv:2606.19544. doi:10.48550/arXiv.2606.19544. [PREPRINT]
- Panickssery, A., Bowman, S. R., & Feng, S. (2024). LLM Evaluators Recognize and Favor Their Own Generations. arXiv:2404.13076. doi:10.48550/arXiv.2404.13076.
- Plank, B. (2022). The 'Problem' of Human Label Variation: On Ground Truth in Data, Modeling and Evaluation. *EMNLP 2022*. arXiv:2211.02570. doi:10.48550/arXiv.2211.02570.
- Podsakoff, P. M., MacKenzie, S. B., Lee, J.-Y., & Podsakoff, N. P. (2003). Common method biases in behavioral research: A critical review of the literature and recommended remedies. *Journal of Applied Psychology*, 88(5), 879-903. doi:10.1037/0021-9010.88.5.879.
- Razin, N., Wang, Z., Strauss, H., Wei, S., Lee, J. D., & Arora, S. (2025). What Makes a Reward Model a Good Teacher? An Optimization Perspective. *NeurIPS 2025*. arXiv:2503.15477. doi:10.48550/arXiv.2503.15477.
- Saito, K., Wachi, A., Wataoka, K., & Akimoto, Y. (2023). Verbosity Bias in Preference Labeling by Large Language Models. arXiv:2310.10076. doi:10.48550/arXiv.2310.10076. [PREPRINT]
- Schroeder, K., & Wood-Doughty, Z. (2024). Can You Trust LLM Judgments? Reliability of LLM-as-a-Judge. arXiv:2412.12509. doi:10.48550/arXiv.2412.12509. [PREPRINT]
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askell, A., Bowman, S. R., Cheng, N., Durmus, E., Hatfield-Dodds, Z., Johnston, S. R., Kravec, S., Maxwell, T., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548. doi:10.48550/arXiv.2310.13548.
- Singh, S., Nan, Y., Wang, A., D'Souza, D., Kapoor, S., Ustun, A., Koyejo, S., Deng, Y., Longpre, S., Smith, N. A., Ermi, B., Fadaee,

- M., et al. (2025). The Leaderboard Illusion. arXiv:2504.20879. doi:10.48550/arXiv.2504.20879.
- Singhal, P., Goyal, T., Xu, J., & Durrett, G. (2023). A Long Way to Go: Investigating Length Correlations in RLHF. COLM 2024. arXiv:2310.03716. doi:10.48550/arXiv.2310.03716.
- Stureborg, R., Alikaniotis, D., & Suhara, Y. (2024). Large Language Models are Inconsistent and Biased Evaluators. arXiv:2405.01724. doi:10.48550/arXiv.2405.01724. [PREPRINT]
- Szymanski, A., Ziems, N., Eicher-Miller, H. A., Li, T. J.-J., Jiang, M., & Metoyer, R. A. (2024). Limitations of the LLM-as-a-Judge Approach for Evaluating LLM Outputs in Expert Knowledge Tasks. arXiv:2410.20266. doi:10.48550/arXiv.2410.20266.
- Tan, S., Zhuang, S., Montgomery, K., Tang, W. Y., Cuadron, A., Wang, C., Popa, R. A., & Stoica, I. (2024). JudgeBench: A Benchmark for Evaluating LLM-based Judges. ICLR 2025. arXiv:2410.12784. doi:10.48550/arXiv.2410.12784.
- Thakur, A. S., Choudhary, K., Ramayapally, V. S., Vaidyanathan, S., & Hupkes, D. (2024). Judging the Judges: Evaluating Alignment and Vulnerabilities in LLMs-as-Judges. GEM 2025. arXiv:2406.12624. doi:10.48550/arXiv.2406.12624.
- Wallach, H., Desai, M., Cooper, A. F., Wang, A., Atalla, C., Barocas, S., Blodgett, S. L., Chouldechova, A., Corvi, E., Dow, P. A., Garcia-Gathright, J., Olteanu, A., et al. (2025). Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. ICML 2025. arXiv:2502.00561. doi:10.48550/arXiv.2502.00561.
- Wang, P., Li, L., Chen, L., Cai, Z., Zhu, D., Lin, B., Cao, Y., Liu, Q., Liu, T., & Sui, Z. (2023). Large Language Models are not Fair Evaluators. arXiv:2305.17926. doi:10.48550/arXiv.2305.17926.
- Wen, J., Zhong, R., Khan, A., Perez, E., Steinhardt, J., Huang, M., Bowman, S. R., He, H., & Feng, S. (2024). Language Models Learn to Mislead Humans via RLHF. arXiv:2409.12822. doi:10.48550/arXiv.2409.12822.
- Wen, X., Lou, J., Lu, Y., Lin, H., Yu, X., Lu, X., He, B., Han, X., Zhang, D., & Sun, L. (2024). Rethinking Reward Model Evaluation: Are We Barking up the Wrong Tree? ICLR 2025. arXiv:2410.05584. doi:10.48550/arXiv.2410.05584.
- Wu, M., & Aji, A. F. (2023). Style Over Substance: Evaluation Biases for Large Language Models. arXiv:2307.03025. doi:10.48550/arXiv.2307.03025. [PREPRINT]
- Yang, J., Feldhus, N., Mohtaj, S., Hennig, L., Wang, Q., Metheniti, E., Hakimov, S., Jakob, C., Solopova, V., Rieck, K., Schlangen, D., Moller, S., et al. (2026). What Are We Measuring in NLG? A Meta-Analysis of Evaluation Trends 2020-2025. arXiv:2601.07648. doi:10.48550/arXiv.2601.07648. [PREPRINT]
- Ye, J., Wang, Y., Huang, Y., Chen, D., Zhang, Q., Moniz, N., Gao, T., Geyer, W., Huang, C., Chen, P.-Y., Chawla, N. V., & Zhang, X. (2024). Justice or Prejudice? Quantifying Biases in LLM-as-a-Judge. arXiv:2410.02736. doi:10.48550/arXiv.2410.02736.
- Zhang, X., Xiong, W., Chen, L., Zhou, T., Huang, H., & Zhang, T. (2024). From Lists to Emojis: How Format Bias Affects Model Alignment. arXiv:2409.11704. doi:10.48550/arXiv.2409.11704. [PREPRINT]
- Zhao, Y., Liu, H., Yu, D., Kung, S., Chen, M., Mi, H., & Yu, D. (2025). One Token to Fool LLM-as-a-Judge. arXiv:2507.08794. doi:10.48550/arXiv.2507.08794.
- Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., Zhang, H., Gonzalez, J. E., & Stoica, I. (2023). Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. NeurIPS 2023 Datasets and Benchmarks. arXiv:2306.05685. doi:10.48550/arXiv.2306.05685.
- Zheng, X., Pang, T., Du, C., Liu, Q., Jiang, J., & Lin, M. (2024). Cheating Automatic LLM Benchmarks: Null Models Achieve High Win Rates. ICLR 2025. arXiv:2410.07137. doi:10.48550/arXiv.2410.07137.