AND A GOOD JUDGE TOO?: METHODOLOGICAL LESSONS FOR AUDITING LLM JUDGES IN RUBRIC-BASED EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges have become core infrastructure in automated AI evaluation, yet their reliability remains underexamined. We audit and extend L2-Bench, a rubric-based benchmark for second language learning design. This is a "hard" domain marked by expert disagreement, where ground truth is contested and LLM judges are most likely to be deployed. Guided by a recent judge-failure taxonomy, we show that L2-Bench's production judge (Claude Sonnet 4.6) is statistically stable, but most importantly we demonstrate that stability alone cannot establish validity because of rater label skew—a problem common to evaluation in "hard" domains. To anchor judge scores to human expertise, we import rater-severity modeling from educational measurement, with a GLMM attributing 10.1% of latent-scale variance to rater severity. Scored as a held-out rater against severity-corrected panels, the judge is consistent with convergent validity ($\kappa = 0.861$ vs. 0.788 human ceiling, overlapping 95% CIs). We argue that valid judge claims in hard domains require both kinds of evidence, and propose the Minimum Viable Judge Assessment (MVJA), a tiered reporting template for standardizing judge analysis across the field.

1 INTRODUCTION

LLMs are increasingly becoming essential components of evaluation pipelines. This is because they purportedly help to scale automated evaluation benchmarks into open-ended domains where fixed answer keys do not provide sufficient signal quality (Zheng et al., 2023; Gu et al., 2024). The need is greatest in what we call "hard" contexts—domains in which experts broadly agree about what judgment criteria should exist, but do not often agree about how to apply those criteria in practice. Evaluations in these domains increasingly rely on rubric-based scoring, and the prevailing design is to employ an LLM judge instead of painstakingly gathering human ratings for each evaluation task every time the evaluation is run.

However, many in the evaluations community caution against the use of LLM judges and are increasingly urging the field to implement judges more intentionally (Thakur et al., 2024; Zhao et al., 2026). A growing literature documents judge failure modes, adding evidence to these concerns, and another stream questions the quality of judge-human agreement in these evaluative tasks (Elangovan et al., 2025; Krumdick et al., 2025). These are pressing questions for the field.

In order to improve the practice of judge integration in evaluation pipelines—and specifically to assess the tractability of improving the measurement validity of existing evaluations that rely upon LLM judges—we audit L2-Bench (Anonymous, 2026). L2-Bench is an ideal candidate: a recent rubric-based evaluation benchmark designed to assess model performance in a "hard" context and one that includes preliminary judge assessment already.

Our **main finding** is that passing the full battery of tests suggested by emerging consensus on judge analysis in the AI evaluations community is insufficient to establish judge convergent validity—the type of validity that matters most for LLM judges in rubric-based evaluation in "hard" contexts. In L2-Bench's case, the production judge's evidence is ultimately consistent with convergent validity, but demonstrating this, and then interpreting it correctly, required methods and techniques the AI evaluations and auditing field has not yet adopted. Finally, we argue that establishing a judge's

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

convergent validity is necessary, but not sufficient, for strong claims about the overall measurement validity of an evaluation design that employs it. We make the following contributions:

1) An empirical contribution: We extend L2-Bench's baseline judge analysis with 32 judge configurations (nine open and closed models from five families, three prompt variants) and a ten-resample stability test over 44 configurations. By these tests, which our rapid review would rate an 'extensive' validation, the production judge is stable and shows no detectable model-family self-preference. We also identify alternatives, including the open-weight Qwen 3.8 Max and GLM-5.2, that match or exceed it on apparent human alignment.

2) A methodological contribution: We import the rater-mediated assessment tradition from educational measurement, which has long modeled rater severity as a distinct facet of scoring (Linacre, 1989; McNamara, 1996), to LLM evaluation, where it has been generally overlooked. Using L2-Bench's human rater data, we show that aggregate judge–human agreement is uninformative under the benchmark's label skew, and that genuinely contested criteria carry the most informative agreement signal. A mixed-effects model's severity-corrected leave-one-out ceiling shows that about a third of the judge's apparent advantage over experts reflects human threshold inconsistency, not superior perception. Against this ceiling, the production judge demonstrates convergent validity ($\kappa = 0.861$ vs. 0.788, overlapping 95% CIs). We argue this method is generalizable.

3) A technical documentation and governance contribution: Arguing that 'extensive' judge analysis should become the standard, we propose a way to standardize its reporting in scientific literature and technical documentation: the "Minimum Viable Judge Assessment" (MVJA) report.

2 Assessing Judges: The State of the Field

LLM-as-a-judge emerged as an evaluation technique after Zheng et al. (2023) showed that GPT-4's judgements agreed with human preference ratings roughly 80% of the time, and judges then proliferated rapidly in evaluation design (Gu et al., 2024; Li et al., 2024a;b). Yet as the typology of judge failure modes first outlined by Zheng et al. (2023) expands (Zheng et al., 2023; Shi et al., 2025; Tripathi et al., 2025; Dubois et al., 2024; Saito et al., 2023; Panickssery et al., 2024; Wataoka et al., 2024; Chen et al., 2025; Tan et al., 2025; Zeng et al., 2024; Thakur et al., 2024; Han et al., 2025; Yamauchi et al., 2026), concerns about judge stability have spread across the evaluation ecosystem. Zhao et al. (2026) show that even state-of-the-art models remain susceptible to verdict flipping, like their predecessors.

The field now also assesses the downstream effects of judge failures (Ye et al., 2024; Koo et al., 2024), e.g., in reward modeling (Lambert et al., 2024; Malik et al., 2026). Many urge more sophisticated statistical approaches to judge validity (Choi et al., 2026; Laban et al., 2023; Sharma et al., 2025; Agarwal & Khanna, 2025) and better disclosure of where judges are more or less reliable (Sushmita et al., 2026). Research has also returned to the root question of what judge "agreement" with human raters entails. Elangovan et al. (2025) show that aggregate rating correlation can mask reliability, especially when human raters disagree, and Krumdick et al. (2025) show that in expert domains judges agree with humans mainly on questions the judge can itself answer. This literature has drawn little on the rater-mediated assessment tradition in educational measurement, which has studied rater severity and what inter-rater agreement actually demonstrates for decades (Linacre, 1989; McNamara, 1996; Myford & Wolfe, 2003).

Has practice improved? Our rapid review of 37 research articles proposing evaluations with LLM judge scoring (methodology in Appendix A) suggests much work remains. Only 5 (13.5%) provided extensive judge analysis (a dedicated section or subsection plus appendix material), while 15 (40.5%) provided minimal (under one paragraph) or none. Though a crude heuristic, this suggests the will for more extensive judge analysis exists but structural and methodological challenges may be obstructing it. A more formal systematic meta-analysis would be beneficial.

3 Auditing a production judge

To test what an extensive judge analysis can and cannot show, we audit a recent, openly released rubric-based benchmark that uses an LLM judge as its scorer. The audit has three phases. First, we perform what our rapid review would classify as an "extensive" judge analysis (Section 3.1). Second,

we examine why that analysis falls short of establishing convergent validity (Sections 3.2–3.3). Third, we reconstruct rater ground truth to test convergent validity directly (Sections 3.4–3.5). Every judge experiment is described in full, in the order it was run, in Appendix D.

3.1 Replicating the state of the art

In this phase we use conventional techniques common in "extensive" judge analysis. Our case study is L2-Bench (Anonymous, 2026), chosen in part for its extensive expert validation dataset. L2-Bench measures LLM capabilities in second language education. An LLM judge scores single-turn task–response pairs against a task rubric, one criterion at a time. At the time of writing it is one of the most extensively validated AIED benchmarks: 221 practitioners from 45 countries and six stakeholder groups rated a stratified random sample of its tasks, mapped to their expertise, in a three-stage workflow (Appendix C). We restrict our analysis to the $N = 142$ practitioners who reached Stage C, in which they scored a revealed "AI response" against the task rubric; we refer to them as "experts" henceforth. These experts rated in line with the full cohort: task authenticity $M = 4.43$ and criteria adequacy $M = 4.22$ (Stage A), and no preference between the expert-written reference answer and a Claude Sonnet 4.6 response in the Stage B blind A/B comparison (reference chosen 50.3% of the time). In Stage C they contributed 17,093 criterion ratings across 442 items, split roughly evenly between the real Sonnet 4.6 response and a reference answer disguised as the "AI response" (Table 5).

L2-Bench's baseline analysis (Appendix D.1) scored three judge models on 48 tasks. We extend it to nine judge models from five families, crossed with three prompt variants that differ in whether the reference answer is shown and whether the judge reasons before its verdict (32 judge configurations; Table 7, prompts in Appendix D.4). We also add a Naive baseline that passes every positively weighted criterion without reading the response. All alignment responses were generated by Claude Sonnet 4.6, so the production judge scores its own model's output. On each axis below, the audited production judge (Sonnet 4.6 with the reference-guided direct-verdict prompt, v1) performs well:

Judge–human alignment (apparent). Each configuration re-scores the same Sonnet 4.6 responses on 98 tasks (2× the baseline sample). We compare its verdicts with the expert majority vote wherever one exists. This yields $n = 650$ gold (task, criterion) verdicts over 41 tasks. The production judge agrees with the majority on 90.9% of criteria ($\kappa = 0.651$, F1 $= 0.946$). It ranks 4th of 32 configurations and 3rd by model family, behind Qwen 3.8 Max (best $\kappa = 0.742$, prompt v3) and GLM-5.2 ($\kappa = 0.691$, v3; Table 1, full leaderboard in Appendix D.5). The Naive baseline beats every judge ($\kappa = 0.842$), for reasons we examine in Section 3.3.

Performance on reference answers. Judges re-score the expert-written reference answers, where near-universal passes are expected ($n = 1{,}863$ criteria on 98 tasks). The two configurations run on all 98 tasks lead: GLM-5.2 | v3 (accuracy 0.916, $\kappa = 0.685$) and Sonnet 4.6 | v3 (0.903, $\kappa = 0.648$; Appendix D.6).

Stability over resampling. We re-score one fixed response (by Claude Haiku 4.5) on each of 24 tasks (381 criteria; +19 tasks vs. baseline) ten times (+7 resamples vs. baseline) under 44 configurations, adding prompt v4 (see Appendix Table 6). The production configuration flips 2.9% of criteria (Wilson 95% CI $[1.6, 5.1]$); three configurations flip fewer. Pooled over all four prompts, 93.2% of Sonnet 4.6 criteria are identical across all ten resamples, third of ten judge models behind Qwen 3.8 27B (96.5%) and Qwen 3.8 Max (93.4%); the least stable is Kimi K2.5 (65.4%). Across configurations, 59.5% of the 2,330 flips move toward Fail, so what instability there is appears as added severity rather than leniency (Appendix D.7).

Self-preference. Because the production judge scores its own model's output, we test for self-preference in two ways. First, on identical text, judges cluster by prompt rather than by model family: the four v3 configurations from three families group together, and the Sonnet 4.6 configurations split across clusters (Appendix D.8). Second, relative to human gold, Claude judges are *harsher* on the Sonnet 4.6 response than on the reference answer, compared with non-Claude judges (difference-in-differences -3.5 pp, 95% CI $[-7.4, -0.1]$). We find no self-preference; if anything, the CI excludes zero in the opposite direction. The evidence is weak, however, because a single model generated every response.

Prompt and score sensitivity. Is "best judge" a property of the model or of the prompt? Across the eight model families run under all three prompts, the production family ranks 2nd (v1), 1st (v2),

Table 1: Judge–human alignment on the $n = 650$ gold verdicts (41 tasks), sorted by Cohen's κ; selected rows of the 32-configuration leaderboard (Appendix D.5). "Pass" = share of criteria the judge passes; the gold pass rate is 86.9%. The Naive baseline (Pass iff criterion weight > 0) does not read response or reference.

Judge configuration	**Accuracy**	**F1**	κ	**Pass**
Naive baseline (weight > 0)	0.963	0.979	0.842	0.860
Qwen 3.8 Max | v3	0.935	0.962	0.742	0.838
GLM-5.2 | v3	0.926	0.957	0.691	0.854
Qwen 3.8 Max | v1	0.928	0.958	0.690	0.862
Sonnet 4.6 | v1 (production)	**0.909**	**0.946**	**0.651**	**0.825**
DeepSeek V3.2 | v1	0.895	0.937	0.638	0.786
Sonnet 4.6 | v3	0.897	0.938	0.630	0.800
GPT-5.6-Terra | v1 (best sampling point)	0.888	0.932	0.606	0.791
Qwen 3.5 9B | v1	0.894	0.942	0.347	0.957
GLM-5.2 | v2 (lowest κ)	0.832	0.906	0.137	0.914

and 4th (v3). GLM-5.2 ranks 4th, 8th, and 2nd. Pairwise Kendall's τ between prompt variants is -0.07–0.43. Judge choice also moves the benchmark score itself: the same Sonnet 4.6 responses score 73.3–88.9% depending on the judge (production: 84.6%; Appendix D.9). "Best judge" is therefore partly a property of the prompt, and should be reported per prompt variant.

3.2 Judge stability is not convergent validity

By the standards the field currently uses for an extensive judge analysis, this is a validated judge. But is statistical stability all that is required? Despite prevailing assumptions, stability alone can be a mirage. In a large-scale evaluation of 21 judge models, Norman et al. (2026) find that judges with high test-retest reliability can simultaneously exhibit severe position bias. Reliability, in short, does not imply validity: it is of little use if it does not converge on the proper target, which is expert human judgment. In this second phase of our audit, we therefore extend L2-Bench's judge analysis to explore this question in greater depth.

The question matters most in rubric-based evaluation, which many evaluations in "hard" contexts use, L2-Bench among them. As originally reported, L2-Bench's experts assign expert-written reference answers and model responses statistically indistinguishable rubric scores. We re-derive the weighted task score from expert judgments: it averages 0.851 on the reference answers ($n = 512$ rater–item pairs) and 0.853 on the AI responses ($n = 505$). The two criterion pass rates are also nearly identical (80.9% compared with 81.1%). In blind A/B testing, the $N = 221$ practitioners preferred the expert-written reference only 51.3% of the time. This raises two questions. Can the rubric itself support the discrimination that LLM judge scoring requires? Or do measures of convergence break down in "hard" contexts? To assess each, we must reconsider the proper convergence target and reconstruct the human ratings themselves.

3.3 Reasoning about target selection in LLM judge analysis

Model evaluation asks which model performs *best*, usually against a human baseline. Judge evaluation asks which judge performs *most like* the human gold standard, whether well or poorly. A judge that "outperforms" its human reference may be miscalibrated, not superhuman. Measuring this is another social-science measurement challenge (Wallach et al., 2025). It is especially difficult where experts largely agree on which criteria should exist but much less on how to apply them. In second-language performance assessment in particular, rater variability has long been treated not as noise to eliminate but as an intrinsic feature of the measurement that must itself be modeled (McNamara, 1996). As might be expected, agreement metrics behave strangely in "hard" contexts.

We hypothesized that the root cause of this difficulty is the label distribution itself. In our audit sample, individual experts vote Pass on 1,269 of 1,481 criteria (85.7%). Among the gold labels (Section 3.4), 565 of 650 are Pass (86.9%). This skew decouples raw agreement from chance-corrected agreement (κ). A leave-one-out (LOO) comparison shows this directly: we hold out each expert's vote and

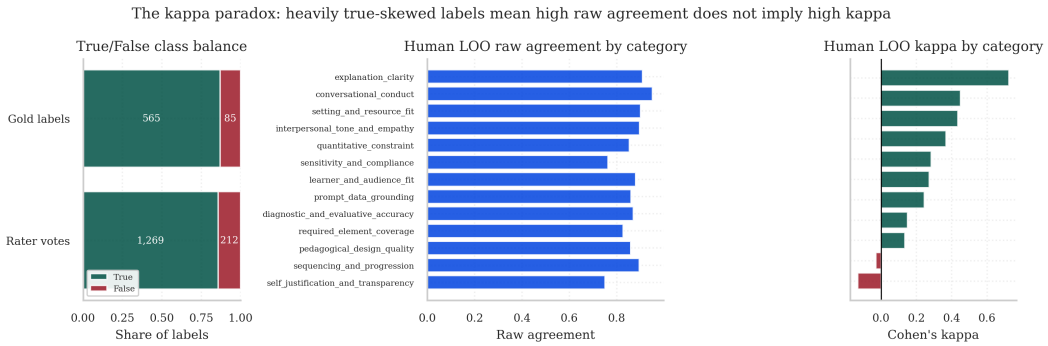

Figure 1: The κ paradox in the human rater data. Left: the gold majority labels (565/650 Pass, 86.9%) and the underlying individual rater votes (1,269/1,481 Pass, 85.7%) are both heavily skewed toward Pass. Centre and right: across the thirteen criterion categories, raw human leave-one-out agreement (centre) is uniformly high, while the corresponding chance-corrected κ (right) ranges from strongly positive (setting and resource fit, 0.724) to below zero (self-justification and transparency, -0.133) or undefined (explanation clarity, where every subpanel majority is Pass). The categories where nearly every vote is Pass are precisely those where κ collapses.

compare it with the majority of the remaining experts on the same unit (Appendix G). On *sequencing and progression* criteria (the order and staging of activities), a held-out expert agrees with the rest of the panel 89.4% of the time, yet $\kappa = -0.029$. On *setting and resource fit* criteria (fit to the teaching environment), agreement is slightly *higher* (90.0%) and κ is 0.724. The difference is that setting-fit criteria have a genuine Fail rate: 30% of gold labels are Fail, so there is something to disagree about. Sequencing criteria have none: every gold label is Pass, so the agreement that exists is agreement by default. On *explanation clarity* (90.9% agreement), the minority Fail votes never form a subpanel majority, so κ is undefined. κ depends on how many genuinely disputed units a stratum contains, so it cannot be compared across strata or read as an absolute level (Figure 1). This is a κ paradox (Feinstein & Cicchetti, 1990; Viera & Garrett, 2005).

The same skew makes aggregate judge–human agreement falsely reassuring, a pattern Norman et al. (2026) also observe across benchmarks. The Naive baseline tops the leaderboard (Table 1) because 76 of the 85 gold Fails are penalty criteria whose behavior did not occur, so the correct verdict is predictable from the rubric alone. Here κ does its job. Qwen 3.5 9B | v1 passes 95.7% of criteria and reaches 89.4% accuracy, close to production, but its κ of 0.347 is about half of production's. Accuracy and prevalence-adjusted agreement measures (PABAK, Gwet's AC1) largely hide this gap: the AC1 gap is 0.004 against a κ gap of 0.304 (Appendix E).

A judge that agrees with human rating may therefore still be unable to discriminate appropriate from inappropriate applications of a rubric. This is not because humans cannot discriminate, but because the label distribution leaves little room to show it. We conjecture that the problem *intensifies as model capabilities improve*: as pass rates rise, the pool of informative Fails shrinks, and κ intervals widen for a fixed annotation budget. This is a reasoned expectation that we have not tested.

Qualitative analysis complicates the picture further. Experts left comments on 387 of 1,447 Stage B rater-items (26.7%, from 103 raters). Even where raters converge on a judgment, their remarks often reveal that agreement is thinner than their vote implies. On tasks where every rater preferred the same response, comments still contain reservations. One rater (task 140) prefers Response A "in general" but notes "it would be even better with the 'Handling the Common Problems' of answer B." Another describes their vote in what became a unanimous preference (task 339) as a "close call… Response B is *slightly* more culturally specific." Others pass a response while critiquing it ("organization could be more detailed", task 433), or note that an activity should be timed at "35 minutes, not 20", as in the reference answer (task 463).

Returning to the task ratings with these comments in mind, we find structured rating variance, as L2-Bench's mixed-effects analysis reported (we replicate and extend that analysis for binary votes in Appendix F). The weighted task score is the passed positive weight divided by the total positive

weight, less any triggered penalties and floored at zero (Eq. 5). On task 62, every rater preferred the reference answer in Stage B, yet the four raters who scored the Sonnet 4.6 response in Stage C gave it 0.21, 0.48, 0.82, and 0.94. Eight items share this pattern: every rater preferred the same response, at least two raters scored the same text, and those scores span 0.4 or more. Apparent consensus can conceal substantive disagreement. What, then, is an LLM judge auditor to do?

3.4 No agreement without disagreement

Rather than look to agreement for a way out, we reason that the better (indeed, the only) place to look is disagreement, because that is where human agreement is actually informative. Doing so requires reconstructing gold human ratings. Throughout, a *unit* is one (task, criterion, scored text) triple, and a unit's *gold* label is its expert majority vote where at least two experts rated it and $\geq$67% agree.

Finding gold. Like all gold, L2-Bench's gold labels require careful extraction. Recall the 650 gold units against which every judge was scored in Section 3.1. Only 10 of them (1.5%) carry any dissenting vote, and these same 10 are contested (human pass rate between 0.2 and 0.8). This is by construction: the $\geq$67% agreement rule removes most disagreement before any judge is scored. The *raw panel* (every individual expert vote on a unit, before any majority is taken) tells a different story. Of the 1,232 units scored by three or more experts, 832 (67.5%) are unanimous, 376 (30.5%) are split, and 24 (1.9%) are tied. The split region is where the overall base-rate skew does not predict rater votes. It is the seam in which κ can properly be computed.

Verifying gold. To test the seam, we must check whether disagreement there reflects a genuine difference of opinion about the response *itself*, rather than differences in where each rater sets the pass threshold. Modeling rater severity as a distinct facet of scoring is long established in performance assessment (Linacre, 1989). A logistic GLMM with crossed random intercepts is the random-effects formulation of this many-facet Rasch model (De Boeck & Wilson, 2004). Following (Muradoglu et al., 2023), we fit one to all 17,093 criterion votes (142 raters, 442 tasks, 10,826 units). It has fixed effects for the scored text (AI response or reference answer) and for whether the rater scored a response other than the one they preferred in Stage B (Eq. 1). Rater identity accounts for 10.1% of latent-scale variance, the unit for 46.3%, and the task for 0.2%. Raters differ systematically in strictness at the level of individual criteria, not tasks. This is the binary-vote counterpart of the 27–33% rater share L2-Bench reported for its Likert-scale ratings: smaller, but clearly present. Rater offsets span -2.54 to $+2.40$ logits. A response that an average rater would pass with probability 0.5 would be passed with probability 0.07 by the strictest rater and 0.92 by the most lenient. We remove each rater's offset with a deliberately conservative rule: a vote flips only if, with the offset removed, it would clearly land on the other side of the pass line (Appendix F). The rule flips 236 votes (1.4%). The result is a severity-corrected consensus, which we maintain is a truer gold set: genuine disagreement, stripped of most thresholding noise.

Human ceiling. Following the alternative annotator test (Calderon et al., 2025) and automated essay scoring practice (Williamson et al., 2012), we score the judge as a held-out rater against the human leave-one-out ceiling, adding severity correction on both sides.(Section 3.3; Eq. 9). Across all 1,232 units with at least three experts (Frame A; 3,527 comparisons), a held-out expert agrees with the rest of the panel 86.4% of the time ($\kappa = 0.490$, CI $[0.428, 0.537]$), rising to 89.3% ($\kappa = 0.559$, CI $[0.491, 0.609]$) on severity-corrected votes. We then restrict to the units a judge can be compared on: votes on the Sonnet 4.6 response in audited tasks, with at least three experts per unit (Frame B; 169 units, 502 comparisons). Here a held-out expert agrees 90.2% of the time ($\kappa = 0.716$, CI $[0.590, 0.815]$), and 93.1% after severity correction ($\kappa = 0.788$, CI $[0.690, 0.862]$). This is L2-Bench's empirical ceiling for human agreement.

Where experts disagree. If disagreement were uniform, the ceiling would be noise. It is not uniform. Human LOO κ ranges from -0.133 for self-justification and transparency to 0.724 for setting and resource fit (Table 2). Criterion type alone does not explain the spread: required element coverage, a checklist-style category, has one of the lowest κ values (0.150). The seam is identified.

3.5 Findings

Leaderboard stability. Removing the 10 contested gold units leaves the judge leaderboard essentially unchanged (Spearman $\rho = 0.9989$ across all 32 configurations; no configuration moves more than

Table 2: The thirteen criterion categories (Appendix E). Share: percentage of the 16,695 rubric criteria in the benchmark. Gold: number of the 650 gold units, and their gold pass rate. Human LOO: Frame A agreement and κ with unit-clustered bootstrap 95% CIs (n = units of comparison). Judge: production judge disagreement rate against gold, and its Frame B judge-as-rater κ with unit count; – = undefined or not estimable.

		Gold		Human LOO (Frame A)			Judge	
Category	**Share**	n	**Pass**	n	**Agr.**	κ **[95% CI]**	**Disagr.**	κ (n)
Learner and audience fit	25.2	175	0.943	333	0.879	0.284 [0.115, 0.410]	0.034	0.897 (46)
Pedagogical design quality	13.9	78	0.987	183	0.859	0.134 [−0.051, 0.333]	0.077	– (16)
Required element coverage	13.6	89	0.966	177	0.827	0.150 [−0.026, 0.323]	0.101	0.594 (31)
Setting and resource fit	10.7	106	0.698	158	0.900	0.724 [0.613, 0.800]	0.075	0.906 (29)
Diagnostic and evaluative accuracy	8.7	63	0.984	74	0.870	0.245 [−0.034, 0.501]	0.079	– (2)
Sensitivity and compliance	7.5	37	0.108	93	0.762	0.368 [0.170, 0.579]	0.297	0.682 (13)
Sequencing and progression	5.4	27	1.000	59	0.894	−0.029 [−0.066, −0.010]	0.000	– (5)
Conversational conduct	3.6	14	0.929	22	0.950	–	0.071	– (6)
Prompt–data grounding	2.5	18	0.889	28	0.860	0.272 [−0.094, 0.634]	0.278	1.000 (5)
Quantitative constraint	2.3	14	0.929	29	0.853	0.435 [−0.094, 0.693]	0.214	1.000 (5)
Self-justification and transparency	2.3	9	0.889	13	0.750	−0.133 [−0.304, −0.042]	0.556	– (1)
Explanation clarity	2.2	13	1.000	43	0.909	–	0.000	– (7)
Interpersonal tone and empathy	2.1	7	1.000	20	0.897	0.449 [0.295, 0.716]	0.000	– (3)

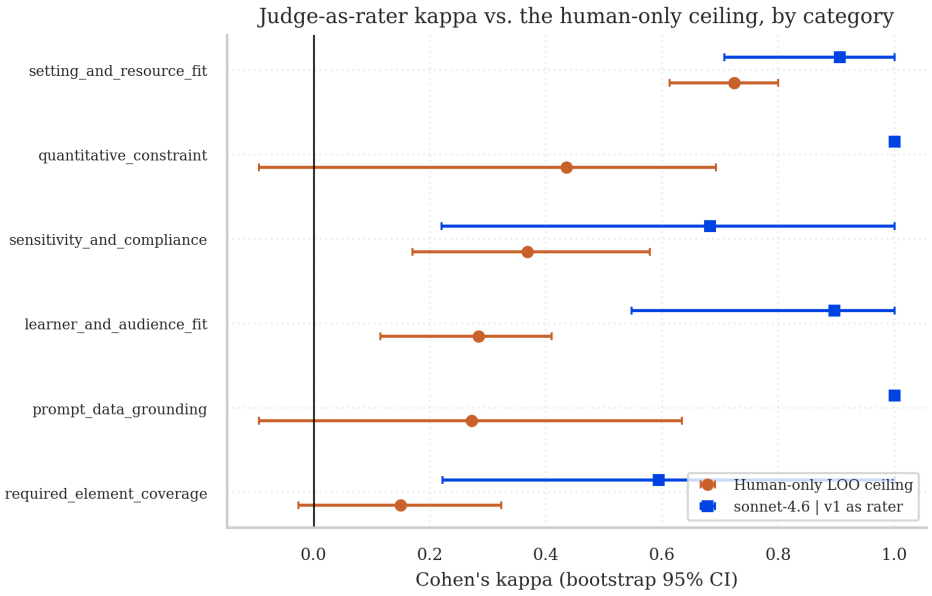

Figure 2: Judge-as-rater leave-one-out κ (Frame B: 169 units, 502 comparisons) versus the human-only ceiling (Table 2), by criterion category, with unit-clustered bootstrap 95% CIs. The production judge (Sonnet 4.6 | v1) matches or exceeds the human ceiling in all six estimable categories; per-category judge samples are small (5–46 units).

two ranks). This is a weak test, because the gold filter had already removed most disagreement (Appendix G.1), but a welcome sanity check.

Judge comparability on raw agreement. The LOO ceiling extends naturally to the judge: we score it exactly as we score a held-out expert, comparing its verdict with the majority of each $k-1$ expert subpanel on Frame B. The production judge agrees 94.4% of the time ($\kappa = 0.826$, CI $[0.708, 0.939]$), *above* the human ceiling of 90.2% ($\kappa = 0.716$) in point estimate. Swapping the judge in for a random expert makes the panel slightly *more* internally consistent. Its point estimate meets or exceeds the human ceiling in all six categories where both are estimable, though the intervals are wide and overlap (Figure 2; full-panel variant in Appendix G).

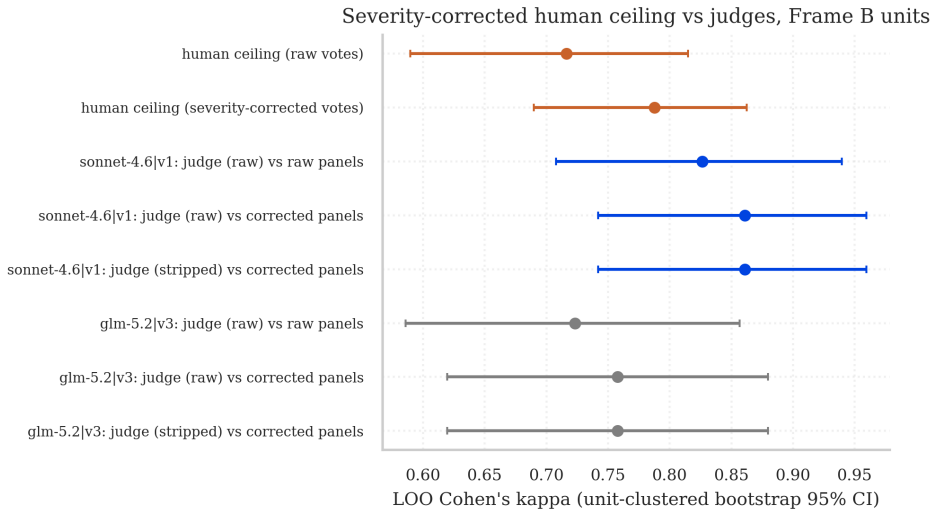

Figure 3: Severity-corrected convergence on Frame B (169 units). Leave-one-out κ with unit-clustered bootstrap 95% CIs for the human ceiling on raw and severity-corrected votes (orange), and for each judge scored against raw and severity-corrected human panels, before and after removing the judge's own severity offset (production judge in blue, GLM-5.2 | v3 in grey).

Judge comparability on severity-corrected agreement. If a good judge should match human performance rather than exceed it, why does L2-Bench's judge appear to beat the ceiling? The severity model answers this. A single judge applies one consistent threshold; a human panel has as many thresholds as it has raters. We therefore re-run the judge-as-rater analysis with severity removed from both sides. The judge needs no correction in practice: its estimated offset (-0.183 logits) is too small for the rule to flip any verdict (Appendix F). The human ceiling, by contrast, rises to $\kappa = 0.788$, and the production judge to 0.861 (CI $[0.742, 0.959]$). The gap narrows from 0.110 to 0.073. About a third of the judge's apparent advantage was human threshold inconsistency, not superior perception of the criteria (Figure 3). GLM-5.2 | v3 lands at 0.758, just below the corrected ceiling, with overlapping intervals.

Altogether, these results establish that L2-Bench's production judge is consistent with convergent validity with expert consensus. It agrees with the severity-corrected panel as closely as an individual expert does (monotrait–heteromethod agreement; Campbell & Fiske, 1959). At the level of criterion categories, it also disagrees where humans disagree (structural validity; Messick, 1995): judge and human disagreement rates are rank-correlated across the thirteen categories (Spearman $\rho = 0.834$, $p < 0.001$). Any further improvements in the benchmark as a measurement instrument must therefore come from the dataset and rubric themselves.

4 Discussion: Transferability and Recommendations

In "hard" contexts, no signal is trustworthy enough on its own. What we have provided here is an illustration of how constructive realism as an evaluation paradigm can help triangulate those signals. As Messick (1989) maintains, constructs are built, often by communities of experts over years of refinement and enforcement, to represent real social facts (Searle, 1995). Although the community itself can be wrong in applying its own construct, community consensus remains our best available estimate as to the construct's proper application. Assessing an LLM judge's performance on an evaluation benchmark in this space must therefore (1) demonstrate its ability to behave as a representative member of that community, not as a wise, all-seeing oracle, and (2) characterize how it errs relative to that community's own errors.

This observation recalls several recent critiques of AI evaluation from the AI in education subfield. Tuomi (2025) has generally called for a "scientific realist" turn in AIED evaluation. Likewise, Thomas et al. (2026) and He et al. (2026a) independently argue that inter-rater reliability is best understood

as a diagnostic tool that helps locate areas of disagreement—the same areas we show here provide informative signal about judge performance relative to human consensus. Yet it is precisely this area of disagreement that much prevailing analysis of rater agreement, and by extension evaluation design and judge analysis, discards as noise. In "hard" contexts, judge validity cannot be established without this information.

We expect these lessons to generalize beyond L2-Bench (see B. Our audit demonstrates that a new standard in judge analysis is tractable. The methods we apply, some well known and others imported from different measurement traditions, can improve the state of judge assessment and often without collecting new data—so long as evaluation developers release the rater-level data they already hold. Many do not. We think this is one reason they should, and all the more so as the use of LLM judges and the demand for the auditability of AI model evaluation grows (McCaslin et al., 2025; Brundage et al., 2026; Ghosh et al., 2026; Jiang et al., 2026a).

These lessons may also generalize to domains outside of evaluation as well. Reward models are trained on preference labels from raters who, like our experts, share a construct but apply it differently. Standard reward models, like consensus gold, often treat disagreement as noise (Lambert et al., 2024; Malik et al., 2026). We submit that a reward model that outperforms human consensus may warrant the same scrutiny as a judge that exceeds the human rater ceiling. Likewise, LLM content moderation constructs (community-built policies; Palla et al., 2025) are designed to decrease the likelihood of violation, which produces a label skew similar to the one we encountered in this audit. Although LLMs in either configuration are not strictly speaking judges, they are deputized with judge-like roles. We make four recommendations:

R1. Before analyzing a judge, identify where human raters genuinely disagree. Report base rates, compute κ within strata of comparable prevalence, and include a content-blind baseline. Recall that under label skew, aggregate agreement mostly measures the base rate without accounting for it.

R2. Treat the human ceiling more like a target, than a floor. Where a construct is defined by expert practice, the best judge may be the one that *matches* (not dramatically exceeds) the community's corrected consensus.

R3. Extensive analysis is the new standard. In "hard" contexts, a judge is validated only when it is stable and when it matches severity-corrected human performance *where signal is strongest*.

R4. Disclose rater data! Evaluation developers should release raw, rater-keyed votes alongside consensus labels, and judge analyses should report the full judge configuration and its human ceilings.

Towards a Minimum Viable Judge Assessment. To make our recommendations portable, we propose a Minimum Viable Judge Assessment (MVJA; Appendix I), a form of technical documentation bridging our methods and the judge disclosure templates now being called for (Sushmita et al., 2026). The MVJA targets rubric-based evaluations in "hard" contexts, where rater agreement is likely low and labels skewed. Six "required" checks, corresponding to R1–R4, report label prevalence and stratified κ, a content-blind baseline, the leave-one-out human ceiling, judge stability and configuration, and rater-level votes. "Recommended" and "as appropriate" checks add severity correction, error co-location, self-preference and robustness tests. Each check maps onto a judge disclosure report field (McCaslin et al., 2025).

5 Conclusion

Our audit of L2-Bench served as a live-fire exercise in judge validation. First, stability is not validity: tests showed the production judge was consistent, but in a "hard" domain, validity requires showing it scores like a qualified expert. Second, standard tools can fail under realistic conditions: label skew rendered aggregate κ uninformative, and experts themselves disagreed. Third, the field need not start from scratch: methods imported from educational measurement (isolating contested criteria, modeling rater severity, and building severity-corrected leave-one-out ceilings) let us show that the judge is consistent with expert-panel performance. The MVJA distills these lessons into a standard for judge assessment and disclosure. As LLM judges increasingly decide what AI evaluations claim to measure, validating them becomes inseparable from the measurement validity of the evaluation field itself. And, as we have demonstrated here, drawing on disciplines that have long grappled with scoring under disagreement is one practical way to strengthen that foundation.

AI USE STATEMENT

(This section is **required** and does not count toward the page limit.)

Generative AI tools assisted with: running judge evaluation pipelines; researching statistical methods; implementing statistical computations from specifications written by the authors; supporting and reviewing data analysis; manuscript preparation (LaTeX table and equation formatting, bibliography and citation formatting, copy-editing). Generative AI was not used to produce human ratings; develop theoretical models or conceptual frameworks; propose or refine hypotheses; design research methodology or experiments from scratch. AI feedback on methodology was treated as advisory and adopted only where the authors judged it sound. All AI-assisted code was checked by the authors, and every reported number was recomputed end-to-end from source data with outputs checked by the authors against those specifications before inclusion. The authors take full responsibility for the content of this paper, including text, claims and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

(This section is **recommended** and does not count toward the page limit.)

Human subjects. We collect no new human data. All human ratings are secondary data from the L2-Bench practitioner validation study (Anonymous, 2026), which was conducted under institutional oversight with informed consent, UK GDPR-compliant data handling, and a supplementary departmental ethics review that ruled the study exempt. Participants were told about the blind A/B design after the study and could withdraw. The rater-level votes we analyze are keyed by pseudonymous rater IDs. We report demographics only in aggregate, and quote free-text comments only by task ID.

Potential harms. Our recommendation to release rater-level data (R4) presumes pseudonymization and consent to secondary analysis. Rater-severity estimates describe how individual practitioners applied a rubric in one study; they are not measures of teaching quality and should not be used to evaluate educators.

REPRODUCIBILITY STATEMENT

(This section is **recommended** and does not count toward the page limit.)

The audited benchmark's task dataset is publicly released under CC BY-SA 4.0 (Anonymous, 2026). Appendix D documents every judge experiment in the order it was run, including the task samples, all 32 model/provider/decoding configurations (Table 7), and the full text of every judge prompt (Appendix D.4). Appendix D.5 describes the re-run that scores every task on the practitioner-rated responses. The gold-label rule is defined in Section 3.4, the severity model and correction rule in Appendix F, the leave-one-out frames in Appendix G, and every estimator in Appendix H. Uncertainty is reported as unit- or task-clustered bootstrap or Wilson intervals, as stated at each use.

ACKNOWLEDGMENTS

[[Redacted for anonymity]]

REFERENCES

Anurag Acharya, Bishal Lakha, Rounak Meyur, Rohan Nuttall, Sarthak Chaturvedi, Anika Halappanavar, Leah Hare, Lin Zeng, Mike Parker, Sai Munikoti, and Sameera Horawalavithana. DraftNEPABench: A benchmark for drafting NEPA document sections with coding agents. In *Proceedings of the ACM Conference on AI and Agentic Systems*, CAIS '26, pp. 674–688, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724152. doi: 10.1145/3786335.3813132. URL https://doi.org/10.1145/3786335.3813132.

Mahak Agarwal and Divyam Khanna. When persuasion overrides truth in multi-agent LLM debates: Introducing a confidence-weighted persuasion override rate (cw-por), 2025.

Mehrnoush Alizade, Tengrui Kong, and Suman Kalyan Maity. SciTables: A dataset and evaluation framework for complex table-to-text generation. *Proc. VLDB Endow.*, 19(11):3400–3412, Septem-

ber 2026. ISSN 2150-8097. doi: 10.14778/3836663.3836697. URL https://doi.org/10.14778/3836663.3836697.

Miles Brundage, Noemi Dreksler, Aidan Homewood, Sean McGregor, Patricia Paskov, Conrad Stosz, Girish Sastry, A. Feder Cooper, George Balston, Steven Adler, Stephen Casper, Markus Anderljung, Grace Werner, Sören Mindermann, Vasilios Mavroudis, Ben Bucknall, Charlotte Stix, Jonas Freund, Lorenzo Pacchiardi, José Hernández-Orallo, Matteo Pistillo, Michael Chen, Chris Painter, Dean W. Ball, Cullen O'Keefe, Gabriel Weil, Ben Harack, Graeme Finley, Ryan Hassan, Scott Emmons, Charles Foster, Anka Reuel, Bri Treece, Yoshua Bengio, Daniel Reti, Rishi Bommasani, Cristian Trout, Ali Shahin Shamsabadi, Rajiv Dattani, Adrian Weller, Robert Trager, Jaime Sevilla, Lauren Wagner, Lisa Soder, Ketan Ramakrishnan, Henry Papadatos, Malcolm Murray, and Ryan Tovcimak. Frontier AI auditing: Toward rigorous third-party assessment of safety and security practices at leading AI companies, 2026. URL https://arxiv.org/abs/2601.11699.

Nitay Calderon, Roi Reichart, and Rotem Dror. The alternative annotator test for LLM-as-a-judge: How to statistically justify replacing human annotators with LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081, Vienna, Austria, July 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.acl-long.782. URL https://aclanthology.org/2025.acl-long.782/.

Donald T. Campbell and Donald W. Fiske. Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56(2):81–105, 1959. doi: 10.1037/h0046016.

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

Yutong Cheng, Yang Liu, Changze Li, Dawn Song, and Peng Gao. CTIConnect: A benchmark for retrieval-augmented LLMs over heterogeneous cyber threat intelligence. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8741–8752, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817527. URL https://doi.org/10.1145/3770855.3817527.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. Diagnosing the reliability of llm-as-a-judge via item response theory, 2026. URL https://arxiv.org/abs/2602.00521.

Preetam Prabhu Srikar Dammu, Arnav Palkhiwala, Tanya Roosta, and Chirag Shah. iAgentBench: Benchmarking sensemaking capabilities of information-seeking agents on high-traffic topics. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3130–3137, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808627. URL https://doi.org/10.1145/3805712.3808627.

Paul De Boeck and Mark Wilson (eds.). *Explanatory Item Response Models: A Generalized Linear and Nonlinear Approach*. Springer, New York, NY, 2004. ISBN 978-1-4419-2323-3. doi: 10.1007/978-1-4757-3990-9.

S. Dev, P. Paskov, A. Sloan, K. Wei, P. Nascimento de Lima, S. Chowdhury, J. Johnson, and W. Marcellino. Simpler is better for autograders: Toward cost-effective llm evaluations for open-ended tasks. Technical report, RAND Corporation, Santa Monica, CA, 2026. URL https://www.rand.org/pubs/research_reports/RRA4618-1.html.

Jinru Ding, Chao Ding, Yidong Jiang, Wenrao Pang, Boyi Xiao, Zhiqiang Liu, Jiayuan Chen, Yun Zhong, Tiantian Yuan, Junming Guan, Dawei Cheng, and Jie Xu. Beyond knowledge to agency: Evaluating expertise, autonomy, and integrity in finance with CNFinBench. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8765–8776, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817482. URL https://doi.org/10.1145/3770855.3817482.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: A simple way to debias automatic evaluators. In *Conference on Language Modeling (COLM)*, 2024. doi: 10.48550/arXiv.2404.04475.

Aparna Elangovan, Lei Xu, Jongwoo Ko, Mahsa Elyasi, Ling Liu, Sravan Bodapati, and Dan Roth. Beyond correlation: The impact of human uncertainty in measuring the effectiveness of automatic evaluation and LLM-as-a-judge. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.03775.

Anushri Eswaran, Oleg Golev, Darshan Tank, Sidhant Rahi, and Himanshu Tyagi. CryptoAnalystBench: Failures in multi-tool long-form LLM analysis. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8813–8822, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817585. URL https://doi.org/10.1145/3770855.3817585.

Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. The problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990. doi: 10.1016/0895-4356(90)90158-L.

Avijit Ghosh, Anka Reuel, Jenny Chim, Wm. Matthew Kennedy, Srishti Yadav, Jennifer Mickel, Yanan Long, Andrew Tran, Anastassia Kornilova, Damian Stachura, Kevin Klyman, Felix Friedrich, Jeba Sania, Jan Batzner, Anoop Mishra, Eliya Habba, Yixiong Hao, Nathan Heath, Shalaleh Rismani, Usman Gohar, Andrea Loehr, David Manheim, Ruchira Dhar, Sree Harsha Nelaturu, Aarush Sinha, Leshem Choshen, Drishti Sharma, Ishan Khire, Amit Saha, Subramanyam Sahoo, Michael Hardy, Michael Alexander Riegler, Kabir Manghnani, Michelle Lin, Yanan Jiang, Yilin Huang, Asaf Yehudai, Jessica Ji, Aris Hofmann, Mubashara Akhtar, Max Lamparth, Nuno Moniz, Yacine Jernite, Stella Biderman, Zeerak Talat, Sanmi Koyejo, Mykel Kochenderfer, and Irene Solaiman. Evaluation cards: An interpretive layer for ai evaluation reporting, 2026. URL https://arxiv.org/abs/2606.09809.

Ozan Gokdemir, Neil Getty, Robert Underwood, Sandeep Madireddy, Franck Cappello, Arvind Ramanathan, Ian T. Foster, and Rick L. Stevens. Automated MCQA benchmarking at scale: Evaluating reasoning traces as retrieval sources for domain adaptation of small language models. In *Proceedings of the SC '25 Workshops of the International Conference for High Performance Computing, Networking, Storage and Analysis*, SC Workshops '25, pp. 545–552, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400718717. doi: 10.1145/3731599.3767410. URL https://doi.org/10.1145/3731599.3767410.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-judge, 2024.

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

Ashutosh Hathidara, Julien Yu, Vaishali Senthil, Sebastian Schreiber, and Anil Babu Ankisettipalli. MirrorBench: A benchmark to evaluate conversational user-proxy agents for human-likeness. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 8978–8988, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817496. URL https://doi.org/10.1145/3770855.3817496.

Peng He, Zhaohui Li, Zeyuan Wang, Jinjun Xiong, and Tingting Li. Judging the judges: Human validation of multi-llm evaluation for high-quality k–12 science instructional materials, 2026a. URL https://arxiv.org/abs/2602.13243.

Qing He, Dongsheng Bi, Jianrong Lu, Minghui Yang, Zixiao Chen, Jiacheng Lu, Jing Chen, Nannan Du, Xiao Cui, Sijing Wu, Peng Xiang, Yingying Hu, Yi Guo, Chunpu Li, Shaoyang Li, Zhuo Dong, Ming Jiang, Shuai Guo, Liyun Feng, Jin Peng, Zhou Yang, Han Ying, Jie Zheng, Yujie Yang, Jian Wang, Jinjie Gu, and Junwei Liu. MLB: A scenario-driven benchmark for evaluating large language models in clinical applications. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9013–9023, New York, NY, USA, 2026b.

Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817448. URL https://doi.org/10.1145/3770855.3817448.

Silan Hu, Shiqi Zhang, Yimin Shi, and Xiaokui Xiao. GEM-Bench: A benchmark for ad-injected response generation within generative engine marketing. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9103–9112, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817474. URL https://doi.org/10.1145/3770855.3817474.

Han Jiang, Susu Zhang, Dongyao Zhu, Yuzhuo Bai, Sang T. Truong, Xiaoyuan Yi, Sanmi Koyejo, Xing Xie, and Ziang Xiao. Ai evaluation should require standardized item-level data releases, 2026a. URL https://arxiv.org/abs/2604.03244.

Yidong Jiang, Junrong Chen, Eftychia Makri, Jialin Chen, Peiwen Li, Ali Maatouk, Leandros Tassiulas, Eliot Brenner, Bing Xiang, and Rex Ying. Fin-RATE: A real-world financial analytics and tracking evaluation benchmark for LLMs on SEC filings. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9183–9194, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817528. URL https://doi.org/10.1145/3770855.3817528.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. *arXiv preprint arXiv:2309.17012*, 2024. URL https://arxiv.org/abs/2309.17012. Findings of the Association for Computational Linguistics: ACL 2024.

Varvara Krechetova and Denis Kochedykov. GeoBenchX: Benchmarking LLMs in agent solving multistep geospatial tasks. In *Proceedings of the 1st ACM SIGSPATIAL International Workshop on Generative and Agentic AI for Multi-Modality Space-Time Intelligence*, GeoGenAgent '25, pp. 27–35, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400722615. doi: 10.1145/3764915.3770721. URL https://doi.org/10.1145/3764915.3770721.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. No free labels: Limitations of LLM-as-a-judge without human grounding, 2025.

Parth Kulshreshtha, Kriti Banka, Naman Khandelwal, Ashi Jain, Keerti Bajaj, Sunil Kothari, and Manish Mehta. Beyond public benchmarks: An LLM-as-judge framework for enterprise code evaluation with margin-driven optimization. In *Proceedings of the 1st International Workshop on AI for Software Quality Evaluation - Judgment, Metrics, Benchmarks, and Beyond*, AI-SQE '26, pp. 10–18, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400724077. doi: 10.1145/3786175.3788345. URL https://doi.org/10.1145/3786175.3788345.

Philippe Laban, Lidiya Murakhovs'ka, Caiming Xiong, and Chien-Sheng Wu. Are you sure? challenging LLMs leads to performance drops in the FlipFlop experiment, 2023.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. *arXiv preprint arXiv:2403.13787*, 2024. URL https://arxiv.org/abs/2403.13787.

Bryan Li, William Walden, Yu Hou, Gabrielle Kaili-May Liu, Dawn Lawrie, James Mayfield, Eugene Yang, Chris Callison-Burch, and Laura Dietz. DoGMaTiQ: Automated generation of question-and-answer nuggets for report evaluation. In *Proceedings of the 2026 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval (ICTIR)*, ICTIR '26, pp. 230–240, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400726002. doi: 10.1145/3805713.3820425. URL https://doi.org/10.1145/3805713.3820425.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge, 2024a.

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

Haitao Li, Yifan Chen, Hu YiRan, Qingyao Ai, Junjie Chen, Xiaoyu Yang, Jianhui Yang, Yueyue Wu, Zeyang Liu, and Yiqun Liu. LexRAG: Benchmarking retrieval-augmented generation in multi-turn legal consultation conversation. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '25, pp. 3606–3615, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400715921. doi: 10.1145/3726302.3730340. URL https://doi.org/10.1145/3726302.3730340.

Shaobo Li, Yirui Zhou, Yuan Xu, Kevin Chen, Daniel Waddington, Swaminathan Sundararaman, Hubertus Franke, and Jian Huang. RAGPerf: An end-to-end benchmarking framework for retrieval-augmented generation systems. *Proc. VLDB Endow.*, 19(11):3689–3703, September 2026b. ISSN 2150-8097. doi: 10.14778/3836663.3836718. URL https://doi.org/10.14778/3836663.3836718.

Yuanxian Li, Siyu Tao, Yimiao Feng, and Jie Zheng. BASIS: An LLM-empowered benchmark and agent system for inferring synthetic lethality. In *Proceedings of the 17th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics*, BCB '26, New York, NY, USA, 2026c. Association for Computing Machinery. ISBN 9798400726538. doi: 10.1145/3807503.3819483. URL https://doi.org/10.1145/3807503.3819483.

Ming Liang, Qingyu Zhang, Zhipeng Zuo, Shaoqiang Zheng, Dajun Chen, Wei Jiang, and Yong Li. FuseApplyBench: Multilingual benchmark for trustworthy code edit applying task. In *Proceedings of the 34th ACM SIGSOFT International Symposium on Software Testing and Analysis*, ISSTA Companion '25, pp. 183–189, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714740. doi: 10.1145/3713081.3732929. URL https://doi.org/10.1145/3713081.3732929.

John Michael Linacre. *Many-Facet Rasch Measurement*. MESA Press, Chicago, IL, 1989. ISBN 0-941938-02-6. No DOI assigned.

Wenxuan Liu, Zixuan Li, Long Bai, Chunmao Zhang, Fenghui Zhang, Zhuo Chen, Wei Li, Yuxin Zuo, Fei Wang, Bingbing Xu, Xuhui Jiang, Jin Zhang, Xiaolong Jin, Jiafeng Guo, Tat-Seng Chua, and Xueqi Cheng. Towards knowledgeable deep research: Framework and benchmark. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 1163–1174, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3809617. URL https://doi.org/10.1145/3805712.3809617.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. Rewardbench 2: Advancing reward model evaluation, 2026. URL https://arxiv.org/abs/2506.01937.

Tegan McCaslin, Jide Alaga, Samira Nedungadi, Seth Donoughe, Tom Reed, Rishi Bommasani, Chris Painter, and Luca Righetti. STREAM (ChemBio): A standard for transparently reporting evaluations in AI model reports, 2025. URL https://arxiv.org/abs/2508.09853.

Tim F. McNamara. *Measuring Second Language Performance*. Applied Linguistics and Language Study. Addison Wesley Longman, London, 1996. ISBN 978-0-582-08907-5. No DOI assigned.

Samuel Messick. Validity. In Robert L. Linn (ed.), *Educational Measurement*, pp. 13–103. American Council on Education and Macmillan, New York, NY, 3rd edition, 1989. No DOI assigned.

Samuel Messick. Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50 (9):741–749, 1995.

Melis Muradoglu, Joseph R. Cimpian, and Andrei Cimpian. Mixed-effects models for cognitive development researchers. *Journal of Cognition and Development*, 24(3):307–340, 2023. doi: 10.1080/15248372.2023.2176856.

Carol M. Myford and Edward W. Wolfe. Detecting and measuring rater effects using many-facet Rasch measurement: Part I. *Journal of Applied Measurement*, 4(4):386–422, 2003. No DOI assigned.

Bang Nguyen, Dominik Soós, Qian Ma, Rochana R. Obadage, Zack Ranjan, Sai Koneru, Timothy M. Errington, Shakhlo Nematova, Sarah Rajtmajer, Jian Wu, and Meng Jiang. ReplicatorBench: Benchmarking LLM agents for replicability in social and behavioral sciences. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 11693–11704, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3819003. URL https://doi.org/10.1145/3770855.3819003.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias, 2026. URL https://arxiv.org/abs/2606.19544.

Konstantina Palla, José Luis Redondo García, Claudia Hauff, Francesco Fabbri, Andreas Damianou, Henrik Lindström, Dan Taber, and Mounia Lalmas. Policy-as-prompt: Rethinking content moderation in the age of large language models. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '25, pp. 840–854, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 979-8-4007-1482-5. doi: 10.1145/3715275.3732054.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. doi: 10.48550/arXiv.2404.13076.

Dhaval Patel, Shuxin Lin, James Rayfield, Nianjun Zhou, Chathurangi Shyalika, Suryanarayana R Yarrabothula, Roman Vaculin, Natalia Martinez, Fearghal O'Donncha, and Jayant Kalagnanam. AssetOpsBench: A real-world evaluation benchmark for AI-driven task automation in industrial asset management. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9649–9660, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817562. URL https://doi.org/10.1145/3770855.3817562.

Valentin Radu, Myles Foley, Maryam Vahdat Pour, Massimo Pegoraro, Frank Faricy, and Mark Girolami. CrossBench: Reproducible LLM memory evaluation across conversational and coding workloads. In *Proceedings of the 24th Annual International Conference on Mobile Systems, Applications and Services Workshops*, MobiSys Workshop '26, pp. 87–92, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400727122. doi: 10.1145/3812836.3814760. URL https://doi.org/10.1145/3812836.3814760.

Shaina Raza, Aravind Narayanan, Vahid Reza Khazaie, Ashmal Vayani, Ahmed Y. Radwan, Mukund S. Chettiar, Amandeep Singh, Mubarak Shah, and Deval Pandya. HumaniBench: A human-centric framework for large multimodal models evaluation. *ACM Trans. Intell. Syst. Technol.*, September 2026. ISSN 2157-6904. doi: 10.1145/3845999. URL https://doi.org/10.1145/3845999. Just Accepted.

Russell Ro and Reza Abbasi-Asl. RECITE: Revising eligibility criteria incorporating textual evidence. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 11924–11935, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3819027. URL https://doi.org/10.1145/3770855.3819027.

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

Saron Samuel, Andrew Yates, Dawn Lawrie, Ian Soboroff, Trevor Adriaanse, Benjamin Van Durme, and Eugene Yang. CoverageBench: Evaluating information coverage across tasks and domains. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3388–3395, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808625. URL https://doi.org/10.1145/3805712.3808625.

John R. Searle. *The Construction of Social Reality*. Free Press, New York, NY, 1995. No DOI assigned.

Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2025. URL https://arxiv.org/abs/2310.13548.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2025)*, pp. 292–314, Mumbai, India, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Shanu Sushmita, Jayash Koshal, Meghana Makhija, Hui Wan, and Amjad Abu-Jbara. JUDGe 2026: Can we trust the judge? Building reliable evaluation for language models. Workshop at the 40th Conference on Neural Information Processing Systems (NeurIPS 2026), Atlanta, GA, 2026. URL https://judge2026.github.io/. Failure-mode taxonomy. Accessed 25 September 2026.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.12784.

Aman Singh Thakur, Kartik Choudhary, Venkat Srinik Ramayapally, Sankaran Vaidyanathan, and Dieuwke Hupkes. Judging the judges: Evaluating alignment and vulnerabilities in LLMs-as-judges, 2024.

Danielle R. Thomas, Conrad Borchers, Kirk P. Vanacore, Kenneth R. Koedinger, and René F. Kizilcec. Modernizing ground truth: Four shifts toward improving reliability and validity in AI in education. In Emmanuel G. Blanchard, Guanliang Chen, Min Chi, and Seiji Isotani (eds.), *Artificial Intelligence in Education (AIED 2026)*, volume 16586 of *Lecture Notes in Computer Science*, pp. 117–131, Cham, 2026. Springer. doi: 10.1007/978-3-032-29773-0_9.

Armin Toroghi, Faeze Moradi Kalarde, and Scott Sanner. CommonWhy: A dataset for evaluating entity-based causal commonsense reasoning in large language models. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3418–3425, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808628. URL https://doi.org/10.1145/3805712.3808628.

Tuhina Tripathi, Manya Wadhwa, Greg Durrett, and Scott Niekum. Pairwise or pointwise? evaluating feedback protocols for bias in LLM-based evaluation. In *Conference on Language Modeling (COLM)*, 2025. URL https://openreview.net/forum?id=uyX5Vnow3U.

Ilkka Tuomi. What counts as evidence in AI & ED: Towards science-for-policy 3.0. *European Journal of Education Policy and Practice*, 1(1):1–31, 2025. doi: 10.5117/EJEP2025.1.001.TUOM.

Anthony J. Viera and Joanne M. Garrett. Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5):360–363, 2005. No DOI assigned.

Hanna Wallach, Meera Desai, A. Feder Cooper, Angelina Wang, Chad Atalla, Solon Barocas, Su Lin Blodgett, Alexandra Chouldechova, Emily Corvi, P. Alex Dow, Jean Garcia-Gathright, Alexandra Olteanu, Nicholas J. Pangakis, Stefanie Reed, Emily Sheng, Dan Vann, Jennifer Wortman Vaughan, Matthew Vogel, Hannah Washington, and Abigail Z. Jacobs. Position: Evaluating generative AI systems is a social science measurement challenge. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 82232–82251. PMLR, 2025. URL https://proceedings.mlr.press/v267/wallach25a.html. PMLR assigns no DOI; preprint at https://doi.org/10.48550/arXiv.2502.00561.

Ziting Wang, Shize Zhang, Haitao Yuan, Jinwei Zhu, Wei Dong, and Gao Cong. FDABench: A benchmark for data agents on analytical queries over heterogeneous data. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 9985–9996, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817454. URL https://doi.org/10.1145/3770855.3817454.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. Self-preference bias in LLM-as-a-judge. In *NeurIPS Safe Generative AI Workshop*, 2024. doi: 10.48550/arXiv.2410.21819.

Martin Weyssow, Aton Kamanda, Xin Zhou, and Houari Sahraoui. CodeUltraFeedback: An LLM-as-a-judge dataset for aligning large language models to coding preferences. *ACM Trans. Softw. Eng. Methodol.*, 35(3), February 2026. ISSN 1049-331X. doi: 10.1145/3736407. URL https://doi.org/10.1145/3736407.

David M. Williamson, Xiaoming Xi, and F. Jay Breyer. A framework for evaluation and use of automated scoring. *Educational Measurement: Issues and Practice*, 31(1):2–13, 2012. doi: 10.1111/j.1745-3992.2011.00223.x.

Yin Wu, Quanyu Long, Jing Li, Jianfei Yu, and Wenya Wang. Visual-RAG: Benchmarking text-to-image retrieval augmented generation for visual knowledge intensive queries. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3473–3480, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808615. URL https://doi.org/10.1145/3805712.3808615.

Jia Xu, Tianyi Wei, Bojian Hou, Patryk Orzechowski, Shu Yang, Ruochen Jin, Rachael Paulbeck, Joost Wagenaar, George Demiris, and Li Shen. MentalChat16K: A benchmark dataset for conversational mental health assistance. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '25, pp. 5367–5378, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400714542. doi: 10.1145/3711896.3737393. URL https://doi.org/10.1145/3711896.3737393.

Yusuke Yamauchi, Taro Yano, and Masafumi Oyamada. An empirical study of LLM-as-a-judge: How design choices impact evaluation reliability. In *Proceedings of the Fifth Workshop on Generation, Evaluation and Metrics (GEM)*, pp. 167–176. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.gem-main.19. URL https://aclanthology.org/2026.gem-main.19/.

Seonghyeon Ye, Doyoung Kim, Sungdong Kim, Hyeonbin Hwang, Seungone Kim, Yongrae Jo, James Thorne, Juho Kim, and Minjoon Seo. FLASK: Fine-grained language model evaluation based on alignment skill sets. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://openreview.net/forum?id=CYmF38ysDa.

Yongan Yu, Xianda Du, Qingchen Hu, Jiahao Liang, Jingwei Ni, Dan Qiang, Kaiyu Huang, Grant Mckenzie, Renée Sieber, and Fengran Mo. WeatherArchive: A benchmark for retrieval-augmented reasoning over historical weather archives. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3514–3522, New York, NY, USA, 2026. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808603. URL https://doi.org/10.1145/3805712.3808603.

Zhengran Zeng, Ruikai Shi, Keke Han, Yixin Li, Kaicheng Sun, Yidong Wang, Zhuohao Yu, Rui Xie, Wei Ye, and Shikun Zhang. SWR-Bench: Assessing LLM performance in real-world code review comment generation. *Proc. ACM Softw. Eng.*, 3(FSE), June 2026. doi: 10.1145/3808144. URL https://doi.org/10.1145/3808144.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations (ICLR 2024)*, 2024. URL https://arxiv.org/abs/2310.07641.

Dake Zhang, Mark D. Smucker, and Charles L. A. Clarke. Resources for automated evaluation of assistive RAG systems that help readers with news trustworthiness assessment. In *Proceedings*

of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '26, pp. 3544–3551, New York, NY, USA, 2026a. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808624. URL https://doi.org/10.1145/3805712.3808624.

Fanjin Zhang, Zhengyang Wang, Ruixuan Huang, Kefan Zhang, Amy Xin, Yuanchun Wang, Shu Zhao, Evgeny Kharlamov, Jie Tang, and Juanzi Li. AISE-Bench: A full-cycle curated benchmark for information seeking on academic knowledge graphs. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 10234–10245, New York, NY, USA, 2026b. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3817492. URL https://doi.org/10.1145/3770855.3817492.

Junkai Zhang, Jingru Gan, Xiaoxuan Wang, Zian Jia, Changquan Gu, Jianpeng Chen, Yanqiao Zhu, Mingyu Derek Ma, Dawei Zhou, Ling Li, and Wei Wang. MatSciBench: Benchmarking the reasoning ability of large language models in materials science. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, KDD '26, pp. 12868–12879, New York, NY, USA, 2026c. Association for Computing Machinery. ISBN 9798400722592. doi: 10.1145/3770855.3818888. URL https://doi.org/10.1145/3770855.3818888.

Xinkai Zhang, Jingtao Zhan, Yiqun Liu, and Qingyao Ai. TEC: A collection of human trial-and-error trajectories for problem solving. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '26, pp. 3559–3567, New York, NY, USA, 2026d. Association for Computing Machinery. ISBN 9798400725999. doi: 10.1145/3805712.3808611. URL https://doi.org/10.1145/3805712.3808611.

Justin Zhao, Himaghna Bhattacharjee, Hannah Korevaar, Bhaktipriya Radharapu, and Khalid El-Arini. Jagged judges: Epistemic stability under silence, pressure, and persistence, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *Advances in Neural Information Processing Systems (NeurIPS) Datasets and Benchmarks Track*, 2023. doi: 10.48550/arXiv.2306.05685.

Fengbin Zhu, Junfeng Li, Liangming Pan, Wenjie Wang, Fuli Feng, Chao Wang, Huanbo Luan, and Tat-Seng Chua. Towards temporal-aware multi-modal retrieval augemented generation in finance. In *Proceedings of the 33rd ACM International Conference on Multimedia*, MM '25, pp. 6289–6297, New York, NY, USA, 2025. Association for Computing Machinery. ISBN 9798400720352. doi: 10.1145/3746027.3755723. URL https://doi.org/10.1145/3746027.3755723.

Appendix

A Rapid review methodology and per-benchmark breakdown
C The L2-Bench practitioner validation study
D Judge experiments, in the order run: baseline (D.1), power (D.2), configurations (D.3), prompts (D.4), alignment on the rated responses (D.5), reference answers (D.6), stability (D.7), self-preference (D.8), prompt and score sensitivity (D.9), web search and cost (D.10)
E Criterion categories and κ-paradox agreement metrics
F Rater severity model
G Leave-one-out ceilings and hard-criteria removal
H Equations

A Rapid Review Methodology and Extended Discussion

Building on prior surveys of the field as a whole (the most recent published in late 2024), we map the current state of judge assessment with a rapid review of 248 papers from the ACM Digital Library, a prominent repository for the proceedings of top AI/ML conferences. Papers were surfaced by the query "LLM judge" and restricted to research articles. We identified 37 articles within this set that proposed new evaluations in which an LLM judge was implemented as an automated scorer. Within this subset, we assessed the depth of judge assessment reported in the main body and the appendix, labeling papers "minimal" (judge analysis of less than one paragraph), "moderate" (a named subsection but no further analysis), or "extensive" (a named subsection and additional analysis in an appendix).

Table 3: Depth of LLM-judge analysis across 37 benchmarks (2024-2026) with LLM judges. Depth is coded as *minimal* (<1 paragraph), *moderate* (a subsection), or *extensive* (a subsection and appendix).

Analysis depth	Benchmarks	Share
Extensive	5	13.5%
Moderate	17	45.9%
Minimal	15	40.5%
Total	37	100.0%

Table 4: Per-benchmark breakdown of judge-performance analysis, grouped by depth.

Benchmark	What the judge-performance analysis consists of
Extensive — subsection *and* appendix analysis ($n = 5$)	
AssetOpsBench (Patel et al., 2026)	Sec. 5.5 Human Validation of LLM Judge (Cohen's κ per dimension, 5 runs) + extended-validation table
FDABench (Wang et al., 2026)	Judge-validation study with 3 experts + appendix Krippendorff α / ICC / τ_b table
RECITE (Ro & Abbasi-Asl, 2026)	"Human Evaluation of the Automated Judge" subsection + appendix quadratic-weighted $\kappa = 0.694$
CTIConnect (Cheng et al., 2026)	Body validation ($\kappa = 0.85$, self-consistency, bias diff.) + Appendix A.1 full protocol
BASIS (Li et al., 2026c)	Robustness statement + Appendix E.2 human audits + F.12 second-judge same-family bias check
Moderate — a subsection, nothing in an appendix ($n = 17$)	
CNFinBench (Ding et al., 2026)	Human-agreement study ($\kappa = 0.74$, 3 experts) + prompt-sensitivity robustness
DraftNEPABench (Acharya et al., 2026)	Sec. 4.4 Human Validation (procedure + secondary independent validation vs. SMEs)
MirrorBench (Hathidara et al., 2026)	Sec. 4.3 Human–Judge alignment + multi-seed robustness + calibration controls
ReplicatorBench (Nguyen et al., 2026)	Human eval of 128 checkpoints, Krippendorff α, GPT-4o vs. GPT-5, self-inflation check
CryptoAnalystBench (Eswaran et al., 2026)	Reliability validation vs. expert subset; judge flags failures but misses precise scores
Fin-RATE (Jiang et al., 2026b)	Human Validation: inter-annotator 87.88%, inter-judge agreement, weights tuned to humans
Knowledgeable Deep Research (Liu et al., 2026)	Sec. 5.4.1 Judge Model Analysis + pairwise-agreement rate across judges
SWR-Bench (Zeng et al., 2026)	Validation subsection: $\sim$90% agreement, κ, agreement ranges across tasks

continued on next page

Table 4 – continued from previous page

Benchmark	What the judge-performance analysis consists of
SciTables (Alizade et al., 2026)	Sec. 4.2.1 Human–LLM Judge Alignment (Spearman/Kendall; base vs. fine-tuned judge)
WeatherArchive (Yu et al., 2026)	Sec. 3.2.3 Oracle Quality & Validation: Fleiss $\kappa = 0.76$, accuracy 0.82, winning rate
Resources for Assistive RAG (Zhang et al., 2026a)	Kendall's τ (run-level) + Cohen's κ / Gwet's AC1 (label-level) vs. human
MLB (He et al., 2026b)	Trained SFT judge validated at 92.1% accuracy, $\kappa = 81.3\%$ against 19k expert annotations
CodeUltraFeedback (Weyssow et al., 2026)	"LLM Judgements Correlate with Human Evaluators" (Cohen's κ, per preference)
MentalChat16K (Xu et al., 2025)	Dual LLM+human eval; Cohen's κ among evaluators; LLM–human consistency
RAGPerf (Li et al., 2026b)	Judge-sensitivity analysis: multiple judges, run-to-run variance, cross-judge agreement
GeoBenchX (Krechetova & Kochedykov, 2025)	Panel of judge evaluators + human-evaluation section testing evaluator performance
DoGMaTiQ (Li et al., 2026a)	Nugget judgments compared to official judgments via rank correlations
Minimal — under one paragraph ($n = 15$)	
CoverageBench (Samuel et al., 2026)	Judge derives qrels; no reliability analysis
CrossBench (Radu et al., 2026)	Judge scores correctness; no analysis
Automated MCQA (Gokdemir et al., 2025)	Arbitrary judge grades; no analysis
Beyond Public Benchmarks (Kulshreshtha et al., 2026)	Describes judge method (Sec. 3.1.3); no performance validation
FuseApplyBench (Liang et al., 2025)	Three judge metrics; no reliability analysis
MatSciBench (Zhang et al., 2026c)	Optional judge for formula questions; no analysis
Visual-RAG (Wu et al., 2026)	One sentence: judge F1$= 0.92$ vs. human
TEC (Zhang et al., 2026d)	One sentence: judge agrees 94.8% with exact match ($\kappa = 0.892$)
iAgentBench (Dammu et al., 2026)	Three-judge panel; no reliability analysis
Temporal-Aware Finance RAG (Zhu et al., 2025)	Inter-annotator agreement for data creation, not the judge
CommonWhy (Toroghi et al., 2026)	GPT-4o judge for equivalence; no validation
GEM-Bench (Hu et al., 2026)	One-line correlation $\rho = 0.73$ with humans
LexRAG (Li et al., 2025)	Generic "rigorous validation" claim; no analysis
HumaniBench (Raza et al., 2026)	LLM-Judge metrics; no judge-vs-human agreement analysis
AISE-Bench (Zhang et al., 2026b)	Selects best-human-aligned judges; minimal analysis

B LIMITATIONS

One solver. Every Part A response was generated by Claude Sonnet 4.6, and the stability responses by Claude Haiku 4.5. The self-preference test therefore cannot separate generator identity from response quality, and the sensitivity of model rankings to judge choice is untested.

One benchmark. The method is applied to one benchmark; we expect, but do not show, that it transfers to other rubric-based benchmarks with rater-level data.

C L2-BENCH PRACTITIONER VALIDATION STUDY

The L2-Bench practitioner study ran online from 2 to 12 March 2026. Participants were recruited through three closed channels (an institutional teacher panel, external practitioner networks, and internal staff), giving 367 registered practitioners, of whom 241 took part. Panel members were paid the standard panel rate for one hour of work. All other participation was voluntary, and anyone contributing 3.5 hours or more was offered acknowledgement. The cohort spanned six stakeholder groups: teachers, content developers, assessment specialists, generalist education professionals, education researchers, and learners. After quality exclusions, 221 practitioners from 45 countries remained. Most (85.5%) reported more than ten years of experience, and the five largest countries, all European, accounted for 55% of the cohort.

Items and allocation. A stratified random sample of 504 items (42 per competency, about half of the public set) was drawn, aiming for three independent ratings per item. Each item combined the task input (with its resources inlined), a Claude Sonnet 4.6 response, the expert reference answer, and the full rubric (task-specific, consensus, and universal criteria). An intake survey recorded 24 binary role and expertise flags. A rule-based allocator used these to assign each practitioner between two and eight competencies, filling the rarest competencies first. Assessment specialists were capped at three competencies. Items were then drawn at random within each practitioner's competencies, in proportion to the time they had offered.

Procedure. A custom web application took each item through time-gated stages, each unlocked only once the previous stage was submitted. In Stage A, practitioners rated task authenticity and criteria adequacy on 5-point scales. In Stage B, they saw two unlabelled responses in random order (the Sonnet 4.6 response and the reference answer) and chose the one they preferred; there was no "no preference" option. In Stage C, the application "revealed" which response had come from the AI, and the practitioner scored that response Pass or Fail on every rubric criterion, with optional comments. The reveal was randomised: for roughly half of rater-items, the text presented as the "AI response" was in fact the reference answer. This let us compare rubric scores on the two texts under the same source label. Practitioners who committed one hour completed Stages A and B for four items. Those who committed 3.5 hours or more also completed Stage C, and anyone could request further blocks of six items, up to 30 in total. Calibration slides and a 20-minute recording were available throughout, and practitioners confirmed at each login that they had reviewed them.

Comments. Comments were optional at every stage. Stage A drew comments on 287 rater-items (100 raters), Stage B on 387 of 1,447 (26.7%; 103 raters), and Stage C on 378 (90 raters).

Quality control and ethics. Every session event was logged (51,975 events). Raters were excluded under a two-tier rule. The first tier was a hard cut: a median time per item under one minute, or a screenshot rate of 50% or more. The second was a weighted score built from timing and screenshot signals, offset by a leniency score reflecting professional background. Twenty raters were excluded. Of the retained cohort, 142 practitioners completed Stage C, contributing 17,093 criterion ratings on 442 items (Table 5). The study was reviewed by an institutional research ethics committee and ruled exempt. Consent was recorded in the application, data were pseudonymised, and participants were debriefed about the randomised reveal after the study and offered the chance to withdraw.

Table 5: Stage C practitioner ratings by the response actually scored as the "AI response" (N = 142 practitioners, after exclusions). A rater-item is one practitioner scoring one item; 201 items were scored in both arms by different practitioners. Gold units are (item, criterion) pairs with $\geq$2 raters and $\geq$67% agreement on the same scored text.

	AI response	**Reference answer**	**Total**
Items	329	314	442
Rater-items	505	512	1,017
Criterion ratings	8,446	8,647	17,093
Criterion pass rate	81.1%	80.9%	81.0%
Mean weighted task score	0.853	0.851	–
Gold units (items)	1,854 (139)	1,802 (136)	3,656 (249)

D JUDGE EXPERIMENTS

This appendix describes every judge experiment in the order it was run: the inherited baseline (D.1), the power analysis that sized the extension (D.2), the configurations and prompts (D.3–D.4), and then each experiment (D.5–D.10).

D.1 L2-BENCH BASELINE JUDGE ANALYSIS

Judge selection. The L2-Bench authors ran two small experiments to select the production judge (Claude Sonnet 4.6) and to validate its stability and agreement with human raters. Both used a stratified 48-task sample rated by language education practitioners. First, they tested three judge models under four prompt variants (Table 6), which vary whether the reference answer is shown, whether PASS/FAIL definitions and worked examples are given, and whether the judge writes a critique before its verdict. Second, they ran an initial stability analysis.

Table 6: L2-Bench judge prompt variants (full text in Appendix D.4). v1 is the only variant without reasoning before the verdict.

Version	Reference	Definitions	Examples	Critique first	Used in
v1 (production)	yes	no	no	no	Part A, stability
v2	yes	yes	3	yes	Part A, stability
v3	no	no	no	yes	Part A, Part B, stability
v4	no	yes	3	yes	Part B, stability

Judge bias. The L2-Bench authors also ran an initial self-preference analysis. Although another model from the same Claude family (Opus 4.7) performed best on the L2-Bench evaluation dataset, the authors note two design mitigations against same-family self-preference in the production judge: (1) judging is a reference-guided, per-criterion binary grading against a concrete expert answer, which constrains the room for stylistic preference; and (2) judge selection weighed both performance and stability across multiple judge model families. The technical report found cross-family agreement comparable to within-judge resampling stability, and categorized judge behavior close to decision boundaries as early signal on systematic failure modes. It lists self-preference as a known limitation.

Judge stability. L2-Bench also ran a 3-resample stability test. Five tasks (76 criteria) were each independently re-scored three times by four models (Sonnet 4.5, Sonnet 4.6, DeepSeek V3.2, and Kimi K2.5) under each of the four prompt versions. Stability was defined as the rate at which all three verdicts were identical, complemented by the flip rate. The production judge (Sonnet 4.6, prompt v1) was 100% stable (0% flip rate) across all 76 criteria. CoT prompting reduced stability by 7.1 pp, as expected from the literature (Dev et al., 2026). The 101 flip cases were coded manually using abductive categories: most were thresholding disagreements, followed by criterion interpretation ambiguity, negation/polarity confusion, implicit versus explicit requirements, scope, and audience or target confusion.

Judge validation against human ratings. Finally, the L2-Bench authors validated the judge against the practitioners. The human inter-rater ceiling was low, which is unsurprising (He et al., 2026a; Thomas et al., 2026): Krippendorff's $\alpha = 0.362$, with 89.8% raw agreement. Against it, the production judge returned Cohen's $\kappa = 0.746$ against the human majority vote, with 92.3% accuracy. Although κ and α are not directly comparable, the authors argue, following Messick (1995), that validity rests on judge–human convergence rather than perfect rater consensus. Our extension reproduces this baseline closely on the original 48-task scope (accuracy 0.926, $\kappa = 0.741$ at $n = 337$, against the published 0.923 and 0.746 at $n = 326$; the count differs because gold units are now keyed on the scored text).

D.2 POWER ANALYSIS

We sized both extensions from the baseline pilot before running them.

Stability. The pilot (5 tasks, 76 criteria, 3 resamples, 16 configurations) gave flip rates of 0–27.6% (pooled 8.3%). With 24 tasks (≈364 criteria), the 95% half-width of a flip rate is ± 1.4 pp at $p = 2\%$

and ±3.1 pp at $p = 10\%$ (Eq. 6). A criterion with per-resample flip probability q is detected as unstable with probability $1 - q^k - (1 - q)^k$ (Eq. 7): at $q = 0.1$ this is 27% at $k = 3$ and 65% at $k = 10$. We recommended 24 tasks (two per competency) at $k \geq 5$; the experiment ran 24 tasks at $k = 10$.

Alignment. From the pilot's task-level standard deviation of score shifts ($s \approx 18$–22 pp), the number of tasks needed for a target standard error is $n = (s/\mathrm{SE})^2$ (Eq. 8): 82–121 tasks for SE ≤ 2.0 pp, and 329–485 for SE ≤ 1.0 pp. The balanced maximum available with $\geq$3 raters per task was 264 tasks (22 per competency), which would give SE ≈1.1–1.4 pp. The experiment ran 98 tasks (37% of that), targeting CIs under 2 pp. The achieved 95% bootstrap half-widths for the production judge at $n = 650$ gold verdicts are ±0.022 (accuracy), ±0.082 (κ), and ±0.014 (F1). This bootstrap resamples criteria and ignores clustering within tasks, so these half-widths are likely too narrow.

D.3 JUDGE CONFIGURATIONS

Table 7: The 32 Part A judge configurations. All judges received the prompt as a single user message with no system prompt. Bedrock judges ran without extended thinking (production uses a 1,024-token thinking budget). Their temperature was 0.0 on all 98 tasks. Fireworks judges used reasoning effort "none", with unparseable outputs re-run at temperature 0.1. The Naive baseline passes a criterion iff its weight is positive.

Family	Model	Provider	Prompts	Temp.	Configs
Claude	Sonnet 4.6 (production), Sonnet 4.5	Bedrock	v1–v3	0.0	6
DeepSeek	V3.2	Bedrock	v1–v3	0.0	3
DeepSeek	V4 Flash	Fireworks	v1–v3	0.0	3
GLM	GLM-5.2	Fireworks	v1–v3	0.0	3
Qwen	Qwen 3.8 Max, Qwen 3.8 27B	Fireworks	v1–v3	0.0	6
Qwen	Qwen 3.5 9B	Fireworks (dedicated)	v1–v3	0.0	3
GPT	GPT-5.6-Terra	Responses API	v1	see caption[a]	8
Total judge configurations					32

[a] A 2 × 4 grid: web search on/off × {temperature 0.0, 0.5, default 1.0 with low reasoning effort, default 1.0 with xhigh reasoning effort}; max 4,096 output tokens.

The stability experiment (Appendix D.7) used a different set of 44 configurations: the nine non-GPT models (adding Kimi K2.5) under prompts v1–v4, plus the same eight-point GPT-5.6-Terra grid under v1.

D.4 JUDGE PROMPTS

All four prompts share a header ("You are an expert evaluator for AI-generated responses in second language (L2) education. Your job is to judge whether an AI response meets a single scoring criterion.") followed by the task, [reference answer], AI response and criterion blocks. v1, the production prompt, reads in full:

```
## Original Task
{task_text}
## Reference Answer
{reference_answer}
## AI Response
{ai_response}
## Criterion
- Description: {criterion_description}
## Instructions
Determine whether the AI response meets the criterion
above. Use the reference answer as a guide for what a
high-quality response looks like, but note that alternative
valid approaches may exist.
- 'true' if the criterion IS met.
- 'false' if the criterion IS NOT met.
```

```
You MUST output either 'true' or 'false' and **nothing
else**.  Otherwise your answer will NOT be parsed correctly.
```

v3 removes the reference answer block and replaces the instructions with:

```
First, write a detailed critique that:
- Identifies whether the specific requirement of the
criterion is present.
- Cites specific evidence from the AI response to support
your judgement.
After your critique, output your final verdict on a new line.
Output ONLY 'true' if the criterion IS met, or ONLY 'false'
if the criterion IS NOT met.  Do not output any other text
after your verdict --- it will NOT be parsed correctly.
```

v2 keeps the reference answer and adds, before the v3 instructions, PASS/FAIL definitions (PASS: the response "meets the core requirement of the specific criterion"; FAIL: the content is "factually incorrect (especially pedagogical content)", "entirely missing", or pitched inappropriately for the target learners). It also adds three worked examples, each with a critique and result, on teaching modal verbs to an A2 class: a clear PASS, a clear FAIL, and a borderline PASS. v4 is v2 without the reference answer. A parse failure triggers one retry with the suffix "Error: You output an invalid value. You MUST output 'true' or 'false' ONLY in your response." Because v1 is the only prompt without a critique, contrasts between v1 and the other prompts confound reference guidance and few-shot examples with reasoning.

D.5 Judge–human alignment (Part A) on the rated responses

Design. Each configuration scored the same Sonnet 4.6 responses on 98 tasks (1,863 criteria): the 48 baseline tasks, plus 50 tasks drawn from the practitioner-rated pool. The 50 were balanced by competency, restricted to published tasks, and ordered by number of human-labelled criteria (so they maximise label density; they are not a random sample). 41 of the 98 tasks carry gold labels on the AI response, giving $n = 650$ gold verdicts per configuration (Table 8).

Rated responses. On all 98 tasks, judges score the exact Sonnet 4.6 response the practitioners rated, so every gold verdict compares judge and humans on the same text. A first run on the 50 new tasks had regenerated the responses (and used provider-default temperature on Bedrock); it was discarded, and the 50 tasks were re-run on the rated responses under the baseline configuration (Table 7). Agreement is lower on the new tasks: the production judge reaches $\kappa = 0.741$ on the 337 baseline-task verdicts (22 tasks) and 0.530 on the 313 new-task verdicts (19 tasks), where the gold pass rate is higher (0.911 vs. 0.831).

Table 8: Full Part A leaderboard, sorted by κ on the 650 gold verdicts. $n = 650$ unless noted (a few configurations have unparseable verdicts).

Configuration	Accuracy	F1	κ	Pass
Naive baseline	0.963	0.979	0.842	0.860
Qwen 3.8 Max | v3	0.935	0.962	0.742	0.838
GLM-5.2 | v3	0.926	0.957	0.691	0.854
Qwen 3.8 Max | v1	0.928	0.958	0.690	0.862
Sonnet 4.6 | v1 (production)	**0.909**	**0.946**	**0.651**	**0.825**
DeepSeek V3.2 | v1	0.895	0.937	0.638	0.786
Qwen 3.8 27B | v3	0.891	0.934	0.632	0.775
Sonnet 4.6 | v3	0.897	0.938	0.630	0.800
DeepSeek V4 Flash | v3	0.902	0.942	0.627	0.820
GPT-5.6-Terra, T1.0 xhigh, web	0.888	0.932	0.606	0.791
DeepSeek V3.2 | v3	0.894	0.937	0.589	0.828
Sonnet 4.6 | v2	0.902	0.943	0.584	0.857
GLM-5.2 | v1	0.895	0.939	0.574	0.845
GPT-5.6-Terra, T1.0 xhigh, no web	0.875	0.925	0.573	0.782
Sonnet 4.5 | v3 ($n = 649$)	0.858	0.913	0.549	0.752

Configuration	Accuracy	F1	κ	Pass
Qwen 3.8 27B | v2	0.874	0.925	0.522	0.820
GPT-5.6-Terra, T1.0 low, no web	0.857	0.913	0.514	0.778
GPT-5.6-Terra, T1.0 low, web	0.842	0.902	0.493	0.754
Sonnet 4.5 | v1	0.854	0.912	0.479	0.797
GPT-5.6-Terra, T0.0, no web	0.812	0.881	0.453	0.712
GPT-5.6-Terra, T0.0, web	0.805	0.876	0.440	0.705
GPT-5.6-Terra, T0.5, no web	0.805	0.876	0.436	0.708
GPT-5.6-Terra, T0.5, web	0.803	0.875	0.434	0.706
DeepSeek V3.2 | v2	0.874	0.928	0.428	0.878
Sonnet 4.5 | v2	0.835	0.902	0.394	0.809
DeepSeek V4 Flash | v2	0.868	0.925	0.374	0.891
Qwen 3.5 9B | v2 ($n = 646$)	0.816	0.888	0.369	0.779
Qwen 3.8 27B | v1	0.874	0.929	0.355	0.912
Qwen 3.5 9B | v1	0.894	0.942	0.347	0.957
Qwen 3.5 9B | v3 ($n = 648$)	0.785	0.867	0.336	0.738
Qwen 3.8 Max | v2	0.865	0.925	0.227	0.940
DeepSeek V4 Flash | v1	0.883	0.937	0.221	0.974
GLM-5.2 | v2	0.832	0.906	0.137	0.914

D.6 Performance on reference answers (Part B)

Judges re-scored the expert reference answers against their own rubrics. The expected verdict is Pass for every positively weighted criterion and Fail for every penalty criterion. Two configurations were run on all 98 tasks; five were run on the 48 baseline tasks only (Table 9). On the shared 48 tasks, GLM-5.2 | v3 scores 0.915 / 0.948 / 0.715 (accuracy / F1 / κ) and Sonnet 4.6 | v3 0.899 / 0.938 / 0.672. The two agree on 93.4% of criteria.

Table 9: Part B: judge verdicts on expert reference answers against the rubric's expected verdict, on each configuration's native scope.

Configuration	Accuracy	F1	κ	Precision	Recall	n
GLM-5.2 | v3	0.916	0.950	0.685	0.963	0.937	1,863
Sonnet 4.6 | v3	0.903	0.942	0.648	0.963	0.922	1,863
DeepSeek V3.2 | v3	0.877	0.923	0.612	0.960	0.889	875
Sonnet 4.6 | v4	0.897	0.939	0.604	0.926	0.953	875
DeepSeek V3.2 | v4	0.878	0.927	0.549	0.923	0.932	875
Sonnet 4.5 | v3	0.843	0.901	0.539	0.958	0.849	875
Sonnet 4.5 | v4	0.853	0.911	0.471	0.913	0.910	875

D.7 Stability over resampling

Design. We kept the four published baseline tasks and added two published tasks per competency, choosing those closest to the median criterion count (16), for 24 tasks and 381 criteria. One fixed response per task (Claude Haiku 4.5) was re-scored ten times by each of 44 configurations: nine models $\times$ prompts v1–v4, plus the eight-point GPT-5.6-Terra grid under v1. A criterion is *stable* if all ten verdicts agree; the flip rate is the share of criteria that are not stable, with Wilson 95% CIs (Eq. 6).

Configuration vs. family. The 2.9% in the main text is the flip rate of the production *configuration* (97.1% stable). The 93.2% is the stable rate of the Sonnet 4.6 *family*, pooled over its four prompts (97.1, 89.8, 94.2, and 91.9%). The two use the same definition at different levels of aggregation. Sonnet 4.6 is the third most stable family, behind Qwen 3.8 27B (96.5%) and Qwen 3.8 Max (93.4%). The production configuration is fourth: Qwen 3.8 27B | v1 (0.3%), DeepSeek V4 Flash | v1 (1.8%) and Qwen 3.5 9B | v1 (2.1%) flip less, and Qwen 3.8 27B | v3 ties it (Table 10, Figure 4).

D.8 Self-preference

Pairwise consensus. Across all 1,847 criteria scored by every one of the 32 configurations, judge pairs agree on 84.4% of verdicts on average (range 73.5–96.6%; κ mean 0.510, range 0.110–0.918).

Table 10: Stable (exact-match across ten resamples) rate by judge family, pooled over prompts v1–v4 ($n = 1{,}524$ criterion records per family; GPT-5.6-Terra: eight v1 configurations, $n = 3{,}048$).

Family	Stable	Family	Stable
Qwen 3.8 27B	96.5%	Sonnet 4.5	88.6%
Qwen 3.8 Max	93.4%	GLM-5.2	87.9%
Sonnet 4.6	93.2%	GPT-5.6-Terra	83.0%
DeepSeek V4 Flash	90.0%	DeepSeek V3.2	76.8%
Qwen 3.5 9B	89.3%	Kimi K2.5	65.4%

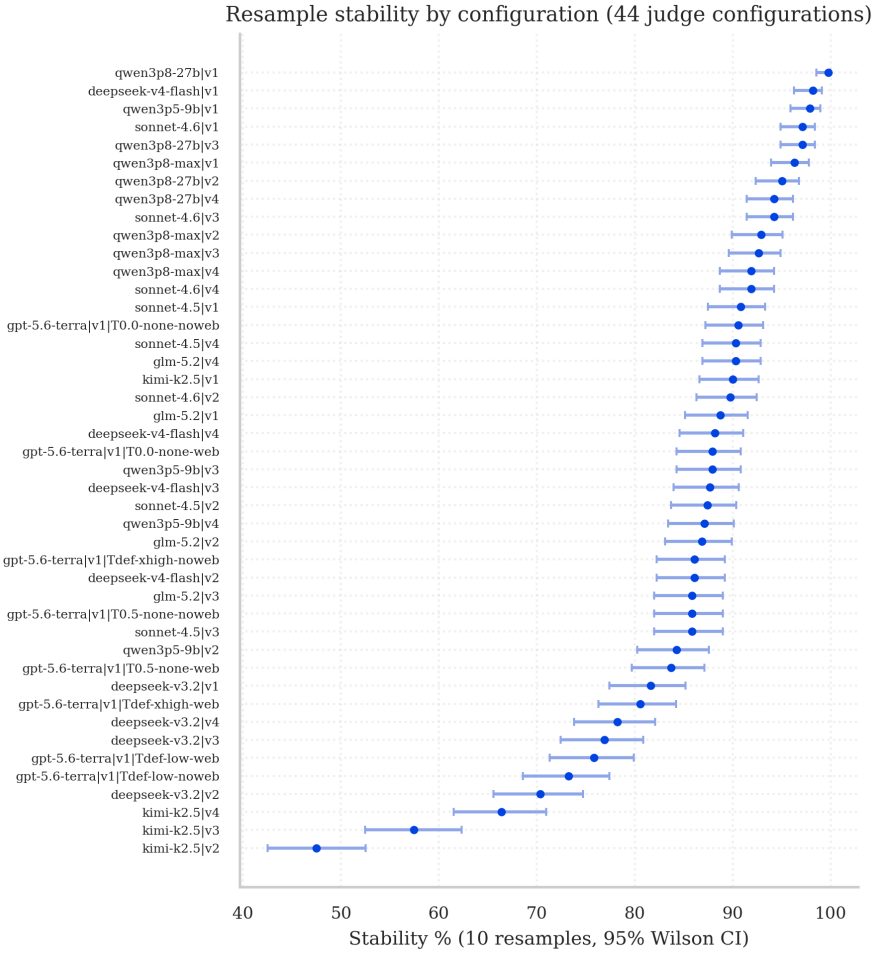

Figure 4: Stability (share of criteria with identical verdicts across ten resamples of 24 tasks, 381 criteria) with Wilson 95% confidence intervals, for all 44 judge configurations. The production judge (Sonnet 4.6 | v1) flips 2.9% of verdicts (CI $[1.6, 5.1]$); the Naive baseline is 0% by construction. Pooled across configurations, 59.5% of flips move toward Fail.

The most similar pair is two GPT-5.6-Terra sampling points (temperature 0.0 and 0.5, no web search; $\kappa = 0.918$). Figure 5 clusters the ten configurations with the highest alignment κ, using $1 - \kappa$ as the distance. GPT-5.6-Terra (xhigh, web) is an outlier, and DeepSeek V3.2 | v1 and v3 pair within family. The rest form one cluster with two branches: the v3 configurations of GLM-5.2, DeepSeek V4 Flash, Qwen 3.8 27B and Sonnet 4.6 (the closest pair, Qwen 3.8 27B and Sonnet 4.6, $\kappa = 0.811$), and Sonnet 4.6 | v1 with Qwen 3.8 Max | v1 and v3 ($\kappa = 0.800$ within Qwen 3.8 Max). Prompt structures the clusters more than family, but not exclusively, and the production family does not form a bloc: its

two configurations sit on different branches ($\kappa = 0.733$). Only the selection of the top ten depends on the gold labels.

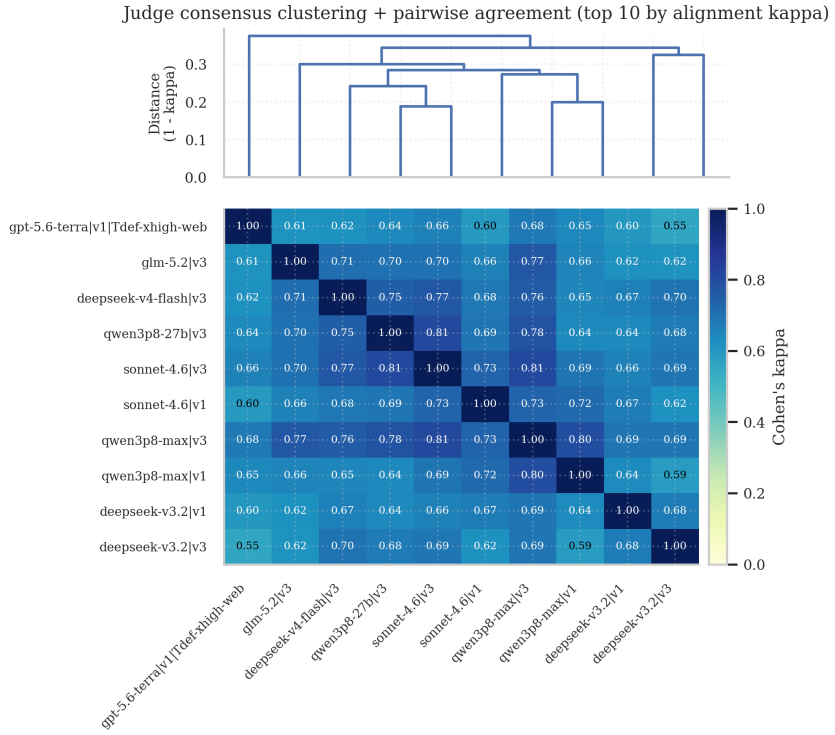

Figure 5: Hierarchical clustering (distance $1 - \kappa$) and pairwise Cohen's κ for the ten Part A configurations with the highest alignment κ, on identical Sonnet 4.6 responses.

Difference-in-differences. Pairwise consensus cannot show self-preference on its own: two judges with opposite source preferences can produce the same aggregate scores. We therefore compare each judge's leniency on the two texts the experts scored, using the 48 baseline tasks, where every judge verdict and human label refer to the same text. The leniency shift is (judge pass rate $-$ gold pass rate) on the Sonnet 4.6 response minus the same quantity on the reference answer (Eq. 10). A self-preferring Claude judge would show a positive shift. Claude judges instead show negative shifts, and the Claude minus non-Claude difference is -3.5 pp (95% CI $[-7.4, -0.1]$, task-cluster bootstrap, 1,000 draws; Table 11). This rules out a large self-preference, but the test is weak: a single model generated every AI response, so generator identity and response quality cannot be separated.

Table 11: Leniency shift (v3 prompt, 48 baseline tasks). Error rates are against gold on the Sonnet 4.6 response (AI) and the reference answer (Ref.).

Judge	Error, AI	Error, Ref.	Shift (pp)
Sonnet 4.6	0.086	0.078	-4.0
Sonnet 4.5	0.113	0.131	-3.3
GLM-5.2	0.050	0.067	-1.0
DeepSeek V3.2	0.086	0.095	$+0.8$

D.9 Prompt and score sensitivity

Prompt. Table 12 ranks the eight families run under v1–v3 by alignment κ. Pairwise Kendall's τ is 0.07 (v1–v2), 0.43 (v1–v3) and -0.07 (v2–v3). No family holds a constant rank; Qwen 3.5 9B varies least (7th, 6th, 8th). For the production family, moving from v1 to v3 costs $\Delta\kappa = 0.021$.

Table 12: Family rank by alignment κ under each prompt.

Family	v1	v2	v3
Qwen 3.8 Max	1	7	1
Sonnet 4.6	2	1	4
DeepSeek V3.2	3	3	6
GLM-5.2	4	8	2
Sonnet 4.5	5	4	7
Qwen 3.8 27B	6	2	3
Qwen 3.5 9B	7	6	8
DeepSeek V4 Flash	8	5	5

Score. The benchmark reports a weighted task score (Eq. 5). Scored by each judge, the same Sonnet 4.6 responses (98 tasks, 1,863 criteria) average between 73.3% (Qwen 3.5 9B | v3) and 88.9% (Qwen 3.8 Max | v3); the production judge gives 84.6%, and the Naive baseline 100%. Among the ten configurations with the highest alignment κ, the range is still 82.8–88.9%. Task-level agreement between judges is moderate: the Spearman correlation of task scores between production and GLM-5.2 | v3 is 0.59, and between production and Qwen 3.8 Max | v3 it is 0.55. This is score sensitivity, not ranking sensitivity. Because every response came from one solver model, these data cannot show whether model *rankings* on the leaderboard would change under a different judge. Testing that requires re-scoring several solver models' responses with alternative judges.

D.10 Web search and cost

Web search. Four GPT-5.6-Terra configurations had a web-search tool enabled. Across 22,692 tool-enabled calls (Part A and stability), the judge never invoked it. Web-enabled configurations still cost more per criterion ($0.0084–0.0091 vs. $0.0047–0.0059), presumably from the tool definition's input tokens.

Cost. From Part A token counts at list prices, the production judge costs $0.0127 per criterion judgment and GLM-5.2 about six times less ($0.0021). These figures exclude reasoning tokens.

E CRITERION CATEGORIES AND κ-PARADOX METRICS

Taxonomy. Every criterion of the 1,000 scored tasks (16,695 instances) carries one of thirteen functional categories in four groups (Table 13). The taxonomy was built bottom-up in two cycles: LLM mining of a stratified sample, a fixed vocabulary, full classification, an independent governance review, one redesign, and full re-classification. Labels are LLM-proposed. A blind cross-check on a stratified sample of 390 criteria agreed on 73.8% ($\kappa = 0.717$), with disagreements concentrated on adjacent categories (e.g. grounding vs. coverage).

Table 13: Criterion category groups, with each category's dominant criterion style.

Group	Category	Style
Contextual fit & compliance	Sensitivity and compliance	mixed
	Learner and audience fit	holistic
	Setting and resource fit	holistic
Expert judgment & accuracy	Self-justification and transparency	checklist
	Diagnostic and evaluative accuracy	mixed
	Explanation clarity	holistic
	Pedagogical design quality	holistic
Content & structure	Quantitative constraint	checklist
	Sequencing and progression	mixed
	Required element coverage	checklist
	Prompt–data grounding	mixed
Interpersonal & dialogic	Conversational conduct	mixed
	Interpersonal tone and empathy	holistic

Agreement metrics under skew. Table 14 compares three configurations on the 650 gold verdicts. Accuracy, PABAK (Eq. 3) and Gwet's AC1 (Eq. 4) all rank Qwen 3.5 9B | v1 close to production. Only κ (Eq. 2) exposes that it reaches its accuracy by passing almost everything. We therefore report κ throughout, and read it within strata rather than across them.

Table 14: Agreement with gold under an 86.9% Pass base rate.

Configuration	Pass rate	Accuracy	PABAK	AC1	κ
Naive baseline	0.860	0.963	0.926	0.952	0.842
Sonnet 4.6 | v1 (production)	0.825	0.909	0.818	0.877	0.651
Qwen 3.5 9B | v1	0.957	0.894	0.788	0.874	0.347

F RATER SEVERITY MODEL

Units. Every analysis in this paper keys a rating unit on the text actually scored: (item, criterion, scored response), where the scored response is either the Sonnet 4.6 response or the reference answer disguised as the "AI response". Votes on different texts are never pooled, so a unit's majority is always a majority about one piece of text.

Model. We fit a binary logistic GLMM (M1) to all 17,093 post-exclusion Stage C criterion votes (142 raters, 442 tasks, 10,826 units), with fixed effects for the scored arm and for "duping" (the response a rater preferred in Stage B is not the one they scored in Stage C), and crossed random intercepts for rater r, task t, and unit u:

$$\operatorname{logit} \Pr(y_{rtu} = 1) = \beta_0 + \beta_{\text{AI}}\, x_u^{\text{AI}} + \beta_{\text{dup}}\, x_r^{\text{dup}} + a_r + b_t + c_u, \tag{1}$$

with $a_r \sim \mathcal{N}(0, \sigma^2_{\text{rater}})$, and likewise for b_t, c_u. The model is fitted by variational Bayes (`statsmodels BinomialBayesMixedGLM`). Variance shares use the logistic residual $\pi^2/3$ on the latent scale (Table 15). The scored arm has no detectable effect ($\hat{\beta}_{\text{AI}} = 0.021$, SD 0.035), consistent with the near-identical pass rates in Table 5. Two intercept-only linear probability models give a cross-check on the probability scale: rater 5.5%, unit 34.9%.

Table 15: Latent-scale variance decomposition of M1 (17,093 votes).

Component	Variance	Share
Rater (a_r)	0.764	10.1%
Unit (c_u)	3.500	46.3%
Task (b_t)	0.012	0.2%
Residual ($\pi^2/3$)	3.290	43.5%

Rater offsets. Posterior-mean rater intercepts $\hat{a}_r$ range from -2.54 to $+2.40$ logits. The strictest rater passes 36.8% of their 68 votes, and the most lenient passes 99.2% of their 127 votes.

Correction rule. For each observed vote we reconstruct its conditional latent residual $\hat{\varepsilon} = \mathbb{E}[\varepsilon \mid y, \eta]$ (a truncated-logistic mean, where η is the linear predictor), and re-threshold the severity-free latent $\eta - \hat{a}_r + \hat{\varepsilon}$ at zero. Because $\hat{\varepsilon}$ keeps each vote on its observed side of the threshold, only a severe rater's Fails can flip to Passes and only a lenient rater's Passes can flip to Fails. The rule flips 236 votes (1.4%). A quantile-matching alternative instead gives each rater their severity-free expected pass count $\sum \sigma(\eta - \hat{a}_r)$ and assigns Passes in descending order of severity-free propensity. It flips about 1,442 votes and over-corrects: it drives the human ceiling to $\kappa = 0.948$ (Frame A) and 0.965 (Frame B). This erases the within-unit variation the ceiling exists to measure, so we do not use it.

Judge offsets. Each judge's verdicts on the 169 judge-comparable units are treated as one additional rater. We take each unit's mean severity-free human score s_u, then solve for the $\hat{u}_J$ at which $\text{mean}_u\, \sigma(s_u + \hat{u}_J)$ equals the judge's observed pass rate. The same latent-residual rule is then applied with $\eta_J = s_u + \hat{u}_J$. Production (Sonnet 4.6 | v1): pass rate 0.799, $\hat{u}_J = -0.183$, 0/169 verdicts flipped. Open-weight (GLM-5.2 | v3): pass rate 0.846, $\hat{u}_J = +0.368$, 0/169 flipped. Stripped and raw judge verdicts are therefore identical.

G LEAVE-ONE-OUT CEILINGS

Procedure. For each unit with $k \geq 3$ raters, we hold out each rater in turn and compare their vote with the majority of the remaining $k - 1$ (Eq. 9). Comparisons where the subpanel ties are excluded. Agreement and Cohen's κ are computed over all remaining comparisons, and 95% CIs come from a unit-clustered bootstrap. κ is reported as undefined when either side is constant within a stratum. A judge enters as a held-out rater: its verdict is compared with every $k - 1$ human subpanel of the unit (*drop-one*), or once with the full k-rater majority (*full-k*).

Frames. Frame A pools all tasks and both scored arms. Frame B keeps only votes on the Sonnet 4.6 response, on the 98 judge-audit tasks, and is the frame on which judges are scored (Table 16).

Table 16: Leave-one-out agreement by frame, correction, and rater. Comparisons are counted after tie exclusion; correction breaks some subpanel ties (Frame A tie rate 13.9% → 11.1%; Frame B 9.4% → 9.0%), so corrected rows carry more comparisons.

Frame	Rater / votes	Units	Comp.	Agreement	κ [95% CI]
A	Human, raw	1,232	3,527	0.864	0.490 [0.428, 0.537]
A	Human, severity-corrected	1,232	3,641	0.893	0.559 [0.491, 0.609]
B	Human, raw	169	502	0.902	0.716 [0.590, 0.815]
B	Human, severity-corrected	169	504	0.931	0.788 [0.690, 0.862]
B	Production judge, drop-one	169	502	0.944	0.826 [0.708, 0.939]
B	Production judge, full-k	169	166	0.958	0.866 [0.767, 0.961]
B	Production judge vs. corrected	169	504	0.956	0.861 [0.742, 0.959]
B	GLM-5.2 | v3, drop-one	169	502	0.920	0.723 [0.586, 0.856]
B	GLM-5.2 | v3, full-k	169	166	0.934	0.767 [0.628, 0.884]
B	GLM-5.2 | v3 vs. corrected	169	504	0.933	0.758 [0.620, 0.880]

Categories. The per-category human ceiling (Frame A) and the production judge's drop-one κ (Frame B) are reported together in Table 2. Because the two columns come from different frames, per-category judge–human gaps are indicative only; per-category judge estimates rest on 5–46 units.

Co-location of disagreement. Structural validity asks whether the judge disagrees with the panel where humans also disagree. At the criterion level, the test has almost no power: only 10 of the 650 judge-matched gold units carry any dissent, and the correlation between practitioner agreement and mean judge error across all 32 judge configurations is null (Spearman $\rho = 0.038$, $p = 0.34$, $n = 650$). At the category level, the production judge's disagreement rate against gold and the human LOO disagreement rate are rank-correlated across the thirteen categories (Spearman $\rho = 0.834$, $p < 0.001$). Judge disagreement is highest for self-justification and transparency (55.6% vs. human 25.0%) and sensitivity and compliance (29.7% vs. 23.8%), which are also the two categories with the highest human disagreement.

G.1 Removing contested gold units

Gold units already require ≥67% agreement among ≥2 raters, so only 10 of the 650 judge-matched units carry any dissent. Removing those 10 leaves the leaderboard essentially unchanged: across all 32 judge configurations, Spearman $\rho = 0.9989$ between the full and filtered κ rankings, no configuration moves more than two ranks, the production judge stays 4th, and the top three are preserved (Qwen 3.8 Max | v3, GLM-5.2 | v3, Qwen 3.8 Max | v1). Because the gold filter has already removed most disagreement, this is a weak test of robustness to contested labels.

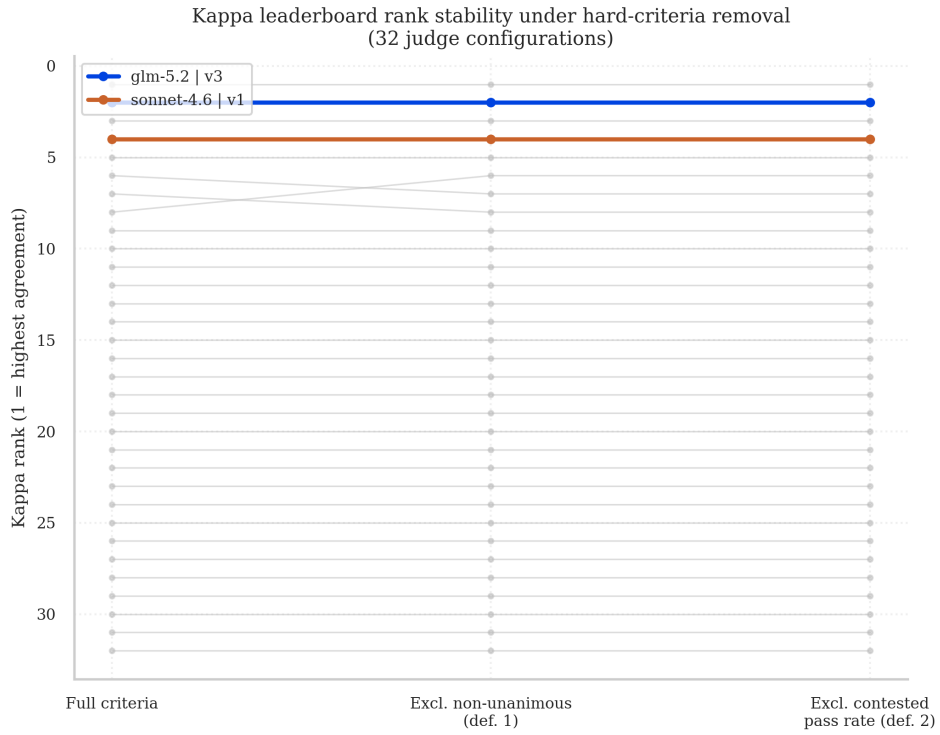

Figure 6: Judge κ rank under the full gold set and two definitions of contested units (non-unanimous; contested pass rate). All 32 judge configurations.

H EQUATIONS

This appendix collects the estimators used in the paper, excluding the severity model (Eq. 1, Appendix F). Throughout, n is the number of gold verdicts compared, p_o the observed proportion of agreement, and π_J, π_H the judge and human Pass rates.

Cohen's κ. Chance-corrected agreement between judge and gold, with chance agreement taken from the two marginals:

$$\kappa = \frac{p_o - p_e}{1 - p_e}, \qquad p_e = \pi_J \pi_H + (1 - \pi_J)(1 - \pi_H). \tag{2}$$

Under a skewed base rate p_e approaches 1, so a small shortfall in p_o produces a large drop in κ (the κ paradox; Appendix E).

PABAK. Prevalence- and bias-adjusted κ fixes chance agreement at $1/2$, removing the dependence on the marginals:

$$\mathrm{PABAK} = 2p_o - 1. \tag{3}$$

Gwet's AC1. AC1 uses a chance term that shrinks, rather than grows, under skew:

$$\mathrm{AC1} = \frac{p_o - p_e^{\gamma}}{1 - p_e^{\gamma}}, \qquad p_e^{\gamma} = 2\bar{\pi}(1 - \bar{\pi}), \quad \bar{\pi} = \tfrac{1}{2}(\pi_J + \pi_H). \tag{4}$$

Weighted task score. L2-Bench scores a response on task t from its criterion verdicts $v_c \in \{0, 1\}$ and integer weights w_c (negative for penalty criteria). Penalties subtract when triggered, the score is normalised by the positive weight available, and it is floored at zero:

$$S_t = \max\left(0,\ \frac{\sum_{c \in t} w_c\, v_c}{\sum_{c \in t} \max(w_c, 0)}\right). \tag{5}$$

Flip rate and Wilson interval. For a configuration, a criterion is unstable if its $k = 10$ resampled verdicts are not all identical; the flip rate is $\hat{p} = m/N$ for m unstable of N criteria. With $z = 1.96$, its 95% Wilson interval is

$$\frac{\hat{p} + \frac{z^2}{2N} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{N} + \frac{z^2}{4N^2}}}{1 + \frac{z^2}{N}}. \tag{6}$$

The normal-approximation half-width $z\sqrt{\hat{p}(1-\hat{p})/N}$ was used for planning (Appendix D.2).

Detection probability. A criterion whose verdict is Pass with probability $1 - q$ on each independent resample is observed as unstable in k resamples with probability

$$P_{\text{detect}}(q, k) = 1 - q^k - (1 - q)^k. \tag{7}$$

Tasks needed for a target standard error. With s the between-task standard deviation of a task-level score shift, the number of tasks needed for standard error SE is

$$n_{\text{tasks}} = \left(\frac{s}{\mathrm{SE}}\right)^2. \tag{8}$$

Leave-one-out agreement. For a unit u with $k_u \geq 3$ votes $y_{u1}, \ldots, y_{uk_u}$, each held-out vote is compared with the majority of the rest, skipping ties:

$$\mathrm{Agr}_{\mathrm{LOO}} = \frac{\sum_u \sum_i \mathbb{1}[y_{ui} = \mathrm{maj}(\mathbf{y}_{u,-i})]}{\sum_u \sum_i \mathbb{1}[\mathrm{maj}(\mathbf{y}_{u,-i}) \text{ defined}]}, \tag{9}$$

with κ computed over the same (held-out, majority) pairs. A judge enters as vote i against each $(k_u - 1)$-rater subpanel (drop-one) or against the full k_u-rater majority (full-k).

Leniency shift (difference-in-differences). For judge J, with π the Pass rate on a given text, the leniency shift is

$$\Delta_J = \left(\pi_J^{\mathrm{AI}} - \pi_H^{\mathrm{AI}}\right) - \left(\pi_J^{\mathrm{Ref}} - \pi_H^{\mathrm{Ref}}\right), \tag{10}$$

and the self-preference contrast is the mean Δ_J over Claude judges minus the mean over non-Claude judges, with a task-cluster bootstrap CI (Appendix D.8).

I MINIMUM VIABLE JUDGE ASSESSMENT

Table 17: Minimum Viable Judge Assessment (MVJA) for **rubric-based evaluations in "hard" contexts**. *Outputs* lists the quantities it contributes to a judge disclosure report.

Tier	Check	Rationale	Outputs
Required	Stratified agreement (R1)	Under skew, aggregate agreement measures the base rate; read κ as diagnostic, not threshold.	Label prevalence; κ by stratum
	Content-blind baseline (R1)	Shows how much apparent alignment the rubric gives away for free.	Baseline accuracy and κ
	Human ceiling (R2)	Judge–human agreement is uninterpretable without it; score the judge as a held-out rater.	LOO κ [95% CI], human and judge
	Judge stability (R3)	Necessary, not sufficient; cannot be assumed.	Flip rate [Wilson CI]; flip direction
	Judge configuration (R3)	"Best judge" is partly a property of the prompt.	Model, prompt, decoding; score under an alternative judge
	Rater-level data (R4)	Neither judge nor gold can be validated without it.	Raw votes with pseudonymous rater IDs
Recommended	Severity correction (R2)	Required if the judge beats the raw ceiling: a flag, not headroom.	Rater variance share; corrected ceiling; judge offset
	Error co-location (R2)	Does the judge err where humans err?	Spearman ρ across categories
	Self-preference (R3)	Inter-judge agreement alone cannot detect it.	Leniency shift (DiD) [CI]
	Per-criterion ceilings (R3)	Criteria experts cannot apply are beyond any judge's repair.	Human κ by criterion type
	Qualitative comparison (R1)	Comments reveal disagreement that votes conceal.	Coded rater comments
As appropriate	Reasoning transparency	Hedging flags the least reliable CoT verdicts.	Hedged-verdict rate
	Ranking sensitivity	Does leaderboard order survive a judge swap?	Kendall's τ across judges
	Adversarial perturbation	Do meaning-preserving rewrites flip verdicts?	Flip rate under rewrite
	Weight sensitivity	Do rankings survive admissible reweightings?	Rank stability