FROM LLM JUDGE SCORES TO EVALUATION DESIGN: A GENERALIZABILITY-THEORY FRAMEWORK

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges provide task-specific feedback for selecting responses, comparing models, and training policies. Their scores combine stable differences between responses with variation introduced by the judging procedure. We present a generalizability-based framework that profiles each fixed judge, projects grading precision from a pilot, and checks whether a protocol improves decisions against an external criterion. We instantiate it with twelve open-weight judges grading the same responses on three sources with correctness, rubric-quality, or human-preference references. The crossed design varies two close prompt paraphrases and three generations. We find a small pilot per source predicts generation-noise rankings on separate queries (correlations 0.93–0.95) and approximate grading disagreement. A second grading under unchanged wording improves aggregate ordering point estimates by 0.89–2.95 percentage points. An independent-call diagnostic shows that averaging can strengthen both reference-consistent and reference-opposed preferences. The framework yields judge-specific precision and validity profiles under declared conditions and separates protocol planning from judge selection. Code repository: https://anonymous.4open.science/r/RV-Judge-3C82/README.md

1 INTRODUCTION

The ability to express evaluation criteria in natural language makes LLM judges a flexible component of model development. The same interface can assess factual accuracy, instruction following, or safety, providing feedback for model comparison, response selection, and policy training (Liu et al., 2023; Wen et al., 2026; Gunjal et al., 2026). Each application requires choosing a judge and a grading protocol. Evaluating a judge therefore requires identifying where score variation comes from and checking whether the resulting decisions agree with the intended criterion.

Recent work measures how judge scores change with prompt wording, repeated sampling, and judge identity (Bhat & Varma, 2026; Verma et al., 2026; Zhao et al., 2026; Choi et al., 2026; Messing, 2026). Generalizability theory puts measured sources of variation on a common scale and projects the precision of alternative designs (Cronbach et al., 1972; Brennan, 2001). The intended decision determines which of those sources count as measurement error in contrast to valid signal.

A practitioner can specify the target score and nuisance conditions, collect repeated gradings of matched responses, and estimate which sources of error a feasible allocation would reduce. Precision alone cannot establish validity: a stable preference can oppose the reference, and variation shared by candidates need not affect their ordering. The framework therefore pairs response-score and within-query contrast profiles with reference-based decision checks. We instantiate it with twelve open-weight judges on identical candidate responses under two template paraphrases and three generations, using source-native correctness, human rubric scores, and human preferences as criteria.

We provide (i) a reusable workflow that links the measurement target, judge-specific variance components, D-study allocations, and external-criterion validation; (ii) an empirical test of whether a small pilot forecasts grading on separate queries; and (iii) matched comparisons of grading allocations and score readouts, with independent-call diagnostics of when averaging helps or harms reference-based ordering.

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.

2 Related Work

Where the criterion comes from. The usual answer is human labels on a sample, with a judge scored by its agreement with them (Zheng et al., 2023; Calderon et al., 2025). Agreement is easy to compute and hard to interpret, and choosing a judge by it can select one that serves the intended decision worse (Guerdan et al., 2025). A second answer is a panel, on the assumption that independent judges cancel each other's errors (Verga et al., 2024). The assumption fails when the errors are correlated, which both theory and measurement suggest they are (Balasubramanian et al., 2026; Kohli, 2026), though not so far as to make additional judges useless (Li, 2026). A third answer is a verifiable reference, available only where correctness can be checked (Krumdick et al., 2025). Reward-model benchmarks also show that offline ranking and downstream policy training require separate validation (**?**). Our fixed-candidate ordering experiment addresses the former use, not policy training.

What moves a judge's score. A large body of work perturbs one thing at a time and reports how often the verdict changes: the order in which candidates are presented (Anghel et al., 2025), the wording of the prompt (Bhat & Varma, 2026), superficial restyling of the response (Eiras et al., 2025), sustained argumentative pressure (Zhao et al., 2026), kinship between judge and candidate (Spiliopoulou et al., 2025), how the score is read off the model (Xie et al., 2026), and the sampling draw itself, which temperature zero does not reliably pin down (Lau, 2026). Larger audits confirm the picture at scale and add that judge rankings do not transfer between benchmarks (Norman et al., 2026). Repeated-voting studies also measure the benefit and limits of additional judgments (**?**).

Measurement models put the effects on one scale. Generalizability theory splits a score's variance among the conditions that produced it, then projects how much error a proposed design would remove (Cronbach et al., 1972; Brennan, 2001). It reached language models through educational measurement, where the object scored is a human essay and the criterion a human rating (Song et al., 2025; Wang et al., 2025a), and related psychometric tools have since been applied to judges directly (Choi et al., 2026; Żatuchin, 2026). Closest to this paper, Messing (2026) runs a crossed decomposition and a decision study on judge pipelines, prescribes averaging over judges as the remedy, and validates the projections against answer keys and human votes.

3 A Framework for Focused Measurement Studies

3.1 Measurement target and experimental design

We treat evaluation as a measurement procedure: specify what a score is meant to express, which conditions can change it, and what decision it will support. Let x denote a query and $c \in \mathcal{C}_x$ a candidate response. The scored item is $i = (x, c)$. A study specifies the construct, such as factual accuracy, the scoring conditions that may vary, and the decision the scores support. A reference answer supplied to the judge is part of its input; it is distinct from the criterion used to assess the judge.

The generalization universe specifies which scoring conditions may vary and how they are weighted. The expected score over those conditions is the universe score; its interpretation depends on the rubric and the chosen conditions (Brennan, 1992; 2003). Generalizability theory (GT) distinguishes a *G-study*, which estimates sources of score variation, from a *D-study*, which projects measurement

Target use	Focused study	Diagnostic output
Compare judges	Grade identical responses under repeated conditions.	Judge-specific variance profiles and criterion accuracy.
Evaluate candidate ordering	Preserve candidate comparisons within queries.	Contrast precision and reference ordering agreement.
Allocate grading calls	Estimate components in a repeated pilot.	Projected error and measured accuracy of feasible allocations.

Table 1: The intended use determines the measurement target, conditions, and validation criterion.

error under a proposed design. A practitioner holds some conditions fixed and varies others to study the relevant sources of error (Table 1). The resulting audit reports a judge-specific variance profile, response-score and decision-contrast precision, projected error under feasible allocations, and accuracy against a task-matched criterion. A new application can change the facets and criterion while following the same sequence; conclusions apply only to conditions actually sampled. A study can vary prompt wording, sampling settings, or other relevant conditions while holding the remaining instrument settings fixed. A nuisance facet changes the scoring conditions without intentionally changing the evaluation criterion; sensitivity to such a facet measures a lack of robustness. We compare judges through separate profiles on matched material.

3.2 Separating stable differences from measurement error

For a fixed judge, let z index the scoring conditions declared by the study, including any generation randomness. The score of item i can be written as

$$s_{iz} = \mu + \underbrace{I_i}_{\text{stable item effect}} + \underbrace{C_z}_{\text{shared condition effects}} + \underbrace{\varepsilon_{iz}}_{\text{interaction and residual}}, \qquad u_i = \mathbb{E}_z[s_{iz}] = \mu + I_i. \quad (1)$$

Effects are centered over the declared item population and condition universe. C_z collects shifts shared across items; ε_{iz} collects item–condition interactions and residual variation. These are groups of effects, not three independent draws: their component structure and dependence follow the experimental design. A persistent item–template interaction, for example, does not become independent when the same template is repeated.

The variance $\sigma_I^2 = \mathrm{Var}_i(u_i)$ captures differences between items that persist across the declared scoring conditions. A consistently high score for a wrong response is therefore stable too: this quantity measures reproducibility, not correctness. The G-study estimates the variance of item effects, condition effects, and their interactions under the specified crossed or nested design. Finite-repeat corrections keep sampling noise from being misread as a stable difference or a persistent wording effect.

For an operational design $\mathcal{D}$, let $\bar{s}_{i\mathcal{D}}$ denote its reported score. Absolute measurement error and dependability are

$$\sigma^2_{\Delta,\mathcal{D}} = \mathbb{E}_{i,\mathcal{D}}(\bar{s}_{i\mathcal{D}} - u_i)^2, \qquad \Phi_{\mathcal{D}} = \frac{\sigma_I^2}{\sigma_I^2 + \sigma^2_{\Delta,\mathcal{D}}}, \qquad \mathrm{SEM}_{\mathcal{D}} = \sqrt{\sigma^2_{\Delta,\mathcal{D}}}. \quad (2)$$

The expectation includes the design's sampling of conditions. We consider designs whose scores are unbiased for the declared universe score in expectation. Dependability expresses error relative to item variation; SEM expresses it in score units. Relative reliability excludes shared score shifts when only ordering matters. For candidate selection, we instead apply the decomposition directly to within-query score differences, so common shifts cancel. Neither coefficient nor SEM measures criterion accuracy.

D-studies compare measurement allocations. A D-study projects how an allocation changes the estimated error components. In a balanced crossed design with mutually uncorrelated random effects and independent sampling of exchangeable facet levels, the standard expression is

$$\sigma^2_{\Delta,\mathcal{D}} = \sum_{H \neq \{I\}} \frac{\sigma_H^2}{\prod_{f \in H \setminus \{I\}} n_f}. \quad (3)$$

Here H indexes an effect or interaction, and n_f is the number of levels averaged for facet f; the sum covers all error components. Each component decreases only with replication of its own facets. Fixed or finite facets, nested draws, and unequal allocations require the corresponding design-specific calculation. Thus a D-study compares feasible allocations at a given budget or precision requirement; it does not prescribe the same averaging rule for every study.

Instantiation in this study. For each fixed judge, the conditions z in Equation 1 are two equally weighted templates and independent generations within each item–template cell. Let $\sigma^2_{\mathrm{word}} = \mathbb{E}_{i,t}(\mathbb{E}_r s_{itr} - u_i)^2$ measure persistent deviation from the two-template universe score, including

both a shift shared across items and item–template interaction. Let $\sigma^2_{\text{gen}} = \mathbb{E}_{i,t}\,\text{Var}_r(s_{itr})$ measure generation variation within a cell. Thus the one-grading absolute error in Equation 2 is $\sigma^2_{\Delta,1} = \sigma^2_{\text{word}} + \sigma^2_{\text{gen}}$, and its dependability is $\Phi_1 = \sigma^2_I/(\sigma^2_I + \sigma^2_{\text{word}} + \sigma^2_{\text{gen}})$. Averaging K independent calls under one uniformly chosen template leaves $\sigma^2_{\text{word}} + \sigma^2_{\text{gen}}/K$; balancing calls across both templates cancels their deviation from the defined two-template mean and leaves σ^2_{gen}/K. These are finite-template instances of the D-study, not claims about unseen wordings. For candidate selection we apply the same calculation to within-query score differences, where shifts shared by candidates cancel. Appendix A gives the estimators and unequal-allocation formula. Stable score variation can still include reproducible mistakes, so criterion accuracy is checked separately.

3.3 Validating precision and decision accuracy

We check two claims separately. A precision claim asks whether independently obtained scores agree as the pilot predicts. A decision claim asks whether those scores order the candidates according to an external correctness or human reference. Judges can agree with themselves or one another while sharing the same mistake, so the first check cannot replace the second. We estimate the components on calibration queries, then predict disagreement between independent gradings on separate test queries. The predicted squared differences are $2\sigma^2_{\text{gen}}$ for the same template, $2\sigma^2_{\text{gen}} + 4\sigma^2_{\text{word}}$ across templates, and σ^2_{gen} for two independent averages each covering both templates. Comparing these forecasts with observed disagreement tests whether the pilot estimates transfer to new queries; it does not test whether the universe score is correct.

External labels answer the separate validity question (Table 2). We first average eligible pairs within each query, then queries within a source. Aggregate figures use the equally weighted mean of the three source accuracies. This is a declared summary of ordering decisions, not a pooled rating scale. Human annotations are operational references, not infallible truth.

To diagnose the effect of replication, we separate the calls that estimate a judge's direction from those that measure the benefit of averaging. Four calls per response estimate whether its within-query score difference agrees with the reference; two unused calls under the same template compare a single grading with their average. The direction estimate weights the two wordings equally. Appendix C specifies the split and integration.

4 Evaluation Protocol

Each fixed judge grades candidates pointwise, one response per call, under two templates and three independent generations per response–template cell. Calibration queries estimate variance components; nonoverlapping test queries assess forecasts and reference-based ordering. The same response set and template allocation are used for every judge.

Source	Available criterion	Validation metric
JudgeBench	Correctness-labelled pair winner	Selection accuracy
BiGGenBench	Human rubric rating, 1–5	Ordering agreement
HelpSteer3	Signed human preference, with ties	Preference agreement

Table 2: Source-native criteria.

JudgeBench supplies two answers per query, BiGGenBench four, and HelpSteer3 two (Tan et al., 2025; Kim et al., 2025; Wang et al., 2025b). Each source contributes 16 calibration and 100 test queries, selected independently of label values. JudgeBench strata are knowledge, LiveBench math, LiveBench reasoning, and LiveCodeBench; BiGGenBench uses eight task categories; HelpSteer3 uses code, general, multilingual, and STEM.

Twelve open-weight judges produce 9,216 calibration and 57,600 test gradings, with no missing or unparsed records. Each call generates a rationale, then a constrained digit from 0 to 5. Sampling and seed rules are fixed across the recorded calls. Scores are normalized to $[0, 1]$. The probability-weighted readout uses the same call and its six recorded digit log probabilities.

The six recorded gradings per candidate let us replay nine distinct allocations of up to three calls per wording. For example, two calls can repeat one wording or cover both. We average the scores within an allocation, then compare candidates. The candidates in a query use the same wording allocation, but their generation draws are independent. We average each possible eligible subset of stored calls for each candidate and then compare all candidate-subset combinations; stored repeat indices do not pair candidates. These replays compare designs using the available outputs and do not create new independent observations.

An exploratory diagnostic partitions the same-wording two-call accuracy gain by the first call's absolute candidate-score difference: an integer-score tie, one rating step, or at least two steps. For each candidate, we integrate distinct first and second generations independently. The three contributions sum to the original two-call gain under the same pair and query weights.

We fit variation components separately for each source. The displayed judge profile averages source-normalized variance shares; aggregate standard error of measurement (SEM) takes the square root after averaging source error variances. For decisions, we average eligible pair outcomes within queries, then queries within sources, and give the three sources equal weight. Calibration and test queries are bootstrapped independently. Highlighted design contrasts use simultaneous intervals across the twelve judges; other intervals are pointwise. All allocation analyses replay saved outputs, so the conclusions are retrospective and conditional on these judges, references, and two wordings.

5 Results

5.1 Judge profiles distinguish accuracy and measurement error

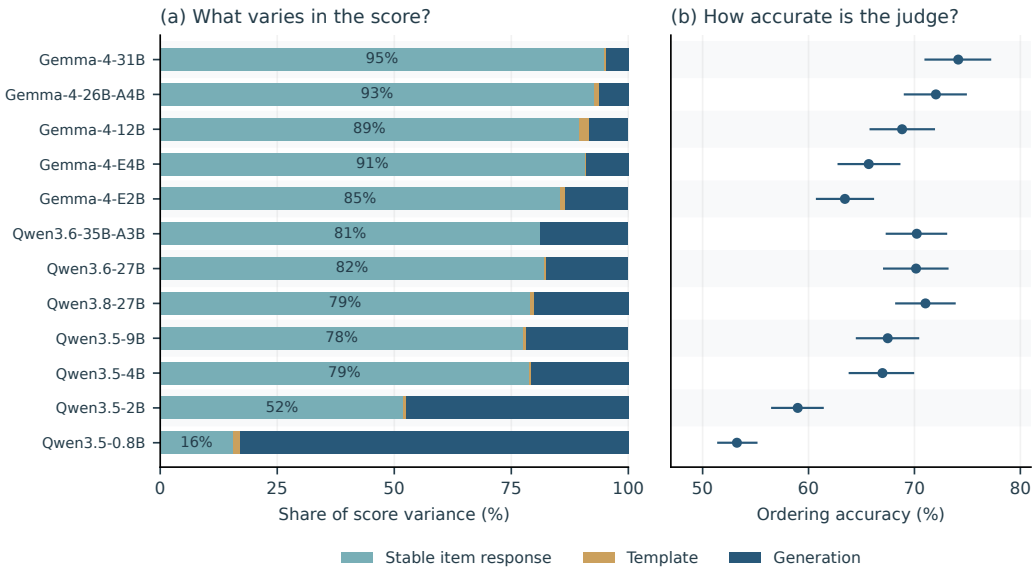


Figure 1: **Compare the instrument before allocating calls.** Left: stable item differences, template sensitivity, and generation variance for twelve judges. Shares are normalized within each source and equally averaged. Right: single-grading ordering accuracy across the three sources, with pointwise 95% query-bootstrap intervals.

Figure 1 shows that judges differ in stable response differentiation and generation noise. The source-specific results in Table 3, rather than the displayed cross-source summaries, determine whether a judge is suitable for a particular criterion.

Judge	JudgeBench Acc. ↑	JudgeBench Φ	BiGGenBench Order ↑	BiGGenBench Φ	HelpSteer3 Agree ↑	HelpSteer3 Φ
Gemma-4-31B	**79.1**	**0.920**	72.0	**0.966**	**71.2**	**0.959**
Gemma-4-26B-A4B	<u>72.8</u>	<u>0.890</u>	**73.2**	<u>0.958</u>	70.0	<u>0.932</u>
Gemma-4-12B	69.5	0.829	71.2	0.955	65.8	0.899
Gemma-4-E4B	64.4	0.865	72.7	0.934	60.0	0.922
Gemma-4-E2B	61.1	0.869	70.6	0.869	58.6	0.824
Qwen3.6-35B-A3B	69.3	0.706	71.1	0.901	70.2	0.829
Qwen3.6-27B	70.3	0.714	70.9	0.922	69.3	0.823
Qwen3.8-27B	68.9	0.683	<u>73.0</u>	0.908	**71.2**	0.779
Qwen3.5-9B	66.3	0.677	71.8	0.895	64.3	0.755
Qwen3.5-4B	61.7	0.717	71.7	0.879	67.5	0.767
Qwen3.5-2B	55.2	0.484	68.1	0.636	53.6	0.441
Qwen3.5-0.8B	54.4	0.186	53.4	0.201	51.9	0.080

Table 3: **Accuracy and dependability answer different questions.** Single-grading accuracy/agreement (percent) and response-score dependability for each source. Bold and underline indicate the best and second-best point estimates. Source-native labels validate accuracy.

Gemma-4-31B has strong agreement and dependability on JudgeBench and HelpSteer3; on BiGGenBench, Gemma-4-26B-A4B and Qwen3.8-27B have the highest ordering point estimates. Gemma-4-E4B has higher response-score dependability than Qwen3.8-27B on every source, yet lower reference ordering agreement on JudgeBench and HelpSteer3. This confirms the previously documented separation of consistency and validity (Norman et al., 2026; **?**). Response-score Φ includes between-query variation; within-query contrast precision is the matched precision measure for candidate selection. Qwen3.5-0.8B is near chance on all three sources. BiGGenBench separates this roster weakly: ten of twelve point estimates lie between 70.6% and 73.2%.

5.2 A SMALL PILOT PREDICTS GRADING VARIABILITY

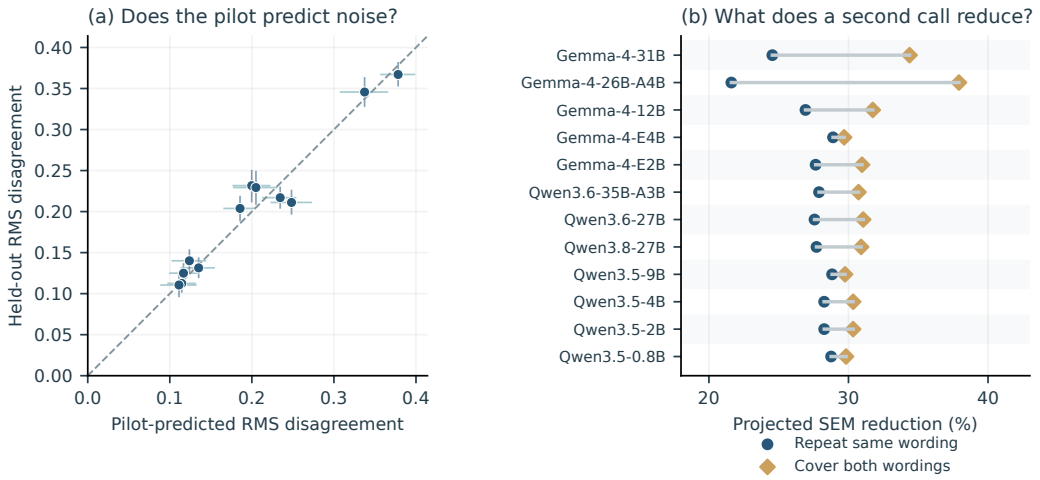


Figure 2: **From a pilot to a grading budget.** Left: pilot-predicted versus held-out RMS disagreement between independent same-template gradings; each point is a judge, with pointwise 95% intervals for both quantities. Right: projected score-SEM reduction from a second grading, either under the same wording or covering both observed wordings.

A 16-query pilot per source broadly ranks generation noise on separate test queries (rank correlations 0.93–0.95 across the full roster). It is less decisive for validity: Qwen3.8-27B ranks eighth in calibration single-call macro accuracy (66.2%) but third on test (71.0%). Pilot precision screening therefore does not replace task-matched reference labels for judge selection.

Pilot forecasts of grading disagreement on test queries are approximate (Figure 2); aggregate observed-to-predicted RMS ratios range from 0.83 to 1.23. Source-specific errors are larger (ratios 0.69–1.54), so the pilot is useful for planning and comparison. Full checks are in Appendix B.

5.3 What a second grading changes—and what it cannot fix

On calibration queries, a second same-template grading projects a 21.6–28.9% reduction in score SEM. On test queries, ordering point estimates improve by 0.89–2.95 points (Figure 3, left).

Repetition is no substitute for judge selection. An independent-call diagnostic estimates each judge's preferred direction with four calls and evaluates the effect of averaging on two separate calls. Averaging improves reference-consistent comparisons but tends to strengthen reference-opposed ones (Figure 3, right).

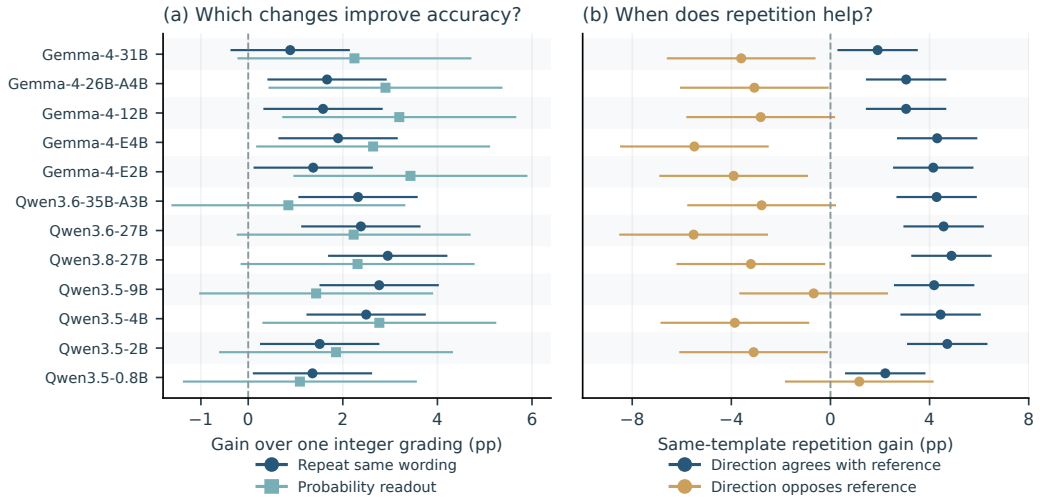


Figure 3: **Reducing variation can strengthen correct and incorrect judgments.** Left: accuracy changes from same-template repetition and from the probability-weighted readout, each relative to one integer-score grading. Right: the repetition effect on strata defined using four separate calls. Both panels evaluate the same two-call intervention; the right panel conditions on an estimated direction. Bars are simultaneous 95% intervals within the stated comparison families (Appendices B–D).

Using both observed wordings cancels their finite-universe wording deviation, whereas repeating one wording only reduces generation noise. Yet the additional ordering-accuracy benefit of wording coverage over same-wording repetition is unresolved: judge-level differences range from -0.41 to $+0.89$ points, and no simultaneous interval excludes zero. At six calls, Qwen3.6-27B, Gemma-4-26B-A4B, and Qwen3.8-27B reach 76.2%, 76.0%, and 75.5% macro ordering accuracy, respectively, versus 74.1% for one Gemma-4-31B call; Gemma-4-31B reaches 76.2% with six calls. The weakest two judges remain at 55.8% and 61.3% after six calls. A probability-weighted readout from the same call improves each judge's exact-tie ordering point estimate by 0.85–3.43 points relative to its integer score.

6 Discussion and Limitations

The framework gives practitioners a sequence for evaluating a judge under declared conditions: define the decision and criterion, measure the relevant variance components, compare feasible protocols, and validate the resulting decisions. In this study, that sequence motivates task-matched references for judge selection and a repeated pilot for precision planning.

The pilot's noise ordering and approximate disagreement forecasts transfer to separate queries, showing that the variance estimates contain information beyond a fitting-sample identity. Yet the

extra accuracy from a second call is small, balanced wording has no established additional ordering benefit, and independent calls show that averaging can reinforce a reference-opposed preference.

The observed hard-score gain is concentrated at first-call ties. This makes score readout and decision margins relevant to a call budget, while score and contrast SEM still address precision rather than correctness. A practitioner choosing between one probability-weighted call, repeated integer calls, and a tie-triggered rule must measure criterion accuracy and call cost at the intended tolerance.

The study covers twelve checkpoints from two model families, two templates, and three generations on fixed benchmark responses. Wording is a nuisance facet, but two templates do not establish robustness to arbitrary prompts. Generation dominance is conditional on the recorded sampling settings, the rationale cap, and fixed pointwise scoring; the high cap-hit rate may alter judge behavior. Source-macro accuracy weights three criteria equally; another application can change the ranking. References are not independently adjudicated, and all query-bootstrap intervals condition on the tested roster and annotations.

Forecast and decision diagnostics use saved outputs on disjoint queries or calls, not a prospective deployment. Generation independence and an unchanged scoring process are modelling assumptions. Call counts are gradings per response, not equal compute across models or thinking settings. No result establishes an optimal temperature, an unsupervised accuracy selector, or downstream robustness under policy optimization. Using these scores for policy training would require an on-policy test of reward utility and of whether optimization changes the judge's error pattern. We leave these extensions for future work.

7 Conclusion

We present a generalizability-based framework for evaluating fixed judges as measurement instruments. It links the declared scoring universe and intended decision to a variance profile, allocation forecasts, and external-criterion validation. In our pointwise instantiation, a small pilot recovered relative generation noise and approximately forecast grading disagreement on separate queries. A second same-template grading increased point-estimate ordering accuracy by 0.89–2.95 points across judges; covering both tested templates reduced projected score error without a demonstrated additional ordering gain. Dependability and criterion agreement diverged, and averaging sometimes reinforced reference-opposed preferences. Task-matched references should guide judge selection, while a repeated pilot can guide precision planning. The framework can accommodate other facets and decisions when they are specified and sampled; these findings concern the judges, templates, and references tested here. Our results show that reliability and validity, while complementing each other, must remain separate assessments. A stable but wrong preference still requires an external criterion to detect.

AI Use Statement

LLM assistance was used for statistical exploration, analysis code, figures, and drafting. The authors are responsible for the final manuscript and its claims.

Ethics Statement

This study analyzes public benchmark responses and existing annotations. It collects no new human data.

Reproducibility Statement

Our code is open source, and the datasets used in this study are public.

REFERENCES

Cătălin Anghel, Andreea Alexandra Anghel, Emilia Pecheanu, Adina Cocu, Adrian Istrate, and Constantin Adrian Andrei. Diagnosing bias and instability in LLM evaluation: A scalable pairwise meta-evaluator. *Information*, 16(8):652, 2025. doi: 10.3390/info16080652.

Krishnakumar Balasubramanian, Aleksandr Podkopaev, and Shiva Prasad Kasiviswanathan. Dependence-aware label aggregation for LLM-as-a-Judge via Ising models, 2026. arXiv:2601.22336.

Savita Bhat and Vasudeva Varma. All prompts are created equal? evaluating robustness of LLM judges against non-adversarial prompt variations. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 38730–38745, 2026. doi: 10.18653/v1/2026.findings-acl.1929. URL https://aclanthology.org/2026.findings-acl.1929/.

Robert L. Brennan. Generalizability theory. *Educational Measurement: Issues and Practice*, 11(4):27–34, 1992. doi: 10.1111/j.1745-3992.1992.tb00260.x. URL https://ncme.org/wp-content/uploads/2025/10/Module-14-Generalizability-Theory-Brennan-Winter-1.pdf.

Robert L. Brennan. *Generalizability Theory*. Springer, New York, 2001.

Robert L. Brennan. Coefficients and indices in generalizability theory. Technical Report 1, Center for Advanced Studies in Measurement and Assessment, University of Iowa, 2003. URL https://education.uiowa.edu/sites/education.uiowa.edu/files/2026-04/casma-research-report-1-archived.pdf.

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 *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 16051–16081, 2025. doi: 10.18653/v1/2025.acl-long.782. URL https://aclanthology.org/2025.acl-long.782/.

Junhyuk Choi, Sohhyung Park, Chanhee Cho, Hyeonchu Park, and Bugeun Kim. Diagnosing the reliability of LLM-as-a-Judge via item response theory. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2026. arXiv:2602.00521.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. Wiley, New York, 1972.

Francisco Eiras, Eliott Zemour, Eric S. Lin, and Vaikkunth Mugunthan. Know thy judge: On the robustness meta-evaluation of llm safety judges, 2025. URL https://openalex.org/W4416113418. arXiv:2503.04474.

Luke Guerdan, Solon Barocas, Kenneth Holstein, Hanna Wallach, Zhiwei Steven Wu, and Alexandra Chouldechova. Validating LLM-as-a-Judge systems under rating indeterminacy. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/a309239c11a28c597d050bd4a1752d32-Abstract-Conference.html.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains. In *International Conference on Learning Representations (ICLR)*, 2026. URL https://arxiv.org/abs/2507.17746.

Seungone Kim, Juyoung Suk, Ji Yong Cho, Shayne Longpre, Chaeeun Kim, Dongkeun Yoon, Guijin Son, Yejin Cho, Sheikh Shafayat, Jinheon Baek, Sue Hyun Park, Hyeonbin Hwang, Jinkyung Jo, Hyowon Cho, Haebin Shin, Seongyun Lee, Hanseok Oh, Noah Lee, Namgyu Ho, Se June Joo, Miyoung Ko, Yoonjoo Lee, Hyungjoo Chae, Jamin Shin, Joel Jang, Seonghyeon Ye, Bill Yuchen Lin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. The BiGGen bench: A principled benchmark for fine-grained evaluation of language models with language models. In *Proceedings of NAACL: Human Language Technologies (Volume 1: Long Papers)*, pp.

5877–5919, 2025. doi: 10.18653/v1/2025.naacl-long.303. URL https://aclanthology.org/2025.naacl-long.303/.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels, 2026.

Michael Krumdick, Charles Lovering, Varshini Reddy, Seth Ebner, and Chris Tanner. No free labels: Limitations of LLM-as-a-Judge without human grounding, 2025. arXiv:2503.05061.

Fiona Lau. Same input, different scores: A multi-model study on the inconsistency of LLM judges, 2026. arXiv:2603.04417.

Yanran Li. Calibrate, don't curate: Label-efficient estimation from noisy LLM judges, 2026. arXiv:2605.09702.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2511–2522, 2023. arXiv:2303.16634.

Solomon Messing. Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking, 2026. URL https://arxiv.org/html/2604.11581v6. arXiv:2604.11581v6.

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. arXiv:2606.19544.

Dan Song, Won-Chan Lee, and Hong Jiao. Exploring LLM autoscoring reliability in large-scale writing assessments using generalizability theory, 2025.

Evangelia Spiliopoulou, Riccardo Fogliato, Hanna Burnsky, Tamer Soliman, Jie Ma, Graham Horwood, and Miguel Ballesteros. Play Favorites: A Statistical Method to Measure Self-Bias in LLM-as-a-Judge, August 2025. URL http://arxiv.org/abs/2508.06709. arXiv:2508.06709 [cs].

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. URL https://openreview.net/forum?id=G0dksFayVq.

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

Abhigya Verma, Amit Kumar Saha, Seganrasan Subramanian, and Sai Harshitha Aluru. Agent-JudgeBench: A multi-difficulty benchmark for evaluating LLM judges on agentic tool-calling, 2026. URL https://arxiv.org/abs/2608.26623. Accepted to EMNLP 2026, as reported in the arXiv record.

Yuehan Wang, Jinyan Huang, Lun Du, Yuxin Guo, Ying Liu, and Rong Wang. Evaluating large language models as raters in large-scale writing assessments: A psychometric framework for reliability and validity. *Computers and Education: Artificial Intelligence*, 9:100481, 2025a. doi: 10.1016/j.caeai.2025.100481.

Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2025b. URL https://arxiv.org/abs/2505.11475v2.

Bosi Wen, Yilin Niu, Cunxiang Wang, Xiaoying Ling, Ying Zhang, Pei Ke, Hongning Wang, and Minlie Huang. IF-RewardBench: Benchmarking judge models for instruction-following evaluation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 23816–23843. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1092. URL https://aclanthology.org/2026.acl-long.1092/.

Fengyu Xie, Yilun Zhao, Bingsen Chen, Arman Cohan, and Chen Zhao. Small language models as judges for rubric-based reinforcement learning, 2026. URL https://arxiv.org/abs/2608.30005. Findings of EMNLP 2026, as reported in the arXiv record.

Dmitrij Żatuchin. Where does the noise come from? a variance-components decomposition of non-determinism in LLM brand answers, 2026. arXiv:2607.13304.

Justin Zhao, Himaghna Bhattacharjee, Hannah Korevaar, Bhaktipriya Radharapu, and Khalid El-Arini. Jagged judges: Epistemic stability under perturbation, pressure, and persistence, 2026. URL https://arxiv.org/html/2608.12645v2. arXiv:2608.12645v2.

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. arXiv:2306.05685.

A Measurement Targets and D-Study Details

Template and generation instance. For a fixed judge, let Y_{itr} score response i under template t and independent generation r:

$$Y_{itr} = u_i + d_{it} + \epsilon_{itr}, \qquad u_i = \mathbb{E}_t \mathbb{E}_r Y_{itr}, \quad \mathbb{E}_t d_{it} = 0, \quad \mathbb{E}_r \epsilon_{itr} = 0. \tag{4}$$

Here d_{it} is the template-specific deviation from the universe score, and ϵ_{itr} is generation noise within that condition. For compact estimator formulas, the appendix abbreviates σ_I^2, σ_{word}^2, and σ_{gen}^2 as S, B, and A, respectively. These are the same components used in Section 3.2. The components are

$$S = \text{Var}_i(u_i), \qquad B = \mathbb{E}_{i,t} d_{it}^2, \qquad A = \mathbb{E}_{i,t} \text{Var}_r(Y_{itr}). \tag{5}$$

S is stable response differentiation; B is template sensitivity, including common shifts and response-specific effects; A is generation variance. Templates have identified levels shared across responses; repetition indexes independent draws within response–template cells, not shared measurement occasions. Stable variation is measurement signal, not correctness: it can contain between-query differences and reproducible mistakes.

D-studies compare grading allocations. A D-study changes only the error components affected by the proposed allocation. Suppose an evaluation averages n_1 gradings under one template and n_2 under the other, with $K = n_1 + n_2$. When unequal, choose the orientation of the allocation uniformly to retain the equal-template target. Then

$$V(n_1, n_2) = \left(\frac{n_1 - n_2}{K} \right)^2 B + \frac{A}{K}, \qquad \text{SEM}_{\mathcal{D}} = \sqrt{V(n_1, n_2)}. \tag{6}$$

Repeating one template leaves error $B + A/K$; balancing both removes finite-template deviation and leaves A/K. These are expectations over independent generations and the specified allocation, allowing heterogeneous cell variances. Covering both templates exhausts this finite facet; additional generations still sample the generation distribution. Reporting only one template's expected score changes the target; its discrepancy from the equal-wording target remains template sensitivity in the robustness study.

A practical D-study compares feasible allocations against a precision requirement. To reach SEM at most e by repeating one randomly chosen template requires $K \geq A/(e^2 - B)$ when $e^2 > B$; otherwise repetition alone cannot attain the target. Balanced coverage requires an even $K \geq A/e^2$. These are precision requirements, not accuracy guarantees. The corresponding contrast calculation follows below.

Finite-repeat estimators. With R repeats, let m_{it} and v_{it} be the cell mean and unbiased variance, and let $\bar{m}_i = (m_{i1} + m_{i2})/2$. We estimate

$$\widehat{A} = \text{avg}_{i,t}\, v_{it}, \qquad \widehat{S} = s_i^2(\bar{m}_i) - \widehat{A}/(2R), \tag{7}$$

$$\widehat{B} = \text{avg}_i \left[\frac{(m_{i1} - m_{i2})^2}{4} - \frac{v_{i1} + v_{i2}}{4R} \right]. \tag{8}$$

The corrections remove finite-repeat noise from stable differentiation and template sensitivity.

Allocation error. The finite-template universe is explicit: u_i is the average of the two template-specific expected scores. For allocation weights w_1, w_2, the template deviation equals $(w_1 - w_2)d_{i1}$ because $d_{i2} = -d_{i1}$. Averaging its square gives the template term in Equation 6. Randomly orienting unequal allocations makes each template's expected grading count $K/2$, yielding generation variance A/K even when the template-specific variances differ. Balanced allocations achieve the same result without random orientation. A globally fixed orientation can have different error and changes the estimand if its template weighting defines the deployed score.

For candidates a, b, the expected contrast is $u_a - u_b$. Its generation component is the average of $v_{at} + v_{bt}$, assuming independent candidate gradings. Its template component is estimated from the squared difference between the two template-specific mean contrasts, divided by four and corrected by $(v_{a1} + v_{b1} + v_{a2} + v_{b2})/(4R)$. Shared shifts cancel before squaring. Pair components are averaged

within query over reference-eligible pairs, then across queries. Contrast SEM measures noise about the expected score difference, not uncertainty about which reference is correct. Absolute score dependability and contrast precision are distinct targets and should not be interchanged.

All raw components are retained. Truncating a negative estimate is a boundary convention, not evidence that the component is exactly zero. Stable response variance is estimated over the observed query–response objects; query-level resampling preserves their clustering. The summary decomposition averages normalized source components rather than concatenating response populations.

There are nine feasible unoriented allocations. For each candidate we enumerate subsets independently without replacement from the three stored generations in each template cell. For a pairwise decision we integrate the Cartesian product of the two candidates' possible averages. Reordering repeat indices independently in every cell leaves all reported ordering estimates unchanged. Extrapolated precision curves beyond six calls are model projections only; no corresponding accuracy is inferred.s

B Forecast Validation and Statistical Comparisons

Calibration and test queries are disjoint within every source. Generation rankings compare the same fixed roster across the two splits. Their 95% query-bootstrap intervals are $[0.755, 0.965]$, $[0.811, 0.972]$, and $[0.804, 0.965]$ for JudgeBench, BiGGenBench and HelpSteer3. RMS forecast checks use a fixed pseudorandom permutation of repeats independently within each response–judge–template cell. For each test response, the same-template check compares two disjoint single generations, averaged over the two templates. The cross-template check compares one generation from each template. The balanced check compares two disjoint groups, each containing one generation from each template. Their expected squared disagreements are respectively $2A$, $2A + 4B$, and A. The checks share some stored gradings with one another and are not independent validation experiments. Repeat-index permutations are an assignment sensitivity check on the same outputs.

The pilot supplies predicted variances; test squared disagreements supply observed quantities. Each is averaged over responses within query and then over queries within source. Macro RMS takes the square root after equally averaging source mean squared disagreements. Its ratio intervals jointly propagate independent pilot and test query bootstraps. Agreement in these observable quantities validates neither reference accuracy nor an unobserved error on each individual response.

Paired design comparisons resample whole queries and recompute eligible denominators. For each stated contrast, let $\widehat{\delta}_j$ be the macro accuracy difference and $\widehat{\delta}_j^{(b)}$ its bootstrap replicate. The simultaneous interval is $\widehat{\delta}_j \pm c$, where c is the 95th percentile of $\max_j |\widehat{\delta}_j^{(b)} - \widehat{\delta}_j|$. This controls the reported family of twelve comparisons approximately under the bootstrap. Best and second-best table entries are descriptive rankings. All intervals are conditional on fixed annotations and the tested roster.

C Independent-Call Direction Diagnostic

For each evaluated template and each candidate, reserve two of its three generations for the outcome. The unused generation under that template and all three under the other template estimate direction. If r_a and r_b denote the unused indices for candidates a, b, the independent direction estimate is

$$\widehat{d}_{ab}^{\,\mathrm{dir}} = \frac{1}{2}\left[Y_{atr_a} - Y_{btr_b} + \frac{1}{3}\sum_{r=1}^{3}(Y_{a,1-t,r} - Y_{b,1-t,r}) \right], \tag{9}$$

with templates indexed here by $t \in \{0, 1\}$. Equal template weights preserve the declared universe despite unequal numbers of unused generations. The sign relative to the reference defines reference-consistent, reference-opposed, or tied strata. Only the reserved calls evaluate the single-grading accuracy and the accuracy of their same-template average. All nine choices of unused candidate indices and both evaluated templates are integrated, with independent candidate draws for the single-grading baseline.

Eligible pairs are weighted equally within query. Stratum masses and accuracy numerators are averaged by source, then equally across sources, before division. A query bootstrap preserves all

calls and assignments together. The three strata exactly partition both unconditional one-grading and two-grading accuracy. Separate simultaneous bands cover the twelve effects within each reported stratum. This is conditional evidence based on a noisy direction estimate, not identification of permanent errors. As a sensitivity, using only the unused same-template call to classify direction gives positive gains for all twelve reference-consistent strata and negative point estimates for eleven reference-opposed strata; ten positive and four negative simultaneous intervals exclude zero.

D MATCHED SCORE READOUT

Let ℓ_k be the stored log probability for rating digit $k \in \{0, \ldots, 5\}$. The probability-weighted readout is $\sum_k k \exp(\ell_k) / [5 \sum_k \exp(\ell_k)]$. It uses the same generated rationale as the integer score. These probabilities describe score tokens, not calibrated probabilities of correctness. All comparisons retain independent candidate integration and query-level reference metrics.

The three readout contrasts are probability-weighted versus integer single gradings, a probability-weighted single grading versus two integer gradings, and two probability-weighted gradings versus one. A joint simultaneous band spans all 36 judge–contrast effects. Five positive effects in the first contrast exclude zero; none in the other two does. The point gain is positive in 35 of the 36 judge–source comparisons, with Qwen3.5-9B on JudgeBench the exception (-0.28 points).

Integer ties account for 29.4–49.9% of decisive-pair outcomes across the aggregate judge summaries. The accuracy contribution from readout changes on tied draws is $+0.85$ to $+3.43$ points; on non-tied draws it is only -0.24 to $+0.17$. Their sum exactly recovers each judge's total gain. This diagnoses lost resolution in the rating interface, without establishing that the readout replaces a second grading at equivalent accuracy.

First-call margin accounting. For each judge, we enumerate independent candidate draws and partition its same-wording two-call gain by the first-call integer-score gap. The contributions sum exactly to that judge's reported gain. First-call ties account for 67.6–97.1% of the gain across judges; the exact-tie convention is material to this interpretation.