ABC: CERTIFIED SELECTIVE EVALUATION WITH A PANEL OF LLM JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-as-a-judge has reshaped evaluation by reducing the human effort required to assess response quality, yet LLM judgments can disagree with human ratings. More capable models may narrow this gap, but they can also make the evaluation substantially more expensive. In practice, users may therefore rely on a panel of affordable judges and, for each response, decide whether to accept the panel's judgment or escalate it to a stronger model or a human. This raises two questions: how should the judges' scores be combined, and which resulting judgments can be accepted automatically at a controlled error rate. In this paper, we propose the ABC panel: *Admit*, *Blend*, and *Certify*. ABC constructs a statistical test based on cross-fitted residual association, admitting judges whose scores remain associated with human ratings beyond observed surface features, such as response length and style. It then blends the admitted judges using penalized logistic regression fitted to human ratings. Finally, it selects and certifies an acceptance rule such that, with a high probability, the population error rate among admitted judgments is below a threshold. Experiments on extensive human-rated benchmarks show that the ABC panel automates a larger share of evaluation than single-judge and ensemble baselines while controlling the error rate.

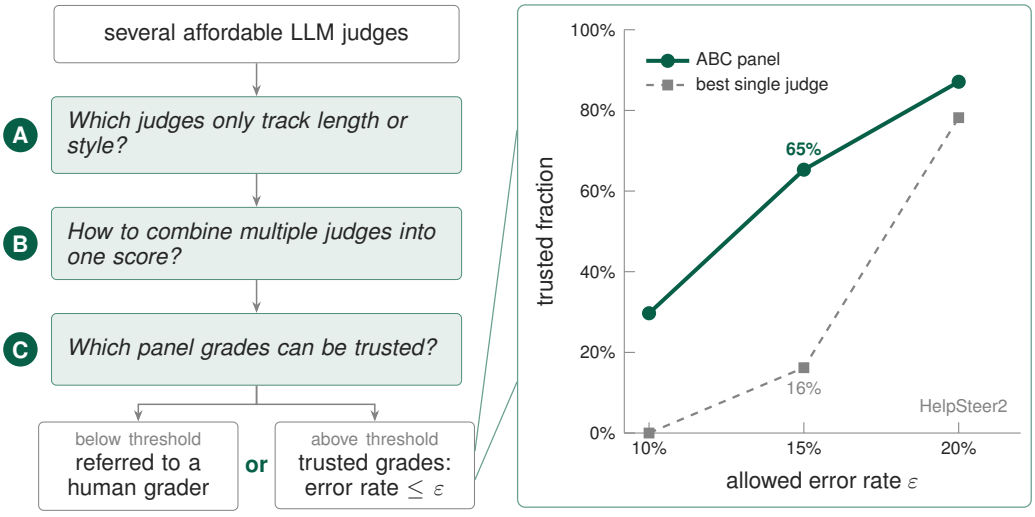

Figure 1: The ABC panel. Left: *Admit* (A), *Blend* (B), and *Certify* (C), select and combine affordable LLM judges, then determine when to trust the panel's grade of a response or refer the response to a human grader. Right: coverage, i.e., the fraction of grades that the panel accepts, against the allowed error rate ε on the benchmark data HelpSteer2 at 95% confidence. At $\varepsilon = 15\%$ the panel accepts 65% of its grades, as compared with 16% for the best single judge.

The new scores from nine LLM judges on HelpSteer2 will be released as a new benchmark for LLM-as-a-judge research after the anonymous review period.

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.

1 INTRODUCTION

With the development of LLMs, large numbers of responses can now be generated easily. LLM-as-a-judge is widely used to evaluate those responses: a judge reads a response and gives a score according to a preset rubric, or reads two responses and judges which is better. For example, a judge rates the helpfulness of a response in the benchmark data HelpSteer2 (Wang et al., 2024), or compares two chatbot responses to the same question in Jung et al. (2025). The responses are typically open-ended, so the evaluation is subjective and human ratings remain the gold standard; no LLM judge matches human ratings everywhere (Tan et al., 2025; Wei et al., 2025). Collecting human ratings, however, is slow and expensive, which makes large-scale evaluation difficult.

Consider Alice, who is in charge of evaluation in a small team. With thousands of responses to grade, she has a few human graders, but her budget does not allow them to read every response. She first asks the human graders to only grade a few hundred responses, and she then turns to several affordable LLM judges, such as small open models running on one GPU or models available through inexpensive APIs. Importantly, Alice needs to know which automated grades she can trust. She wants a statement of the following form:

> *The system accepts 80% of the grades and refers the other 20% to a human grader.*
> *With 95% confidence, the error rate among the accepted grades is at most 15%.*

In general, Alice wants to accept as many automated grades as possible while ensuring that the error rate among the accepted grades exceeds ε with probability at most δ.

There are three main challenges. First, some judges are deceptively powerful: they read only surface features of a response, such as its length, formatting, and style (Dubois et al., 2024; Li et al., 2024; Hu et al., 2025). For example, longer responses often receive higher human grades, so a judge that reads nothing but length may still be correlated with the human grade. We call such a judge a surface judge, which adds nothing beyond the surface features. We need a formal statistical test that removes it. Second, even after removing surface judges, the judges can still disagree with each other and overlap (Kohli, 2026). We need an automatic way to select and to merge them. The common practice is to let the judges vote (Verga et al., 2024), which uses the judges equally and ignores the human grades, and it accepts far fewer grades than the panel (Figure 1). Third, the accepted grades must be certified: with the stated confidence, we need to know that their error rate is below ε. The unknown error rate has to be verified from a limited number of human grades, for a panel that was itself built from human grades, which makes a naive estimate of the error rate too optimistic.

In this paper, we propose the ABC panel, an end-to-end system that addresses Alice's three challenges in order. We split her human-graded prompts into training and certification data. In Step A (Admit), we test each judge on the training data and admit it only if there is evidence that its score is correlated with the human grade after accounting for the surface features. In Step B (Blend), we fit an ℓ_1-penalized logistic regression of the human grade on the admitted scores and a prediction based on the surface features to obtain the panel's score. In Step C (Certify), we use fixed-sequence testing on the certification data to choose a threshold for accepting grades, with the guarantee Alice seeks. The three steps also apply when a prompt has several responses.

To our best knowledge, the ABC panel is the first to admit, blend, and certify multiple LLM judges with a finite-sample guarantee on the error rate among accepted grades. Existing guarantees cover a single judge (Jung et al., 2025; Badshah et al., 2026a), or bound the risk of an averaged panel only on average over calibration sets (Cheng et al., 2026), so their acceptance rate at a level α is not comparable with coverage at a certified error rate. Beyond the combination, three parts are new: Step A is the first formal statistical test to qualify a judge by what its score adds beyond the surface features; the certificate in Step C stays valid even when several dependent responses share a prompt, and its starting rule is chosen in a data-driven manner. On three benchmarks, the ABC panel accepts more grades than the best single judge, majority vote, and a cascade of judges, while its error rate on held-out test data stays below ε.

In addition, we contribute new benchmark data for LLM-as-a-judge research: the scores of nine LLM judges on HelpSteer2 and of eight additional judges on the preference data of Jung et al. (2025), along with the surface features of every response. The data and the code will be released.

2 RELATED WORK

LLM-as-a-judge has become a widely used approach for evaluating open-ended model outputs, with early systems such as MT-Bench and G-Eval demonstrating the use of LLMs for scoring, ranking, and rubric-based assessment (Zheng et al., 2023; Liu et al., 2023). Recent work has increasingly focused on judge reliability. JudgeBench evaluates LLM judges on objectively verifiable tasks in knowledge, reasoning, mathematics, and coding, while Zhou et al. (2025) study their effectiveness as evaluators in test-time scaling. Other studies examine judge reliability across models, prompts, and tasks (Wei et al., 2025; Tan et al., 2025; Xu et al., 2025; Li et al., 2025), including long-form evaluation (Chen et al., 2026). These studies show that agreement with human judgments can vary substantially across settings and may partly reflect response characteristics. Systematic biases have also been documented, including sensitivity to response length, position, and style (Dubois et al., 2024; Li et al., 2024; Hu et al., 2025; Shi et al., 2025). Recent surveys synthesize these reliability and bias concerns and emphasize the need for systematic evaluation of LLM judges (Gu et al., 2026). Statistical approaches have begun to address this problem more directly. Maia Polo et al. (2025), for example, model judge-specific deviations from human judgments using response-level covariates and develop asymptotic inference. Recent work has also considered selective use of LLM judges, with formal guarantees for the error rate of accepted evaluations (Jung et al., 2025; Badshah et al., 2026a;b). These methods, however, do not address the construction and certification of a multi-judge grading rule.

Combining predictions from multiple models is classical in statistics and machine learning. Forecast combination, stacking, and ensemble methods provide general frameworks for aggregating heterogeneous predictors (Clemen, 1989; Wolpert, 1992; Breiman, 1996; Dietterich, 2000). For LLM evaluation, multiple judges may provide complementary information, motivating panel construction. Their agreement with human grades may also reflect response characteristics, motivating covariate adjustment. Residualized associations and orthogonalized estimation provide precedents for such adjustment (Robinson, 1988; Chernozhukov et al., 2018; Shah & Peters, 2020). Existing work leaves open how to screen judges for task-specific information, combine qualified judges, and certify the resulting grading rule.

Certifying a grading rule requires deciding when an automated grade should be accepted rather than deferred to a human. Selective prediction formalizes this decision through the trade-off between coverage and prediction error. The reject-option framework of Chow (1970) and subsequent work on selective classification (Geifman & El-Yaniv, 2017) characterize this trade-off. More recent work provides finite-sample or high-probability control of user-specified risks, including risk-controlling prediction sets (Bates et al., 2021) and the Learn then Test framework (Angelopoulos et al., 2025). Multiple-testing methods provide tools for selecting among candidate procedures while controlling an overall error level (Maurer et al., 1995; Genovese et al., 2006; Roeder & Wasserman, 2009). These results provide a statistical basis for certifying a data-dependent selective rule, but do not directly address multi-judge grading with dependent responses within prompts.

A key challenge in applying selective-prediction tools to LLM grading is that individual responses may not be independent. For independent responses, exact binomial inference provides a direct finite-sample approach to controlling a binary error probability; when multiple responses are generated for the same prompt, however, response-level binomial inference is inappropriate. Recent work addresses this dependence through hierarchical models, cluster-aware inference, and methods tailored to cluster dependence in LLM evaluation (Zhang et al., 2025; Chacón Sartori & García, 2026; Zeng et al., 2026). Studies of uncertainty and judge reliability have also used multiple generations or repeated judgments (Gupta & Kumar, 2026; Chen et al., 2026; Longjohn et al., 2025; Saha et al., 2026). Recent work on anytime-valid inference for LLM generation and evaluation (Akter et al., 2025; Gao, 2026; Li, 2026) develops finite-sample-valid sequential procedures based on e-statistics, conformal test martingales, and e-based sequential testing (Pérez-Ortiz et al., 2024; Lardy & Pérez-Ortiz, 2025; Wang et al., 2025; Wang & Zhong, 2026). These developments motivate treating prompts, rather than individual responses, as the independent sampling units and aggregating responses within each prompt when certifying a selective grading rule.

Our framework brings these components together: task-specific judge qualification after adjustment for response characteristics, panel construction from the qualified judges, and prompt-level finite-sample certification of the resulting selective grading rule.

3 THE ABC PANEL

3.1 OVERVIEW

Suppose we have N prompts. For each prompt $i = 1, \ldots, N$, a human either evaluates the quality of one LLM-generated response or compares two responses generated by two LLMs. In the two cases, we set $Y_i = 1$ if its human grade reaches a specified cutoff or the human prefers the first response, respectively, and $Y_i = 0$ otherwise. We also collect J LLMs to evaluate the same response or pair of responses, giving scores $Z_{ij} \in \mathbb{R}$, where for each j, $Z_{ij} > 0$ means the jth judge predicts $Y_i = 1$.

To account for surface characteristics in the LLM responses, we compute d *surface features*, $X_i \in \mathbb{R}^d$, from the response text, such as length, formatting, and readability. For simplicity, we consider binary human grades and one evaluation per prompt in the main text, and assume that prompts are independent and identically distributed. Section 3.5 and Supplement B extend the method to prompts with multiple responses.

Our goal is to build a panel of judges and decide when to trust its grade. This requires choosing (1) a subset of judges $\mathcal{A} \subseteq \{1, \ldots, J\}$, (2) a score $\eta(Z_{\mathcal{A}}, X)$ that combines their scores $Z_{\mathcal{A}} = \{Z_j\}_{j \in \mathcal{A}}$ with the surface features X, and (3) a threshold τ. The panel assigns the grade $\widehat{Y} = \mathbb{I}\{\eta(Z_{\mathcal{A}}, X) > 0\}$. We accept this grade when $|\eta(Z_{\mathcal{A}}, X)| \geq \tau$ and otherwise refer the prompt to a human grader.

To measure the performance, we define its coverage $C(\mathcal{A}, \eta, \tau)$ and selective risk $\pi(\mathcal{A}, \eta, \tau)$ by

$$C(\mathcal{A}, \eta, \tau) = \mathbb{P}\{|\eta(Z_{\mathcal{A}}, X)| \geq \tau\} \text{ and } \pi(\mathcal{A}, \eta, \tau) = \mathbb{P}\left\{\widehat{Y} \neq Y \mid |\eta(Z_{\mathcal{A}}, X)| \geq \tau\right\}.$$

Given a pre-specified error level $\varepsilon \in (0, 1)$, we determine the three parts as follows: (1) the **A**dmit step, which determines the set $\mathcal{A}$, admitting a judge only if its agreement with the human grades goes beyond the surface features; (2) the **B**lend step, which computes the score η by combining the admitted judges and the surface features into a single score by a penalized logistic regression on the human grades; (3) the **C**ertify step, which determines the threshold τ through

$$\max_{\tau > 0} C(\mathcal{A}, \eta, \tau), \text{ subject to } \pi(\mathcal{A}, \eta, \tau) \leq \varepsilon. \tag{1}$$

We refer to the three steps as Step A (admit), Step B (blend), and Step C (certify), thereby the ABC panel. We randomly split the N prompts into a training set $\mathcal{I}$, a certification set $\mathcal{I}_{\mathrm{c}}$, and a test set $\mathcal{I}_{\mathrm{te}}$. We fit the panel on the training set in Steps A and B, certify the rule on the certification set in Step C, and evaluate it on the test set. Let $n = |\mathcal{I}|$ be the number of training prompts.

3.2 STEP A: ADMISSION

Step A screens out what we call *surface judges*, whose agreement with human grades comes only from a shared dependence on surface features such as response length, and admits the rest.

The main purpose of Step A is not to raise coverage but to certify which judges carry signal beyond the surface features, before Step B can select a surface judge by chance. We make this precise using the conditional correlation between the grade and the judge's score after accounting for X.

For each judge j, let $\mu(X) = \mathbb{E}[Y \mid X]$ and $\nu_j(X) = \mathbb{E}[Z_j \mid X]$ be the conditional means of the human grade and the judge's score given X. The residuals $U = Y - \mu(X)$ and $V_j = Z_j - \nu_j(X)$ both have conditional mean zero given X, so their covariance is

$$C_j = \mathbb{E}[UV_j] = \mathbb{E}[\mathrm{Cov}(Y, Z_j \mid X)].$$

We standardize this covariance to obtain the conditional correlation,

$$\rho_j = \frac{C_j}{\sqrt{A_j B}}, \qquad A_j = \mathbb{E}[V_j^2], \qquad B = \mathbb{E}[U^2].$$

Definition 1. *Judge j is a surface judge if $\rho_j = 0$.*

It is important to note that a surface judge with $\rho_j = 0$ may still have a large marginal correlation with the human grade because both depend on the same surface features. Screening by marginal

correlation could therefore admit such a judge, so Step A uses conditional correlation; see Supplement C.6 for examples.

We next estimate ρ_j. First, we estimate μ by logistic regression of Y on X, and estimate ν_j by linear regression of Z_j on X. To avoid overfitting, we use K-fold cross-fitting. Specifically, for each fold k, $\widehat{\mu}^{(-k)}$ and $\widehat{\nu}_j^{(-k)}$ are fitted using the other $K-1$ folds. For each i in fold $k(i)$, define

$$\widehat{U}_i = Y_i - \widehat{\mu}^{(-k(i))}(X_i), \quad \text{and} \quad \widehat{V}_{ij} = Z_{ij} - \widehat{\nu}_j^{(-k(i))}(X_i).$$

We center $\widehat{U}_i$ and $\widehat{V}_{ij}$ at their training-sample means, and estimate the conditional correlation by

$$\widehat{\rho}_j = \frac{\widehat{C}_j}{\sqrt{\widehat{A}_j \widehat{B}}}, \text{ where } \widehat{A}_j = \frac{1}{n}\sum_{i=1}^n \widehat{V}_{ij}^2,\ \widehat{B} = \frac{1}{n}\sum_{i=1}^n \widehat{U}_i^2, \text{ and } \widehat{C}_j = \frac{1}{n}\sum_{i=1}^n \widehat{U}_i \widehat{V}_{ij}.$$

This uses the residual construction of Robinson (1988), with cross-fitting following the double machine learning framework of Chernozhukov et al. (2018).

By the delta method applied to $\widehat{\rho}_j$, we estimate its standard error,

$$\widehat{\mathrm{se}}_j = \Big\{ \frac{1}{n(n-1)} \sum_{i=1}^n \widehat{\psi}_{ij}^2 \Big\}^{1/2}. \tag{2}$$

where

$$\widehat{\psi}_{ij} = \frac{\widehat{U}_i \widehat{V}_{ij} - \widehat{C}_j}{\sqrt{\widehat{A}_j \widehat{B}}} - \frac{\widehat{\rho}_j}{2}\left\{ \frac{\widehat{V}_{ij}^2 - \widehat{A}_j}{\widehat{A}_j} + \frac{\widehat{U}_i^2 - \widehat{B}}{\widehat{B}} \right\} \text{ and } \bar{\psi}_j = \frac{1}{n}\sum_{i=1}^n \widehat{\psi}_{ij} = 0.$$

Proposition 1. *Under the conditions of Proposition 4 in Supplement C.6, for each judge j,*

$$(\widehat{\rho}_j - \rho_j)/\widehat{\mathrm{se}}_j \text{ converges to } N(0,1) \text{ in distribution.}$$

Admission rule. We admit a judge only when its conditional correlation exceeds a minimum magnitude $\rho_{\min} > 0$. With $\alpha = 0.05$ and z_q denoting the q-quantile of the standard normal, let

$$I_j = \big(\widehat{\rho}_j - z_{1-\alpha/(2J)}\widehat{\mathrm{se}}_j,\ \widehat{\rho}_j + z_{1-\alpha/(2J)}\widehat{\mathrm{se}}_j\big)$$

be the Bonferroni-adjusted confidence interval for ρ_j. We admit judge j if I_j lies entirely outside $[-\rho_{\min}, \rho_{\min}]$, with $\rho_{\min} = 0.05$. If no judge is admitted, we refer every prompt to a human grader.

3.3 Step B: blend

Let $\mathcal{A}$ denote the set of judges admitted by Step A, and let $\widetilde{Z}_j$ denote judge j's score standardized using its mean and standard deviation on the training set $\mathcal{I}$. The use of surface features X follows the bridge method (Maia Polo et al., 2025), which models a single judge's score. In our method, Step B combines the scores of the admitted judges $Z_{\mathcal{A}}$ with X into a single logistic score,

$$\eta(Z, X) = \theta_0 + \sum_{j \in \mathcal{A}} \theta_j \widetilde{Z}_j + \theta_h\, h(X), \text{ where } h(X) = \gamma_0 + X^\top \gamma.$$

The sign of η determines the panel's grade, and its magnitude $|\eta|$ measures the panel's confidence. We estimate the model in two stages: Stage 1 estimates the surface score h, and Stage 2 estimates the weights θ collecting $(\theta_0, \theta_j, j \in \mathcal{A}, \theta_h)$.

Stage 1: the surface score. For a grade y and log-odds t, let $\ell(y,t) = \log(1+e^t) - yt$ denote the logistic loss. We first fit the surface component by ridge-penalized logistic regression,

$$\widehat{h}(x) = \widehat{\gamma}_0 + x^\top \widehat{\gamma}, \text{ where } (\widehat{\gamma}_0, \widehat{\gamma}) = \arg\min_{\gamma_0, \gamma}\ \frac{1}{n}\sum_{i=1}^n \ell\big(Y_i, \gamma_0 + X_i^\top \gamma\big) + \frac{\lambda_h}{2}\|\gamma\|^2.$$

The ridge penalty is used only for numerical stability; we fix $\lambda_h = 0.01$ throughout.

To avoid using the same observations both to fit and evaluate the surface component, we use a K-fold data split and use out-of-fold surface predictions. Let $I_1, \ldots, I_K$ partition the training set $\mathcal{I}$, with $I_k^c = \mathcal{I} \setminus I_k$, and let $k(i)$ denote the fold containing observation i. A superscript $(-k)$ indicates that the corresponding fit is obtained without fold I_k. For each fold k, fit

$$(\widehat{\gamma}_0^{(-k)}, \widehat{\gamma}^{(-k)}) = \arg\min_{\gamma_0,\gamma} \frac{1}{|I_k^c|} \sum_{i \notin I_k} \ell\big(Y_i, \gamma_0 + X_i^\top \gamma\big) + \frac{\lambda_h}{2}\|\gamma\|^2,$$

and set $\widehat{h}^{(-k(i))}(X_i) = \widehat{\gamma}_0^{(-k(i))} + X_i^\top \widehat{\gamma}^{(-k(i))}$. Thus, each training data point i enters Stage 2 through $\widehat{h}^{(-k(i))}(X_i)$, whose regression coefficients γ and γ_0 are fitted without its label Y_i.

Stage 2: the judge weights. We then combine the admitted judges with the surface prediction using an ℓ_1 logistic regression,

$$\widehat{\theta} = \arg\min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \ell\Big(Y_i,\ \theta_0 + \sum_{j \in \mathcal{A}} \theta_j \widetilde{Z}_{ij} + \theta_h \widehat{h}^{(-k(i))}(X_i)\Big) + \lambda\Big(\sum_{j \in \mathcal{A}} |\theta_j| + |\theta_h|\Big),$$

where the surface prediction h is evaluated out of fold. Due to the ℓ_1 penalty, θ_j can be zero if the jth judge contributes little to Y after accounting for the other judges and the surface features. The tuning parameter λ is selected by five-fold cross-validation over the same fold split.

3.4 Step C: certification

So far, we have selected $\mathcal{A}$ and fitted η on the training set $\mathcal{I}$. For a pre-specified error level ε, Step C chooses the threshold τ with the objective of solving problem (1). Since the selective risk $\pi(\mathcal{A}, \eta, \tau)$ is unknown, we cannot check the constraint in problem (1) directly. We instead use the certification set $\mathcal{I}_c$ to choose a rule whose risk exceeds ε with probability at most δ. This guarantee is called (ε, δ)-risk control (Bates et al., 2021; Angelopoulos et al., 2025).

In particular, Step C selects a threshold $\widehat{\tau}$ and accepts the panel's grade when $|\eta(Z_{\mathcal{A}}, X)| \geq \widehat{\tau}$; otherwise, it refers the prompt to a human grader. The selected rule satisfies

$$\mathbb{P}\{\pi(\mathcal{A}, \eta, \widehat{\tau}) > \varepsilon\} \leq \delta, \tag{3}$$

where the probability is taken over the certification data; we set $\delta = 0.05$. Among the thresholds certified by its testing procedure, Step C chooses the one with the largest coverage.

For $i \in \mathcal{I} \cup \mathcal{I}_c$, write $\eta_i = \eta(Z_i, X_i)$ for the score of prompt i under the fitted panel. The collection $\{|\eta_i| : i \in \mathcal{I}\}$ over the training set gives the candidate thresholds: with a pre-specified sequence $0 < q_1 < \cdots < q_L \leq 1$, compute $\tau_l = \widehat{Q}_{1-q_l}\{|\eta_i| : i \in \mathcal{I}\}$, where $\widehat{Q}_u$ is the empirical u-quantile.

To certify a candidate τ_l, we ask whether the risk $\pi(\mathcal{A}, \eta, \tau_l) < \varepsilon$. The accepted count is $m_l = \sum_{i \in \mathcal{I}_c} \mathbb{I}\{|\eta_i| \geq \tau_l\}$ and the error count is $e_l = \sum_{i \in \mathcal{I}_c} \mathbb{I}\{|\eta_i| \geq \tau_l\}\mathbb{I}\{\mathbb{I}\{\eta_i > 0\} \neq Y_i\}$. For candidates with positive coverage, we conduct a hypothesis test:

$$H_{0l} :\ \pi(\mathcal{A}, \eta, \tau_l) \geq \varepsilon \text{ v.s. } H_{1l} :\ \pi(\mathcal{A}, \eta, \tau_l) < \varepsilon.$$

Since the certification prompts are independent draws from the population, the error count satisfies $e_l \mid m_l \sim \mathrm{Binomial}(m_l, \pi(\mathcal{A}, \eta, \tau_l))$, giving the exact binomial p-value (Clopper & Pearson, 1934)

$$p_l = \sum_{k=0}^{e_l} \binom{m_l}{k} \varepsilon^k (1-\varepsilon)^{m_l - k}. \tag{4}$$

Accordingly, when $p_l \leq \delta$, we reject H_{0l}, conclude that $\pi(\mathcal{A}, \eta, \tau_l) < \varepsilon$, and claim that τ_l is certified. The following lemma shows that this p-value is valid in finite samples.

Lemma 1. *Given the estimated $\mathcal{A}$ and η, suppose that $\pi(\mathcal{A}, \eta, \tau_l) \geq \varepsilon$. Then $\mathbb{P}\{p_l \leq \delta\} \leq \delta$.*

Therefore, with probability at least $1 - \delta$, the candidate τ_l is either not certified or satisfies the constraint of problem (1).

The test above certifies a single candidate. Among the certified candidates, the next step is to pick the smallest threshold, which gives the largest coverage, the objective of problem (1). A naive approach tests every candidate at level δ and picks the smallest certified one, but this inflates the error. The Bonferroni correction fixes this by testing at level δ/L, but it is too conservative and certifies much less. We instead use fixed-sequence testing (Bauer, 1991), which keeps the full level for each test.

Fixed-sequence testing Starting at a well-chosen strict threshold τ_{l_0}, we test at level δ and loosen the threshold one level at a time until the first failure to certify. We then deploy the last certified threshold $\widehat{\tau} = \tau_{\widehat{l}}$, where

$$\widehat{l} = \max\big\{l \in \{l_0, \ldots, L\} : \; p_{l_0} \le \delta, \; p_{l_0+1} \le \delta, \; \ldots, \; p_l \le \delta\big\},$$

and set $\widehat{\tau} = \infty$ if $p_{l_0} > \delta$.

Theorem 1. *Given the estimated $\mathcal{A}$ and η, and a starting level l_0, the threshold $\widehat{\tau}$ selected by the fixed-sequence testing satisfies the constraint in inequality* (3)*.*

Note that the guarantee does not depend on Steps A and B being right. A wrong admission or a poorly fitted score lowers the coverage the panel can reach, but the certificate is computed on the certification data alone, so the error bound still holds.

Theorem 1 provides the risk guarantee for any starting level l_0. The choice of l_0, however, affects coverage. Since $\tau_1 \ge \cdots \ge \tau_L$, a small l_0 with a large threshold τ_l may accept too few prompts to claim the error rate below ε. A large l_0 with a small threshold τ_l may accept too many incorrect grades. In either case, the sequence may stop early, ending with low coverage.

We choose l_0 by bootstrap. For each of B resamples of the training prompts, we compute p-values $p_l^{(b)}$ as in expression (4). Starting at s, the last certified level is

$$d_s^{(b)} = \max\left\{l \in \{s, \ldots, L\} : p_s^{(b)} \le \delta, \ldots, p_l^{(b)} \le \delta\right\}.$$

The coverage on the original training set $\mathcal{I}$ is

$$S_s^{(b)} = \frac{1}{n} \sum_{i \in \mathcal{I}} \mathbb{I}\left\{|\eta_i| \ge \tau_{d_s^{(b)}}\right\},$$

taken as zero if nothing is certified. We select $l_0 = \arg\max_{s \in 1,\ldots,L} \frac{1}{B} \sum_{b=1}^{B} S_s^{(b)}$.

Returning to problem (1), Step C selects the smallest threshold $\widehat{\tau}$, which gives the largest coverage among the thresholds certified by the procedure, with the risk guarantee in (3). We report its coverage on the certification set as

$$\widehat{C}(\mathcal{A}, \eta, \tau) = \frac{1}{|\mathcal{I}_{\mathrm{c}}|} \sum_{i \in \mathcal{I}_{\mathrm{c}}} \mathbb{I}\{|\eta_i| \ge \widehat{\tau}\}.$$

If no threshold is certified, all prompts are referred to a human and the reported coverage is zero.

3.5 Extension: prompts with multiple responses

So far, we have assumed one evaluation per prompt: a human grade of a single LLM-generated response, or a comparison of two responses. When a prompt has several evaluations, for example grades of several responses to the same prompt, these evaluations may be dependent because they share the prompt and the grading instructions.

We extend the ABC panel with three changes. First, we split the data by prompt, keeping all evaluations of each prompt in the same training, certification, or test set and the same cross-fitting fold. Second, in Step A, we compute the standard error in (2) by summing the contributions within each prompt before squaring. Third, in Step C, we replace the binomial p-value in (4) with a betting p-value (Shafer, 2021; Waudby-Smith & Ramdas, 2024) based on the counts of accepted evaluations and errors within each prompt. The bootstrap for choosing l_0 also resamples whole prompts. With these changes, the analogs of Lemma 1 and Theorem 1 hold. Supplement B gives the details and states the general results, Lemma 2, and Theorem 2 with proofs given in Supplement C.

4 Experiments

The objectives of the numerical experiments are threefold.

- The ABC panel accepts more grades than any single judge or simple combination at the same ε.

Table 1: Certified coverage and test error, averaged over 20 random partitions. Cascade is the cascade of Jung et al. (2025). Bold marks the best coverage among the rules whose test error is at most ε, italics mark a rule whose test error is above ε, and a dash marks a rule that certified nothing.

	$\varepsilon = 0.10$		$\varepsilon = 0.15$		$\varepsilon = 0.20$	
Rule	coverage	test error	coverage	test error	coverage	test error
BiGGen Bench						
ABC panel	**0.014** $\pm$ 0.000	0.092 $\pm$ 0.000	**0.450** $\pm$ 0.050	0.115 $\pm$ 0.009	**0.769** $\pm$ 0.017	0.161 $\pm$ 0.007
Best single judge	–	–	0.229 $\pm$ 0.057	0.118 $\pm$ 0.012	0.640 $\pm$ 0.048	0.154 $\pm$ 0.009
Majority vote	–	–	*0.077* $\pm$ *0.035*	*0.157* $\pm$ *0.013*	0.541 $\pm$ 0.051	0.152 $\pm$ 0.010
Cascade	*0.022* $\pm$ *0.000*	*0.129* $\pm$ *0.000*	0.102 $\pm$ 0.043	0.124 $\pm$ 0.004	0.582 $\pm$ 0.036	0.155 $\pm$ 0.006
Chatbot Arena						
ABC panel	**0.566** $\pm$ 0.008	0.078 $\pm$ 0.003	**0.798** $\pm$ 0.005	0.128 $\pm$ 0.003	**0.956** $\pm$ 0.005	0.175 $\pm$ 0.003
Best single judge	0.460 $\pm$ 0.025	0.077 $\pm$ 0.002	0.687 $\pm$ 0.007	0.128 $\pm$ 0.004	0.915 $\pm$ 0.006	0.179 $\pm$ 0.003
Majority vote	–	–	0.418 $\pm$ 0.030	0.117 $\pm$ 0.004	0.827 $\pm$ 0.008	0.169 $\pm$ 0.004
Cascade	0.503 $\pm$ 0.034	0.079 $\pm$ 0.003	0.757 $\pm$ 0.004	0.128 $\pm$ 0.003	0.866 $\pm$ 0.003	0.176 $\pm$ 0.004
HelpSteer2						
ABC panel	**0.297** $\pm$ 0.030	0.083 $\pm$ 0.002	**0.653** $\pm$ 0.012	0.130 $\pm$ 0.002	**0.871** $\pm$ 0.006	0.181 $\pm$ 0.003
Best single judge	–	–	0.162 $\pm$ 0.041	0.136 $\pm$ 0.004	0.782 $\pm$ 0.010	0.181 $\pm$ 0.003
Majority vote	–	–	0.229 $\pm$ 0.052	0.133 $\pm$ 0.003	0.696 $\pm$ 0.032	0.167 $\pm$ 0.005
Cascade	–	–	0.200 $\pm$ 0.037	0.129 $\pm$ 0.005	0.619 $\pm$ 0.012	0.176 $\pm$ 0.003

- Step A removes the surface judges that only track surface features and keeps the others.
- The certificate holds on held-out test data: the error among accepted grades stays below ε.

We evaluated ABC on three benchmarks with human grades. BiGGen Bench has 2,704 responses to 677 prompts, up to four per prompt, each scored by a human from 1 to 5. HelpSteer2 has 14,318 responses to 7,159 prompts, two per prompt, each with a human helpfulness score from 0 to 4. Chatbot Arena has 5,169 comparisons, each a single-turn prompt with two responses and one human vote; it is the subset released with the judges of Jung et al. (2025). For each benchmark, the panel holds the judge scores released with it and the scores of additional judges that we collected: 11 judges on BiGGen, six from Bridge and five of ours; 11 on Chatbot Arena, the three released by Jung et al. and eight of ours; and 9 on HelpSteer2, all ours. The small open models SmolLM2-135M, SmolLM2-360M, Qwen2.5-0.5B, and Llama-3.2-1B are the judges we added.

We made the grades binary: on BiGGen the grade is 1 when the human score is at least 4, on HelpSteer2 when the helpfulness score is at least 3, and on Chatbot Arena when the human prefers response A. A judge's score is the log-odds of grade 1 computed from the probabilities it reports. We used exactly the 76 surface features of the bridge method, 38 for the prompt and 38 for the response. They cover length, formatting, readability, sentiment, and discourse counts; see (Maia Polo et al., 2025) for the full list and more details in the supplement.

We used 20 random partitions of the prompts into 30% training, 50% certification, and 20% test data, with $\varepsilon \in \{0.10, 0.15, 0.20\}$ and $\delta = 0.05$. Table 1 reports certified coverage, averaged over the 20 partitions, and the error among the accepted grades on the test data. The three baselines keep Steps A and C and change the score of Step B. Specifically, *the best single judge* uses the log-odds of the admitted judge with the largest out-of-fold coverage at ε on the training data. In *majority vote*, each admitted judge votes by the sign of its log-odds, and the score is the number of votes for grade 1 minus the number for grade 0. For cascade of Jung et al. (2025): the admitted judges in cost order, each with its own cutoff on its log-odds; the first judge that passes its cutoff gives the grade, and the cutoffs are set on the training data for each level of a ladder of training error rates, which Step C then tests in place of the panel's menu.

Table 1 shows higher coverage for ABC than for every baseline that keeps its test error below ε, at every allowed error rate. At $\varepsilon = 0.10$ on Chatbot Arena, coverage is 0.566 against 0.503 for the cascade and 0.460 for the best single judge. At the same rate on HelpSteer2, ABC accepts 0.297 of the grades, while no baseline certifies anything. On BiGGen at $\varepsilon = 0.15$, ABC reaches 0.450 against 0.229 for the best single judge and 0.102 for the cascade; at $\varepsilon = 0.10$, with only 677 prompts, ABC

Table 2: Step A on the partitions of Table 1: judges per benchmark, judges admitted in every partition, and each judge screened out in at least one partition, with the number of partitions admitting it, its mean $\widehat{\rho}$, and its mean Bonferroni-adjusted 95% confidence interval. Full lists in Supplement E.

Benchmark	judges	admitted	screened-out judge	admitted in	$\widehat{\rho}$	95% interval
BiGGen Bench	11	11	none			
Chatbot Arena	11	7	SmolLM2-135M	0 of 20	0.002	[−0.072, 0.075]
			SmolLM2-360M	0 of 20	0.005	[−0.068, 0.078]
			Qwen2.5-0.5B	6 of 20	0.119	[0.047, 0.191]
			Llama-3.2-1B	15 of 20	0.145	[0.068, 0.221]
HelpSteer2	9	5	SmolLM2-135M	0 of 20	0.056	[0.013, 0.100]
			SmolLM2-360M	0 of 20	0.011	[−0.034, 0.056]
			Qwen2.5-0.5B	0 of 20	−0.013	[−0.057, 0.032]
			Llama-3.2-1B	0 of 20	0.030	[−0.015, 0.075]

Table 3: ABC on the held-out test data: the test error, mean with standard error over the certifying partitions, and the number of the 20 partitions in which a one-sided binomial test does not reject that it is at most ε. The count shows the confidence level $1 - \delta = 0.95$ of the certificate.

	$\varepsilon = 0.10$		$\varepsilon = 0.15$		$\varepsilon = 0.20$	
Benchmark	test error	at most ε	test error	at most ε	test error	at most ε
BiGGen Bench	0.092 ± 0.000	20 of 20	0.115 ± 0.009	19 of 20	0.161 ± 0.007	19 of 20
Chatbot Arena	0.078 ± 0.003	20 of 20	0.128 ± 0.003	20 of 20	0.175 ± 0.003	20 of 20
HelpSteer2	0.083 ± 0.002	20 of 20	0.130 ± 0.002	20 of 20	0.181 ± 0.003	20 of 20

and the cascade each certify a rule in one of the 20 partitions. The allowed error rate is the user's choice. When the judges or the labeled sample cannot support it, the certificate accepts nothing, and every response goes to a human grader.

In Table 2, Step A admits every judge on BiGGen in every partition. On HelpSteer2 it admits five of the nine judges in every partition. On Chatbot Arena it admits seven of the eleven in every partition, screens out the two SmolLM2 models in all 20, and admits Qwen2.5-0.5B in 6 and Llama-3.2-1B in 15 of the 20. In Supplement E, we also included two synthetic judges that read only the response length. They are admitted by a plain correlation test in up to all 20 partitions but are screened out by Step A in every partition.

In Table 1, the test error of ABC is below ε on every benchmark, while two baseline cells exceed ε on BiGGen. We next examine the confidence level $1 - \delta = 0.95$. For every partition we check whether its test error is above ε via a one-sided binomial test. Table 3 shows that, on every benchmark and at every ε, 19 or 20 of the 20 partitions pass, which aligns with the confidence level of 95%.

5 Scope and limitations

We return to Alice. The ABC panel gives her an end-to-end procedure that she can implement. Step A screens the judges that only track surface features, Step B combines the affordable judges that remain and delivers grades, and Step C tells her which grades to trust, with a guarantee that their error rate is at most ε at confidence $1 - \delta$. ABC performed well on three benchmarks, for which we also collected new judge scores. We will release these scores with the paper as a benchmark for future work on judge aggregation.

Some limitations remain. The panel currently handles binary grades. Extending it to grades with several ordered levels remains future work. In addition, the performance still depends on the available judges. If every judge predicts human grades poorly or reads only surface features, the panel may certify nothing, and every response goes to a human grader. Better judges could allow the panel to accept more grades at the same allowed error rate.

6 AI DISCLOSURE

The ideas, methods, experiments, and conclusions in this paper were developed by the authors. We used ChatGPT to help polish the writing and correct grammatical errors. The code was written by the authors, with ChatGPT used only for debugging and for improving comments and documentation. We also used ChatGPT to help identify potentially relevant literature that we might have missed; we then read the original papers ourselves and cited them only when they were relevant.

REFERENCES

Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Anytime-valid answer sufficiency certificates for LLM generation via sequential information lift. *arXiv preprint arXiv:2510.06478*, 2025.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025.

Sher Badshah, Ali Emami, and Hassan Sajjad. Judge, retrieve, or abstain: Uncertainty-guarded LLM judging with provable risk guarantees. *arXiv preprint arXiv:2608.17994*, 2026a. Accepted at COLM 2026.

Sher Badshah, Ali Emami, and Hassan Sajjad. SCOPE: Selective conformal optimized pairwise LLM judging. *arXiv preprint arXiv:2602.13110*, 2026b. Accepted at ICML 2026.

Stephen Bates, Anastasios Angelopoulos, Lihua Lei, Jitendra Malik, and Michael Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6), 2021. Article 43.

Peter Bauer. Multiple testing in clinical trials. *Statistics in Medicine*, 10(6):871–890, 1991.

Leo Breiman. Stacked regressions. *Machine Learning*, 24(1):49–64, 1996.

Camilo Chacón Sartori and José H García. A fixed-budget, cluster-aware standard for LLM-as-a-judge evaluation: A multi-hop RAG stress test. *arXiv preprint arXiv:2605.27789*, 2026.

Junjie Chen, Yuxi Dong, Haitao Li, Weihang Su, Yujia Zhou, Min Zhang, Yiqun Liu, and Qingyao Ai. Benchmarking LLM-as-a-judge for long-form output evaluation. *arXiv preprint arXiv:2606.01629*, 2026.

Ming Cheng, Yusheng Dai, Qiuhong Ke, Zhaolin Chen, and Lizhen Qu. Multi-expert conformal risk control for pairwise LLM judging in open-ended dialogue. *arXiv preprint arXiv:2608.26529*, 2026. Accepted to EMNLP 2026 Main Conference.

Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal*, 21(1):C1–C68, 2018.

C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

Robert T. Clemen. Combining forecasts: A review and annotated bibliography. *International Journal of Forecasting*, 5(4):559–583, 1989.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934.

Thomas G Dietterich. Ensemble methods in machine learning. In *International Workshop on Multiple Classifier Systems*, pp. 1–15. Springer, 2000.

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

Rui Gao. Deadline-valid replay: Anytime-valid testing for LLM inference policies. In *The ICML 2026 Workshop on Hypothesis Testing*, 2026.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*, pp. 4878–4887, 2017.

Christopher R. Genovese, Kathryn Roeder, and Larry Wasserman. False discovery control with p-value weighting. *Biometrika*, 93(3):509–524, 2006.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, Saizhuo Wang, Kun Zhang, Zhouchi Lin, Bowen Zhang, Lionel Ni, Wen Gao, Yuanzhuo Wang, and Jian Guo. A survey on LLM-as-a-judge. *The Innovation*, 7 (6):101253, 2026. ISSN 2666-6758.

Manan Gupta and Dhruv Kumar. Diagnosing LLM judge reliability: Conformal prediction sets and transitivity violations. In *ICML 2026 Workshop on Combining Theory and Benchmarks: Towards A Virtuous Cycle to Understand and Guarantee Foundation Model Performance*, 2026.

Zhengyu Hu, Linxin Song, Jieyu Zhang, Zheyuan Xiao, Tianfu Wang, Zhengyu Chen, Nicholas Jing Yuan, Jianxun Lian, Kaize Ding, and Hui Xiong. Explaining length bias in LLM-based preference evaluations. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 6763–6794, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *International Conference on Learning Representations (ICLR)*, 2025.

Guneet Kohli. Nine judges, two effective votes: Correlated errors undermine LLM evaluation panels. *arXiv preprint arXiv:2605.29800*, 2026.

Tyron Lardy and Muriel F. Pérez-Ortiz. Anytime-valid tests of group invariance through conformal prediction. In *Proceedings of the Fourteenth Symposium on Conformal and Probabilistic Prediction with Applications*, volume 266 of *Proceedings of Machine Learning Research*, pp. 645–665. PMLR, 10–12 Sep 2025.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, 2025.

Tianle Li, Anastasios Angelopoulos, and Wei-Lin Chiang. Does style matter? disentangling style and substance in Chatbot Arena. LMSYS Org blog, `https://www.lmsys.org/blog/2024-08-28-style-control/`, 2024.

Yitao Li. Who drifted: the system or the judge? anytime-valid attribution in LLM evaluation pipelines. *arXiv preprint arXiv:2606.15474*, 2026.

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*, pp. 2511–2522, 2023.

Rachel Longjohn, Shang Wu, Saatvik Kher, Catarina Belém, and Padhraic Smyth. Bayesian evaluation of large language model behavior. *arXiv preprint arXiv:2511.10661*, 2025.

Felipe Maia Polo, Xinhe Wang, Mikhail Yurochkin, Gongjun Xu, Moulinath Banerjee, and Yuekai Sun. Bridging human and LLM judgments: Understanding and narrowing the gap. In *Advances in Neural Information Processing Systems 38 (NeurIPS 2025)*, 2025.

Willi Maurer, Ludwig Hothorn, and Walter Lehmacher. Multiple comparisons in drug clinical trials and preclinical assays: a-priori ordered hypotheses. In *Biometrie in der chemisch-pharmazeutischen Industrie*, volume 6, pp. 3–18. Fischer Verlag, 1995.

Muriel Felipe Pérez-Ortiz, Tyron Lardy, Rianne de Heide, and Peter D Grünwald. E-statistics, group invariance and anytime-valid testing. *The Annals of Statistics*, 52(4):1410–1432, 2024.

Peter M. Robinson. Root-N-consistent semiparametric regression. *Econometrica*, 56(4):931–954, 1988.

Kathryn Roeder and Larry Wasserman. Genome-wide significance levels and weighted hypothesis testing. *Statistical Science*, 24(4):398–413, 2009.

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. LLM-as-a-judge on a budget. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026.

Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society Series A*, 184(2):407–431, 2021.

Rajen D. Shah and Jonas Peters. The hardness of conditional independence testing and the generalised covariance measure. *The Annals of Statistics*, 48(3):1514–1538, 2020. doi: 10.1214/19-AOS1857.

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*, pp. 292–314, 2025.

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.

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. *arXiv preprint arXiv:2404.18796*, 2024.

Bengt von Bahr and Carl-Gustav Esseen. Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$. *The Annals of Mathematical Statistics*, 36(1):299–303, 1965.

Bin Wang and Yan Zhong. NxN e-valuation: Hypothesis certification via a conformal CRT null. *arXiv preprint arXiv:2608.06621*, 2026.

Hongjian Wang, Sanjit Dandapanthula, and Aaditya Ramdas. Anytime-valid FDR control with the stopped e-BH procedure. *Statistics & Probability Letters*, pp. 110512, 2025.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models. *arXiv preprint arXiv:2406.08673*, 2024.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B*, 86(1):1–27, 2024.

Hui Wei, Shenghua He, Tian Xia, Fei Liu, Andy Wong, Jingyang Lin, and Mei Han. Systematic evaluation of LLM-as-a-judge in LLM alignment tasks: Explainable metrics and diverse prompt templates. In *ICLR 2025 Workshop on Building Trust in Language Models and Applications*, 2025.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992.

Tianruo Rose Xu, Vedant Gaur, Liu Leqi, and Tanya Goyal. The progress illusion: Revisiting meta-evaluation standards of LLM evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 19033–19043. Association for Computational Linguistics, 2025.

Zhenghao Zeng, David Arbour, Avi Feller, Ishita Dasgupta, Atanu Sinha, and Edward Kennedy. Handling missing responses under cluster dependence with applications to language model evaluation. *Advances in Neural Information Processing Systems*, 38:56527–56556, 2026.

Wenbo Zhang, Hengrui Cai, and Wenyu Chen. Beyond the singular: The essential role of multiple generations in effective benchmark evaluation and analysis. *arXiv preprint arXiv:2502.08943*, 2025.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Yilun Zhou, Austin Xu, Peifeng Wang, Caiming Xiong, and Shafiq Joty. Evaluating judges as evaluators: The JETTS benchmark of LLM-as-judges as test-time scaling evaluators. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 79436–79471. PMLR, 13–19 Jul 2025.

SUPPLEMENTARY MATERIAL

A THE ABC ALGORITHM

Algorithm 1 details the implementation of the ABC panel for one response per prompt. With several responses as introduced in Section 3.5, the split, the folds and the bootstrap resample prompts, the standard error in line 4 sums $\widehat{\psi}_{ij}$ within each prompt, and the p-values in lines 15 and 20 are the betting p-values.

Algorithm 1 The ABC panel with one evaluation per prompt.

Require: Independent training set $\mathcal{I}$ (n prompts) and certification set $\mathcal{I}_{\mathrm{c}}$, each with judge scores Z, surface features X, and human grades Y; error level ε; $\delta = \alpha = 0.05$; $\rho_{\min} = 0.05$; $K = 5$; $\lambda_h = 0.01$; penalty grid $\Lambda = \{10^{-3}, 3 \cdot 10^{-3}, 10^{-2}, 3 \cdot 10^{-2}\}$; $L = 60$ targets q_l evenly spaced from 0.05 to 1; $B = 100$.

Step A: admit

1: Split $\mathcal{I}$ at random into folds $I_1, \ldots, I_K$; standardize X on $\mathcal{I}$.

2: For each k, fit $\widehat{\mu}^{(-k)}$ by logistic regression of Y on X and $\widehat{\nu}_j^{(-k)}$ by linear regression of Z_j on X, $j = 1, \ldots, J$, using $\mathcal{I} \setminus I_k$, each with ridge penalty 0.01.

3: For $i \in I_k$, set $\widehat{U}_i = Y_i - \widehat{\mu}^{(-k)}(X_i)$ and $\widehat{V}_{ij} = Z_{ij} - \widehat{\nu}_j^{(-k)}(X_i)$; center both over $\mathcal{I}$.

4: For each j, compute $\widehat{\rho}_j$, $\widehat{\psi}_{ij}$, and $\widehat{\mathrm{se}}_j$ as in Section 3.2; set $\mathcal{C}_j = [\widehat{\rho}_j - z_{1-\alpha/(2J)}\widehat{\mathrm{se}}_j, \widehat{\rho}_j + z_{1-\alpha/(2J)}\widehat{\mathrm{se}}_j]$.

5: $\mathcal{A} \leftarrow \{j : \mathcal{C}_j \cap [-\rho_{\min}, \rho_{\min}] = \varnothing\}$; if $\mathcal{A} = \varnothing$, return $\widehat{\tau} = \infty$ and zero coverage, referring every prompt to a human grader.

Step B: blend

6: Standardize Z_j, $j \in \mathcal{A}$, on $\mathcal{I}$ to obtain $\widetilde{Z}_j$.

7: Stage 1: fit $\widehat{h}$ by ridge logistic regression of Y on X using $\mathcal{I}$, with penalty λ_h; for each k, fit $\widehat{h}^{(-k)}$ in the same way using $\mathcal{I} \setminus I_k$.

8: Stage 2: choose $\lambda \in \Lambda$ by K-fold cross-validated logistic loss for the ℓ_1 logistic regression of Y on $(\widetilde{Z}_j)_{j \in \mathcal{A}}$ and $\widehat{h}^{(-k(i))}(X_i)$; fit $\widehat{\theta}$ on $\mathcal{I}$ using the selected penalty.

9: $\eta(z, x) \leftarrow \widehat{\theta}_0 + \sum_{j \in \mathcal{A}} \widehat{\theta}_j \widetilde{z}_j + \widehat{\theta}_h \widehat{h}(x)$.

Step C: certify

10: $\eta_i \leftarrow \eta(Z_i, X_i)$ for $i \in \mathcal{I} \cup \mathcal{I}_{\mathrm{c}}$.

11: Set $\tau_l \leftarrow \widehat{Q}_{1-q_l}\{|\eta_i| : i \in \mathcal{I}\}$, $l = 1, \ldots, L$; merge equal values, reindex so that $\tau_1 > \cdots > \tau_L$, and update L to the number of distinct thresholds.

12: For each l, set $a_{il} \leftarrow \mathbb{I}\{|\eta_i| \geq \tau_l\}$ and $w_{il} \leftarrow a_{il}\mathbb{I}\{\mathbb{I}\{\eta_i > 0\} \neq Y_i\}$ for $i \in \mathcal{I}$; set $c_l \leftarrow n^{-1} \sum_{i \in \mathcal{I}} a_{il}$.

13: **for** $b = 1, \ldots, B$ **do**

14: Draw n training prompts with replacement from $\mathcal{I}$.

15: Compute $p_l^{(b)}$ from the resampled (a_{il}, w_{il}) using (4), $l = 1, \ldots, L$.

16: For each start $s \in \{1, \ldots, L\}$, let $d_s^{(b)}$ be the last $l \in \{s, \ldots, L\}$ with $p_s^{(b)}, \ldots, p_l^{(b)} \leq \delta$; set $S_s^{(b)} \leftarrow c_{d_s^{(b)}}$, or zero if no such level exists.

17: **end for**

18: $l_0 \leftarrow \arg\max_{s \in \{1, \ldots, L\}} B^{-1} \sum_{b=1}^{B} S_s^{(b)}$; initialize $\widehat{\tau} \leftarrow \infty$.

19: **for** $l = l_0, \ldots, L$ **do**

20: Compute m_l and e_l on $\mathcal{I}_{\mathrm{c}}$, and p_l using (4); if $p_l \leq \delta$, set $\widehat{\tau} \leftarrow \tau_l$; otherwise exit the loop.

21: **end for**

22: **return** $\widehat{\tau}$ and the observed coverage $|\mathcal{I}_{\mathrm{c}}|^{-1} \sum_{i \in \mathcal{I}_{\mathrm{c}}} \mathbb{I}\{|\eta_i| \geq \widehat{\tau}\}$, taken as zero if $\widehat{\tau} = \infty$.

B PROMPTS WITH MULTIPLE RESPONSES

We treat the case of several graded responses per prompt (Section 3.5); several comparisons per prompt are handled in the same way. Let $g = 1, \ldots, G_{\mathrm{tr}}$ index the training prompts, with prompt g

contributing m_g graded responses, indexed by $i = 1, \ldots, m_g$. Responses within a prompt share the question and grading instructions and may therefore be dependent, whereas different prompts are independent draws from a common law P. Let $K_{\max}$ denote the largest number of graded responses that a prompt can contribute, so that $1 \le m_g \le K_{\max}$. Write g_i for the prompt containing training response i, let $n = \sum_{g=1}^{G_{\mathrm{tr}}} m_g$ be the total number of training responses, and let G_{c} denote the number of certification prompts. For response (g, i), write $O_{gi} = (Z_{gi}, X_{gi}, Y_{gi})$. The value of $K_{\max}$ is fixed by the sampling protocol and is listed for each dataset in Table 4.

Because responses within a prompt may be dependent, we split the data by prompt and keep all responses from a prompt together. At the same time, coverage and selective risk are defined over individual responses. For a function f of one response, define the response-weighted law R by

$$\mathbb{E}_R f(O) = \frac{\mathbb{E}_P\big[\sum_{i=1}^{m_g} f(O_{gi})\big]}{\mathbb{E}_P[m_g]}. \tag{5}$$

It is the law of a response drawn uniformly from the pooled responses of many independent prompts. Population expectations and probabilities involving individual responses in Steps A to C are taken under R, whereas expectations involving whole prompts are taken under P. When $m_g \equiv 1$, the two laws coincide.

Step A. The parameter ρ_j and its estimator remain as in Section 3.2, with population expectations under R and sample averages over the n training responses. Both regressions are cross-fitted by prompt. To account for dependence within prompts, sum the terms $\widehat{\psi}_{ij}$ within each prompt before computing the variance,

$$\widehat{S}_{gj} = \sum_{i \in g} \widehat{\psi}_{ij}, \qquad \text{and} \qquad \widehat{\mathrm{se}}_j^{\,2} = \frac{G_{\mathrm{tr}}}{G_{\mathrm{tr}} - 1} \, \frac{\sum_{g=1}^{G_{\mathrm{tr}}} \widehat{S}_{gj}^2}{n^2},$$

where G_{tr} is the number of training prompts. The prompt sums $\widehat{S}_{gj}$ are independent across prompts and sum to zero, so this is their sample variance, scaled to the mean over responses. When $m_g \equiv 1$, so that $G_{\mathrm{tr}} = n$, this reduces to (2). Under the assumptions of Proposition 4, $\sqrt{G_{\mathrm{tr}}}(\widehat{\rho}_j - \rho_j) \Rightarrow N(0, \sigma_j^2)$ and $G_{\mathrm{tr}}\, \widehat{\mathrm{se}}_j^{\,2} \to_p \sigma_j^2$.

Step B. All folds are formed by prompt, so the cross-fitted surface prediction $\widehat{h}^{(-k(i))}(X_i)$ of a response is fitted without any response to the same prompt. The fitting objectives are unchanged.

Step C. Certification must account for dependence within prompts while controlling the error rate over accepted responses. For candidate l, let $A_{g,l}$ count the responses accepted from certification prompt g, and let $E_{g,l}$ count those graded incorrectly. Then

$$C(\mathcal{A}, \eta, \tau_l) = \frac{\mathbb{E}_P A_{g,l}}{\mathbb{E}_P m_g}, \qquad \pi(\mathcal{A}, \eta, \tau_l) = \frac{\mathbb{E}_P E_{g,l}}{\mathbb{E}_P A_{g,l}},$$

where the risk is defined when $\mathbb{E}_P A_{g,l} > 0$. This agrees with the response-level definition $C(\mathcal{A}, \eta, \tau_l) = \mathbb{P}_R\{|\eta(Z_{\mathcal{A}}, X)| \ge \tau_l\}$ and $\pi(\mathcal{A}, \eta, \tau_l) = \mathbb{P}_R\{\widehat{Y} \ne Y \mid |\eta(Z_{\mathcal{A}}, X)| \ge \tau_l\}$. It is the long-run error rate among accepted responses, not the average of the prompt-specific ratios $E_{g,l}/A_{g,l}$. For positive coverage, $H_{0l} : \pi(\tau_l) \ge \varepsilon$ is equivalent to $\vartheta_l = \mathbb{E}_P(E_{g,l} - \varepsilon A_{g,l}) \ge 0$. We test this latter inequality, which remains defined at zero coverage. Dependence within prompts means that the error count conditional on the accepted count need not be binomial.

Let $\mathcal{T}$ collect everything fixed before the certification data are seen: the training data and everything computed from them, including $\mathcal{A}$, η, the menu and the starting level l_0; the target ε; and the random draws made independently of the data, such as the split, the folds, the bootstrap stream and the order of the certification prompts used below. The random split makes the certification prompts independent of $\mathcal{T}$, and the probabilities below are conditional on $\mathcal{T}$.

For a fixed candidate l, define

$$V_g = \varepsilon + \frac{E_{g,l} - \varepsilon A_{g,l}}{K_{\max}}.$$

The bounds $0 \le E_{g,l} \le A_{g,l} \le K_{\max}$ give $V_g \in [0,1]$, and $\mathbb{E}_P V_g - \varepsilon = \vartheta_l / K_{\max}$. Thus the null is $\mathbb{E}_P V_g \ge \varepsilon$. A prompt with no accepted response has $V_g = \varepsilon$. Each correct accepted grade lowers V_g by $\varepsilon / K_{\max}$, and each error raises it by $(1-\varepsilon)/K_{\max}$; these contributions balance at error rate ε.

Independent prompts make this bounded-mean hypothesis testable by betting (Waudby-Smith & Ramdas, 2024). Before fitting, we draw a data-independent order of the certification prompts and include it in $\mathcal{T}$. In that order, define

$$\mathcal{K}_0 = 1, \qquad \mathcal{K}_t = \prod_{u=1}^{t} \big(1 + \lambda_u(\varepsilon - V_u)\big), \qquad p_l = \Big(\max_{0 \le t \le G_c} \mathcal{K}_t\Big)^{-1},$$

where G_c denotes the number of certification prompts. The stake $\lambda_u \in [0, 0.75/(1-\varepsilon)]$ is chosen from $V_1, \dots, V_{u-1}$ using the formula in Supplement C.1, with tuning level fixed at the target certification level δ.

Conditional on $\mathcal{T}$, under the null, $\mathbb{E}\{1 + \lambda_t(\varepsilon - V_t) \mid \mathcal{T}, V_1, \dots, V_{t-1}\} = 1 + \lambda_t(\varepsilon - \mathbb{E}_P V_g) \le 1$. Each factor is at least $1/4$, so $\mathcal{K}_t$ is a nonnegative supermartingale starting at one. Ville's inequality (Shafer, 2021) gives $\mathbb{P}\{\max_{0 \le t \le G_c} \mathcal{K}_t \ge 1/a \mid \mathcal{T}\} \le a$, or equivalently $\mathbb{P}\{p_l \le a \mid \mathcal{T}\} \le a$, for every $a \in (0,1)$ (Lemma 2, Supplement C.2). Validity uses the stake bounds and the fact that the stake is chosen before the next prompt is observed; the particular choice affects power. The sampling protocol fixes the test before certification: binomial when $K_{\max} = 1$, betting when $K_{\max} > 1$.

The threshold menu and fixed-sequence argument are unchanged. To choose the starting level, however, we perform the bootstrap at the prompt level: a_{il} and w_{il} are summed within each training prompt, the prompts are resampled with replacement, and each replicate computes the betting p-values in a fresh random order of its prompts. Certified coverage remains the fraction of certification responses accepted, or zero if no rule is deployed.

C THEORY

Write $O_{gi} = (Z_{gi}, X_{gi}, Y_{gi})$ for response i of prompt g, $O_g = (m_g, O_{g1}, \dots, O_{gm_g})$ for the prompt, P for the law of a prompt, $\bar{m} = \mathbb{E}_P m_g$ and $n = \sum_g m_g$. The response-weighted law R of (5), with $\bar{m} = \mathbb{E}_P m_g$, coincides with P when $m_g \equiv 1$. Throughout, $\mathcal{A}$ and η are the admitted set and the panel score fitted on the training data, and $\eta(z,x)$ abbreviates $\eta(z_{\mathcal{A}}, x)$. For candidate τ_l let $a_l(z,x) = \mathbb{I}\{|\eta(z,x)| \ge \tau_l\}$, let $d(z,x) = \mathbb{I}\{\eta(z,x) > 0\}$ be the panel's grade $\widehat{Y}$, and let

$$A_{g,l} = \sum_{i=1}^{m_g} a_l(Z_{gi}, X_{gi}), \qquad E_{g,l} = \sum_{i=1}^{m_g} a_l(Z_{gi}, X_{gi}) \mathbb{I}\{d(Z_{gi}, X_{gi}) \ne Y_{gi}\},$$

$C_l = C(\mathcal{A}, \eta, \tau_l) = \mathbb{E}_P A_{g,l}/\bar{m}$, $M_l = \mathbb{E}_P E_{g,l}/\bar{m}$, $\vartheta_l = \mathbb{E}_P(E_{g,l} - \varepsilon A_{g,l}) = \bar{m}(M_l - \varepsilon C_l)$ (the excess-error mass of Supplement B), and $\pi_l = \pi(\mathcal{A}, \eta, \tau_l) = M_l / C_l$ when $C_l > 0$. Write G_{tr} and G_c for the numbers of training and certification prompts. Throughout this appendix, prompts are i.i.d. draws from P and $1 \le m_g \le K_{\max}$ almost surely, with $K_{\max}$ fixed by the sampling protocol (Appendix B).

Let $\mathcal{T}$ collect everything fixed before the certification data are seen: the training data; everything fitted from them, namely the admitted set $\mathcal{A}$, the score η, the menu $\tau_1 > \cdots > \tau_L$ and the starting level l_0; the target ε; and the random draws made independently of the data, namely the split, the folds, the bootstrap stream and the certification order of Supplement B. The random split makes the certification prompts independent of $\mathcal{T}$ (Remark 3), and conditioning on $\mathcal{T}$ is what "given the estimated $\mathcal{A}$ and η" means in Lemma 1 and Theorem 1. All population quantities hold fixed the rules determined by $\mathcal{T}$. Below, H_{0l} denotes the null $\vartheta_l \ge 0$; at positive coverage it is the null $\pi(\mathcal{A}, \eta, \tau_l) \ge \varepsilon$ of Section 3.4, and it remains defined at zero coverage (Supplement B).

The two results below are the general forms of Lemma 1 and Theorem 1; they cover one or several responses per prompt.

Lemma 2 (Valid p-values)**.** *Fix a $\mathcal{T}$-measurable $\varepsilon \in (0,1)$. For every candidate l, the p-value p_l, exact binomial* (4) *when $K_{\max} = 1$ and betting on prompt sums when $K_{\max} > 1$ (Appendix B), satisfies*

$$\mathbb{P}(p_l \le a \mid \mathcal{T}) \le a \qquad \text{for all } a \in (0,1) \text{ whenever } \vartheta_l \ge 0.$$

In particular $\mathbb{P}(p_l \le \delta \mid \mathcal{T}) \le \delta$ under H_{0l}, which is Lemma 1 when $K_{\max} = 1$.

Theorem 2 (Safety of Step C)**.** *Given $\mathcal{T}$, test the candidates $l_0, l_0+1, \dots, L$ in this order at level $\delta \in (0,1)$, stop at the first nonrejection, and deploy the last rejected candidate $\widehat{l}$, so that $\widehat{\tau} = \tau_{\widehat{l}}$, or nothing, so that $\widehat{\tau} = \infty$. Let $\widehat{\mathcal{R}}$ be the set of rejected candidates. Then*

$$\mathbb{P}\{\exists\, l \in \widehat{\mathcal{R}} : \vartheta_l \geq 0 \mid \mathcal{T}\} \leq \delta, \tag{6}$$

which is control of the family-wise error rate, and the deployed rule satisfies

$$\mathbb{P}\{\widehat{\tau} < \infty,\ \pi(\mathcal{A}, \eta, \widehat{\tau}) > \varepsilon \mid \mathcal{T}\} \leq \delta. \tag{7}$$

Both bounds hold marginally over $\mathcal{T}$. Inequality (7)*, averaged over $\mathcal{T}$, is* (3) *with its convention that the event is empty when nothing is deployed; with $K_{\max} = 1$ it is Theorem 1.*

The theorem needs no calibrated probabilities, no correct model, no successful admission test, no monotone risks and no independence between the p-values. Bootstrap selection of l_0 is a function of the training data and an independent random stream, so it affects coverage, not validity. The bound is on the joint event of deployment and excess risk, not conditional on deployment, and it concerns one candidate family and one target ε at a time (Remark 5).

C.1 The betting stake

Fix a candidate and suppress l. The certification prompts are taken in the order contained in $\mathcal{T}$, and $\mathcal{F}_t = \sigma(\mathcal{T}, V_1, \dots, V_t)$. Start from $\widehat{\mu}_0 = 1/2$, $\widehat{\sigma}_0^2 = 1/4$ and $Q_0 = 0$. Before observing V_t, bet

$$\lambda_t = \min\Big\{\frac{0.75}{1-\varepsilon},\ \sqrt{\frac{2\log(1/\delta)}{\widehat{\sigma}_{t-1}^2\, t\log(1+t)}}\Big\}; \tag{8}$$

after observing it, update

$$Q_t = Q_{t-1} + (V_t - \widehat{\mu}_{t-1})^2, \qquad \widehat{\mu}_t = \frac{1/2 + \sum_{u\leq t} V_u}{t+1}, \qquad \widehat{\sigma}_t^2 = \frac{1/4 + Q_t}{t+1} \geq \frac{1}{4(t+1)}.$$

Thus $\widehat{\sigma}_{t-1}^2$ is a strictly positive predictable variance proxy, λ_t is $\mathcal{F}_{t-1}$-measurable, $0 \leq \lambda_t \leq 0.75/(1-\varepsilon)$, and $1 + \lambda_t(\varepsilon - V_t) \geq 1/4$ for $V_t \in [0,1]$. The proofs use only these properties, not consistency of the proxy. The tuning level is fixed at δ, and each candidate uses its own observations.

C.2 Proof of Lemma 2

Proof. Condition on $\mathcal{T}$ throughout, so the candidate rules, the menu and the certification order are fixed, and the G_c certification prompts are i.i.d. from P (Remark 3). Fix a candidate and suppress l; throughout, $a \in (0,1)$ denotes a level.

Exact-binomial branch ($K_{\max} = 1$). Let $N = \sum_g A_g$, $W = \sum_g E_g$ and $F_{m,r}(w) = \mathbb{P}\{\mathrm{Binomial}(m,r) \leq w\}$ for $r \in [0,1]$; the p-value is $p = F_{N,\varepsilon}(W)$ for $N > 0$ and $p = 1$ for $N = 0$. If $C = 0$, then $N = W = 0$ almost surely and $p = 1$. Let $C > 0$. Each prompt is not accepted, accepted and correct, or accepted and wrong, with probabilities $1-C$, $C(1-\pi)$ and $C\pi$, independently across prompts. Hence $(G_c - N, N - W, W)$ is multinomial, and

$$W \mid (\mathcal{T}, N = m) \sim \mathrm{Binomial}(m, \pi) \tag{9}$$

for every m with positive probability. For $m \geq 1$ let $k_m(a) = \max\big(\{-1\} \cup \{w \in \{0,\dots,m\} : F_{m,\varepsilon}(w) \leq a\}\big)$, with $F_{m,p}(-1) = 0$, so that $\{p \leq a\} = \{W \leq k_m(a)\}$ on $\{N = m\}$. Under H_{0l}, $\pi \geq \varepsilon$; the binomial law is stochastically increasing in its success probability (couple the indicators through common uniforms), so

$$\mathbb{P}(p \leq a \mid \mathcal{T}, N = m) = F_{m,\pi}(k_m(a)) \leq F_{m,\varepsilon}(k_m(a)) \leq a.$$

For $m = 0$ the event is empty. Averaging over N proves the lemma for this branch.

Betting branch. Let $V_g = \varepsilon + (E_g - \varepsilon A_g)/K_{\max}$. Since $0 \leq E_g \leq A_g \leq K_{\max}$, $V_g \in [0,1]$ and

$$\mathbb{E}_P V_g - \varepsilon = \vartheta / K_{\max} \tag{10}$$

at positive and at zero coverage, so H_{0l} is exactly $\mathbb{E}_P V_g \geq \varepsilon$. With the stakes of Supplement C.1, let $\mathcal{K}_0 = 1$, $\mathcal{K}_t = \prod_{u \leq t}\{1 + \lambda_u(\varepsilon - V_u)\}$ and $p = (\max_{0 \leq t \leq G_c} \mathcal{K}_t)^{-1}$. The next prompt is independent of $\mathcal{F}_{t-1}$ and λ_t is $\mathcal{F}_{t-1}$-measurable, so under H_{0l}

$$\mathbb{E}\{1 + \lambda_t(\varepsilon - V_t) \mid \mathcal{F}_{t-1}\} = 1 + \lambda_t(\varepsilon - \mathbb{E}_P V_g) \leq 1.$$

Each factor is at least $1/4$ by the stake bound, so $(\mathcal{K}_t)_{t \leq G_c}$ is a nonnegative supermartingale with $\mathcal{K}_0 = 1$; only the nonnegativity of the factors is used below, and the value $1/4$ plays no role. Let T_a be the first t with $\mathcal{K}_t \geq 1/a$, or G_c if there is none. This stopping time is bounded, so $\mathbb{E}(\mathcal{K}_{T_a} \mid \mathcal{T}) \leq \mathcal{K}_0 = 1$ by optional stopping, while $\mathcal{K}_{T_a} \geq 1/a$ on $\{p \leq a\}$ and $\mathcal{K}_{T_a} \geq 0$ elsewhere. Hence $\mathbb{P}(p \leq a \mid \mathcal{T}) \leq a$, the finite-horizon maximal inequality of testing by betting (Shafer, 2021; Waudby-Smith & Ramdas, 2024). When $G_c = 0$, or when $C = 0$ (every factor is then one), $p = 1$. □

Remark 1 (What is and is not conditioned on). Equation (9) conditions on the accepted count of one candidate, not on the scores and covariates of the accepted responses; given those, the error count would be Poisson–binomial with unequal success probabilities. Acceptance and error need not be independent: the argument uses only i.i.d. prompts and one rule fixed by $\mathcal{T}$. "Exact" means valid in finite samples; the nonrandomized binomial p-value is conservative rather than uniform. The branch, binomial or betting, is fixed by $K_{\max}$ before the certification data are seen. The bootstrap p-values computed on the training data only choose l_0 and need not be valid p-values.

C.3 Proof of Theorem 2 and its scope

Proof of Theorem 2. Condition on $\mathcal{T}$; the menu, the starting level l_0 and the order are then fixed. If none of $l_0, \ldots, L$ is a true null, no false rejection is possible. Otherwise let l^* be the first true null in the order; its index is a function of P and $\mathcal{T}$, not of the certification data. Testing stops at the first nonrejection, so a true null can be rejected only if the procedure reaches and rejects l^*:

$$\{\exists\, l \in \widehat{\mathcal{R}} : \vartheta_l \geq 0\} \subseteq \{p_{l^*} \leq \delta\},$$

and Lemma 2 gives (6). The inclusion need not be an equality: an earlier false null that is not rejected stops the sequence before l^*. For (7), note first that a zero-coverage candidate has $p = 1$ and is never rejected, and that an empty menu or an empty certification set deploys nothing; hence on $\{\widehat{\tau} < \infty\}$ the deployed candidate is rejected and has $C_{\widehat{l}} > 0$, so $\pi_{\widehat{l}}$ is defined. A deployed candidate with $\pi_{\widehat{l}} > \varepsilon$ then has $\vartheta_{\widehat{l}} = \bar{m} C_{\widehat{l}}(\pi_{\widehat{l}} - \varepsilon) > 0$ and is a rejected true null, so the event in (7) is contained in the event in (6). Iterated expectation over $\mathcal{T}$ gives the marginal bounds, and (3) follows with its convention for $\widehat{\tau} = \infty$. □

This is the fixed-sequence argument (Bauer, 1991; Maurer et al., 1995), used for risk control in Learn then Test (Angelopoulos et al., 2025).

Remark 2 (Selection before and after testing). Validity rests on two facts: an unsafe deployment requires a false rejection, and everything the procedure uses besides the certification data belongs to $\mathcal{T}$. The first gives (3) for any measurable choice among the rejected candidates; Step C takes the last one, which has the largest acceptance set. The second lets the menu, the target and the starting level be arbitrary measurable functions of the training data and of independent randomization, including the bootstrap choice of l_0 in Section 3.4, whose consistency is never assumed; their quality affects coverage, not validity. What the argument excludes is a starting level, an order or a decision to continue that depends on the certification data.

Remark 3 (Random split under the sampling model). Let $O_1, \ldots, O_N$ be i.i.d. from P, and let a data-independent partition assign N_T prompts to training and N_C to certification. Given the partition, the two samples are independent with laws $P^{\otimes N_T}$ and $P^{\otimes N_C}$. Conditioning further on the training data, on the objects fitted from them and on the independent design randomization, including the certification order, that is, conditioning on $\mathcal{T}$, therefore leaves the certification prompts i.i.d. from P. The guarantee is a statement under this sampling model, not one conditional on the realized benchmark.

Remark 4 (Filtering on Arena). On Arena each comparison carries one human verdict, so $m_g = K_{\max} = 1$, and comparisons are modeled as i.i.d. draws from one law even when they share prompt text. Let $D_g = D(O_g)$ indicate that comparison g passes the eligibility and decisive filters, which are fixed functions of that comparison alone, and let $P_D = P(\cdot \mid D_g = 1)$ with $\mathbb{P}(D_g = 1) > 0$.

Given $\mathcal{T}$ and the number M of retained certification comparisons, the retained comparisons are i.i.d. from P_D in their original order, whichever comparisons were retained, and the menu and the starting level do not use M. Lemma 2 and Theorem 2 therefore hold given M, and averaging over M gives the unconditional bounds; $M = 0$ deploys nothing. The certified quantities are $\mathbb{P}(d \neq Y \mid \text{accept}, D = 1)$ and $\mathbb{P}(\text{accept} \mid D = 1)$; they say nothing about comparisons a human would call a tie. A filter that looked at other comparisons in the split would break this argument.

Remark 5 (Scope). The theorem covers one candidate family and one $\mathcal{T}$-measurable target $\varepsilon \in (0, 1)$. A relative target $\varepsilon = re_0$ computed on the training data is therefore allowed, but separate targets, families or partitions carry separate guarantees rather than a simultaneous one. The bound is on the joint event of deployment and excess risk, not on excess risk given deployment. It bounds neither the coverage from below, so the reported certification coverage is descriptive, nor the error on a particular batch, on a subgroup, or under distribution shift.

C.4 Oracle coverage and calibrated forecasts

Here a score η is fixed and all probabilities are under R. The calibrated forecast is $\widetilde{p}(\eta) = \mathbb{P}_R(Y = 1 \mid \eta)$, and $q(\eta) = \min\{\widetilde{p}(\eta), 1 - \widetilde{p}(\eta)\}$ is the Bayes error given the score.

Definition 2 (Oracle coverage)**.** *For $\varepsilon \in [0, 1]$, the randomized oracle coverage is*

$$V_\varepsilon(\eta) = \sup\big\{\mathbb{E}_R a(\eta) :\ a \text{ measurable},\ 0 \le a \le 1,\ \mathbb{E}_R[a(\eta)\{q(\eta) - \varepsilon\}] \le 0\big\}, \tag{11}$$

and, with $a_\tau = \mathbb{I}\{|\eta| \ge \tau\}$ and $d = \mathbb{I}\{\eta > 0\}$, the deterministic threshold coverage is

$$D_\varepsilon(\eta) = \sup\Big(\{0\} \cup \big\{\mathbb{E}_R a_\tau :\ \mathbb{E}_R a_\tau > 0,\ \mathbb{E}_R[a_\tau \mathbb{I}\{d \neq Y\}] \le \varepsilon\, \mathbb{E}_R a_\tau\big\}\Big). \tag{12}$$

V_ε allows any acceptance probability measurable in η with the Bayes grade; D_ε is the largest coverage of a population-safe deterministic threshold on $|\eta|$. Neither is claimed to be attained by the finite menu or certified in finite samples.

Proposition 2. *For every score and $\varepsilon \in [0, 1]$,*

$$D_\varepsilon(\eta) \le V_\varepsilon(\eta) = \inf_{\lambda \ge 0} \mathbb{E}_R\big[\big(1 - \lambda\{q(\eta) - \varepsilon\}\big)_+\big]. \tag{13}$$

Proof. Suppress η and R. Every grade measurable in η has conditional error at least q, so a feasible threshold satisfies $\mathbb{E}[a_\tau q] \le \mathbb{E}[a_\tau \mathbb{I}\{d \neq Y\}] \le \varepsilon \mathbb{E} a_\tau$ and is feasible for V_ε; hence $D_\varepsilon \le V_\varepsilon$. For feasible a and $\lambda \ge 0$, $\mathbb{E}a \le \mathbb{E}[a\{1 - \lambda(q - \varepsilon)\}] \le \mathbb{E}[(1 - \lambda(q - \varepsilon))_+]$, which is weak duality. We show equality, allowing atoms in the law of q.

If $\mathbb{E}q \le \varepsilon$, then $a \equiv 1$ is feasible and $\lambda = 0$ gives one on both sides. If $\mathbb{E}q > \varepsilon$ and $q \ge \varepsilon$ almost surely, a feasible a can be positive only on $\{q = \varepsilon\}$, so the primal value is $\mathbb{P}(q = \varepsilon)$, and the dual integrand decreases to $\mathbb{I}\{q = \varepsilon\}$ as $\lambda \to \infty$ (dominated convergence).

Otherwise $\mathbb{E}q > \varepsilon$ and $\mathbb{P}(q < \varepsilon) > 0$, so $\varepsilon < 1/2$. For $c \in [\varepsilon, 1/2]$ let $h(c) = \mathbb{E}[(q - \varepsilon)\mathbb{I}\{q \le c\}]$, which is nondecreasing and right-continuous with $h(\varepsilon) < 0 < h(1/2)$, and let $c^* = \inf\{c : h(c) \ge 0\}$. Then $c^* > \varepsilon$ and $h_- := \mathbb{E}[(q - \varepsilon)\mathbb{I}\{q < c^*\}] \le 0 \le h(c^*)$. If $p_* := \mathbb{P}(q = c^*) > 0$, set $\beta = -h_-/\{(c^* - \varepsilon)p_*\} \in [0, 1]$; otherwise $h_- = h(c^*) = 0$ and $\beta = 0$. The acceptance $a^* = \mathbb{I}\{q < c^*\} + \beta \mathbb{I}\{q = c^*\}$ satisfies $\mathbb{E}[a^*(q - \varepsilon)] = 0$. With $\lambda^* = 1/(c^* - \varepsilon)$ the factor $1 - \lambda^*(q - \varepsilon)$ is positive below c^*, zero at c^* and negative above it, so $\mathbb{E}[(1 - \lambda^*(q - \varepsilon))_+] = \mathbb{E}[a^*\{1 - \lambda^*(q - \varepsilon)\}] = \mathbb{E}a^*$, which proves equality. □

Lemma 3 (Sharper forecasts cover more)**.** *Let η_1, η_2 be scores for the same outcome under R, with $\widetilde{p}_1 = \widetilde{p}(\eta_1)$ and $\widetilde{p}_2 = \widetilde{p}(\eta_2)$. If $\widetilde{p}_1 \preceq_{\mathrm{cx}} \widetilde{p}_2$, that is, $\mathbb{E}\phi(\widetilde{p}_1) \le \mathbb{E}\phi(\widetilde{p}_2)$ for every convex ϕ on $[0, 1]$, then $V_\varepsilon(\eta_1) \le V_\varepsilon(\eta_2)$ for every ε. This holds in particular when $\sigma(\eta_1) \subseteq \sigma(\eta_2)$.*

Proof. For each $\lambda \ge 0$, $\phi_\lambda(p) = (1 - \lambda\{\min(p, 1 - p) - \varepsilon\})_+$ is convex, because $\min(p, 1 - p)$ is concave and the positive part of a convex function is convex. So $\mathbb{E}\phi_\lambda(\widetilde{p}_1) \le \mathbb{E}\phi_\lambda(\widetilde{p}_2)$ for every λ; take the infimum and use (13). Under nested information, $\widetilde{p}_1 = \mathbb{E}(\widetilde{p}_2 \mid \eta_1)$ by the tower property, and conditional Jensen gives the convex order. □

C.5 A FIXED-CLASS INEQUALITY

Proposition 3. *Let $\mathcal{F}$ be a fixed class of rules, $C(f)$ and $M(f)$ their coverage and accepted-error mass under R, and $C_n(f)$, $M_n(f)$ the empirical versions on one sample. Suppose that, on an event of probability at least $1-\delta'$, $\sup_{f\in\mathcal{F}}|C_n(f)-C(f)| \le r_n$ and $\sup_{f\in\mathcal{F}}|M_n(f)-M(f)| \le r_n$. Let $\widehat{f} \in \mathcal{F}$ satisfy $M_n(\widehat{f}) \le \varepsilon C_n(\widehat{f})$ and let $\Delta_n = \sup\{C_n(f) : f \in \mathcal{F},\ M_n(f) \le \varepsilon C_n(f)\} - C_n(\widehat{f}) \ge 0$. On that event,*

$$M(\widehat{f}) \le \varepsilon C(\widehat{f}) + (1+\varepsilon) r_n,$$

and for every $f_\circ \in \mathcal{F}$ and $\gamma > 0$ with $M(f_\circ) \le (\varepsilon-\gamma)C(f_\circ)$ and $\gamma C(f_\circ) \ge (1+\varepsilon)r_n$,

$$C(\widehat{f}) \ge C(f_\circ) - 2r_n - \Delta_n.$$

Proof. First, $M(\widehat{f}) \le M_n(\widehat{f}) + r_n \le \varepsilon C_n(\widehat{f}) + r_n \le \varepsilon C(\widehat{f}) + (1+\varepsilon)r_n$. Second, $M_n(f_\circ) - \varepsilon C_n(f_\circ) \le M(f_\circ) - \varepsilon C(f_\circ) + (1+\varepsilon)r_n \le -\gamma C(f_\circ) + (1+\varepsilon)r_n \le 0$, so $f_\circ$ is empirically feasible, and $C(\widehat{f}) \ge C_n(\widehat{f}) - r_n \ge C_n(f_\circ) - \Delta_n - r_n \ge C(f_\circ) - 2r_n - \Delta_n$. □

C.6 THE ADMISSION TEST

We extend the admission test in Section 3.2 to allow dependence among responses from the same prompt. Throughout this subsection, expectations for individual responses are under R, while expectations over prompts are under P. The regressions of Y and Z_j on X and all cross-fitting folds are constructed at the prompt level, so that all responses from a given prompt are assigned to the same fold. For judge j, define

$$\mu(x) = \mathbb{E}_R(Y \mid X = x), \qquad \nu_j(x) = \mathbb{E}_R(Z_j \mid X = x),$$

and the residuals

$$U = Y - \mu(X), \qquad V_j = Z_j - \nu_j(X).$$

As in Section 3.2, define

$$A_j = \mathbb{E}_R(V_j^2), \qquad B = \mathbb{E}_R(U^2), \qquad C_j = \mathbb{E}_R(UV_j), \qquad \rho_j = \frac{C_j}{\sqrt{A_j B}}.$$

Define

$$\psi_j(O) = \frac{UV_j - C_j}{\sqrt{A_j B}} - \frac{\rho_j}{2}\left\{\frac{V_j^2 - A_j}{A_j} + \frac{U^2 - B}{B}\right\}.$$

For prompt g, define its prompt-level sum by

$$S_{gj} = \sum_{i=1}^{m_g} \psi_j(O_{gi}).$$

Since prompts are independent under P, the asymptotic variance of the estimator is determined by the variability of these prompt-level contributions. In particular, define

$$\sigma_j^2 = \frac{\mathbb{E}_P(S_{gj}^2)}{\bar{m}^2}.$$

Let $I_1, \ldots, I_K$ partition the training prompts independently of their data. For $g \in I_k$, define the cross-fitted residuals

$$\widehat{U}_{gi} = Y_{gi} - \widehat{\mu}^{(-k)}(X_{gi}), \qquad \widehat{V}_{gi,j} = Z_{gi,j} - \widehat{\nu}_j^{(-k)}(X_{gi}),$$

where $\widehat{\mu}^{(-k)}$ and $\widehat{\nu}_j^{(-k)}$ are fitted on the prompts outside I_k. The residuals are centered by their means over all n training responses. Let $\widehat{A}_j$, $\widehat{B}$, and $\widehat{C}_j$ denote the resulting sample moments, and define

$$\widehat{\rho}_j = \frac{\widehat{C}_j}{(\widehat{A}_j \widehat{B})^{1/2}}.$$

Let

$$\widehat{\psi}_{gi,j} = \frac{\widehat{U}_{gi}\widehat{V}_{gi,j} - \widehat{C}_j}{\sqrt{\widehat{A}_j \widehat{B}}} - \frac{\widehat{\rho}_j}{2}\left\{\frac{\widehat{V}_{gi,j}^2 - \widehat{A}_j}{\widehat{A}_j} + \frac{\widehat{U}_{gi}^2 - \widehat{B}}{\widehat{B}}\right\}, \qquad \widehat{S}_{gj} = \sum_{i=1}^{m_g} \widehat{\psi}_{gi,j},$$

which is the plug-in version of ψ_j, written $\widehat{\psi}_{ij}$ in Section 3.2. All quantities involving $\widehat{A}_j$ and $\widehat{B}$ are considered on the event $\widehat{A}_j, \widehat{B} > 0$, whose probability tends to one.

Proposition 4 (Clustered inference for conditional correlation)**.** *Suppose that the training prompts are i.i.d. under P, with $1 \le m_g \le K_{\max} < \infty$, where $K_{\max}$ is fixed, and that $G_{\mathrm{tr}} \to \infty$. Suppose that K and J are fixed and that, for every j, $A_j > 0$, $B > 0$, $\mathbb{E}_R(U^4 + V_j^4) < \infty$ and $\sigma_j^2 > 0$. Assume that for each fold k, the regression fits $\widehat{\mu}^{(-k)}$ and $\widehat{\nu}_j^{(-k)}$ use only data from the other folds, with any randomization independent of the data in fold k, and satisfy*

$$\max_{k \le K}\left\{\|\widehat{\mu}^{(-k)} - \mu\|_{2,R}^2 + \|\widehat{\nu}_j^{(-k)} - \nu_j\|_{2,R}^2\right\} = o_p(G_{\mathrm{tr}}^{-1/2}), \tag{14}$$

where the norms are taken with respect to an independent new response drawn from R, conditional on the fitted functions. Then, for every j,

$$\widehat{\rho}_j - \rho_j = \frac{1}{n}\sum_{g=1}^{G_{\mathrm{tr}}} S_{gj} + o_p(G_{\mathrm{tr}}^{-1/2}), \qquad \sqrt{G_{\mathrm{tr}}}\,(\widehat{\rho}_j - \rho_j) \Rightarrow N(0, \sigma_j^2).$$

Moreover,

$$G_{\mathrm{tr}}\,\widehat{\mathrm{se}}_j^{\,2} \to_p \sigma_j^2, \qquad \textit{where } \widehat{\mathrm{se}}_j^{\,2} = \frac{G_{\mathrm{tr}}}{G_{\mathrm{tr}} - 1}\frac{\sum_{g=1}^{G_{\mathrm{tr}}} \widehat{S}_{gj}^2}{n^2}.$$

Consequently, the Bonferroni confidence intervals

$$\left[\widehat{\rho}_j - z_{1-\alpha/(2J)}\widehat{\mathrm{se}}_j,\ \widehat{\rho}_j + z_{1-\alpha/(2J)}\widehat{\mathrm{se}}_j\right], \qquad j = 1, \ldots, J,$$

simultaneously cover all ρ_j with probability at least $1 - \alpha - o(1)$. Hence, if judge j is admitted only when its confidence interval lies entirely outside $[-\rho_{\min}, \rho_{\min}]$, then

$$\mathbb{P}\{\textit{any false admission}\} \le \alpha + o(1).$$

With one response per prompt ($m_g \equiv 1$), $G_{\mathrm{tr}} = n$, $\sigma_j^2 = \mathbb{E}_R\psi_j^2$, and $\widehat{\mathrm{se}}_j$ is (2); with Slutsky's theorem this gives Proposition 1.

Proof. We prove the result for a fixed judge j. For $g \in I_k$, write

$$\Delta_{\mu,gi} = \widehat{\mu}^{(-k)}(X_{gi}) - \mu(X_{gi}), \qquad \Delta_{\nu,gi,j} = \widehat{\nu}_j^{(-k)}(X_{gi}) - \nu_j(X_{gi}).$$

Then

$$\widehat{U}_{gi} = U_{gi} - \Delta_{\mu,gi}, \qquad \widehat{V}_{gi,j} = V_{gi,j} - \Delta_{\nu,gi,j}.$$

For notational simplicity, let

$$D_{\mu,k} = \|\widehat{\mu}^{(-k)} - \mu\|_{2,R}^2, \qquad D_{\nu,k} = \|\widehat{\nu}_j^{(-k)} - \nu_j\|_{2,R}^2.$$

By (14),

$$\max_{k \le K}(D_{\mu,k} + D_{\nu,k}) = o_p(G_{\mathrm{tr}}^{-1/2}).$$

We first establish the expansions of the three sample moments. For a fixed fold I_k, conditional on the fitted functions $\widehat{\mu}^{(-k)}$ and $\widehat{\nu}_j^{(-k)}$, the prompts in I_k remain independent. For $Q_{\nu,g} = \sum_{i=1}^{m_g} U_{gi}\Delta_{\nu,gi,j}$, $g \in I_k$, we have

$$\mathbb{E}_R(U\Delta_{\nu,j} \mid \widehat{\nu}_j^{(-k)}) = \mathbb{E}_R\Big[\mathbb{E}_R(U \mid X)\Delta_{\nu,j}(X) \mid \widehat{\nu}_j^{(-k)}\Big] = 0,$$

so that, given $\widehat{\nu}_j^{(-k)}$, the prompt sums $Q_{\nu,g}$, $g \in I_k$, are independent with mean $\mathbb{E}_P(Q_{\nu,g} \mid \widehat{\nu}_j^{(-k)}) = \bar{m}\,\mathbb{E}_R(U\Delta_{\nu,j} \mid \widehat{\nu}_j^{(-k)}) = 0$, by the definition (5) of R. No independence within a prompt is used. Moreover, since $m_g \le K_{\max}$,

$$|Q_{\nu,g}|^{4/3} \le m_g^{1/3} \sum_{i=1}^{m_g} |U_{gi}\Delta_{\nu,gi,j}|^{4/3} \le K_{\max}^{1/3} \sum_{i=1}^{m_g} |U_{gi}\Delta_{\nu,gi,j}|^{4/3}.$$

By Hölder's inequality,

$$\mathbb{E}_R\,|U\Delta_{\nu,j}|^{4/3} \le \left\{\mathbb{E}_R(U^4)\right\}^{1/3} \left\{\mathbb{E}_R(\Delta_{\nu,j}^2)\right\}^{2/3} = \left\{\mathbb{E}_R(U^4)\right\}^{1/3} D_{\nu,k}^{2/3}.$$

Thus, by the von Bahr–Esseen inequality for independent mean-zero variables with exponent $r = 4/3 \in [1,2]$ (von Bahr & Esseen, 1965), applied conditionally on $\widehat{\nu}_j^{(-k)}$,

$$\mathbb{E}\left[\left|\sum_{g\in I_k} Q_{\nu,g}\right|^{4/3} \;\middle|\; \widehat{\nu}_j^{(-k)}\right] \le 2^{4/3} \sum_{g\in I_k} \mathbb{E}\left(|Q_{\nu,g}|^{4/3} \;\middle|\; \widehat{\nu}_j^{(-k)}\right) \le 2^{4/3} K_{\max}^{1/3}\,\bar{m}\left\{\mathbb{E}_R(U^4)\right\}^{1/3} |I_k|\, D_{\nu,k}^{2/3}.$$

Conditional Markov's inequality then gives, for every $c > 0$,

$$\mathbb{P}\left\{\left|\sum_{g\in I_k} Q_{\nu,g}\right| > c\, G_{\mathrm{tr}}^{3/4} D_{\nu,k}^{1/2} \;\middle|\; \widehat{\nu}_j^{(-k)}\right\} \le \frac{2^{4/3} K_{\max}^{1/3} \bar{m} \{\mathbb{E}_R(U^4)\}^{1/3}\, |I_k|\, D_{\nu,k}^{2/3}}{c^{4/3}\, G_{\mathrm{tr}}\, D_{\nu,k}^{2/3}} \le \frac{2^{4/3} K_{\max}^{1/3} \bar{m} \{\mathbb{E}_R(U^4)\}^{1/3}}{c^{4/3}},$$

since $|I_k| \le G_{\mathrm{tr}}$ (on $\{D_{\nu,k} = 0\}$ the sum is zero). Hence

$$\sum_{g\in I_k} Q_{\nu,g} = O_p\left(G_{\mathrm{tr}}^{3/4} D_{\nu,k}^{1/2}\right) = o_p(G_{\mathrm{tr}}^{1/2}),$$

because $D_{\nu,k}^{1/2} = o_p(G_{\mathrm{tr}}^{-1/4})$ by (14). Summing over the fixed number K of folds gives

$$\sum_{g=1}^{G_{\mathrm{tr}}} \sum_{i=1}^{m_g} U_{gi}\Delta_{\nu,gi,j} = o_p(G_{\mathrm{tr}}^{1/2}).$$

The same argument, with $\mathbb{E}_R(V_j \mid X) = 0$ and $\mathbb{E}_R(V_j^4) < \infty$ in place of $\mathbb{E}_R(U \mid X) = 0$ and $\mathbb{E}_R(U^4) < \infty$, gives

$$\sum_{g=1}^{G_{\mathrm{tr}}} \sum_{i=1}^{m_g} V_{gi,j}\Delta_{\mu,gi} = o_p(G_{\mathrm{tr}}^{1/2}),$$

and likewise

$$\sum_{g=1}^{G_{\mathrm{tr}}} \sum_{i=1}^{m_g} V_{gi,j}\Delta_{\nu,gi,j} = o_p(G_{\mathrm{tr}}^{1/2}), \qquad \sum_{g=1}^{G_{\mathrm{tr}}} \sum_{i=1}^{m_g} U_{gi}\Delta_{\mu,gi} = o_p(G_{\mathrm{tr}}^{1/2}),$$

which are the cross terms needed for $\widehat{A}_j$ and $\widehat{B}$ below.

For the product of the regression errors, conditional Cauchy–Schwarz gives

$$\mathbb{E}_R\left[|\Delta_\mu \Delta_{\nu,j}|\right] \le \left\{\mathbb{E}_R(\Delta_\mu^2)\right\}^{1/2} \left\{\mathbb{E}_R(\Delta_{\nu,j}^2)\right\}^{1/2} = D_{\mu,k}^{1/2} D_{\nu,k}^{1/2} \le \frac{D_{\mu,k} + D_{\nu,k}}{2} = o_p(G_{\mathrm{tr}}^{-1/2}).$$

Hence, by Markov's inequality within each fold,

$$\frac{1}{n} \sum_{g=1}^{G_{\mathrm{tr}}} \sum_{i=1}^{m_g} |\Delta_{\mu,gi}\Delta_{\nu,gi,j}| = o_p(G_{\mathrm{tr}}^{-1/2}).$$

Therefore,

$$\widehat{C}_j - C_j = \frac{1}{n} \sum_{g=1}^{G_{\mathrm{tr}}} \sum_{i=1}^{m_g} (U_{gi}V_{gi,j} - C_j) + o_p(G_{\mathrm{tr}}^{-1/2}). \tag{15}$$

Similarly, using

$$\widehat{V}_{gi,j}^2 = V_{gi,j}^2 - 2V_{gi,j}\Delta_{\nu,gi,j} + \Delta_{\nu,gi,j}^2$$

and

$$\frac{1}{n}\sum_{g,i}\Delta_{\nu,gi,j}^2 = O_p\left(\max_{k\le K} D_{\nu,k}\right) = o_p(G_{\mathrm{tr}}^{-1/2}),$$

which again follows from conditional Markov's inequality within each fold, since $\mathbb{E}\{\sum_{g\in I_k}\sum_i \Delta_{\nu,gi,j}^2 \mid \widehat{\nu}_j^{(-k)}\} = |I_k|\,\bar{m}\,D_{\nu,k}$ and $n \ge G_{\mathrm{tr}}$, together with the cross term $\sum_{g,i} V_{gi,j}\Delta_{\nu,gi,j} = o_p(G_{\mathrm{tr}}^{1/2})$ established above, we obtain

$$\widehat{A}_j - A_j = \frac{1}{n}\sum_{g=1}^{G_{\mathrm{tr}}}\sum_{i=1}^{m_g}(V_{gi,j}^2 - A_j) + o_p(G_{\mathrm{tr}}^{-1/2}). \tag{16}$$

Likewise,

$$\widehat{B} - B = \frac{1}{n}\sum_{g=1}^{G_{\mathrm{tr}}}\sum_{i=1}^{m_g}(U_{gi}^2 - B) + o_p(G_{\mathrm{tr}}^{-1/2}). \tag{17}$$

The centering of the cross-fitted residuals does not affect these expansions. Let $\bar{\widehat{U}}$ and $\bar{\widehat{V}}_j$ be the means of the uncentered residuals over the n training responses; centering replaces $\widehat{C}_j$, $\widehat{A}_j$ and $\widehat{B}$ by $\widehat{C}_j - \bar{\widehat{U}}\bar{\widehat{V}}_j$, $\widehat{A}_j - \bar{\widehat{V}}_j^{\,2}$ and $\widehat{B} - \bar{\widehat{U}}^{\,2}$. Indeed,

$$\bar{\widehat{U}} = \frac{1}{n}\sum_{g,i} U_{gi} - \frac{1}{n}\sum_{g,i}\Delta_{\mu,gi} = O_p(G_{\mathrm{tr}}^{-1/2}) + o_p(G_{\mathrm{tr}}^{-1/4}) = o_p(G_{\mathrm{tr}}^{-1/4}),$$

where the first term is the prompt-level central limit theorem applied to the i.i.d. sums $\sum_i U_{gi}$, which have mean $\bar{m}\,\mathbb{E}_R U = 0$ and finite variance, and the second follows from the Cauchy–Schwarz inequality, $|\frac{1}{n}\sum_{g,i}\Delta_{\mu,gi}| \le (\frac{1}{n}\sum_{g,i}\Delta_{\mu,gi}^2)^{1/2} = o_p(G_{\mathrm{tr}}^{-1/4})$, and similarly $\bar{\widehat{V}}_j = o_p(G_{\mathrm{tr}}^{-1/4})$. Consequently,

$$\bar{\widehat{U}}^{\,2} + \bar{\widehat{V}}_j^{\,2} + |\bar{\widehat{U}}\bar{\widehat{V}}_j| = o_p(G_{\mathrm{tr}}^{-1/2}),$$

which shows that centering contributes only $o_p(G_{\mathrm{tr}}^{-1/2})$ to the three sample moments.

Now let $f(C, A, B) = C(AB)^{-1/2}$. Since $A_j > 0$ and $B > 0$, the gradient is well defined in a neighborhood of the population parameter. A first-order Taylor expansion around (C_j, A_j, B), together with (15)–(17), gives

$$\widehat{\rho}_j - \rho_j = \frac{\widehat{C}_j - C_j}{\sqrt{A_j B}} - \frac{\rho_j}{2}\left\{\frac{\widehat{A}_j - A_j}{A_j} + \frac{\widehat{B} - B}{B}\right\} + o_p(G_{\mathrm{tr}}^{-1/2}),$$

because the second-order remainder is $O_p(G_{\mathrm{tr}}^{-1})$; the Hessian of f is bounded near (C_j, A_j, B), and $(\widehat{C}_j - C_j, \widehat{A}_j - A_j, \widehat{B} - B) = O_p(G_{\mathrm{tr}}^{-1/2})$ by (15)–(17) and the prompt-level central limit theorem applied to the i.i.d. sums $\sum_i(U_{gi}V_{gi,j} - C_j)$, $\sum_i(V_{gi,j}^2 - A_j)$ and $\sum_i(U_{gi}^2 - B)$, which have mean zero and finite variance under the fourth-moment assumption. Substituting (15)–(17) yields

$$\widehat{\rho}_j - \rho_j = \frac{1}{n}\sum_{g=1}^{G_{\mathrm{tr}}} S_{gj} + o_p(G_{\mathrm{tr}}^{-1/2}). \tag{18}$$

Because the prompts are i.i.d. under P, $m_g \le K_{\max}$ and $\mathbb{E}_R(U^4 + V_j^4) < \infty$, the prompt-level sums S_{gj} are i.i.d. with finite second moments. Moreover,

$$\frac{n}{G_{\mathrm{tr}}} = \frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}} m_g \to_p \bar{m}.$$

Therefore, by the central limit theorem,

$$\sqrt{G_{\mathrm{tr}}}(\widehat{\rho}_j - \rho_j) = \frac{\sqrt{G_{\mathrm{tr}}}}{n}\sum_{g=1}^{G_{\mathrm{tr}}} S_{gj} + o_p(1) = \frac{1}{\bar{m}\sqrt{G_{\mathrm{tr}}}}\sum_{g=1}^{G_{\mathrm{tr}}} S_{gj} + o_p(1) \Rightarrow N(0, \sigma_j^2),$$

where $\sigma_j^2 = \frac{\mathbb{E}_P(S_{gj}^2)}{\bar{m}^2}$.

It remains to establish consistency of the variance estimator. Since $A_j, B > 0$ and $|\widehat{A}_j - A_j| + |\widehat{B} - B| + |\widehat{C}_j - C_j| = O_p(G_{\mathrm{tr}}^{-1/2})$, the denominators in $\widehat{\psi}_{gi,j}$ are bounded away from zero with probability tending to one, and $|\widehat{\rho}_j| \le 1$. We show that

$$\frac{1}{n}\sum_{g=1}^{G_{\mathrm{tr}}}\sum_{i=1}^{m_g}(\widehat{\psi}_{gi,j} - \psi_j(O_{gi}))^2 = o_p(1). \tag{19}$$

The difference $\widehat{\psi}_{gi,j} - \psi_j(O_{gi})$ has two sources. The first is the replacement of (C_j, A_j, B, ρ_j) by $(\widehat{C}_j, \widehat{A}_j, \widehat{B}, \widehat{\rho}_j)$ in the coefficients of ψ_j; these coefficients are continuous at the population values, the replacements are $O_p(G_{\mathrm{tr}}^{-1/2})$, and they multiply $U_{gi}V_{gi,j}$, $V_{gi,j}^2$, U_{gi}^2 or constants, whose empirical mean squares are $O_p(1)$ by the law of large numbers under the fourth-moment assumption (applied to the i.i.d. prompt sums, with $m_g \le K_{\max}$); their contribution to (19) is therefore $o_p(1)$. The second is the replacement of $(U_{gi}, V_{gi,j})$ by the centered cross-fitted residuals $(\widehat{U}_{gi} - \bar{\widehat{U}}, \widehat{V}_{gi,j} - \bar{\widehat{V}}_j)$, which changes the products UV_j, V_j^2 and U^2 by terms of the form $U\Delta_\nu$, $V_j\Delta_\mu$, $U\Delta_\mu$, $V_j\Delta_\nu$, $\Delta_\mu\Delta_\nu$, Δ_μ^2, Δ_ν^2, and by terms involving the centering constants $\bar{\widehat{U}}, \bar{\widehat{V}}_j = o_p(G_{\mathrm{tr}}^{-1/4})$. It suffices to show that each of these has empirical mean square $o_p(1)$. For the terms linear in a nuisance error, the empirical Cauchy–Schwarz inequality gives

$$\frac{1}{n}\sum_{g,i} U_{gi}^2\Delta_{\nu,gi,j}^2 \le \Big(\frac{1}{n}\sum_{g,i} U_{gi}^4\Big)^{1/2}\Big(\frac{1}{n}\sum_{g,i}\Delta_{\nu,gi,j}^4\Big)^{1/2},$$

where the first factor is $O_p(1)$ by the law of large numbers. For the second factor no L_4 rate for the regression fits is assumed; instead, since all summands are nonnegative,

$$\frac{1}{n}\sum_{g,i}\Delta_{\nu,gi,j}^4 \le \frac{1}{n}\Big(\sum_{g,i}\Delta_{\nu,gi,j}^2\Big)^2 = n\Big(\frac{1}{n}\sum_{g,i}\Delta_{\nu,gi,j}^2\Big)^2 = n\cdot o_p(G_{\mathrm{tr}}^{-1}) = o_p(1),$$

using $\frac{1}{n}\sum_{g,i}\Delta_{\nu,gi,j}^2 = o_p(G_{\mathrm{tr}}^{-1/2})$ from above and $n \le K_{\max}G_{\mathrm{tr}}$. The same bound holds for Δ_μ, and it also covers the quadratic terms: $\frac{1}{n}\sum_{g,i}\Delta_{\mu,gi}^2\Delta_{\nu,gi,j}^2 \le \frac{1}{2n}\sum_{g,i}(\Delta_{\mu,gi}^4 + \Delta_{\nu,gi,j}^4) = o_p(1)$ and $\frac{1}{n}\sum_{g,i}\Delta_{\nu,gi,j}^4 = o_p(1)$. The centering constants contribute terms such as $\bar{\widehat{U}}^2\frac{1}{n}\sum_{g,i}\widehat{V}_{gi,j}^2$ and $\bar{\widehat{U}}^2\bar{\widehat{V}}_j^2$, which are $o_p(1)$ because the empirical second moments of the uncentered residuals are $O_p(1)$ by the bounds just established. Combining these bounds with the bounded coefficients proves (19).

Since $m_g \le K_{\max}$,

$$\frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}}(\widehat{S}_{gj} - S_{gj})^2 \le K_{\max}\frac{n}{G_{\mathrm{tr}}}\frac{1}{n}\sum_{g=1}^{G_{\mathrm{tr}}}\sum_{i=1}^{m_g}(\widehat{\psi}_{gi,j} - \psi_j(O_{gi}))^2 = o_p(1),$$

because $n/G_{\mathrm{tr}} = O_p(1)$ and by (19). Hence,

$$\left|\frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}}\widehat{S}_{gj}^2 - \frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}}S_{gj}^2\right| \le \left\{\frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}}(\widehat{S}_{gj} - S_{gj})^2\right\}^{1/2} \times \left\{\frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}}(\widehat{S}_{gj} + S_{gj})^2\right\}^{1/2} = o_p(1),$$

by Cauchy–Schwarz. Since $\mathbb{E}_P(S_{gj}^2) < \infty$, the law of large numbers gives

$$\frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}} S_{gj}^2 \to_p \mathbb{E}_P(S_{gj}^2).$$

Table 4: Benchmark sizes and human grades. Counts are for eligible responses or comparisons; prompts are the independent units, with $K_{\max}$ evaluations per prompt at most. On HelpSteer2, a prompt is the released source pair. Chatbot Arena is the subset of single-turn comparisons released with the judges of Jung et al. (2025).

Benchmark	Items	Prompts	Human label	$K_{\max}$	Judges
BiGGen	2,704	677	Response quality	4	11
HelpSteer2	14,318	7,159	Helpfulness	2	9
Chatbot Arena	5,169	5,169	Preference	1	11

Consequently,

$$G_{\mathrm{tr}}\widehat{\mathrm{se}}_j^{\,2} = \frac{G_{\mathrm{tr}}^2}{G_{\mathrm{tr}}-1}\frac{1}{n^2}\sum_{g=1}^{G_{\mathrm{tr}}}\widehat{S}_{gj}^2 = \frac{G_{\mathrm{tr}}}{G_{\mathrm{tr}}-1}\left(\frac{G_{\mathrm{tr}}}{n}\right)^2\left\{\frac{1}{G_{\mathrm{tr}}}\sum_{g=1}^{G_{\mathrm{tr}}}\widehat{S}_{gj}^2\right\} \to_p \frac{\mathbb{E}_P(S_{gj}^2)}{\bar{m}^2} = \sigma_j^2.$$

Since $\sigma_j^2 > 0$ by assumption, Slutsky's theorem gives

$$\frac{\widehat{\rho}_j - \rho_j}{\widehat{\mathrm{se}}_j} \Rightarrow N(0,1).$$

Since J is fixed, Bonferroni's inequality yields simultaneous coverage of the J confidence intervals with probability at least $1-\alpha-o(1)$.

On the simultaneous coverage event, if judge j is admitted because its confidence interval lies entirely outside $[-\rho_{\min}, \rho_{\min}]$, then necessarily $|\rho_j| > \rho_{\min}$. Thus a false admission can occur only when the simultaneous coverage event fails, and hence

$$\mathbb{P}\{\text{any false admission}\} \le \alpha + o(1). \qquad \square$$

Remark 6 (Rates and the implemented fits). Write $e_\mu = \widehat{\mu}^{(-k)} - \mu$ and $e_\nu = \widehat{\nu}_j^{(-k)} - \nu_j$. Condition (14) is stated in squared L_2 rates because the variance estimates carry the biases $\mathbb{E}_R e_\mu^2$ and $\mathbb{E}_R e_\nu^2$; it implies the product rate $\|e_\mu\|_{2,R}\|e_\nu\|_{2,R} \le \frac{1}{2}(\|e_\mu\|_{2,R}^2 + \|e_\nu\|_{2,R}^2) = o_p(G_{\mathrm{tr}}^{-1/2})$ that the covariance needs, so no separate product condition is required. The proof is pointwise in P and uses orthogonality and cross-fitting in the manner of Chernozhukov et al. (2018) and Shah & Peters (2020), applied directly to the centered, prompt-clustered correlation. The implementation departs from the proposition in two ways: the fixed ridge penalty 0.01 can converge to a penalized population fit rather than the conditional mean, even in a correct model class, and standardizing X on the whole training set before the cross-fitted fits lets the held-out fold enter through the scaling, since a ridge fit is not scale invariant. Neither is shown to satisfy (14), so the admission intervals are approximate diagnostics. Theorem 2 does not depend on this condition.

D Details of the setup

Table 4 lists the three benchmarks. On BiGGen, $Y_i = \mathbb{I}\{\text{human score} \ge 4\}$ on the 1–5 scale; on HelpSteer2, $Y_i = \mathbb{I}\{\text{helpfulness} \ge 3\}$ on the released 0–4 scale, corresponding to a judge score ≥ 4 on Bridge's 1–5 scale. On Chatbot Arena, each comparison had a decisive human verdict, and $Y_i = \mathbb{I}\{\text{the human prefers A}\}$. A judge reporting probabilities over scores entered as $Z_{ij} = \log P_{ij}(\text{score} \ge 4) - \log P_{ij}(\text{score} \le 3)$, and on Chatbot Arena as $Z_{ij} = \log P_{ij}(\text{A}) - \log P_{ij}(\text{B})$. Five BiGGen judges returned only a score $s \in [1,5]$; we used $Z_{ij} = \mathrm{logit}\{a + (1-2a)(s-1)/4\}$, which is zero at $s = 3$, with $a = 0.025$ for the three single-digit judges and $a = 1/202$ for the two Prometheus judges, whose scores lie on a grid of step 0.2. Every coding has decision boundary zero; none is assumed to give calibrated log-odds of the human grade.

We used the same 38 lightweight features from the bridge method (Maia Polo et al., 2025), including length, formatting, readability, sentiment, and discourse counts, for the prompt and the response, giving 76 columns on each benchmark. On Chatbot Arena, the features entered as differences and sums over the two responses; model identity was excluded. We kept responses or comparisons from

Table 5: Admissions out of 20 partitions, with the mean and range of the conditional correlation, on the partitions of Table 1 (30% training data); the Bonferroni family level is 0.05 over the judges of the benchmark. BiGGen contains scores from two sources, so a model may appear in both.

Judge	adm.	mean $\widehat{\rho}$	range of $\widehat{\rho}$
BiGGen Bench			
Prometheus (ours)	20	0.361	[0.298, 0.427]
Prometheus/BiGGen (ours)	20	0.449	[0.399, 0.486]
GPT-4 (ours)	20	0.432	[0.392, 0.460]
GPT-4-turbo (ours)	20	0.474	[0.434, 0.516]
Claude (ours)	20	0.411	[0.307, 0.496]
GPT-4-turbo (Bridge)	20	0.382	[0.331, 0.421]
GPT-4.1-nano (Bridge)	20	0.325	[0.255, 0.367]
GPT-4.1 (Bridge)	20	0.447	[0.406, 0.522]
GPT-4o-mini (Bridge)	20	0.382	[0.315, 0.464]
Llama-3.1-8B (Bridge)	20	0.294	[0.237, 0.372]
Selene-1-Mini (Bridge)	20	0.389	[0.330, 0.434]
HelpSteer2			
GPT-4.1-nano	20	0.238	[0.185, 0.269]
GPT-4o-mini	20	0.320	[0.286, 0.353]
GPT-4.1	20	0.404	[0.378, 0.429]
Llama-3.1-8B	20	0.279	[0.232, 0.311]
Selene-1-Mini	20	0.356	[0.325, 0.375]
SmolLM2-135M	0	0.056	[0.024, 0.079]
SmolLM2-360M	0	0.011	[−0.011, 0.040]
Qwen2.5-0.5B	0	−0.013	[−0.037, 0.019]
Llama-3.2-1B	0	0.030	[0.012, 0.047]

the same prompt together in every split; on HelpSteer2, a prompt was the released source pair. Each panel contained every judge collected for that benchmark, with Step A determining which were used. Algorithm 1 in Supplement A gives the algorithm settings. The results came from gated runs on one machine on 25 September 2026, each with a launch snapshot of every input and a completion check of every result file; we will release the code and data after the anonymous review period.

E Admissions and controls

Table 5 gives the admissions of every judge; all eleven BiGGen judges are admitted throughout.

We compared the conditional correlation test with a marginal correlation test using the same Bonferroni correction (Table 6). We used two synthetic scores, $0.5X_{\mathrm{char}}/s_{\mathrm{tr}}(X_{\mathrm{char}}) + \xi$ and $X_{\mathrm{rel}}/s_{\mathrm{tr}}(X_{\mathrm{rel}}) + \zeta$, where X_{char} and X_{rel} measure character count and relative length, s_{tr} is the training standard deviation, and $\xi \sim N(0, 0.5^2)$ and $\zeta \sim N(0, 0.05^2)$ are independent noise. The marginal test admits these surface judges in as many as 20 of 20 partitions; the conditional test admits neither in any partition on any benchmark. The separate negative control, $0.5X_{\mathrm{char}}/s_{\mathrm{tr}}(X_{\mathrm{char}}) + \xi$, was screened with the panel and is also admitted in 0 of 20 partitions on every benchmark.

F The certificate is conservative

Over the partitions with a certified rule, the mean test error of ABC lies between 0.76 and 0.90 of ε in every cell of Table 3 with more than one such partition, and the mean test and certification coverage differ by at most 0.011. The margin comes from the certificate itself: to reject at level δ, the test needs the error on the certification data to fall below ε by a margin of order $\sqrt{\varepsilon(1-\varepsilon)/m}$, where m is the number of accepted certification grades, and with several responses per prompt the dependence within prompts adds to it. Certification and test coverage correlate at 0.983 on BiGGen, 0.996 on Chatbot Arena, and 0.998 on HelpSteer2 across the certifying partitions, with largest absolute gaps of 0.067, 0.030, and 0.041.

Table 6: Marginal and conditional correlation tests on the same 20 training partitions, each using 50% of the prompts. Cells give admissions and mean correlation. Each panel was screened together with both synthetic judges, using Bonferroni correction at family level 0.05. Admissions (adm.) are out of 20 partitions.

	Marginal		Conditional	
Judge	adm.	mean $\widehat{\rho}$	adm.	mean $\widehat{\rho}$
HelpSteer2				
GPT-4.1-nano	20	0.288	20	0.241
GPT-4o-mini	20	0.373	20	0.325
GPT-4.1	20	0.450	20	0.412
Llama-3.1-8B	20	0.341	20	0.283
Selene-1-Mini	20	0.401	20	0.361
SmolLM2-135M	19	0.106	0	0.059
SmolLM2-360M	0	0.021	0	0.012
Qwen2.5-0.5B	0	0.007	0	−0.010
Llama-3.2-1B	1	0.073	0	0.036
Synthetic verbosity	0	0.041	0	−0.001
Synthetic relative length	20	0.114	0	0.008
Chatbot Arena				
GPT-4-turbo	20	0.647	20	0.569
GPT-3.5-turbo	20	0.584	20	0.480
Mistral-7B	20	0.561	20	0.432
GPT-4.1-nano	20	0.507	20	0.402
GPT-4o-mini	20	0.625	20	0.508
Llama-3.1-8B	20	0.543	20	0.402
Selene-1-Mini	20	0.564	20	0.455
SmolLM2-135M	1	−0.095	0	0.007
SmolLM2-360M	0	−0.041	0	0.006
Qwen2.5-0.5B	3	0.091	15	0.116
Llama-3.2-1B	20	0.286	20	0.147
Synthetic verbosity	20	0.221	0	−0.001
Synthetic relative length	0	0.027	0	0.013
BiGGen				
Prometheus (ours)	20	0.407	20	0.358
Prometheus/BiGGen (ours)	20	0.511	20	0.459
GPT-4 (ours)	20	0.499	20	0.444
GPT-4-turbo (ours)	20	0.529	20	0.477
Claude (ours)	20	0.469	20	0.420
GPT-4-turbo (Bridge)	20	0.458	20	0.391
GPT-4.1-nano (Bridge)	20	0.414	20	0.324
GPT-4.1 (Bridge)	20	0.521	20	0.458
GPT-4o-mini (Bridge)	20	0.465	20	0.389
Llama-3.1-8B (Bridge)	20	0.385	20	0.309
Selene-1-Mini (Bridge)	20	0.468	20	0.394
Synthetic verbosity	13	0.137	0	0.002
Synthetic relative length	2	0.135	0	−0.025

G SENSITIVITY TO THE SPLIT

On HelpSteer2, we kept 20% of the prompts for testing and varied the training/certification/test allocation over 20/60/20, 30/50/20, 40/40/20, and 50/30/20. ABC has the highest certified coverage among the reported rules in every setting, and its pooled test error is below ε in every cell (Table 7). At $\varepsilon \leq 0.15$, coverage increases with the certification set; at $\varepsilon = 0.10$, it rises from 0.226 under 50/30/20 to 0.330 under 20/60/20. At $\varepsilon = 0.20$, the four certified coverages differ by less than 0.03.

Table 7: HelpSteer2 under four training/certification/test allocations, 20 partitions, $\delta = 0.05$; entries as in Table 1.

	$\varepsilon = 0.10$		$\varepsilon = 0.15$		$\varepsilon = 0.20$	
Rule	coverage	test error	coverage	test error	coverage	test error
training/certification/test = 20/60/20						
ABC panel	**0.330** $\pm$ 0.021	0.082 $\pm$ 0.003	**0.676** $\pm$ 0.013	0.135 $\pm$ 0.002	**0.882** $\pm$ 0.008	0.184 $\pm$ 0.002
Best single judge	–	–	0.237 $\pm$ 0.041	0.137 $\pm$ 0.004	0.796 $\pm$ 0.010	0.184 $\pm$ 0.003
Majority vote	–	–	0.183 $\pm$ 0.051	0.135 $\pm$ 0.004	0.744 $\pm$ 0.022	0.173 $\pm$ 0.003
Cascade	*0.010* $\pm$ *0.000*	*0.139* $\pm$ *0.000*	0.249 $\pm$ 0.032	0.131 $\pm$ 0.004	0.643 $\pm$ 0.011	0.180 $\pm$ 0.003
training/certification/test = 30/50/20						
ABC panel	**0.297** $\pm$ 0.030	0.083 $\pm$ 0.002	**0.653** $\pm$ 0.012	0.130 $\pm$ 0.002	**0.871** $\pm$ 0.006	0.181 $\pm$ 0.003
Best single judge	–	–	0.162 $\pm$ 0.041	0.136 $\pm$ 0.004	0.782 $\pm$ 0.010	0.181 $\pm$ 0.003
Majority vote	–	–	0.229 $\pm$ 0.052	0.133 $\pm$ 0.003	0.696 $\pm$ 0.032	0.167 $\pm$ 0.005
Cascade	–	–	0.200 $\pm$ 0.037	0.129 $\pm$ 0.005	0.619 $\pm$ 0.012	0.176 $\pm$ 0.003
training/certification/test = 40/40/20						
ABC panel	**0.267** $\pm$ 0.037	0.083 $\pm$ 0.003	**0.619** $\pm$ 0.035	0.130 $\pm$ 0.003	**0.865** $\pm$ 0.009	0.179 $\pm$ 0.003
Best single judge	–	–	0.174 $\pm$ 0.045	0.145 $\pm$ 0.003	0.765 $\pm$ 0.017	0.180 $\pm$ 0.004
Majority vote	–	–	0.184 $\pm$ 0.052	0.138 $\pm$ 0.002	0.680 $\pm$ 0.034	0.164 $\pm$ 0.006
Cascade	–	–	0.223 $\pm$ 0.031	0.131 $\pm$ 0.003	0.614 $\pm$ 0.015	0.175 $\pm$ 0.004
training/certification/test = 50/30/20						
ABC panel	**0.226** $\pm$ 0.035	0.080 $\pm$ 0.003	**0.585** $\pm$ 0.036	0.126 $\pm$ 0.004	**0.857** $\pm$ 0.011	0.176 $\pm$ 0.004
Best single judge	–	–	0.151 $\pm$ 0.044	0.137 $\pm$ 0.007	0.624 $\pm$ 0.074	0.184 $\pm$ 0.005
Majority vote	–	–	0.170 $\pm$ 0.062	0.150 $\pm$ 0.010	0.586 $\pm$ 0.036	0.152 $\pm$ 0.005
Cascade	–	–	0.126 $\pm$ 0.035	0.131 $\pm$ 0.007	0.597 $\pm$ 0.028	0.175 $\pm$ 0.006