CALIBRATE, DON'T CURATE: LABEL-EFFICIENT ESTIMATION FROM NOISY LLM JUDGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Multi-judge evaluation increasingly needs calibrated probabilities, not only verdicts. With a small labeled set, a natural practice is to *curate*: rank judges by individual accuracy and keep the best few. We argue that the labels are better spent to *calibrate* the panel: model the judges jointly and choose the panel by held-out proper-scoring risk, which may still select a subset, but never by marginal accuracy. We explain why curation appears to work. A conditionally independent aggregator double counts redundant judges, and in a two-block model we prove that this misspecification alone can make an accuracy-curated panel beat the full panel, while the correctly specified full panel beats both. On four labeled pairwise benchmarks with 38–144 judges, removing only the modeled dependence shifts the comparison against the full panel on every benchmark and split, and two pre-registered tests, a controlled simulation and three held-out panels, agree with this mechanism. Under conditionally independent aggregation the five most accurate judges beat the full panel by 0.09–0.13 NLL on both reward-model benchmarks; dependence-aware aggregation shrinks the gap to at most 0.015. Selecting panels by calibrated risk stays within 0.017 NLL of the best accuracy-ranked rule on every benchmark, without choosing the panel size by hand; the aggregator matters more than the panel rule. JudgeBench remains a counterexample: we recommend never pruning by accuracy alone, not keeping every judge.

1 INTRODUCTION

Panels of LLM judges and reward models are a standard instrument for evaluating language models (Zheng et al., 2023; Chiang et al., 2024; Lambert et al., 2025). Increasingly, the required output is not a verdict but a probability: how likely is it that response A is better than response B? Such probabilities drive selective evaluation and escalation to human raters (Jung et al., 2025), and they are judged by strictly proper scoring rules (Gneiting & Raftery, 2007).

When a small labeled set is available, a natural way to use it is to *curate*: estimate each judge's accuracy on the labels, keep the few best judges, and aggregate them. Curation is intuitive and cheap to deploy, and it is well motivated when judges err independently and point accuracy is the target. We argue that for probabilistic evaluation it answers the wrong question. Individual accuracy measures how good a judge is alone. Calibrated risk depends on what a judge adds *given the others* and on whether the aggregator can learn that contribution from the available labels. Two judges that share a backbone may each be accurate yet jointly redundant; a judge that is mostly wrong in a stable way, or right mainly where the others fail, can be worth more than its accuracy suggests. We therefore advocate spending the labels to *calibrate* the panel: fit an aggregator that models the judges jointly, and choose the panel by held-out proper-scoring risk. Throughout, *curate* means accuracy-ranked pruning and *calibrate* means fitting and selecting by calibrated probabilistic risk, which may still select a subset of judges.

Contributions.

1. **Why curation appears to work (Section 3).** Enlarging the observed panel cannot worsen proper-loss Bayes risk even for dependent judges, and the fitted risk difference splits exactly into an information gain and an excess-risk difference. In a two-block model we prove that a conditionally independent aggregator alone can make an accuracy-curated panel beat the full panel after any

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.

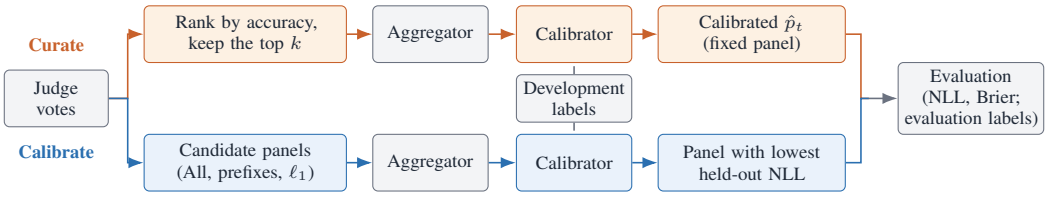

Figure 1: Curation versus calibration. Both arms use the same labels, aggregator and calibrator; they differ only in whether the panel is chosen by individual accuracy or by held-out calibrated risk.

calibration, while the correctly specified full panel beats both, and that accuracy ranking can select exactly the redundant judges. A finite-candidate guarantee licenses choosing panels by validation risk plus inference cost.

2. **The aggregator decides whether more judges help (Section 4.2).** On JudgeBench, RewardBench, RewardBench 2 and LLMBar (38–144 judges), with a leakage-free protocol, one-coin and full two-coin Dawid–Skene aggregators favor the five most accurate judges, by 0.09–0.13 NLL on the two reward-model benchmarks. Shrinkage LDA and ridge remove this advantage (to within 0.015 NLL) and favor the full panel on RewardBench and LLMBar. JudgeBench is a counterexample.
3. **Removing modeled dependence reproduces the reversal (Sections 4.3–4.4).** Removing only the off-diagonal covariance of a fixed discriminant model shifts the full-versus-top-five contrast against the full panel on all four benchmarks and on all 21 splits of each. Two pre-registered tests agree: a controlled simulation reproduces the reversal close to where the theory predicts it, and the intervention has the same sign on three held-out panels. Across seven real panels, the effect size tracks redundancy only loosely.
4. **Choosing panels by risk rather than accuracy (Section 4.5).** Across four benchmarks and two downstream model families, risk-based selection over a library containing the full panel and the accuracy-ranked panels stays within 0.017 NLL of the best accuracy-ranked rule and is best on LLMBar, without fixing the panel size in advance. Its gains are small: the aggregator matters more than the panel rule, and the value of risk-based selection is that it does not commit to the wrong objective.

Every estimator we use is standard, and we do not claim that the full panel is uniformly best. FCPS (Zhu et al., 2026) already selects panels and aggregator families by validation on a seven-judge pool; our contribution is an empirical diagnosis for large, dependent LLM-judge and reward-model panels: we identify, prove and test the mechanism that makes accuracy curation look right. In the sense of our title, a panel chosen by calibrated risk is *calibration*, not curation, even when it is a subset, and "label-efficient" means that one small labeled set is spent on fitting and selecting the panel's predictive model rather than on ranking judges; how many labels that needs depends on panel size and redundancy (Sections 4.4 and 4.6).

2 Setup and Estimators

Data and label boundary. Each item t has a binary target z_t (e.g., response A is better than response B). Judge $k \in [K]$ returns a vote $y_{tk} \in \{0, 1\}$ or a missing value. A labeled development set $\mathcal{D}$ is split into disjoint slices for fitting aggregators, fitting calibrators, and selecting panels; an evaluation set $\mathcal{E}$ is used only to compute metrics. The feature vector of every item is a function of its votes and missingness pattern alone, and evaluation labels enter only the metric computation; the code that builds votes from released judge outputs does not read evaluation labels. Appendix C.1 describes two correctness-encoding pitfalls that this boundary rules out and why they systematically distort panel comparisons.

Target and metrics. For a panel $S \subseteq [K]$ we estimate $\hat{p}_t = \hat{P}(z_t = 1 \mid \mathbf{y}_{t,S})$ and score it with negative log-likelihood (NLL) and Brier score, both strictly proper. We also report the binning-free cumulative calibration errors ECCE-MAD and ECCE-R of Arrieta-Ibarra et al. (2022), accuracy, and, for the optional set-valued output, conformal coverage and set size. Binned ECE is descriptive only.

Panel rules. Figure 1 contrasts the two ways of spending labels. *All* uses $S = [K]$. *Accuracy top-k* (curation) ranks judges by development accuracy and keeps the first k; our default is $k = 5$ ("Top-5"). The *accuracy path* ranks by accuracy but chooses $k \in \{3, 5, 10, 20, 40, 80\}$ by development calibrated NLL, so it is the strongest accuracy-ranked curator we consider. *Risk-based selection* evaluates a library $\mathcal{S}$ of candidate panels (All, the accuracy prefixes, and the supports of an ℓ_1-logistic path over its regularization grid), refits each with the same aggregator and calibrator, and returns

$$\hat{S} \in \arg\min_{S \in \mathcal{S}} \ \widehat{R}_{\mathrm{NLL}}(S) + \lambda\, |S|/K, \tag{1}$$

where $\widehat{R}_{\mathrm{NLL}}$ is calibrated NLL on a development selection slice disjoint from the fitting and calibration slices and $\lambda \geq 0$ prices judge calls. Because the library contains the accuracy prefixes, risk-based selection can reproduce curation whenever curation has lower estimated risk; it differs only in the objective.

Aggregators. Each aggregator maps the selected votes to a raw probability $A_S(\mathbf{y}_t)$. The two *conditionally independent* (CI) aggregators assume that votes are independent given z_t. The one-coin model gives each judge a symmetric reliability a_k with posterior $\mathrm{Beta}(\alpha_k, \beta_k)$ from development counts. Over the set J_t of judges with parseable votes, its posterior-predictive log odds is a sum of per-judge log-likelihood ratios,

$$\operatorname{logit} A_S(\mathbf{y}_t) = \operatorname{logit} \pi + \sum_{k \in S \cap J_t} (2y_{tk} - 1) \log \frac{\alpha_k}{\beta_k}, \tag{2}$$

because $P(y_{tk} = 1 \mid z_t = 1, \mathcal{D})/P(y_{tk} = 1 \mid z_t = 0, \mathcal{D}) = \mathbb{E}[a_k \mid \mathcal{D}]/\mathbb{E}[1 - a_k \mid \mathcal{D}] = \alpha_k/\beta_k$ and a vote of 0 contributes the reciprocal (Appendix A). Equation (2) assumes that availability J_t is uninformative about z_t. The full two-coin Dawid–Skene model (Dawid & Skene, 1979) replaces a_k by class-conditional rates $s_k = P(y_{tk} = 1 \mid z_t = 1)$ and $r_k = P(y_{tk} = 0 \mid z_t = 0)$, estimated from the labeled fitting slice with add-one smoothing. The two *dependence-aware* aggregators model the votes jointly. Shrinkage linear discriminant analysis (LDA) fits class means μ_0, μ_1 and a shared Ledoit–Wolf covariance Σ (Ledoit & Wolf, 2004) to the mean-imputed vote vector, so its log odds is linear in the votes with weights $\Sigma^{-1}(\mu_1 - \mu_0)$: correlated judges share rather than add their evidence. Ridge logistic regression on the same mean-imputed votes is the discriminative counterpart; its penalty is chosen on development data (fixed at $C = 1$ in the label-budget sweep). Setting the off-diagonal entries of the fitted LDA covariance to zero yields a CI Gaussian model with identical means, priors, imputation and shrinkage; we use this as a matched intervention on dependence (Section 4.3).

Calibration and the composed predictor. Raw scores are mapped by beta calibration (Kull et al., 2017), $g(p) = \sigma(a \log p + b \log(1 - p) + c)$, fitted on the calibration slice; Platt scaling (Platt, 2000) is the tied subfamily $b = -a$. In the repeated-split experiments (a, b, c) is shrunk toward the identity with an elastic-net penalty ($\lambda = 0.01$), and temperature scaling replaces it when fewer than 30 calibration labels are available, when the fit fails, or when a fitted coefficient exceeds 20 in magnitude. On the frozen split, in the covariance intervention and in the stage ablation we constrain $a \geq 0$, $b \leq 0$ so that g is monotone, with an ℓ_2 penalty (0.01) on (a, b) toward zero, and fit it on 5-fold out-of-fold aggregator scores of the development set. Aggregator outputs are clipped to $[10^{-3}, 1 - 10^{-3}]$ (which bounds a raw NLL by 6.9). The deployed predictor is the single composed map $F_S(\mathbf{y}_t) = g\{A_S(\mathbf{y}_t)\}$, with stages fitted sequentially on disjoint slices or by cross-fitting as just described; we make no claim about joint convergence. Split conformal prediction (Angelopoulos & Bates, 2023) optionally turns F_S into sets with marginal coverage $1 - \alpha$.

3 Oracle Information, Fitted Risk, and Selection

We separate three objects that the curation debate tends to conflate: the information in a panel, the risk of a fitted predictor built from it, and the rule used to choose among fitted predictors. Proofs are in Appendix A.

Proposition 1 (Oracle monotonicity for dependent judges)**.** *Let $Y \in \{0, 1\}$ and let $X_1, \ldots, X_K$ be arbitrarily dependent judge signals. For $S \subseteq T \subseteq [K]$ let $\eta_S = P(Y = 1 \mid \mathbf{X}_S)$. Then $R_{\mathrm{Brier}}(T) = \mathbb{E}[\mathrm{Var}(Y \mid \mathbf{X}_T)] \leq R_{\mathrm{Brier}}(S)$ and $R_{\mathrm{NLL}}(T) = H(Y \mid \mathbf{X}_T) \leq R_{\mathrm{NLL}}(S)$, with either*

inequality strict if and only if $I(Y; \mathbf{X}_{T\setminus S} \mid \mathbf{X}_S) > 0$. The same monotonicity holds for every strictly proper binary loss with finite Bayes risk.

Proposition 1 needs no conditional independence: an oracle never loses by seeing more judges. It says nothing about a fitted predictor, and the gap is exact.

Proposition 2 (Information gain versus excess risk)**.** *Let $\hat{\eta}_U$ be any predictor for panel U fitted on development data D independent of a fresh item, and let $e_U^L(D) = R_L(\hat{\eta}_U \mid D) - R_L(\eta_U) \geq 0$ be its excess risk. If the NLL risks are finite (clipping predictions to $[\epsilon, 1-\epsilon]$ suffices), then for $S \subseteq T$,*

$$R_{\mathrm{NLL}}(\hat{\eta}_T \mid D) - R_{\mathrm{NLL}}(\hat{\eta}_S \mid D) = -I(Y; \mathbf{X}_{T\setminus S} \mid \mathbf{X}_S) + e_T^{\mathrm{NLL}}(D) - e_S^{\mathrm{NLL}}(D),$$
$$R_{\mathrm{Brier}}(\hat{\eta}_T \mid D) - R_{\mathrm{Brier}}(\hat{\eta}_S \mid D) = -\mathbb{E}[(\eta_T - \eta_S)^2] + e_T^{\mathrm{Brier}}(D) - e_S^{\mathrm{Brier}}(D).$$

Adding judges helps a fitted predictor exactly when their conditional information exceeds the extra excess risk they induce. Accuracy ranking measures neither term: a highly accurate judge may be nearly redundant given the others (small information), and a panel fitted with a misspecified likelihood may pay a large excess risk. The decomposition is an identity, not a bound; nothing in our setting controls $e_T - e_S$ without assuming correct specification, which shared-backbone panels violate.

When curation looks right, and why. Conditionally independent aggregators such as Eq. (2) add the log-likelihood ratios of correlated judges as if they were independent evidence. A monotone post-hoc map can rescale a score but cannot reorder items, and the best risk that any monotone map achieves depends on the score only through the ordering it induces (Lemma 1). The following model makes the consequence exact. A *redundant* block of m judges shares an item-level error factor; a *diverse* block of r judges errs independently. With $s = 2Y - 1$ and U, ξ_k, ξ'_j i.i.d. $N(0,1)$,

$$X_k = \tfrac{\Delta_R}{2} s + \sqrt{\rho}\, U + \sqrt{1-\rho}\, \xi_k \ (k \leq m), \qquad X'_j = \tfrac{\Delta_D}{2} s + \xi'_j \ (j \leq r), \tag{3}$$

so every judge has unit noise variance and individual accuracy $\Phi(\Delta/2)$. A linear score $T = \sum_k w_k X_k + \sum_j w'_j X'_j$ is Gaussian given Y, so its *separation* $d = \mathbb{E}[T \mid s = 1] - \mathbb{E}[T \mid s = -1]$ divided by $\mathrm{sd}(T \mid Y)$ determines it.

Proposition 3 (Redundancy makes curation look right)**.** *In model (3), (i) the smallest NLL achievable by any calibration map of a linear score is a strictly decreasing function of its separation d. (ii) The full-covariance (Bayes, population LDA) score on all judges has $d_\star^2 = \Delta_R^2/(\rho + (1-\rho)/m) + r\Delta_D^2$, which is nondecreasing in m. (iii) The conditionally independent score on all judges has*

$$d_{\mathrm{CI}}^2 = \frac{(m\Delta_R^2 + r\Delta_D^2)^2}{m^2\rho\Delta_R^2 + m(1-\rho)\Delta_R^2 + r\Delta_D^2} \xrightarrow[m\to\infty]{} \frac{\Delta_R^2}{\rho}.$$

Hence if $\Delta_D > \Delta_R$, accuracy ranking selects the diverse block (for which the CI score is Bayes, with $d_D^2 = r\Delta_D^2$), and for every $\rho > \Delta_R^2/(r\Delta_D^2)$ and all large m the curated panel beats the CI score on all judges after any calibration, whereas the full-covariance score on all judges beats the curated panel by $d_\star^2 - d_D^2 = \Delta_R^2/(\rho + (1-\rho)/m) > 0$ for every m. (iv) If instead $\Delta_R > \Delta_D$ and $\rho > 0$, accuracy ranking selects $k \leq m$ redundant judges, whose Bayes score has $d^2 = k\Delta_R^2/(1 + (k-1)\rho) < \Delta_R^2/\rho$, while the Bayes score of one redundant and $k-1$ diverse judges has $d^2 = \Delta_R^2 + (k-1)\Delta_D^2$, which is larger whenever $(k-1)\Delta_D^2 \geq \Delta_R^2(1-\rho)/\rho$.

Under a conditionally independent aggregator, pruning to accurate judges looks right because of misspecification, not weak judges (ii–iii), and accuracy ranking can pick exactly the redundant judges (iv). Section 4.4 tests these predictions with binary votes.

A guarantee for risk-based selection. Since fitted monotonicity fails, the panel must be chosen empirically. For a library of M candidate panels fitted and calibrated independently of an i.i.d. selection sample of size n, choosing the candidate with the smallest clipped validation NLL plus call cost (Eq. (1)) is within $|\log\epsilon|\sqrt{2\log(2M/\delta)/n}$ of the best candidate with probability $1-\delta$ (Theorem 1, Appendix A). Unlike the jury theorems behind curation, this permits arbitrary dependence among judges within an item; it certifies the procedure, not individual numbers. At our sizes it is loose: with $\epsilon = 10^{-3}$, $\delta = 0.05$, at most $M = 70$ candidate fits and selection slices of 35–299 items, the bound is 1.6–4.7 nats, far larger than any difference we report. The classical motivation for

Table 1: $\Delta_{\mathrm{A-T}}$ after beta calibration at the largest label budget (50 repeated splits; mean and 2.5/97.5 split percentiles). Negative favors the full panel. Bold: range excludes zero.

	Conditionally independent		Dependence-aware	
Benchmark (K)	One-coin	Two-coin (DS)	Shrinkage LDA	Ridge
JudgeBench (38)	$+0.105$ $[-0.004, +0.358]$	$+0.104$ $[-0.000, +0.275]$	$+0.037$ $[-0.016, +0.144]$	$+0.064$ $[-0.002, +0.139]$
RewardBench (104)	$\mathbf{+0.089}$ $[+0.060, +0.121]$	$\mathbf{+0.089}$ $[+0.060, +0.118]$	-0.005 $[-0.013, +0.004]$	$\mathbf{-0.016}$ $[-0.031, -0.001]$
RewardBench 2 (144)	$\mathbf{+0.133}$ $[+0.104, +0.163]$	$\mathbf{+0.134}$ $[+0.103, +0.164]$	$+0.013$ $[-0.006, +0.048]$	$+0.015$ $[-0.010, +0.048]$
LLMBar (53)	$+0.035$ $[-0.063, +0.239]$	$+0.031$ $[-0.087, +0.179]$	-0.034 $[-0.127, +0.059]$	-0.056 $[-0.111, +0.005]$

keeping weak judges, a weighted Condorcet bound under which even below-chance judges help after sign reversal (Nitzan & Paroush, 1982; Berend & Paroush, 1998), requires known reliabilities and conditional independence. We state it, together with a bound showing that near-chance reliability signs cannot be estimated reliably, in Appendix A.1; neither result carries our empirical claims.

4 Experiments

4.1 Benchmarks, panels, and protocol

Benchmarks and panels. We use four labeled pairwise benchmarks. JudgeBench (Tan et al., 2025) has 350 pairs with objectively correct answers; its panel is 32 released judges plus 6 API judges ($K{=}38$). RewardBench (Lambert et al., 2025) has 2,985 chosen/rejected pairs scored by 100 released reward models plus 4 open-weight LLM judges ($K{=}104$). RewardBench 2 (Malik et al., 2026) is best-of-4: each prompt has one correct and three incorrect responses. From the released per-response scores of 174 reward models we form one correct-versus-incorrect pair per prompt, drawn once for all judges, so that chance accuracy is $1/2$ (1,845 pairs). We exclude 30 checkpoints of a label-flipping ablation, trained on deliberately corrupted preferences, which leaves $K{=}144$; keeping them changes no conclusion (Appendix C). LLMBar (Zeng et al., 2024) has 419 instruction-following pairs, including adversarial subsets; its panel is 44 released evaluator configurations plus 5 API and 4 open-weight judges ($K{=}53$). Reward models score each response separately, so we form their A/B votes under an exogenous, balanced random presentation order; the vote is then exactly what the model outputs for that order, and the label records which position holds the preferred response. Added LLM judges were queried once per item with a randomized order and a fixed pairwise prompt (Zheng et al., 2023); unparseable outputs are missing. The four LLM judges added to RewardBench were stored as agreement with the preferred response, so they enter the same random order as the reward models, without their position bias. Appendix C lists judges, prompts, parsing and coverage. We also use an external panel, the objective-correctness subset of PPE (Frick et al., 2025) (12,775 pairs in 2,555 question groups, 19 released reward models and LLM-judge configurations).

Protocol. All fitting, tuning, calibration and selection use development data only. We use two resampling schemes and never mix them in one number. *Repeated splits*: 50 random 50/50 development/evaluation splits; the development half is divided into fitting and calibration slices (and, for selection, a third selection slice). Ranges in brackets after table entries are 2.5/97.5 percentiles of the paired contrast across splits; they describe split-to-split variability and we read them as direction only. *Frozen split*: one stratified 60/40 split, with 95% paired item-bootstrap intervals conditional on that split. We write $\Delta_{\mathrm{A-T}} = \mathrm{NLL}(\mathrm{All}) - \mathrm{NLL}(\text{Top-5})$, so negative values favor the full panel. The simulation and held-out panels of Section 4.4 were specified before they were run. The other analyses were designed after inspecting earlier results on the same benchmarks; they are exploratory, and no interval adjusts for the number of analyses.

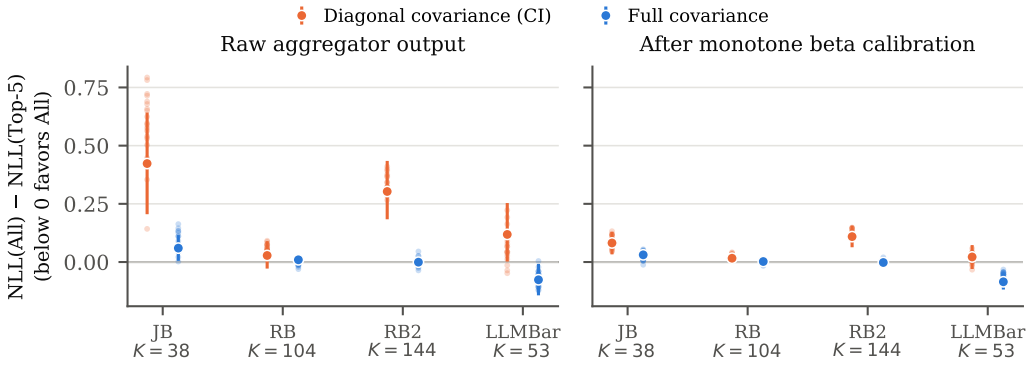

Figure 2: Matched covariance intervention. Each point pair is one fitted LDA per panel, scored with full versus zeroed off-diagonal covariance (large markers: frozen split with 95% item-bootstrap interval; small markers: 20 further splits). Removing dependence moves $\Delta_{\mathrm{A-T}}$ against the full panel on every benchmark and split.

4.2 The aggregator decides whether more judges help

Table 1 holds the panel, items, calibrator and label budget fixed and changes only the aggregator. Under both conditionally independent models, the five most accurate judges beat the full panel on RewardBench and RewardBench 2, with ranges excluding zero, and point estimates favor them on JudgeBench and LLMBar. Under both dependence-aware models the full panel's deficit disappears: on RewardBench 2 it shrinks from 0.13 to 0.013–0.015 (ranges include zero), and on RewardBench and LLMBar the sign reverses; JudgeBench keeps the sign favoring Top-5. The pattern is present before calibration (raw $\Delta_{\mathrm{A-T}}$ on RewardBench 2: $+0.38$ one-coin, $+0.02$ ridge; Appendix D), so it is not produced by the calibration map. The full two-coin Dawid–Skene model tracks one-coin to within 0.005 everywhere: class-conditional asymmetry is not what separates the families. The frozen split agrees. On LLMBar, ridge gives $\Delta_{\mathrm{A-T}} = -0.086\,[-0.137, -0.035]$ and LDA $-0.085\,[-0.117, -0.053]$, whereas one-coin gives $+0.028\,[-0.019, +0.074]$; on JudgeBench, ridge gives $+0.026\,[-0.001, +0.053]$ and one-coin $+0.078\,[+0.029, +0.127]$. On the external PPE panel, ridge favors the full panel in raw NLL, $\Delta_{\mathrm{A-T}} = -0.009\,[-0.015, -0.003]$ (frozen split, question-group bootstrap), but the effect shrinks to $-0.003\,[-0.005, +0.000]$ when each model family keeps one representative, so we treat it as weak support (per-stratum LLMBar results: Appendix D). Accuracy curation thus looks right under the CI aggregators whose assumptions these panels violate. Under models that account for dependence the comparison is mixed: neutral or favoring the full panel on RewardBench and LLMBar, while Top-5 stays ahead on JudgeBench and, narrowly, on RewardBench 2. The RewardBench 2 reward models are highly redundant (effective panel size $K_{\mathrm{eff}} \approx 6$ of 144; Section 4.4), so five of them already carry most of the panel's usable information. LLMBar differs in composition: 25 of its 53 judges are below chance, and the dependence-aware aggregators give many of them negative weight. Removing judges below $1/2$ on the fitting slice within each split (29 of 53 remain on average) shrinks ridge $\Delta_{\mathrm{A-T}}$ from -0.056 to $-0.022\,[-0.059, +0.027]$ and leaves the other benchmarks essentially unchanged (Appendix D): about half of LLMBar's full-panel gain comes from exploiting judges that are wrong in a stable way. The dependence effect itself survives the removal: the calibrated covariance interaction of Section 4.3 is $-0.102\,[-0.153, -0.052]$ and negative on all 21 splits.

4.3 Removing modeled dependence reproduces the reversal

The four aggregators differ in more than dependence: generative versus discriminative fitting, functional form and regularization all change. To isolate dependence we fit one shrinkage LDA per panel and score it twice, once with its full covariance and once with the off-diagonal entries set to zero, keeping means, priors, imputation and shrinkage fixed. The diagonal variant is a conditionally independent Gaussian model. We report the *interaction* $\Delta^{\mathrm{full}}_{\mathrm{A-T}} - \Delta^{\mathrm{diag}}_{\mathrm{A-T}}$ on the frozen split and on 20 additional splits (Figure 2).

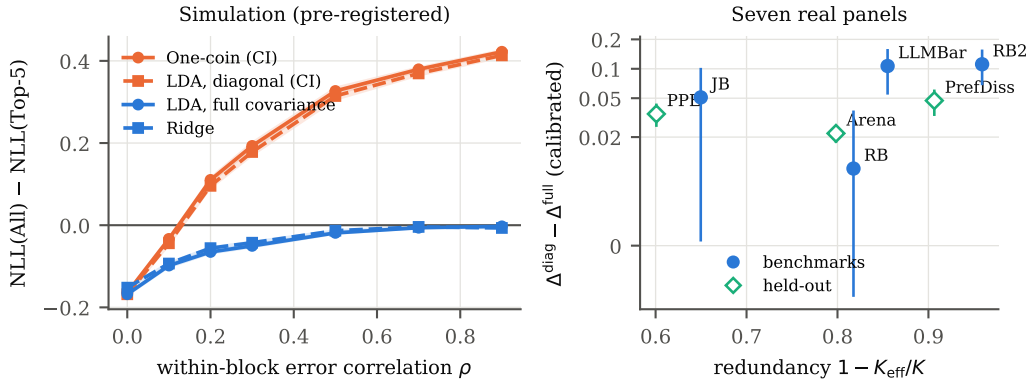

Figure 3: Left: pre-registered simulation with 50 redundant judges (accuracy 0.70, shared error factor ρ) and 5 diverse judges (accuracy 0.80); means and 95% intervals over 30 replicates. Right: size of the calibrated covariance effect $\Delta^{\text{diag}}_{\text{A}-\text{T}} - \Delta^{\text{full}}_{\text{A}-\text{T}}$ against panel redundancy $1 - K_{\text{eff}}/K$ for the four benchmarks (frozen split) and three held-out panels (group-bootstrap intervals).

After calibration the interaction is negative on all four benchmarks and on all 21 splits of each: $-0.051\,[-0.102, -0.001]$ on JudgeBench, $-0.014\,[-0.037, +0.009]$ on RewardBench (split mean -0.028, range $[-0.042, -0.013]$), $-0.111\,[-0.157, -0.068]$ on RewardBench 2, and $-0.107\,[-0.159, -0.054]$ on LLMBar. On RewardBench 2 the same fitted model ties Top-5 with its covariance (-0.002) and loses to it by 0.109 NLL without; raw interactions are larger (up to -0.36), consistent with calibration partially masking but not undoing the double counting. On JudgeBench, modeling dependence shrinks the full panel's deficit from $+0.082$ to $+0.031$ but does not remove it, consistent with the excess-risk term of Proposition 2 dominating for 38 judges fitted on 210 items.

4.4 A PRE-REGISTERED TEST: CONTROLLED SIMULATION AND HELD-OUT PANELS

Because the analyses above were designed on benchmarks we had already examined, we tested the mechanism with two further analyses whose hypotheses and decision rules were committed to our repository before any of them was run (Appendix B).

Simulation. Binary votes come from a redundant block of m judges (accuracy 0.70; errors share an item factor with weight ρ) and a diverse block of five independent judges (accuracy 0.80), evaluated with the Table 1 protocol. All four hypotheses held (Figure 3, left). At $\rho = 0$ the full panel beats Top-5 under every aggregator (one-coin $\Delta_{\text{A}-\text{T}} = -0.164\,[-0.170, -0.158]$). Under the one-coin model the sign flips between $\rho = 0.1$ and 0.2, close to the crossing at $\rho \approx 0.10$ that Proposition 3 predicts for $m = 50$, and reaches $+0.42$ at $\rho = 0.9$. Across the ρ sweep, dependence-aware aggregators never let the full panel fall behind ($\Delta_{\text{A}-\text{T}}$ between -0.17 and -0.003), the matched interaction grows monotonically with ρ, and the one-coin reversal grows with the size m of the redundant block. The limits show too: with 200 redundant judges and 700 fitting items the dependence-aware full panel loses to Top-5 by 0.05 (the estimation term of Proposition 2), and at $\rho = 0.5$ its advantage is resolved only from 1,000–2,000 items (Appendix B).

Held-out panels. We ran the covariance intervention on three panels on which it had never been run, using a group partition frozen two months earlier: PPE, Preference-Dissection (32 LLM judges of human preferences) and Chatbot-Arena (17 position-averaged LLM judges). The panels are held out for this analysis only: two months earlier we had compared ridge on All and Top-5 on all three, and All had the lower NLL each time. The calibrated interaction is negative on all three: $-0.035\,[-0.044, -0.025]$, $-0.047\,[-0.061, -0.033]$ and $-0.022\,[-0.025, -0.019]$. Each repeats the benchmark pattern. With zeroed off-diagonal covariance the full panel loses to Top-5 ($+0.025$, $+0.042$, $+0.021$); with the full covariance it does not (-0.010, -0.005, -0.001). Across all seven real panels the size of the dependence effect tracks redundancy only loosely (Figure 3, right; redundancy is $1 - K_{\text{eff}}/K$ with effective panel size $K_{\text{eff}} = K^2 / \sum_{j,k} R_{jk}^2$, where R is the pooled

Table 2: Panel selection under a fixed downstream family (aggregator + beta calibration; 50 repeated splits). Cells give evaluation NLL with the mean number of judges called in parentheses. The last column is the paired difference, risk-based minus accuracy path (negative favors risk-based selection), with its 2.5/97.5 split range. Bold: lowest NLL in the row.

Family	Benchmark	All	Top-5	Accuracy path	Risk-based	Risk − acc. path
Ridge	JudgeBench	0.514 (38)	**0.476**	0.477 (10.0)	0.492 (13.9)	$+0.015$ $[-0.015, +0.158]$
	RewardBench	**0.113** (104)	0.134	0.116 (65.7)	0.116 (58.6)	-0.000 $[-0.012, +0.014]$
	RewardBench 2	0.169 (144)	**0.160**	**0.160** (18.2)	0.162 (25.9)	$+0.002$ $[-0.015, +0.035]$
	LLMBar	0.298 (53)	0.347	0.318 (31.3)	**0.295** (18.0)	-0.022 $[-0.071, +0.010]$
LDA	JudgeBench	0.537 (38)	**0.506**	**0.506** (9.0)	0.509 (13.2)	$+0.003$ $[-0.034, +0.052]$
	RewardBench	0.149 (104)	0.152	**0.148** (38.2)	**0.148** (33.9)	-0.001 $[-0.010, +0.005]$
	RewardBench 2	0.186 (144)	0.175	0.175 (26.7)	**0.174** (33.4)	-0.001 $[-0.013, +0.013]$
	LLMBar	0.331 (53)	0.360	0.334 (27.5)	**0.315** (16.3)	-0.019 $[-0.106, +0.059]$

within-class correlation matrix of the development votes): the pre-registered criterion, a positive Spearman correlation, is met, but at 0.5 with seven panels (exact permutation $p = 0.27$; the same value as in the pre-registered evaluation, which used an earlier, flawed RewardBench 2 encoding; Appendix C.1). RewardBench 2, the most redundant panel ($K_{\text{eff}} = 5.9$ of 144 judges), has the largest effect, but LLMBar's is nearly as large at lower redundancy. We read the real panels as consistent with the mechanism, not as a test of its dose–response.

4.5 Choosing panels by calibrated risk

Table 2 (and Figure 5 in Appendix D) compares four panel rules under two downstream families. The accuracy path is the strongest accuracy-ranked curator: it searches the panel size by development risk but ranks judges by accuracy. Risk-based selection (Eq. (1) with $\lambda = 0$) searches a library that contains the full panel, those same prefixes and the ℓ_1-path supports, and it scores them by calibrated risk rather than accuracy. Across the eight cells it stays within 0.017 NLL of the better accuracy-ranked rule. It has the lowest NLL on LLMBar under both families (0.02 below the accuracy path and 0.05 below Top-5, with ranges crossing zero; 0.005 below the accuracy path once below-chance judges are removed) and ties on RewardBench and RewardBench 2. It trails on JudgeBench, where Top-5 is best and the 35-item selection slice makes risk estimates noisy (ridge: $+0.017$ against Top-5). A sign-aware curator that ranks judges by $|\text{accuracy} - 1/2|$ picks the same five judges as Top-5 on all 50 splits of every benchmark. The gains from risk-based selection are therefore small. Its value is that it does not commit to the wrong objective: the best rule differs across benchmarks (All on RewardBench, Top-5 on JudgeBench, about a third of the judges on LLMBar), and risk-based selection comes within 0.017 NLL of the best one on each without a hand-set panel size (call pricing and ℓ_1 sparsification: Appendix D). In our data the choice of aggregator matters more than the choice of panel rule.

4.6 What calibration does and does not do

"Calibrate" in our title refers to the objective, calibrated risk, not to a particular post-hoc map. In a leakage-free stage ablation (Appendix D), a Platt map repairs the overconfident one-coin full panel (-0.48 NLL on JudgeBench, -0.36 on LLMBar) but is nearly neutral after ridge (-0.029, $+0.011$). After beta calibration every paired ECCE-MAD range between All and Top-5 includes zero, for every benchmark and aggregator, so the NLL differences reflect ranking and sharpness rather than calibration error, as the ordering argument of Section 3 predicts. With a fixed evaluation set and nested budgets, the CI aggregators favor Top-5 on the two reward-model benchmarks from 100–250 labels, whereas the dependence-aware comparisons are unresolved at every budget except ridge on RewardBench, which favors the full panel only with all 1,492 development labels. Appendix D reports position effects and conformal sets.

5 Related Work

Aggregating annotators, weak sources and LLM judges. Dawid–Skene (Dawid & Skene, 1979) and its descendants estimate annotator reliability from disagreement (Raykar et al., 2009), including for pairwise comparisons (Chen et al., 2013), mostly under conditional independence. Weak supervision models dependencies among labeling functions because correlated sources double count evidence (Ratner et al., 2017; Bach et al., 2017). LLM errors are substantially correlated across models, more so for larger and more accurate ones (Kim et al., 2025; Goel et al., 2025), and panels of diverse LLM judges can beat a single large judge (Verga et al., 2024); our intervention measures what this correlation costs when it is left unmodeled and shows that it decides whether accuracy curation appears to help. For LLM judges, CARE (Zhao et al., 2026) models correlated errors through shared latent confounders, judge-aware Bradley–Terry variants add per-judge discrimination or rater quality (Xu et al., 2026; Aczel et al., 2025; Qian et al., 2026), and Wang (2025) decomposes evaluation noise; these target rankings or latent labels without supervision, whereas we ask how a small labeled set should be spent. Stacking and ensemble selection (Wolpert, 1992; Caruana et al., 2004) choose members by validation performance of the combined predictor, as does FCPS (Zhu et al., 2026) for judge panels.

Calibrating judges and label-efficient evaluation. LLM judges are often overconfident (Tian et al., 2025). Single-judge calibration uses hidden-state probes (Radharapu et al., 2025), repeated samples (Dadkhahi et al., 2025), regressors on a judge's rationale and score (Sahoo et al., 2025), or subgroup-aware recalibration (Manggala et al., 2025); none observes a heterogeneous black-box vote panel. Our maps are standard: Platt scaling (Platt, 2000), beta calibration (Kull et al., 2017), proper scoring rules (Gneiting & Raftery, 2007), binning-free calibration metrics (Arrieta-Ibarra et al., 2022) and conformal sets (Angelopoulos & Bates, 2023). Prediction-powered ranking (Chatzi et al., 2024), bias-corrected estimation from noisy judges (Chen et al., 2026), limits on replacing human labels with judges (Dorner et al., 2025) and selective escalation with guarantees (Jung et al., 2025) use small labeled sets to correct or certify judge-based estimates; they take the judge output as given, whereas we ask which judges should produce it.

6 Discussion and Limitations

For probabilistic evaluation with a small labeled set, marginal judge accuracy is the wrong criterion for building a panel. An oracle never loses by seeing more judges, but a fitted panel predictor can lose when its aggregator mis-specifies how the judges depend on one another. On our benchmarks, in a controlled simulation and on held-out panels, that misspecification, rather than weak judges, is what made accuracy curation look right. Once dependence is modeled, the choice between a large and a small panel is a trade-off between estimation risk and inference cost, to be settled by held-out proper-scoring risk. On our benchmarks the best panel is the full one on RewardBench and about a third of the judges on LLMBar, while on JudgeBench and the highly redundant RewardBench 2 no rule beats the five most accurate judges by more than 0.005 NLL. The practical gains of selecting by risk over the best accuracy-ranked rule are small; the gain is in not relying on an objective that is right only when the aggregator is wrong.

Limitations. Our targets are binary. Reward-model votes use a random presentation order and so carry no position effects, and RewardBench 2 enters as one sampled correct-versus-incorrect pair per best-of-4 prompt. Outside the two pre-registered analyses, our analyses reuse benchmarks we had examined before, so their intervals are descriptive. Proposition 3 is proved for Gaussian scores; the simulation checks it with binary votes but only for one two-block structure, and across real panels redundancy explains the size of the dependence effect only partly (rank correlation 0.5, $p = 0.27$); on LLMBar about half of the full panel's gain comes from exploiting below-chance judges. Finally, label efficiency is benchmark-specific: conditionally independent aggregators favor Top-5 from 100–250 labels on the reward-model benchmarks, but no budget resolves the dependence-aware comparison except on RewardBench with every development label.

AI USE STATEMENT

LLMs are the object of study in this paper: LLM judges and reward models produce the votes we aggregate. We did not use generative AI to generate synthetic datasets, for translation, or to produce any label, and no benchmark label was created or modified with AI assistance. Among the tasks with recommended disclosure, we used generative AI to draft and edit parts of the paper, to search and summarize literature, to verify and format references against Crossref, arXiv and Semantic Scholar records, and to create figures.

ETHICS STATEMENT

This work uses only publicly released benchmarks, judge verdicts and reward-model outputs; no new human-subject data were collected and no model was trained. Panels of automated judges can inherit shared biases; our results show that such dependence can make aggregated confidence misleading, which is a reason to validate judge panels against human labels rather than to replace them. We declare no conflicts of interest.

REPRODUCIBILITY STATEMENT

Section 2 and Appendix C define every estimator, panel rule, split and metric; Appendix A contains complete proofs with all assumptions stated. The benchmarks and released judge outputs are public, and Appendix C lists the judges, prompts, parser and coverage. The anonymized supplementary material contains the analysis code; the vote matrix of every panel, including the votes of the LLM judges we added, with the frozen-split and group-partition indices; the result files behind every table and figure, with their SHA-256 hashes; and the scripts that regenerate the tables and figures from those files. The pre-registration protocol of Appendix B, committed before the corresponding analyses were run, is included with its hash, which every simulation and held-out result file records. The supplement also contains the script that rebuilds RewardBench 2 pairs from the public per-response scores. Appendix C.1 documents two encoding pitfalls that are the most consequential failure modes for this kind of study.

REFERENCES

Till Aczel, Lucas Theis, and Roger Wattenhofer. Efficient Bayesian inference from noisy pairwise comparisons. *arXiv preprint arXiv:2510.09333*, 2025. doi: 10.48550/arXiv.2510.09333.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi: 10.1561/2200000101.

Anthropic. The Claude 3 model family: Opus, Sonnet, Haiku. Model card, 2024. URL `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`.

Imanol Arrieta-Ibarra, Paman Gujral, Jonathan Tannen, Mark Tygert, and Cherie Xu. Metrics of calibration for probabilistic predictions. *Journal of Machine Learning Research*, 23(351):1–54, 2022. doi: 10.48550/arXiv.2205.09680.

Stephen H. Bach, Bryan He, Alexander Ratner, and Christopher Ré. Learning the structure of generative models without labeled data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 273–282, 2017. doi: 10.48550/arXiv.1703.00854. arXiv:1703.00854.

Daniel Berend and Jacob Paroush. When is Condorcet's jury theorem valid? *Social Choice and Welfare*, 15(4):481–488, 1998. doi: 10.1007/s003550050118.

Niko Brummer and Johan du Preez. The PAV algorithm optimizes binary proper scoring rules. *arXiv preprint arXiv:1304.2331*, 2013. doi: 10.48550/arXiv.1304.2331.

Rich Caruana, Alexandru Niculescu-Mizil, Geoff Crew, and Alex Ksikes. Ensemble selection from libraries of models. In *Proceedings of the Twenty-First International Conference on Machine Learning*, pp. 18, 2004. doi: 10.1145/1015330.1015432.

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Prediction-powered ranking of large language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 113096–113133, 2024. doi: 10.52202/079017-3594.

Xi Chen, Paul N. Bennett, Kevyn Collins-Thompson, and Eric Horvitz. Pairwise ranking aggregation in a crowdsourced setting. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pp. 193–202, 2013. doi: 10.1145/2433396.2433420.

Yiqun T Chen, Sizhu Lu, Sijia Li, Moran Guo, and Shengyi Li. Efficient inference for noisy LLM-as-a-Judge evaluation. *arXiv preprint arXiv:2601.05420*, 2026. doi: 10.48550/arXiv.2601.05420.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024. doi: 10.48550/arXiv.2403.04132. arXiv:2403.04132.

Hamid Dadkhahi, Firas Trabelsi, Parker Riley, Juraj Juraska, and Mehdi Mirzazadeh. Distribution-calibrated inference time compute for thinking LLM-as-a-Judge. *arXiv preprint arXiv:2512.03019*, 2025. doi: 10.48550/arXiv.2512.03019.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806.

Florian E. Dorner, Vivian Y. Nastl, and Moritz Hardt. Limits to scalable evaluation at the frontier: LLM as judge won't beat twice the data. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.13341. arXiv:2410.13341.

Evan Frick, Tianle Li, Connor Chen, Wei-Lin Chiang, Anastasios N. Angelopoulos, Jiantao Jiao, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. How to evaluate reward models for RLHF. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.14872. arXiv:2410.14872.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, et al. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*, 2024. doi: 10.48550/arXiv.2408.00118.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437.

Shashwat Goel, Joschka Strüber, Ilze Amanda Auzina, Karuna K. Chandra, Ponnurangam Kumaraguru, Douwe Kiela, Ameya Prabhu, Matthias Bethge, and Jonas Geiping. Great models think alike and this undermines AI oversight. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 19621–19678, 2025. doi: 10.48550/arXiv.2502.04313. arXiv:2502.04313.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. doi: 10.48550/arXiv.2407.21783.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023. doi: 10.48550/arXiv.2310.06825.

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. doi: 10.48550/arXiv.2407.18370. arXiv:2407.18370.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066, 2025. doi: 10.48550/arXiv.2506.07962. arXiv:2506.07962.

Meelis Kull, Telmo M. Silva Filho, and Peter Flach. Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2): 5052–5080, 2017. doi: 10.1214/17-EJS1338SI.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, LJ Miranda, Bill Yuchen Lin, Khyathi Chandu, Nouha Dziri, Sachin Kumar, Tom Zick, Yejin Choi, Noah A. Smith, and Hannaneh Hajishirzi. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797, 2025. doi: 10.18653/v1/2025.findings-naacl.96.

Olivier Ledoit and Michael Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004. doi: 10.1016/S0047-259X(03)00096-4.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations (ICLR)*, 2026. doi: 10.48550/arXiv.2506.01937. arXiv:2506.01937.

Putra Manggala, Atalanti Mastakouri, Elke Kirschbaum, Shiva Prasad Kasiviswanathan, and Aaditya Ramdas. QA-Calibration of language model confidence scores. In *International Conference on Learning Representations (ICLR)*, 2025. doi: 10.48550/arXiv.2410.06615. arXiv:2410.06615.

Meta AI. The Llama 4 herd: The beginning of a new era of natively multimodal AI innovation. Blog post, 2025. URL `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

Shmuel Nitzan and Jacob Paroush. Optimal decision rules in uncertain dichotomous choice situations. *International Economic Review*, 23(2):289–297, 1982. doi: 10.2307/2526438.

OpenAI. GPT-4o mini: Advancing cost-efficient intelligence. Blog post, 2024. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`.

John C. Platt. Probabilities for SV machines. In Alexander J. Smola, Peter Bartlett, Bernhard Schölkopf, and Dale Schuurmans (eds.), *Advances in Large-Margin Classifiers*, pp. 61–74. MIT Press, 2000. doi: 10.7551/mitpress/1113.003.0008.

Mengjie Qian, Guangzhi Sun, Mark J. F. Gales, and Kate M. Knill. Who can we trust? LLM-as-a-jury for comparative assessment. *arXiv preprint arXiv:2602.16610*, 2026. doi: 10.48550/arXiv.2602.16610.

Qwen Team, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. doi: 10.48550/arXiv.2412.15115.

Bhaktipriya Radharapu, Eshika Saxena, Kenneth Li, Chenxi Whitehouse, Adina Williams, and Nicola Cancedda. Calibrating LLM judges: Linear probes for fast and reliable uncertainty estimation. *arXiv preprint arXiv:2512.22245*, 2025. doi: 10.48550/arXiv.2512.22245.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797.

Vikas C. Raykar, Shipeng Yu, Linda H. Zhao, Anna Jerebko, Charles Florin, Gerardo Hermosillo Valadez, Luca Bogoni, and Linda Moy. Supervised learning from multiple experts: Whom to trust when everyone lies a bit. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 889–896, 2009. doi: 10.1145/1553374.1553488.

Aishwarya Sahoo, Jeevana Kruthi Karnuthala, Tushar Parmanand Budhwani, Pranchal Agarwal, Sankaran Vaidyanathan, Alexa Siu, Franck Dernoncourt, Jennifer Healey, Nedim Lipka, Ryan Rossi, Uttaran Bhattacharya, and Branislav Kveton. Quantitative LLM judges. *arXiv preprint arXiv:2506.02945*, 2025. doi: 10.48550/arXiv.2506.02945.

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

Zailong Tian, Zhuoheng Han, Yanzhe Chen, Haozhe Xu, Xi Yang, Richeng Xuan, Houfeng Wang, and Lizi Liao. Overconfidence in LLM-as-a-Judge: Diagnosis and confidence-driven solution. *arXiv preprint arXiv:2508.06225*, 2025. doi: 10.48550/arXiv.2508.06225.

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. doi: 10.48550/arXiv.2404.18796.

Sida Wang. Measuring all the noises of LLM evals. *arXiv preprint arXiv:2512.21326*, 2025. doi: 10.48550/arXiv.2512.21326.

David H. Wolpert. Stacked generalization. *Neural Networks*, 5(2):241–259, 1992. doi: 10.1016/S0893-6080(05)80023-1.

Mingyuan Xu, Xinzi Tan, Jiawei Wu, and Doudou Zhou. A judge-aware ranking framework for evaluating large language models without ground truth. *arXiv preprint arXiv:2601.21817*, 2026. doi: 10.48550/arXiv.2601.21817.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. doi: 10.48550/arXiv.2505.09388.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *International Conference on Learning Representations (ICLR)*, 2024. doi: 10.48550/arXiv.2310.07641. arXiv:2310.07641.

Jitian Zhao, Changho Shin, Tzu-Heng Huang, Satya Sai Srinath Namburi GNVV, and Frederic Sala. CARE: Confounder-aware aggregation for reliable LLM evaluation. *arXiv preprint arXiv:2603.00039*, 2026. doi: 10.48550/arXiv.2603.00039.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020.

Bin Zhu, Yi Xie, and Yanghui Rao. A finite-calibration regime map for LLM judge panels. *arXiv preprint arXiv:2606.01034*, 2026. doi: 10.48550/arXiv.2606.01034.

A Proofs and Supporting Results

Equation (2) as a sum of log-likelihood ratios. Under the one-coin model, judge k reports the correct side with probability a_k independently of the other judges given z_t, and $a_k \mid \mathcal{D} \sim \mathrm{Beta}(\alpha_k, \beta_k)$ with $\alpha_k = c_k{+}1$, $\beta_k = n_k{-}c_k{+}1$ for c_k correct votes out of n_k. Integrating a_k against its posterior, which is independent across judges, gives the posterior-predictive probabilities $P(y_{tk} = 1 \mid z_t = 1, \mathcal{D}) = \mathbb{E}[a_k \mid \mathcal{D}] = \alpha_k/(\alpha_k{+}\beta_k)$ and $P(y_{tk} = 1 \mid z_t = 0, \mathcal{D}) = \beta_k/(\alpha_k{+}\beta_k)$, so a vote of 1 has likelihood ratio α_k/β_k and a vote of 0 its reciprocal. Bayes' rule over the judges in $S \cap J_t$ yields Eq. (2), provided that the availability set J_t is independent of z_t given $\mathcal{D}$; otherwise the log odds gain the term $\log\{P(J_t \mid z_t = 1, \mathcal{D})/P(J_t \mid z_t = 0, \mathcal{D})\}$. With $q_k = \alpha_k/(\alpha_k + \beta_k)$, the expected contribution of judge k under $z_t = 1$ is $(2q_k{-}1)\log\{q_k/(1{-}q_k)\} = \mathrm{KL}\{\mathrm{Bern}(q_k) \,\|\, \mathrm{Bern}(1{-}q_k)\}$. This is a different divergence from the Bhattacharyya exponent of Appendix A.1.

Proof of Proposition 1. Since $S \subseteq T$, $\sigma(\mathbf{X}_S) \subseteq \sigma(\mathbf{X}_T)$. For Brier loss the Bayes risk is $\mathbb{E}[\mathrm{Var}(Y \mid \mathbf{X}_S)]$ and the law of total variance gives $\mathrm{Var}(Y \mid \mathbf{X}_S) = \mathbb{E}[\mathrm{Var}(Y \mid \mathbf{X}_T) \mid \mathbf{X}_S] + \mathrm{Var}(\eta_T \mid \mathbf{X}_S)$; taking expectations proves monotonicity, with equality iff $\eta_T = \eta_S$ almost surely. For NLL the Bayes risk is the conditional entropy and $H(Y \mid \mathbf{X}_S) - H(Y \mid \mathbf{X}_T) = I(Y; \mathbf{X}_{T\setminus S} \mid \mathbf{X}_S) \geq 0$, which for binary Y vanishes iff $\eta_T = \eta_S$ almost surely. For a general strictly proper binary loss L with finite Bayes risk, the generalized entropy $H_L(q) = \inf_p\{qL(1,p) + (1-q)L(0,p)\}$ is strictly concave, and $\eta_S = \mathbb{E}[\eta_T \mid \mathbf{X}_S]$ by the tower property, so conditional Jensen gives $\mathbb{E}[H_L(\eta_T)] \leq \mathbb{E}[H_L(\eta_S)]$ with equality iff $\eta_T = \eta_S$ almost surely. No independence assumption is used. □

Proof of Proposition 2. Condition on D, so the fitted predictors are fixed functions. For log loss, $R_{\mathrm{NLL}}(\hat{\eta}_U \mid D) = H(Y \mid \mathbf{X}_U) + \mathbb{E}[\mathrm{KL}\{\mathrm{Bern}(\eta_U) \| \mathrm{Bern}(\hat{\eta}_U)\} \mid D]$, and the second term is $e_U^{\mathrm{NLL}}(D)$. Subtracting the identities for T and S and applying the chain rule for entropy gives the first display. For Brier loss, $R_{\mathrm{Brier}}(\hat{\eta}_U \mid D) = \mathbb{E}[\mathrm{Var}(Y \mid \mathbf{X}_U)] + \mathbb{E}[(\hat{\eta}_U - \eta_U)^2 \mid D]$, and $\mathbb{E}[\mathrm{Var}(Y \mid \mathbf{X}_S)] - \mathbb{E}[\mathrm{Var}(Y \mid \mathbf{X}_T)] = \mathbb{E}[(\eta_T - \eta_S)^2]$ by total variance, which gives the second display. Both are identities; neither bounds $e_T - e_S$. □

Lemma 1 (Monotone calibration depends on the score only through its ordering)**.** *Let s be a real-valued score, h a strictly increasing map, and $\mathcal{G}$ the class of non-decreasing maps into $[0,1]$. For any proper loss L, $\inf_{g\in\mathcal{G}} \mathbb{E}[L(Y, g(s))] = \inf_{g\in\mathcal{G}} \mathbb{E}[L(Y, g(h(s)))]$, and for Brier loss and NLL the infimum is attained by the isotonic regression of Y on s.*

Proof. If $g \in \mathcal{G}$, then $g \circ h^{-1}$ is non-decreasing on the range of h and extends to a non-decreasing map, so every risk attained with s is attained with $h(s)$; the converse uses $g \circ h$. Hence the infimum depends on s only through the ordering it induces. The isotonic regression $\mathbb{E}[Y \mid \text{level sets of the isotonic projection}]$ minimizes squared error over $\mathcal{G}$ and, being a conditional mean of Y on a monotone partition, simultaneously minimizes every Bregman loss, including log loss, over $\mathcal{G}$ (Brummer & du Preez, 2013). □

A consequence used in Section 3: two aggregators whose scores induce the same ordering of items have the same best achievable monotone-calibrated risk, and an aggregator whose score orders items worse cannot be rescued by any monotone post-hoc map, including monotone beta and Platt scaling.

Proof of Proposition 3. (i) Given s, a linear score T is Gaussian with mean $s\mu$ and variance τ^2, where 2μ is the mean difference and $\tau^2 = w^\top \Sigma w$ does not depend on Y because the noise is independent of Y. The posterior log odds given T is $(2\mu/\tau^2)T$, a monotone function of T, so the smallest NLL over all calibration maps is $H(Y \mid T)$ and is attained by a monotone map. With $V = T/\tau$ and $d = 2\mu/\tau$, $V \mid s \sim N(sd/2, 1)$ and $H(Y \mid T) = h(d) = \mathbb{E}\log\{1 + \exp(-d^2/2 - dG)\}$ with $G \sim N(0,1)$. For $d' > d > 0$, $V_d \overset{d}{=} (d/d')V_{d'} + \{1 - (d/d')^2\}^{1/2} G'$ with independent G', a garbling of $V_{d'}$. The data-processing inequality gives $h(d) \geq h(d')$, with equality only if the posterior given $V_{d'}$ were a function of V_d, which it is not; hence h is strictly decreasing. (ii) Let S_R and S_D be the block means. Given s they are independent, $S_R \sim N(s\Delta_R/2, v_R)$ with $v_R = \rho + (1-\rho)/m$ and $S_D \sim N(s\Delta_D/2, 1/r)$, and within each block the judges are exchangeable, so the Bayes score is linear in (S_R, S_D) with $d_\star^2 = \Delta_R^2/v_R + r\Delta_D^2$ (the Mahalanobis separation). It is nondecreasing in m because v_R is nonincreasing in m. (iii) The conditionally independent score uses the per-judge log-likelihood ratio weights Δ_R and Δ_D, so $T = m\Delta_R S_R + r\Delta_D S_D$, whose mean difference is $m\Delta_R^2 + r\Delta_D^2$ and whose variance is $m^2\Delta_R^2 v_R + r\Delta_D^2$; the displayed d_{CI}^2 follows, and so does its limit. Every diverse judge is more accurate than every redundant one when $\Delta_D > \Delta_R$, so accuracy ranking selects the diverse block. Its judges are independent, so the conditionally independent score is the Bayes score for that block, with $d_D^2 = r\Delta_D^2$. If $\rho > \Delta_R^2/(r\Delta_D^2)$ the limit Δ_R^2/ρ is below d_D^2, so $d_{\mathrm{CI}} < d_D$ for all large m, and by (i) the curated panel has strictly smaller calibrated NLL. Finally $d_\star^2 - d_D^2 = \Delta_R^2/v_R > 0$. (iv) For k exchangeable redundant judges the Bayes score is their sum, with $d^2 = (k\Delta_R)^2/(k^2\rho + k(1-\rho)) = k\Delta_R^2/(1 + (k-1)\rho)$, which increases to Δ_R^2/ρ and never reaches it for $\rho > 0$. One redundant judge and $k-1$ diverse judges are independent given Y, so their Bayes separation is $\Delta_R^2 + (k-1)\Delta_D^2$. This is at least Δ_R^2/ρ whenever $(k-1)\Delta_D^2 \geq \Delta_R^2(1-\rho)/\rho$, and then it strictly exceeds the accuracy-ranked panel's separation. □

Theorem 1 (Finite-candidate selection)**.** *Conditional on development data used for fitting and calibration, let predictors $f_1, \ldots, f_M$ (e.g., fitted panels) with costs $c_m \geq 0$ be fixed before observing*

an independent i.i.d. selection sample of size n, with predictions clipped to $[\epsilon, 1-\epsilon]$. For $\lambda \geq 0$ let $\hat{m} \in \arg\min_m \{\widehat{R}_{\mathrm{NLL}}(f_m) + \lambda c_m\}$. With probability at least $1-\delta$,

$$R_{\mathrm{NLL}}(f_{\hat{m}}) + \lambda c_{\hat{m}} \leq \min_{m \leq M} \{R_{\mathrm{NLL}}(f_m) + \lambda c_m\} + |\log \epsilon| \sqrt{2 \log(2M/\delta)/n}.$$

Proof of Theorem 1. Conditional on the fitting and calibration data, each clipped log loss is a fixed function taking values in $[0, B]$ with $B = |\log \epsilon|$. By Hoeffding's inequality and a union bound, with probability at least $1-\delta$, $|\widehat{R}(f_m) - R(f_m)| \leq B\sqrt{\log(2M/\delta)/(2n)}$ simultaneously for all m. Let $m^\star$ minimize $R(f_m) + \lambda c_m$. Then $R(f_{\hat{m}}) + \lambda c_{\hat{m}} \leq \widehat{R}(f_{\hat{m}}) + \lambda c_{\hat{m}} + B\sqrt{\log(2M/\delta)/(2n)} \leq \widehat{R}(f_{m^\star}) + \lambda c_{m^\star} + B\sqrt{\log(2M/\delta)/(2n)} \leq R(f_{m^\star}) + \lambda c_{m^\star} + 2B\sqrt{\log(2M/\delta)/(2n)}$, and $2\sqrt{x/2} = \sqrt{2x}$. Within-item dependence among judges is irrelevant because the argument concerns i.i.d. items. The guarantee fails if the candidate library or its tuning depends on the selection sample. □

A.1 Idealized jury bounds and sign estimation

Theorem 2 (Oracle weighted-jury bound)**.** *Assume equal class priors, known one-coin reliabilities $p_k \in (0,1)\setminus\{1/2\}$ and conditionally independent votes $X_k \in \{-1, +1\}$ with $P(X_k = Y \mid Y) = p_k$. Let δ_K threshold $\sum_k \log\{p_k/(1-p_k)\}X_k$ with any fixed or randomized tie rule. Then $P(\delta_K \neq Y) \leq \frac{1}{2}\exp(-\sum_k C_k)$ with $C_k = -\frac{1}{2}\log\{1-(2p_k-1)^2\} \geq \frac{1}{2}(2p_k-1)^2$.*

Proof. The rule is the likelihood-ratio (Bayes) rule. The Bhattacharyya coefficient of judge k's two class-conditional channels is $2\sqrt{p_k(1-p_k)} = \sqrt{1-(2p_k-1)^2}$; independence makes the joint coefficient the product $\exp(-\sum_k C_k)$, and the Bhattacharyya bound for equal priors gives the claim. The lower bound on C_k follows from $-\log(1-u) \geq u$. □

Below-chance judges ($p_k < 1/2$) contribute to the exponent after sign reversal (Nitzan & Paroush, 1982; Berend & Paroush, 1998). The theorem needs the sign of $p_k - 1/2$ to be known. For n_k i.i.d. labeled votes and the sample accuracy $\hat{p}_k$ (ties counted as errors), Hoeffding's inequality gives $P\{\mathrm{sign}(\hat{p}_k - \frac{1}{2}) \neq \mathrm{sign}(p_k - \frac{1}{2})\} \leq \exp\{-2n_k(p_k - \frac{1}{2})^2\}$, which is uninformative when $|p_k - 1/2| \lesssim n_k^{-1/2}$. Near-chance judges should therefore be shrunk toward zero weight rather than confidently reversed; the Beta(1,1) posterior in Eq. (2) does this, and ridge and shrinkage LDA do so through their penalties. Neither result is used to support an empirical claim.

B Pre-registered Analyses

The hypotheses, decision rules, generative model, grids and seeds below were written in a protocol file that was committed to our repository before any of the corresponding jobs was submitted; every result file records the protocol's SHA-256. The protocol and all outputs are in the supplementary material. Table 3 lists the outcomes exactly as specified.

C Benchmarks, Panels, and Protocol Details

Panels. Table 6 summarizes the panels. JudgeBench and LLMBar ship verdicts from released judges; we add API-hosted and open-weight LLM judges to widen backbone coverage. RewardBench and RewardBench 2 ship reward-model scores for each candidate response; we add four open-weight LLM judges to RewardBench only. Coverage is the number of items with a parseable vote; missing votes are mean-imputed from the fitting slice by the LDA and ridge aggregators and skipped by the one- and two-coin models.

Prompting and parsing. Added LLM judges receive an MT-Bench-style pairwise prompt (Zheng et al., 2023) with the two responses labeled Assistant A and Assistant B in an order randomized independently per item, and must end with `[[A]]` or `[[B]]`. API judges use temperature 0; open-weight judges use temperature 0.01. After removing any `<think>` block, a parser tries, in order, the bracketed token, a bold `**A**`/`**B**`, a trailing "Assistant/Output/Solution A" in the last 80 characters, and a bare trailing letter; if none matches the vote is missing. The parsed verdict is

Table 3: Pre-registered hypotheses and outcomes.

	Prediction (decision rule)	Outcome
H1	One-coin $\Delta_{\mathrm{A-T}}$ interval below 0 at $\rho = 0$ and above 0 at every $\rho \geq 0.5$	held
H2	LDA and ridge: cell mean $\Delta_{\mathrm{A-T}} \leq 0.01$ at every ρ, and ≥ 0.02 below one-coin for $\rho \geq 0.3$	held
H3	Interaction interval below 0 for $\rho \geq 0.2$; Spearman$(\rho, \lvert\text{interaction}\rvert) \geq 0.9$	held (Spearman 1.0)
H4	Spearman$(m$, one-coin $\Delta_{\mathrm{A-T}}) \geq 0.9$ over the m grid	held (Spearman 1.0)
B1	Calibrated interaction < 0 on PPE, Preference-Dissection and Chatbot-Arena	held (all intervals < 0)
B2	Spearman(redundancy, $\lvert\text{interaction}\rvert) > 0$ over the seven real panels	held (Spearman 0.5, $p = 0.27$)†

†Exact two-sided permutation p-value, added after the fact. The pre-registered evaluation used the earlier RewardBench 2 encoding of Appendix C.1; with the corrected 144-judge panel the Spearman correlation is again 0.5 ($p = 0.27$).

Table 4: Simulation, varying the redundant-block size m ($\rho = 0.5$, $N = 2000$) and the number of items N ($m = 50$, $\rho = 0.5$). Entries are $\Delta_{\mathrm{A-T}}$ means over 30 replicates.

m	5	10	25	50	100	200
One-coin	+0.009	+0.091	+0.260	+0.327	+0.345	+0.356
LDA (full)	−0.028	−0.035	−0.029	−0.020	−0.004	+0.047
Ridge	−0.028	−0.034	−0.028	−0.015	+0.005	+0.048
Redundancy	0.17	0.38	0.67	0.82	0.91	0.95
N	200	500	1,000	2,000	4,000	
One-coin	+0.294	+0.283	+0.317	+0.311	+0.319	
LDA (full)	−0.007	+0.012	−0.012	−0.023	−0.029	
Ridge	+0.011	+0.011	−0.004	−0.016	−0.023	

mapped to the canonical A/B coordinates and stored as a vote; it is never compared with the label when features are built (Figure 4).

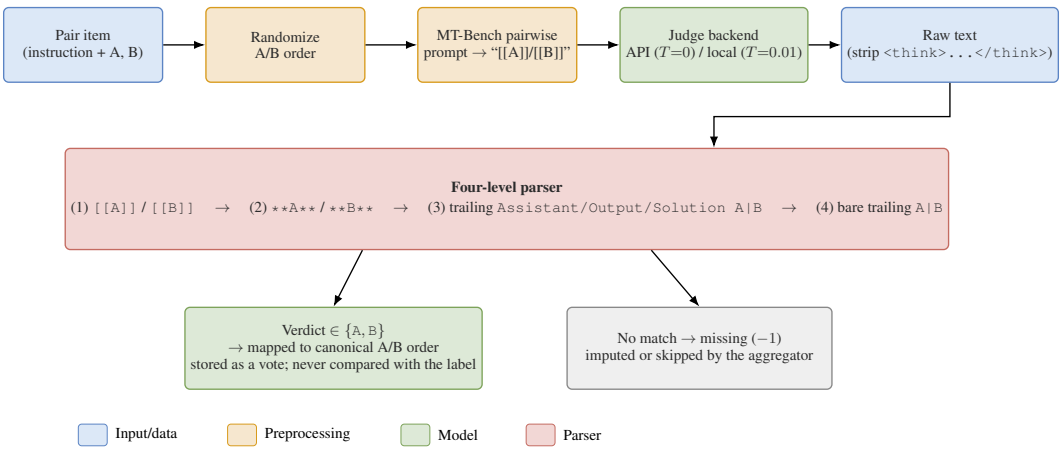

Figure 4: Inference pipeline for the added LLM judges. Each item is shown in a random A/B order inside a fixed pairwise prompt; the judge's text is parsed by a four-level fallback parser, and unparseable outputs are recorded as missing. Parsed verdicts are stored as votes in canonical A/B coordinates; labels are never used to build them.

Table 5: Redundancy of the seven real panels on development items and the calibrated covariance interaction (four benchmarks: frozen split, item bootstrap; three held-out panels: frozen group partition, group bootstrap). $\bar{\rho}$: mean off-diagonal within-class vote correlation.

Panel	K	K_{eff}	$1 - K_{\text{eff}}/K$	$\bar{\rho}$	Interaction [95% CI]
JudgeBench	38	13.3	0.65	0.20	$-0.051\,[-0.102, -0.001]$
RewardBench	104	19.0	0.82	0.16	$-0.014\,[-0.037, +0.009]$
RewardBench 2	144	5.9	0.96	0.37	$-0.111\,[-0.157, -0.068]$
LLMBar	53	7.7	0.86	0.26	$-0.107\,[-0.159, -0.054]$
PPE (held out)	19	7.6	0.60	0.13	$-0.035\,[-0.044, -0.025]$
Preference-Dissection (held out)	32	3.0	0.91	0.54	$-0.047\,[-0.061, -0.033]$
Chatbot-Arena (held out)	17	3.4	0.80	0.47	$-0.022\,[-0.025, -0.019]$

Table 6: Panel composition. Added LLM judges: GPT-4o-mini (OpenAI, 2024), Claude 3 Haiku (Anthropic, 2024), Qwen3-32B (Yang et al., 2025), Llama 3.1 8B and Llama 3.3 70B (Grattafiori et al., 2024), Llama 4 Scout (Meta AI, 2025) (API); Qwen2.5-7B-Instruct (Qwen Team et al., 2024), Gemma-2-9B-it (Gemma Team et al., 2024), Llama-3.1-8B-Instruct, Mistral-7B-Instruct-v0.3 (Jiang et al., 2023) (open-weight, 4-bit).

Benchmark	Items	K	Released judges	Added judges (coverage)
JudgeBench	350	38	32 LLM judges	6 API (164–309 of 350)
RewardBench	2,985	104	100 reward models	4 open-weight (93–100%)
RewardBench 2	1,845	144	144 reward models*	none
LLMBar	419	53	44 evaluator configurations	5 API (≈100%), 4 open-weight (227–419 of 419)
PPE (external)	12,775	19	13 reward models, 6 LLM-judge configs	none

*Of the 174 models with released per-response scores; the 30 excluded checkpoints come from a label-flipping ablation (1–100% of training preferences flipped).

Reward-model votes. A reward model scores each response on its own, so its preference between two responses does not depend on presentation order. For each item we draw a balanced random assignment of the chosen response to position A or B, independent of all scores (seed fixed across analyses), set $z_t = 1$ if the chosen response is in position A, and record the vote $y_{tk} = 1$ iff the model scores position A higher. This is the vote the model would output under that presentation; it has no position bias by construction. RewardBench 2 releases, for each prompt, one correct and three incorrect responses (more on its Ties subset) and each model's score for every response. We draw, once for all judges (seed 20260925), one correct and one incorrect response per prompt and treat that pair as the item; tied or missing scores are missing votes (0.05% of entries). Twenty rows whose prompt identifier occurs twice in the release cannot be aligned across models and are dropped. On the remaining 1,845 items, the best-of-4 outcomes we had stored agree exactly with the per-item `results` field of the release for all 174 models, which fixes the correspondence between our judge names and the released files. Pairwise accuracy ranges from 0.36 to 0.96 (median 0.85) on the 144-judge panel; with the 30 label-flipping checkpoints it falls to 0.17, and 5 judges are below $1/2$. The four LLM judges added to RewardBench were stored as agreement with the preferred response; we place them in the same random order as the reward models, so their votes carry no position bias.

RewardBench 2 with all 174 judges. Keeping the 30 label-flipping checkpoints changes no conclusion. At the largest budget, beta-calibrated $\Delta_{\text{A}-\text{T}}$ is $+0.148$ (one-coin), $+0.148$ (two-coin), $+0.015\,[-0.007, +0.053]$ (LDA) and $+0.016\,[-0.006, +0.042]$ (ridge). The calibrated covariance interaction is $-0.116\,[-0.160, -0.073]$ on the frozen split and negative on all 21 splits, and the redundancy index is 0.97 ($K_{\text{eff}} = 5.9$). In the selection experiment, ridge gives NLL 0.164 (All), 0.160 (Top-5), 0.160 (accuracy path) and 0.164 (risk-based, 38.8 judges); LDA gives 0.191, 0.175, 0.175 and 0.174.

PPE. We use the objective-correctness conflict pairs of PPE (Frick et al., 2025) for MMLU-Pro, MATH, GPQA, IFEval and MBPP-Plus, with 13 reward models and 6 released LLM-judge configurations (a hand-coded verbosity baseline is excluded before any outcome is computed). Pairs sharing a question stay in the same split and bootstrap cluster. The pooled five-domain endpoint was declared primary before fitting; domain results are secondary (Table 11).

C.1 Two correctness-encoding pitfalls

Pairwise benchmarks often store each judge's output as *correct* or *incorrect* relative to the known winner, because that is what accuracy leaderboards need. Using that indicator as the aggregator's input and scoring against the known winner leaks the evaluation label into the features: every column is then a noisy copy of a constant target, every added judge looks informative, and larger panels win almost mechanically. Sparse selectors also fail to sparsify, because no column is uninformative. An earlier version of our evaluation harness had exactly this defect. Every result in this paper is computed from vote-level features in canonical A/B coordinates, and the code that builds votes from the released judge outputs does not read evaluation labels. For reward models the vote is computed from the released scores of the two responses under the drawn presentation order, which is exactly the information a judge sees in a randomized-order query; the label records that order and is otherwise used only for metrics.

The second pitfall concerns best-of-N benchmarks. RewardBench 2 reports whether a model scores the correct response above all three incorrect ones. Recoding that outcome as a pairwise vote under a random order produces a "judge" whose chance accuracy is $1/4$, not $1/2$: weak but informative rankers then look like below-chance anti-judges, sign-flipping aggregators exploit them, and full panels appear to gain large amounts over accuracy-ranked subsets. An earlier version of this analysis used that recoding and reported large RewardBench 2 gains for the full panel and for risk-based selection; with genuine pairwise votes (Appendix C) those gains disappear, while the covariance interaction keeps its sign. We recommend building pairwise votes from per-response scores and checking that chance accuracy is $1/2$.

D Additional Results

Table 7: Raw (uncalibrated) $\Delta_{\mathrm{A-T}}$ at the largest budget, 50 repeated splits. The aggregator-driven pattern is already present before calibration.

Benchmark	One-coin	Two-coin	Shrinkage LDA	Ridge
JudgeBench	$+0.620$ $[+0.397, +0.882]$	$+0.630$ $[+0.398, +0.901]$	$+0.107$ $[+0.012, +0.235]$	$+0.067$ $[+0.028, +0.111]$
RewardBench	$+0.210$ $[+0.155, +0.279]$	$+0.211$ $[+0.158, +0.266]$	-0.008 $[-0.025, +0.013]$	-0.014 $[-0.030, +0.002]$
RewardBench 2	$+0.376$ $[+0.298, +0.444]$	$+0.376$ $[+0.301, +0.448]$	$+0.021$ $[-0.024, +0.094]$	$+0.016$ $[-0.009, +0.049]$
LLMBar (53)	$+0.135$ $[-0.068, +0.273]$	$+0.139$ $[-0.063, +0.281]$	-0.045 $[-0.134, +0.051]$	-0.060 $[-0.107, -0.017]$

Label budget. Within each of 50 repetitions the evaluation half is fixed and the calibration budget grows through nested prefixes of the development half (70% of each budget fits the aggregator, 30% the calibrator). Table 8 gives the smallest budget from which the beta-calibrated split range of $\Delta_{\mathrm{A-T}}$ stays on one side of zero at every larger budget. On a 10-judge LLMBar subpanel of the added API and open-weight judges, no aggregator resolves at any budget.

Matched covariance intervention: full results. Table 9 reports both arms and the interaction on the frozen 60/40 split, with split summaries over the frozen split and 20 further splits. The reward-model orientation seed matches the selection experiments.

Below-chance judges. On all labeled items, 1 of 38 JudgeBench judges, 6 of 104 RewardBench judges, 1 of 144 RewardBench 2 judges and 25 of 53 LLMBar judges are below $1/2$ (LLMBar median 0.507, minimum 0.39). We reran the budget sweep and the selection experiment after removing, within each split, every judge below $1/2$ on the fitting slice; evaluation labels play no role

Table 8: Smallest calibration budget from which the 2.5/97.5 split range of beta-calibrated $\Delta_{\mathrm{A-T}}$ excludes zero at all larger budgets. "Unresolved": no such budget. Budget grids: JudgeBench 20–175, RewardBench 20–1,492, RewardBench 2 20–922, LLMBar 20–209.

Benchmark	One-coin	Two-coin	Shrinkage LDA	Ridge
JudgeBench	unresolved	unresolved	unresolved	unresolved
RewardBench	Top-5 from 100	Top-5 from 100	unresolved	All at 1,492
RewardBench 2	Top-5 from 250	Top-5 from 150	unresolved	unresolved
LLMBar (53)	unresolved	unresolved	unresolved	unresolved

Table 9: Matched covariance intervention. $\Delta^{\mathrm{diag}}_{\mathrm{A-T}}$ and $\Delta^{\mathrm{full}}_{\mathrm{A-T}}$: All minus Top-5 under the diagonal and full covariance of one LDA fit; interaction $= \Delta^{\mathrm{full}}_{\mathrm{A-T}} - \Delta^{\mathrm{diag}}_{\mathrm{A-T}}$. Frozen split: 95% item-bootstrap interval. Splits: mean, 2.5/97.5 range, and fraction negative over 21 splits.

		$\Delta^{\mathrm{diag}}_{\mathrm{A-T}}$	$\Delta^{\mathrm{full}}_{\mathrm{A-T}}$	Interaction (frozen)	Interaction (21 splits)
Raw	JudgeBench	$+0.423$	$+0.060$	$-0.363\,[-0.564, -0.166]$	$-0.490\,[-0.732, -0.178]$, 1.00
	RewardBench	$+0.028$	$+0.010$	$-0.019\,[-0.075, +0.038]$	$-0.058\,[-0.089, -0.020]$, 1.00
	RewardBench 2	$+0.303$	-0.001	$-0.304\,[-0.430, -0.183]$	$-0.331\,[-0.406, -0.268]$, 1.00
	LLMBar	$+0.119$	-0.076	$-0.195\,[-0.344, -0.059]$	$-0.146\,[-0.273, -0.037]$, 1.00
Beta	JudgeBench	$+0.082$	$+0.031$	$-0.051\,[-0.102, -0.001]$	$-0.065\,[-0.111, -0.013]$, 1.00
	RewardBench	$+0.016$	$+0.002$	$-0.014\,[-0.037, +0.009]$	$-0.028\,[-0.042, -0.013]$, 1.00
	RewardBench 2	$+0.109$	-0.002	$-0.111\,[-0.157, -0.068]$	$-0.120\,[-0.147, -0.094]$, 1.00
	LLMBar	$+0.022$	-0.085	$-0.107\,[-0.159, -0.054]$	$-0.073\,[-0.123, -0.038]$, 1.00

in the removal. On average 36.2, 97.8, 142.7 and 29.2 judges remain. At the largest budget, beta-calibrated $\Delta_{\mathrm{A-T}}$ becomes $+0.105$, $+0.096$, $+0.034$, $+0.056$ on JudgeBench (one-coin, two-coin, LDA, ridge), $+0.090$, $+0.091$, -0.005, -0.017 on RewardBench, $+0.134$, $+0.134$, $+0.013$, $+0.014$ on RewardBench 2, and $+0.082$, $+0.085$, $-0.018\,[-0.079, +0.045]$, $-0.022\,[-0.059, +0.027]$ on LLMBar. In the selection experiment on LLMBar, ridge gives NLL 0.339 (All), 0.347 (Top-5), 0.342 (accuracy path) and 0.338 (risk-based), and LDA gives 0.346, 0.360, 0.347 and 0.343; every rule except Top-5, which is unchanged, is worse than with the below-chance judges, which therefore carry information that accuracy ranking discards. In the matched covariance intervention on the frozen LLMBar split, with below-chance judges removed from All on the data each model is fitted on (29 of 53 remain), the calibrated interaction is $-0.102\,[-0.153, -0.052]$ (split mean -0.085, negative on all 21 splits), with $\Delta^{\mathrm{full}}_{\mathrm{A-T}} = -0.040$ and $\Delta^{\mathrm{diag}}_{\mathrm{A-T}} = +0.062$. On the other benchmarks, risk-based selection stays within 0.018 NLL of the accuracy path.

Other RewardBench 2 pair draws. Each RewardBench 2 item is one drawn correct/incorrect pair. With two further draws (seeds 20260926 and 20260927), beta-calibrated $\Delta_{\mathrm{A-T}}$ at the largest budget is $+0.145$ and $+0.123$ (one-coin), $+0.010$ and $+0.011$ (LDA), and $+0.013$ and $+0.014$ (ridge); the calibrated covariance interaction is $-0.120\,[-0.166, -0.073]$ and $-0.116\,[-0.166, -0.069]$, negative on all 21 splits. With ridge, NLL is 0.162 and 0.193 for All, 0.163 and 0.191 for Top-5, 0.163 and 0.188 for the accuracy path, and 0.159 and 0.187 for risk-based selection.

Cost-aware selection. Adding a call price $\lambda|S|/K$ to Eq. (1) moves along the frontier. With ridge on RewardBench 2, $\lambda \in \{0, 0.005, 0.01, 0.02, 0.05\}$ selects 25.9, 18.7, 15.5, 12.9 and 8.4 judges on average, all at NLL 0.162. On the 53-judge LLMBar panel the same range selects 18.0 down to 10.5 judges with NLL 0.295–0.298, against 0.298 for all 53. On RewardBench, $\lambda = 0.05$ selects 17.3 judges at NLL 0.118 versus 0.113 for all 104; on JudgeBench it selects 6.3 judges at 0.479. An ℓ_1-logistic path fitted on the frozen development split with a one-standard-error rule selects $C = 0.3$ for all three fit seeds, keeping 8 of 38 JudgeBench judges (test raw NLL 0.421) and 11 of 53 LLMBar judges (0.251).

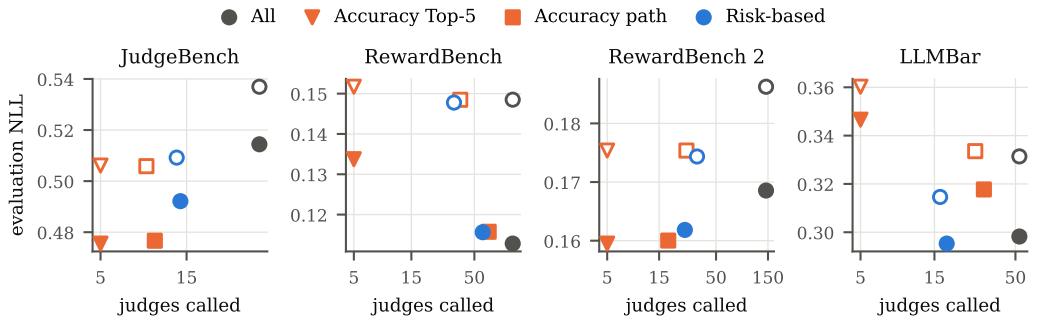

Figure 5: Evaluation NLL against the mean number of judges called (log scale), averaged over 50 splits. Filled markers: ridge + beta; hollow markers: shrinkage LDA + beta. Orange rules rank judges by accuracy; blue selects by held-out calibrated risk from a library that contains All and the orange prefixes.

Which stage matters. Table 10 runs every stage subset on 30 disjoint model/calibration/conformal partitions with a fixed test set. Platt and beta maps are near-substitutes; both repair the one-coin aggregator and neither helps much after ridge. The conformal stage leaves NLL unchanged by construction.

Table 10: Test NLL of the full panel by stage subset (30 partitions).

Stages	JB one-coin	JB ridge	LLMBar one-coin	LLMBar ridge
Aggregation only	1.000	0.498	0.729	0.234
+ Platt	0.516	0.468	0.373	0.245
+ monotone beta only	0.538	0.471	0.386	0.245
+ Platt then beta	0.523	0.471	0.377	0.248

Conformal sets. With ridge and strict split conformal on disjoint slices (50 repetitions, nominal 90%), realized coverage is 0.935 (All) and 0.931 (Top-5) on JudgeBench with mean set sizes 1.41 and 1.35, and 0.937 and 0.938 on LLMBar with sizes 1.10 and 1.25. On LLMBar the full panel yields singleton sets for 89% of items versus 74% for Top-5.

Position effects. On the 44 released LLMBar configurations, which were run in both presentation orders, adding position and vote-by-position terms to ridge does not reduce NLL ($+0.005\,[-0.002, +0.013]$), while averaging the two orders reduces it by $0.017\,[0.010, 0.026]$ at twice the calls. Under a position-aware model the full panel still beats Top-5 ($-0.068\,[-0.112, -0.028]$).

Where the full panel helps on LLMBar. On the 168 frozen LLMBar test items, raw ridge on All has lower NLL than raw ridge on Top-5 for 78% of items overall, for 97% of the adversarial GPTInst items and 96% of the adversarial Neighbor items, but only for 40% of the Natural items, where All is worse on average ($+0.093$ NLL). Grouping items by the majority votes of the top-5 and of the remaining 48 judges, the largest average gains occur when both majorities are wrong (-0.44 NLL, 22 items; the full panel is less confidently wrong) and when the top-5 majority is right but the remaining majority is wrong (-0.18, 61 items).

Table 11: External PPE panel: raw ridge $\Delta_{\mathrm{A-T}}$ on the frozen split with question-group bootstrap intervals. The pooled endpoint was pre-declared as primary.

Endpoint	Items / groups	$\Delta_{\mathrm{A-T}}$ [95% CI]
PPE pooled (primary)	12,775 / 2,555	$-0.0088\,[-0.0146, -0.0033]$
family-deduplicated		$-0.0026\,[-0.0055, +0.0004]$
IFEval	2,560 / 512	$-0.0151\,[-0.0288, -0.0013]$
MBPP-Plus	2,535 / 507	$-0.0248\,[-0.0589, +0.0117]$
MMLU-Pro	2,560 / 512	$-0.0089\,[-0.0183, +0.0002]$
MATH	2,560 / 512	$-0.0053\,[-0.0142, +0.0036]$
GPQA	2,560 / 512	$+0.0039\,[-0.0038, +0.0120]$