THE EFFECT SIZE IS A FREE PARAMETER OF THE JUDGE PANEL: BUY THE SCALE, NOT THE SAMPLE

Anonymous authors
Paper under double-blind review

ABSTRACT

A panel of LLM judges differs from its gold standard in two ways that behave differently. Writing the panel score of system s as $\bar{\mu}_s = \alpha + \beta\theta_s + \tilde{c}_s$, the slope β is a pure scale error that multiplies every reported effect size and by itself reorders nothing, while the residual $\tilde{c}_s$ can reorder a leaderboard and, by a classical aliasing argument, is beyond the reach of gold-free aggregation. We measure both across three settings and 99,487 judge verdicts from up to 26 judges in 14 model families. The central finding is that β is a free parameter of the panel: on identical items under an identical protocol, it runs from 0.45 under an open-weight roster to 1.74 under a frontier one on Chatbot Arena, tracking the panel's mean member accuracy, not the task. Judged effect sizes therefore have no unit until the roster is reported. The fix is cheap: two gold-anchored systems with no residual of their own identify (α, β), and a closed-form rule says when rescaling helps; two anchors decide its $\hat{\beta} \geq 1$ branch, eight buy its full form. Fed an estimated $\hat{\beta}$, the rule's mean regret was never positive, and on a miscalibrated panel 32 anchor labels (64 with the debiased four-anchor estimator) beat 512 spent uniformly under prediction-powered inference on effect-size error (CIs exclude a tie). Anchors cannot buy the ranking: the rank-relevant residual clears a conservative matched null on a deepened open-weight-leaning panel (flip risk $p = 0.02$) but not on a commercial one at the same depth, where it is 18% of a median gold gap; and it is neither predictable from surface features nor transferable across panels. Code and every judge verdict are released in the supplementary material.

1 INTRODUCTION

Panels of LLM judges score each system's outputs, the scores are averaged, and the result is a leaderboard (Zheng et al., 2023; Verga et al., 2024; Gera et al., 2024). The literature has catalogued judge biases (Ye et al., 2024; Panickssery et al., 2024; Feuer et al., 2024) and shown that inter-judge agreement does not certify validity (Norman et al., 2026; Yang et al., 2026; Kohli, 2026; Rao & Callison-Burch, 2026); it has not separated the two ways a judged leaderboard can be wrong, which have opposite consequences and opposite remedies.

Let θ_s be the gold value of system s and $\bar{\mu}_s$ the panel's mean score. Regressing panel on gold splits the discrepancy into an affine part and a residual,

$$\bar{\mu}_s \;=\; \alpha + \beta\,\theta_s + \tilde{c}_s. \tag{1}$$

An increasing affine map preserves order, so β on its own *cannot change a single ranking*: it is a pure scale error multiplying every reported effect size by one factor, while $\tilde{c}_s$ is everything that can reorder a leaderboard (β sets only the margin a residual must clear, §7). The distinction matters because the field's outputs are of both kinds: "model A beats model B" is a claim about order, "by 4 points" one about scale, and meta-analyses, reward thresholds and router margins consume the scale; conflating the two overstates the danger of judge bias (much of it is scale) or understates it (what is left is fixed by nothing).

This paper measures both components at leaderboard scale, and asks what a practitioner can actually buy. The central measurement concerns β:

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.

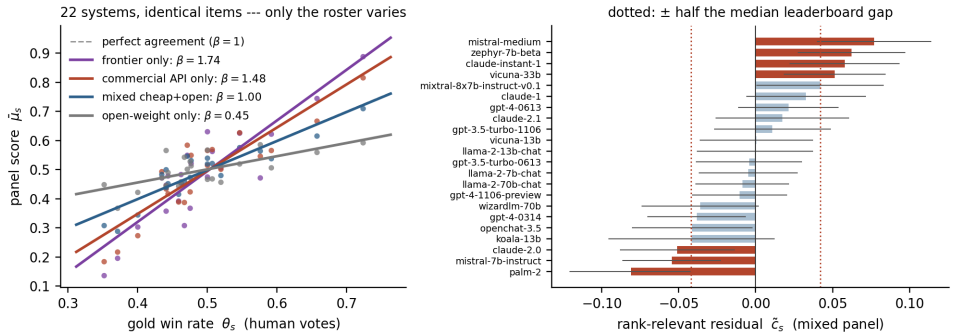


Figure 1: The affine split at leaderboard scale (Chatbot Arena, M=22 systems, 846 human-voted battles). **Left:** panel score against human gold for four rosters from the same enrolled judges on identical items under an identical protocol. The slope, the unit in which every effect size is reported, is set by the roster: commercial judges inflate every gap by 48%, open-weight judges deflate by 55%. **Right:** the rank-relevant residual $\tilde{c}_s$ of the mixed panel with bootstrap standard errors; dotted lines mark $\pm$ half the median leaderboard gap. This is the component that no roster choice, anchor budget, or agreement statistic repairs.

The effect size an LLM-judge benchmark reports is a free parameter of the judge panel.

Varying only which judges are enrolled, on identical items under an identical protocol, the panel-vs-gold slope runs from 0.452 $[0.18, 0.76]$ for an open-weight panel to 1.740 $[1.53, 1.93]$ for a frontier one on Arena, with 0.41–1.27 on MT-Bench and 0.13–0.71 on verifiable grading (Table 1); across random sub-panels the central 90% spans $[0.88, 1.45]$ on Arena (Appendix K). It is not a property of "LLM judges" and not a property of the task: it tracks the panel's mean member accuracy ($r = +0.89$ on Arena), as the Rogan–Gladen identity $\beta = \mathrm{TPR} - \mathrm{FPR}$ predicts in the verifiable regime (Rogan & Gladen, 1978). The 13-family union of the two wings lands at $\beta = 0.996$ by cancellation, a lucky equilibrium rather than a property of breadth. Two papers using different panels on the same benchmark therefore report effect sizes in different, unstated units.

The second contribution is that this is cheap to fix, and we say exactly when fixing it helps. Two gold-anchored systems identify (α, β) (Proposition 1), but rescaling by $1/\beta$ also divides the residual by β, so it is not always a good idea. Proposition 2 makes this a closed-form test: rescaling lowers expected squared effect-size error always when $\beta \geq 1$, and when $\beta < 1$ only if $\sigma_{\tilde{c}}^2/\sigma_\theta^2 < \beta^2(1-\beta)/(1+\beta)$ (the across-system variances of residual and gold), a threshold that never exceeds 0.0902, so in practice weak panels should be left alone; two anchors decide the $\hat{\beta} \geq 1$ branch, and the threshold needs $K \geq 8$ anchored systems (§3). Against the gold-purchasing literature (Angelopoulos et al., 2023a; Chatzi et al., 2024; Divekar, 2026) the practical statement is sharper: on the miscalibrated commercial panel, 32 labels on two anchors, or 64 on the selection-debiased four-anchor estimator, beat 512 spent uniformly under PPI++ on effect-size error (§6).

What the anchors cannot buy is the ranking. On the one Arena panel deep enough to resolve it (11 judges, 357 battles per system, $\beta = 0.625$), 32% of pairs carry a post-calibration residual difference large enough to reverse their true order, against a 24% null ($p = 0.02$); at our headline density nothing separates from noise. That residual is not the bias the taxonomies describe: surface features (length, markdown, hedging, refusals) predict it with leave-one-out $R^2 < 0$, and it transfers weakly at best between disjoint panels ($r = 0.15$), so gold spent measuring it under one panel does not carry.

We are deliberate about what is old: the decomposition is the Social Relations Model (Kenny & La Voie, 1984), non-identifiability of shared bias without gold is classical (Cronbach, 1955; Kenny & Albright, 1987; Della Penna & Reid, 2012) and was restated for LLM judges by Vossler et al. (2026), and attenuation is Rogan & Gladen (1978). We claim the measurement and the budget rule, not the algebra.

Contributions. (1) The dose-response measurement: β spans $[0, 1.8]$ as a function of panel composition alone, with the mechanism identified (§5). (2) Proposition 2, a closed-form rule for when anchor rescaling helps, verified on three panels and shown to have non-positive regret when fed a two-anchor estimate (§3, §6). (3) A head-to-head against PPI++ on real human votes: on a miscalibrated panel, 32 anchor labels (64 with the debiased estimator) beat 512 spent uniformly on effect-size error, with bootstrap CIs that exclude a tie (§6). (4) An audit of the residual — flip risk, statistical status, unpredictability, non-transfer — and three negative results (§7, Appendix B).

2 Related work

Classical measurement. Everything structural in §3 predates LLMs. Cronbach (1955) split rating accuracy into four terms that the Social Relations Model (Kenny & La Voie, 1984) re-derives; Kenny & Albright (1987) warn that rater consensus does not imply accuracy when raters share stereotypes, and a panel of LLM judges is a monomethod block (Campbell & Fiske, 1959). The psychometric repair is anchoring (Kopf et al., 2015) and the epidemiological name for our slope is Rogan & Gladen (1978) attenuation. Intercept–slope–residual against a reference is the classical method-comparison problem, whose standard tools, Bland–Altman (Bland & Altman, 1986) and Deming over ordinary regression (Deming, 1943; Cornbleet & Gochman, 1979), exist because the reference is measured with error (Appendices G, D). In crowdsourcing Della Penna & Reid (2012) prove shared labeller bias is unidentifiable without gold, as Albert & Dodd (2004) document for diagnostic tests; latent-class identifiability (Allman et al., 2009; Dawid & Skene, 1979) turns on the conditional independence shared bias violates, and Prelec et al. (2017) show escaping it needs extra elicited information, not more raters. Generalizability theory (Brennan, 2001) supplies our random-facet accounting.

Identifiability for LLM judges. Vossler et al. (2026) prove that under Strong Constancy (judge confusion matrices identical across systems; in our notation β fixed and $\tilde{c} \equiv 0$) rankings are identifiable from two or more judges, and that true scores are not without gold "for any number of judges"; they parameterise violations by a prior magnitude. We measure the violation and split it into the part that cannot reorder (β) and the part that can ($\tilde{c}$); Xu et al. (2026a) give the item-level analogue. Gold-free aggregators make the aliasing concrete: CyclicJudge (Zhu et al., 2026) absorbs panel-wide shared bias into its estimand, judge-aware ranking (Xu et al., 2026b; Yu et al., 2026) recovers scores only up to affine transformations of a consensus that contains it, and CARE (Zhao et al., 2026) resolves the resulting label ambiguity heuristically. Panel practice descends from Verga et al. (2024), whose diversity rationale Acharya et al. (2026) and Balasubramanian et al. (2026) weaken only for independent or item-level contamination, not the system-level component we audit.

Agreement is not validity. Large-scale audits document high reliability with invalidity (Norman et al., 2026), judge-error correlations up to 0.97 within a single judge family and 0.66–0.71 across families (Yang et al., 2026), nine judges collapsing to two effective votes (Kohli, 2026), consensus reflecting shared bias (Jain et al., 2025; Song et al., 2026), and measurement-theoretic critiques of agreement coefficients (Rao & Callison-Burch, 2026; Chehbouni et al., 2025; Amidei et al., 2019); §3 explains why agreement cannot signal validity *in either direction*. Bias taxonomies (Ye et al., 2024; Feuer et al., 2024; Panickssery et al., 2024; Li et al., 2026; Tan et al., 2025) catalogue mechanisms that could generate $\tilde{c}$, but their surface features do not explain what survives calibration (§7).

Buying gold. Lee et al. (2025) apply Rogan–Gladen correction for one judge grading one system; our β is the panel-level, cross-system version, with a preference-task analogue where no confusion-matrix story applies. Feng et al. (2026) certify claims from gold-estimated judge error rates, Fiedler (2026) documents wrong-signed single-judge estimates, and Chen et al. (2026) give semiparametric-efficient estimators unifying measurement error with prediction-powered inference (Angelopoulos et al., 2023a;b); Divekar (2026), Chatzi et al. (2024) and Boyeau et al. (2025) buy gold subsamples to correct rankings from model-labelled data, the closest existing analogue of our §6 comparison. All allocate labels across systems; our contribution is that the allocation is the decision. A parallel literature budgets over *items* (Maia Polo et al., 2024; Vivek et al., 2024; Perlitz et al., 2024), choosing which examples to run so a ranking survives; it composes with ours, choosing the battles our anchors are labelled on, and cannot identify β, since subsetting items never separates θ from $\tilde{c}$. Length-controlled AlpacaEval (Dubois et al., 2024) is the closest deployed analogue of buying the unit, regressing out a style covariate and reporting the adjusted effect size, where we show the residual

scale is a property of the *roster*, not of any covariate. Arena publishes Bradley–Terry ratings, on which a roster-dependent scale is a discrimination parameter; we use win rates because that is where the affine split is exact, and replicate the finding on the BT scale (§5, Appendix M). Conformal Elo (Kargi & Salinas, 2026) calibrates judge-vs-human residuals over held-out models assuming exchangeability; their own diagnostic (residuals correlating $+0.49$ to $+0.90$ with human Elo) is our $\beta \neq 1$. Neuhof & Benjamini (2026) and Yauney et al. (2026) quantify ranking uncertainty from item sampling alone; our variance floor adds the judge-family facet that item resampling cannot see.

3 FRAMEWORK

Items $x \sim P$ are evaluated for systems $s \in \{1, \dots, M\}$ by judges $l \in \{1, \dots, L\}$ in model families $\phi(l) \in \{1, \dots, F\}$. The gold functional is $\theta_s = \mathbb{E}[Y_s]$; judge l's score is $\mu_{ls} = \mathbb{E}[Z_l^s]$, with bias $b_{ls} = \mu_{ls} - \theta_s$ and panel mean $\bar{\mu}_s = \frac{1}{L}\sum_l \mu_{ls}$.

Definition 1 (Bias decomposition; Social Relations Model)**.** *$b_{ls} = g + a_l + c_s + d_{ls}$ with sum-to-zero identification over l and s: g grand bias, a_l judge leniency, c_s common-mode per-system bias, d_{ls} the judge×system interaction.*

Lemma 1 (Aliasing; classical)**.** *From $\{\mu_{ls}\}$ alone, a_l and d_{ls} are identified and $\theta_s + c_s$ is identified up to a constant, but θ_s and c_s are not separately identified. Every aggregator that depends on the verdicts only through the score matrix is therefore a function of $\theta + c$, and is ranking-consistent for every θ iff c is constant; item-level aggregators escape only through assumptions (conditional independence of judges given the item) that shared bias violates by construction (Allman et al., 2009; Della Penna & Reid, 2012).*

Proposition 1 (Affine split)**.** *Let $\bar{\mu}_s = \alpha + \beta\theta_s + \tilde{c}_s$ with $\sum_s \tilde{c}_s = \sum_s \tilde{c}_s\theta_s = 0$ and $\beta > 0$. Then **(i)** $\bar{\mu}_s - \bar{\mu}_t = \beta(\theta_s - \theta_t) + (\tilde{c}_s - \tilde{c}_t)$, so with $\tilde{c} \equiv 0$ the panel ranking equals the gold ranking for any β while every gap is scaled by β; **(ii)** if θ is known for two systems $u \neq v$ with $\theta_u \neq \theta_v$ and $\tilde{c}_u = \tilde{c}_v = 0$ then $\beta = (\bar{\mu}_u - \bar{\mu}_v)/(\theta_u - \theta_v)$ identifies the scale; **(iii)** $\tilde{c}_s$ is not identifiable from $\{\mu_{ls}\}$ without an anchor for system s itself.*

Part (ii) is the cheap fix and part (iii) is its limit. But rescaling by $1/\beta$ rescales the residual too, so it is not free:

Proposition 2 (When to rescale)**.** *Write $\sigma_\theta^2 = \mathrm{Var}_s(\theta_s)$ and $\sigma_{\tilde{c}}^2 = \mathrm{Var}_s(\tilde{c}_s)$, and note $\tilde{c} \perp \{1, \theta\}$ by construction. Dividing estimated gaps by β lowers the expected squared effect-size error iff*

$$\sigma_{\tilde{c}}^2\,(1-\beta^2)/\beta^2 \;<\; (1-\beta)^2\sigma_\theta^2, \tag{2}$$

which holds for every $\beta > 1$ (at $\beta = 1$ rescaling is the identity and the error is unchanged), and for $\beta < 1$ reduces to $\sigma_{\tilde{c}}^2/\sigma_\theta^2 < \beta^2(1-\beta)/(1+\beta)$. The right-hand side is maximised at $\beta \approx 0.618$ with value 0.0902.

Minimising over *all* constants would give $\gamma^\star = \beta\sigma_\theta^2/(\beta^2\sigma_\theta^2 + \sigma_{\tilde{c}}^2)$, weakly dominating both options but needing $\sigma_{\tilde{c}}^2$, exactly what Proposition 1(iii) says two anchors cannot supply. The second branch needs the same quantity, so its scope is narrow: *with two anchors the rule is exactly "rescale iff $\hat{\beta} \geq 1$"*; at $K \geq 8$ anchored systems the OLS residual about the fitted line estimates $\sigma_{\tilde{c}}^2$ and the anchor spread σ_θ^2, both debiasable by their own sampling variance, with the cautions of Appendix J. So: *two anchors buy the scale, eight buy the decision.*

Why agreement cannot substitute for gold. The ambiguity decomposition (Krogh & Vedelsby, 1995; Wood et al., 2023) gives $\bar{E} = E_{\text{panel}} + A$: observable disagreement A lower-bounds *member* error and bounds nothing about panel error, and the sign of $\mathrm{corr}(A, E_{\text{panel}})$ is forced by which term varies more: we measure -0.72 in one setting and $+0.60$ in another (cf. Kuncheva & Whitaker, 2003; Rao & Callison-Burch, 2026). High agreement means members are similar, not right.

Inference. Judge errors correlate within model families (Yang et al., 2026; Kohli, 2026), so we treat the family as the exchangeable unit: intervals come from a hierarchical bootstrap over items, judges within family, then families, and $\mathrm{Var}(\bar{\mu}_s) = \sigma_{\text{item}}^2/n + \sigma_{\text{within}}^2/L + \sigma_{\text{fam}}^2/F$ is floored by the family term (Brennan, 2001). Residual statistics are reported against a *matched null* in which $\tilde{c} \equiv 0$ and only sampling noise with the estimated per-system covariance is present, projected off $\mathrm{span}\{1, \theta\}$ exactly as the estimator does. Without this, $\mathrm{RMS}(\tilde{c})$ looks large when it is merely noisy (Appendix B).

Table 1: Panel-vs-gold slope by composition; every row scores *identical* items under an identical protocol. β is the OLS slope against gold *as reported*, the estimand throughout; β_{dec} recomputes both scales on human-decided battles with judge ties dropped (undefined on MATH, whose exact-match gold has no ties). CIs are the hierarchical bootstrap of §3; sub-panel ranges are 5th–95th percentiles. Correcting for regressor noise targets the latent-preference slope instead and multiplies a whole column by one constant (Appendix D).

	MT-Bench (M=6)		MATH (M=8)	Arena (M=22)	
composition	β [95% CI]	β_{dec}	β [95% CI]	β [95% CI]	β_{dec}
open-weight only	0.414 [0.21, 0.62]	0.431	0.134 [-0.01, 0.31]	0.452 [0.18, 0.76]	0.407
mixed cheap+open	0.872 [0.56, 1.08]	0.832	0.443 [0.24, 0.64]	0.996 [0.62, 1.30]	0.842
commercial API only	1.205 [1.06, 1.34]	1.114	0.633 [0.44, 0.84]	1.480 [1.23, 1.71]	1.246
frontier only	**1.265** [1.11, 1.41]	1.139	**0.707** [0.50, 0.90]	**1.740** [1.53, 1.93]	1.463
over random sub-panels of every size (5th–95th percentile)					
β range	[0.75, 1.15]		[0.37, 0.63]	[0.88, 1.45]	
corr(β, member acc.)	+0.94		+0.45	+0.89	

4 Experimental design

Settings. Three settings, chosen so that estimand, gold type and leaderboard density all vary; Table 3 in Appendix S gives the full summary. **(A) MT-Bench**: pairwise battles between M=6 systems with 3.3k human votes as gold (Zheng et al., 2023). **(B) MATH grading**: solutions to 84 MATH-500 problems from M=8 commercial systems graded against exact answer match, so θ_s is noiseless *per item* though it still carries item-sampling variance. **(C) Chatbot Arena**: human-voted battles over the M=22 most-battled systems of the public 55k release (Chiang et al., 2024), stratified over system pairs; the snapshot is the public 2024 release, a vintage choice §8 returns to. A and C share an estimand but differ 2.8× in density, which is where the residual bites.

The panel. We enrol judges in three tiers: *cheap API* (13 judges, 5 families: OpenAI, Google, DeepSeek, Alibaba, ByteDance), *open-weight* run locally (8 judges, 8 families: Phi-4-mini, Llama-3.2-3B, Mistral-7B, Falcon3-3B, Granite-3.3-2B, OLMo-2-1B, StableLM-2, Yi-1.5-6B), and *frontier* (5 judges, 3 families: GPT-5.4, GPT-5.1, Claude-Sonnet-4.6, Claude-Opus-4.8, Qwen-Max). All judges see the same MT-Bench-style pairwise prompt with hash-balanced presentation order (measured balance 47.8–51.5%) or the same grading prompt, at temperature 0. Composition sweeps partition the roster; Appendix T has the full list.

The study comprises 99,487 verdicts: 43,366 in complete-cell blocks, 35,723 in the deeper slice of §7, the rest outside the completeness filters (Appendix T reconciles this).

Hygiene, fixed in advance. A judge is dropped, never re-prompted, if format compliance falls below 60% or item coverage below 90%, except that a family's last judge is kept, since a lost family is a lost degree of freedom in the variance floor above. On Arena both Google judges fall below the floor (859 and 886 of 1,600), so that clause retains one and costs 45% of the complete cells; dropping the family instead leaves every conclusion intact (Appendix R). Every verdict is released, so every number reproduces from the shipped panels at zero API cost (the study cost $33.3 in API fees plus GPU hours).

5 The slope is a property of the panel

The slope is large and its direction is set by the panel. The commercial-API-only panel, the composition almost every published evaluation uses, is *over-decisive* on preference tasks: every reported Arena gap is inflated by 48% ($\beta = 1.480$) and every MT-Bench gap by 21%, while the same panel on verifiable grading *attenuates* ($\beta = 0.633$). Widening to 13 families moves β down in all three settings; an open-weight-only panel compresses gaps by 55–87%. Over 2,520–3,000 random sub-panels per setting, β correlates +0.89 (Arena), +0.94 (MT-Bench) and +0.45 (MATH) with the panel's *mean member accuracy* $\frac{1}{L}\sum_l(1 - \mathbb{E}[(Z_l - Y)^2])$, one minus each judge's mean squared deviation from gold, ties counting as a half-unit.

Controls for the three artefacts that could explain this. First, gold is itself estimated from finitely many votes, and noise in a regressor attenuates OLS; correcting for it (Appendix D) multiplies a whole column by one constant and so changes no comparison. That correction assumes gold noise is independent of panel noise, which it is not, since judges and the human voter score the same battle. We therefore also report the assumption-free version: panel on one half of the battles, gold on the *disjoint* other half. That gives $\beta_{\text{split}} = 1.05$, 0.70 and 0.31 for the commercial, mixed and open-weight Arena panels. That looks like a refutation, but halving the gold doubles its sampling noise and OLS attenuates by the regressor's reliability: the estimates form a ladder predicted with no free parameter, $\beta_{\text{split}}/\beta = (1-q')/(1-q)$, $q' = 2q/(1+q)$, matching measurement to a median 2.2% over all twelve cells, the same slope through a noisier regressor, and the ordering is preserved exactly ($\text{corr}(\beta_{\text{split}}, \beta) = +1.000$; Appendix H). Second, humans tie far more than judges, and scoring a tie as half a win compresses gold toward 0.5; recomputing both scales on human-decided battles with judge ties dropped shrinks every slope (Arena's commercial panel to 1.246, so about half the excess above one is convention) but leaves the roster dependence intact, β_{dec} spanning 0.407 to 1.463 (Appendix F). The tier separation does not rest on intervals — on MATH the open-weight and mixed intervals overlap, and resampling three frontier families cannot give a 95% interval — but on an *exact* permutation over the two tiers' 11 families: the observed split is the most extreme of all $\binom{11}{3} = 165$ assignments in all three settings, and no single family drives it (Appendix L). Third, Arena publishes Bradley–Terry ratings, not win rates, and a linear fit reaching 1.74 on a bounded scale must eventually break. Refitting both sides as BT models by MM (Bradley & Terry, 1952; Hunter, 2004) reproduces the ordering ($r = +0.993$) and *widens* the spread to $5.06\times$ against $3.85\times$. In published Elo the 22 systems have a standard deviation of 39 points under an open-weight roster and 151 under a frontier one, against 61 for the humans (Appendix M).

Mechanism: decisiveness inflates, noise attenuates. Two forces set β and our data separate them. Refusing the middle ground stretches the win-rate scale and pushes β up; verdict noise pushes it down, the panel analogue of Rogan–Gladen attenuation, which in the verifiable regime would set β to the panel mean of $\text{TPR}_l - \text{FPR}_l$. The two wings make the tension visible: commercial judges are decisive *and* accurate so decisiveness wins; open-weight judges tie least of all (4.7%) yet have the lowest β, because they are much noisier; the mixed panel lands near $\beta = 1$ on Arena because the two wings cancel, not because breadth is magic.

Buying better judges makes it worse on preferences, and better on grading. The obvious objection is that we used inexpensive judges and a frontier panel would simply be calibrated. It is not, and the mechanism predicts *opposite* directions in the two regimes, the sharpest test we can put it to. On preferences the frontier panel has the highest mean member accuracy of any tier (0.752 on Arena) and the *worst* calibration, $\beta = 1.740$ against 1.480 and 0.452, with MT-Bench replicating the ordering ($0.414 < 1.205 < 1.265$). On verifiable grading there are no ties to refuse and a better judge must move β *towards* one. It does, monotonically: $0.134 \to 0.633 \to 0.707$. A stronger panel therefore does not calibrate the scale; it moves β further in whichever direction the protocol already points. MATH gold is an exact answer match, so TPR_l and FPR_l are observable and the Rogan–Gladen identity can be checked rather than invoked. It is close but not exact (panel means 0.111, 0.490, 0.503 against β of 0.134, 0.633, 0.707), and since the identity needs the Strong Constancy of Vossler et al. (2026), that gap *is* a measurement of the assumption's failure rather than the prior magnitude they place on it, and it widens with capability ($+0.02$, $+0.14$, $+0.20$).

Effect sizes have a missing unit. A reported 4-point Arena win-rate gap is a true gap of $4/1.48 = 2.7$ points under the commercial panel and $4/0.45 = 8.8$ under the open-weight one: a $3.3\times$ ratio between two reasonable rosters on the same benchmark with the same prompts. Pooling judged effect sizes across papers pools numbers with different units, and any consumer calibrated in gap units (reward thresholds, router margins) inherits an invisible rescaling whenever the roster changes; a single-panel study cannot tell this panel property from a task property (Appendix B).

6 Spending a gold budget

Proposition 1(ii) says two gold-evaluated systems identify the scale; the gold-purchasing literature instead spreads labels over systems and rectifies each with its own subsample (Angelopoulos et al., 2023a; Chatzi et al., 2024; Divekar, 2026). We run the head-to-head on real human votes: hide the gold, hand each strategy n labelled battles, and score the recovered gold vector against the

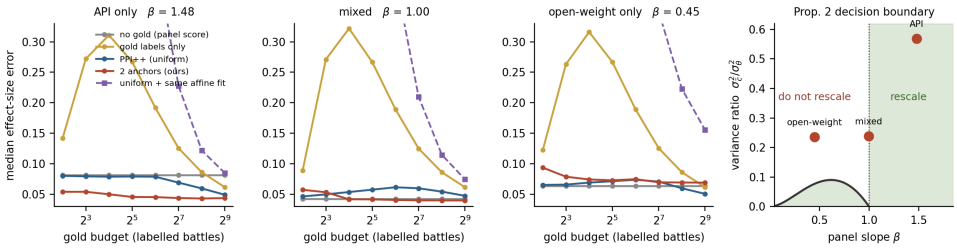


Figure 2: Where a gold budget should go (Arena, M=22, gold = human votes, 120 draws per budget). Anchor spending dominates on the miscalibrated commercial panel at every budget, is neutral on the mixed panel, and *hurts* the weak open-weight panel. Right: the Proposition 2 boundary with the three measured panels on it; both axis and curve are in *variance* ratios, as in equation 2.

full-sample gold vector. PPI++ (Angelopoulos et al., 2023b) gets its tuned estimator and shrinkage toward the panel prior for systems with few labels, so it is not a strawman at small n, and the shrinkage is not a handicap at large n (Appendix C).

A handful of anchor labels does the work of hundreds spent uniformly, for effect sizes and only for effect sizes. Rescaling cannot reorder: pair accuracy stays at the no-gold 0.8225 at every budget, while PPI++ reaches 0.8355 at $n = 512$ (Appendix O). On the commercial panel ($\beta = 1.48$), median effect-size error is 0.0819 with no gold and 0.0803 under PPI++ at $n = 4$; two anchors reach 0.0542 at $n = 4$ and 0.0457 at $n = 32$, better than PPI++ achieves at $n = 512$ (0.0493) — and that crossing is not noise: the bootstrap 95% CI for the n=32-vs-n=512 difference is $[-0.0042, -0.0005]$. Anchors win at *every* budget tested (Figure 2, left). The frontier panel, which the rule most wants to rescale ($\beta = 1.740$), behaves the same way: two anchors cut the error from 0.104 with no gold to 0.047 at $n = 32$, against 0.095 for PPI++ at the same budget, though there the crossing against PPI++ at $n = 512$ is within noise. The reason is structural: the scale is two parameters shared by all M systems, while uniform allocation buys M rectifiers from n/M labels each. **Allocation, not the estimator, does the work**, separated as a 2×2 of {anchors, uniform} $\times$ {widest pair, median over pairs} at matched budget on all three panels (Appendix I). Changing the *allocation* costs 7.0–10.3$\times$ at $n = 16$ and 1.7–1.9$\times$ at $n = 512$; changing the *estimator* costs 1.1–1.3$\times$ at $n = 512$, and within the anchor row the two differ by at most 0.006 past $n = 64$; regressing on 22 noisy $\hat{\theta}_s$ re-introduces the attenuation concentration avoids and fails outright far more often (46% degenerate draws at $n = 16$ against 7%). Against PPI++, a far better uniform estimator, the honest allocation margin is $1.7\times$ at $n = 16$ and $1.1\times$ at $n = 512$; what survives is the *shape*: anchors reach their floor by $n = 32$ where PPI++ needs 512 to approach it. Routing labels adaptively instead, to the least-resolved adjacent pair, helps on neither currency: it costs on gap error at every budget and ties uniform on pair accuracy, because one crowd vote per battle leaves an undecided pair irreducibly rather than resolvably noisy (Appendix I). Figure 2 uses the widest pair, needing two anchored systems; the median over pairs of §8 needs four, removes a $+0.22$ selection bias large enough to flip the *decision* (Appendix O), and crosses PPI++'s $n = 512$ error at $n = 64$ rather than 32 (CI $[-0.0044, -0.0016]$; Appendix I). We lean on $n = 32$ because the anchor denominator is degenerate in 85% of draws at $n = 4$ but 2% at $n = 32$.

And the rule says when this transfers. On the mixed panel anchoring is neutral, and on the open-weight panel ($\beta = 0.452$) it is actively harmful: 0.0636 with no gold versus 0.0815 after two-anchor rescaling, and even the oracle affine correction is worse than nothing (0.0962), and PPI++, which does not rescale, draws level with anchors at small budgets and beats them outright from $n = 256$ on (0.051 vs 0.070 at $n = 512$). This is exactly Proposition 2: dividing by 0.452 amplifies the residual $2.2\times$, and $\sigma_{\tilde{c}}^2/\sigma_{\theta}^2 = 0.24$ far exceeds the threshold 0.077 (Figure 2, right).

What the rule is worth, and whether it survives estimation. All three observed panels sit at $\beta \geq 1$ or far above the threshold, so the data alone exercise only the branch that, since the bound never exceeds 0.0902, collapses to *rescale iff* $\hat{\beta} \geq 1$. We probe the other branch on semi-synthetic panels built from the real gold vector and residual *shape*; since $\tilde{c}$ is orthogonal to θ by construction, agreement there is an implementation check, not evidence, and what the sweep buys is the branch's

Table 2: The rank-relevant residual against a matched null for *all three* statistics: $\tilde{c} \equiv 0$, sampling noise only, projected off $\mathrm{span}\{1, \theta\}$ as the estimator does, resampling items, judges within family, families, and gold on the same item weights. A naive item-only null is 8–32% tighter and reports $p = 2 \cdot 10^{-4}$ for flip risk on the deep row where this one reports 0.021 (Appendix B). Flip risk is the share of pairs with $|\tilde{c}_s - \tilde{c}_t| > \beta|\theta_s - \theta_t|$; *wrong* is the share the panel actually misorders. Only the deep row separates from noise, and only on the first two statistics.

	L/F	btl/sys	$\mathrm{RMS}(\tilde{c})$	null	p	flip	null	p	wrong	null	p
MT-Bench	24/14	130	0.0294	0.0196	0.10	6.7%	4.4%	0.65	6.7%	2.2%	0.33
MATH-500	26/14	76	0.0090	0.0094	0.52	10.7%	11.8%	0.68	0.0%	5.8%	1.00
Arena	22/14	79	0.0440	0.0430	0.43	32.2%	29.3%	0.28	16.1%	14.6%	0.33
same 11-judge / 9-family panel ($\beta = 0.625$), only battles per system change											
Arena, shallow	11/9	143	0.0331	0.0284	0.15	37.7%	33.3%	0.19	20.3%	16.6%	0.14
Arena, deep	11/9	**357**	0.0248	**0.0191**	**0.039**	**32.0%**	**23.5%**	**0.021**	13.9%	11.8%	0.24

price. In all 18 cells where Proposition 2 and the $\hat{\beta} \geq 1$ heuristic disagree the heuristic is the one that is wrong, by a median 218% and up to 30× in squared gap error. The rule run on $K = 8$ *estimated* anchors, which is what a practitioner would have, recovers the right decision in 79% of draws over those cells against the heuristic's 0%, but wrongly rescales in 46% of draws over the five cells where the heuristic is right: a usable, noisy instrument (Appendix J). Fed an *estimated* $\hat{\beta}$ from two anchors, the rule's regret against an oracle is ≤ 0 on every panel at every budget $n \in \{32, 128, 512\}$, in the *mean* over replicates rather than the median, which would be identically zero whenever the decision is right more than half the time. On the mixed panel $\beta = 0.996$ makes the rule "wrong" 98.0% of the time at $n = 512$ and yet *lowers* mean realised error by 0.0024: the decision is hardest exactly where it is worth least. The conservative variant, rescaling only if a bootstrap lower bound clears one, is the harmful one, costing $+0.0124$ on the commercial panel at $n = 32$ (Appendix N). Over-decisive panels are cheap to repair; weak panels are not repairable by rescaling at all.

Two further results, in Appendix P. Removing the slope before conformalising halves leave-one-system-out interval width ($0.300 \to 0.151$) at 0.86 coverage against a 0.90 target; and reweighting judges to $\beta(w) = 1$, a linear constraint unlike rescaling, does not beat it (0.0746 vs 0.0451 median gap error, winning on 20.3% of pairs).

7 THE RESIDUAL: WHAT NO BUDGET BUYS

After the best affine calibration what remains is $\tilde{c}_s$: the only part of common-mode bias that can reorder a leaderboard and, by Proposition 1(iii), the part no gold-free method identifies. Four measured properties.

(R1) Establishing it needs depth, and it is significant but not overwhelmingly so. At our headline panels' per-system sampling the residual is not separable from noise: against a family-inclusive null, $\mathrm{RMS}(\tilde{c})$ gives $p = 0.10, 0.52, 0.43$ on MT-Bench, MATH, Arena. A variance decomposition puts 62% of the residual's sampling variance on the item facet, so we collected a disjoint slice of 2,400 further Arena battles on the same 22 systems and re-ran on the 11 judges (9 families) scoring both, composition fixed, battles per system rising 143 to 357. The noise floor falls $0.0284 \to 0.0191$ while the noise-debiased residual is stable ($0.0164 \to 0.0155$) and p moves $0.15 \to \mathbf{0.039}$: noise shrinks with depth, the residual does not. But eight of those eleven judges are small open-weight models, at $\beta = 0.625$, so we ran six commercial judges over both slices too, same design and null, 368 battles per system: there the residual does *not* clear the null ($p = 0.077$ on $\mathrm{RMS}(\tilde{c})$, 0.12 on flip risk). It is bounded rather than established there, and the two rosters compare only with the scale divided out, since β stretches the residual and its own noise floor together: in *gold* units the debiased residual is 0.0147, 18% of the median gold gap, against 0.0565 and 71% open-weight (Appendix Q). A decisive panel carries the *smaller* rank-relevant distortion once its unit is corrected, and is the harder place to establish one.

(R2) It is large relative to reported gaps. On the deep panel the residual leaves 32.0% of pairs reversible against a 23.5% null ($p = 0.021$), while the 13.9% it *actually* misorders does not clear its own null (11.8%, $p = 0.24$). Across settings $\mathrm{RMS}(\tilde{c})$ stays in 0.009–0.044 while the median

panel-score gap falls from 0.277 (M=6) to 0.099 (M=22), so the same residual reaches far more comparisons, and public leaderboards are denser still. A $\max - \min$ spread ratio would overstate this, since its growth in M is partly mechanical; flip risk is not (Appendix B).

(R3) It is not the bias the taxonomies describe. Eight gold-free system features (length, markdown, bullets, code, refusals, hedging, exclamations, kinship) give in-sample $R^2 = 0.45$ but *negative* leave-one-out R^2 for every model tried, permutation $p = 0.17$; and 0.45 is about what 8 features over 22 systems buy from noise, so this excludes *strong* predictability, not moderate (Appendix E). Style biases (Ye et al., 2024; Feuer et al., 2024) are real, but load on the *slope*, which calibration absorbs, so a bias taxonomy and an unpredictable residual are not in conflict.

(R4) It does not transfer across panels, and provider kinship is why, in part. The 22-system $\tilde{c}$ vectors of two *disjoint* panels, commercial-only versus open-weight-only on the same items and battles, correlate at only $r = 0.15$ (95% CI $[-0.24, 0.48]$, bounding transfer loosely rather than excluding it), against $r = 0.89$ between nested panels. So $\tilde{c}$ is a property of the (system, panel) pair, not of the system: gold spent measuring it under one panel does not correct the next. The frontier tier names one ingredient. Adding it puts *Anthropic* judges in the panel for the first time and adds further *OpenAI* ones, so the nine Arena systems from those providers gain kinship and the other thirteen (whose Meta- and Mistral-derived kinship sits in the open-weight tier, unchanged) do not. That is a controlled, within-system, between-panel contrast on identical items, and it moves exactly the systems it should. *All nine* gain rank-relevant credit; the contrast against the other thirteen is $+0.0156$, or 18.3% of the median *panel-score* gap, permutation $p = 0.013$ (the label is permuted because $\tilde{c}$ sums to zero; the sign test alone gives $p = 0.023$). Kin systems are also stronger ($\mathrm{corr}(\mathrm{kin}, \theta) = +0.53$), and a regression control for θ is empty here: the contrast is a difference of affine residuals, orthogonal to gold by construction, so adding θ only rescales the kinship coefficient by $1/(1 - r^2)$, to $+0.022$. The control that carries weight is non-parametric, permuting kinship within gold-adjacent pairs: $p = 0.031$. MT-Bench agrees (all three kin systems gain, contrast 8.9%) though at M=6 the test has no power, and on verifiable grading, where there is no style to prefer, the effect vanishes (1.6%, $p = 0.48$). This does not contradict (R3): predicting $\tilde{c}$ from a static feature fails where intervening on the composition succeeds, and it sharpens the problem, because the one mechanism we can identify is one that adding judges cannot remove.

8 Discussion

The auditor's protocol. Four practices, costing two to eight gold evaluations and no new judging. *Report the roster*: a judged number has no unit without it. *Buy the scale first*: two anchored systems give $\hat{\beta}_{uv} = (\bar{\mu}_u - \bar{\mu}_v)/(\theta_u - \theta_v)$ and the decision "rescale by $1/\hat{\beta}$ iff $\hat{\beta} \geq 1$" (Proposition 2), but the widest pair selects on a positive residual contrast ($+0.22$ here, Appendix O), so pre-specify the pair or, with four, take the *median* $\hat{\beta}_{uv}$ over well-separated pairs; eight buy the full rule, because only then is $\sigma^2_{\tilde{c}}/\sigma^2_{\theta}$ estimable (Appendix J). *Use the family as the exchangeable unit* in nulls as well as intervals, since family-level intervals are 25–32% wider here and item-only resampling (Neuhof & Benjamini, 2026; Yauney et al., 2026) misses the dependence entirely. *Abstain below the residual floor*: report a pair as ordered only if its panel gap $|\bar{\mu}_s - \bar{\mu}_t|$ exceeds $2\sqrt{2}\,\hat{\sigma}_{\tilde{c}}$ (two standard deviations of a residual difference; $\hat{\sigma}_{\tilde{c}}$ from $K \geq 8$ anchors or a deep slice): 0.044 on our deep panel, which 95% of adjacent and 41% of all pairs fail. Two anchors restore the unit, but no gold-free procedure restores the ranking.

Which slope is *the* slope. β is the slope against gold *as reported*; dividing by the gold's reliability gives the latent slope (Arena commercial panel: $1.48 \to 2.20$, Appendix H), and every cross-roster comparison is invariant to the choice. **On a live leaderboard** the protocol is prospective, the anchors gold-evaluated once; since $\tilde{c}$ does not transfer (R4), the scale is re-bought at every roster change.

Limitations. The frontier tier spans three families; $M = 22$ is dense by audit standards and sparse by leaderboard standards; Arena gold is single crowd votes; all data are English, under one pairwise and one grading protocol. The Arena snapshot is the public 2024 release, so the audited systems predate the judges. The kinship contrast is nine kin against thirteen non-kin; the residual is established on a deepened open-weight-leaning panel and bounded, not established, on the commercial one, whose three families also limit its null; and a quadratic term takes 20% of $\mathrm{Var}(\tilde{c})$ on two Arena rosters (Appendix R).

AI USE STATEMENT

In this work, we used generative AI tools to assist in writing, editing, and debugging code, including code used to implement our methods and experiments. We did not use generative AI tools for any of the other tasks that require disclosure: in particular, we did not use them to generate synthetic data, to develop theoretical models or conceptual frameworks, to formulate mathematical claims or assist in proofs, to propose or refine hypotheses, to design the research methodology or experiments, to assist with translation, to clean or reformat datasets, or to support data analysis or the interpretation of results. The generative-model outputs that make up the measurements here (the graded responses in the verifiable setting and every judge verdict) are the object of study rather than data generated in place of measurement, and the semi-synthetic panels of Appendix N are built by the released analysis code from the measured gold vector and the measured residual shape. Additionally, we used generative AI tools to edit the manuscript to improve readability and clarity, and to obtain simulated peer-review-style feedback on drafts, which informed revisions to the writing and presentation. All AI-assisted code was reviewed and tested for correctness by the authors, and all AI-suggested edits and feedback were reviewed and adopted at the authors' discretion; the scientific content, claims, and results are entirely the authors' own. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

All code, prompts, judge verdicts, panel scores and analysis outputs are released (anonymised archive in the supplementary material). Every quantitative claim in this paper was checked against the released code and result files by a scan that recomputes or string-checks each number and asserts it appears in the compiled PDF (`scripts/consistency_scan.py`, 331 checks); each proof was checked by hand; every reference for which we hold a source PDF was checked against it mechanically by `scripts/32_verify_bibliography.py` (34 of 64 entries, forward on all 34 and in reverse where the source lists one e-mail per author; the remainder were checked against the publisher record); and every figure is rendered by the released matplotlib scripts from the released result files. Every analysis reproduces from the shipped panels at zero API cost; a response cache covering the MT-Bench and MATH calls also ships, and re-judging the Arena slices from scratch costs money (the study cost $33.3 at public list prices, judges absent from the price table being charged at a deliberately high default, plus GPU hours for the open-weight tier). Seeds fix item sampling, presentation-order balancing, sub-panel draws and all bootstraps. Appendix U lists the commands.

ETHICS STATEMENT

This work audits evaluation methodology using only public datasets (MT-Bench with its released human votes, MATH-500, and the public Chatbot Arena 55k release) under their licenses; no new human data was collected. Findings that particular systems are over- or under-credited concern historical snapshots and are not claims about current products.

REFERENCES

Anish Acharya, Kris W. Pan, and Brian Verkhovsky. RoPoLL: robust panel of LLM judges. *arXiv preprint arXiv:2606.30931*, 2026.

Paul S. Albert and Lori E. Dodd. A cautionary note on the robustness of latent class models for estimating diagnostic error without a gold standard. *Biometrics*, 60(2):427–435, 2004.

Elizabeth S. Allman, Catherine Matias, and John A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *The Annals of Statistics*, 37(6A):3099–3132, 2009.

Jacopo Amidei, Paul Piwek, and Alistair Willis. Agreement is overrated: a plea for correlation to assess human evaluation reliability. In *Proceedings of the 12th International Conference on Natural Language Generation (INLG)*, 2019.

Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023a.

Anastasios N. Angelopoulos, John C. Duchi, and Tijana Zrnic. PPI++: efficient prediction-powered inference. *arXiv preprint arXiv:2311.01453*, 2023b.

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

Rina Foygel Barber, Emmanuel J. Candès, Aaditya Ramdas, and Ryan J. Tibshirani. Predictive inference with the jackknife+. *The Annals of Statistics*, 49(1):486–507, 2021.

J. Martin Bland and Douglas G. Altman. Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet*, 327(8476):307–310, 1986.

Pierre Boyeau, Anastasios Nikolas Angelopoulos, Tianle Li, Nir Yosef, Jitendra Malik, and Michael I. Jordan. AutoEval done right: Using synthetic data for model evaluation. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pp. 5276–5290. PMLR, 2025.

Ralph Allan Bradley and Milton E. Terry. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345, 1952.

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

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

Ivi Chatzi, Eleni Straitouri, Suhas Thejaswi, and Manuel Gomez Rodriguez. Prediction-powered ranking of large language models. In *Advances in Neural Information Processing Systems 37*, 2024.

Khaoula Chehbouni, Mohammed Haddou, Jackie Chi Kit Cheung, and Golnoosh Farnadi. Neither valid nor reliable? Investigating the use of LLMs as judges. *arXiv preprint arXiv:2508.18076*, 2025.

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.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: an open platform for evaluating LLMs by human preference. In *International Conference on Machine Learning (ICML)*, 2024.

Carlos Cinelli and Chad Hazlett. Making sense of sensitivity: extending omitted variable bias. *Journal of the Royal Statistical Society: Series B*, 82(1):39–67, 2020.

P. J. Cornbleet and N. Gochman. Incorrect least-squares regression coefficients in method-comparison analysis. *Clinical Chemistry*, 25(3):432–438, 1979.

Lee J. Cronbach. Processes affecting scores on "understanding of others" and "assumed similarity". *Psychological Bulletin*, 52(3):177–193, 1955.

Alexander P. Dawid and Allan 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.

Nicolás Della Penna and Mark D. Reid. Crowd & prejudice: an impossibility theorem for crowd labelling without a gold standard. In *Collective Intelligence*, 2012. arXiv:1204.3511.

W. Edwards Deming. *Statistical Adjustment of Data*. John Wiley & Sons, New York, 1943.

Abhishek Divekar. Statistically reliable LLM-based ranking evaluation via prediction-powered inference. *arXiv preprint arXiv:2606.05308*, 2026.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B. Hashimoto. Length-controlled AlpacaEval: a simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*, 2024.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987.

Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: robust statistical evaluation of LLMs with imperfect judges. In *International Conference on Learning Representations (ICLR)*, 2026.

Benjamin Feuer, Micah Goldblum, Teresa Datta, Sanjana Nambiar, Raz Besaleli, Samuel Dooley, Max Cembalest, and John P. Dickerson. Style outweighs substance: failure modes of LLM judges in alignment benchmarking. *arXiv preprint arXiv:2409.15268*, 2024.

James Fiedler. Bias and uncertainty in LLM-as-a-judge estimation. *arXiv preprint arXiv:2605.06939*, 2026.

Ariel Gera, Odellia Boni, Yotam Perlitz, Roy Bar-Haim, Lilach Eden, and Asaf Yehudai. JuStRank: benchmarking LLM judges for system ranking. *arXiv preprint arXiv:2412.09569*, 2024.

David R. Hunter. Mm algorithms for generalized Bradley–Terry models. *The Annals of Statistics*, 32(1):384–406, 2004.

Suryaansh Jain, Umair Z. Ahmed, Shubham Sahai, and Ben Leong. Beyond consensus: mitigating the agreeableness bias in LLM judge evaluations. *arXiv preprint arXiv:2510.11822*, 2025.

Bora Kargi and David Salinas. From uncertain judgments to calibrated rankings: conformal Elo estimation for LLM evaluation. *arXiv preprint arXiv:2606.13221*, 2026.

David A. Kenny and Linda Albright. Accuracy in interpersonal perception: A social relations analysis. *Psychological Bulletin*, 102(3):390–402, 1987.

David A. Kenny and Lawrence La Voie. The social relations model. *Advances in Experimental Social Psychology*, 18:141–182, 1984.

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

Julia Kopf, Achim Zeileis, and Carolin Strobl. Anchor selection strategies for DIF analysis: review, assessment, and new approaches. *Educational and Psychological Measurement*, 75(1):22–56, 2015.

Anders Krogh and Jesper Vedelsby. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems 7*, pp. 231–238, 1995.

Ludmila I. Kuncheva and Christopher J. Whitaker. Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. *Machine Learning*, 51(2):181–207, 2003.

Chungpa Lee, Thomas Zeng, Jongwon Jeong, Jy yong Sohn, and Kangwook Lee. How to correctly report LLM-as-a-judge evaluations. *arXiv preprint arXiv:2511.21140*, 2025.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. Preference leakage: a contamination problem in LLM-as-a-judge. In *International Conference on Learning Representations (ICLR)*, 2026.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235, pp. 34303–34326. PMLR, 2024.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. *arXiv preprint arXiv:2607.16259*, 2026.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: a systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 37*, 2024.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking (of language models). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 2519–2536, 2024.

Dražen Prelec, H. Sebastian Seung, and John McCoy. A solution to the single-question crowd wisdom problem. *Nature*, 541:532–535, 2017.

Delip Rao and Chris Callison-Burch. Agreement metrics for LLM-as-judge evaluation: What to report and why. *arXiv preprint arXiv:2606.00093v2*, 2026.

Walter J. Rogan and Beth Gladen. Estimating prevalence from the results of a screening test. *American Journal of Epidemiology*, 107(1):71–76, 1978.

Mingyang Song, Mao Zheng, and Chenning Xu. Beyond the illusion of consensus: from surface heuristics to knowledge-grounded evaluation in LLM-as-a-judge. *arXiv preprint arXiv:2603.11027*, 2026.

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

Pat Verga, Sebastian Hofstätter, 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.

Rajan Vivek, Kawin Ethayarajh, Diyi Yang, and Douwe Kiela. Anchor points: Benchmarking models with much fewer examples. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 1576–1601, 2024.

Patrick Vossler, Fan Xia, Yifan Mai, Adarsh Subbaswamy, and Jean Feng. LLMs judging LLMs: a simplex perspective. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2026. arXiv:2505.21972.

Danny Wood, Tingting Mu, Andrew M. Webb, Henry W. J. Reeve, Mikel Luján, and Gavin Brown. A unified theory of diversity in ensemble learning. *Journal of Machine Learning Research*, 24 (359):1–49, 2023.

Jian Xu, Delu Zeng, John Paisley, and Qibin Zhao. When can you debias an LLM judge? Identifiability limits, a test, and designs for top-k ranking. *arXiv preprint arXiv:2607.02104*, 2026a.

Mingyuan Xu, Xinzi Tan, Jiawei Wu, and Doudou Zhou. A judge-aware ranking framework for evaluating large language models without ground truth. In *International Conference on Machine Learning (ICML)*, 2026b. arXiv:2601.21817.

Zongyou Yang, Yinghan Hou, and Xiaokun Yang. When the judge changes, so does the measurement: auditing LLM-as-judge reliability. *arXiv preprint arXiv:2607.08535*, 2026.

Gregory Yauney, Shahzaib Saqib Warraich, Swabha Swayamdipta, and Kyle Lo. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? Quantifying biases in LLM-as-a-judge. *arXiv preprint arXiv:2410.02736*, 2024.

Shibo Yu, Yingzhou Wang, Yan Chen, Guodong Li, and Jin-Hong Du. Heterogeneous judge-aware ranking with sensitivity, disagreement, and confidence. *arXiv preprint arXiv:2605.05073*, 2026.

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.

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 36, Datasets and Benchmarks Track*, 2023.

Ziyi Zhu, Olivier Tieleman, Alexey Bukhtiyarov, and Jinghong Chen. CyclicJudge: mitigating judge bias efficiently in LLM-based evaluation. *arXiv preprint arXiv:2603.01865*, 2026.

A PROOFS

Lemma 1. Fix κ and set $\theta'_s = \theta_s + c_s - \kappa$, $g' = g + \kappa$, $c'_s = 0$. Then $b'_{ls} = \mu_{ls} - \theta'_s = g' + a_l + d_{ls}$ reproduces the identical observed $\{\mu_{ls}\}$ with no common-mode bias, and both parameterisations satisfy Definition 1, including $\sum_s c'_s = 0$; hence only $\theta_s + c_s$ is identified, up to the constant. a_l and d_{ls} are contrasts of μ_{ls} across l within s, hence identified. An aggregator that sees the verdicts only through $\{\mu_{ls}\}$ is a function of them, hence of $\theta + c$, and ranks consistently with θ for all θ iff c is constant (if c is not constant, $\theta = -c$ makes $\theta + c$ constant while θ is not, so no ranking read off $\theta + c$ agrees with θ). □

Proposition 1. (i) is the displayed gap identity with $\beta > 0$. (ii) Subtracting the two anchor equations eliminates α and the (zero) residuals. (iii) By Lemma 1 the observations constrain only $\theta + c$, so θ is free up to any vector v: $(\theta - v,\ c + v)$ reproduces every μ_{ls}. The affine residual is $\tilde{c} = (I - P_\theta)\bar{\mu}$ with P_θ the projection onto $\mathrm{span}\{1, \theta\}$, and for generic v the projection onto $\mathrm{span}\{1, \theta - v\}$ differs, so $\tilde{c}$ changes; even with (α, β) pinned by anchors, $\tilde{c}_s = \bar{\mu}_s - \alpha - \beta\theta_s$ takes a different value for every candidate θ_s. So no functional of $\{\mu_{ls}\}$ identifies $\tilde{c}_s$ unless gold pins θ_s itself. □

Proposition 2. For a random pair (s, t) write $\Delta\theta = \theta_s - \theta_t$, $\Delta\tilde{c} = \tilde{c}_s - \tilde{c}_t$, so $\mathbb{E}[\Delta\theta^2] = 2\sigma^2_\theta$, $\mathbb{E}[\Delta\tilde{c}^2] = 2\sigma^2_{\tilde{c}}$ and $\mathbb{E}[\Delta\theta\Delta\tilde{c}] = 0$ because $\tilde{c} \perp \theta$ by construction. The uncorrected gap error is $(\beta - 1)\Delta\theta + \Delta\tilde{c}$ and the rescaled one is $\Delta\tilde{c}/\beta$, so rescaling helps iff $2\sigma^2_{\tilde{c}}/\beta^2 < 2(\beta - 1)^2\sigma^2_\theta + 2\sigma^2_{\tilde{c}}$, i.e. iff $\sigma^2_{\tilde{c}}(1 - \beta^2)/\beta^2 < (1 - \beta)^2\sigma^2_\theta$. For $\beta \geq 1$ the left side is ≤ 0 and the right side ≥ 0. For $\beta < 1$, divide by $(1 - \beta) > 0$ to get $\sigma^2_{\tilde{c}}(1 + \beta)/\beta^2 < (1 - \beta)\sigma^2_\theta$. Maximising $\beta^2(1 - \beta)/(1 + \beta)$ over $(0, 1)$ gives 0.0902 at $\beta = 0.6180$. □

B CLAIMS REVISED OR WITHDRAWN, AND WHY

Each entry below is a claim this study made at some point and then narrowed or withdrew, with all analyses re-run after each revision. The table records what changed and what the change did to the claim; several of the entries are recoverable from the released artefacts by any reviewer who re-runs them.

withdrawn / corrected	why
"Non-identification of c is new"	classical (Cronbach, 1955; Kenny & Albright, 1987; Della Penna & Reid, 2012); Vossler et al. (2026) §4.2 states it for LLM judges. Kept as attributed Lemma 1.
"Judge Robustness Value" as a new statistic	it is the robustness value of Cinelli & Hazlett (2020).
κ triages wrong claims (AUC 0.93)	with a global $\mathrm{RMS}(d)$, κ is a monotone rescaling of the gap: $\mathrm{Spearman}(\kappa, \mathrm{gap}) = 1.000000$ in our own data.
"The panel flips a real ranking"	that pair's gold ordering is itself unestablished ($p = 0.47$); with 14/15 established gold pairs MT-Bench cannot support a precision/recall claim.
Anchor-calibrated certification beats naive ranking	dominated by a no-anchor gap threshold (precision 1.0 at recall 0.80 vs 0.32–0.47); reported as a negative result.
"$\beta > 1$ on preferences, $\beta < 1$ on grading is a task property"	adding open-weight families moved both slopes down monotonically; β tracks panel competence (Table 1). A one-panel study cannot distinguish task from panel effects.
$\mathrm{corr}(\theta, c) = +0.71$ as the headline	opposite sign in the grading setting; the correlation is dominated by the slope, which is why the affine split exists.
"The residual grows from $0.35\times$ to $1.85\times$ the median gap as M grows"	**withdrawn.** $\max - \min$ over M points grows like $\sqrt{2 \log M}$ while the median gap shrinks, so the ratio is partly mechanical: sweeping M inside Arena, the statistic climbs $0.84 \to 1.85$ while flip risk against a matched null moves only $0.303 \to 0.313$ with the excess over the null *falling* from $+0.058$ to $+0.005$. Replaced by flip risk (Table 2).
Reweighting judges to $\beta = 1$ improves on rescaling	tested and negative (§6); reported rather than dropped.
Percentile intervals for β on Arena	they resampled gold, which re-attenuates every replicate, so all four Arena point estimates sat at or above their own upper bound. The estimand is the slope against gold *as reported*, so the regressor is conditioned on (Appendix G).
"The panel gets 14% of pairs wrong"	**withdrawn.** It was the one statistic in Table 2 with no null. Given one, 13.9% against 11.8% is $p = 0.24$: not an established excess. Flip risk and $\mathrm{RMS}(\tilde{c})$ do clear the same null on the deep row.
Table 2's nulls	the flip column used an item-only null while the RMS column used the family-inclusive one, so the table was internally inconsistent and the flip p was too small ($2 \cdot 10^{-4}$ against 0.021). Every column now uses the conservative null, which also moves the headline rows from $p = 0.035/0.23/0.40$ to $0.10/0.52/0.43$.
Proposition 2's second branch as a usable rule	it needs $\sigma^2_{\tilde{c}}/\sigma^2_{\tilde{\theta}}$, which two anchors cannot supply, the same objection the paper raises against $\gamma^\star$. Scope narrowed: two anchors give the $\hat{\beta} \geq 1$ decision, eight give the branch (Appendix J).
"Allocation, not the estimator" at $10\times$	the factorial's two estimators were both weak. Against PPI++ the margin is $1.7\times$ at $n = 16$ and $1.1\times$ at $n = 512$; on the open-weight panel PPI++ wins outright (Appendix I).
"32 anchor labels beat 512" unqualified	true for the widest pair, which §8 tells practitioners not to use; the debiased median over four anchors crosses at $n = 64$ (Appendix I). Stated with the estimator and the currency (effect-size error) everywhere.
A joint regression on kinship and gold as a control	the contrast is orthogonal to gold by construction, so the "control" only inflated the coefficient by $1/(1 - r^2)$; replaced by the within-pair permutation (§7).
"No curvature loaded onto $\tilde{c}$"	a quadratic term explains 20% of $\mathrm{Var}(\tilde{c})$ on two Arena rosters, 5% on the deep slice (Appendix R). Stated as measured.

C GOLD-BUDGET EXPERIMENT DETAILS

Gold is the released human vote on each battle, and the target every strategy is scored against is the full-sample gold vector over all 846 battles, of which each labelled draw is a with-replacement subsample (a draw of $n = 512$ covers about 45% of the battles), so the overlap favours the uniform strategies at large n, not the anchors. For budget n we draw n battles with replacement: uniformly for the uniform strategies, and uniformly from the battles involving either anchor system for the anchor strategies, without stratification over opponents. Anchors are the two systems the *panel* puts furthest apart, so no gold is needed to choose them; this is the widest-pair heuristic, whose $+0.22$ selection bias Appendix O measures and whose debiased four-anchor alternative Appendix I runs at the same budgets. Estimators: *naive* $= \bar{\mu}_s$; *gold-only* the labelled mean with $\bar{\mu}_s$ as fallback; ***PPI++*** $\hat{\theta}_s = \bar{y}_s^{\text{lab}} + \lambda(\bar{z}_s^{\text{all}} - \bar{z}_s^{\text{lab}})$ with power tuning $\lambda = \widehat{\text{Cov}}(y, z)/[\widehat{\text{Var}}(z)(1 + n/N)]$ pooled across systems, shrunk toward the fallback by $m/(m+10)$ for a system with m labels (without this shrinkage the uniform strategies are strawmen at small n, since a one-label mean would overwrite a usable prior). At $n = 512$ a system holds about 46 labels and the estimate still carries 18% of a panel score whose scale is off by β; that is not a handicap: re-running PPI++ with no shrinkage, with $m/(m + 2)$, with none from 20 labels on, or shrunk toward the panel's grand mean gives median gap errors of 0.057, 0.054, 0.057 and 0.049 at $n = 512$ on the commercial panel against 0.049 as run, and the n=32-vs-512 crossing CI excludes a tie against every variant (`results/ppi_shrinkage_control.json`). *anchor* $= (\bar{\mu}_s - \hat{\alpha})/\hat{\beta}$, undefined when either anchor has fewer than two labels or the two gold means coincide, in which case the draw is dropped from the anchor arm's median (Appendix O reports the share). Reported numbers are medians over 120 draws per budget. Results for all three metrics (score RMSE, gap error, pair accuracy) and all three compositions are in the released JSON.

D THE ERRORS-IN-VARIABLES FACTOR

Gold is estimated from finitely many votes, so the regressor carries sampling noise and OLS attenuates. Writing $\hat{\theta} = \theta + u$, the slope of $\bar{\mu}$ on $\hat{\theta}$ is $\beta\,\sigma_\theta^2/(\sigma_\theta^2 + \sigma_u^2)$, so the latent-preference slope is recovered by multiplying by $\hat{\sigma}_{\hat{\theta}}^2/(\hat{\sigma}_{\hat{\theta}}^2 - \hat{\sigma}_u^2)$, with $\hat{\sigma}_u^2$ the mean bootstrap sampling variance of $\hat{\theta}_s$ over items. That gives $\times 1.021$ (MT-Bench), $\times 1.062$ (MATH) and $\times 1.483$ (Arena): Arena's single crowd vote per battle makes 33% of its gold variance sampling noise. MATH's gold is an exact answer match and so is noiseless *per item*, but θ_s is still a mean over 84 problems and inherits that item-sampling variance, which is what the $\times 1.06$ removes.

This correction assumes u is independent of the panel noise, which it is not: judges and the human voter see the same battle. The assumption-free estimate is the disjoint-half one in §5. We report the raw slope throughout because the estimand that matters operationally is the slope against the gold standard *as published*; the correction is a sensitivity, and because it multiplies a whole column by one constant it changes no comparison between compositions.

E PREDICTING THE RESIDUAL FROM OBSERVABLES

(R3) It is not the bias the taxonomies describe. We regressed $\tilde{c}_s$ (Arena, mixed panel) on eight gold-free system features: log length, markdown density, bullet count, code rate, refusal rate, hedging, exclamation density, and kinship. In-sample $R^2 = 0.45$, but leave-one-out R^2 is *negative* for the full model (-0.37), the best ridge (-0.29) and the best single feature (-0.06), with permutation $p = 0.17$; and 0.45 is about what 8 features over 22 systems buy from noise alone, so this excludes *strong* predictability, not moderate. Style biases (Ye et al., 2024; Feuer et al., 2024) are real but load on the *slope*, which calibration absorbs.

F TIE CONVENTIONS

Second, humans tie far more than judges (29.8% vs 16.7% commercial, 12.8% frontier, 4.7% open-weight on Arena), and scoring a tie as half a win compresses gold toward 0.5. Recomputing both scales on human-decided battles with judge ties dropped shrinks the slopes (Arena's commercial

panel falls to 1.246, so about half the excess above one is convention) but the roster dependence is untouched: β_{dec} still spans 0.407 to 1.463.

G INTERVAL CONSTRUCTION

An earlier version of Table 1 reported percentile intervals from a hierarchical bootstrap that resampled families, judges within family, items *and* gold. On Arena that produced intervals whose upper limit every point estimate met or exceeded (the frontier panel read 1.740 with $[1.109, 1.733]$), while MT-Bench and MATH were centred normally. The cause is a property of the estimator, not a coding slip. β is an OLS slope on a regressor $\hat{\theta}$ that is itself a mean over finitely many gold labels, and OLS attenuates by that regressor's reliability. A replicate resamples items, so its $\hat{\theta}$ carries *more* noise than the full-sample one and every replicate is more attenuated than $\hat{\beta}$; the bootstrap distribution shifts down and the percentile interval inherits the shift. The shift must then scale with the share q of observed gold variance that is sampling noise, and it does, but only where q is material: on MT-Bench ($q = 0.020$) and MATH ($q = 0.061$) the per-composition biases scatter within $\pm 2.3\%$ of zero, while on Arena ($q = 0.327$) every composition shifts down, by 17.4% to 19.4%, a mean of -18.3%.

The fix follows from the estimand rather than from a more elaborate interval. Our β is the slope against the gold standard *as reported*: a published leaderboard is a fixed artefact that a practitioner compares against, so we condition on the regressor and resample only the panel side: families, then judges within family, then items. Table 1's intervals are the 2.5th and 97.5th percentiles of that bootstrap, $B = 2000$. Under it $\hat{\beta}$ lands 43–60% of the way up in all twelve cells and the bootstrap bias is at most 1.9% of β (against 18.3% before). The Arena extremes still separate: open-weight $[0.18, 0.76]$ against frontier $[1.53, 1.93]$.

We keep the alternative rather than deleting it. BCa (Efron, 1987), with acceleration from a leave-one-family-out jackknife, applied to the resampled-gold bootstrap gives Arena frontier $[1.71, 2.11]$. It corrects the median bias and the skew, but it *re-centres on* the attenuation artefact instead of removing it, and it targets the latent-preference slope rather than the reported one. That estimand is the subject of Appendix D, where the same q appears as the errors-in-variables factor. Both are in `results/beta_intervals.json`.

H THE ESTIMAND LADDER

Three estimates of the commercial Arena panel's slope appear in this paper, $\beta = 1.480$, $\beta_{\mathrm{eiv}} = 2.201$ and $\beta_{\mathrm{split}} = 1.052$ and a reader is entitled to ask which is the answer. They are one quantity seen through regressors of different reliability $R = 1 - q$:

panel	q	β (all gold)	β_{eiv} (R=1)	β_{split} (half gold)	predicted
commercial	0.327	1.480	2.201	1.052	1.115
mixed	0.327	0.996	1.481	0.705	0.751
open-weight	0.327	0.452	0.672	0.314	0.340
frontier	0.327	1.740	2.587	1.241	1.310

The predicted column is $\beta \cdot (1 - q')/(1 - q)$ with $q' = 2q/(1 + q)$, the noise share when gold comes from half the battles. It contains no fitted parameter. Across all twelve setting×composition cells the prediction matches the measurement to a median of 2.2%; the largest error anywhere is 7.6%, on Arena. So the split-half control is not a competing estimate of the headline number, and quoting $\beta_{\mathrm{split}} = 1.05$ as "the real inflation is 5%" would be an error of the same kind as quoting an uncorrected OLS slope as a latent effect. We report β against gold as reported because that is the quantity a practitioner compares against; the other two rungs are recoverable from q, which we publish.

What survives every rung is the ordering, which is what the paper claims: $\mathrm{corr}(\beta, \beta_{\mathrm{split}}) = +1.000$ in all three settings.

I ALLOCATION VERSUS ESTIMATOR

The claim that concentration of labels rather than the estimator carries §6 needs a factorial, not an aside. Every arm gets the same budget n and the same scoring as §6 (median |gap error| against the full-sample gold vector, 200 draws per cell for the four factorial arms; the PPI++ row is the 120-draw run of §6).

panel	arm	n=16	32	64	128	256	512
commercial	anchor, widest pair	0.047	0.044	0.045	0.044	0.043	0.044
	anchor, median pair	0.071	0.053	0.046	0.046	0.046	0.045
	uniform, affine fit	0.829	0.754	0.443	0.228	0.122	0.085
	uniform, median pair	0.327	0.251	0.167	0.119	0.089	0.071
	uniform, PPI++	0.079	0.079	0.079	0.069	0.060	0.049
mixed	anchor, widest pair	0.042	0.041	0.041	0.040	0.040	0.040
	anchor, median pair	0.059	0.053	0.042	0.042	0.041	0.041
	uniform, affine fit	0.745	0.660	0.404	0.210	0.115	0.074
	uniform, median pair	0.299	0.239	0.161	0.118	0.082	0.067
	uniform, PPI++	0.054	0.058	0.062	0.060	0.054	0.048
open-weight	anchor, widest pair	0.074	0.068	0.071	0.070	0.070	0.070
	anchor, median pair	0.086	0.075	0.074	0.067	0.065	0.065
	uniform, affine fit	0.825	0.798	0.539	0.349	0.223	0.155
	uniform, median pair	0.304	0.247	0.184	0.150	0.120	0.104
	uniform, PPI++	0.069	0.072	0.074	0.071	0.060	0.051

The 2×2 is the four non-italic rows; PPI++ is added as a fifth arm because a factorial containing only estimators we beat would prove nothing. Averaging the two declared estimators, moving from anchors to uniform costs 7.0–10.3$\times$ at $n = 16$ and 1.7–1.9$\times$ at $n = 512$ (the ranges are over the three panels). Averaging the two allocations, moving between estimators costs 2.2–2.3$\times$ at $n = 16$ and 1.1–1.3$\times$ at $n = 512$, and essentially all of it sits in the uniform rows: within the anchor rows the two estimators are within 0.006 of each other from $n = 64$ on. The uniform affine fit also fails outright far more often, degenerate in 46% of draws at $n = 16$ against 7% for anchors, and 73% on the open-weight panel: the same attenuation showing up as instability rather than bias.

Against PPI++ the honest margin is smaller. PPI++ is a much better uniform estimator than either declared level, so the allocation advantage against it is $1.7\times$ at $n = 16$ on the commercial panel and $1.1\times$ at $n = 512$, not $10\times$. What survives is the shape: anchors reach their floor by $n = 32$ while PPI++ needs 512 to approach it, which is the label-efficiency claim, and on the open-weight panel PPI++ *beats* anchors at every budget (0.051 vs 0.070 at $n = 512$), which is Proposition 2 telling the truth about a panel that should not be rescaled at all. The headline's 32 labels are specific to the widest pair: the debiased median over four anchored systems crosses PPI++ at $n = 512$ (0.049) at $n = 64$ (0.046; bootstrap CI for the difference $[-0.0044, -0.0016]$) and not at $n = 32$, where the CI $[0.002, 0.015]$ favours PPI++; on the frontier panel the same estimator crosses at $n = 64$ with CI $[-0.0064, -0.0028]$ (`results/ppi_shrinkage_control.json`).

An earlier revision quoted this control as "0.320 at $n = 16$" from an ad-hoc run that was non-monotone in n and is superseded by the table above; the $n = 512$ value (0.084) reproduces. That run ships as `results/superseded_allocation_vs_estimator_adhoc.csv`; `allocation_control.json` is the authoritative file.

Adaptive allocation does not rescue the uniform side. Concentration is not the only alternative to spreading labels evenly: active testing routes them where the ranking is least resolved. We add that arm, keeping PPI++ as the estimator so only the allocation changes. Labels are drawn one at a time; at each step we take the adjacent pair of the current estimated ranking with the smallest $|\Delta\hat{\theta}|/\mathrm{SE}$ (with $\mathrm{SE}_s = \sqrt{0.25/(m_s + 10)}$, the shrinkage prior counted as ten pseudo-labels) and label a battle involving either system, so the design is nested and one run reports every budget.

		median |gap error|		median pair accuracy	
panel	arm	n=128	512	n=128	512
commercial	anchors (widest pair)	0.044	0.044	0.823	0.823
	uniform, PPI++	0.071	0.047	0.814	0.840
	adaptive, PPI++	0.076	0.055	0.818	0.836
open-weight	anchors (widest pair)	0.071	0.070	0.727	0.727
	uniform, PPI++	0.071	0.049	0.727	0.797
	adaptive, PPI++	0.071	0.054	0.727	0.792

Adaptive routing is *worse* than uniform on gap error on all four panels at $n = 512$ (paired bootstrap CIs $[+0.0021, +0.0063]$ open-weight, $[+0.0038, +0.0075]$ mixed, $[+0.0056, +0.0097]$ commercial, $[+0.0073, +0.0120]$ frontier, all excluding a tie) and ties it on pair accuracy, its own target (every CI covers zero; commercial $[-0.0130, +0.0043]$). The arm is not crippled: it labels 16.6 effective systems of 22 at $n = 32$ and 20.5 at $n = 512$, so it spreads rather than collapsing onto one pair. The mechanism is the gold: an Arena battle carries a single crowd vote, so an adjacent pair that the panel leaves unresolved is mostly *irreducibly* noisy, and labels spent there buy less than the coverage they take from the rest of the board, which the gap metric averages over. Both uniform arms still beat anchors on pair accuracy at $n = 512$ where β is far from one (commercial $+0.013$, CI $[+0.0087, +0.0216]$), and lose to them on gap error, which is the division of labour §6 claims.

J BUYING THE RESIDUAL RATIO

Proposition 2's second branch needs $\sigma_{\tilde{c}}^2/\sigma_\theta^2$, and Proposition 1(iii) says two anchors cannot supply it. That is a real limit, not a presentational one: with $K = 2$ the rule *is* the $\hat{\beta} \geq 1$ heuristic, and we say so in §3 rather than leaving the branch as decoration.

K anchors do supply it. Fitting $\bar{\mu}_k = \alpha + \beta\theta_k + \tilde{c}_k$ by OLS over K gold-evaluated systems, the residual variance around the line estimates $\sigma_{\tilde{c}}^2$ and the spread of the $\hat{\theta}_k$ estimates σ_θ^2; each carries the sampling variance of its own gold mean, which is computable from the labelled battles themselves, so both debias: $\hat{\sigma}_{\tilde{c}}^2 = \max\{0,\ \frac{K}{K-2}\mathrm{Var}(r) - \hat{\beta}^2\bar{v}\}$ and $\hat{\sigma}_\theta^2 = \mathrm{Var}_k(\hat{\theta}_k) - \bar{v}$, with $\bar{v}$ the mean sampling variance. The truncation is not cosmetic: at $n = 512$ the untruncated residual variance is negative in up to 20% of draws, and a truncated estimate returns ratio 0, hence "rescale". A residual exists from $K = 3$ ($K - 2$ degrees of freedom); we require $K \geq 4$ so that its variance has at least two. Two cautions bound the estimator. Anchors chosen by panel spread are not a random sample of the leaderboard, so $\mathrm{Var}_k(\hat{\theta}_k)$ over-states σ_θ^2, by 1.5–1.7× at $K = 8$ and 2.9–3.5× at $K = 4$ on our panels, which biases the ratio downward, toward rescaling; a spread-free alternative reads σ_θ^2 off the panel scores of all M systems, $\hat{\sigma}_\theta^2 = (\mathrm{Var}_s(\bar{\mu}_s) - \hat{\sigma}_{\tilde{c}}^2)/\hat{\beta}^2$, since $\mathrm{Var}_s(\bar{\mu}_s) = \beta^2\sigma_\theta^2 + \sigma_{\tilde{c}}^2$ exactly. And a small K leaves the residual variance itself noisy, whichever denominator is used.

The question is operational: does the rule fed $(\hat{\beta}, \hat{\rho})$ decide as the rule fed the truth? On the three real panels, which sit far from the boundary, $K = 8$ at $n = 512$ agrees with the truth in 97% (commercial) and 99% (open-weight) of draws (98% and 100% with the board-wide denominator), and $K = 4$ is too noisy to use: it wrongly triggers rescaling on the open-weight panel in 82% of draws (58% board-wide), where the heuristic is right almost always. The mixed panel is the exception in both directions (51–72%), and for the reason §6 already gives: at $\beta = 0.996$ the decision is a coin flip whose cost is nil.

Where the branch actually bites (the 11 semi-synthetic cells with $\beta < 1$ and a residual small enough that rescaling helps, built from the real gold vector and the real residual shape), the estimated rule at $K = 8$, $n = 512$ recovers the correct decision in 79% of draws against the $\hat{\beta} \geq 1$ heuristic's 0% (the heuristic is wrong in every such cell by construction). The other side of that number: in the 5 cells with $\beta < 1$ where the heuristic is right, the estimated rule wrongly triggers a harmful rescale in 46% of draws. The board-wide denominator trades one error for the other (71% and 42%). So the branch is not decoration, but it is a noisy instrument that costs six more anchored systems than the scale does, and a practitioner who buys only two should read Proposition 2 as the heuristic.

K THE DOSE-RESPONSE OVER RANDOM SUB-PANELS

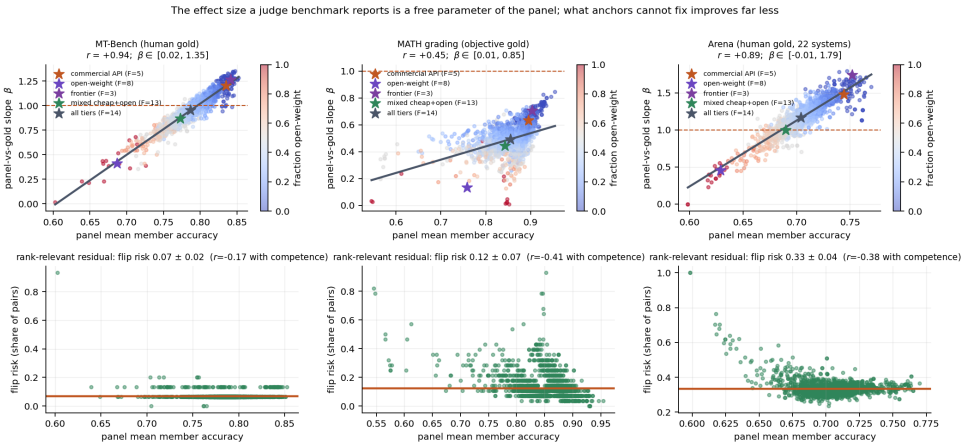


Figure 3: The dose-response. Each point is one random sub-panel on identical items; colour is the fraction of open-weight judges. The slope (top) spans $[-0.01, 1.79]$ on Arena and tracks mean member accuracy ($r = +0.89$). Flip risk (bottom) improves with competence only weakly and from a high floor ($r = -0.38$): the best sub-panels still leave 30% of pairs reversible.

Table 1 reports four named compositions; this is the same quantity over 2,520–3,000 random sub-panels of every size, which is what establishes that β moves continuously with the roster rather than jumping between three tiers we happened to choose.

L SEPARATING THE TIERS WITHOUT RESAMPLING THREE FAMILIES

The frontier tier spans 3 families. A percentile interval whose outer bootstrap stage draws 3 units with replacement has only 10 distinct multisets to work with, so it cannot really be a 95% interval over the facet we argue is binding, and the abstract's "non-overlapping" claim should not rest on it.

Two tests avoid the problem entirely. The first is an *exact permutation* over families: pool the 11 families making up the frontier and open-weight tiers, assign 3 to a pseudo-frontier and 8 to a pseudo-open-weight, recompute the slope difference, and enumerate all $\binom{11}{3} = 165$ assignments. The observed split is the most extreme of the 165 in every setting, so $p = 1/165 = 0.006$, the floor: Arena $\Delta\beta = +1.288$, MT-Bench $+0.851$, MATH $+0.574$. The second is leave-one-family-out: dropping each family in turn leaves frontier in $[1.729, 1.758]$ and open-weight in $[0.388, 0.530]$ on Arena (MT-Bench $[1.207, 1.342]$ vs $[0.359, 0.478]$; MATH $[0.686, 0.722]$ vs $[0.070, 0.156]$), so no single family drives the separation in any setting.

Both are reported in preference to the interval, which we retain in Table 1 for comparability across compositions rather than as the evidence for separation.

M THE BRADLEY–TERRY SCALE

Chatbot Arena reports Bradley–Terry (Bradley & Terry, 1952) ratings on the Elo scale, not raw win rates, so it is fair to ask whether a roster-dependent slope is an artefact of the win-rate parameterisation, a bounded quantity that a linear fit with $\beta = 1.740$ must eventually violate. We fit BT by the standard MM algorithm (Hunter, 2004) separately to (a) the human votes and (b) each panel's mean verdict on the *same* battles, ties split half-and-half on both sides as in the win-rate analysis, and regress panel log-strengths on gold log-strengths across the 22 systems. β_{BT} is then the panel's discrimination relative to the humans', expressed in the units the leaderboard publishes.

composition	L	F	β_{BT}	β (win rate)	panel Elo s.d.
open-weight only	8	8	0.443	0.452	39.5
mixed cheap+open	17	13	1.035	0.996	70.1
commercial API only	9	5	1.720	1.480	116.8
frontier only	5	3	2.240	1.740	150.6
human gold	—	—	1	1	60.9

The ordering is identical to the win-rate ordering, $\mathrm{corr}(\beta_{\mathrm{BT}}, \beta) = +0.993$ across compositions, and the spread is *larger* on the BT scale ($5.06\times$ against $3.85\times$), so the parameterisation is not generating the effect but damping it. Read as Elo, the four rosters compress or stretch the same 22-system leaderboard from a standard deviation of 39 Elo points to one of 151 while the humans give 61: an open-weight panel would report roughly two-thirds of the true Elo gaps and a frontier panel roughly two and a half times them.

N RULE VALIDATION

Proposition 2 is a decision rule, so it owes two things beyond a proof: evidence that its non-trivial branch is worth having, and evidence that it survives being fed an estimate rather than β.

The branch is worth having. We build semi-synthetic panels from the real gold vector ($M = 22$) and the real residual *shape* taken from the mixed panel, rescaled to a chosen $\sigma_{\tilde{c}}^2/\sigma_\theta^2$, and sweep $\beta \in [0.30, 1.40]$ against ratio $\in [0.002, 0.60]$, 56 cells. For each we evaluate both decisions directly, with no modelling step, and compare with what the rule predicts. Since $\tilde{c}$ is defined as the affine residual it is orthogonal to θ by construction (measured $r = 5.6 \times 10^{-16}$), and the pairwise algebra then gives $\mathbb{E}[e_{\mathrm{raw}}] = 2(\beta-1)^2\sigma_\theta^2 + 2\sigma_{\tilde{c}}^2$ against $\mathbb{E}[e_{\mathrm{res}}] = 2\sigma_{\tilde{c}}^2/\beta^2$, which is the threshold. Agreement (56/56) is therefore an implementation check and we do not present it as evidence. The informative quantity is the *regret* of the obvious alternative, "rescale iff $\beta \geq 1$", in the 18 cells where the two differ: a median 218% and a maximum 2983% of excess squared gap error. The branch bites hardest where a weak panel is nonetheless clean: at $\beta = 0.45$, ratio 0.002, ignoring the rule costs $30\times$.

It survives estimation. The protocol never sees β; it sees $\hat{\beta}$ from two anchors. We resample anchor labels at $n \in \{32, 128, 512\}$ (400 outer draws, 60 inner bootstrap replicates for the lower bound) and compare four rules by realised median |gap error|, scored identically to §6 with truth defined in the same currency so the two cannot drift apart, and by regret against an oracle that takes the correct decision using the *same* $\hat{\beta}$.

			decision wrong		mean loss		mean regret	
panel	β	n	point	lower-CI	point	never	point	lower-CI
commercial	1.480	32	6.9%	52.7%	0.0504	0.0819	−0.0005	+0.0124
		128	0.2%	6.5%	0.0443	0.0819	−0.0000	+0.0013
		512	0.0%	0.0%	0.0433	0.0819	0.0000	0.0000
mixed	0.996	32	64.1%	11.8%	0.0422	0.0427	−0.0005	+0.0008
		128	83.2%	29.0%	0.0405	0.0427	−0.0022	−0.0011
		512	98.0%	62.3%	0.0403	0.0427	−0.0024	−0.0017
open-weight	0.452	32	17.1%	0.3%	0.0636	0.0636	−0.0000	+0.0000
		128	4.0%	0.2%	0.0635	0.0636	−0.0001	+0.0000
		512	0.0%	0.0%	0.0636	0.0636	0.0000	0.0000

Entries printed 0.0000 are exactly zero; a signed ± 0.0000 denotes a non-zero value below display precision. Regret is *paired*: within a replicate we compare the rule's realised loss against the oracle's on the same $\hat{\beta}$, then average the differences. The median of those differences would be identically zero whenever the decision is right more than half the time, which is most of this table, so we report the mean, and note that the 90th percentile of the paired difference is ≤ 0 in every cell as well. The oracle takes the loss-minimising decision under the *true* β and executes it with the rule's $\hat{\beta}$, so it isolates decision error from estimation error; it is not an oracle over $\hat{\beta}$, which is why a rule can beat it on an individual replicate and the mean regret can come out slightly negative.

Three things follow. (i) The point rule's mean regret is never positive: at the budgets a practitioner would actually spend, following it is free. (ii) Its decision-error rate is high only on the mixed panel, where $\beta = 0.996$ places the truth on the boundary, and there being "wrong" 98% of the time *lowers* the loss, because dividing by $\hat{\beta} \approx 1$ is nearly the identity and the small correction it does make points the right way. Decision accuracy is therefore the wrong summary near $\beta = 1$; regret is the right one, and it is what we report. (iii) The conservative variant, demanding that a 5th-percentile bootstrap bound clear one before rescaling, is the rule that actually costs something, forgoing most of the improvement on the commercial panel at $n = 32$ (mean regret $+0.0124$), which is the budget where the improvement is largest. Adding a safety margin here is the wrong instinct, because the asymmetry in Proposition 2 already makes the failure mode one-sided.

O Anchor estimator diagnostics

Two caveats bound it. The estimator is a ratio whose denominator is degenerate in 85% of draws at $n = 4$ (falling to 2% at $n = 32$), so the $n = 4$ median rests on a surviving minority and we lean on $n = 32$, where median, mean and 90th percentile are 0.0455, 0.0510 and 0.0764. And picking the pair the panel puts furthest apart selects on a positive residual contrast: $\hat{\beta} = 1.697$ against an OLS 1.480, a $+0.22$ bias that on the mixed panel reads 1.196 where the truth is 0.996, enough to flip our own recommendation. The median $\hat{\beta}$ over well-separated pairs removes most of it (1.000 on the mixed panel). The other two metrics are exactly as unflattering as Proposition 1(i) requires, which is why we report them. Anchor rescaling is an increasing affine map, so it *cannot* change a ranking: pair accuracy is 0.8225 at every budget, identical to spending nothing, while PPI++, which can reorder, reaches 0.8355 at $n = 512$. And on absolute score error a noisy $\hat{\beta}$ makes things worse before better (0.134 at $n = 4$ against 0.076 for the raw panel). Two anchors buy the unit, nothing else.

P Conformal intervals and judge reweighting

Slope removal halves conformal intervals. Following Kargi & Salinas (2026) we form leave-one-system-out conformal intervals for θ_s given $\bar{\mu}_s$ at $\alpha_{\text{conf}} = 0.1$. With raw residuals the commercial panel's intervals are honest but wide (coverage 0.91, width 0.300) because the width is mostly unmodelled slope; fitting the affine map on calibration systems and conformalising its *deleted* residuals (Barber et al., 2021) halves the width to 0.151, but at 0.86 empirical coverage against a 0.90 target, so it is not a free win. The jackknife+ construction guarantees only $1 - 2\alpha$ in the worst case, and at $M = 22$ the gap between 0.91 and 0.86 is one system. The defensible statement is narrow: when $\beta \neq 1$ most of a raw-residual interval's width is unmodelled slope. When $\beta \approx 1$ the two coincide; for the weak panel slope removal buys nothing. Post-affine exchangeability is simply more plausible than raw exchangeability, and their own residual-vs-Elo diagnostic is the symptom.

A negative result: reweighting does not beat rescaling. Because $\beta(w) = \sum_l w_l g_l$ is *linear* in judge weights, with g_l, judge l's own anchor-pair slope, "$\beta = 1$" is a linear constraint and one can solve $\min \|w - \text{uniform}\|^2$ subject to it and to $\sum_l w_l = 1$ (closed form; no sign constraint is needed, the weights stay near uniform), making the panel calibrated by construction and, unlike rescaling, able to change the ranking. It does not pay: across the 138 admissible anchor pairs on the commercial panel (of 231, those whose gold values differ by at least 0.05; each pair is scored on the other 20 systems), median gap error is 0.0451 for rescaling and 0.0746 for reweighting, which wins on only 20.3% of pairs. With one constraint and min-norm regularisation the weights stay near uniform, so the residual barely moves. The one suggestive signal is that reweighting nudges pair accuracy up (0.831 vs 0.820), the only intervention we tested that can move a ranking at all; at our M it is not significant.

Q The residual on a commercial roster

§7 establishes the rank-relevant residual on the one panel deep enough to resolve it, and that panel is eight-elevenths small open-weight judges at $\beta = 0.625$. The obvious question is what it looks like on a roster a practitioner would field, so we ran the commercial tier over the deep slice too and

repeated the audit unchanged. Both panels below are pinned by judge name, and within each the only thing that changes between rows is battles per system.

		btl/		RMS($\tilde{c}$)			flip risk			gold units	
panel (L/F)	depth	sys	β	obs	null	p	obs	null	p	deb.	/gap
commercial (6/3)	shallow	149	1.430	0.0504	0.0442	0.20	29.9%	24.1%	0.07	0.0162	20%
	deep	368	1.472	0.0382	0.0310	0.08	21.2%	17.3%	0.12	0.0147	18%
open-weight (8/8)	shallow	142	0.332	0.0323	0.0280	0.18	54.5%	50.2%	0.24	0.0462	56%
	deep	356	0.341	0.0269	0.0185	**0.005**	48.5%	37.6%	**0.014**	0.0565	71%
pooled (14/11)	shallow	142	0.816	0.0336	0.0296	0.21	30.3%	27.3%	0.25	0.0185	22%
	deep	356	0.829	0.0230	0.0207	0.25	23.8%	19.7%	0.14	0.0114	14%

Nulls are the family-inclusive ones of §3. "Debiased" is $\sqrt{\mathrm{RMS}(\tilde{c})^2 - \mathbb{E}[\text{null}^2]}$ divided by β, and "/gap" expresses that as a share of the median gold gap, the currency a practitioner's effect sizes are in after rescaling.

The three readings below follow.

The residual does not clear the null on a commercial panel, even at this depth. At 368 battles per system, $p = 0.077$ on $\mathrm{RMS}(\tilde{c})$ and 0.12 on flip risk. Depth is not the binding constraint there: the panel's own noise floor is 0.031 against the open-weight panel's 0.019 at the same battle count, because decisiveness inflates the sampling variance of $\bar{\mu}_s$ along with β. A $\beta = 1.47$ panel has a larger residual *and* a larger floor, and the ratio of the two does not improve.

In gold units the ordering reverses, and that is the comparison that means something. $\tilde{c}$ lives on the panel's scale, so panels with different β cannot be compared in panel units at all. Dividing by β, the commercial panel's rank-relevant distortion is 0.0147, 18% of a median gold gap, against the open-weight panel's 0.0565 and 71% — a factor of 3.8. The pooled 14-judge panel, whose $\beta = 0.83$ sits nearest one, is lower still (14%) and also fails to separate ($p = 0.25$). So a competent, decisive roster does not merely report inflated gaps; after the inflation is corrected, it also carries the least rank-relevant bias of the three. That is the good news the audit can offer, and it is the opposite of what a reading of §7 alone would suggest.

The statistical status of the residual is itself roster-dependent. It clears the null on the two low-β panels (open-weight-only $p = 0.005$; the 11-judge panel of Table 2, $p = 0.039$) and on neither high-β nor pooled one. We report all six rows rather than the one that separates: the honest statement is that a rank-relevant residual is *established* at leaderboard scale on weak panels, *bounded but not established* on strong ones, and small in gold units exactly where it is hardest to detect. Non-transfer survives at depth: across the two disjoint deep panels the $\tilde{c}$ vectors correlate $r = +0.08$ over the 22 systems, tighter than the $r = 0.15$ (R4) reports at shallow density.

Two caveats. The commercial deep panel spans three families, against eight for the open-weight one, which bounds the resolution of a null that resamples families — the same objection §5 raises against a percentile interval over three frontier families. And the tier's Google judge is not in it: its coverage on the shallow slice is 886 of 1600, below the pre-registered 90% floor and retained in Table 1's commercial tier only as the last judge of its family, so including it would have cut this comparison's shallow row from 149 to about 82 battles per system, which is the one quantity the comparison varies.

R Robustness checks

Presentation order and gold depth (MT-Bench). Hash-balanced A/B presentation measured 47.8–51.5% per system; restricting to battles with ≥ 2 or ≥ 3 human votes moves the c-spread from 0.193 to 0.189–0.193. **Leave-one-out.** Leave-one-judge spread range $[0.176, 0.212]$; leave-one-family $[0.179, 0.208]$: no single judge or family produces the common-mode component. **Split-half.** $\tilde{c}$ split-half $r = 0.84$ over item halves on the 5-family panel; over the turn-1/turn-2 split $r = 0.40$, and -0.12 without claude-v1: the residual is partly task-composition dependent, which we report as both finding and limitation. **Variance components (Arena).** Resampling items, judges within family, and families separately attributes 62%, 19% and 19% of the residual's sampling variance respectively, which is why depth in battles (not more systems) is the lever on its statistical status.

Judge main effects. In pairwise data $z_A + z_B = 1$ forces $\mathrm{RMS}(a) \approx 0$ (measured 0.0003–0.008); only the grading setting identifies leniency, where it is large ($\mathrm{RMS}(a) = 0.18$, $\max|a_l| = 0.74$), one more reason pointwise scores from different judges are not comparable. **Family-level error bars.** Bootstrapping families rather than items widens CIs by 25–32%. **The coverage floor's family-preservation clause.** Both Google judges cover less than the 90% floor on Arena (859 and 886 of 1,600), so the clause that never drops a family's last judge retains one of them, and the complete-cell block is capped by its 886 items: 846 complete items against the 1,527 available if the family is dropped outright, a 45% cost. That is a real trade, so we re-fit every slope on the 21-judge, 1,527-item panel. The roster ordering is identical and the effect is *larger* without the family: β reads 0.336, 0.896, 1.456, 1.657 against 0.452, 0.996, 1.480, 1.740, a max/min spread of 4.93× against 3.85×, with corr $= +0.999$ across the four compositions and a largest single shift of 0.116. The clause costs battles and buys a family; it does not manufacture the dose-response, which is stronger without it (`results/coverage_floor_sensitivity.json`).

Curvature. The affine split is a choice, so we add a centred quadratic in θ to the fit and ask how much of $\mathrm{Var}(\tilde{c})$ it absorbs. On MATH, under 1% in every composition; on the deep 11-judge Arena panel of §7, 5% ($p = 0.33$); on Arena's mixed and commercial rosters, 20% ($p = 0.041$ and 0.044; open-weight 3%, frontier 6%), with a negative coefficient that survives the Bradley–Terry transform of Appendix M (21–23%, $p = 0.03$–0.04), so it is not win-rate saturation; on MT-Bench $M = 6$ leaves three degrees of freedom and nothing separates. So up to a fifth of the residual variance on two Arena rosters is smooth curvature rather than per-system bias, and the residual audit of §7 rests on the deep panel, where it is 5% (`results/curvature_check.json`).

S SETTING SUMMARY

Table 3: The three settings. Judges are counted after the pre-registered compliance and coverage filters; the analysis uses complete cells (every retained judge scores every retained item).

	MT-Bench	**MATH-500**	**Chatbot Arena**
estimand	human win rate	verifiable accuracy	human win rate
gold standard	3.3k human votes	exact-answer match	crowd votes
systems M	6	8	22
items (raw / complete cells)	401 / 389	90 / 84	1,600 / 846
median panel-score gap over all pairs	0.277	0.059	0.099
judges L / families F	24 / 14	26 / 14	22 / 14
graded (item, system) cells	389	593	846
judgments (complete cells)	9,336	15,418	18,612

MATH cells. $8 \times 84 = 672$ cells were designed and 593 are graded: 12 have no response from one generator and 67 lack a parsable verdict from at least one of the 26 judges, and the complete-cell filter drops them. The dropped cells are harder (mean level 3.7 against 2.8, 1,296 against 614 solution characters, gold accuracy 0.28 against 0.84), so θ_s is a per-system mean over 64–80 problems rather than one common set. On the 52 problems complete for every system the four MATH slopes are 0.169, 0.447, 0.618, 0.707 against Table 1's 0.134, 0.443, 0.633, 0.707, and no ordering changes (`results/kinship_cells_floor.json`).

T PANEL ROSTER

Cheap API (5 families, 13 judges; 9 on Arena, 11 on MT-Bench): OpenAI ×4, Google ×2, DeepSeek ×2, Alibaba ×3, ByteDance ×2, all inexpensive non-reasoning tiers. *Open-weight* (8 families, run locally in bfloat16, greedy, $\leq$ 24 new tokens): Phi-4-mini-instruct, Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, Falcon3-3B-Instruct, Granite-3.3-2B-Instruct, OLMo-2-1B-Instruct, StableLM-2-1.6B-Chat, Yi-1.5-6B-Chat. Four further open-weight families (InternLM2.5-1.8B, SmolLM3-3B, LFM2-2.6B, Aya-8B) were staged in the sweep script but never run, an exclusion made before any of them was judged, on the same in-advance basis as the excluded frontier candidates; the released `gpu_job.sh` still lists them, which is why the script enumerates twelve

and the panel has eight. The deep-slice panel of §7 is the 11 judges that scored both Arena slices: the three OpenAI cheap-tier judges plus all eight open-weight judges, 9 families, at $\beta = 0.625$; the deep slice's 35,723 verdicts are fifteen per-judge shards over the extra battles: those eleven judges, one OpenAI judge the compliance filter later dropped, and three further commercial judges (two ByteDance, one DeepSeek) that make up the commercial deep panel of Appendix Q, with per-judge attrition of up to 80 battles (130 for the dropped judge). Every panel over the deep slice is pinned by judge name in the analysis scripts, so adding a judge to the slice cannot move a published row. *Frontier* (3 families, 5 judges): GPT-5.4, GPT-5.1, Claude-Sonnet-4.6, Claude-Opus-4.8, Qwen-Max, selected by a 10-item pilot on fresh uncached items measuring format compliance and cost; candidates whose endpoints returned persistent errors, and those that truncated before the verdict, were excluded before any analysis, as was a Gemini judge whose sweep could not be completed — no partially-scored judge enters any analysis, which the complete-cell filter enforces mechanically. Two reasoning-mode API judges were dropped by the compliance floor. On Arena both Google judges fell below the coverage floor, at 859 and 886 of 1,600 items; the family-preservation clause of §4 keeps the better-covered one, which is why the commercial tier counts five families there and the complete-cell block is capped at 846 items. No judge's prompt was ever tuned individually. Exact endpoint identifiers are in `src/judges.py`; per-judge verdict counts, before and after the complete-cell filter, are in `results/judge_call_costs.json`.

U Reproduction commands

```
# data and judging (API keys needed only here; every later step is $0)
python3 scripts/00_fetch_data.py        # datasets + SHA-256 manifest
python3 scripts/02_mtbench_panel.py     # judge MT-Bench battles
python3 scripts/03_math_panel.py        # generate + grade MATH-500
python3 scripts/10_arena_panel.py       # judge Arena battles (M=22)
python3 scripts/17_arena_deepen.py --n-items 2400    # the deep slice
bash scripts/gpu_job.sh                 # open-weight tier, all shards
python3 scripts/09_local_judge.py --panel arena \    # ... or one shard;
    --judge-id loc-phi4-mini --model-dir models/loc-phi4-mini
                                        # panel in {mtbench, math, arena}
python3 scripts/14_frontier_panel.py    # frontier tier / extra API items
python3 scripts/08_merge_panels.py      # merge shards, <name>_panel_full
# analyses, in order; <name> in {mtbench, math, arena}
python3 scripts/07_corrected.py --panel <name>_full
python3 scripts/11_beta_dose_response.py --panel <name>_full
python3 scripts/12_method_analyses.py   # predict / conformal / anchors
python3 scripts/19_frontier_analysis.py --panel arena  # tiers, kinship
python3 scripts/20_deep_residual.py     # residual on the deep slice
python3 scripts/21_canonical_table.py   # Table 1 (single source)
python3 scripts/22_residual_table.py    # Table 2
python3 scripts/23_shared_noise_controls.py   # family-inclusive nulls
python3 scripts/24_bt_scale.py          # Appendix M
python3 scripts/25_rule_validation.py   # Appendix N
python3 scripts/26_beta_intervals.py    # Appendix G
python3 scripts/27_estimand_ladder.py   # Appendix H
python3 scripts/15_gold_budget.py       # PPI++ head-to-head (Section 6)
python3 scripts/28_allocation_control.py --reps 200  # Appendix I
python3 scripts/34_ppi_shrinkage_control.py   # Appendix C control
python3 scripts/30_ratio_from_anchors.py      # Appendix J
python3 scripts/31_family_separation.py       # Appendix L
python3 scripts/16_reweight_density.py  # reweighting + flip risk
python3 scripts/35_curvature_check.py   # Appendix R curvature
python3 scripts/36_kinship_cells_floor.py   # kinship, cells, floor
python3 scripts/37_adaptive_allocation.py --reps 250  # Appendix I
python3 scripts/38_commercial_deep.py   # Appendix Q (commercial roster)
python3 scripts/29_count_judgments.py   # the 99,487 reconciliation
python3 scripts/33_judge_call_costs.py  # per-judge verdict counts
python3 scripts/05_figures.py           # Figure 3 (Appendix K), rest
python3 scripts/13_fig_arena_affine.py  # Figure 1
python3 scripts/18_fig_budget.py        # Figure 2
# verification (the reproducibility statement)
```

```
python3 scripts/consistency_scan.py     # every quoted number vs results/
python3 scripts/32_verify_bibliography.py
```