Calibrating Rubric Partition Sensitivity in LLM Judges

Anonymous authors
Paper under double-blind review

Abstract

Equivalent rubric content need not produce equivalent LLM-judge scores when its criteria are grouped differently. Raw differences confound this interface effect with repeat variability, score discreteness, and the number of partitions tested. We introduce an estimator-matched conditional randomization calibration and apply it to 780 items in seven dataset cohorts, five corpus families, and 96,120 successful calls across DeepSeek v4 Pro, Qwen3.8 27B, and Selene-1-Mini. Direct grouped scoring has calibrated drift ratios CDR = 1.259, 1.239, and 2.355 (95% intervals reported in that model order). Item-level excess remains positive after 10% within-cohort trimming for all three models, and no single item changes the corresponding pooled direct-excess estimate by more than 10%. Post-hoc mechanism diagnostics show signed residual centralization across the score range, additional multi-atom contextual residual on common support, and conditional associations with block size and atom disagreement. A partition-independent atom-fixed control has no comparable positive score-unit excess for any model. The holistic single-block contrast is model-dependent, underscoring that the result is a measurement property of the scoring interface rather than a correction coefficient that transfers unchanged across evaluators.

1 Introduction

LLM judges turn open-ended responses into the scalar scores used to compare systems, select training data, and define rewards. Their rubrics are often documented as lists of criteria. The scoring interface is less often treated as part of the instrument: the same ordered criteria may be presented separately, paired into blocks, or judged holistically. A contextual model can use those interfaces differently even when no criterion is added, removed, reworded, or reweighted.

Prompt and rubric sensitivity are already established concerns. Small formatting changes can produce large model-dependent performance shifts (Sclar et al., 2023); evidence-grounded rubric execution improves auditability and stability (Hong et al., 2026); criterion and score-option order can move rubric judgments (Xu et al., 2026); and rubric aggregation choices matter when scores become learning rewards (Gunjal et al., 2025). The open measurement problem is narrower. If we observe dispersion across several content-equivalent rubric partitions, how much exceeds the dispersion expected from repeated calls under the same estimator? Raw standard deviations cannot answer this across datasets whose score distributions, repeat counts, and numbers of legal boundaries differ.

We address that problem with an estimator-matched conditional randomization test and its normalized summary, the calibrated drift ratio (CDR). Here drift has one specific meaning: dispersion across partition means. It is nonnegative and does not imply that one partition is directionally biased. The matched null preserves each item's observed scores and paired display orders while randomizing their boundary labels. CDR divides the observed estimator by the same estimator under exchangeability.

The intervention in Figure 1 uses every internal boundary of an ordered atomic rubric while holding the displayed block count at two. Each boundary is shown in forward and reversed block order, with four repeats. The primary score averages the two orders by canonical block identity. In the illustration, Fine denotes the partition-independent atom-fixed con-

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.

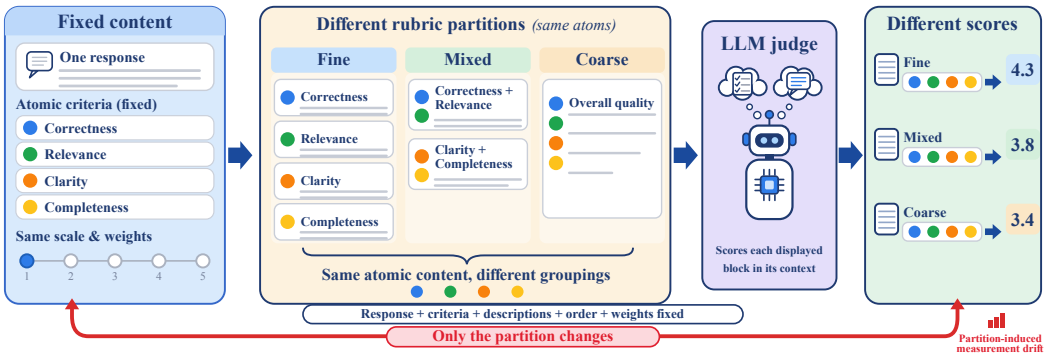

Figure 1: Only the primary boundary moves. Three interfaces over the same ordered atoms: Fine is the partition-independent atom-fixed control, Mixed the primary sweep over legal boundaries with exactly two displayed blocks, and Coarse the single-block holistic arm. Only the boundary changes; atom text, order, weights, response, 0–5 block scale, and output schema stay fixed. "Overall quality" in the Coarse drawing is illustrative shorthand: the implemented prompt keeps every atom description under Block 1. Orders are paired and totals reconstructed by block identity; scores are illustrative.

trol, Mixed denotes the primary fixed-two-block boundary sweep, and Coarse denotes the secondary single-block holistic arm. Fine and Coarse are controls rather than additional partitions in the primary estimand. We apply this protocol to seven non-overlapping cohorts nested in five corpus families and to three model configurations. Previously analyzed and newly sampled HelpSteer2 and UltraFeedback cohorts are rescored under this protocol; no old judge output enters the new estimates.

Our results support three claims. First, direct grouped scoring exceeds its matched null for all three models, and the item distribution is broad enough that the pooled effect survives the declared tail diagnostic. Second, the atom-fixed control does not show comparable positive score-unit excess, while the holistic arm and position diagnostics reveal additional interface effects. Third, post-hoc mechanism analysis is consistent with context-dependent compression: raw block residuals centralize toward the interior of the scale, and multi-atom residuals vary with block size and atom disagreement after singleton scale calibration. These are conditional associations, not causal claims.

- An estimator-matched item-conditional null for dispersion across content-equivalent rubric partitions.
- An auditable seven-strata, three-model study with item distributions, tail sensitivity, and cross-model matching.
- Singleton scale geometry separated from multi-atom context, with the remaining associations quantified by grouped validation.

2 Related Work

LLM-as-a-judge systems range from general pairwise and pointwise evaluators (Zheng et al., 2023) to criteria-based prompting (Liu et al., 2023), learned judges (Zhu et al., 2023; Kim et al., 2023), multidimensional calibrated evaluation (Hashemi et al., 2024), and atomic factual scoring (Min et al., 2023). Bias studies document inconsistent outputs and cognitive or positional effects (Stureborg et al., 2024; Koo et al., 2024). We ask whether a fixed set of atomic criteria yields an invariant scalar when only its scoring boundaries change.

Table 1 summarizes the closest work, which leaves a clear methodological gap. Sclar et al. (2023) measure performance ranges over meaning-preserving prompt formats. Rulers (Hong et al., 2026) operationalizes human rubrics through fixed specifications, evidence, and score calibration and evaluates stability under semantically equivalent perturbations. Xu et al. (2026) identify score-option and criterion-order biases and study order permutation. Rubrics

Table 1: Positioning relative to the most closely related studies. "Matched null" means that the null recomputes the same across-partition dispersion statistic while preserving item-level observations.

Work	Main variable	Main target	Matched null
Sclar et al.	Prompt format	Task performance range	No
Hong et al.	Rubric execution	Human-score reliability	No
Xu et al.	Option/criterion order	Position bias	No
Gunjal et al.	Reward aggregation	Policy learning	No
Ours	Two-block boundary	Partition-mean dispersion	Yes

as Rewards (Gunjal et al., 2025) compares ways to aggregate rubric feedback for reinforcement learning. These studies establish that representation and aggregation matter, but none supplies a null matched to the across-partition dispersion estimator needed to compare rubric geometries with different finite-sample baselines.

Our "invariance" follows the measurement principle that irrelevant instrument changes should not alter the compared quantity (Meredith, 1993; Vandenberg & Lance, 2000).

3 Estimand and Calibration

3.1 Content-equivalent partitions

Let an ordered rubric contain atoms $A = ((a_1, w_1), \ldots, (a_m, w_m))$, where $w_j \geq 0$ and $\sum_j w_j = 1$. Boundary $b \in \{1, \ldots, m-1\}$ creates the two blocks $B_b^L = (a_1, \ldots, a_b)$ and $B_b^R = (a_{b+1}, \ldots, a_m)$. The direct judge returns one score for each displayed block. We reconstruct the total by block identity,

$$T_{ibr}^o = W_b^L q_{ibr}^o(B_b^L) + W_b^R q_{ibr}^o(B_b^R), \tag{1}$$

where $W_b^L = \sum_{j \leq b} w_j$ and $W_b^R = \sum_{j>b} w_j = 1 - W_b^L$, for item i, repeat r, and display order $o \in \{F, R\}$. This prevents the reversed display from changing weights. The order-symmetrized boundary score is $Z_{ibr} = (T_{ibr}^F + T_{ibr}^R)/2$.

For each item we average repeats within boundary, $\bar{Z}_{ib} = R^{-1} \sum_r Z_{ibr}$, and define the item-level dispersion $d_i = \mathrm{SD}_b(\bar{Z}_{ib})$. Here SD_b uses the $N-1$ (Bessel) denominator. With only two legal boundaries, as in AlpacaEval and MT-Bench, the item-level dispersion reduces to $|\bar{Z}_{i1} - \bar{Z}_{i2}|/\sqrt{2}$ rather than the half-range. Thus drift is exactly the dispersion across partition means. A large d_i means that legal boundary choices disagree for that item; it does not say which boundary raises the score. For cohort c in family f, with item set $\mathcal{I}_{fc}$, the cohort-balanced observed drift is $D_{\mathrm{obs},fc} = |\mathcal{I}_{fc}|^{-1} \sum_{i \in \mathcal{I}_{fc}} d_i$.

Let $\mathcal{F}$ be the corpus families and $K = |\mathcal{F}|$; this study has five families, so $K = 5$. The family and primary pools are:

$$D_{\mathrm{obs},f} = \frac{1}{|\mathcal{C}_f|} \sum_{c \in \mathcal{C}_f} D_{\mathrm{obs},fc}, \qquad D_{\mathrm{obs}}^{\mathrm{fam}} = \frac{1}{K} \sum_{f \in \mathcal{F}} D_{\mathrm{obs},f}. \tag{2}$$

The same cohort-then-family operation defines $D_{0,f}$ and D_0^{fam} under the matched null. Thus d_i is a within-cohort building block.

3.2 Estimator-matched null

Under the null, boundary labels are exchangeable within an item. A draw permutes paired observations (T^F, T^R) across boundary labels, preserving per-item scores, repeat pairing, order exposure, and boundary multiplicities. We recompute d_i exactly and apply the same pooling. For each family, let $D_{0,f}$ be the resulting Monte Carlo expectation. We report

$$\mathrm{CDR}_f = \frac{D_{\mathrm{obs},f}}{D_{0,f}}, \qquad \mathrm{CDR} = \frac{1}{K} \sum_{f \in \mathcal{F}} \mathrm{CDR}_f. \tag{3}$$

The score-point excess is $E = D^{\text{fam}}_{\text{obs}} - D^{\text{fam}}_{0}$. The headline CDR is the equal-family mean of family-level CDRs; the observed/null columns give family-balanced drifts whose difference is E. CDR is dimensionless with null value 1; E stays in score points. Randomization p-values use add-one correction; intervals resample items and recompute family-level quantities.

Seven strata are reduced to five family estimates before pooling: the two HelpSteer2 cohorts receive equal weight, as do the two UltraFeedback cohorts, and each family then receives weight $1/K$. This is the primary estimand per judge; an item-weighted pool is a sensitivity analysis. Holm correction covers three judge-level primary tests and, separately, fifteen judge-by-family, three aggregation, and six cohort-stability tests.

3.3 Controls and secondary estimands

The primary intervention fixes the displayed block count at two. Fixing output cardinality, block-score scale, and prompt scaffold isolates boundary placement from the number of requested scores and the amount of text scored together. Fine and Coarse therefore serve as controls for atom-wise and holistic interfaces; they are not folded into the primary two-block CDR.

The atom-fixed protocol requests one score per atom under a prompt that is byte-identical across nominal boundary labels, aggregates totals with the same atom weights, and calibrates them with the same statistic. Positive excess there would indicate that nominal labels or repeat noise, not displayed grouping, reproduce the primary pattern.

The single-block arm presents all atoms together and yields the secondary contrast $H - S$, where H is the holistic score and S is the mean order-symmetrized two-block total. Position is analyzed at block level through a first-position contrast and at total level through $T^F - T^R$, and a separate calibrated statistic tests whether the order contrast varies by boundary. These quantities diagnose the scoring interface; none replaces the primary estimator.

The distribution and mechanism analyses are post-hoc diagnostics of the completed run; we selected them after inspecting aggregate residual patterns, so they describe compatible structure rather than test a preregistered mechanism. Singleton direct blocks calibrate the scale. We subtract a cross-fitted singleton baseline from multi-atom residuals on common atomic-mean support and summarize the contextual residual by block size and atom disagreement. All mechanism statements below are conditional associations.

4 Unified Experiment

4.1 Data and intervention

The study contains 780 items from seven dataset-cohort strata: two non-overlapping HelpSteer2 samples (Wang et al., 2024), two non-overlapping UltraFeedback samples (Cui et al., 2023), and one sample each from MT-Bench (Zheng et al., 2023), AlpacaEval (Li et al., 2023), and GSM8K (Cobbe et al., 2021). They form five corpus families. HelpSteer2 has five atoms, UltraFeedback and GSM8K-v2 four, and MT-Bench and AlpacaEval three. GSM8K-v2 retains a binary final-answer atom as an objective anchor, normalized to the common 0–5 scale for atom-fixed aggregation. The direct grouped contract nevertheless requests a common ordinal 0–5 score for every displayed block, including a block containing that objective atom. This scale asymmetry limits the GSM8K comparison. Appendix A.1 lists every atom and cohort.

Every legal internal boundary is evaluated: four for HelpSteer2, three each for UltraFeedback and GSM8K-v2, and two each for MT-Bench and AlpacaEval. Every direct boundary has forward and reversed displays and four repeats in each order. Atom-fixed has four repeats for every nominal boundary. Single-block has four repeats per item. This yields 32,040 scored cells per model configuration and 96,120 overall.

Table 2: Dense headline summary. Observed/null are the family-balanced direct drift and its estimator-matched null in score points; CDR is dimensionless; excess, holistic–split, and first-position entries are in score points; the boundary × order entry is a calibrated interaction CDR. Intervals are 95% item-cluster bootstrap intervals.

Judge	Observed / null	Direct CDR	Direct excess	Atom-fixed CDR	Holistic − split	First-position	Boundary × order
DeepSeek v4 Pro	0.143 / 0.114	1.259 [1.123, 1.404]	0.029 [0.017, 0.042]	0.988 [0.854, 1.135]	-0.103 [-0.131, -0.077]	0.031 [0.020, 0.042]	1.087 [0.956, 1.227]
Qwen3.8 27B	0.190 / 0.154	1.239 [1.135, 1.343]	0.036 [0.020, 0.053]	0.915 [0.837, 0.995]	-0.198 [-0.242, -0.156]	0.055 [0.039, 0.071]	1.185 [1.061, 1.322]
Selene-1-Mini	0.139 / 0.060	2.355 [2.139, 2.579]	0.079 [0.068, 0.090]	1.041 [0.849, 1.264]	0.264 [0.222, 0.309]	0.040 [0.016, 0.064]	2.464 [2.238, 2.693]

4.2 Evaluator setup

We evaluated three fixed model configurations: DeepSeek v4 Pro, Qwen3.8 27B, and Selene-1-Mini. All configurations received the same stimulus and scoring contract; we compare interfaces within each configuration and treat none as representative of a model family.

The task grid was completed before computing the estimands. We retained one valid observation per pre-specified cell and paired forward/reversed displays before computing dispersion.

4.3 Inference

We use the estimator-matched conditional randomization test and item-cluster bootstrap defined above. The primary analysis averages within-cohort estimates and weights families equally; all control and diagnostic analyses follow their stated estimands. The released code exposes the exact offline computation for confirmatory_v1, and the same pipeline reproduces the full unified run.

5 Results

5.1 The primary drift estimate replicates across models

DeepSeek v4 Pro gives CDR = 1.259 (95% CI [1.123, 1.404]) and excess $E = 0.029$ ([0.017, 0.042]). Qwen3.8 27B gives CDR = 1.239 (95% CI [1.135, 1.343]) and excess $E = 0.036$ ([0.020, 0.053]). Selene-1-Mini gives CDR = 2.355 (95% CI [2.139, 2.579]) and excess $E = 0.079$ ([0.068, 0.090]). The randomization value is $p < 10^{-4}$ for each model and the three model-level Holm-adjusted values are 0.00015. The pooled CDRs are above the matched null for all three models, with Selene-1-Mini showing the largest calibrated excess. The headline CDR is the equal-family mean of family-level CDRs; it is therefore not the ratio of the two pooled observed/null columns in the dense table.

The direct item excess is a per-item score-point comparison between observed partition dispersion and the matched null; it is not a raw evaluator score. The untrimmed pooled excess is positive for all three models, and it remains positive after 10% within-cohort symmetric trimming: 0.022 for DeepSeek v4 Pro, 0.030 for Qwen3.8 27B, and 0.068 for Selene-1-Mini. The largest leave-one-item changes are 4.5%, 3.1%, and 1.0%, respectively. All three satisfy the declared tail criterion, so the result is distributed across items rather than driven by one extreme item; the trimming analysis is diagnostic because it is conditioned on the outcome.

We do not collapse pairwise item matching into a single cross-model correlation, so cross-configuration association stays a descriptive appendix diagnostic and the replication claim rests on model-specific estimates. A calibration coefficient fitted for one evaluator should not be transferred to another.

5.2 Family profiles recover replication and measurement context

Selene-1-Mini's direct-scoring CDR is elevated in every family, while DeepSeek v4 Pro and Qwen3.8 27B show the clearest elevation in AlpacaEval, HelpSteer2, and UltraFeedback (Table 3, Figure 3). MT-Bench stays close to its calibrated null for DeepSeek v4 Pro and Qwen3.8 27B but is elevated for Selene-1-Mini. GSM8K-v2 is less decisive for the first

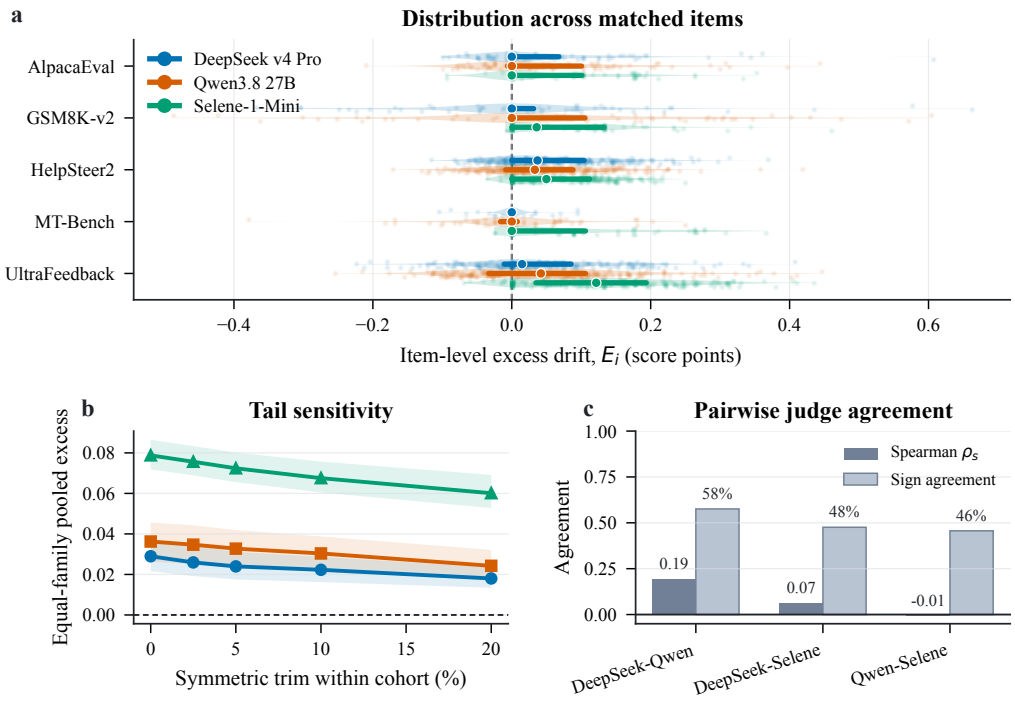

Figure 2: Item-level distribution and tail sensitivity. Panel (a) shows the direct item excess $E_i = d_i^{\text{obs}} - d_i^{\text{null}}$ per corpus family; the point and interval inside each raincloud are the median and interquartile range. Panel (b) re-pools item excess after symmetric within-cohort trimming. Panel (c) reports pairwise matched-item sign agreement across the three deployments. Bands are 95% item-cluster bootstrap intervals.

two models: their family-level CDR intervals include 1, whereas Selene-1-Mini's interval is above 1. Endpoint mass and atom disagreement vary substantially across families, so raw dispersion is not comparable without matched calibration.

The two independently sampled cohorts within HelpSteer2 and UltraFeedback are stable: all six new-minus-old intervals contain zero and every Holm-adjusted comparison is $p = 1.0$ (Appendix Table 6), so the pooled effects are not carried by one reused cohort.

As a sensitivity check, removing GSM8K-v2 and averaging the other four family CDRs gives 1.309 for DeepSeek v4 Pro, 1.254 for Qwen3.8 27B, and 2.345 for Selene-1-Mini.

5.3 Partition-independent scoring removes positive score-unit excess

For DeepSeek v4 Pro, atom-fixed CDR is 0.988 ([0.854, 1.135]) and excess is 0.003 ([-0.011, 0.018]). For Qwen3.8 27B, CDR is 0.915 ([0.837, 0.995]) and excess is -0.007 ([-0.024, 0.010]).

Table 3: Family profile and replication map. Slash-separated entries follow the judge order of the main results. Cohorts/items counts item rows across non-overlapping cohorts; atom SD is the item-average SD of stable atom-fixed profiles; endpoint percentage uses canonical direct block scores of 0 or 5; positive-item percentage is the fraction of item excess values above zero.

Family	Cohorts / items	Atoms / boundaries	Atom SD	Endpoint %	Direct CDR	Positive item %
AlpacaEval	1 / 100	3 / 2	0.530 / 0.543 / 0.380	76.9 / 66.9 / 9.5	1.617 / 1.472 / 2.476	42.0 / 50.0 / 44.0
GSM8K-v2	1 / 100	4 / 3	0.975 / 0.384 / 1.193	77.2 / 80.0 / 0.7	1.058 / 1.177 / 2.393	28.0 / 37.0 / 52.0
HelpSteer2	2 / 250	5 / 4	0.716 / 0.530 / 0.592	24.1 / 9.9 / 0.6	1.377 / 1.279 / 2.025	71.0 / 69.8 / 69.8
MT-Bench	1 / 80	3 / 2	0.045 / 0.207 / 0.310	95.8 / 80.6 / 6.2	1.008 / 1.004 / 2.630	10.0 / 27.5 / 48.8
UltraFeedback	2 / 250	4 / 3	0.551 / 0.575 / 0.506	44.5 / 15.7 / 16.2	1.234 / 1.261 / 2.251	56.2 / 63.3 / 82.3

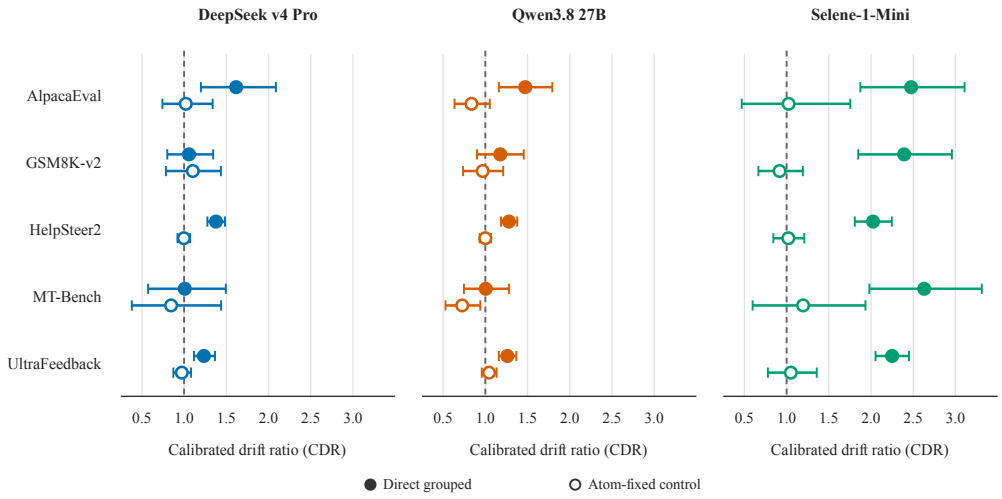

Figure 3: Observed partition dispersion relative to its matched null. Filled markers show direct grouped scoring; open markers show the partition-independent atom-fixed control. Bars are 95% item-cluster bootstrap intervals. HelpSteer2 and UltraFeedback first average their two cohorts. The dashed line is the calibrated null CDR = 1.

Selene-1-Mini's CDR is 1.041 ([0.849, 1.264]) and excess is 0.001 ([-0.005, 0.007]). Qwen3.8 27B's ratio interval lies just below 1, while its score-unit excess interval contains zero; Selene-1-Mini's ratio interval includes 1 and its excess interval is centered near zero. A ratio interval under equal-family weighting need not agree with a difference interval, so we report both scales and claim only that no model shows positive pooled atom-fixed excess comparable to direct grouped scoring.

The single-block contrast is model-dependent. The pooled holistic-minus-split contrast is -0.103 ([-0.131, -0.077]) for DeepSeek v4 Pro, -0.198 ([-0.242, -0.156]) for Qwen3.8 27B, and 0.264 ([0.222, 0.309]) for Selene-1-Mini. Grouping all criteria into one score is therefore consequential, not a neutral shortcut.

5.4 Order is a separate, model-dependent interface effect

DeepSeek v4 Pro's first-position advantage is 0.031 points ([0.020, 0.042]), Qwen3.8 27B's is 0.055 ([0.039, 0.071]), and Selene-1-Mini's is 0.040 ([0.016, 0.064]). Once totals are reconstructed by block identity, the forward-minus-reversed contrast is -0.007 ([-0.024, 0.011]) for DeepSeek v4 Pro, 0.047 ([0.020, 0.074]) for Qwen3.8 27B, and 0.007 ([-0.018, 0.033]) for Selene-1-Mini. The boundary-by-order interaction CDR is 1.087 ([0.956, 1.227]), 1.185 ([1.061, 1.322]), and 2.464 ([2.238, 2.693]), respectively. Qwen3.8 27B shows both a total-order effect and an interaction; Selene-1-Mini shows a strong boundary interaction despite a near-zero total-order contrast.

These diagnostics agree with prior evidence that criterion order matters (Xu et al., 2026). They do not invalidate the main CDR, which averages paired orders before computing dispersion, but they show why a one-order design would mix partition and position effects.

Table 4: Partition-independent and aggregation-level controls, pooled with equal family weights. Atom-fixed entries are CDRs; holistic entries are single-block minus mean split total in score points.

	DeepSeek		Qwen		Selene	
Estimand	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
Atom-fixed	0.988	[0.854, 1.135]	0.915	[0.837, 0.995]	1.041	[0.849, 1.264]
Holistic − split	−0.103	[−0.131, −0.077]	−0.198	[−0.242, −0.156]	0.264	[0.222, 0.309]

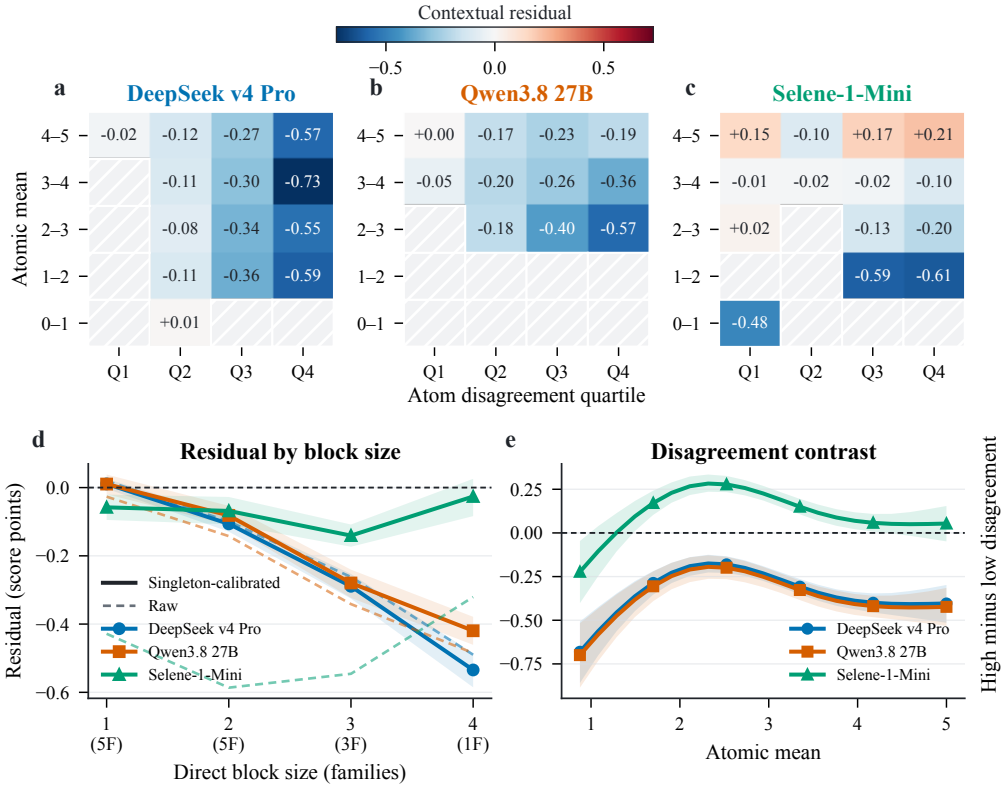

Figure 4: Post-hoc mechanism landscape. Panels (a–c): singleton-calibrated contextual residual heatmaps for the three evaluators over common support; hatched cells contain fewer than 30 items. Panel (d): raw and singleton-calibrated residuals by direct block size, with family support falling from five families at sizes one and two to one at size four. Panel (e): model-based high-minus-low disagreement contrast over common support. Bands are item-cluster bootstrap intervals; all panels are descriptive conditional associations.

5.5 Mechanism diagnostics support context-dependent compression

The direct-minus-atomic residual is not a universal downward shift: item-level scatter changes sign across the score range. With block size fixed, singleton rows show the clearest bounded-scale pattern, positive away from the upper endpoint and negative near it. We call this signed centralization.

To separate that scale geometry from context, we fit a natural cubic spline to singleton direct blocks within each judge and cross-fit it by (dataset, item). For each multi-atom row, the contextual residual is the observed residual minus the singleton prediction, evaluated only on the overlap of singleton and multi-atom atomic-mean support. The full item-centered model explains 0.365 of held-out within-item variation, compared with 0.010 for judge indicators alone, 0.288 after adding scale, and 0.312 after adding block size. The Shapley increments are 0.257, 0.025, and 0.072 for scale, size, and disagreement, respectively (Appendix Table 9).

Contextual residual is close to zero for singletons and generally becomes more negative for larger multi-atom blocks, with a support-dependent recovery at the largest sizes for Selene-1-Mini. Block size is tied to family and boundary structure, and the four-atom endpoint comes only from HelpSteer2. The disagreement contrast is negative over the common prediction grid for DeepSeek v4 Pro and Qwen3.8 27B, while Selene-1-Mini's contrast is mixed across the score range. The leave-one-family-out checks qualify the generalization: the verified GSM8K-v2-held-out fit is $R^2 = 0.298$, and all five family-held-out fits are positive but range down to about 0.198.

On balanced subsets in which every retained family contributes every retained block size, the Shapley size increment falls from 0.025 to 0.009 for sizes one and two across all five families, and to 0.021 for sizes one through three in the three supporting families: it stays positive but smaller. Because the model is item-centered, family fixed effects are absorbed by item identity.

The single-block arm gives a separate aggregation-level comparison. The holistic-minus-split contrast is negative for DeepSeek v4 Pro and Qwen3.8 27B but positive for Selene-1-Mini, while the atom-fixed control has no comparable positive score-unit excess. Together, these controls distinguish three sources of variation: repeat noise captured by the matched null, bounded-scale geometry captured by singleton calibration, and additional context associated with jointly scoring multiple atoms.

6 Limitations

We report three fixed evaluator deployments; the published names do not characterize their broader model families. Our estimand is dispersion across content-equivalent partitions, so the results establish interface non-invariance rather than which interface is most valid against human ratings, and no human, ranking, or policy-learning experiment is included. The mechanism results are post-hoc conditional associations, not identified mediators, and restricting them to the balanced block-size sets keeps the size increment positive but smaller (0.009 and 0.021). Scope is limited to English prompts, a 0–5 scale, one prompt family, and five corpus families, and atom-fixed scoring trades more score objects for recomputability.

7 Discussion and Conclusion

Calibration changes what a partition sweep can support. Raw across-partition SD is not portable: it varies with repeat noise, the score lattice, endpoint mass, and boundary count. CDR gives a common reference by replaying the same estimator under an item-conditional null; the score-unit excess keeps the original scale.

For practice, a rubric specification should record four objects: the ordered atoms and their weights, the partition shown to the judge, the aggregation code, and a repeat-calibrated uncertainty analysis. Atom-level outputs allow transparent re-aggregation later and make the aggregation rule explicit.

The three configurations separate shared from model-specific behavior. DeepSeek v4 Pro and Qwen3.8 27B have close pooled CDRs, whereas Selene-1-Mini has a larger direct excess and a strong boundary interaction. Qwen3.8 27B shows a lower atom-fixed ratio and a larger holistic penalty; Selene-1-Mini's holistic contrast has the opposite sign. The calibration method is portable, but fitted effects must be re-estimated for each deployed judge and interface.

Partition sensitivity also matters when scores become rewards (Gunjal et al., 2025): a pipeline that changes boundaries changes its reward landscape without changing criterion prose. A scalar shift matters for policy learning only if it reorders candidate completions, which a common shift need not do. We measure dispersion, not downstream ranking.

Across 96,120 successful calls, moving only a fixed-two-block boundary produces more dispersion across partition means than matched repeat variability for all three evaluators, and the finding survives equal-family and item weighting, recurs in independent cohorts, and is absent as positive score-unit excess under partition-independent atom scoring. Estimator-matched calibration thereby turns rubric partition sensitivity from a raw prompt comparison into a comparable, auditable measurement diagnostic.

AI use statement

In this work, we used generative AI tools for tasks that require disclosure: implementing the experimental methods, including the analysis, estimation, and dataset-preparation code, and translating our draft text into English. We did not use generative AI to generate synthetic datasets, propose or refine hypotheses, design the study, develop the conceptual framework, formulate or prove the mathematical claims, or interpret the results; the estimands, experimental design, and interpretation are the authors'. We also used generative AI tools for tasks with recommended disclosure: writing and editing software code, generating the reported tables and figures from that code, and editing the manuscript for readability. All AI-assisted work was reviewed by the authors: we ran and audited the model evaluations, recomputed the reported statistics from the frozen records, checked the code against the protocol, and verified all cited metadata against primary records and an independent bibliographic index. We take responsibility for the final content of this work, including all text, claims, and artifacts.

Reproducibility statement

The experiment and analysis code is released with the anonymized review artifact, and the method is specified in full in Section 3 and Appendix A: the atom sets, the boundary intervention, the prompt contract, the estimator-matched null, the pooling rules, and the inference procedure. Running the released pipeline against the same three evaluation models reproduces the reported results end to end, and every reported statistic can also be recomputed offline from the released derived scores without model calls.

References

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021. URL https://arxiv.org/abs/2110.14168v1.

Ganqu Cui, Lifan Yuan, Ning Ding, Guanming Yao, Bingxiang He, Wei Zhu, Yuan Ni, Guotong Xie, Ruobing Xie, Yankai Lin, Zhiyuan Liu, and Maosong Sun. UltraFeedback: Boosting language models with scaled AI feedback, 2023. URL https://arxiv.org/abs/2310.01377.

Anisha Gunjal, Anthony Wang, Elaine Lau, Vaskar Nath, Yunzhong He, Bing Liu, and Sean M. Hendryx. Rubrics as rewards: Reinforcement learning beyond verifiable domains, 2025. URL https://arxiv.org/abs/2507.17746.

Helia Hashemi, Jason Eisner, Corby Rosset, Benjamin Van Durme, and Chris Kedzie. LLM-Rubric: A multidimensional, calibrated approach to automated evaluation of natural language texts. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 13806–13834. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.745. URL https://aclanthology.org/2024.acl-long.745/.

Yihan Hong, Huaiyuan Yao, Bolin Shen, Wanpeng Xu, Hua Wei, and Yushun Dong. From rubrics to reliable scores: Evidence-grounded text evaluation with LLM judges, 2026. URL https://arxiv.org/abs/2601.08654. EMNLP 2026 Main Conference.

Seungone Kim, Jamin Shin, Yejin Cho, Joel Jang, Shayne Longpre, Hwaran Lee, Sangdoo Yun, Seongjin Shin, Sungdong Kim, James Thorne, and Minjoon Seo. Prometheus: Inducing fine-grained evaluation capability in language models, 2023. URL https://arxiv.org/abs/2310.08491v1.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In Findings of the Association for Computational Linguistics: ACL 2024, pp. 517–545. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.29. URL https://aclanthology.org/2024.findings-acl.29/.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. AlpacaEval: An automatic evaluator of instruction-following models, May 2023. URL https://github.com/tatsu-lab/alpaca_eval. GitHub repository.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.153. URL https://aclanthology.org/2023.emnlp-main.153/.

William Meredith. Measurement invariance, factor analysis and factorial invariance. Psychometrika, 58(4):525–543, 1993. doi: 10.1007/BF02294825. URL https://doi.org/10.1007/BF02294825.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long form text generation. In Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 12076–12100. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I Learned to Start Worrying about Prompt Formatting, 2023. URL https://arxiv.org/abs/2310.11324.

Rickard Stureborg, Dimitris Alikaniotis, and Yoshi Suhara. Large language models are inconsistent and biased evaluators, 2024. URL https://arxiv.org/abs/2405.01724v1.

Robert J. Vandenberg and Charles E. Lance. A review and synthesis of the measurement invariance literature: Suggestions, practices, and recommendations for organizational research. Organizational Research Methods, 3(1):4–70, 2000. doi: 10.1177/109442810031002. URL https://doi.org/10.1177/109442810031002.

Zhilin Wang, Yi Dong, Olivier Delalleau, Jiaqi Zeng, Gerald Shen, Daniel Egert, Jimmy J. Zhang, Makesh Narsimhan Sreedhar, and Oleksii Kuchaiev. HelpSteer2: Open-source dataset for training top-performing reward models, 2024. URL https://arxiv.org/abs/2406.08673v1.

Yuzheng Xu, Tosho Hirasawa, Tadashi Kozuno, and Yoshitaka Ushiku. Am I More Pointwise or Pairwise? Revealing Position Bias in Rubric-Based LLM-as-a-Judge, 2026. URL https://arxiv.org/abs/2602.02219.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena. In Advances in Neural Information Processing Systems, volume 36, pp. 46595–46623, 2023. doi: 10.52202/075280-2020. URL https://arxiv.org/abs/2306.05685v1.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. JudgeLM: Fine-tuned large language models are scalable judges, 2023. URL https://arxiv.org/abs/2310.17631v1.

A Reproducibility and Additional Results

A.1 Scope and model configurations

The broader study contains 780 items in seven non-overlapping dataset strata nested in five corpus families. HelpSteer2 and UltraFeedback each contribute an old and a new cohort; MT-Bench, AlpacaEval, and GSM8K-v2 contribute one cohort each. The cohort labels describe sampling provenance, not evidential rank.

Table 5: Seven dataset-cohort strata nested in five corpus families. Boundaries are the structurally distinct fixed-two-block partitions analyzed.

Dataset stratum	Family	Cohort	Items	Atoms	Boundaries
AlpacaEval	AlpacaEval	corrected benchmark	100	3	2
GSM8K-v2	GSM8K-v2	re-atomized objective anchor	100	4	3
HelpSteer2 (new)	HelpSteer2	prior fixed boundary	150	5	4
HelpSteer2 (old)	HelpSteer2	previously analyzed	100	5	4
MT-Bench	MT-Bench	corrected official	80	3	2
UltraFeedback (new)	UltraFeedback	prior fixed boundary	150	4	3
UltraFeedback (old)	UltraFeedback	previously analyzed	100	4	3

The ordered atoms are fixed within each family. HelpSteer2 uses helpfulness, correctness, coherence, complexity fit, and verbosity. UltraFeedback uses instruction following, truthfulness, honesty, and helpfulness. MT-Bench uses instruction following, accuracy, and style/clarity. AlpacaEval uses reference coverage, instruction fit, and fluency. GSM8K-v2 uses problem understanding, solution strategy, reasoning and computation, and a binary final-answer atom. The final atom is normalized to the common 0–5 scale only for atom-fixed aggregation; this scale difference is a limitation of that comparison. Atoms have equal intended weight within a family, so a direct block receives the sum of the weights of its atoms.

For all three evaluator configurations—DeepSeek v4 Pro, Qwen3.8 27B, and Selene-1-Mini—we report the public model names. The names support within-model interface comparisons, not claims that a particular deployment represents its broader model family.

A.2 Boundary intervention

For an ordered rubric with m atoms, boundary b divides the sequence into a prefix and suffix. We evaluate every legal boundary with two displayed blocks, both forward and reversed order, and four repeats per order. Totals are reconstructed by block identity so that reversing the display does not change the atom weights. The two order observations are averaged before boundary dispersion is computed. The resulting statistic is therefore sensitive to grouping while cancelling a constant first-position effect.

The atom-fixed control requests one score per atom under the same nominal boundary grid and applies the same atom weights. The single-block control requests one score for all atoms together. These controls change the scoring interface in known ways and are reported separately from the primary fixed-two-block boundary sweep.

Each direct condition uses the same stimulus, atom descriptions, score scale, and structured output contract; only the boundary and display order vary. The public artifact documents the corresponding confirmatory contract and user-message prompt templates without including source item text, responses, or service records.

A.3 Estimator-matched inference

The item-level drift is the sample standard deviation across boundary means, using the $N - 1$ denominator. The null permutes complete forward/reversed pairs among boundary labels within an item, preserving the observed score multiset, repeat counts, and number of observations per boundary. CDR is observed drift divided by the mean drift under this estimator-matched null; the score-unit excess is the corresponding difference.

We first average item estimates within each cohort and then give equal weight to corpus families. Confidence intervals use item-cluster bootstrap samples; randomization p-values use add-one correction. Holm correction is applied within each declared testing family. The broader-study aggregate is thus a summary of a fixed estimand rather than a claim that all items or all evaluator configurations receive equal raw-record weight.

A.4 Additional results

The family decomposition and control estimates are reported in the main text. The following table gives the new-cohort minus old-cohort CDR comparisons for the two families with independent old and new cohorts.

Table 6: Cohort stability for the two families with old and new non-overlapping samples. Differences are new-cohort minus old-cohort CDR; Holm correction covers the judge-by-family comparisons.

Family	Judge	ΔCDR	95% CI	Holm p
HelpSteer2	DeepSeek v4 Pro	-0.042	[-0.251, 0.162]	1.0
HelpSteer2	Qwen3.8 27B	-0.006	[-0.201, 0.184]	1.0
HelpSteer2	Selene-1-Mini	0.043	[-0.403, 0.486]	1.0
UltraFeedback	DeepSeek v4 Pro	0.115	[-0.131, 0.354]	1.0
UltraFeedback	Qwen3.8 27B	-0.058	[-0.267, 0.148]	1.0
UltraFeedback	Selene-1-Mini	0.070	[-0.321, 0.467]	1.0

The position table separates first-position score contrasts, canonical total contrasts, and the boundary-by-order interaction. These are diagnostics: the primary estimator averages the paired orders before comparing boundaries.

The item-weighted pool, which weights items rather than families, gives direct CDRs 1.286, 1.255, and 2.268 in the main-results judge order. Holm-adjusted family-level p-values over the five families in profile-table row order are 0.00075, 0.5935, 0.00075, 0.9813, 0.00075 for DeepSeek v4 Pro, 0.00075, 0.0170, 0.00075, 0.9813, 0.00075 for Qwen3.8 27B, and 0.00075, 0.00075, 0.00075, 0.00075, 0.00075 for Selene-1-Mini.

Table 7: Position and boundary-by-order diagnostics with equal family weights. The first two rows are score-point contrasts; the final row is a calibrated dispersion ratio.

	DeepSeek		Qwen		Selene	
Estimand	Estimate	95% CI	Estimate	95% CI	Estimate	95% CI
First-position advantage	0.031	[0.020, 0.042]	0.055	[0.039, 0.071]	0.040	[0.016, 0.064]
Forward − reversed total	-0.007	[-0.024, 0.011]	0.047	[0.020, 0.074]	0.007	[-0.018, 0.033]
Boundary × order CDR	1.087	[0.956, 1.227]	1.185	[1.061, 1.322]	2.464	[2.238, 2.693]

The repeat-variability table reports exact-match rates and within-cell standard deviations by stratum and protocol, including the single-block arm.

Table 8: Repeat variability by stratum and protocol. For direct and atom-fixed rows, Exact is the fraction of item–boundary–order cells whose four canonical totals match exactly; for single-block rows, it is the fraction of item cells whose four totals match. SD is the mean within-cell sample standard deviation.

	DeepSeek		Qwen		Selene	
Stratum, protocol	Exact	SD	Exact	SD	Exact	SD
AlpacaEval, direct	0.545	0.160	0.403	0.236	0.880	0.041
AlpacaEval, atom-fixed	0.440	0.131	0.370	0.166	0.860	0.042
AlpacaEval, single-block	0.690	0.192	0.480	0.324	0.890	0.063
GSM8K-v2, direct	0.548	0.477	0.507	0.664	0.942	0.018
GSM8K-v2, atom-fixed	0.720	0.321	0.510	0.672	0.670	0.096
GSM8K-v2, single-block	0.600	0.395	0.520	0.704	0.980	0.022
HelpSteer2 (new), direct	0.197	0.357	0.107	0.378	0.888	0.037
HelpSteer2 (new), atom-fixed	0.095	0.302	0.022	0.359	0.660	0.063
HelpSteer2 (new), single-block	0.393	0.353	0.333	0.396	0.947	0.027
HelpSteer2 (old), direct	0.194	0.349	0.104	0.381	0.868	0.047
HelpSteer2 (old), atom-fixed	0.105	0.297	0.030	0.349	0.630	0.072
HelpSteer2 (old), single-block	0.370	0.366	0.420	0.361	0.980	0.012
MT-Bench, direct	0.859	0.049	0.522	0.236	0.894	0.047
MT-Bench, atom-fixed	0.863	0.035	0.338	0.216	0.762	0.054
MT-Bench, single-block	0.863	0.085	0.613	0.308	0.925	0.051
UltraFeedback (new), direct	0.303	0.391	0.047	0.498	0.858	0.053
UltraFeedback (new), atom-fixed	0.198	0.386	0.033	0.451	0.776	0.063
UltraFeedback (new), single-block	0.447	0.361	0.247	0.537	0.927	0.044
UltraFeedback (old), direct	0.272	0.394	0.057	0.478	0.830	0.065
UltraFeedback (old), atom-fixed	0.193	0.359	0.027	0.458	0.680	0.080
UltraFeedback (old), single-block	0.400	0.418	0.280	0.527	0.900	0.055

A.5 Post-hoc diagnostics

The tail analysis is outcome-dependent. It uses symmetric trimming within cohorts and a leave-one-item influence check to describe whether the aggregate excess is concentrated in a small number of observations. These checks are descriptive and do not change the primary estimand.

The mechanism analysis is also post-hoc. Singleton rows provide a scale calibration, and multi-atom residuals are summarized only on common support. The fitted surface is assessed with grouped cross-validation by complete item; leave-one-family-out fits describe sensitivity to semantic coverage. These diagnostics quantify conditional associations with score level, block size, and atom disagreement. They do not identify a latent evaluator operation or establish mediation or causality.

A.6 Data ethics and privacy

All analyses use existing public evaluation corpora. We collect no human-subject responses and no personal data, and no rubric was optimized against a target system. The released artifact contains the analysis code and derived scores, and no personal or human-subject data.

Table 9: Mechanism model diagnostics. The first four rows report grouped five-fold within-item held-out R^2 and RMSE; the Shapley rows report average held-out R^2 increments for the three feature groups; LOFO rows summarize held-out family robustness, with the per-family breakdown reported separately. These post-hoc descriptive models do not predict absolute item scores or establish causality.

Model or diagnostic	Within-item R^2 / increment	RMSE
Judge only	0.010	0.646
Scale	0.288	0.548
Scale + size	0.312	0.539
Scale + size + conflict	0.365	0.518
Shapley: scale	0.257	–
Shapley: size	0.025	–
Shapley: conflict	0.072	–
LOFO median R^2	0.267	–
LOFO minimum R^2	0.198	–

A.7 Review artifact and software boundary

The anonymized review artifact is available at https://anonymous.4open.science/r/offline-eval-7q2m/. It contains the frozen confirmatory_v1 contract, the exact user-message prompt templates with item placeholders, the derived score table, a deterministic offline analysis, verification scripts, and tests. Section 3 and this appendix specify the estimator, the matched null, and the pooling rules in full, so the reported results are reproducible by running the pipeline against the same three evaluation models. Evaluated outputs are experimental data and are separate from the assistance used to prepare the manuscript.