DECISION GEOMETRY OF PAIRWISE LLM EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Changing the labels around fixed answers can change an LLM judge's decisions. Interpreting the resulting score gain requires understanding which choices are made and how they combine across candidate orders. Our label-intervention studies show that similar correctness gains can arise from different changes in decisive coverage and alignment with the correct answer, with different implications as the cost of errors changes. We trace these effects from paired verdicts to their decision geometry. An invertible decomposition gives an exact attribution of correctness changes to coverage and signed alignment. A threshold model then explains how position preference and tying interact: a tie can reveal a preference that opposing votes would cancel, while a higher threshold can remove it again. The model yields a non-monotone selection boundary and a sign-invariance prediction under shared evidence. Response fits favor label-dependent position and tie parameters, and paired diagnostics distinguish observed cross-label patterns from both deterministic sign invariance and the higher reversal frequencies predicted by independent completion. Sharp coupling bounds characterize the paired energies compatible with each fitted marginal response law. This analysis follows an interface change through selection, aggregation and joint behavior to explain what changes behind its reported score. An anonymized code snapshot is available at `https://anonymous.4open.science/r/dgpllme/`.

1 INTRODUCTION

A pairwise LLM judge receives a question and two answers, then expresses a preference through a prescribed interface (Zheng et al., 2023; Tan et al., 2024). In our studies, the question and both answer strings remain unchanged. Replacing the external labels and allowed verdict tokens, A/B, with FIRST/SECOND, changes the judge's decisions and its reported correctness. This fixed-answer intervention lets us examine how the response process shapes a score: which comparisons receive a decisive choice, which candidate is selected, and how choices combine across orders.

Figure 1 shows why this interpretation matters. Qwen's correctness increases on both answer sets, yet the gains have different compositions. On GPT pairs, increased decisiveness and stronger alignment with the correct answer both contribute. On Claude pairs, expanded coverage accounts for the gain while signed alignment changes slightly in the opposite direction. Similar score improvements therefore have different values when errors carry a cost. We ask: *how does a response interface reshape selection and aggregation, and what do those changes mean for the resulting score?*

Prompt formatting can change model performance even when meaning is preserved (Sclar et al., 2024). Multiple-choice studies show effects of option order (Pezeshkpour & Hruschka, 2024) and token priors over option IDs (Zheng et al., 2024); PriDe uses permutations to estimate and remove those priors. Pairwise-judge studies establish sensitivity to presentation and answer order (Chen et al., 2024; Shi et al., 2025), motivating evaluation in both orders (Wang et al., 2024). These observations locate our question at the response interface. We follow a label change through individual judgments, paired records and aggregate preferences: tying changes which views contribute a vote; position sensitivity changes how those votes combine.

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.

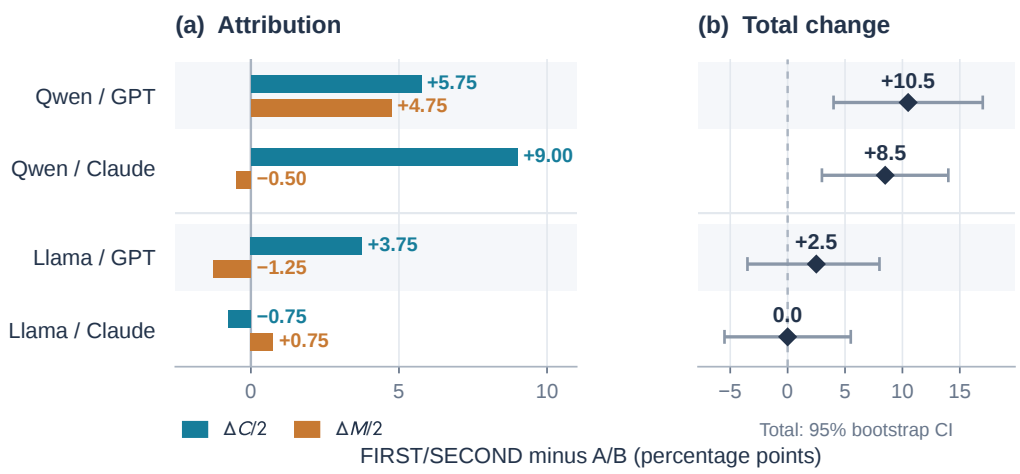

Figure 1: **Similar score gains can arise from different decisions.** (a) contributions from decisive coverage ($\Delta C/2$) and the correct-minus-incorrect rate ($\Delta M/2$). (b) their sum gives the change in correct/total judgments, with paired question-bootstrap 95% intervals. All changes are FIRST/SECOND minus A/B, in percentage points. The two Qwen gains have different alignment contributions.

Mapping verdicts back to candidate identity makes these effects visible. Choosing the first answer twice selects opposite candidates after an order swap. The average and half-difference separate stable preference from position sensitivity. The objective label supplies correctness; separate tie and validity records preserve selection on the full trial denominator.

To explain how these choices arise, we combine scalar evidence with a position offset and a symmetric tie threshold. Opposite choices can cancel across orders. Raising the threshold first turns the weaker choice into a tie, revealing a preference, then makes both responses tie. This geometry also predicts that shared evidence preserves the sign of a nonzero aggregate preference across protocols, even with question-specific offsets and thresholds.

The frequency of each choice also leaves open how choices occur together. Sharp extremal-dependence bounds (Puccetti & Wang, 2015) characterize the compatible paired energies. Fitted response laws favor label-dependent position and tie parameters. Paired observations then test deterministic sign invariance and the reversal frequencies predicted by independent completion, connecting score attribution to testable accounts of the response process.

Our main contributions are:

- We decompose correctness changes into decisive coverage and signed alignment, revealing their implications for utility.
- We derive a non-monotone selection boundary and a sign-invariance prediction from a threshold model of paired judgments.
- We characterize paired energies through sharp coupling bounds and evaluate joint predictions against observed reversals.

2 From paired verdicts to decision geometry

2.1 Coordinates and observable accounting

Fix canonical candidates A and B for question i and an objective correctness label $g_i \in \{-1, +1\}$, with $+1$ identifying A. Let $x = +1$ display A first and $x = -1$ display A second. Under protocol a, a valid presented response $z_{i,a,x} \in \{-1, 0, +1\}$ selects the second candidate, ties, or selects the

first. The canonical decision $d_{i,a,x} = xz_{i,a,x}$ recovers identity. Suppressing i, a, write the paired decisions as d_+, d_-.

Let $V \in \{0, 1\}$ indicate a valid output. For scoring over all requested trials, we extend $d = 0$ to invalid outputs; they remain distinct from valid TIE responses in validity statistics and response fitting. Mechanism tests use questions with valid responses in both orders under both protocols.

Define the order-even and order-odd components

$$R = \frac{d_+ + d_-}{2}, \qquad P = \frac{d_+ - d_-}{2}. \tag{1}$$

Exchanging the two observations preserves R and reverses P. Their sum and difference recover the original observations. Positive P indicates a first-position tendency; gR records the orientation of R toward the correct candidate. A single tie transition contributes to both squared components.

Let C denote decisive coverage, A correct/total trials, and W incorrect/total trials. Expectations average questions with equal weight and both orders with weight $1/2$. Put $Q = \mathbb{E}[R^2]$, $B = \mathbb{E}[P^2]$ and $M = \mathbb{E}[gR]$, the signed alignment with the correctness label. Then

$$C = Q + B, \qquad A = \frac{C + M}{2}, \qquad W = \frac{C - M}{2}. \tag{2}$$

Using $R^2 + P^2 = (d_+^2 + d_-^2)/2$ and $\mathbf{1}\{d = g\} = (d^2 + gd)/2$ gives

$$\Delta A = \frac{\Delta C}{2} + \frac{\Delta M}{2}. \tag{3}$$

This identity attributes observed correctness changes to coverage and signed alignment on any paired response distribution.

With reward $+1$ for a correct choice, penalty $-\lambda$ for an incorrect choice and zero for TIE or invalid output, utility is

$$U_\lambda = A - \lambda W = \frac{1 - \lambda}{2} C + \frac{1 + \lambda}{2} M. \tag{4}$$

At equal error cost, $U_1 = M$; a separate invalid-output cost adds a validity term. Always choosing the first candidate gives $(d_+, d_-) = (+1, -1)$, $C = 1$, $A = W = 1/2$ and $M = U_1 = 0$, separating coverage from alignment.

2.2 A SCALAR-EVIDENCE MECHANISM

The accounting identifies how paired choices contribute to a score. To explain how those choices arise, let q_i be the judge's latent evidence for canonical A over B, with $g_i q_i$ its orientation toward the correct candidate. A position offset $b_{i,a}$ and symmetric threshold $\tau_{i,a} \geq 0$ turn this evidence into a verdict in each order:

$$T_\tau(s) = \begin{cases} +1, & s > \tau, \\ 0, & |s| \leq \tau, \\ -1, & s < -\tau, \end{cases} \qquad \begin{aligned} d_{i,a,+} &= T_{\tau_{i,a}}(q_i + b_{i,a}), \\ d_{i,a,-} &= T_{\tau_{i,a}}(q_i - b_{i,a}). \end{aligned} \tag{5}$$

The shared-evidence hypothesis keeps q_i fixed across orders and protocols. Offsets and thresholds may vary by question; the response law is deterministic conditional on these quantities.

Proposition 1 (Response regions)**.** *Away from $|q \pm b| = \tau$, both responses select the same canonical identity if $|q| > \tau + |b|$, select the same displayed position if $|b| > \tau + |q|$, and both tie if $\tau > |q| + |b|$. The remaining regions have one decisive response.*

Proposition 2 (Cross-protocol sign invariance)**.** *Under Equation 5, $R \neq 0$ implies $\operatorname{sign}(R) = \operatorname{sign}(q)$. Thus two protocols sharing q satisfy $R_a R_{a'} \geq 0$.*

Oddness and monotonicity preserve the sign of q in every nonzero R. A cross-protocol reversal directly tests this shared-evidence prediction.

Proposition 3 (Selection boundary)**.** *Away from threshold equalities,*

$$R \neq 0 \quad \Longleftrightarrow \quad |q| > \big||b| - \tau\big|. \tag{6}$$

A nonzero vote escapes either two ties or opposing decisive choices. Appendix A proves both cases and specifies the boundaries.

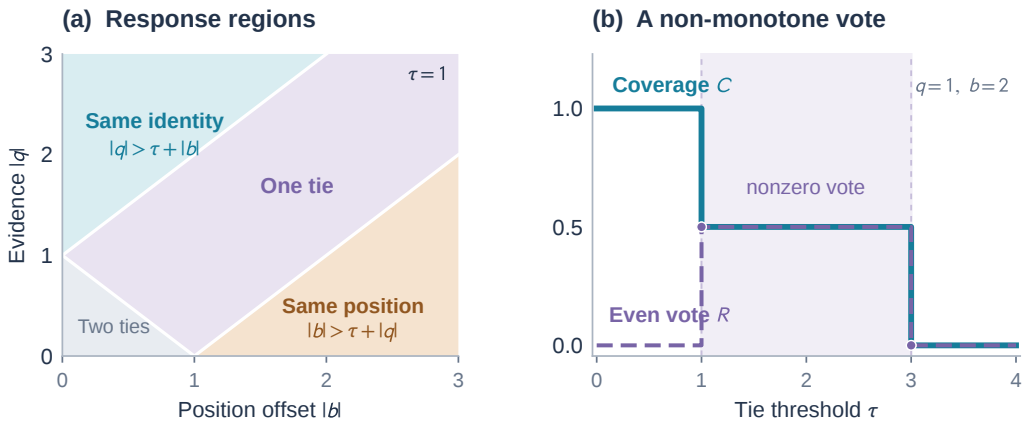

Figure 2: **Position and tying interact before aggregation.** Left: paired-response regions at $\tau = 1$. Right: at $q = 1, b = 2$, raising τ first recovers a nonzero order-even vote and later removes it. The panels visualize the response regions and selection law of the deterministic mechanism.

At $q = 1, b = 2$, thresholds 0, 1.5 and 4 give $(+1, -1)$, $(+1, 0)$ and $(0, 0)$: selection first appears and then disappears (Figure 2). At threshold zero, averaging the quantized scores yields zero, while quantizing their average yields $+1$. Thus the order of quantization and averaging determines which evidence survives aggregation.

2.3 Marginal scores and paired response coupling

The same single-order choice frequencies can yield different paired patterns. Their joint law contributes a cross-order moment $H = \mathbb{E}[d_+ d_-]$, giving

$$Q = \frac{C + H}{2}, \qquad B = \frac{C - H}{2}. \tag{7}$$

At fixed label orientation, each order's marginal response distribution fixes C, M, A and U_λ. The coupling supplies H, allocating coverage between even and odd energy. For example, let $g = +1$ and each order choose A or B with equal probability. Matching the two choices gives $Q = 1, B = 0$; always opposing them gives $Q = 0, B = 1$. Both have $C = 1, A = 1/2, M = 0$.

Proposition 4 (Sharp bounds under fixed marginals)**.** *Let F_+^{-1}, F_-^{-1} be the generalized quantile functions of the two canonical score marginals, with invalid output assigned zero only for scoring. Over all joint distributions with these marginals,*

$$\int_0^1 F_+^{-1}(u) F_-^{-1}(1-u)\, du \;\le H \le\; \int_0^1 F_+^{-1}(u) F_-^{-1}(u)\, du. \tag{8}$$

Both bounds are attainable and give sharp bounds for Q, B through Equation 7.

This applies classical extremal-dependence results (Puccetti & Wang, 2015); Appendix F gives the finite transport proof and computed ranges. Conditional independence supplies one possible joint distribution inside this family. Section 4.5 evaluates that completion against observed paired events.

2.4 Reporting profile and utility ordering

Let $v = \mathbb{E}[V]$ denote the validity rate. The profile (v, C, M, H) recovers correct and incorrect rates as $(C + M)/2$ and $(C - M)/2$, the valid tie rate as $v - C$, and invalidity as $1 - v$. Conditional accuracy is $(C + M)/(2C)$ for $C > 0$, while the paired energies are $(C + H)/2$ and $(C - H)/2$. The profile obeys $|M| \le C \le v \le 1$ and $|H| \le C$. The first bounds follow from $M = A - W$, $C = A + W$ and the inclusion of decisive responses among valid responses; the second follows

from $|d_+ d_-| \leq (d_+^2 + d_-^2)/2$ and the triangle inequality. At fixed coverage, increasing M transfers mass from incorrect to correct decisions; increasing H transfers energy from the order-odd to the order-even component.

A protocol has at least as much utility as another for every $\lambda \geq 0$ exactly when $\Delta A \geq 0$ and $\Delta W \leq 0$, or equivalently $\Delta M \geq |\Delta C|$. To see this, write $\Delta U_\lambda = \Delta A - \lambda \Delta W$: nonnegativity for every nonnegative cost requires a nonnegative intercept and slope. Substituting $\Delta A = (\Delta C + \Delta M)/2$ and $\Delta W = (\Delta C - \Delta M)/2$ gives the stated criterion. For a specified finite cost interval, the affine utility difference attains its minimum at an endpoint. Applied to empirical counts, these relations compare estimated utility curves; paired question resampling supplies the sampling assessment.

3 Study design and quantitative response models

Data and intervention. We reuse response pairs and objective correctness labels from JudgeBench (Tan et al., 2024). The main study uses all 100 GPT-generated pairs remaining after a 50-question pilot and 200-question expansion. A second study selects 100 Claude-generated pairs with question text disjoint from all 350 GPT pairs. Both input sets and their respective protocols were fixed before the corresponding study. Public parent judgments informed the preceding exploratory phase. Appendix B gives the selection chronology and dataset construction.

Each question receives A/B and First/Second candidate/verdict labels in both orders. Only external candidate headers and allowed verdict vocabulary change; question and answer strings, including embedded task letters, are preserved. Qwen3-8B and Llama-3.2-3B-Instruct each produce 400 judgments per set, for 1,600 total. Each judge uses the same response protocol across label arms. All requested trials enter the analysis; responses outside the prescribed verdict format form a separate invalid category. Earlier five-view experiments appear in Appendix G.

Primary analysis and chronology. We report correctness over all requested trials, decisive coverage, validity, and accuracy conditional on a decisive response. Paired intervals resample questions 5,000 times. Exact sign-flip tests operate on each question's two-order counts. Four primary tests per study—two judges by validity and correctness—receive Holm correction within that study. The decomposition, utility sensitivity and response-model family were designed after inspecting both datasets. The following retrospective explanatory comparisons use the same definitions across studies.

A marginal response model. The deterministic model gives question-level structural predictions. To compare population response patterns, we fit ordered logistic regression (McCullagh, 1980) to valid presented verdicts. With $e = xg$ indicating whether the correct candidate is first, let

$$z = T_{\tau_a}(\alpha_a e + \beta_a + \epsilon), \qquad \epsilon \sim \text{Logistic}(0, 1), \quad \tau_a > 0. \tag{9}$$

Writing $\mu = \alpha_a e + \beta_a$ and F for the logistic CDF, probabilities of second, Tie and first are $F(-\tau_a - \mu)$, $F(\tau_a - \mu) - F(-\tau_a - \mu)$ and $1 - F(\tau_a - \mu)$. Alpha summarizes label alignment, beta is a first-position intercept, and tau defines a symmetric tie interval on a fixed scale. This pooled model describes conditional marginal responses through population coefficients on a fixed logistic scale. Section 4.5 adds an explicit response coupling.

A separate validity gate $\pi_{a,m} = (n_{\text{valid}} + 1/2)/(n+1)$ uses label arm a and MMLU-Pro versus other sources m, estimated on training data only. It multiplies the three valid-response probabilities by π and assigns invalid probability $1 - \pi$. This gate is identical across five response structures: shared α, β, τ (S); arm-specific β (B); arm-specific τ (T); bias-plus-threshold (BT), with arm-specific β, τ; and all three arm-specific (F). The shared response structure thus includes label-dependent validity through the gate.

Five-fold cross-validation keeps all four responses to a question together. Held-out four-category negative log likelihood (NLL) scores every requested trial, and bootstrap intervals for loss differences condition on the fitted models. Gold orientation serves as a known explanatory covariate; NLL measures response fit conditional on that orientation. We report all structures, fold diagnostics and transfer directions. The retrospective comparison completes the fitted marginal distributions by conditional independence and evaluates paired events using the same fits (Appendix F).

4 WHAT CHANGES WHEN CANDIDATE LABELS CHANGE?

4.1 A REPLICATED SCORE DIRECTION WITH DIFFERENT COMPONENTS

Judge / pairs	Correct	Decisive	ΔA	95% CI	Holm p
Qwen / GPT	78→99	129→152	$+10.5$	$[4.0, 17.0]$	0.0117
Qwen / Claude	76→93	122→158	$+8.5$	$[3.0, 14.0]$	0.0213
Llama / GPT	41→46	75→90	$+2.5$	$[-3.0, 8.5]$	1.0000
Llama / Claude	34→34	71→68	$+0.0$	$[-5.5, 5.5]$	1.0000

Table 1: **Label-intervention results.** AB → FS counts use 200 trials per arm (100 questions, two orders). Differences and intervals are percentage points; tests preserve question pairing. Qwen denotes Qwen3-8B and Llama denotes Llama-3.2-3B-Instruct.

Changing A/B to FIRST/SECOND increases Qwen's all-trial correctness by 10.5 points on GPT pairs and 8.5 points on Claude pairs (Table 1). Question-bootstrap intervals are $[4, 17]$ and $[3, 14]$ points, with Holm-adjusted $p = .0117$ and $.0213$. Llama's corresponding changes are $+2.5$ points and zero, with intervals $[-3, 8.5]$ and $[-5.5, 5.5]$ points.

Equation 3 separates these effects (Figure 1). For Qwen/GPT, $\Delta A = 5.75 + 4.75 = 10.5$ points: increased coverage and stronger signed alignment both contribute. For Qwen/Claude, $\Delta A = 9.00 - 0.50 = 8.5$ points: coverage contributes more than the total gain. Conditional accuracy moves from 60.5% to 65.1% in the former and from 62.3% to 58.9% in the latter.

The second Qwen set produces 17 additional correct and 19 additional incorrect judgments. Its utility difference is therefore $\Delta U_\lambda = (17-19\lambda)/200$, crossing zero at the point estimate $\lambda = 17/19$. At equal error cost, the changes in $U_1 = M$ are $+9.5$ and -1.0 points across the two sets. The estimated utility curve translates these counts into a protocol comparison at each error cost; its sampling interpretation follows the paired design. For Llama/GPT the accuracy decomposition is $3.75 - 1.25 = 2.5$ points; for Llama/Claude it is $-0.75 + 0.75 = 0$.

4.2 EVEN ENERGY AND VALIDITY SUPPORT REFINE THE EXPLANATION

For Qwen/GPT, even energy Q rises from $.2175$ to $.4150$, while odd energy B falls from $.4275$ to $.3450$. For Qwen/Claude, both increase: Q from $.2800$ to $.4000$ and B from $.3300$ to $.3900$, even as M falls from $.150$ to $.140$. Even energy measures the magnitude of order-stable decisions; M supplies their signed orientation toward the correct label. Llama's even energy decreases and odd energy increases in both sets (Appendix C).

Full-denominator scoring can include asymmetric invalidity, which we separate through common support. The four groups, Qwen/GPT, Qwen/Claude, Llama/GPT and Llama/Claude, have 98, 99, 54 and 34 questions with all four responses valid. Splitting the accuracy effect into contributions from these questions and the complement gives $9.5 + 1.0$, $8.0 + 0.5$, $5.0 - 2.5$ and $4.5 - 4.5$ points, respectively, always using the full 100-question denominator. For Llama/Claude, the 4.5-point contribution from jointly valid questions and the -4.5-point contribution from the complement exactly balance. This partition attributes the full protocol effect to explicit validity supports.

4.3 POSITION AND TIE PARAMETERS CAPTURE MARGINAL RESPONSE CHANGES

Allowing both β and τ to vary improves grouped predictive NLL over the shared valid-response law in every group (Table 2). The improvements are $.0235$, $.0205$, $.0290$ and $.0772$ nats per judgment, with fixed-fit bootstrap intervals $[.0079, .0387]$, $[.0041, .0378]$, $[.0010, .0578]$ and $[.0451, .1094]$. Full flexibility is slightly better for Qwen/GPT. Figure 3 displays the fitted decision regions.

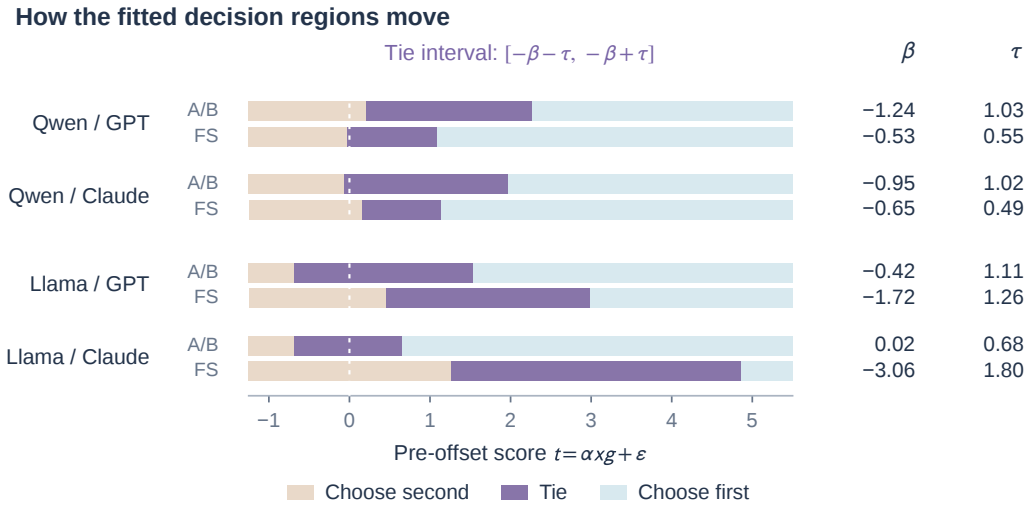

Figure 3: **Label effects shift and resize the tie interval.** Bands show the fitted response rule on the score $t = \alpha xg + \epsilon$. The tie interval $[-\beta - \tau, -\beta + \tau]$ has center $-\beta$ and width 2τ; scores below/above it select the second/first candidate. Qwen's band contracts; Llama's shifts right and widens. All bands use the same logistic scale.

Judge / pairs	Shared	Bias	Threshold	BT	Full
Qwen / GPT	1.0155	1.0089	1.0059	0.9920	0.9916
Qwen / Claude	1.0315	1.0323	1.0119	1.0109	1.0122
Llama / GPT	1.0932	1.0649	1.0904	1.0643	1.0647
Llama / Claude	0.8860	0.8221	0.8876	0.8088	0.8096

Table 2: **Grouped held-out response NLL** (nats/judgment; lower is better). BT allows arm-specific β, τ with shared α. All structures use the same label-dependent validity gate.

Qwen's fitted threshold contracts from 1.032 to $.554$ on GPT pairs and 1.019 to $.487$ on Claude pairs. Llama instead exhibits a stronger second-position intercept: β changes from $-.421$ to -1.722 and $.016$ to -3.060, while τ increases from 1.108 to 1.263 and $.675$ to 1.799 (Figure 3). Tie frequency is jointly determined by the evidence distribution, position offset and interval width; a shift away from the interval can reduce ties while the threshold expands.

Full-data fits transferred between generators favor the BT structure over shared responses in all four directions, with some simpler structures performing better still. Because both sets were inspected when choosing the family, these are retrospective transfer diagnostics. All losses, coefficients and fold ranges appear in Appendix D, resolving how each parameter contributes across sources.

4.4 Cross-protocol signs distinguish response mechanisms

Proposition 2 applies on common-valid questions with nonzero R under both protocols. The sign-reversal counts are 2 among 28 assessable Qwen/GPT questions, 1/36 for Qwen/Claude, 0/4 for Llama/GPT and 1/3 for Llama/Claude. One Qwen/GPT pair changes from (Tie, A) to (Tie, B), giving $R = +1/2$ and $R = -1/2$. Appendix E describes all four cases and their paired choices.

The observed reversals distinguish these questions from the predictions of deterministic shared evidence, including its version with question-specific offsets and thresholds. They provide concrete targets for mechanisms with protocol-dependent evidence, asymmetric thresholds or view-specific variation. The diagnostic records event prevalence on each explicit eligible support, including the

four Llama/GPT questions. Equation 9 assigns probability to each category; the paired record additionally specifies which responses occur together on the same question.

4.5 JOINT DIAGNOSTICS CHARACTERIZE CROSS-VIEW RESPONSE STRUCTURE

We next ask whether the fitted choices combine into the observed cross-label reversals. Completing each fitted marginal model by conditional independence, given covariates and fitted parameters, specifies its four-response joint law. Enumerating all 4^4 outcomes per question gives exact expectations. This retrospective check evaluates the marginal predictions together with the added dependence assumption.

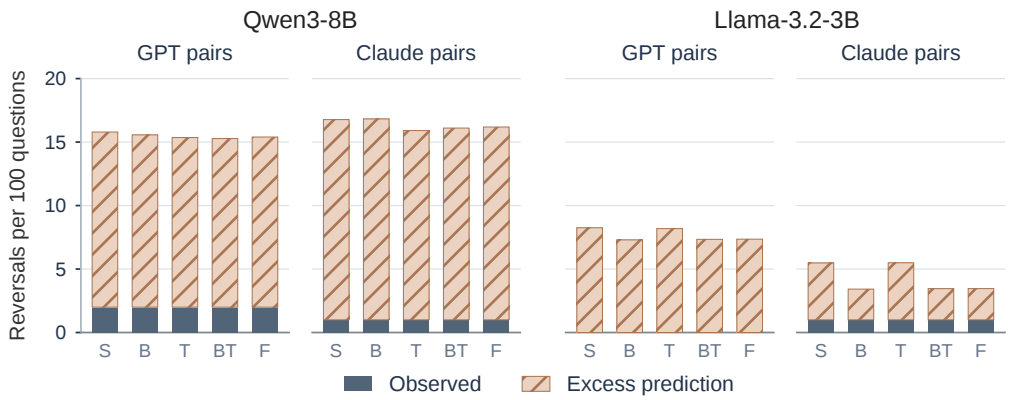

Figure 4: **All five independent completions predict more reversals than observed.** Full bars give expected events per 100 questions; solid bases give observed counts $(2, 1, 0, 1)$, and hatching gives the prediction excess. S/B/T/BT/F denote shared, bias, threshold, bias-plus-threshold and full models. Events require four valid responses and opposite nonzero R signs across labels. Exact values are in Table 13.

The BT completion predicts 15.28 and 16.10 cross-label sign reversals per 100 Qwen questions, versus the observed 2 and 1 (Figure 4). Its expected assessable supports are 36.15 and 36.41, compared with 28 and 36 observed. In the Claude set, the eligible count is closely matched, and the reversal count provides an additional, discriminating check of joint structure. The corresponding Llama expectations are 7.34 and 3.46, versus 0 and 1. For Qwen, predicted reversals remain 15.28–15.79 and 15.91–16.82 across all five marginal structures. The prediction excess is consistent across the five response structures.

The diagnostic positions the observed pattern between two concrete reference accounts: sign invariance under shared deterministic evidence and higher reversal frequencies under fitted independent completion. It evaluates the combined marginal and dependence model, with item heterogeneity, cross-view dependence and marginal fit all contributing to that description. Equation 7 connects this joint perspective to the paired energies. Table 3 shows the coupling freedom of the BT marginals. Each row fixes the fitted category probabilities, then varies how the two orders are paired. The product law gives one value of Q inside the sharp feasible interval. For Qwen/GPT under FIRST/SECOND, for example, that value is $.3677$, while the compatible range is $[.097, .635]$. Observed Q is $.4150$. The table supplies the corresponding odd-energy ranges, with their allocation governed by the cross-order moment. These bounds describe paired energies within each protocol; the reversal diagnostic evaluates a completion across protocols. Both distinguish what the marginals determine from what the response coupling supplies. Appendix F gives the construction, all completions and uncertainty conventions.

Judge / pairs	Arm	Observed Q	Product Q	Feasible Q	Feasible B
Qwen / GPT	AB	0.2175	0.2330	$[0.073, 0.378]$	$[0.261, 0.566]$
Qwen / GPT	FS	0.4150	0.3677	$[0.097, 0.635]$	$[0.117, 0.655]$
Qwen / Claude	AB	0.2800	0.2495	$[0.066, 0.429]$	$[0.176, 0.538]$
Qwen / Claude	FS	0.4000	0.3639	$[0.080, 0.605]$	$[0.178, 0.703]$
Llama / GPT	AB	0.1025	0.1797	$[0.016, 0.317]$	$[0.057, 0.358]$
Llama / GPT	FS	0.0650	0.1329	$[0.011, 0.206]$	$[0.243, 0.438]$
Llama / Claude	AB	0.1025	0.1777	$[0.010, 0.341]$	$[0.015, 0.346]$
Llama / Claude	FS	0.0250	0.0682	$[0.004, 0.086]$	$[0.253, 0.335]$

Table 3: Paired-energy compatibility under fixed fitted marginals of the bias-plus-threshold model. The mathematical ranges cover all two-order couplings of those fixed probabilities; bounds are computed per question and then averaged.

5 Related work

LLM Judge Protocols. Balanced Position Calibration averages scores before choosing a winner (Wang et al., 2024); JudgeBench combines signed decisions (Tan et al., 2024). Our quantization result explains why aggregation order matters. Shi et al. (2025) distinguish repetition stability, position consistency and preference fairness, explicitly include ties, and relate position sensitivity to answer-quality gaps. Their quality proxy uses aggregate win rates; our q represents latent judge evidence in a threshold response law. Length-Controlled AlpacaEval adjusts preferences for response length (Dubois et al., 2024). We hold response content fixed and trace how the interface changes coverage and the alignment of decisive choices. Prometheus 2 learns direct and pairwise evaluation (Kim et al., 2024), and OffsetBias improves robustness through debiased preference data (Park et al., 2024). Our fixed-answer intervention examines the response interface, connecting instruction adherence (Murugadoss et al., 2024) and protocol robustness (Li et al., 2025) to selection and aggregation.

Consistency and Selection. PoLL aggregates diverse model judges (Verga et al., 2024); FESTA estimates uncertainty across equivalent and complementary multimodal inputs (Bhattacharya et al., 2025). Selective evaluation calibrates trust against human agreement (Jung et al., 2025), and active calibration uses human clarification (Li et al., 2026). We study paired views of one judge against objective labels, making joint dependence and the contribution of ties explicit. Our utility accounting assigns each decisive choice its value at a specified error cost, connecting the selection of comparisons to the interpretation of a score.

Statistical Foundations. Reject-option theory (Chow, 1970), selective classification (Geifman & El-Yaniv, 2017) and LLM abstention (Wen et al., 2025) connect coverage to utility. Paired-comparison and tie models (Bradley & Terry, 1952; Davidson, 1970), ordinal regression (McCullagh, 1980) and extremal coupling (Puccetti & Wang, 2015) provide the mathematical foundations for our attribution, threshold geometry and joint-response analysis.

6 Discussion

The label intervention leaves the answers fixed while changing the choices made around them. Coverage and signed alignment explain why similar correctness gains carry different utility at different error costs. The threshold geometry then shows how a preference can emerge from cancellation as one response becomes a tie. Paired observations test this geometry's sign prediction and the reversal frequencies of fitted independent responses.

Following the same answers across labels and orders makes these connections visible. Candidate identity links the views, the objective label orients their choices, and separate records of valid ties and invalid outputs show which comparisons receive a decisive judgment. The score at the end of this process can then be read through the decisions that compose it: when the judge chooses, which candidate it favors, and which preferences survive aggregation.

Reproducibility statement

Section 3 specifies the paired design, score denominators and response-model comparison. Appendices A and F give the proofs; Appendices B and D define the study samples, response categories, estimation objective and uncertainty calculations. Complete counts, parameter estimates and paired-response examples accompany the main results in Appendices C–F.

AI use statement

We used generative AI tools to aid or polish writing. The authors take responsibility for the final paper.

References

Debarpan Bhattacharya, Apoorva Kulkarni, and Sriram Ganapathy. FESTA: Functionally equivalent sampling for trust assessment of multimodal LLMs. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 12277–12295, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.657. URL https://aclanthology.org/2025.findings-emnlp.657/.

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. doi: 10.1093/biomet/39.3-4.324. URL https://doi.org/10.1093/biomet/39.3-4.324.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. Humans or LLMs as the judge? a study on judgement bias. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8301–8327, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.474. URL https://aclanthology.org/2024.emnlp-main.474/.

C. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970. doi: 10.1109/TIT.1970.1054406. URL https://doi.org/10.1109/TIT.1970.1054406.

Roger R. Davidson. On extending the bradley-terry model to accommodate ties in paired comparison experiments. *Journal of the American Statistical Association*, 65(329):317–328, 1970. doi: 10.1080/01621459.1970.10481082. URL https://doi.org/10.1080/01621459.1970.10481082.

Yann Dubois, Percy Liang, and Tatsunori Hashimoto. Length-controlled alpacaeval: A simple debiasing of automatic evaluators. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=CybBmzWBX0.

B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/4a8423d5e91fda00bb7e46540e2b0cf1-Paper.pdf.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979. URL https://www.jstor.org/stable/4615733.

Jaehun Jung, Faeze Brahman, and Yejin Choi. Trust or escalate: LLM judges with provable guarantees for human agreement. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UHPnqSTBPO.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 4334–4353, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.248. URL https://aclanthology.org/2024.emnlp-main.248/.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 2757–2791, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.138. URL https://aclanthology.org/2025.emnlp-main.138/.

Pengqi Li, Lizhong Ding, Zhehao Zhou, Chunhui Zhang, Jiarun Fu, Hao Li, Ye Yuan, and Guoren Wang. Demystifying uncertainty in LLMs: Active calibration between concepts and human evaluations. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 5726–5759, San Diego, California, United States, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.259. URL https://aclanthology.org/2026.acl-long.259/.

Peter McCullagh. Regression models for ordinal data. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 42(2):109–127, 1980. doi: 10.1111/j.2517-6161.1980.tb01109.x. URL https://doi.org/10.1111/j.2517-6161.1980.tb01109.x.

Bhuvanashree Murugadoss, Christian Poelitz, Ian Drosos, Vu Le, Nick McKenna, Carina Suzana Negreanu, Chris Parnin, and Advait Sarkar. Evaluating the evaluator: Measuring LLMs' adherence to task evaluation instructions, 2024. URL https://arxiv.org/abs/2408.08781.

Junsoo Park, Seungyeon Jwa, Meiying Ren, Daeyoung Kim, and Sanghyuk Choi. OffsetBias: Leveraging debiased data for tuning evaluators. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1043–1067, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.57. URL https://aclanthology.org/2024.findings-emnlp.57/.

Pouya Pezeshkpour and Estevam Hruschka. Large language models sensitivity to the order of options in multiple-choice questions. In Kevin Duh, Helena Gomez, and Steven Bethard (eds.), *Findings of the Association for Computational Linguistics: NAACL 2024*, pp. 2006–2017, Mexico City, Mexico, June 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-naacl.130. URL https://aclanthology.org/2024.findings-naacl.130/.

Giovanni Puccetti and Ruodu Wang. Extremal dependence concepts. *Statistical Science*, 30(4), 2015. doi: 10.1214/15-STS525. URL https://doi.org/10.1214/15-STS525.

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. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2310.11324.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In Kentaro Inui, Sakriani Sakti, Haofen Wang, Derek F. Wong, Pushpak Bhattacharyya, Biplab Banerjee, Asif Ekbal, Tanmoy Chakraborty, and Dhirendra Pratap Singh (eds.), *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314, Mumbai, India, December

2025. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics. ISBN 979-8-89176-298-5. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, Willian Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges, 2024. URL https://arxiv.org/abs/2410.12784.

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, 2024. URL https://arxiv.org/abs/2404.18796.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi: 10.1080/01621459.1927.10502953. URL https://doi.org/10.1080/01621459.1927.10502953.

Chujie Zheng, Hao Zhou, Fandong Meng, Jie Zhou, and Minlie Huang. Large language models are not robust multiple choice selectors. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=shr9PXz7T0.

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, 2023. URL https://arxiv.org/abs/2306.05685.

A PROOFS AND BOUNDARY CONVENTIONS

A.1 OBSERVABLE IDENTITIES

For $d \in \{-1, 0, +1\}$ and $g \in \{-1, +1\}$,

$$\mathbf{1}\{d = g\} = (d^2 + gd)/2, \qquad \mathbf{1}\{d = -g\} = (d^2 - gd)/2. \tag{10}$$

Averaging over the two orders gives correct rate $(C + M)/2$ and incorrect rate $(C - M)/2$. Expanding the squares in Equation 1 gives

$$R^2 + P^2 = \frac{(d_+ + d_-)^2 + (d_+ - d_-)^2}{4} = \frac{d_+^2 + d_-^2}{2}. \tag{11}$$

The right-hand side is the fraction of the two trials with a decisive output. Averaging over questions proves Equation 2; substitution in $U_\lambda = A - \lambda W$ proves Equation 4. The proof also holds under the stated scoring extension for invalid outputs because their score is zero. Valid ties and invalid outputs retain separate response categories.

A.2 THRESHOLD RESPONSE REGIONS

Proof of Proposition 1. The threshold is odd and monotone. Two positive canonical choices require $q + b > \tau$ and $q - b > \tau$, equivalent to $q > \tau + |b|$. Two negative choices similarly require $q < -\tau - |b|$. Thus both orders make the same decisive canonical choice exactly when $|q| > \tau + |b|$.

Opposing decisive canonical choices require one of $q + b, q - b$ to exceed τ and the other to fall below $-\tau$. If $b > 0$ this is $b > \tau + q$ and $b > \tau - q$, or $b > \tau + |q|$. The negative-b case gives $|b| > \tau + |q|$. Mapping back to presentation shows that these choices select the same displayed position.

Two ties occur if and only if $\max(|q + b|, |q - b|) \le \tau$. Since this maximum is $|q| + |b|$, the exact closed region is $\tau \ge |q| + |b|$. Outside these regions and the threshold-equality lines, exactly one response is decisive. □

The open-region statement in the main text excludes the lines $|q \pm b| = \tau$. At such a line, any response exactly at threshold is a tie, including the case $\tau = 0, s = 0$. At $\tau = |q| + |b|$ both responses are ties.

A.3 SIGN INVARIANCE AND NON-MONOTONE SELECTION

Proof of Proposition 2. For $q \ge 0$, $q + b \ge b - q$, so monotonicity and oddness imply $T_\tau(q + b) \ge T_\tau(b - q) = -T_\tau(q - b)$, hence $R \ge 0$. For $q \le 0$ the inequality reverses and $R \le 0$. At $q = 0$ oddness yields $R = 0$. Thus nonzero R has the sign of q. For two protocols sharing q, any nonzero R values share its sign, independent of their offsets and nonnegative thresholds. □

The sign criterion holds for item-specific b and τ under common evidence and deterministic symmetric thresholding. It is evaluated directly from canonical verdicts, making the prediction observable.

Proof of Proposition 3. Set $h = |q|$ and $k = |b|$. If $k \le \tau$, two position-dominated decisive choices are impossible. The zero-vote region is then two ties, $h + k < \tau$ away from equality. Its complement gives $h > \tau - k$. If $k > \tau$, two ties are impossible in the open region. A zero vote occurs when the two choices follow the same displayed position, $k > \tau + h$. Its complement gives $h > k - \tau$. These cases combine into $h > |k - \tau|$, excluding threshold equalities as stated. □

At $q = 1, b = 2$, the average of latent canonical scores is $((q + b) + (q - b))/2 = 1$. With $\tau = 0$, their quantized average is $(T_0(3) + T_0(-1))/2 = 0$, while the quantized average score is $T_0(1) = 1$. At $\tau = 1.5$ the two quantized responses become $(1, 0)$; at $\tau = 4$ they become $(0, 0)$. The nonzero symmetrized vote consequently appears and disappears as the tie threshold increases, illustrating the selection boundary in Proposition 3.

B Data construction and protocol details

B.1 Study samples and analysis chronology

JudgeBench (Tan et al., 2024) supplies the response pairs and objective labels. The main intervention uses the 100 GPT-4o pairs remaining after the pilot and exploratory expansion: 40 MMLU-Pro, 31 LiveBench-reasoning, 19 LiveBench-math and 10 LiveCodeBench questions. The replication uses Claude-3.5-Sonnet pairs. Excluding every question in the full 350-pair GPT collection and retaining one pair per question leaves 178 distinct questions, from which 100 are sampled without stratification or selection by response validity or correctness. Thus the two studies have disjoint question text and jointly vary the answer generator and question distribution.

Public parent judgments informed the exploratory phase. The response format was chosen after the pilot; the first label intervention followed the exploratory expansions, and the replication followed the first intervention. Each intervention was specified before its corresponding evaluations. The explanatory model family was specified after both intervention datasets were available, so its comparisons are retrospective.

B.2 Label intervention and response categories

The two label arms differ only in the external candidate and verdict vocabulary; question and answer content is held fixed. Each question is presented in both orders. Responses are assigned to the three valid categories (first, second, and tie) when they satisfy the prespecified response format; all other outputs form a separate invalid category. Presented positions are then mapped back to canonical candidate identity before computing order-paired statistics. A valid response consists of one verdict tag containing A, B or TIE under A/B, and FIRST, SECOND or TIE under FIRST/SECOND. Exterior whitespace and capitalization do not affect validity; additional text, multiple verdicts, unrecognized labels and empty responses are invalid. Both arms use the same instruction to choose the more factually and logically correct answer and to tie when the answers are equally correct. Question and answer text, including embedded task letters, is presented in full. The same evaluation rule is used for both judges and both studies.

B.3 Primary uncertainty calculations

Question-level bootstrap intervals (Efron, 1979) use 5,000 resamples, keeping both orders and both label arms together. Percentile intervals are unadjusted and recompute conditional denominators within each resample. For validity and correctness, exact two-sided sign-flip tests use per-question differences in the two-order counts. Under the exchangeability null, the two arms' score vectors can be swapped within each question. Holm correction (Holm, 1979) applies to the four primary tests within each study. Source and secondary comparisons are descriptive. The retrospective decomposition and response diagnostics use the same question-level resampling principle.

C Complete accounting and common-support results

Tables 4–7 retain all denominators and both judges. Define common support S by validity in both orders of both arms. For any question-level statistic f with paired difference Δf_i,

$$\Delta\overline{f} = \frac{1}{n}\sum_i \mathbf{1}\{i \in S\}\Delta f_i + \frac{1}{n}\sum_i \mathbf{1}\{i \notin S\}\Delta f_i. \tag{12}$$

The two contributions share the full denominator $n = 100$. A mean conditional on S has denominator $|S|$ and is a different quantity. We report both conditional means and full-denominator contributions. Under zero scoring of invalid responses, the odd component can reflect asymmetric validity as well as position preference and tie transitions. The common-support partition locates the contributions associated with validity, while P retains its definition as an observable order-odd component.

Judge / pairs	Arm	Valid	Correct	Decisive	Cond. accuracy
Qwen / GPT	AB	198	78	129	60.5%
Qwen / GPT	FS	200	99	152	65.1%
Qwen / Claude	AB	199	76	122	62.3%
Qwen / Claude	FS	200	93	158	58.9%
Llama / GPT	AB	146	41	75	54.7%
Llama / GPT	FS	136	46	90	51.1%
Llama / Claude	AB	105	34	71	47.9%
Llama / Claude	FS	86	34	68	50.0%

Table 4: Complete response counts, each with 200 requested trials per arm. Conditional accuracy divides correct by decisive outputs; valid TIEs contribute to validity.

Judge / pairs	ΔA	$\Delta C/2$	$\Delta M/2$	ΔU_1
Qwen / GPT	$+10.50$	$+5.75$	$+4.75$	$+9.50$
Qwen / Claude	$+8.50$	$+9.00$	-0.50	-1.00
Llama / GPT	$+2.50$	$+3.75$	-1.25	-2.50
Llama / Claude	$+0.00$	-0.75	$+0.75$	$+1.50$

Table 5: Exact attribution in percentage points of requested judgments. The middle two columns sum to the accuracy change.

Judge / pairs	Arm	C	$Q = \mathbb{E}[R^2]$	$B = \mathbb{E}[P^2]$	M
Qwen / GPT	AB	0.6450	0.2175	0.4275	0.1350
Qwen / GPT	FS	0.7600	0.4150	0.3450	0.2300
Qwen / Claude	AB	0.6100	0.2800	0.3300	0.1500
Qwen / Claude	FS	0.7900	0.4000	0.3900	0.1400
Llama / GPT	AB	0.3750	0.1025	0.2725	0.0350
Llama / GPT	FS	0.4500	0.0650	0.3850	0.0100
Llama / Claude	AB	0.3550	0.1025	0.2525	-0.0150
Llama / Claude	FS	0.3400	0.0250	0.3150	0.0000

Table 6: Full-denominator paired components. Invalid responses receive score zero and form a separate response category.

Judge / pairs	Common	Other	Common ΔA	Other ΔA
Qwen / GPT	98	2	$+9.50$	$+1.00$
Qwen / Claude	99	1	$+8.00$	$+0.50$
Llama / GPT	54	46	$+5.00$	-2.50
Llama / Claude	34	66	$+4.50$	-4.50

Table 7: Common-four-valid support and its complement. Contributions are percentage points using the full 100-question denominator, so they sum without reweighting.

D Response fitting, uncertainty and transfer

D.1 Estimation and grouped prediction

The five response structures have 3, 4, 4, 5 and 6 response parameters, respectively, together with the same four validity probabilities. They minimize the mean negative log likelihood of valid training

responses plus $0.001\|\theta\|_2^2$, where θ contains the free α, β and $\log\tau$ coefficients. Estimation restricts alignment and position coefficients to $[-12, 12]$ and $\log\tau$ to $[-6, 4]$. All fitted coefficients lie in the interior of these ranges.

Five folds each contain 80 training questions and 20 held-out questions. All four responses to a question stay together. Validity probabilities use only the corresponding training observations, with the smoothing rule in Section 3. Held-out likelihood includes all four response categories and is conditional on the known gold orientation and source group. Full-study coefficients use the same estimation objective; transfer evaluates them on the other response collection.

Judge / pairs	Shared α	β: AB $\to$ FS	τ: AB $\to$ FS
Qwen / GPT	0.450	$-1.239 \to -0.534$	$1.032 \to 0.554$
Qwen / Claude	0.375	$-0.952 \to -0.647$	$1.019 \to 0.487$
Llama / GPT	0.118	$-0.421 \to -1.722$	$1.108 \to 1.263$
Llama / Claude	-0.075	$0.016 \to -3.060$	$0.675 \to 1.799$

Table 8: Full-data coefficients for the bias-plus-threshold working model; residual logistic scale is fixed to one.

Judge / pairs	BT $-$ shared	Paired 95% CI	Fold NLL range
Qwen / GPT	-0.02353	$[-0.03875, -0.00786]$	$[0.9498, 1.0362]$
Qwen / Claude	-0.02053	$[-0.03776, -0.00406]$	$[0.9608, 1.0820]$
Llama / GPT	-0.02896	$[-0.05781, -0.00099]$	$[0.9644, 1.2342]$
Llama / Claude	-0.07718	$[-0.10935, -0.04511]$	$[0.6489, 0.9647]$

Table 9: Loss differences and paired 95% intervals from question bootstrap, conditional on the out-of-fold predictions and chosen model family. Fitted parameters are held fixed during resampling.

Judge / pairs	Arm	α range	β range	τ range
Qwen / GPT	AB	$[0.379, 0.561]$	$[-1.280, -1.207]$	$[0.985, 1.074]$
Qwen / GPT	FS	$[0.379, 0.561]$	$[-0.615, -0.470]$	$[0.519, 0.588]$
Qwen / Claude	AB	$[0.230, 0.507]$	$[-1.029, -0.851]$	$[0.921, 1.093]$
Qwen / Claude	FS	$[0.230, 0.507]$	$[-0.717, -0.583]$	$[0.460, 0.541]$
Llama / GPT	AB	$[0.020, 0.183]$	$[-0.689, -0.227]$	$[1.031, 1.247]$
Llama / GPT	FS	$[0.020, 0.183]$	$[-1.867, -1.569]$	$[1.169, 1.347]$
Llama / Claude	AB	$[-0.140, 0.036]$	$[-0.260, 0.156]$	$[0.569, 0.831]$
Llama / Claude	FS	$[-0.140, 0.036]$	$[-3.080, -3.042]$	$[1.754, 1.853]$

Table 10: Empirical parameter ranges across five training folds for the bias-plus-threshold model. Each interval records the minimum and maximum of the corresponding five fitted coefficients.

Train $\to$ test	Shared	Bias	Threshold	BT	Full
Qwen: Gpt $\to$ Claude	1.0187	1.0241	0.9992	1.0008	1.0023
Qwen: Claude $\to$ Gpt	1.0070	1.0051	0.9989	0.9890	0.9900
Llama: Gpt $\to$ Claude	0.9157	0.8604	0.9175	0.8561	0.8563
Llama: Claude $\to$ Gpt	1.1239	1.1036	1.1218	1.1115	1.1118

Table 11: Retrospective transfer NLL. Fit coefficients and validity gates on one set, then score all responses in the other. Both datasets were inspected before selecting the model family.

The loss intervals in Table 9 bootstrap questions with the out-of-fold predictions held fixed. They characterize question-sampling variation conditional on the fitted models and the chosen model family. Fold parameter ranges in Table 10 give the empirical minima and maxima of the five training fits. Table 11 reports all structures in both directions. Threshold-only attains the best transfer NLL for Qwen GPT-to-Claude, and bias-only does so for Llama Claude-to-GPT, resolving the parameter contributions by transfer direction.

D.2 Parameter interpretation and paired dependence

The unit logistic scale and centered symmetric threshold interval fix the coordinate system of the ordered model. Its coefficients summarize the conditional marginal response population, including the contributions of item difficulty and validity selection. Each judge is evaluated as a complete response source under the study protocol. The intervention jointly changes lexical form, ordinal meaning and overlap with embedded task letters; the protocol fixes the correspondence between verdict tokens and displayed positions. Thus the fitted coefficients describe the effect of this complete interface. The gold coefficient measures association with objective correctness orientation, and logistic residuals represent variation across the observed trials. Strict validity records compliance with the output contract; correctness records agreement with the inherited objective label.

Grouped validation keeps all four responses to a question in one fold. The response likelihood evaluates their marginal category probabilities. Joint completion adds a specified dependence structure and evaluates the resulting paired disagreements and cross-protocol sign events. This distinction assigns a concrete predictive target to each level of the model; Appendix F gives the completion and its evaluation.

E Complete cross-protocol sign diagnostics

Judge / pairs	Four valid	Assessable	Violations	Wilson 95% CI
Qwen / GPT	98	28	2	$[0.020, 0.226]$
Qwen / Claude	99	36	1	$[0.005, 0.142]$
Llama / GPT	54	4	0	$[0.000, 0.490]$
Llama / Claude	34	3	1	$[0.061, 0.792]$

Table 12: Cross-protocol sign-invariance diagnostics. Assessable questions have four valid responses and nonzero R in both arms. Violations count opposite signs; Wilson intervals quantify their prevalence on each reported assessable support.

The following lists every question with four valid responses and opposite nonzero order-even signs under the two protocols. Outputs refer to canonical candidate identity, listed as (original, swapped).

1. Qwen/GPT, LiveCodeBench, gold B: AB (Tie, A), FS (Tie, B); R changes from $+1/2$ to $-1/2$.
2. Qwen/GPT, MMLU-Pro biology, gold B: AB (Tie, B), FS (Tie, A); R changes from $-1/2$ to $+1/2$.
3. Qwen/Claude, MMLU-Pro business, gold B: AB (A, Tie), FS (B, B); R changes from $+1/2$ to -1.
4. Llama/Claude, LiveBench math, gold A: AB (A, Tie), FS (B, Tie); R changes from $+1/2$ to $-1/2$.

Wilson intervals (Wilson, 1927) in Table 12 quantify sign-reversal prevalence on each assessable support, including the Llama denominators of four and three. Each listed event distinguishes its question from the deterministic shared-evidence law. The complete event list provides reproducible targets for protocol-dependent evidence, asymmetric thresholds and stochastic response models.

F JOINT-RESPONSE ADEQUACY AND COUPLING BOUNDS

F.1 SHARP FINITE-SUPPORT BOUNDS

For fixed marginals on canonical scores $\{-1, 0, 1\}$, let γ_{jk} be a joint probability matrix with the specified row and column sums. The cross moment is $H = \sum_{jk} jk\gamma_{jk}$. Suppose $a < b$ and $c < d$. Moving mass ε from cells $(a, d), (b, c)$ to $(a, c), (b, d)$ preserves both marginals and changes the product moment by

$$\varepsilon(ac + bd - ad - bc) = \varepsilon(b - a)(d - c) \geq 0. \tag{13}$$

Repeated uncrossing attains the maximum by matching the two marginal quantiles in the same order. Matching them in opposite order attains the minimum. Both constructions have the required marginals and establish Proposition 4. This is the finite-support rearrangement argument underlying classical extremal-dependence bounds (Puccetti & Wang, 2015). Linear transformation through Equation 7 gives sharp bounds for Q and B.

The bounds follow the monotone and antitone rearrangements of the finite score distributions. They are computed question by question from the held-out marginal probabilities before averaging. Invalid probability is merged with zero only for these score bounds. Each range is the mathematically feasible set of a two-order energy conditional on the fitted marginals, with those probabilities held fixed. The next subsection specifies the four-response completion used for cross-arm events.

F.2 AN EXPLICIT INDEPENDENT COMPLETION

Let $p_{ij}(y)$ denote the fitted four-category predictive distribution for question i and view/arm $j \in \{1, 2, 3, 4\}$. We add the working assumption

$$p_i(y_1, y_2, y_3, y_4) = \prod_{j=1}^{4} p_{ij}(y_j). \tag{14}$$

Conditional independence here is given the explanatory covariates and fitted training-fold parameters. For each question, the product law yields exact expectations for common validity, eligibility, reversals, paired moments and valid disagreements. All five response structures are retained.

A reversal event requires four valid outputs and $R_{\text{A/B}} R_{\text{FIRST/SECOND}} < 0$. Its rate uses all 100 questions as denominator, unlike the earlier fraction of sign reversals among assessable questions. The BT model's expected common-four-valid counts are $95.59, 96.38, 47.56, 30.22$, compared with observed $98, 99, 54, 34$ in Qwen/GPT, Qwen/Claude, Llama/GPT and Llama/Claude. Its expected assessable counts are $36.15, 36.41, 14.91, 6.98$, compared with $28, 36, 4, 3$ observed. All reversal expectations are in Table 13.

For each event, we bootstrap the question-level observed indicator minus its fixed held-out predicted probability, using 5,000 resamples. For the BT model's reversal event, the observed-minus-expected residuals are $-13.28, -15.10, -7.34, -2.46$ percentage points, with percentile intervals $[-15.33, -10.25]$, $[-16.19, -13.04]$, $[-8.46, -6.19]$ and $[-4.06, -.09]$. These residual intervals quantify question-sampling variation conditional on the fitted probabilities and the retrospective model family. Every resample uses the same probabilities and event definition; results are reported for all five structures.

The diagnostic tests the combination of the fitted marginals and Equation 14. Its residuals summarize the joint contribution of item variation, cross-view dependence and marginal approximation to the predicted event rates. Evaluating these rates alongside marginal NLL gives response models explicit targets at both the individual-view and paired levels.

Judge / pairs	Observed	Shared	Bias	Threshold	BT	Full
Qwen / GPT	2	15.79	15.57	15.36	15.28	15.40
Qwen / Claude	1	16.77	16.82	15.91	16.10	16.19
Llama / GPT	0	8.25	7.29	8.18	7.34	7.35
Llama / Claude	1	5.48	3.42	5.49	3.46	3.46

Table 13: Cross-label sign reversals per 100 questions: observed counts and expectations under conditionally independent completions of all five fitted marginal structures. Each event requires four valid responses and opposite nonzero even votes.

G Earlier multi-view experiments and controls

These exploratory controls motivated the label intervention by locating changes in selection and the verdict vocabulary. We first report repeated and reordered judgments, then compare aggregation rules on the same observed responses.

The earlier 200-question expansion contains 95 MMLU-Pro, 52 LiveBench-reasoning, 28 LiveBench-math and 25 LiveCodeBench pairs. Each judge receives original, exact-repeat, swapped, JSON and JSON-swapped views. Repeat is excluded from aggregation. Of 1,000 outputs per judge, Qwen has 976 valid and Llama 703. Original and repeat responses agree on every question, including invalid outputs. These exact repeats establish an empirical control for the study protocol.

Among jointly valid original/swapped views, ternary judgments differ on 119/194 Qwen questions and 61/118 Llama questions. Decisive A/B reversal uses the narrower jointly decisive support: 63/86 and 39/60. JSON affects selection as well as winner identity; 36 of 38 Qwen original-to-JSON changes are tie transitions. Of 321 invalid outputs, 308 are exact XML tags containing C–J. This concentration identifies external verdict vocabulary as a concrete intervention target and motivated the candidate-alphabet studies.

Judge	Method	Calls	Correct/200	Decisive/200	Cond. acc.
Qwen	Original	1	73	113	64.6%
Qwen	Two-order vote	2	60	79	75.9%
Qwen	Four-view vote	4	73	97	75.3%
Llama	Original	1	47	87	54.0%
Llama	Two-order vote	2	23	43	53.5%
Llama	Four-view vote	4	25	49	51.0%

Table 14: Earlier expansion results, on a distinct 200-question set shared between judges. Any invalid required view makes the aggregate abstain.

For signed voting, A contributes $+1$, B -1 and TIE zero. Zero total votes produce a tie, and any invalid required view produces an abstention. Four-view Qwen voting preserves total correctness while reducing coverage; Llama's correct count changes from 47 to 25 (Table 14). A shared/new/dropped decisive-support partition attributes these changes as $+4 + 27 - 31 = 0$ and $-2+5-25 = -22$. For two equally weighted votes, every decisive aggregate preserves the identity of a decisive first vote, a structural property of the rule.

A uniform selector among four acquired views gives expected correct counts 67.75 and 47.50, against voting's 73 and 25. Differences are $+2.625$ points with interval $[-2.875, 8.0]$ and -11.25 with interval $[-15.0, -7.5]$. These are exploratory equal-call comparisons. Expected-coverage random thinning supplies a selection reference; its Qwen interval conditions on the 4,806 of 5,000 resamples with feasible coverage matching. A cascade applied to the same observed responses uses 3.21/3.57 views per question and obtains 76/26 correct. These controls characterize aggregation and selection in the preceding exploratory phase.