# SECOND-ORDER RESPONSE LAWS FOR LLM JUDGES: FROM PROMPT STABILITY TO DECISION QUALITY

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

When an LLM judge changes its verdict after a prompt is reworded, the change may come from the wording or the randomness of another call. Confusing these sources can mislead model comparisons and response selection, yet separating them with few repeated calls is difficult. Plug-in prompt-dispersion scores combine both sources. We develop a framework that connects measurements of judge variability to decision quality. Second-order response laws describe how verdict distributions vary across prompts; crossing prompts with candidate orders separates prompt, position, interaction, and call variation. Agreement between distinct calls removes finite-call bias. We derive the prompt estimator's exact variance and establish matching upper and lower bounds for detecting weak prompt effects under a fixed, balanced design. Identical-input controls and an independent reference with more repeated calls validate the correction. On a matched panel, correction reverses the Qwen–Gemma-3 ranking by prompt sensitivity. We analyze over one million recorded categorical judgments across six judge models and three benchmarks. Cooling makes Qwen's repeated calls more consistent while its corrected prompt estimate rises nearly fivefold, from .011 to .054; the plug-in estimate rounds to .05456 at both temperatures. To examine what greater repeatability means for decisions, we evaluate independent groups of votes balanced across candidate orders. In Qwen, cooling produces more ties; changing the tie rule on the same votes shifts the estimated accuracy effect of cooling by 9.84 percentage points. An analytical model of position dependence explains how balanced votes can concentrate near a tie boundary and lower accuracy when ties lead to abstention, even when the judge favors the correct answer on average. The value of consistency therefore depends on alignment with correct answers, how votes are combined, and the cost of abstention.

## 1 INTRODUCTION

Consider two LLM judges that each select candidate A half the time. For one, every prompt gives an equal chance of A or B. For the other, half the prompts always select A and half always select B. Their average judgments are the same. Repeating a prompt reveals randomness in the first judge; changing the prompt reveals systematic dependence in the second (Figure 1). Which kind of stability matters when a judge compares models or selects responses?

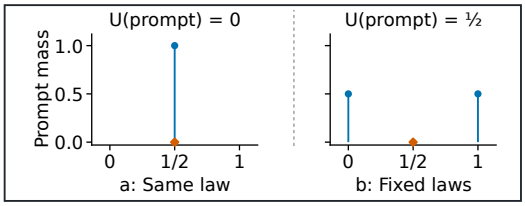


Figure 1: The same marginal $\Pr(A) = 1/2$, two mechanisms: (a) every prompt randomizes; (b) each prompt fixes a verdict, with half choosing each candidate. Horizontal: $\Pr(A \mid x)$; vertical: prompt mass.

This question arises in instruction-following, preference, and correctness-based evaluation (Zheng et al., 2023; Zeng et al., 2024; Tan et al., 2025). A changed verdict can reflect wording, position, or another random call. With few repeated calls, sampling noise inflates observed prompt dispersion and can distort comparisons of judges by prompt sensitivity.

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.

We develop a framework that connects measurements of judge variability to decision quality. Its basis is the *second-order response law*, which describes how verdict distributions vary across prompts. Crossing prompts with candidate orders separates prompt, position, interaction, and call variation. Agreement between distinct calls removes the inflation caused by limited repeats, revealing how prompt dispersion changes.

The correction reveals changes in prompt dependence that the plug-in score masks. Cooling improves repeatability in Qwen and Gemma while corrected prompt estimates increase. Qwen's prompt estimate rises 4.79-fold; the plug-in score rounds to .05456 at both endpoints. Table 1 separates call-level uncertainty from item-composition sensitivity.

We then collect independent groups of balanced votes to study how these changes affect decisions through gold alignment, voting rules, and abstention cost (Figure 2).

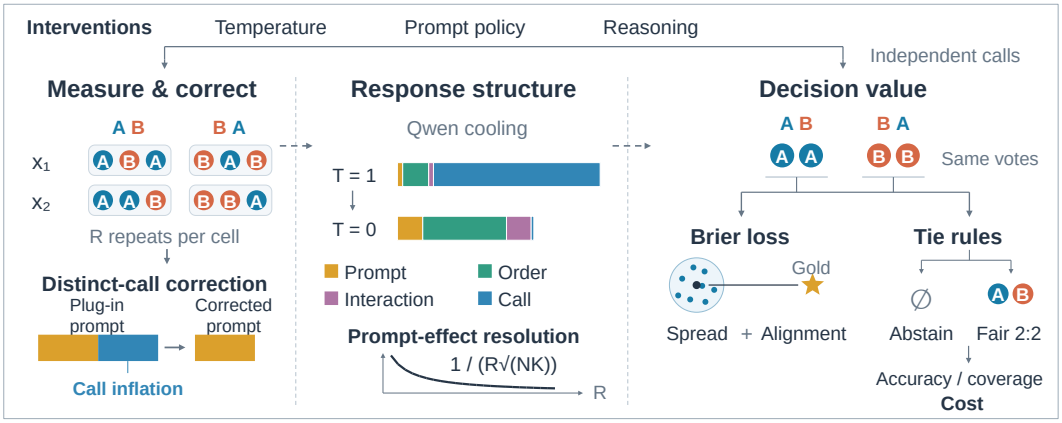


Figure 2: From response structure to decision value. Left: repeated calls for each prompt and order let us correct finite-call inflation using agreement between distinct calls. Middle: Qwen's corrected components under cooling ($N = 40$ items, $K = 48$ prompts, $R = 4$ repeats), shown on a common absolute scale, and the detection limit near a zero prompt effect under a balanced design with fixed items and prompts. Cooling reduces call variation while increasing estimated prompt dispersion. Right: independent groups of balanced votes let us decompose Brier loss and compare how tie rules affect accuracy, coverage, and cost on the same votes. AABB illustrates one group; the cloud represents multiple deployment vectors and their mean relative to gold. Temperature, prompt policy, and reasoning are studied separately. Colors track candidate identity; other symbols are schematic. Dashed arrows connect results from separate call samples.

**Relation to existing work.** Prompt-sensitivity studies measure performance variation across formats and templates (Sclar et al., 2024; Mizrahi et al., 2024), response-likelihood changes (Chatterjee et al., 2024), and within-instance correctness or quality-score differences (Zhuo et al., 2024). PromptEval estimates template-level performance distributions (Polo et al., 2024); CALM compares modification robustness with repeated-call consistency (Ye et al., 2025), and JudgeSense measures verdict agreement under paraphrase (Bellibatlu et al., 2026). Other studies examine position bias (Wang et al., 2024; Shi et al., 2025), repeatability and bias (Norman et al., 2026), tie instructions (Usami et al., 2026), judgment-distribution inference (Wang et al., 2025), and calibrated A/B/tie aggregation (Dadkhahi et al., 2025).

Crossed variance-component analysis (Cronbach et al., 1972; Brennan, 2001; Searle et al., 1992) and second-order uncertainty measures (Sale et al., 2023) provide the basis for our decomposition; distinct-call correction uses within-cell quadratic U-statistics (Hoeffding, 1948). Related LLM work decomposes correctness variation (Romanou et al., 2026), separates sampling noise from system-prompt reward variation (Gao et al., 2026), projects precision across evaluation configurations (Messing, 2026), and links variance to balanced judge allocation (Zhu et al., 2026) or adaptive querying (Saha et al., 2026). We decompose per-item verdict distributions, remove finite-call bias, and give finite-sample guarantees for the corrected measurement.

**Contributions.** Across six judge models and three benchmarks, over one million recorded judgments support three connected parts of the framework (Figure 2; Table S1; Section K.1).

*Measurement and correction.* The plug-in prompt estimate has exact bias $(K-1)U_{\text{call}}/(2KR)$; distinct-call correction removes this decoding-dependent term with two independent calls per cell. We derive its exact variance and matching weak-effect detection bounds under a fixed crossed design, including nulls with nonzero prompt–order interaction. Identical-input controls, an independent reference with more repeats, and known-law simulations validate the measurement; on a matched panel, correction reverses the Qwen–Gemma-3 ranking by prompt sensitivity (Section 3).

*Response structure.* One analytical family shows how sharpening separates repeatability from prompt dispersion; another explains tie concentration at a fixed gold margin. In matched experiments, cooling separates repeatability from estimated prompt dependence, examples shift dependence toward position, and thinking increases call variation (Sections 4–6).

*Decision value.* At a fixed positive gold margin, our position-dependence model shows that lower vector risk and Brier loss can coexist with lower strict-vote accuracy. Relative to abstention, resolving pure-2:2 ties by a fair coin changes an accuracy contrast by half the change in pure ties; for Qwen's cooling contrast, the shift is 9.84 pp. A collision-moment criterion, based on polynomial unbiased estimation (Kimpara et al., 2023), identifies when the risk difference between equal-budget designs can be estimated without bias from one call per cell (Propositions S6–S7). For a fixed target, a variance identity gives a continuous oracle allocation criterion for items and repeats (Proposition S10). Independent deployments test measurement-based risk predictions across three models and two benchmarks (Table S32).

## 2 Response Structure and Finite-Call Measurement

We separate verdict variation by source and correct for finite-call bias. Fix an item and a judge configuration, including the checkpoint, rendered instructions, decoding, stopping, serving system, and parser. After restoring candidate identity across presentation orders, a completed verdict lies in $\mathcal{Z} = \{\text{CANDIDATE_1}, \text{CANDIDATE_2}, \text{TIE}, \perp\}$, where $\perp$ is an invalid output. The target is therefore the full distribution of generated and parsed verdicts, including invalid outcomes. For prompt $X \sim Q$, write $P_X(c) = \mathbb{P}(Z = c \mid X)$ and $\bar{p} = \mathbb{E}_Q P_X$. For the law $\mathcal{L}_Q(P_X)$, define

$$U_{\text{prompt}}(Q) = \mathbb{E}_Q\|P_X - \bar{p}\|^2, \qquad U_{\text{within}}(Q) = \mathbb{E}_Q[1 - \|P_X\|^2], \qquad U_{\text{total}}(Q) = 1 - \|\bar{p}\|^2. \tag{1}$$

The first quantity measures how verdict distributions vary across prompts; the second is the probability that two independent calls under the same prompt disagree.

**Crossing prompts with candidate order.** For a fixed set of $K$ equally weighted prompts and two equally weighted orders, let $P_{ko} = \mathcal{L}(Z \mid x_k, o)$, $P_k^q = (P_{k,AB} + P_{k,BA})/2$, $P^o = K^{-1}\sum_k P_{ko}$, and $\bar{p} = K^{-1}\sum_k P_k^q$. Define

$$U_{\text{prompt}}^{(q)} = \frac{1}{K}\sum_k \|P_k^q - \bar{p}\|^2, \qquad U_{\text{order}} = \frac{1}{2}\sum_o \|P^o - \bar{p}\|^2,$$
$$U_{\text{int}} = \frac{1}{2K}\sum_{k,o} \|P_{ko} - P_k^q - P^o + \bar{p}\|^2, \qquad U_{\text{call}} = \frac{1}{2K}\sum_{k,o}(1 - \|P_{ko}\|^2). \tag{2}$$

**Proposition 1** (Disagreement decomposition)**.** *For any finite categorical alphabet,*

$$U_{\text{total}}(Q) = U_{\text{within}}(Q) + U_{\text{prompt}}(Q), \qquad U_{\text{total}} = U_{\text{prompt}}^{(q)} + U_{\text{order}} + U_{\text{int}} + U_{\text{call}}, \tag{3}$$

*with all four population components nonnegative.*

Prompt dependence averaged over orders differs from prompt–order interaction: even when $U_{\text{prompt}}^{(q)} = 0$, the effect of position can vary across prompts. Their sum is the average prompt dispersion within orders. The crossed profile shows how interventions redistribute variation among these sources.

**Removing finite-call inflation.** Observe $R \geq 2$ independent calls per cell, with independence across cells and items conditional on their fixed laws. For cell $a$, let $\widehat{p}_a$ be its empirical categorical vector. Estimate inner products by

$$\widehat{A}_{aa} = \frac{\sum_{r \neq r'} \mathbf{1}\{Z_{ar} = Z_{ar'}\}}{R(R-1)}, \qquad \widehat{A}_{ab} = \widehat{p}_a^{\top}\widehat{p}_b \quad (a \neq b). \tag{4}$$

These are unbiased estimates of $P_a^\top P_b$. Substituting these agreements into the quadratic component formulas gives estimates using all distinct call pairs and preserves Equation (3) for every sample. The corrected and plug-in prompt estimates satisfy

$$\mathbb{E}\widehat{U}^{(q)}_{\text{plug,prompt}} - U^{(q)}_{\text{prompt}} = \frac{K-1}{2KR} U_{\text{call}}, \tag{5}$$

$$\widehat{U}^{(q)}_{\text{plug,prompt}} - \widehat{U_{\text{prompt}}}^{(q)} = \frac{K-1}{2KR} \widehat{U_{\text{call}}}. \tag{6}$$

At the same repeat budget, a larger call component produces greater finite-call inflation of the plug-in prompt estimate. We retain raw unbiased estimates, including negative values near zero. On this balanced design the estimator is algebraically equivalent to trace-mean-square correction. Our analysis establishes its exact variance and matching weak-effect detection bounds; complete estimators and proofs appear in the supplement.

**Resolution of weak prompt dependence.** The next result quantifies repeated-call uncertainty for fixed items and prompt laws. Let $\theta = N^{-1}\sum_i U^{(q)}_{\text{prompt}}(s_i)$ and let $\widehat{\theta}$ average the corrected estimates over $N$ items.

**Theorem 1** (Finite-call measurement and resolution)**.** *For $N \geq 1$, $K \geq 2$, $R \geq 2$, and mutually independent calls from arbitrary fixed categorical cell laws,*

$$\mathbb{E}\widehat{\theta} = \theta, \qquad \text{Var}(\widehat{\theta}) \leq \frac{\theta}{NKR} + \frac{K-1}{8NK^2R(R-1)}. \tag{7}$$

*Write $b = (K-1)/[8NK^2R(R-1)]$. For fixed $\alpha, \beta \in (0,1)$, the test $\widehat{\theta} > \sqrt{(1-\alpha)b/\alpha}$ has type-I error at most $\alpha$ at $\theta = 0$ and type-II error at most $\beta$ when $\theta \geq C_{\alpha,\beta}/(R\sqrt{NK})$. The constant depends only on the error levels. For even $K$ and $\alpha+\beta < 3/4$, no test attains these levels uniformly at separations $0 < \theta \leq \sqrt{\log(5/4)}/(4R\sqrt{NK})$. Under the null, order-averaged prompt laws are equal within each item, and prompt–order interaction may be nonzero.*

When prompt dispersion is zero, only the quadratic variance term remains. The matching bounds show how repeats and prompt coverage determine the smallest detectable effect. The exact variance, constants, test, and lower-bound construction are given in Theorems S1, S2, and S3. The categorical covariance inequality $\text{tr}(\Sigma_a\Sigma_b) \leq 1/4$, attained by identical fair binary laws, sharpens the variance envelope. For $(N, K, R) = (40, 48, 4)$, the explicit power bound guarantees 95% power at 5% type-I error when prompt dispersion is at least .0226; at $(152, 6, 16)$, the sufficient level is .00732. These are sufficient effect sizes for detecting prompt dependence within a single condition.

## 3 Validating the Measurement

We validate the correction against a zero-effect control and an independent reference, then compare judges. Throughout, $N$ counts items, $K$ prompts, and $R$ calls per item–prompt–order cell. Call-level intervals describe repeated-call uncertainty for fixed items and prompts; paired whole-item intervals describe variation in benchmark composition. Each result states its uncertainty target.

**A zero-effect control.** All hidden prompt labels produce byte-identical model inputs, so true prompt dispersion is zero. At $R = 2$, correction changes Qwen's estimate from .0406 to $-.0053$. Across $R = 2, 4, 8$, corrected estimates remain near zero while the plug-in excess shrinks as calls are repeated, as Equation (5) predicts (Figure 3a).

**Correction improves agreement with the reference.** Correction improves agreement with an independently collected $R = 16$ reference at every tested repeat depth (Figure 3b).

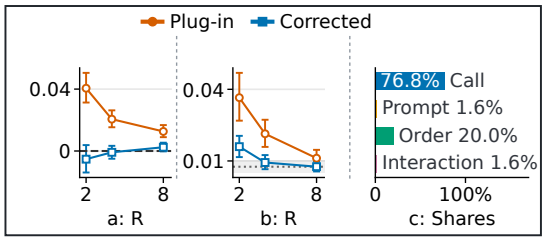


Figure 3: Measurement validation: (a) zero-effect control; (b) MAE to an independent $R = 16$ reference; (c) reference component shares. $R$: calls per cell. Gray: reference split-half disagreement. Bands summarize stratified item composition.

At $R = 2$, the mean absolute discrepancy across items falls from .03650 to .01599. Section L.3 reports reference split-half diagnostics.

**Correction changes judge comparisons.** Correction reverses the Qwen3.6–Gemma-3 ranking by estimated prompt dependence. On a matched four-configuration panel ($N = 50$, $K = 6$, $R = 4$), their plug-in estimates of .02512 versus .01931 become corrected estimates of .00521 versus .01774. Call variation accounts for 76.3% and 88.5% of total disagreement for Qwen and GPT-OSS, whereas order has the largest point estimate for Gemma-3-27B-IT and Claude Haiku 4.5. Thus, at a fixed call budget, the bias depends on how each judge's variation is distributed across components.

The effect persists on a common five-model panel with 40 LLMBar and 40 JudgeBench items, 24 shared rewordings, both orders, and $R = 2$. On LLMBar, correction reduces Qwen's estimate from .0718 to .0065; Figure 4 shows both datasets and the changed rankings. Claude Haiku's JudgeBench result is analyzed separately to examine its invalid outputs (Section D.2).

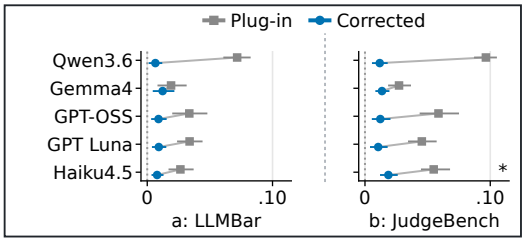


Figure 4: Prompt estimates on (a) LLMBar and (b) JudgeBench: gray plug-in, blue corrected. Each panel has 40 items, 24 prompts, two orders, $R = 2$. Bars: 95% item-composition intervals. Haiku* has 81.74% invalid outputs and is a separate protocol case.

**Calibration under known laws.** Known-law simulations test coverage and certification of small prompt effects across 170 settings with 5,000 independent datasets each. With zero prompt dispersion, $N = 152$, $K = 6$, $R = 8$ and interaction .08, the sharp covariance bound raises the rate of certifying prompt dispersion below .01 from 0% to 97.10% (Monte Carlo interval [96.60%,97.53%]). At the true .01 boundary, no false certification occurs in the tested settings. The sharper bound can therefore certify that prompt dispersion is small without increasing the call budget in this setting.

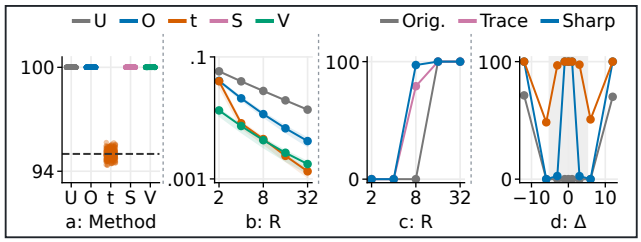


Figure 5: Known-law validation: (a) coverage; (b) mean interval width across eight data-shaped laws (bands: range); (c) prompt-stability certification; (d) informative design comparisons at $R = 32$ (gray: $\pm.005$). U/O/t/S/V: universal, observed variation, block-$t$, oracle S, oracle variance. $R$: calls/cell; $\Delta$: design contrast ($10^{-3}$). Section D.3 gives laws and Monte Carlo uncertainty.

Coverage and precision answer different questions in this experiment: approximate independent-block $t$ intervals cover 94.34–95.66% on this grid, while categorical bounds provide finite-sample guarantees. Section E.4 checks the separate two-panel temperature approximation under different patterns of variance concentration.

## 4 Cooling Separates Repeatability from Prompt Dependence

We use cooling to test whether greater repeatability also means less prompt dependence. We evaluate Qwen3.6-27B and Gemma4-31B-IT on the same $N = 40$ items (20 per LLMBar/JudgeBench), $K = 48$ semantic prompts, both orders, and $T \in \{1, .2, 0\}$. Both use native nonthinking generation, a 512-token cap, and unconstrained verdict output, retaining invalid outcomes. Components use $R = 4$ calls per cell; independent four-call deployments use $J = 8$ groups per item and temperature with matched prompt pairs. Each model contributes 49,920 analyzed calls (Section E).

**Repeatability and prompt dependence diverge.** Cooling makes repeated calls more consistent in both models while their corrected prompt estimates increase. From $T = 1$ to 0, Qwen's call

estimate falls from .35386 to .00569, while its corrected prompt component rises from .01125 to .05387, a factor of 4.79. The plug-in prompt estimate rounds to .05456 at both endpoints: the falling finite-call contribution offsets the growing prompt estimate (Equation (6); Figure 6). Gemma shows the same pattern with a smaller prompt increase; Table 1 reports both models' absolute contrasts and intervals.

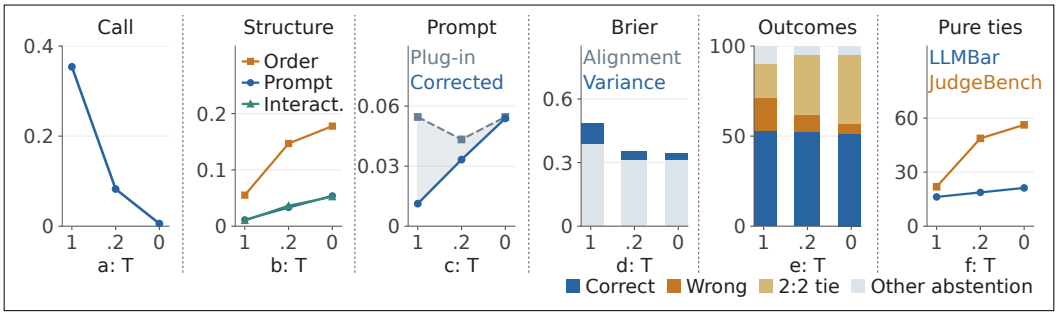


Figure 6: Qwen across temperature $T$: (a) call variation; (b) prompt, order, and interaction components; (c) plug-in and corrected prompt estimates, with their gap shaded; (d) Brier loss decomposed into deployment variance and corrected mean-to-gold squared error; (e) correct, wrong, pure-tie, and other-abstention percentages; (f) pure candidate 2:2 ties by dataset. Components use $R = 4$; independent four-vote deployments use $J = 8$. Protocols and inference: Sections E and E.2.

**Effect sizes and uncertainty.** Table 1 reports absolute contrasts with finite-sample call-level intervals and paired item-composition intervals. Sections E.2 and E.4 give the full results and calibration.

Table 1: Endpoint contrasts ($T = 0$ minus $T = 1$), using complete-pair $R = 4$ estimates. Call-level intervals have simultaneous coverage of at least 95% for each model's four contrasts (two cooling levels, two components); the endpoint subset is shown. Composition intervals are separate pointwise 95% paired item-resampling summaries.

| Model | Component | Estimate | Call-level interval | Composition interval |
|---|---|---|---|---|
| Qwen | Call | $-.34818$ | $[-.37388, -.32247]$ | $[-.37860, -.32010]$ |
| Qwen | Prompt | $+.04262$ | $[-.05699, .12589]$ | $[.03065, .05448]$ |
| Gemma | Call | $-.04488$ | $[-.07059, -.01917]$ | $[-.06736, -.02526]$ |
| Gemma | Prompt | $+.01421$ | $[-.07987, .09869]$ | $[.00731, .02261]$ |

**Other metrics summarize different properties.** Agreement, correctness spread, and prompt dispersion capture different properties of the same verdicts. On these Qwen labels, across-prompt agreement rises from .57 to .71 and the ProSA-style binary correctness gap falls from .32 to .10 (Zhuo et al., 2024), while the corrected prompt estimate rises from .011 to .054. Agreement also depends on how concentrated the mean verdict distribution is; correctness spread records only whether each verdict is correct. Section E.6 gives their definitions and dataset-specific results.

**A sharpening mechanism.** Take two equally weighted binary prompts, identical across orders, with fixed opposite logit gaps $\pm a$, $a > 0$. For $T > 0$, let $p_{\pm}(T) = \sigma(\pm a/T)$, where $\sigma$ is the logistic function. Then

$$U^{(q)}_{\text{prompt}}(T) = \tfrac{1}{2}\tanh^2\left(\frac{a}{2T}\right), \qquad U_{\text{call}}(T) = \frac{1}{2\cosh^2(a/(2T))}. \tag{8}$$

Their sum remains $1/2$: cooling converts within-prompt randomness into between-prompt dispersion while keeping the underlying preferences fixed. Prompts that share a preferred candidate instead converge toward the same verdict as $T \to 0$. This binary family shows how changing response probabilities can redistribute variation. Qwen and Gemma show the same opposing changes in call and prompt estimates when measured after generation, stopping, and verdict parsing.

The 48 prompts fall into four predefined families: lexical, syntactic, discourse, and clause-order variation. Across all four families, call estimates fall and corrected prompt estimates rise in both models. The clause-order subset uses the same five sentences in 12 orders. Section E.6 reports this exploratory reanalysis of the same item panel and its uncertainty.

## 5 From Response Structure to Final Judgments

Decision quality depends on response variation, gold alignment, and the voting rule. We examine their roles using independent four-call deployments. Design D draws two distinct prompts, presents each in both orders, and forms a verdict-frequency vector $V$. The original rule selects a candidate only if it is the unique most frequent category. It abstains if categories share the maximum count or if the most frequent category is tie or invalid.

**Gold alignment.** For $\mathbb{E}V = \bar{p}$ and gold candidate $y$, squared loss separates into

$$\mathbb{E}\|V - e_y\|^2 = \underbrace{\mathbb{E}\|V - \bar{p}\|^2}_{\text{vector risk}} + \underbrace{\|\bar{p} - e_y\|^2}_{\text{gold alignment error}} . \tag{9}$$

The first term measures how tightly vote vectors cluster around the judge's mean; the second measures how close that mean is to the gold answer. Cooling lowers Qwen's vector-risk estimate from .09643 to .02924 and Brier estimate from .48516 to .34141. In Gemma, estimated vector risk also falls, but Brier loss changes little because alignment responds differently. These estimates use the same independent votes; Table S16 reports Qwen's paired item-composition intervals.

**A fixed-margin position mechanism.** Section 4 models the sharpening of opposing prompt preferences under cooling. We now vary position dependence in balanced voting at a fixed mean gold preference. Consider a binary family in which every prompt chooses the gold candidate with probability $\pi + \delta$ in AB order and $\pi - \delta$ in BA order. Here $1/2 \le \pi < 1$ and $0 \le \delta \le 1 - \pi$. Its mean gold margin $m = 2\pi - 1$ stays fixed, $U^{(q)}_{\text{prompt}} = U_{\text{int}} = 0$, and $U_{\text{order}} = 2\delta^2$. Independent balanced votes have vector risk $[\pi(1-\pi) - \delta^2]/2$. Their pure-2:2 probability $T_{22}$ and strict accuracy follow from the sum of independent $\text{Bin}(2, \pi + \delta)$ and $\text{Bin}(2, \pi - \delta)$ variables.

**Proposition 2** (Position dependence and the tie rule)**.** *For $1/2 < \pi < 2/3$, increasing $\delta^2$ decreases vector risk, gold Brier loss, and strict accuracy, while increasing pure ties and fair-coin accuracy. The mean gold margin remains positive and fixed.*

For example, at $\pi = .6$, increasing $\delta$ from 0 to .4 reduces vector risk from .12 to .04 while strict accuracy falls from .4752 to .36. Figure 7 shows the accompanying loss and tie-rule curves. The proof and zero-margin endpoint are in the supplement.

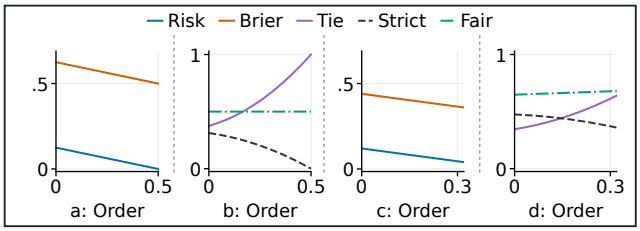


Figure 7: Exact binary families: (a,b) $m = 0$; (c,d) $m = .2$. Horizontal: order component $U_{\text{order}}$. Loss panels show vector risk and gold Brier loss; decision panels show pure ties and strict/fair-coin accuracy.

**Changing the rule on the same independent votes.** A fair pure-2:2 rule chooses either candidate with probability $1/2$ only at a pure candidate tie, leaving all other outcomes unchanged. We compute its expected score exactly from the stored vote groups. If $A$ is accuracy over all deployments,

$$A_{\text{coin22}} = A_{\text{strict}} + \tfrac{1}{2}T_{22}, \qquad \Gamma_t := \Delta_t A_{\text{coin22}} - \Delta_t A_{\text{strict}} = \tfrac{1}{2}\Delta_t T_{22}. \tag{10}$$

Here $\Delta_t$ compares $T = t$ with $T = 1$. Qwen's pure ties increase from 19.06% to 38.75% at $t = 0$, producing a direct rule interaction of 9.84 pp [1.68,18.01]. The interval is simultaneous across both cooling contrasts and both models, conditional on the fixed 40 items and assigned prompt pairs. Each contrast uses 320 paired independent vote groups. Rescoring the same responses directly establishes rule dependence; both rule-specific accuracy contrasts are reported in Table 2. Gemma's smaller tie change, 6.88% to 7.81%, corresponds to an estimated interaction of .47 pp [−1.54,2.48].

The item-level results link position dependence and baseline preferences to ties. The estimated order component rises from .055 to .178 in Qwen, versus .058 to .063 in Gemma. In Qwen, order-component changes correlate with changes in independently sampled pure ties ($r = .75$ after centering within each dataset). The baseline margin $m = \bar{p}_{\text{gold}} - \bar{p}_{\text{other}}$ averages $T = 1$ measurement

Table 2: Cooling contrasts ($T = 0$ minus $T = 1$, pp). Decision intervals jointly cover twelve endpoints within each model; rule-interaction intervals cover four interactions across both models. All are finite-sample 95% intervals.

| Model | Endpoint / rule | Estimate | Simultaneous interval |
|---|---|---|---|
| Qwen | Accuracy / original | $-1.88$ | $[-14.43, 10.68]$ |
| Qwen | Accuracy / fair 2:2 | $+7.97$ | $[-.87, 16.81]$ |
| Qwen | Wrong decision / original | $-12.81$ | $[-23.07, -2.56]$ |
| Qwen | Accuracy rule interaction | $+9.84$ | $[1.68, 18.01]$ |
| Gemma | Accuracy / original | $-.31$ | $[-5.26, 4.63]$ |
| Gemma | Accuracy / fair 2:2 | $+.16$ | $[-4.52, 4.83]$ |
| Gemma | Accuracy rule interaction | $+.47$ | $[-1.54, 2.48]$ |

calls over prompts and orders. Questions with $|m| \leq .25$ show a mean 36.8-pp tie increase, versus .7 pp when $m > .25$. Gemma's tie rate changes on four questions. Section E gives the full item-level results.

**The cost of abstention.** When abstentions increase, both wrong decisions and overall accuracy can decrease. With cost one for a wrong candidate, zero for a correct one, and $c \in [0, 1]$ for abstention,

$$L_c = \mathbb{P}(\text{wrong}) + c\mathbb{P}(\text{abstain}). \tag{11}$$

For Qwen at $T = 0$, the original rule's estimated loss change is $-.128125 + .146875c$: fewer mistakes offset extra abstentions when abstention is cheap. The estimated losses are equal at .872. A finite-sample band covering both rules, both cooling contrasts, and every $c \in [0, 1]$ simultaneously supports lower original-rule loss for $c < .165$. Deployment choice thus depends on the rule and abstention cost; the supplement gives both models' outcomes and interval constructions.

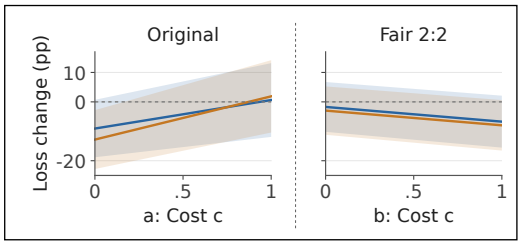


Figure 8: Qwen loss change (pp) relative to $T = 1$: (a) original rule; (b) fair 2:2. Blue: $T = .2$; orange: $T = 0$. Bands jointly cover both rules, both temperatures, and all costs at 95%, given items and prompt pairs.

## 6 Prompting, Reasoning, and Deployment Design

We next examine how prompt policies and reasoning change response variation and final decisions.

**Examples redirect dependence toward position.** In an exploratory comparison of Qwen's four six-prompt policies, replacing paraphrasing with examples lowers the prompt estimate from .00339 to .00014 and raises the order estimate from .09784 to .16677. Under D, pure candidate ties rise from 27.88% to 36.85%. Fair pure-2:2 rescoring changes the original-rule accuracy contrast of $-6.36$ pp $[-8.99, -3.73]$ to $-1.88$ pp $[-4.51, .76]$, showing that tie handling accounts for much of the observed gap (Section G).

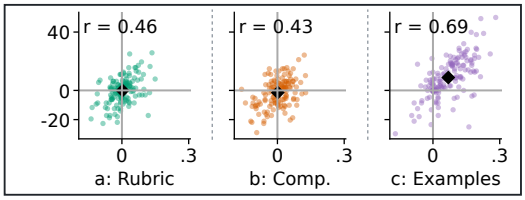


Figure 9: Changes from paraphrasing on 152 items. Horizontal: corrected order component ($R = 16$); vertical: independent pure-2:2 ties (pp; D/E mean, 64 groups/item/design). Diamonds: paired means; $r$: descriptive correlation.

Changes in position dependence correlate with changes in independently sampled ties across all three policies: $r = .69$ for examples, .46 for rubrics, and .43 for comparison (Figure 9). When order is reversed, a preferred position corresponds to the opposite candidate, so balanced votes can cluster near a tie. Section J gives complete response and outcome profiles.

**Reasoning increases call variation.** Thinking increases call variation under both instructions, from .01302 to .09277 for the base instruction, while Brier point estimates also fall. On 32 RewardBench 2 items, the Gemma4-31B-IT study combines base/example instructions with either nonthinking generation or a 1,024-token thinking budget (Malik et al., 2026). Each arm uses 48 prompts, two orders, $R = 2$, eight independent D deployments per item, and the same constrained A/B/TIE final decoder. Figure 10 separates the lower Brier point estimates into increased deployment variation and a larger estimated reduction in mean-to-gold error. Adjusted composition intervals support both call increases in the exploratory twelve-contrast analysis.

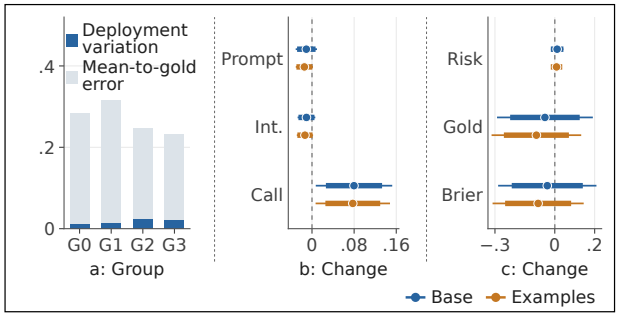


Figure 10: Gemma: (a) Brier decomposition; (b,c) thinking-minus-nonthinking contrasts. G0/G1: nonthinking base/examples; G2/G3: thinking base/examples. Int./Risk/Gold: interaction, deployment variance, squared mean-to-gold error. Thick/thin intervals: 95%/99.583% centered whole-item bootstrap intervals, paired across arms and stratified by domain; the latter target approximate 95% coverage over twelve endpoints. These quantify item composition conditional on recorded calls.

Qwen's intervals are approximate simultaneous 95% item-composition intervals. Gemma's paired component and loss contrasts appear in Figure 10; Section G.2 gives the prespecified tests and exploratory analyses.

**Response components also determine sampling tradeoffs.** Design D uniformly samples two distinct prompts and queries both orders; E uniformly samples four prompts and randomly assigns two to each order. For $K \geq 4$ equally weighted prompts, both estimate the same response mean. Their exact squared-error risk contrast is

$$\mathcal{R}_E - \mathcal{R}_D = \frac{K}{4(K-1)}\left(U_{\text{int}} - U^{(q)}_{\text{prompt}}\right). \tag{12}$$

At the same four-call budget, E has lower population vector risk precisely when prompt dispersion exceeds prompt–order interaction. Both designs cancel the order main effect. Because the same-cell terms also cancel in their risk difference, it can be estimated without bias from one measurement call per cell. More generally, this comparison is possible for designs with the same budget and target weights exactly when their within-cell collision moments are equal (Propositions S6–S7). For a fixed benchmark target, an exact variance identity gives a continuous oracle solution for dividing the budget between items and repeats (Proposition S10). Table S32 compares predicted and observed risk differences across three models and two benchmarks, using separate measurement and deployment calls on the same test pool. Table S34 reports simultaneous 95% call-level intervals for eight Qwen conditions; Section K.2 compares raw, fixed-shrinkage, held-out shrinkage and zero predictors.

## 7 Conclusion

Finite-call correction separates prompt dependence from call randomness, and detection bounds show how small a prompt effect we can detect. Analytical models and matched interventions show how repeatability, prompt dependence, and position effects change together. Their value for final judgments depends on gold alignment, voting rules, and abstention cost; the same components give exact risk comparisons between deployment designs and guide how to divide a call budget between items and repeats.

**Limitations and future work.** Our guarantees apply to independent calls that return categorical verdicts for a fixed set of prompts. Future work can study generated prompts, dependent calls, and continuous scores. We also aim to test more judges and use human review to check that prompts preserve meaning. Other directions are to estimate intervention and design differences more precisely and allocate calls based on the observed variation.

### AI USE STATEMENT

AI tools helped polish the writing, analyze experimental results, and check and correct mathematical formulations. The authors led the formulation and description of the problem, designed and validated the experiments, carried out the theoretical analysis, and prepared the initial and final manuscripts. They take responsibility for the complete work.

### ETHICS STATEMENT

This work studies LLM evaluation using public benchmarks and model-generated judgments, without human-subject experiments. The framework supports transparent analysis of judge variability, including effects of prompt wording, candidate order, and decision rules that can influence model comparison and response selection.

### REPRODUCIBILITY STATEMENT

The main paper and appendix give the estimators, proofs, and experimental protocols. A compact data package provides recorded categorical judgments, prompt and item metadata, gold labels, independent vote groups, and analysis scripts for reproducing the main numerical results.

## REFERENCES

Rohith Reddy Bellibatlu, Edward Raff, and Wenbin Zhang. JudgeSense: A benchmark for prompt sensitivity in LLM-as-a-judge systems. *arXiv preprint arXiv:2604.23478*, 2026. URL https://arxiv.org/abs/2604.23478v2.

Robert L. Brennan. *Generalizability Theory*. Statistics for Social and Behavioral Sciences. Springer, New York, NY, 2001. URL https://link.springer.com/book/10.1007/978-1-4757-3456-0.

Anwoy Chatterjee, H S V N S Kowndinya Renduchintala, Sumit Bhatia, and Tanmoy Chakraborty. POSIX: A prompt sensitivity index for large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 14550–14565, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.852/.

Lee J. Cronbach, Goldine C. Gleser, Harinder Nanda, and Nageswari Rajaratnam. *The Dependability of Behavioral Measurements: Theory of Generalizability for Scores and Profiles*. John Wiley & Sons, New York, 1972. URL https://archive.org/details/dependabilityofb0000unse.

Hamid Dadkhahi, Firas Trabelsi, Parker Riley, Juraj Juraska, and Mehdi Mirzazadeh. Distribution-calibrated inference time compute for thinking LLM-as-a-judge. *arXiv preprint arXiv:2512.03019*, 2025. URL https://arxiv.org/abs/2512.03019v2.

Zhaolin Gao, Yu (Sid) Wang, Bo Liu, Thorsten Joachims, Kianté Brantley, and Wen Sun. $p1$: Better prompt optimization with fewer prompts. *arXiv preprint arXiv:2604.08801*, 2026. URL https://arxiv.org/abs/2604.08801v1.

Wassily Hoeffding. A class of statistics with asymptotically normal distribution. *The Annals of Mathematical Statistics*, 19(3):293–325, September 1948. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-19/issue-3/A-Class-of-Statistics-with-Asymptotically-Normal-Distribution/10.1214/aoms/1177730196.full.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Dhamma Kimpara, Rafael Frongillo, and Bo Waggoner. Proper losses for discrete generative models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 17015–17040. PMLR, 2023. URL https://proceedings.mlr.press/v202/kimpara23a.html.

Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. RewardBench 2: Advancing reward model evaluation. In *International Conference on Learning Representations*, volume 2026, pp. 144839–144866, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/ea4fe0a56d02c93401902b5b4c6b12da-Abstract-Conference.html.

Solomon Messing. Hidden measurement error in LLM pipelines distorts annotation, evaluation, and benchmarking. *arXiv preprint arXiv:2604.11581*, 2026. URL https://arxiv.org/abs/2604.11581v6.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. State of what art? a call for multi-prompt LLM evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949, 2024. URL https://aclanthology.org/2024.tacl-1.52/.

Justin D. Norman, Michael U. Rivera, and D. Alex Hughes. Reliability without validity: A systematic, large-scale evaluation of LLM-as-a-judge models across agreement, consistency, and bias. *arXiv preprint arXiv:2606.19544*, 2026. URL https://arxiv.org/abs/2606.19544v1.

Felipe Maia Polo, Ronald Xu, Lucas Weber, Mírian Silva, Onkar Bhardwaj, Leshem Choshen, Allysson Flavio Melo de Oliveira, Yuekai Sun, and Mikhail Yurochkin. Efficient multi-prompt evaluation of LLMs. In *Advances in Neural Information Processing Systems*, volume 37, pp. 22483–22512. Curran Associates, Inc., 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/28236482f64a72eec43706b6f3a6c511-Abstract-Conference.html.

Angelika Romanou, Mark Ibrahim, Candace Ross, Chantal Shaib, Kerem Oktar, Samuel J. Bell, Anaelia Ovalle, Jesse Dodge, Antoine Bosselut, Koustuv Sinha, and Adina Williams. Brittlebench: Quantifying LLM robustness via prompt sensitivity. *arXiv preprint arXiv:2603.13285*, 2026. URL https://arxiv.org/abs/2603.13285v2.

Aadirupa Saha, Aniket Wagde, and Branislav Kveton. LLM-as-Judge on a budget. *arXiv preprint arXiv:2602.15481*, 2026. URL https://arxiv.org/abs/2602.15481v2.

Yusuf Sale, Paul Hofman, Lisa Wimmer, Eyke Hüllermeier, and Thomas Nagler. Second-order uncertainty quantification: Variance-based measures. *arXiv preprint arXiv:2401.00276*, 2023. URL https://arxiv.org/abs/2401.00276.

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*, volume 2024, pp. 25055–25083, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/6c0e99d736da621403018ca7b32b1a4d-Abstract-Conference.html.

Shayle R. Searle, George Casella, and Charles E. McCulloch. *Variance Components*. John Wiley & Sons, New York, 1992. URL https://onlinelibrary.wiley.com/doi/book/10.1002/9780470316856.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. Judging the judges: A systematic study of position bias in LLM-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pp. 292–314. The Asian Federation of Natural Language Processing and The Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.ijcnlp-long.18. URL https://aclanthology.org/2025.ijcnlp-long.18/.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y. Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations*, volume 2025, pp. 63277–63303, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/9e720fce64f91114c49cfd640d821da3-Abstract-Conference.html.

Hiroyasu Usami, Keisuke Hara, Ayato Tsuboi, and Naohiko Matsuda. LLM judges have dark current: A psychometric datasheet for LLM-as-a-judge evaluation. *arXiv preprint arXiv:2606.15610*, 2026. URL https://arxiv.org/abs/2606.15610v1.

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 *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9440–9450. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.511. URL https://aclanthology.org/2024.acl-long.511/.

Victor Wang, Michael JQ Zhang, and Eunsol Choi. Improving LLM-as-a-judge inference with the judgment distribution. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 23173–23199, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.findings-emnlp.1259. URL https://aclanthology.org/2025.findings-emnlp.1259/.

Jiayi Ye, Yanbo Wang, Yue Huang, Dongping Chen, Qihui Zhang, Nuno Moniz, Tian Gao, Werner Geyer, Chao Huang, Pin-Yu Chen, Nitesh V. Chawla, and Xiangliang Zhang. Justice or prejudice? quantifying biases in LLM-as-a-Judge. In *International Conference on Learning Representations*, volume 2025, pp. 102351–102390, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/fdca08d371e4b6c031397909e20043bd-Abstract-Conference.html.

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. Evaluating large language models at evaluating instruction following. In *The Twelfth International Conference on Learning Representations*, volume 2024, pp. 40193–40219, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/afc8b034823271816d14f7c1aefe1dff-Abstract-Conference.html.

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. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/91f18a1287b398d378ef22505bf41832-Abstract-Datasets_and_Benchmarks.html.

Ziyi Zhu, Olivier Tieleman, Alexey Bukhtiyarov, and Jinghong Chen. CyclicJudge: Mitigating judge bias efficiently in LLM-based evaluation. *arXiv preprint arXiv:2603.01865*, 2026. URL https://arxiv.org/abs/2603.01865v3.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. ProSA: Assessing and understanding the prompt sensitivity of LLMs. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 1950–1976, Miami, Florida, USA, November 2024. Association for Computational Linguistics. URL https://aclanthology.org/2024.findings-emnlp.108/.

# SUPPLEMENTARY MATERIAL

## A RESPONSE LAWS AND COMPLETE-PAIR ESTIMATION

The primary analysis gives equal weight to all prompts in the declared census. Calls are independent conditional on the fixed item–prompt–order laws, with iid repetitions within each cell. Before analysis, we map verdicts back to the original candidate identities and assign invalid completed outputs to a fourth category $\perp$. We report raw estimates, including negative values when a component is near zero.

### A.1 POPULATION DECOMPOSITION

Let $e_Z$ be the one-hot vector of the categorical response and $P_X = \mathbb{E}[e_Z \mid X]$. The conditional-mean residual is orthogonal to every function of $X$, hence

$$\mathbb{E}\|e_Z - \bar{p}\|_2^2 = \mathbb{E}\|e_Z - P_X\|_2^2 + \mathbb{E}\|P_X - \bar{p}\|_2^2. \tag{S1}$$

Because $\|e_Z\|_2^2 = 1$,

$$\mathbb{E}[\|e_Z - P_X\|_2^2 \mid X] = 1 - \|P_X\|_2^2, \tag{S2}$$

$$\mathbb{E}\|e_Z - \bar{p}\|_2^2 = 1 - \|\bar{p}\|_2^2. \tag{S3}$$

Substitution proves $U_{\text{total}} = U_{\text{within}} + U_{\text{prompt}}$. The same proof applies when the categorical alphabet includes invalid completed outputs $\perp$. Conditioning on validity defines a separate response law, as specified below.

Let $X \sim Q$ and $O \sim q$ be independent by design. Put

$$P_{XO} = \mathcal{L}(Z \mid X, O), \quad P_X^q = \mathbb{E}[P_{XO} \mid X],$$
$$P^O = \mathbb{E}[P_{XO} \mid O], \qquad \bar{p} = \mathbb{E}P_{XO},$$

and

$$H_{XO} = \mathbb{E}_{X,O}\|P_{XO}\|_2^2, \qquad H_X = \mathbb{E}_X\|P_X^q\|_2^2, \tag{S4}$$

$$H_O = \mathbb{E}_O\|P^O\|_2^2, \qquad H_0 = \|\bar{p}\|_2^2. \tag{S5}$$

The components are

$$U_{\text{call}} = 1 - H_{XO}, \tag{S6}$$

$$U_{\text{prompt}}^{(q)} = H_X - H_0, \tag{S7}$$

$$U_{\text{order}} = H_O - H_0, \tag{S8}$$

$$U_{\text{int}} = H_{XO} - H_X - H_O + H_0. \tag{S9}$$

Write

$$P_{XO} - \bar{p} = (P_X^q - \bar{p}) + (P^O - \bar{p}) + (P_{XO} - P_X^q - P^O + \bar{p}).$$

The three centered terms are mutually orthogonal in $L^2(Q \otimes q)$: the interaction has conditional mean zero given $X$ or $O$, and the main effects are orthogonal by independence. Therefore

$$H_{XO} - H_0 = (H_X - H_0) + (H_O - H_0) + (H_{XO} - H_X - H_O + H_0).$$

Adding $1 - H_{XO}$ proves

$$U_{\text{total}} = U_{\text{call}} + U_{\text{prompt}}^{(q)} + U_{\text{order}} + U_{\text{int}}.$$

More explicitly,

$$U_{\text{prompt}}^{(q)} = \mathbb{E}_X\|P_X^q - \bar{p}\|_2^2,$$
$$U_{\text{order}} = \mathbb{E}_O\|P^O - \bar{p}\|_2^2,$$
$$U_{\text{int}} = \mathbb{E}_{X,O}\left\|P_{XO} - P_X^q - P^O + \bar{p}\right\|_2^2.$$

Moreover,

$$U_{\text{call}} = \mathbb{E}_{X,O}\left[1 - \|P_{XO}\|_2^2\right].$$

Hence every population component is nonnegative.

### A.2 Unbiased complete-pair estimator for a fixed census

Let $Q_K = K^{-1}\sum_k \delta_{x_k}$ and let every positive-weight prompt–order cell contain $R_{ko} \geq 2$ calls. Assume all distinct calls are mutually independent conditional on the fixed array $\{P_{ko}\}$. Define the same-cell agreement

$$\widehat{A}_{ko,ko} = \frac{1}{R_{ko}(R_{ko}-1)} \sum_{r \neq r'} \mathbf{1}\{Z_{kor} = Z_{kor'}\}. \tag{S10}$$

For distinct cells, define

$$\widehat{A}_{ko,\ell o'} = \widehat{p}_{ko}^{\top}\widehat{p}_{\ell o'}, \qquad (k,o) \neq (\ell,o'). \tag{S11}$$

Set

$$\widehat{H}_{XO} = \frac{1}{K}\sum_k \sum_o q_o \widehat{A}_{ko,ko}, \tag{S12}$$

$$\widehat{H}_X = \frac{1}{K}\sum_k \sum_{o,o'} q_o q_{o'} \widehat{A}_{ko,ko'}, \tag{S13}$$

$$\widehat{H}_O = \sum_o q_o \frac{1}{K^2}\sum_{k,\ell} \widehat{A}_{ko,\ell o}, \tag{S14}$$

$$\widehat{H}_0 = \frac{1}{K^2}\sum_{k,\ell}\sum_{o,o'} q_o q_{o'} \widehat{A}_{ko,\ell o'}. \tag{S15}$$

Each $\widehat{H}$ is unbiased for its fixed-census counterpart. The component estimators

$$\widehat{U_{\text{call}}} = 1 - \widehat{H}_{XO}, \tag{S16}$$

$$\widehat{U_{\text{prompt}}}^{(q)} = \widehat{H}_X - \widehat{H}_0, \tag{S17}$$

$$\widehat{U_{\text{order}}} = \widehat{H}_O - \widehat{H}_0, \tag{S18}$$

$$\widehat{U_{\text{int}}} = \widehat{H}_{XO} - \widehat{H}_X - \widehat{H}_O + \widehat{H}_0, \tag{S19}$$

$$\widehat{U_{\text{total}}} = 1 - \widehat{H}_0 \tag{S20}$$

are therefore unbiased and satisfy the exact decomposition in every sample.

To express the census correction, define

$$\widehat{A}_X = \frac{1}{K}\sum_k \sum_{o,o'} q_o q_{o'} \widehat{A}_{ko,ko'}, \tag{S21}$$

$$\widehat{A}_{\neq X} = \frac{1}{K(K-1)}\sum_{k\neq\ell}\sum_{o,o'} q_o q_{o'} \widehat{A}_{ko,\ell o'}. \tag{S22}$$

Then

$$\widehat{U_{\text{prompt}}}^{(q)} = \frac{K-1}{K}(\widehat{A}_X - \widehat{A}_{\neq X}). \tag{S23}$$

The inclusive $K^{-2}$ sums target the realized fixed census.

### A.3 Exact order mixture and plug-in excess

For one prompt with order laws $P^o$, the declared mixture is $P^q = \sum_o q_o P^o$, so

$$\|P^q\|_2^2 = \sum_o q_o^2 \|P^o\|_2^2 + \sum_{o\neq o'} q_o q_{o'} (P^o)^{\top} P^{o'}. \tag{S24}$$

Agreement between distinct calls in the same order estimates the diagonal terms; agreement between calls in different orders estimates the off-diagonal terms. Pooling equal numbers of AB and BA calls

into one unordered sample introduces finite-$R$ bias in mixture agreement. If $a = P^{AB}$, $b = P^{BA}$, and there are $n$ calls in each order, ordinary pooled distinct-pair agreement has expectation

$$\mathbb{E}\widehat{A}_{\text{pool}} = \left\| \frac{a+b}{2} \right\|_2^2 - \frac{\|a-b\|_2^2}{4(2n-1)}. \tag{S25}$$

For the plug-in prompt estimate in the fixed crossed design,

$$\widehat{U}^{(q)}_{\text{plug,prompt}} = \frac{1}{K}\sum_k \|\widehat{P}^q_k - \bar{\widehat{P}}^{\,q}\|_2^2, \qquad \widehat{P}^q_k = \sum_o q_o \widehat{p}_{ko},$$

with $R_o$ calls in each prompt cell of order $o$,

$$\begin{aligned}\mathbb{E}\widehat{U}^{(q)}_{\text{plug,prompt}} = U^{(q)}_{\text{prompt}} &+ \frac{K-1}{K^2} \\ &\times \sum_k \sum_o \frac{q_o^2}{R_o}(1 - \|P_{ko}\|_2^2).\end{aligned} \tag{S26}$$

For two orders with $q_o = 1/2$ and equal $R$,

$$\mathbb{E}\widehat{U}^{(q)}_{\text{plug,prompt}} = U^{(q)}_{\text{prompt}} + \frac{K-1}{2KR} U_{\text{call}}. \tag{S27}$$

The balanced two-order design gives this simplified expression. At the cell level, the empirical plug-in concentration and the distinct-call agreement satisfy

$$\|\widehat{p}_{ko}\|_2^2 - \widehat{A}_{ko,ko} = \frac{1 - \widehat{A}_{ko,ko}}{R}.$$

This follows directly by writing both quantities in terms of the categorical cell counts. Averaging this identity over the two equally weighted orders and the $K$ prompts, and then applying the fixed-census factor $(K-1)/K$, gives the exact sample identity

$$\widehat{U}^{(q)}_{\text{plug,prompt}} - \widehat{U^{(q)}_{\text{prompt}}} = \frac{K-1}{2KR}\widehat{U_{\text{call}}}. \tag{S28}$$

At $K = 6$, the $R$-scaled excess equals $(5/12)\widehat{U_{\text{call}}}$ in every sample. The simulation with known cell laws, the control with byte-identical inputs, and the independent reference with more repeated calls test the estimator against separately known or measured targets.

**Deterministic and near-deterministic runtimes.** If every fixed prompt–order cell is deterministic, each $P_{ko}$ is a point mass, $U_{\text{call}} = 0$, and the finite-repeat excess vanishes for every $R$. More generally, Equation (S27) shows that the excess shrinks in proportion to $U_{\text{call}}/R$. At temperature zero, any remaining variability of the configured serving system contributes to $P_{ko}$ and is measured by the call component.

**Fixed census versus a prompt generator.** For a single-order response law, write $\widehat{a}_k$ for within-cell distinct-call agreement and $\widehat{b}_{k\ell}$ for cross-cell agreement, and average them as $\widehat{A}_{\text{within}} = K^{-1}\sum_k \widehat{a}_k$ and $\widehat{A}_{\text{across}} = [K(K-1)]^{-1}\sum_{k\neq\ell}\widehat{b}_{k\ell}$. The fixed-census estimators are $\widehat{U}_w = 1 - \widehat{A}_{\text{within}}$, $\widehat{U}_p = (K-1)(\widehat{A}_{\text{within}} - \widehat{A}_{\text{across}})/K$, and $\widehat{U}_t = 1 - \widehat{A}_{\text{within}}/K - (K-1)\widehat{A}_{\text{across}}/K$. Their expectations follow from $\mathbb{E}\widehat{a}_k = \|P_k\|^2$ and $\mathbb{E}\widehat{b}_{k\ell} = P_k^\top P_\ell$. If prompts are instead iid draws from a generator, $\widehat{A}_{\text{within}} - \widehat{A}_{\text{across}}$ targets its prompt variance, and the usual sample-variance plug-in has expectation $U_p + U_w/R$. A declared census uses denominator $K$ and inflation $(K-1)U_w/(KR)$. These are different sampling targets; the model studies use fixed censuses.

### A.4 What the classical mean-square baseline estimates

The classical fixed-effects method estimates the same quadratic dispersion over the fixed prompt census. Let $Y_{kor}$ be the one-hot outcome for one item, $\widehat{P}_{ko}$ its cell mean, and $\widehat{P}_k$ the order-averaged

mean. With $O = 2$, define trace sums of squares

$$SS_p = OR \sum_k \|\widehat{P}_k - \widehat{\widehat{P}}\|^2, \tag{S29}$$

$$SS_{\text{int}} = R \sum_{k,o} \|\widehat{P}_{ko} - \widehat{P}_k - \widehat{P}^o + \widehat{\widehat{P}}\|^2, \tag{S30}$$

$$MS_e = \frac{\sum_{k,o,r} \|Y_{kor} - \widehat{P}_{ko}\|^2}{KO(R-1)} = \widehat{U}_{\text{call}}. \tag{S31}$$

The numerator degrees of freedom are $d_p = K - 1$ and $d_{\text{int}} = (K-1)(O-1)$. The corresponding corrections for the finite census satisfy

$$\frac{d_p}{KOR}\left(\frac{SS_p}{d_p} - MS_e\right) = \widehat{U}^{(q)}_{\text{plug,prompt}} - \frac{K-1}{2KR}\widehat{U}_{\text{call}} = \widehat{U^{(q)}_{\text{prompt}}}, \tag{S32}$$

$$\frac{d_{\text{int}}}{KOR}\left(\frac{SS_{\text{int}}}{d_{\text{int}}} - MS_e\right) = \widehat{U}_{\text{plug,int}} - \frac{K-1}{2KR}\widehat{U}_{\text{call}} = \widehat{U_{\text{int}}}. \tag{S33}$$

Thus the fixed-effects trace-mean-square correction and the complete-pair estimator coincide in every sample in this balanced design, even when categorical cell variances differ. Subtracting the two component estimates cancels their common correction, so the plug-in contrast also equals the unbiased contrast. These identities hold for the fixed-census target. Random-effects REML instead estimates a quantity defined by a hierarchical population model. Comparing it with the fixed-census estimate requires specifying that model and the relation between the two targets. We treat methods that agree in every sample as the same estimator.

## B Fixed-Census Inference Near Zero

Fix $N \geq 1$ benchmark items, a uniform census of $K \geq 2$ prompts, a categorical alphabet with $|\mathcal{Z}| \geq 2$, and order weights $q_o > 0$ with $\sum_o q_o = 1$. Each item–prompt–order cell contains $R \geq 2$ iid categorical calls, and all calls are independent conditional on the fixed cell laws. Define

$$\theta_i = U^{(q)}_{\text{prompt}}(s_i), \qquad \theta = \frac{1}{N}\sum_i \theta_i, \qquad \widehat{\theta} = \frac{1}{N}\sum_i \widehat{U^{(q)}_{\text{prompt}}}(s_i),$$

where the estimate uses every eligible pair of distinct calls. The following results quantify uncertainty from repeated calls while holding the items and prompt census fixed.

Index a prompt–order cell by $a = (k, o)$, write its law on item $i$ as $p_{ia}$, and put

$$W_{(k,o),(\ell,o')} = \left(\frac{\mathbf{1}\{k=\ell\}}{K} - \frac{1}{K^2}\right) q_o q_{o'}, \tag{S34}$$

$$\Sigma_{ia} = \text{diag}(p_{ia}) - p_{ia}p_{ia}^\top, \qquad d_{ia} = 2\sum_b W_{ab} p_{ib}. \tag{S35}$$

Thus $\theta_i = \sum_{a,b} W_{ab} p_{ia}^\top p_{ib}$.

**Theorem S1** (Exact fixed-census variance and a uniform bound)**.** *Under the stated fixed-cell design,* $\mathbb{E}\widehat{\theta} = \theta$ *and* $\text{Var}(\widehat{\theta}) = N^{-2}\sum_i V_i$*, where*

$$\begin{aligned} V_i = {} & \frac{1}{R}\sum_a d_{ia}^\top \Sigma_{ia} d_{ia} + \frac{2}{R(R-1)}\sum_a W_{aa}^2 \,\text{tr}(\Sigma_{ia}^2) \\ & + \frac{4}{R^2}\sum_{a<b} W_{ab}^2 \,\text{tr}(\Sigma_{ia}\Sigma_{ib}). \end{aligned} \tag{S36}$$

*Here $a < b$ uses any fixed ordering of the cells. With $C = |\mathcal{Z}|$ and $s_q = \sum_o q_o^2$,*

$$\text{Var}(\widehat{\theta}) \leq a_{N,R}\theta + b_{N,R}, \tag{S37}$$

$$a_{N,R} = \frac{2s_q}{NKR}, \qquad b_{N,R} = \frac{(K-1)s_q^2}{2NK^2R(R-1)}. \tag{S38}$$

*For six prompts, two balanced orders, and four verdict categories, $a_{N,R} = 1/(6NR)$ and $b_{N,R} = 5/[288NR(R-1)]$.*

Write $P^q_{ik} = \sum_o q_o p_{i,(k,o)}$ and $\bar{p}_i = K^{-1}\sum_k P^q_{ik}$. The first term in Equation (S36) vanishes when $\theta = 0$: indeed, $d_{i,ko} = 2q_o(P^q_{ik} - \bar{p}_i)/K$. Under this null, the order-averaged prompt laws are the same within each item, while prompt–order interaction is allowed. The remaining two terms retain the call variation within each cell.

**Theorem S2** (Finite-sample test and upper confidence bound)**.** *Let $0 < \alpha < 1$, $\lambda_\alpha = (1-\alpha)/\alpha$, and abbreviate $a = a_{N,R}$, $b = b_{N,R}$. For any prespecified $0 \le \tau \le 1 - 1/|\mathcal{Z}|$, the test*

$$\text{reject } H_0 : \theta \le \tau \quad \text{if} \quad \widehat{\theta} > \tau + \sqrt{\lambda_\alpha(a\tau + b)} \tag{S39}$$

*has type-I error at most $\alpha$. Define $x_+ = \max\{\widehat{\theta}, 0\}$ and*

$$\mathcal{U}_\alpha = \min\left\{1 - \frac{1}{|\mathcal{Z}|},\, x_+ + \frac{\lambda_\alpha a}{2} + \sqrt{\lambda_\alpha(ax_+ + b) + \frac{\lambda_\alpha^2 a^2}{4}}\right\}. \tag{S40}$$

*Then $\mathbb{P}(\theta \le \mathcal{U}_\alpha) \ge 1 - \alpha$ for every fixed array of cell laws in this design.*

The categorical covariance inequalities $\|\Sigma\|_{\mathrm{op}} \le 1/2$ and $\mathrm{tr}(\Sigma_a\Sigma_b) \le 1/4$ control the linear and quadratic terms, respectively. The latter constant is attained by two identical laws assigning probability $1/2$ to each of two categories, regardless of the alphabet size. For $C = 4$, the resulting bound on the quadratic terms is one third smaller than the bound obtained from $\mathrm{tr}(\Sigma_a\Sigma_b) \le (1 - 1/C)/2$. Combining this sharp covariance bound with the inequalities for the coefficients and repeat counts gives the variance envelope. Cantelli's inequality converts it into a distribution-free bound at a prespecified sample size. Equation (S40) uses the positive part of the estimate to construct the bound; we report the raw point estimate. An upper bound below a prespecified tolerance certifies that the target is below that tolerance at the stated confidence level. Supplementary Section B.1 also gives an explicit power bound.

**Theorem S3** (A near-zero testing lower bound)**.** *Consider the same balanced observation design with two equally weighted orders, an even $K \ge 2$, $R \ge 2$, $N \ge 1$, and $|\mathcal{Z}| \ge 2$. Let $\mathcal{F}_0$ and $\mathcal{F}_\delta$ be the classes of fixed cell-law arrays with $\theta = 0$ and $\theta = \delta$, respectively. If*

$$0 < \delta \le \delta_{N,K,R} := \frac{\sqrt{\log(5/4)}}{4R\sqrt{NK}}, \tag{S41}$$

*every test $\phi \in [0,1]$ based on all the calls satisfies*

$$\sup_{P \in \mathcal{F}_0} \mathbb{E}_P \phi + \sup_{P \in \mathcal{F}_\delta} \mathbb{E}_P(1 - \phi) \ge \frac{3}{4}. \tag{S42}$$

*Here $\delta$ is the prompt-dispersion target itself. For $K = 6$, $\delta_{N,6,R} = \sqrt{\log(5/4)}/(4R\sqrt{6N})$.*

The construction permits different response contrasts across items and prompt pairs. Together with the power bound for Equation (S39), it gives the near-zero separation order $1/(R\sqrt{NK})$ for fixed error probabilities whose sum is below $3/4$ in this balanced design.

### B.1 Fixed-Census Inference Near Zero: Proofs

Throughout this section, the benchmark items and all prompt–order cell laws are fixed. The randomness comes from the categorical calls, which are mutually independent, with iid repetitions within each cell.

#### Orthogonal expansion and exact variance

We first prove a slightly more general identity allowing different fixed repeat counts $R_a \ge 2$ across cells of one item. Suppress the item index, and let $Y_{ar} = e_{Z_{ar}}$, $\xi_{ar} = Y_{ar} - p_a$, and

$$\bar{\xi}_a = \frac{1}{R_a}\sum_r \xi_{ar}, \qquad \mathbb{E}\xi_{ar} = 0, \qquad \mathbb{E}\xi_{ar}\xi_{ar}^\top = \Sigma_a.$$

For a symmetric deterministic matrix $W$, define

$$T_W = \sum_{a,b} W_{ab} p_a^\top p_b.$$

Estimate a diagonal product with same-cell distinct-call agreement and an off-diagonal product with the dot product of the two empirical distributions. This gives the complete-pair estimator $\widehat{T}_W$. Both estimates are unbiased for the corresponding population products.

For a diagonal product, direct expansion gives

$$\widehat{A}_{aa} = \|p_a\|^2 + 2p_a^\top \bar{\xi}_a + D_a, \qquad D_a = \frac{1}{R_a(R_a-1)} \sum_{r \neq s} \xi_{ar}^\top \xi_{as}. \tag{S43}$$

For $a \neq b$,

$$\widehat{A}_{ab} = p_a^\top p_b + p_b^\top \bar{\xi}_a + p_a^\top \bar{\xi}_b + \bar{\xi}_a^\top \bar{\xi}_b. \tag{S44}$$

Set $d_a = 2\sum_b W_{ab} p_b$ and $Q_{ab} = \bar{\xi}_a^\top \bar{\xi}_b$. Symmetry of $W$ yields

$$\widehat{T}_W - T_W = \sum_a d_a^\top \bar{\xi}_a + \sum_a W_{aa} D_a + 2 \sum_{a<b} W_{ab} Q_{ab}. \tag{S45}$$

All terms in this expansion have mean zero and are pairwise uncorrelated. The linear summands are independent across cells. A linear term and a $D_a$ term have zero covariance because every product contributing to the covariance contains a centered observation occurring only once; conditioning on the other observations gives zero. The same argument applies to a linear term and $Q_{ab}$. Distinct $D_a$ terms are independent. A $D_a$ term and any $Q_{bc}$ are uncorrelated, since either their cells are disjoint or the other cell of $Q_{bc}$ has conditional mean zero. Finally, distinct $Q_{ab}$ terms are independent if their cell pairs are disjoint; if their pairs share one cell, conditioning on that cell leaves the two other centered cell means independent with expectation zero.

The linear variance is

$$\mathrm{Var}(d_a^\top \bar{\xi}_a) = \frac{d_a^\top \Sigma_a d_a}{R_a}.$$

In $\mathbb{E}D_a^2$, a term formed from two ordered repeat pairs contributes only when the second pair is identical to the first or reverses its order. There are $2R_a(R_a-1)$ such terms. Since two distinct repeats are independent,

$$\mathbb{E}(\xi_{ar}^\top \xi_{as})^2 = \mathrm{tr}(\Sigma_a^2), \qquad r \neq s,$$

and hence

$$\mathrm{Var}(D_a) = \frac{2\,\mathrm{tr}(\Sigma_a^2)}{R_a(R_a-1)}.$$

Independence of the two cell means also gives

$$\mathrm{Var}(Q_{ab}) = \frac{\mathrm{tr}(\Sigma_a \Sigma_b)}{R_a R_b}.$$

Taking variances in Equation (S45) proves

$$\begin{aligned} \mathrm{Var}(\widehat{T}_W) = &\sum_a \frac{d_a^\top \Sigma_a d_a}{R_a} + 2\sum_a \frac{W_{aa}^2\,\mathrm{tr}(\Sigma_a^2)}{R_a(R_a-1)} \\ &+ 4\sum_{a<b} \frac{W_{ab}^2\,\mathrm{tr}(\Sigma_a \Sigma_b)}{R_a R_b}. \end{aligned} \tag{S46}$$

For the prompt matrix in Equation (S34), $T_W = \theta_i$. Setting all $R_a = R$ gives Equation (S36). Independence across fixed items then gives $\mathrm{Var}(\widehat{\theta}) = N^{-2}\sum_i V_i$ and $\mathbb{E}\widehat{\theta} = \theta$.

### A UNIFORM VARIANCE ENVELOPE

For any categorical probability vector $p$, $\Sigma = \mathrm{diag}(p) - pp^\top$ is positive semidefinite, $\mathrm{tr}(\Sigma) = 1 - \|p\|^2 \le 1 - 1/C$, where $C = |\mathcal{Z}|$, and $\|\Sigma\|_{\mathrm{op}} \le 1/2$. For the last bound, take any vector $v$ and note that $v^\top \Sigma v = \mathrm{Var}(v_Z)$. A real random variable in $[m, M]$ has variance at most $(M-m)^2/4$: indeed, $\mathbb{E}[(X-m)(M-X)] \ge 0$ implies $\mathrm{Var}(X) \le (M - \mathbb{E}X)(\mathbb{E}X - m) \le (M-m)^2/4$. Apply this with $m = \min_c v_c$ and $M = \max_c v_c$, and use $(M-m)^2 \le 2\|v\|^2$. To control the quadratic terms, write $s_2 = \sum_c p_c^2$ and $s_3 = \sum_c p_c^3$. Cauchy–Schwarz gives $s_2^2 \le (\sum_c p_c^3)(\sum_c p_c) = s_3$, and hence

$$\mathrm{tr}(\Sigma^2) = s_2 - 2s_3 + s_2^2 \le s_2(1 - s_2) \le \frac{1}{4}.$$

For any two categorical covariance matrices, Frobenius Cauchy–Schwarz therefore implies

$$\mathrm{tr}(\Sigma_a^2) \le \frac{1}{4}, \qquad 0 \le \mathrm{tr}(\Sigma_a \Sigma_b) \le \frac{1}{4}. \tag{S47}$$

Both upper bounds are attained when the two laws equal $(1/2, 1/2, 0, \ldots, 0)$.

For the prompt target, let $\Delta_{ik} = P_{ik}^q - \bar{p}_i$. Summing the coefficients in Equation (S34) gives

$$d_{i,ko} = \frac{2q_o}{K}\Delta_{ik}, \qquad \theta_i = \frac{1}{K}\sum_k \|\Delta_{ik}\|^2.$$

Therefore the first variance term obeys

$$\frac{1}{R}\sum_{k,o} d_{i,ko}^\top \Sigma_{i,ko} d_{i,ko} \le \frac{2s_q}{KR}\theta_i. \tag{S48}$$

Write $H_K = I_K - \mathbf{1}\mathbf{1}^\top/K$. The coefficient matrix is $W = (H_K/K) \otimes (qq^\top)$. Since $H_K$ is a rank-$K-1$ orthogonal projection,

$$\|W\|_F^2 = \frac{K-1}{K^2}\left(\sum_o q_o^2\right)^2 = \frac{K-1}{K^2}s_q^2. \tag{S49}$$

Using Equation (S47) and $R^{-2} \le [R(R-1)]^{-1}$, the two quadratic terms in Equation (S36) are at most

$$\begin{aligned}\frac{1}{2R(R-1)}\sum_a W_{aa}^2 + \frac{1}{R^2}\sum_{a<b} W_{ab}^2 \\ \le \frac{\|W\|_F^2}{2R(R-1)} = \frac{(K-1)s_q^2}{2K^2R(R-1)}.\end{aligned} \tag{S50}$$

Equations (S48)–(S50), summed over items and divided by $N^2$, prove Equation (S37) and complete the proof of Theorem S1.

Because every $\theta_i$ is nonnegative, $\theta = 0$ implies $\Delta_{ik} = 0$ for all $i, k$. The individual $P_{iko}$ may vary across prompts at each order, so the null bound covers arrays with nonzero prompt–order interaction. For two orders, $s_q = 1/2$ gives the constants stated for $K = 6$.

### CANTELLI TEST, UPPER BOUND, AND POWER

For a centered real random variable $Y$ with variance $v$, Cantelli's inequality states

$$\mathbb{P}(Y \ge t) \le \frac{v}{v + t^2}, \qquad t > 0. \tag{S51}$$

Applying the inequality to $-Y$ gives the same bound for the lower tail.

Let $v_\tau = a\tau + b$ and $h_\tau = \sqrt{\lambda_\alpha v_\tau}$. Under $\theta \le \tau$, rejection in Equation (S39) implies $\widehat{\theta} - \theta > h_\tau$, and $\mathrm{Var}(\widehat{\theta}) \le a\theta + b \le v_\tau$. The function $v/(v + h_\tau^2)$ increases in $v$, so Cantelli gives

$$\mathbb{P}_\theta(\text{reject}) \le \frac{v_\tau}{v_\tau + h_\tau^2} = \alpha.$$

This proves the test guarantee for the full composite null.

To prove the upper confidence bound, let $x = \widehat{\theta}$, $y = x_+$, and define the untruncated bound

$$u(y) = y + \frac{\lambda_\alpha a}{2} + \sqrt{\lambda_\alpha(ay+b) + \frac{\lambda_\alpha^2 a^2}{4}}.$$

It is the larger root of $(u-y)^2 = \lambda_\alpha(au+b)$ and exceeds $y$. If $\theta > u(y)$, then

$$\theta - x \geq \theta - y > \sqrt{\lambda_\alpha(a\theta+b)}.$$

The strict inequality follows by considering the upward-opening quadratic $(t-y)^2 - \lambda_\alpha(at+b)$ beyond its larger root. Cantelli's lower-tail bound and $\mathrm{Var}(\widehat{\theta}) \leq a\theta + b$ imply $\mathbb{P}\{\theta > u(x_+)\} \leq \alpha$. Finally,

$$0 \leq \theta_i = \frac{1}{K}\sum_k \|P_{ik}^q\|^2 - \|\bar{p}_i\|^2 \leq 1 - \frac{1}{|\mathcal{Z}|},$$

so intersecting the bound with this known maximum cannot remove a valid parameter value. This proves Equation (S40) and Theorem S2. In particular, if a tolerance $\tau$ is prespecified and the true $\theta > \tau$, the event $\mathcal{U}_\alpha \leq \tau$ has probability at most $\alpha$.

**Proposition S1** (Explicit sufficient separation for power)**.** *For the test in Equation* (S39)*, fix a desired type-II error $0 < \beta < 1$, set $\lambda_\beta = (1-\beta)/\beta$, and retain $v_\tau = a\tau + b$ and $h_\tau = \sqrt{\lambda_\alpha v_\tau}$. Define*

$$D_{\alpha,\beta}(\tau) = h_\tau + \frac{\lambda_\beta a}{2} + \sqrt{\lambda_\beta v_\tau + \lambda_\beta a h_\tau + \frac{\lambda_\beta^2 a^2}{4}}. \tag{S52}$$

*Every alternative with $\theta \geq \tau + D_{\alpha,\beta}(\tau)$ has rejection probability at least $1-\beta$.*

**Proof.** Write $\theta = \tau + \delta$. Failure to reject implies $\widehat{\theta} - \theta \leq -(\delta - h_\tau)$. The number $D_{\alpha,\beta}(\tau)$ is the larger root of

$$(\delta - h_\tau)^2 = \lambda_\beta(v_\tau + a\delta),$$

and exceeds $h_\tau$. For $\delta$ at least this root, the lower-tail Cantelli bound is at most

$$\frac{v_\tau + a\delta}{v_\tau + a\delta + (\delta - h_\tau)^2} \leq \frac{1}{1+\lambda_\beta} = \beta.$$

This proves the proposition. If the sufficient separation exceeds the parameter range, no alternative satisfies the power condition.

For two equally weighted orders, let $L = NK$. Since $K \geq 2$ and $R \geq 2$,

$$a = \frac{1}{LR}, \qquad b = \frac{K-1}{8NK^2R(R-1)} \leq \frac{1}{4LR^2}.$$

At $\tau = 0$, write $h_0 = \sqrt{\lambda_\alpha b}$ and use $\sqrt{u+v+w} \leq \sqrt{u} + \sqrt{v} + \sqrt{w}$ and $\sqrt{\lambda_\beta a h_0} \leq (\lambda_\beta a + h_0)/2$ in Equation (S52). This gives

$$\begin{aligned} D_{\alpha,\beta}(0) &\leq \left(\frac{3}{2}\sqrt{\lambda_\alpha} + \sqrt{\lambda_\beta}\right)\sqrt{b} + \frac{3}{2}\lambda_\beta a \\ &\leq \frac{C_{\alpha,\beta}}{R\sqrt{NK}}, \qquad C_{\alpha,\beta} = \frac{\frac{3}{2}\sqrt{\lambda_\alpha} + \sqrt{\lambda_\beta}}{2} + \frac{3}{2}\lambda_\beta. \end{aligned} \tag{S53}$$

The last step uses $a = 1/(LR) \leq 1/(R\sqrt{L})$. This explicit bound establishes the stated separation order. The exact expression (S52) is smaller and is preferable for numerical planning.

### HETEROGENEOUS-ITEM TESTING LOWER BOUND

We prove Theorem S3 directly for every even $K \geq 2$. Use two categories, embedded in the full alphabet by assigning zero probability to the other categories. Let $P_0$ be the fixed-law array with success probability $1/2$ in every cell. Pair the prompts as $(1,2),(3,4),\ldots,(K-1,K)$.

For every fixed sign array $\boldsymbol{\sigma} = (\sigma_{ij}) \in \{-1,1\}^{N\times K/2}$, define an alternative $P_{\boldsymbol{\sigma},\varepsilon}$ by

$$\mathbb{P}(Z = 1 \mid i, 2j-1, o) = \tfrac{1}{2} + \sigma_{ij}\varepsilon, \tag{S54}$$

$$\mathbb{P}(Z = 1 \mid i, 2j, o) = \tfrac{1}{2} - \sigma_{ij}\varepsilon, \qquad o \in \{AB, BA\}. \tag{S55}$$

For each fixed sign array, observations are independent calls from fixed cell laws. The two orders have identical laws, the mean prompt law is $(1/2, 1/2)$, and each prompt contributes $2\varepsilon^2$ to its squared distance from that mean. Thus

$$\theta_i = \theta = 2\varepsilon^2 = \delta \tag{S56}$$

for every alternative. Each item has $K/2$ independent prompt-pair signs; opposite signs within each pair keep the mean prompt law fixed exactly.

For the lower-bound argument only, put the uniform prior on all sign arrays, so the $NK/2$ signs are independent Rademacher variables. Denote the resulting mixture of observation laws by $\overline{P}_\varepsilon$. The prior used in this proof averages observation laws over fixed parameter arrays.

Consider one item and one prompt pair. It contributes $4R$ Bernoulli observations: two prompts, two orders, and $R$ repetitions. Under $P_0$, write each Bernoulli observation as a symmetric variable $Y \in \{-1, 1\}$. If the prompt's fixed orientation within its pair is $t \in \{-1, 1\}$ and the pair sign is $\sigma$, the one-observation likelihood ratio is $1 + 2\sigma t\varepsilon Y$. For two candidate signs $\sigma, \sigma'$, independence of observations therefore gives

$$\mathbb{E}_0[L_\sigma L_{\sigma'}] = (1 + 4\sigma\sigma'\varepsilon^2)^{4R}.$$

Averaging over the two independent signs in the mixture's second moment yields the exact identity

$$1 + \chi^2(\overline{P}_{\varepsilon,\mathrm{pair}}, P_{0,\mathrm{pair}}) = \frac{(1 + 4\varepsilon^2)^{4R} + (1 - 4\varepsilon^2)^{4R}}{2}. \tag{S57}$$

For $0 \le 4\varepsilon^2 < 1$, using $1 + x \le e^x$ and $1 - x \le e^{-x}$ bounds the right side by $\cosh(16R\varepsilon^2)$. The inequality $\cosh u \le e^{u^2/2}$ follows from $\log\cosh u = \int_0^u \tanh v \, \mathrm{d}v \le u^2/2$ for $u \ge 0$, and symmetry for $u < 0$. Thus the one-pair second moment is at most $\exp(128R^2\varepsilon^4)$.

The $NK/2$ item–pair groups are independent under both the prior and the null observation law, so their likelihood-ratio second moments multiply. Substituting $\varepsilon^2 = \delta/2$ from Equation (S56) gives

$$\chi^2(\overline{P}_\varepsilon, P_0) \le \exp(64NKR^2\varepsilon^4) - 1 = \exp(16NKR^2\delta^2) - 1. \tag{S58}$$

Under Equation (S41), the final expression is at most $1/4$. The proposed $\delta$ also satisfies $\delta < 1/2$ for $N \ge 1$, $K \ge 2$, and $R \ge 2$, so $4\varepsilon^2 = 2\delta < 1$ and all Bernoulli probabilities used above are strictly between zero and one.

If $L = \mathrm{d}\overline{P}_\varepsilon/\mathrm{d}P_0$, Cauchy–Schwarz gives

$$\|\overline{P}_\varepsilon - P_0\|_{\mathrm{TV}} = \tfrac{1}{2}\mathbb{E}_0|L - 1| \le \tfrac{1}{2}\sqrt{\mathbb{E}_0(L-1)^2} \le \tfrac{1}{4}.$$

For a possibly randomized test $0 \le \phi \le 1$,

$$\begin{aligned}\mathbb{E}_0\phi + \mathbb{E}_{\overline{P}_\varepsilon}(1 - \phi) &= 1 - \left(\mathbb{E}_{\overline{P}_\varepsilon}\phi - \mathbb{E}_0\phi\right) \\ &\ge 1 - \|\overline{P}_\varepsilon - P_0\|_{\mathrm{TV}} \ge \tfrac{3}{4}.\end{aligned}$$

The supremum over $\mathcal{F}_0$ is at least the error at $P_0$, and the supremum over $\mathcal{F}_\delta$ is at least the prior-average type-II error, because every sign-array alternative has target $\delta$. This proves Equation (S42).

The lower-bound subclass has zero order effect and interaction. A lower bound on this subclass applies to the full fixed-cell class; the upper bound also covers arrays with zero prompt dispersion and nonzero interaction.

For the balanced fixed-repeat design with even $K$, the upper and lower bounds therefore match at order $1/(R\sqrt{NK})$ for fixed error probabilities whose sum is below $3/4$.

# C Acquisition, Uncertainty, and Reproducibility

## C.1 Observation units and study design

For measurement, we use $N$ items, a fixed census of $K$ prompts, two candidate orders, and $R$ repeated calls in each cell. We collect separate calls to form $J$ independent four-vote deployments per item and condition. D selects two distinct prompts and queries both orders for each; E selects four distinct prompts and assigns two to each order. Measurement and deployment use separate sets of calls. Results give equal weight to each item, except for conditional rates with explicitly stated denominators.

Table S1: Distinct study populations and acquired observations. Each study retains its own benchmark population; shared calls are counted once in the controlled-formatting total.

| Study | Measurement / independent deployment | Retained calls |
|---|---|---|
| Controlled formatting | 50 matched items; $K = 6$; null/evaluation $R = 8$, reference $R = 16$; 369 further items at $R = 4$ | 44,112 unique calls across Qwen and the four-configuration panel |
| Semantic comparison | Common 80 items, $K = 24$, $R = 2$; open models additionally use 120 test items and unseen-prompt deployments | 162,816 calls including development/canaries |
| Matched temperatures | Same 40 items, $K = 48$, $R = 4$, $T \in \{1, .2, 0\}$; $J = 8$ D groups | 49,920 calls per model |
| Qwen policies | 152 test items, $K = 6$, $R = 16$; $J = 64$ groups per D/E design and policy | 428,032 test calls |
| Gemma reasoning | 32 RewardBench 2 pairs, $K = 48$, $R = 2$; $J = 8$ D groups in each of four arms | 28,672 formal judgments |
| Independent JudgeBench | 100 development, 200 test items; $K = 6$, $R = 8$ test measurement; $J = 16$ per D/E | 49,600 formal calls |

**Uncertainty targets.** Theorems in Section B and the temperature bounds in Section E.2 concern fresh independent calls conditional on fixed items and prompt assignments. Whole-item bootstrap intervals describe sensitivity to which questions appear in the sample: all prompts, orders, repeats, conditions and independent groups for a selected item are resampled together. The families of temperature decision comparisons condition on the assigned prompt pairs. Deployment trace variance measures variation across deployments. Block-$t$ and bootstrap intervals use approximate calibration; the finite-call bounds have distribution-free guarantees.

### C.2 Rendering, outcome parsing, and model configurations

All pairwise calls preserve the candidate texts. After reversing their order, we map verdicts back to the original candidate identities. The common parser reads A, B or TIE from the first nonempty final line; other completed outputs map to $\perp$. Completed invalid or truncated outputs are retained. Only infrastructure failures without a returned completion are retried under the original conditions. Conditioning within each cell on valid outputs defines a different response law; pooling surviving calls additionally changes cell weights. Requiring at least two valid calls per cell can change which items enter the panel. The primary analysis uses the full four-category law.

The six formatting wrappers use the same judging criterion and required output format while changing headings or delimiters. The shared criterion judges instruction adherence and correctness, forbids rewarding length, style or confidence on their own, and allows TIE. Exact strings, separators and candidate fields are given in Section L.1. The semantic bank has 48 variants, 12 each emphasizing lexical, syntactic, clause-order or discourse edits; six formatting controls are retained. A set of 216 automated checks, including controls, documents their intended equivalence. The four procedural policy banks apply the criterion through distinct judging procedures.

The original four-configuration panel uses the same 50 LLMBar items, six wrappers and $R = 4$ (2,400 calls/configuration); Gemma-3 and the later Gemma4 studies are distinct configurations. Generation settings are recorded in Table S2. The analysis assumes independent calls under stationary cell laws across runtime periods. Each configuration is defined by a fixed model ID and fixed decoding and rendering settings. The cross-model comparison therefore includes both model weights and runtime settings.

### C.3 Data and analysis materials

The accompanying data package provides recorded categorical judgments, prompt and item metadata, gold labels, model settings, independent vote groups, and analysis code for the main numerical results. Fixed probability arrays and trial summaries support the known-law simulations. The analyses retain all conditions, invalid outputs and negative component estimates.

Table S2: Judge configurations in the matched comparison. Each configuration includes its model, renderer, decoder, stopping rule, and categorical parser. All item–prompt–order cells are shared across configurations.

| Configuration | Model | Renderer | Decoding and terminal rule |
|---|---|---|---|
| Qwen3.6-27B | `Qwen/Qwen3.6-27B` | Qwen chat renderer | $T = 1$, top-$p = 1$, 8-token cap; categorical parser |
| GPT-OSS-120B | `openai/gpt-oss-120b` | Harmony renderer | $T = 1$, top-$p = 1$, 512-token cap; final channel only |
| Gemma-3-27B-IT | `google/gemma-3-27b-it` | official chat template | $T = 1$, top-$p = 1$, top-$k = -1$, 32-token cap; newline stop |
| Claude Haiku 4.5 | `claude-haiku-4-5-20251001` | Messages request | $T = 1$; top-$p$ omitted (provider default); no provider seed; 512-token cap; provider `end_turn` |

Table S3: Fixed configurations in the five-model extension. Output caps specify token limits; computational budgets vary across models. Hosted GPT sampling uses its supported reasoning-effort setting; requests omit temperature and top-$p$.

| Configuration | Recorded model ID | Cap | Generation settings |
|---|---|---|---|
| Qwen3.6 | `Qwen/Qwen3.6-27B` | 512 | Thinking off; $T = 1$ |
| Gemma4 | `google/gemma-4-31B-it` | 512 | Thinking off; $T = 1$ |
| GPT-OSS | `openai/gpt-oss-120b` | 4096 | Low effort; $T = 1$ |
| GPT Luna | `gpt-5.6-luna` | 4096 | Low effort |
| Haiku4.5 | `claude-haiku-4-5-20251001` | 512 | $T = 1$ |

# D Measurement Validation and Cross-Model Evidence

## D.1 Identical-input null and independent reference

The controlled study uses 50 LLMBar items, ten per stratum. The Qwen corpus contains 4,800 null-control calls with byte-identical inputs, 4,800 matched $R = 8$ evaluation calls, a disjoint 9,600-call $R = 16$ reference, and 17,712 calls on the remaining 369 items at $R = 4$. A matched four-configuration panel reuses 2,400 Qwen calls and adds 7,200 calls, giving 44,112 unique outcomes. Within each item–order–repeat unit, the six null labels share exactly the same model-visible input but have distinct generation seeds. The evaluation and reference samples share items, prompts, orders, checkpoint and parser but use disjoint calls and seeds. Analyses at lower repeat depths use fixed prefixes of the recorded calls.

Table S4: Prompt-instability estimates in the live experiments. Entries are means across items, with 95% percentile bands for benchmark composition from 20,000 stratified whole-item bootstrap draws. Each resample retains all six hidden labels or templates and both answer orders. The hidden-factor null has an exactly zero target because its labels induce byte-identical inputs. The other rows concern the fixed six-prompt census.

| Setting | $R$ | Plug-in | Debiased all-$R$ |
|---|---|---|---|
| Byte-identical hidden-factor null | 2 | 0.0406 [0.0313, 0.0505] | -0.0053 [-0.0139, 0.0038] |
| | 4 | 0.0206 [0.0153, 0.0263] | -0.0008 [-0.0050, 0.0033] |
| | 8 | 0.0126 [0.0089, 0.0168] | 0.0024 [-0.0002, 0.0054] |
| Matched evaluation | 2 | 0.0392 [0.0289, 0.0501] | 0.0007 [-0.0056, 0.0072] |
| | 4 | 0.0251 [0.0191, 0.0313] | 0.0052 [0.0011, 0.0096] |
| | 8 | 0.0143 [0.0107, 0.0179] | 0.0039 [0.0013, 0.0066] |
| Broader item sample | 4 | 0.0245 [0.0223, 0.0268] | 0.0026 [0.0010, 0.0044] |

All bands in this subsection use 20,000 stratified whole-item resamples to describe variation in item composition. Compared with the independently acquired unbiased $R = 16$ reference, complete-pair correction lowers mean absolute discrepancy at $R = 2, 4, 8$. After acquisition, contiguous

Table S5: Itemwise MAE against the independent $R = 16$ Qwen reference on the same 50 items, six prompts and two orders. The evaluation and reference samples use disjoint calls; corrected estimates use all distinct-call pairs. Reduction is the paired difference $\mathrm{MAE}_{\mathrm{plug}} - \mathrm{MAE}_{\mathrm{corr}}$, with a pointwise 95% composition interval from 20,000 stratified whole-item resamples.

| Evaluation $R$ | Plug-in MAE | Corrected MAE | Reduction [95% interval] |
|---|---|---|---|
| 2 | 0.0365 | 0.0160 | 0.0205 [0.0131, 0.0288] |
| 4 | 0.0214 | 0.0093 | 0.0120 [0.0077, 0.0165] |
| 8 | 0.0112 | 0.0076 | 0.0036 [0.0009, 0.0064] |

Table S6: Raw, unclipped complete-pair decomposition for the matched $R = 4$ panel. Values are macroaverages over the same 50 items. Every row obeys the sample decomposition identity up to display rounding.

| Judge configuration | Call | Prompt | Answer order | Interaction | Total |
|---|---|---|---|---|---|
| Qwen3.6-27B | .19111 | .00521 | .04694 | .00719 | .25046 |
| GPT-OSS-120B | .08028 | −.00123 | .00935 | .00228 | .09068 |
| Gemma-3-27B-IT | .01500 | .01774 | .11017 | .01733 | .16024 |
| Claude Haiku 4.5 | .05917 | .00986 | .08826 | .00799 | .16528 |

and interleaved splits of the reference into two independent $R = 8$ halves give MAEs of .00758 and .00819 (correlations .150 and .197). For evaluation estimates $A_i, C_i$ and independent unbiased reference $T_i$, $\mathbb{E}[(A_i - T_i)^2 - (C_i - T_i)^2] = \mathbb{E}(A_i - \theta_i)^2 - \mathbb{E}(C_i - \theta_i)^2$. Observed squared-loss improvements are .002311, .000716 and .000179. The choice to report complete all-$R$ estimates was made after acquisition. Consecutive-two-repeat estimates give an estimated reference-MAE reduction of $-.0005$ at $R = 8$, with composition band [$-.0041$,.0029]; complete pairs use additional cross-repeat information.

The independent reference gives mean components of .1957 for call variation, .0040 for prompts, .0511 for order and .0040 for interaction. The disjoint 369-item extension also has small corrected prompt dispersion (.0026, composition band [.0010,.0044]), with component means of .2098 for call variation, .0521 for order and .0028 for interaction. This extends the profile within LLMBar under the same wrappers and configuration. The independent $R = 16$ reference covers the initial 50-item panel.

### D.2 Matched model configurations and invalid outputs

The separate semantic comparison uses the same 80 test items (40/dataset), 24 prompts, both orders and $R = 2$ across five configurations. Each open model's larger acquisition includes 120 test items, 40 development items, four canaries and a four-item identical-input diagnostic, with 48 semantic prompts plus six controls on the test sample. Closed models use the nested common panel. The complete extension has 162,816 outcomes. Unseen-prompt deployments for the open models are analyzed separately in Section H.

Claude's order-minus-call point difference in the matched formatting panel is .02910, with a composition band of [$-.01493$,.07514]. The matched formatting panel contains 10 Qwen and 64 GPT-OSS invalid outputs and none for Gemma-3 or Haiku. Recoding $\perp$ as TIE changes the panel's prompt means by at most .00049 at $R = 2$ and .00036 at $R = 4$. Haiku's semantic JudgeBench condition has 81.74% invalid completions and 56.35% cap stops. Its response profile therefore primarily measures output usability. Retaining $\perp$ keeps these configuration-dependent outcomes in the measured law.

**Dependence checks.** We assess monotone drift in the seven original acquisition streams using agreement slopes against rank gap and elapsed time. Whole-item resampling preserves the dependence among agreement pairs that share calls. All outputs remain included; complete-case filtering at $R = 2$ would retain 42/50 null, 45/50 matched and 329/369 broader items.

Table S7: Common semantic-prompt panel. P is the plug-in prompt component; $\widehat{U}_p$, $\widehat{U}_o$, and $\widehat{U}_{po}$ are corrected prompt, order, and interaction components. Accuracy, $\perp$, and truncation are percentages of all completed calls. Values are configuration-specific point estimates. The Haiku JudgeBench condition is reported separately as an output-protocol case: invalid outputs dominate its response profile. The remaining rows support comparisons of judge prompt sensitivity.

| Model | Dataset | P | $\widehat{U}_p$ | $\widehat{U}_o$ | $\widehat{U}_{po}$ | Acc. | $\perp$ | Trunc. |
|---|---|---|---|---|---|---|---|---|
| Qwen3.6 | LLMBar | 0.0718 | 0.0065 | 0.0224 | 0.0116 | 72.27 | 0.36 | 0.05 |
| Qwen3.6 | JudgeBench | 0.0965 | 0.0118 | 0.1102 | 0.0108 | 53.12 | 11.02 | 8.70 |
| Gemma4 | LLMBar | 0.0190 | 0.0123 | 0.0410 | 0.0143 | 82.60 | 0.03 | 0.21 |
| Gemma4 | JudgeBench | 0.0272 | 0.0136 | 0.1110 | 0.0140 | 64.11 | 1.04 | 1.72 |
| GPT-OSS | LLMBar | 0.0336 | 0.0090 | 0.0271 | 0.0027 | 79.35 | 0.00 | 0.00 |
| GPT-OSS | JudgeBench | 0.0587 | 0.0123 | 0.0070 | 0.0024 | 63.46 | 0.05 | 0.03 |
| GPT Luna | LLMBar | 0.0338 | 0.0092 | 0.0512 | 0.0118 | 75.03 | 0.00 | 0.00 |
| GPT Luna | JudgeBench | 0.0454 | 0.0106 | 0.0174 | 0.0132 | 71.46 | 0.00 | 0.00 |
| Haiku4.5 | LLMBar | 0.0265 | 0.0079 | 0.0652 | 0.0064 | 69.71 | 0.94 | 0.00 |
| *Output-protocol case: 81.74% invalid completed outputs* | | | | | | | | |
| Haiku4.5 | JudgeBench | 0.0549 | 0.0187 | 0.0304 | 0.0092 | 11.15 | 81.74 | 56.35 |

### D.3 Known-law validation

The basic bias experiment crosses $K \in \{5, 6\}$, $R \in \{2, 4, 8\}$, two balanced orders and prompt targets zero or .03, with 200,000 independent items per setting. The base law is $(.4, .3, .2, .1)$; the alternative moves mass along $(1, -1, 0, 0)/\sqrt{2}$ with a centered, unit-variance census score. At $K = 6$, the null plug-in biases of .1459, .0730 and .0365 match the $1/R$ identity. Corrected bias lies within 1.44 Monte Carlo standard errors of zero across all 12 settings. Negative null estimates (52.7–56.1% itemwise) are retained.

A separate validation uses 170 fixed settings with 5,000 independent datasets each: 13 analytical families at $N = 32$ and 152, plus eight data-shaped arrays at $N = 152$, crossed with $R = 2, 4, 8, 16, 32$, $K = 6$ and two balanced orders. The analytical grid includes homogeneous, heterogeneous and interaction-only prompt nulls, a stability-tolerance boundary, and signed D/E contrasts. Data-shaped laws use categorical probabilities $(n_c + .5)/18$ from 16-repeat measurements. Treating these arrays as fixed synthetic laws gives the simulations exact known targets. Stored arrays, seeds and trial summaries permit exact reconstruction of each scenario.

The universal and observed-variation design intervals cover their targets in all 5,000 datasets in every setting. The Wilson 95% Monte Carlo lower bound of 99.923% quantifies uncertainty in each observed coverage rate. The approximate independent-block $t$ comparison has coverage of 94.34–95.66% on this grid. Section E.4 separately evaluates the two-panel temperature approximation. At the interaction-only prompt null with $U_{\text{int}} = .08$, $N = 152$ and $R = 8$, the fraction of datasets that certify the fixed tolerance $\theta < .01$ is 0%, 79.06% and 97.10% for the original, trace-refined and sharp categorical covariance bounds. No tested boundary case $\theta = .01$ falsely certifies the tolerance (per-setting Wilson upper bound .077%). In the lower-bound family, four designs $(N, K, R) = (16, 4, 4), (64, 4, 4), (16, 8, 16), (64, 8, 16)$ use $\delta = c/(R\sqrt{NK})$ and 5,000 datasets/setting. Power is .04–.12% at $c = .5$, 56.92–65.16% at $c = 2$, and 100% at $c = 8$ for the trace-envelope test. These power simulations show how the proven separation rate translates into finite-call resolution.

### D.4 Explicit laws and matched estimator comparisons

The 170-setting study tests inference across heterogeneous fixed laws. A separate ten-law study isolates estimator efficiency and equal-budget design selection. Both studies generate independent categorical calls with known targets. Section E.4 reports the separate temperature-panel calibration.

**The 170-setting probability arrays.** For $k = 1, \ldots, 6$, define $h = (-1, -1, -1, 1, 1, 1)$, $g = \sqrt{3/2}(-1, 0, 1, -1, 0, 1)$ and $u = (1, -1, 0, 0)/\sqrt{2}$. The prompt directions are centered, orthogonal and have unit mean square. For item $i = 0, \ldots, N - 1$, let $b_i = (.44 + t_i, .36 - t_i, .15, .05)$, where

$t_i = .04\sin(1.61803398875\, i)$. Nine families use

$$\begin{aligned} P_{ik,AB} &= b_i + \sqrt{S}\, h_k u, \\ P_{ik,BA} &= b_i + \sqrt{S}(\rho h_k + \sqrt{1-\rho^2} g_k) u, \end{aligned} \qquad S = .04, \quad \rho = -\frac{\overline{\Delta}}{\gamma S}, \quad \gamma = \frac{6}{20}, \tag{S59}$$

with $\overline{\Delta} \in \{-.012, -.006, -.003, -.001, 0, .001, .003, .006, .012\}$. Hence

$$\theta = \frac{S - \overline{\Delta}/\gamma}{2}, \qquad \overline{U}_{\text{int}} = \frac{S + \overline{\Delta}/\gamma}{2},$$

so $\overline{\Delta} = .006$ puts prompt dispersion at the tolerance boundary $\theta = .01$. Four further families have $\theta = 0$: the constant binary law $(.5, .5, 0, 0)$; laws $b_i \pm (.1 + .03\cos i)u$ constant across prompts; and interaction-only laws $b_i \pm \sqrt{S} h_k u$ for $S = .005, .08$, with signs denoting AB/BA. Each of these 13 families uses $N = 32$ and 152.

The other eight arrays use the four output-cap and four policy conditions, each with $N = 152$. A cell's 16 observed category counts define $P_c = (n_c + .5)/18$. We treat these probabilities as fixed synthetic laws, retaining the dispersion present in the finite source sample. Combining the 34 arrays with $R = 2, 4, 8, 16, 32$ gives 170 settings, each with 5,000 independent datasets and its own random stream. Methods share calls within each dataset; random streams are independent across settings.

Figure 5a reports coverage for each setting. Panel (b) averages full interval widths over the eight data-shaped laws at $N = 152$, with shading from the minimum to the maximum law-specific mean. Panel (c) uses the interaction-only null with $S = .08, N = 152$. Panel (d) uses the nine contrast families at $N = 152, R = 32$; gray marks the practical tolerance $\pm.005$. Rate bars are 95% Wilson intervals over 5,000 datasets per setting.

**Point estimators compared.** Write $\widehat{\theta} = N^{-1}\sum_i \widehat{U}^{(q)}_{p,i}$ for the complete-pair macroaverage. The five prompt estimates are the plug-in macroaverage, $\widehat{\theta}$, $\max(0, \widehat{\theta})$, $\widehat{\theta}/2$, and zero, all scored by $\mathbb{E}(\widehat{t} - \theta)^2$. The classical trace-mean-square estimate equals $\widehat{\theta}$ sample by sample (Section A.4); the maximum numerical discrepancy was $2.78 \times 10^{-16}$. For $\overline{\Delta}$, we compare the pooled unbiased estimate, fixed half shrinkage, zero and the independent-panel mean. The shrinkage factor is fixed before generating test draws.

At the interaction-only null with $U_{\text{int}} = .08$, $N = 152, R = 8$, prompt MSE is $8.82 \times 10^{-4}$ for plug-in, $1.07 \times 10^{-6}$ for complete pairs, and $5.28 \times 10^{-7}$ for the macro positive part; zero has zero error at this null. For $S = .04$, $\overline{\Delta} = .001$ and $N = 152, R = 8$, the unbiased and half-shrunk contrast estimates have MSEs of $5.57/3.86 \times 10^{-7}$, respectively, versus $10^{-6}$ for zero. At $R = 32$ they become $1.03/2.78 \times 10^{-7}$; at $\overline{\Delta} = .006, R = 8$ they are $5.64 \times 10^{-7}$ and $9.09 \times 10^{-6}$. These comparisons quantify how signal size and precision affect estimation error.

**Ten-law efficiency and allocation study.** Within each scenario, all items share the same law. Set $a_k = (k - 3.5)/\sqrt{35/12}$ and $z_k = (a_k^2 - 1)/\{6^{-1}\sum_\ell (a_\ell^2 - 1)^2\}^{1/2}$. The centered, orthogonal directions have unit mean square. With the same $u$ as above,

$$P_{k,AB} = p_0 + A a_k u, \qquad P_{k,BA} = p_0 + (B a_k + C z_k) u.$$

Table S8 specifies all ten arrays. The prompt module uses $N = 50$, $R = 2, 4, 8$ and 5,000 independent outer datasets per law. Repeat depths use nested prefixes within a dataset. Disjoint-panel estimates average the complete-pair estimates from consecutive two-repeat blocks, whereas all-$R$ estimates also use cross-block pairs. Monte Carlo uncertainty is calculated across independent outer datasets.

At the pure-interaction prompt null ($U_{\text{int}} = .0196$), $R = 8$ aggregate MSEs are $3.96 \times 10^{-6}$ for all-$R$, $2.13 \times 10^{-5}$ for disjoint panels and $1.03 \times 10^{-3}$ for plug-in. Itemwise clipping, $N^{-1}\sum_i \max(0, \widehat{U}^{(q)}_{p,i})$, reduces mean itemwise MSE from $1.99 \times 10^{-4}$ to $1.23 \times 10^{-4}$. Its positive bias raises aggregate MSE to $3.22 \times 10^{-5}$. Clipping *after* averaging projects $\widehat{\theta}$ onto $[0, \infty)$ and weakly reduces squared error for every nonnegative macro target. Across 27 positive-variance law/depth settings, empirical/exact variance ratios are .963–1.051 and oracle-normal coverage is 94.3–95.6%; these intervals use variances calculated from the known laws.

Table S8: Complete ten-law generator. The default $p_0 = (.50, .30, .15, .05)$ lists probabilities for candidate 1, candidate 2, TIE and $\perp$, in that order. Only the two indicated bases differ. All probabilities are valid.

| Family | $A$ | $B$ | $C$ | Alternative $p_0$ |
|---|---|---|---|---|
| No crossed effects | 0 | 0 | 0 | — |
| Deterministic null | 0 | 0 | 0 | $(1, 0, 0, 0)$ |
| Small prompt | .025 | .025 | 0 | — |
| Prompt boundary | .10 | .10 | 0 | — |
| Pure prompt | .14 | .14 | 0 | — |
| Pure interaction | .14 | $-.14$ | 0 | — |
| Regular design tie | .14 | 0 | .14 | — |
| Near tie, E favored | .14 | .01 | .14 | — |
| Near tie, D favored | .14 | $-.01$ | .14 | — |
| High invalid probability | .14 | .14 | 0 | $(.38, .27, .10, .25)$ |

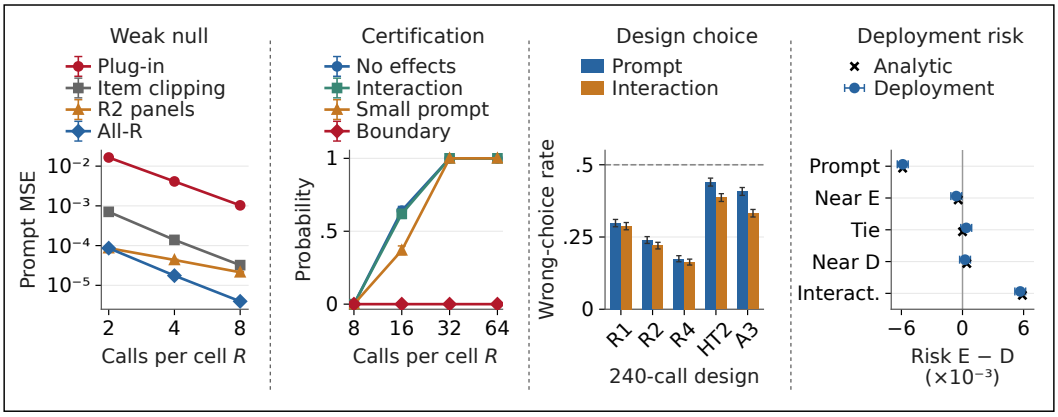


Figure S1: Distinct checks under controlled categorical laws. From left to right: weak-null prompt MSE versus repeat depth; certification of $\theta < .01$ versus depth; wrong sign-based design choices at 240 measurement calls; independently observed versus analytic D/E risk contrasts. MSE and risk bars are $\pm 1.96$ Monte Carlo standard errors over 5,000 outer datasets. Rate bars are 95% Wilson intervals, using 1,000 datasets/law for the separate depth diagnostic and 5,000 for design choice. R1/R2/R4 denote census repeat depths; HT2 (Pair2) and A3 denote the two-call pair and three-call anchor. Near E/D are the corresponding nonzero near-tie laws. The bars quantify Monte Carlo uncertainty under fixed known laws.

The separate depth diagnostic uses four fixed laws (no effects, pure interaction, small prompt and boundary), $N = 50$, 1,000 datasets/law, and $R = 8, 16, 32, 64$. It uses the original Cantelli constant $b^{\text{orig}} = (K-1)/[4NK^2R(R-1)]$, whose minimum 95% upper endpoint is .01981 at $R = 8$. Pure-interaction certification rates for $\theta < .01$ are 0/61.9/100/100%; the small-prompt law reaches 36.9% at $R = 16$ and 100% at $R = 32$. No boundary case falsely certifies at any depth (per-setting Wilson upper endpoint .38%). The fixed grid measures the original bound's resolution at prespecified budgets.

For equal-budget selection, 240 calls permit census $R = 1, 2, 4$ on 20, 10 and 5 items, or 120 independent Pair2 units or 80 Anchor3 units (Equations S75–S76). All methods estimate the same $\Delta$ under the homogeneous scenario law. Forced sign choice selects E for a negative estimate and D otherwise. Under pure prompt variation, wrong-choice rates are 29.8/23.9/17.5% for the census designs, 44.0% for Pair2 and 40.8% for Anchor3; census $R = 4$ has wrong-choice rates of 46.4–46.8% near a regular tie. The universal-bound commitment rule abstains throughout the tested 120- and 240-call configurations. Measurement draws are independent across methods; all methods share a separate 100-item deployment evaluation within each outer dataset. Observed $E - D$ vector-risk differences are $-.00587$ and .00567 under pure prompt and pure interaction, respectively (MCSE .00027 each), versus analytic $-.00588$ and .00588. These homogeneous laws quantify the precision gain from additional repeats; Section I.4 evaluates allocation under item heterogeneity.

# E MATCHED TEMPERATURE INTERVENTIONS

## E.1 SHARED ACQUISITION AND OUTCOME DEFINITIONS

Both Qwen/Qwen3.6-27B and `google/gemma-4-31B-it` use the same fixed 40 items (20 LLMBar, 20 JudgeBench), all 48 prompts, two orders, and temperatures $1, .2, 0$. Four repeats give 46,080 measurement calls per model. Eight independent D deployment groups per item and temperature add 3,840 calls, for 49,920 formal calls per model. Prompt-pair assignments are matched across temperatures and models; measurement and deployment calls are generated separately. All invalid and capped completions enter the four-category law.

Both configurations use bfloat16, native nonthinking, a 512-token output cap, top-$p = 1$, top-$k$ disabled, unconstrained final generation and the fixed first-nonempty-line parser. Qwen runs on one H100 PCIe with tensor parallelism one; Gemma runs on one H200 NVL with a 16,384-token context. The serving stack is vLLM 0.25.1, Torch 2.11.0+cu130, Transformers 5.14.1 and tokenizers 0.22.2. Temperature conditions are interleaved within logical units with distinct recorded generation seeds. Gemma's 681 technical calls are separate from the formal sample.

Under the original rule, a deployment returns the unique plurality winner among candidate 1, candidate 2, TIE and $\perp$; a tied maximum is TIE. A unique invalid winner remains invalid. Fair resolution changes only a pure candidate 2:2 vote vector, assigning exact expected accuracy $1/2$, wrong-candidate frequency $1/2$ and coverage one. All other outcomes retain their original verdict; candidate selection is independent of the gold label. Thus $A_f = A_o + T_{22}/2$, $W_f = W_o + T_{22}/2$, $C_f = C_o + T_{22}$, using exact expectations. Every deployment is included in the denominator of each decision rate.

**Analysis timing and endpoint families.** The four component contrasts compare call and prompt variation between each cooling temperature and $T = 1$. Complete-pair $R = 4$ and complete-group $J = 8$ analyses use all acquired observations. Their finite-sample constructions were selected after collection. Qwen rule rescoring was developed after inspecting descriptive outcomes; Gemma's rule and cost families were specified before its collection. The direct rule interaction is a secondary analysis of the stored groups. Each coverage guarantee applies to its stated model, reporting pool and endpoint family.

## E.2 USING ALL FOUR REPEATS

For each temperature, the complete-pair estimator uses all six distinct-call pairs within each cell. Write $x_t$ and $y_t$ for its prompt and call estimates. Conditional on the fixed items and prompts, Theorem S1 gives

$$\mathrm{Var}(x_t) \le a\theta_t + b, \qquad a = \frac{1}{4NK}, \qquad b = \frac{K-1}{96NK^2},$$

where $\theta_t$ is the population prompt component for the four-category response law. Using all four repeats leaves the linear coefficient unchanged and reduces the quadratic coefficient by a factor of three relative to averaging two $R = 2$ panels.

Set $\eta = .025/6$, $\lambda = (1-\eta)/\eta$, and $x_+ = \max(x, 0)$. Inverting the two one-sided Cantelli inequalities gives

$$U(x) = \min\left\{\frac{3}{4}, x_+ + \frac{\lambda a}{2} + \sqrt{\lambda(ax_+ + b) + \frac{\lambda^2 a^2}{4}}\right\},$$

$$L(x) = \begin{cases} 0, & x \le \sqrt{\lambda b}, \\ \max\left\{0, x + \dfrac{\lambda a}{2} - \sqrt{\lambda(ax+b) + \dfrac{\lambda^2 a^2}{4}}\right\}, & x > \sqrt{\lambda b}. \end{cases}$$

Each limit has one-sided error at most $\eta$. The prompt-contrast interval for $t \in \{.2, 0\}$ is $[L(x_t) - U(x_1), U(x_t) - L(x_1)]$.

For call variation, partitioning the four repeats into two disjoint pairs produces $M = 4NK$ independent mismatch indicators at each temperature. The complete-pair estimate is the conditional mean of

this partitioned estimate given the cell counts. Jensen's inequality gives the same Hoeffding bound for the complete-pair estimate:

$$\Pr(|(y_t - y_1) - (U_{\text{call},t} - U_{\text{call},1})| \geq h) \leq 2e^{-Mh^2}.$$

We use $h = \sqrt{\log(2/.0125)/M}$ for each call contrast. The six prompt limits and two call intervals have total error $6\eta + 2(.0125) = .05$, giving simultaneous coverage at least 95% for the four contrasts within each model under independent calls with fixed categorical laws. Table S10 reports these intervals using pooled $R = 4$ point estimates throughout. Paired item-composition intervals use the same estimates and 20,000 resamples that preserve dataset and stratum; they describe changes in the questions represented by the panel.

Table S9: Full-repeat finite-sample intervals in all ten calibration settings, with 5,000 simulated datasets per setting. Coverage concerns all four contrasts jointly. FWER counts exclusion of zero for any true-null contrast. Width is the mean width of the two prompt intervals; the split-panel column uses finite-sample bounds on the same calls.

| Setting | Coverage (%) | FWER (%) | Full $R = 4$ width | Split-panel width |
|---|---|---|---|---|
| Homogeneous binary null | 100.00 | 0.00 | .1113 | .1619 |
| Interaction-only weak null | 100.00 | 0.00 | .1107 | .1610 |
| Heterogeneous null | 100.00 | 0.00 | .1190 | .1678 |
| Four-category null | 100.00 | 0.00 | .1210 | .1695 |
| Near-deterministic null | 100.00 | 0.00 | .1089 | .1584 |
| Temperature-shaped signal | 100.00 | – | .1689 | .2218 |
| Prompt null; call signal | 100.00 | 0.00 | .1108 | .1611 |
| 10 stochastic items | 100.00 | 0.00 | .1101 | .1601 |
| 2 stochastic items | 100.00 | 0.00 | .1094 | .1592 |
| 1 stochastic item | 100.00 | 0.00 | .1093 | .1590 |

All intervals were available in every setting. The Wilson 95% Monte Carlo interval for each 100% coverage estimate is [99.923%,100%]; for each zero FWER estimate it is [0%,.077%].

Table S10: Temperature contrasts using all four repeats. Call-level intervals have simultaneous coverage of at least 95% for the four contrasts of each model under independent fixed cell laws. Composition intervals are separate pointwise 95% paired resampling summaries. Both interval columns use the same complete-pair $R = 4$ point estimate.

| Model | Contrast | Component | Estimate | Call-level interval | Composition interval |
|---|---|---|---|---|---|
| Qwen | $.2 - 1$ | Call | $-.27135$ | $[-.29706, -.24565]$ | $[-.29228, -.25265]$ |
| Qwen | $.2 - 1$ | Prompt | .02200 | $[-.06996, .09928]$ | $[.01514, .02901]$ |
| Qwen | $0 - 1$ | Call | $-.34818$ | $[-.37388, -.32247]$ | $[-.37860, -.32010]$ |
| Qwen | $0 - 1$ | Prompt | .04262 | $[-.05699, .12589]$ | $[.03065, .05448]$ |
| Gemma | $.2 - 1$ | Call | $-.03659$ | $[-.06230, -.01088]$ | $[-.05482, -.02036]$ |
| Gemma | $.2 - 1$ | Prompt | .00999 | $[-.07987, .09314]$ | $[.00488, .01592]$ |
| Gemma | $0 - 1$ | Call | $-.04488$ | $[-.07059, -.01917]$ | $[-.06736, -.02526]$ |
| Gemma | $0 - 1$ | Prompt | .01421 | $[-.07987, .09869]$ | $[.00731, .02261]$ |

### E.3 Finite-sample intervals for independent deployments

The decision contrasts use all eight independently acquired groups per item. We condition on the item census and assigned prompt pairs, averaging over fresh generation at those assignments. Groups may have different response laws; the two temperatures within a group are paired. Write $D_j \in [-1, 1]$ for a paired difference in accuracy, wrong-decision frequency, or coverage, and $n = 8N$. For $\lambda = 1/2$ and $\psi = \log 2 - 1/2$, define

$$\widehat{\Delta} = \frac{1}{n}\sum_{j=1}^{n} D_j, \qquad h_M = \frac{\log(2M/.05) + \psi \sum_{j=1}^{n} D_j^2}{\lambda n}. \tag{S60}$$

The intervals $[\widehat{\Delta} - h_M, \widehat{\Delta} + h_M] \cap [-1, 1]$ have simultaneous coverage at least 95% for a family of $M$ endpoints. This fixed-time empirical-Bernstein construction uses a zero predictable center for each contrast (Howard et al., 2021). The value of $\lambda$ was fixed for the post-acquisition analysis and

used across all endpoints, models and reporting pools. The interval has positive width even when the observed second moment is zero.

For $x \geq -1$, $\exp(\lambda x - \psi x^2) \leq 1 + \lambda x$. With $\mu_j = \mathbb{E}D_j$,

$$\mathbb{E}\exp\{\lambda(D_j - \mu_j) - \psi D_j^2\} \leq e^{-\lambda\mu_j}(1 + \lambda\mu_j) \leq 1.$$

Independence across groups lets us multiply these bounds. Markov's inequality for both signs and a union bound over the $M$ endpoints prove the claim with unequal group means and variances. This fixed-$\lambda$ construction yields the reported intervals.

The two-rule family has $M = 12$: two cooling contrasts, two rules, and three scores. Adding candidate-only majority gives an $M = 18$ family; its complete bounds are included in the accompanying numerical results. Each family is defined within one model and reporting pool. Item-composition intervals describe a separate source of variation.

**Direct interaction between temperature and the tie rule.** For each cooling level $t \in \{.2, 0\}$, write $\Delta_t$ for the contrast with $T = 1$. The rule interaction is

$$\Gamma_t = \Delta_t A_{\text{coin22}} - \Delta_t A_{\text{strict}} = \tfrac{1}{2}\Delta_t T_{22}. \tag{S61}$$

This direct contrast was specified for a secondary analysis of the stored groups. For paired group $j$, let $X_j = \mathbf{1}\{\text{pure tie at } t\} - \mathbf{1}\{\text{pure tie at } 1\} \in [-1, 1]$. Apply Equation (S60) to $X_j$ with the same $\lambda = 1/2$, then divide the estimate and radius by two. The family has $M = 4$: both cooling contrasts in both models. Thus $\widehat{\Gamma}_t = \sum_j X_j/(2n)$ and the half-width is

$$\frac{\log(8/.05) + (\log 2 - 1/2)\sum_j X_j^2}{n}, \qquad n = 320.$$

The intervals are clipped to $[-1/2, 1/2]$ and condition on the fixed items and assigned prompt pairs. They quantify the change between rules directly, retaining the pairing of the same votes. Their simultaneous guarantee covers these four interactions as a separate family.

Table S11: Temperature-by-rule interactions, in percentage points. A positive value means fair resolution of pure 2:2 ties increases the accuracy effect of cooling relative to the original rule. The finite-sample 95% intervals are simultaneous across all four rows. Each row uses 320 paired independent deployment groups.

| Model | Temperature contrast | $\widehat{\Gamma}_t$ | Simultaneous interval |
|---|---|---|---|
| Qwen | $.2 - 1$ | 7.34 | $[-.34, 15.03]$ |
| Qwen | $0 - 1$ | 9.84 | $[1.68, 18.01]$ |
| Gemma | $.2 - 1$ | .63 | $[-1.44, 2.69]$ |
| Gemma | $0 - 1$ | .47 | $[-1.54, 2.48]$ |

The groupwise identity in Equation (S61) holds for all 256 categorical four-vote patterns with either candidate as gold, and for all 1,920 stored groups. Tables S17 and S18 report the accuracy contrasts under each rule.

For abstention costs, $\Delta L(c) = (1 - c)\Delta W + c(-\Delta A)$. Applying Equation (S60) jointly to the eight endpoints of the four cooling/rule curves gives simultaneous bands for every $c \in [0, 1]$ by linear interpolation. Crossover sets are obtained by inverting these bands, retaining full, empty, and boundary-truncated sets.

All finite-sample intervals were available. The separate eight-endpoint cost family had joint coverage of 99.96–100% across the same settings. The Wilson 95% Monte Carlo interval for 5,000 successes is [99.923%,100%]. The coverage guarantee follows from Equation (S60).

### E.4 Calibration under concentrated call variation

An approximate alternative partitions $R = 4$ into two independent $R = 2$ panels and compares their studentized difference with a $t_{39}$ reference distribution. Calibration uses the exact $N = 40, K = 48$, two-order design and 5,000 independent datasets in each of ten fixed-law settings. The settings include interaction-only weak nulls, invalid-output laws, near-deterministic laws, temperature signals

Table S12: Decision-interval calibration with 5,000 datasets per setting. Coverage requires all 12 intervals to be available and cover their targets. Unavailable intervals under the two-panel approximation are reported separately. Width is the mean full interval width in percentage points; other entries are percentages.

| Scenario | Finite coverage | Width | Panel coverage | Panel FWER | Unavailable |
|---|---|---|---|---|---|
| Distributed, null | 100.00 | 37.07 | 96.46 | 3.54 | 0.00 |
| Distributed, shift | 100.00 | 35.59 | 96.88 | 0.32 | 0.00 |
| Heterogeneous, null | 99.98 | 23.05 | 0.00 | 2.48 | 100.00 |
| Heterogeneous, shift | 100.00 | 23.12 | 0.00 | 0.58 | 100.00 |
| 10 stochastic items | 100.00 | 15.06 | 89.12 | 10.88 | 0.00 |
| 2 stochastic items | 100.00 | 9.19 | 41.74 | 29.68 | 42.68 |
| 1 stochastic item | 100.00 | 8.45 | 6.92 | 21.82 | 91.68 |
| Rare events, null | 100.00 | 8.74 | 0.00 | 0.22 | 100.00 |
| Rare events, shift | 100.00 | 8.84 | 0.00 | 1.30 | 100.00 |
| Shared randomness, shift | 99.94 | 16.09 | 0.00 | 0.00 | 100.00 |

Table S13: Calibration of the two-panel procedure. Simultaneous coverage counts a dataset as successful only when all four intervals are available and cover their targets. Parentheses give Wilson 95% Monte Carlo intervals. FWER is the probability that Holm rejects at least one true null. All settings have 5,000 draws.

| Setting | Four-interval coverage (%) | FWER (%) | Availability (%) |
|---|---|---|---|
| Homogeneous binary null | 96.26 (95.70, 96.75) | 3.74 | 100.00 |
| Interaction-only weak null | 95.40 (94.78, 95.95) | 4.60 | 100.00 |
| Heterogeneous null | 96.18 (95.61, 96.68) | 3.82 | 100.00 |
| Four-category null | 96.36 (95.80, 96.84) | 3.64 | 100.00 |
| Near-deterministic null | 97.00 (96.49, 97.44) | 3.00 | 100.00 |
| Temperature-shaped signal | 96.02 (95.44, 96.53) | – | 100.00 |
| Prompt null; call signal | 96.12 (95.55, 96.62) | 4.58 | 100.00 |
| 10 stochastic items | 92.22 (91.44, 92.93) | 7.78 | 100.00 |
| 2 stochastic items | 65.70 (64.37, 67.00) | 34.26 | 99.92 |
| 1 stochastic item | 42.48 (41.12, 43.86) | 55.82 | 95.86 |

and variance concentrated in 10, 2 or 1 items. All calls are independent conditional on the fixed laws; the scenarios were specified before simulation.

Coverage is 92.22%, 65.70% and 42.48% in the concentrated cases. With one stochastic item, 4.14% of datasets have at least one unavailable interval because the empirical panel variance is zero; availability is counted in simultaneous coverage. The descriptive ranges for effective item counts are 11.35–16.06 for Qwen and 4.23–5.17 for Gemma, with largest contributions 17.4% and 43.5%. These quantities summarize realized variance concentration. Table S9 reports the complete-pair bounds' coverage and width under concentrated variance.

### E.5 Complete response and decision results

The primary component table reports fixed-call intervals and separate pointwise composition intervals for the same $R = 4$ estimates. The realized prompt inflation is $47\widehat{U}_c/384$ at $K = 48, R = 4$.

We estimate deployment vector risk using the unbiased trace sample variance across the eight independent groups. Gold Brier loss is the average squared distance of the four-category vote vector to the gold one-hot vector. Subtracting trace variance from Brier loss estimates squared mean-to-gold distance; we retain the raw estimates. In Qwen, cooling decreases both terms. Gemma's vector risk changes from .02494 to .01959 while mean-to-gold error changes from .35592 to .35932, so Brier loss changes from .38086 to .37891. Gold loss combines vote variation and mean-to-gold alignment.

**Gold-loss decomposition contrasts.** Table S16 gives the paired Qwen loss changes on all 40 items. We estimate vector risk with the unbiased trace sample variance of the eight independent deployment vectors. Subtracting vector risk from Brier loss estimates squared mean-to-gold distance. Both terms decrease under cooling in these composition summaries.

Table S14: Qwen temperature component profiles. Entries are pooled $R = 4$ complete-pair estimates, averaged equally over items; plug-in prompt dispersion is shown alongside the corrected components. Each dataset has 20 items.

| Panel | $T$ | Prompt | Order | Interaction | Call | Total | Plug-in |
|---|---|---|---|---|---|---|---|
| All 40 | 1 | 0.01125 | 0.05516 | 0.01018 | 0.35386 | 0.43045 | 0.05456 |
| All 40 | 0.2 | 0.03325 | 0.14688 | 0.03637 | 0.08251 | 0.29902 | 0.04335 |
| All 40 | 0 | 0.05387 | 0.17773 | 0.05231 | 0.00569 | 0.28959 | 0.05456 |
| LLMBar | 1 | 0.00694 | 0.02216 | 0.00916 | 0.29366 | 0.33193 | 0.04288 |
| LLMBar | 0.2 | 0.01715 | 0.07345 | 0.02439 | 0.06753 | 0.18252 | 0.02541 |
| LLMBar | 0 | 0.03520 | 0.10318 | 0.03615 | 0.00208 | 0.17661 | 0.03546 |
| JudgeBench | 1 | 0.01556 | 0.08816 | 0.01119 | 0.41406 | 0.52897 | 0.06624 |
| JudgeBench | 0.2 | 0.04936 | 0.22032 | 0.04835 | 0.09748 | 0.41551 | 0.06129 |
| JudgeBench | 0 | 0.07253 | 0.25228 | 0.06847 | 0.00929 | 0.40256 | 0.07367 |

Table S15: Gemma response components, pooled $R = 4$ descriptive means over all 40 items. The four corrected components sum to corrected total disagreement.

| Component | $T = 1$ | $T = .2$ | $T = 0$ |
|---|---|---|---|
| Prompt | .01859 | .02858 | .03280 |
| Order | .05810 | .06369 | .06280 |
| Interaction | .01734 | .02690 | .03062 |
| Call | .04501 | .00842 | .00013 |
| Total disagreement | .13904 | .12759 | .12636 |
| Plug-in prompt | .02410 | .02961 | .03282 |
| Finite-call prompt excess | .00551 | .00103 | .00002 |

Table S16: Qwen loss-decomposition contrasts using all $J = 8$ independent four-vote groups per item and temperature. Entries are changes with pointwise 95% paired item-composition intervals from 20,000 resamples that preserve dataset and stratum. Units are squared one-hot-vector distance; these intervals describe item composition conditional on the recorded calls.

| Metric | $\Delta_{.2-1}$ | $\Delta_{0-1}$ |
|---|---|---|
| Vector risk | $-.05709$ $[-.07042, -.04364]$ | $-.06719$ $[-.07958, -.05458]$ |
| Gold Brier loss | $-.13125$ $[-.16641, -.09766]$ | $-.14375$ $[-.18086, -.10547]$ |
| Squared mean-to-gold | $-.07416$ $[-.10246, -.04632]$ | $-.07656$ $[-.10865, -.04425]$ |

Tables S17 and S18 report all twelve endpoints spanning the two rules and cooling levels, in the pooled sample and each dataset. Each pool has its own simultaneous endpoint family. Qwen's original-rule wrong-decision change at $T = 0$ is $-12.81$ pp $[-23.07, -2.56]$. Gemma's pure-tie increase is .94 pp (6.88% to 7.81%), and its original-rule coverage is 83.13% at both endpoints.

Table S17: Qwen cooling contrasts relative to $T = 1$ (percentage points). Brackets are finite-sample 95% intervals simultaneous over the twelve endpoints within each reporting pool. All eight groups per item are used; items and assigned prompt pairs are fixed.

| Pool | $T$ | Rule | Accuracy | Wrong | Coverage |
|---|---|---|---|---|---|
| All 40 | .2 | Original | $-.63$ $[-13.42, 12.17]$ | $-9.06$ $[-19.08, .95]$ | $-9.69$ $[-25.50, 6.12]$ |
| All 40 | .2 | Fair 2:2 | $6.72$ $[-2.36, 15.80]$ | $-1.72$ $[-10.44, 7.00]$ | $5.00$ $[-.79, 10.79]$ |
| All 40 | 0 | Original | $-1.88$ $[-14.43, 10.68]$ | $-12.81$ $[-23.07, -2.56]$ | $-14.69$ $[-31.70, 2.33]$ |
| All 40 | 0 | Fair 2:2 | $7.97$ $[-.87, 16.81]$ | $-2.97$ $[-11.44, 5.51]$ | $5.00$ $[-1.03, 11.03]$ |
| LLMBar | .2 | Original | $5.63$ $[-10.06, 21.31]$ | $-5.63$ $[-17.45, 6.20]$ | $.00$ $[-16.89, 16.89]$ |
| LLMBar | .2 | Fair 2:2 | $6.88$ $[-5.19, 18.94]$ | $-4.38$ $[-15.71, 6.96]$ | $2.50$ $[-6.18, 11.18]$ |
| LLMBar | 0 | Original | $6.88$ $[-8.33, 22.08]$ | $-9.38$ $[-21.20, 2.45]$ | $-2.50$ $[-19.39, 14.39]$ |
| LLMBar | 0 | Fair 2:2 | $9.38$ $[-2.45, 21.20]$ | $-6.88$ $[-17.97, 4.22]$ | $2.50$ $[-6.18, 11.18]$ |
| JudgeBench | .2 | Original | $-6.88$ $[-24.49, 10.74]$ | $-12.50$ $[-28.43, 3.43]$ | $-19.38$ $[-41.82, 3.07]$ |
| JudgeBench | .2 | Fair 2:2 | $6.56$ $[-7.25, 20.38]$ | $.94$ $[-12.88, 14.75]$ | $7.50$ $[-3.11, 18.11]$ |
| JudgeBench | 0 | Original | $-10.63$ $[-28.24, 6.99]$ | $-16.25$ $[-32.66, .16]$ | $-26.88$ $[-51.73, -2.02]$ |
| JudgeBench | 0 | Fair 2:2 | $6.56$ $[-7.01, 20.13]$ | $.94$ $[-12.63, 14.51]$ | $7.50$ $[-3.60, 18.60]$ |

Table S18: Gemma cooling contrasts relative to $T = 1$ (percentage points). Brackets are finite-sample 95% intervals simultaneous over the twelve endpoints within each reporting pool. All eight groups per item are used; items and assigned prompt pairs are fixed.

| Pool | $T$ | Rule | Accuracy | Wrong | Coverage |
|---|---|---|---|---|---|
| All 40 | .2 | Original | $-.94\,[-5.64, 3.77]$ | $.31\,[-3.91, 4.53]$ | $-.63\,[-5.69, 4.44]$ |
| All 40 | .2 | Fair 2:2 | $-.31\,[-4.77, 4.15]$ | $.94\,[-3.16, 5.04]$ | $.63\,[-3.72, 4.97]$ |
| All 40 | 0 | Original | $-.31\,[-5.26, 4.63]$ | $.31\,[-3.91, 4.53]$ | $.00\,[-5.31, 5.31]$ |
| All 40 | 0 | Fair 2:2 | $.16\,[-4.52, 4.83]$ | $.78\,[-3.29, 4.85]$ | $.94\,[-3.52, 5.40]$ |
| LLMBar | .2 | Original | $-1.25\,[-9.93, 7.43]$ | $.63\,[-7.33, 8.58]$ | $-.63\,[-9.55, 8.30]$ |
| LLMBar | .2 | Fair 2:2 | $-.94\,[-9.32, 7.44]$ | $.94\,[-6.96, 8.84]$ | $.00\,[-8.20, 8.20]$ |
| LLMBar | 0 | Original | $-1.25\,[-9.93, 7.43]$ | $.63\,[-7.33, 8.58]$ | $-.63\,[-9.55, 8.30]$ |
| LLMBar | 0 | Fair 2:2 | $-.94\,[-9.32, 7.44]$ | $.94\,[-6.96, 8.84]$ | $.00\,[-8.20, 8.20]$ |
| JudgeBench | .2 | Original | $-.63\,[-9.07, 7.82]$ | $.00\,[-8.20, 8.20]$ | $-.63\,[-9.55, 8.30]$ |
| JudgeBench | .2 | Fair 2:2 | $.31\,[-7.95, 8.57]$ | $.94\,[-7.08, 8.96]$ | $1.25\,[-6.95, 9.45]$ |
| JudgeBench | 0 | Original | $.63\,[-8.30, 9.55]$ | $.00\,[-8.20, 8.20]$ | $.63\,[-8.78, 10.03]$ |
| JudgeBench | 0 | Fair 2:2 | $1.25\,[-7.43, 9.93]$ | $.63\,[-7.33, 8.58]$ | $1.88\,[-6.57, 10.32]$ |

Table S19: Qwen cost contrasts. Endpoint bands are in percentage points and jointly cover both rules, both cooling levels, and all $c \in [0, 1]$ with probability at least 95%, separately within each pool. Crossover sets are obtained by inverting the simultaneous loss bands.

| Pool | $T$ | Rule | Band at $c = 0$ | Band at $c = 1$ | Crossover set |
|---|---|---|---|---|---|
| All 40 | .2 | Original | $[-18.82, .70]$ | $[-11.91, 13.16]$ | $[.000, 1.000]$ |
| All 40 | .2 | Fair 2:2 | $[-10.18, 6.75]$ | $[-15.54, 2.11]$ | $[.000, 1.000]$ |
| All 40 | 0 | Original | $[-22.82, -2.81]$ | $[-10.42, 14.17]$ | $[.165, 1.000]$ |
| All 40 | 0 | Fair 2:2 | $[-11.19, 5.25]$ | $[-16.55, .62]$ | $[.000, 1.000]$ |
| LLMBar | .2 | Original | $[-16.94, 5.69]$ | $[-20.80, 9.55]$ | $[.000, 1.000]$ |
| LLMBar | .2 | Fair 2:2 | $[-15.21, 6.46]$ | $[-18.43, 4.68]$ | $[.000, 1.000]$ |
| LLMBar | 0 | Original | $[-20.69, 1.94]$ | $[-21.57, 7.82]$ | $[.000, 1.000]$ |
| LLMBar | 0 | Fair 2:2 | $[-17.47, 3.72]$ | $[-20.69, 1.94]$ | $[.000, 1.000]$ |
| JudgeBench | .2 | Original | $[-27.92, 2.92]$ | $[-10.23, 23.98]$ | $[.000, 1.000]$ |
| JudgeBench | .2 | Fair 2:2 | $[-12.37, 14.24]$ | $[-19.87, 6.74]$ | $[.000, 1.000]$ |
| JudgeBench | 0 | Original | $[-32.15, -.35]$ | $[-6.48, 27.73]$ | $[.012, 1.000]$ |
| JudgeBench | 0 | Fair 2:2 | $[-12.13, 14.00]$ | $[-19.63, 6.50]$ | $[.000, 1.000]$ |

Table S20: Gemma cost contrasts. Endpoint bands are in percentage points and jointly cover both rules, both cooling levels, and all $c \in [0, 1]$ with probability at least 95%, separately within each pool. Crossover sets are obtained by inverting the simultaneous loss bands.

| Pool | $T$ | Rule | Band at $c = 0$ | Band at $c = 1$ | Crossover set |
|---|---|---|---|---|---|
| All 40 | .2 | Original | $[-3.65, 4.28]$ | $[-3.51, 5.39]$ | $[.000, 1.000]$ |
| All 40 | .2 | Fair 2:2 | $[-2.91, 4.78]$ | $[-3.90, 4.52]$ | $[.000, 1.000]$ |
| All 40 | 0 | Original | $[-3.65, 4.28]$ | $[-4.38, 5.00]$ | $[.000, 1.000]$ |
| All 40 | 0 | Fair 2:2 | $[-3.04, 4.60]$ | $[-4.58, 4.26]$ | $[.000, 1.000]$ |
| LLMBar | .2 | Original | $[-6.83, 8.08]$ | $[-6.93, 9.43]$ | $[.000, 1.000]$ |
| LLMBar | .2 | Fair 2:2 | $[-6.45, 8.33]$ | $[-6.94, 8.81]$ | $[.000, 1.000]$ |
| LLMBar | 0 | Original | $[-6.83, 8.08]$ | $[-6.93, 9.43]$ | $[.000, 1.000]$ |
| LLMBar | 0 | Fair 2:2 | $[-6.45, 8.33]$ | $[-6.94, 8.81]$ | $[.000, 1.000]$ |
| JudgeBench | .2 | Original | $[-7.69, 7.69]$ | $[-7.31, 8.56]$ | $[.000, 1.000]$ |
| JudgeBench | .2 | Fair 2:2 | $[-6.57, 8.45]$ | $[-8.07, 7.44]$ | $[.000, 1.000]$ |
| JudgeBench | 0 | Original | $[-7.69, 7.69]$ | $[-9.04, 7.79]$ | $[.000, 1.000]$ |
| JudgeBench | 0 | Fair 2:2 | $[-6.83, 8.08]$ | $[-9.43, 6.93]$ | $[.000, 1.000]$ |

For Qwen under the original rule, the pooled point estimates of the crossover cost are .935 for $T = .2$ and .872 for $T = 0$; its $T = 0$ finite-sample band favors cooling for $c < .165$. Tables S19 and S20 give the fair-rule and Gemma bands over $[0, 1]$. Band inversion retains full, empty and boundary-truncated crossover sets.

**Output validity and length stops.** Table S21 reports completed outputs at each temperature. The estimates use every completed response, including invalid and capped outputs. A capped response with a valid first final line keeps its verdict: Gemma at $T = 0$ has 163 cap stops and no invalid output. Its three empty thought markers at $T = 1$ contain no reasoning text.

Table S21: Formal temperature-study outputs. Qwen rows separate measurement and individual deployment calls, reporting the percentages of invalid and capped outputs and the total output tokens. Gemma rows combine both phases and report counts of invalid returns, cap stops and empty thought markers. Technical calls are excluded.

| Qwen Phase | $T$ | Calls | Invalid (%) | Capped (%) | Output tokens |
|---|---|---|---|---|---|
| Measurement | 1 | 15,360 | 6.48 | 5.00 | 433,666 |
| Measurement | .2 | 15,360 | 3.78 | 3.72 | 323,504 |
| Measurement | 0 | 15,360 | 3.54 | 3.48 | 305,618 |
| Deployment | 1 | 1,280 | 7.42 | 5.63 | 39,506 |
| Deployment | .2 | 1,280 | 3.44 | 3.36 | 24,994 |
| Deployment | 0 | 1,280 | 3.52 | 3.52 | 25,513 |
| Gemma: both phases | $T$ | Calls | Invalid (count) | Capped (count) | Empty markers |
| | 1 | 16,640 | 85 | 138 | 3 |
| | .2 | 16,640 | 4 | 169 | 0 |
| | 0 | 16,640 | 0 | 163 | 0 |

**Item-composition sensitivity.** Separate bootstrap analyses use 20,000 paired whole-item draws that preserve dataset and stratum; singleton strata remain fixed. The rule-rescoring draws use a centered maximum absolute deviation over twelve endpoints within a pool. Endpoint maxima also give composition bands for affine cost curves. These summaries describe item composition conditional on recorded calls; the group guarantees condition on the fixed assignments and average over fresh generation. The reproducible numerical results include complete pointwise and simultaneous composition tables and every secondary deployment and output metric.

### E.6 Prompt families and mechanism checks

The four families each contain 12 prompts and use the correction for their actual $K = 12$. The clause-order subset reorders exactly the same five sentences, including the output-format line, so it holds wording fixed. All eight model–family comparisons show descriptive prompt increases and call decreases under cooling. This exploratory within-panel analysis uses complete-pair estimates on the shared items; Table S10 gives the model-level finite-call intervals.

Table S22: Pooled $R = 4$ components within each predefined 12-prompt family. All outputs, including invalid ones, are retained. Ratios are descriptive.

| Configuration | Family | Call $T = 1$ | Call $T = 0$ | Prompt $T = 1$ | Prompt $T = 0$ | Ratio |
|---|---|---|---|---|---|---|
| Qwen | Clause order | 0.3594 | 0.0064 | 0.0072 | 0.0456 | 6.30 |
| Qwen | Discourse | 0.3604 | 0.0030 | 0.0098 | 0.0496 | 5.08 |
| Qwen | Lexical | 0.3387 | 0.0068 | 0.0069 | 0.0415 | 6.02 |
| Qwen | Syntactic | 0.3569 | 0.0066 | 0.0141 | 0.0534 | 3.80 |
| Gemma | Clause order | 0.0398 | 0.0000 | 0.0171 | 0.0299 | 1.74 |
| Gemma | Discourse | 0.0483 | 0.0000 | 0.0193 | 0.0340 | 1.77 |
| Gemma | Lexical | 0.0467 | 0.0005 | 0.0123 | 0.0254 | 2.07 |
| Gemma | Syntactic | 0.0453 | 0.0000 | 0.0160 | 0.0288 | 1.80 |

At the item level, measurement calls estimate baseline gold margin and order-component change; separate deployments estimate tie changes. Qwen's within-dataset-centered order–tie association is .75 with a pointwise composition interval [.62,.85]; split measurement halves give .75 and .74. Eighteen weak-margin questions ($|m| \leq .25$) have a mean tie increase of 36.81 pp, compared with .66 pp among 19 positive-margin questions. Gemma has 29 positive-margin questions and four with any tie change. Constant tie changes make some Gemma bootstrap correlations undefined. The exploratory correlations are descriptive associations with estimated covariates; the reported interval is pointwise.

**Baseline-margin groups.** Table S23 reports every fixed margin group, including empty and small groups. In Qwen, the five positive-margin JudgeBench questions have a mean tie increase of 22.50 pp, whereas the 14 positive-margin LLMBar questions have a 7.14-pp decrease. Their pooled positive-margin mean is $+.66$ pp. The JudgeBench increase also includes new ties on questions with a positive baseline gold margin.

Table S23: Descriptive groups defined by baseline gold margin $m$ at $T = 1$. Each cell is $(n, \Delta U_o, \Delta T_{22})$, comparing $T = 0$ with $T = 1$; tie changes are in percentage points. Margin and order estimates use $R = 4$ measurement calls/cell; ties use separate $J = 8$ four-vote deployments. A dash indicates that the group is empty, so its mean is undefined.

| Dataset | Model | $m < -.25$ | $\lvert m\rvert \le .25$ | $m > .25$ |
|---|---|---|---|---|
| All 40 | Qwen | $(3, .133, +37.50)$ | $(18, .241, +36.81)$ | $(19, .009, +.66)$ |
| All 40 | Gemma | $(5, .017, +2.50)$ | $(6, .016, +.00)$ | $(29, .000, +.86)$ |
| LLMBar | Qwen | $(0, —, —)$ | $(6, .293, +33.33)$ | $(14, -.010, -7.14)$ |
| LLMBar | Gemma | $(0, —, —)$ | $(3, .004, -4.17)$ | $(17, -.001, +1.47)$ |
| JudgeBench | Qwen | $(3, .133, +37.50)$ | $(12, .214, +38.54)$ | $(5, .063, +22.50)$ |
| JudgeBench | Gemma | $(5, .017, +2.50)$ | $(3, .028, +4.17)$ | $(12, .002, +.00)$ |

**Agreement and correctness estimands.** Across-prompt categorical agreement is

$$\frac{1}{K(K-1)} \sum_{k \neq \ell} \langle P_k^q, P_\ell^q \rangle = \|\bar{p}\|^2 - \frac{U_{\text{prompt}}^{(q)}}{K-1},$$

so it depends on marginal concentration as well as prompt dispersion.

For one item with gold candidate $y$, let $Y_{kor} = \mathbf{1}\{Z_{kor} = y\}$, where $k = 1, \ldots, K$, $o \in \{\text{AB}, \text{BA}\}$, and $r = 1, \ldots, R$. TIE and invalid verdicts have $Y_{kor} = 0$. The binary correctness gap applies the PSS absolute-difference formula (Zhuo et al., 2024) within each order and repeat:

$$\widehat{S}_{\text{binary}} = \frac{1}{2R\binom{K}{2}} \sum_o \sum_{r=1}^{R} \sum_{k<\ell} |Y_{kor} - Y_{\ell or}|. \tag{S62}$$

For the mean-correctness gap, first set $\widehat{a}_k = (2R)^{-1} \sum_o \sum_{r=1}^{R} Y_{kor}$ and then compute

$$\widehat{S}_{\text{mean}} = \frac{1}{\binom{K}{2}} \sum_{k<\ell} |\widehat{a}_k - \widehat{a}_\ell|. \tag{S63}$$

The first statistic takes absolute differences before averaging over orders and repeated calls; the second averages correctness first. For independent Bernoulli correctness with probabilities $a, b$, the pairwise expectation in the first statistic is $a + b - 2ab$; the absolute difference between the mean correctness probabilities is $|a - b|$. The dataset summaries average these statistics with equal weight for each item.

In Qwen, the binary gap falls from .319 to .101 and across-prompt agreement rises from .569 to .709 as corrected prompt dispersion rises from .011 to .054; Gemma shows smaller changes in the same directions. These summaries measure correctness spread, verdict agreement and prompt dispersion, respectively.

# F Decision Geometry: Derivations

For any random verdict vector with mean $\bar{p}$ and fixed target $t$, expanding the square yields $\mathbb{E}\|V - t\|^2 = \mathbb{E}\|V - \bar{p}\|^2 + \|\bar{p} - t\|^2$, because the cross term has mean zero. This proves Equation (9). The following family evaluates discrete accuracy under the original and fair tie rules at fixed gold alignment.

## F.1 Position dependence at fixed positive margin

We isolate the role of position by holding the mean preference for the gold candidate fixed. Consider a binary family with gold candidate 1. Every prompt chooses candidate 1 with probability $\pi + \delta$ in

Table S24: Six stored-label summaries at all three temperatures. Each row uses the identical item/prompt census for its model. The corrected component and the agreement/correctness summaries have different estimands.

| | | Qwen | | | Gemma | | |
|---|---|---|---|---|---|---|---|
| Dataset | Summary | 1 | .2 | 0 | 1 | .2 | 0 |
| All 40 items | Plug-in prompt | 0.055 | 0.043 | 0.055 | 0.024 | 0.030 | 0.033 |
| All 40 items | Corrected prompt | 0.011 | 0.033 | 0.054 | 0.019 | 0.029 | 0.033 |
| All 40 items | Same-cell agreement | 0.646 | 0.917 | 0.994 | 0.955 | 0.992 | 1.000 |
| All 40 items | Across-prompt agreement | 0.569 | 0.700 | 0.709 | 0.861 | 0.872 | 0.873 |
| All 40 items | Binary correctness gap | 0.319 | 0.139 | 0.101 | 0.059 | 0.049 | 0.048 |
| All 40 items | Mean-correctness gap | 0.160 | 0.106 | 0.093 | 0.049 | 0.046 | 0.046 |
| LLMBar | Plug-in prompt | 0.043 | 0.025 | 0.035 | 0.026 | 0.031 | 0.035 |
| LLMBar | Corrected prompt | 0.007 | 0.017 | 0.035 | 0.022 | 0.030 | 0.035 |
| LLMBar | Same-cell agreement | 0.706 | 0.932 | 0.998 | 0.969 | 0.992 | 1.000 |
| LLMBar | Across-prompt agreement | 0.668 | 0.817 | 0.823 | 0.891 | 0.898 | 0.898 |
| LLMBar | Binary correctness gap | 0.290 | 0.109 | 0.074 | 0.055 | 0.049 | 0.050 |
| LLMBar | Mean-correctness gap | 0.148 | 0.082 | 0.073 | 0.043 | 0.044 | 0.045 |
| JudgeBench | Plug-in prompt | 0.066 | 0.061 | 0.074 | 0.023 | 0.028 | 0.031 |
| JudgeBench | Corrected prompt | 0.016 | 0.049 | 0.073 | 0.015 | 0.027 | 0.031 |
| JudgeBench | Same-cell agreement | 0.586 | 0.903 | 0.991 | 0.941 | 0.991 | 1.000 |
| JudgeBench | Across-prompt agreement | 0.471 | 0.583 | 0.596 | 0.830 | 0.845 | 0.848 |
| JudgeBench | Binary correctness gap | 0.349 | 0.169 | 0.127 | 0.064 | 0.049 | 0.047 |
| JudgeBench | Mean-correctness gap | 0.171 | 0.130 | 0.114 | 0.054 | 0.049 | 0.047 |

AB order and $\pi - \delta$ in BA order, where $1/2 \le \pi < 1$ and $0 \le \delta \le 1 - \pi$. Calls are independent. The order-averaged law and gold margin are

$$\bar{p} = (\pi, 1-\pi, 0, 0), \qquad m = 2\pi - 1.$$

All prompts have this same averaged law: $U^{(q)}_{\text{prompt}} = U_{\text{int}} = 0$, while $U_{\text{order}} = 2\delta^2$. Varying $\delta$ isolates position dependence while holding the mean verdict distribution fixed.

For either balanced four-call design, the number of votes for candidate 1 is a sum of independent $\text{Bin}(2, \pi+\delta)$ and $\text{Bin}(2, \pi-\delta)$ variables. Write $h = \pi^2 - \delta^2$. This gives the deployment risk, Brier loss, tie probability and strict accuracy:

$$\mathcal{R}_d := \mathbb{E}\|V_d - \bar{p}\|^2 = \frac{\pi(1-\pi) - \delta^2}{2}, \tag{S64}$$

$$B_d := \mathbb{E}\|V_d - e_y\|^2 = \mathcal{R}_d + 2(1-\pi)^2,$$

$$T_{22} = 4\pi^2 + (2 - 12\pi)h + 6h^2, \tag{S65}$$

$$A_{\text{strict}} = 4\pi h - 3h^2.$$

Here strict accuracy assigns zero credit to a tied vote. Fair resolution of the tie gives

$$A_{\text{coin}} = A_{\text{strict}} + \tfrac{1}{2}T_{22} = 3\pi^2 - 2\pi^3 + (2\pi - 1)\delta^2. \tag{S66}$$

**Proof of Proposition 2.**

*Proof.* The four-vote probability generating function is $[1-\pi-\delta+(\pi+\delta)z]^2[1-\pi+\delta+(\pi-\delta)z]^2$. Its $z^2$ coefficient gives $T_{22}$; the $z^3$ and $z^4$ coefficients give strict accuracy. Setting $t = \delta^2$,

$$\frac{dT_{22}}{dt} = 12\pi(1-\pi) - 2 + 12t, \quad \frac{dA_{\text{strict}}}{dt} = 6(\pi^2 - t) - 4\pi, \quad \frac{dA_{\text{coin}}}{dt} = 2\pi - 1.$$

Their signs and Equation (S64) give the claim. □

**Proposition S2** (The zero-margin endpoint)**.** *At $\pi = 1/2$, put $u = U_{\text{order}} = 2\delta^2$. Then*

$$\mathcal{R}_d = \frac{1}{8} - \frac{u}{4}, \quad B_d = \frac{5}{8} - \frac{u}{4}, \quad T_{22} = \frac{3}{8} + \frac{u}{2} + \frac{3}{2}u^2, \quad A_{\text{strict}} = \frac{1 - T_{22}}{2}. \tag{S67}$$

*Fair-coin accuracy is $1/2$ throughout this family. At maximal position dependence, every balanced deployment ties and its vector risk is zero.*

The probability of choosing the first-presented candidate in this special case is $a = 1/2 + \delta$. Substituting it in the four-vote law gives the stated endpoint formulas.

For example, $\pi = .6$ and $\delta : 0 \to .4$ keep the gold margin fixed at .2, reduce vector risk from .12 to .04 and Brier loss from .44 to .36, and raise the tie rate from .3456 to .64. Strict accuracy changes from .4752 to .36 while fair accuracy changes from .648 to .68. This family has zero prompt and interaction components, isolating the effects of position dependence and the tie rule.

### F.2 Abstention cost and the full categorical rule

If choosing a wrong candidate costs one and abstaining costs $c$, the binary family has $L_c = 1 - A_{\text{strict}} - (1 - c)T_{22}$. At zero margin, $L_c = 1/2 + (c - 1/2)T_{22}$; the positive-margin example's endpoint laws have equal loss at $c = 14/23$. In the full four-category response law, fair resolution of pure candidate 2:2 ties gives the exact identities in Equation (10); mixed ties, TIE votes and invalid outcomes retain their original handling. The interaction identity holds groupwise for all 256 four-vote patterns and either candidate as gold. Its inferential family is stated in Section E.3.

## G Prompt Policies and Budgeted Reasoning

### G.1 Four Qwen judging policies

The policy study compares four banks of six equally weighted prompts: criterion paraphrases, rubrics, comparison procedures and labeled examples. Each example prompt contains three independently written demonstrations, one per verdict, written without using the test labels. Each bank defines a distinct judging policy. The Qwen checkpoint runs in native nonthinking mode with temperature and top-$p$ set to 1, an 8,192-token context, a 128-token cap, batch size 128 and disabled prefix caching.

The JudgeBench Claude response file supplies 178 distinct questions after excluding 91 normalized matches to the earlier GPT file and one duplicate. These comprise two canaries, 24 development items and 152 test items. Conditions share these items but use independent generation. For each test item and condition, measurement uses $K = 6$ prompts, two orders and $R = 16$ calls per cell; deployments use $J = 64$ independent four-call groups for each D/E design. Eight blocks interleave conditions, measurement calls and deployments. The selector is fixed from development data before test collection. There are 432,736 calls (96 canary, 4,608 development, 428,032 test), of which 311,296 test calls form deployments. The analysis assumes independent calls under stationary cell laws.

Table S25: Qwen judging-policy study. Component estimates pool all 16 measurement repeats. $\perp$ and length-stop percentages use all test calls, including measurement and deployment; their denominator differs from measurement-only validity. Length stops and invalid verdicts are distinct, potentially overlapping outcomes.

| Study | Policy / cap | $\widehat{U}_p$ | $\widehat{U}_{po}$ | $\widehat{U}_c$ | $\perp$ (%) | Length (%) |
|---|---|---|---|---|---|---|
| Policy | Paraphrase / 128 | 0.0034 | 0.0089 | 0.3076 | 2.49 | 1.31 |
| Policy | Rubric / 128 | 0.0035 | 0.0068 | 0.3273 | 3.88 | 1.66 |
| Policy | Comparison / 128 | 0.0086 | 0.0112 | 0.3241 | 4.44 | 2.09 |
| Policy | Examples / 128 | 0.0001 | 0.0010 | 0.2627 | 2.02 | 1.09 |

Table S26: Performance against gold labels in independent four-call deployments. Brier loss uses the averaged verdict vector; error and abstention use the fixed aggregate-verdict rule. Overall error is one minus accuracy and includes abstentions. Both ties and $\perp$ count as abstentions. All deployments, including invalid outputs, contribute. Entries are point estimates, with percentages for error and abstention.

| Study | Policy / cap | Brier D | Brier E | Error D | Error E | Abst. D | Abst. E |
|---|---|---|---|---|---|---|---|
| Policy | Paraphrase / 128 | 0.4662 | 0.4662 | 49.99 | 49.50 | 29.53 | 29.15 |
| Policy | Rubric / 128 | 0.4791 | 0.4821 | 51.44 | 50.57 | 31.84 | 29.91 |
| Policy | Comparison / 128 | 0.4872 | 0.4811 | 51.18 | 50.11 | 30.54 | 29.33 |
| Policy | Examples / 128 | 0.4707 | 0.4718 | 56.35 | 55.93 | 39.34 | 38.80 |

Table S27: Exploratory decision summaries for all four policies. All entries are percentages. $C$ is coverage, $A^{\text{dec}}$ is accuracy among decisive deployments, and $T_{22}$ is the pure candidate 2:2 tie rate. D and E each use four calls. These are point estimates; conditional accuracy uses policy-specific decisive subsets.

| Study | Policy / cap | $C_D$ | $C_E$ | $A_D^{\text{dec}}$ | $A_E^{\text{dec}}$ | $T_{22,D}$ | $T_{22,E}$ |
|---|---|---|---|---|---|---|---|
| Policy | Paraphrase / 128 | 70.47 | 70.85 | 70.97 | 71.29 | 27.88 | 27.38 |
| Policy | Rubric / 128 | 68.16 | 70.09 | 71.24 | 70.53 | 28.47 | 26.64 |
| Policy | Comparison / 128 | 69.46 | 70.67 | 70.28 | 70.59 | 26.53 | 25.81 |
| Policy | Examples / 128 | 60.66 | 61.20 | 71.95 | 72.00 | 36.85 | 36.07 |

Relative to paraphrases, examples reduce prompt dispersion by .00325 (pointwise composition interval [−.00463,−.00186]) and interaction by .00792 ([−.00977,−.00618]); their corrected prompt estimate is .00014. At the same time, the order component rises from .09784 to .16677 (difference .06893, pointwise paired composition interval [.05662,.08121]), and pure ties increase. Under D, Brier loss with examples is .47069 versus about .46622 for paraphrases, alongside the decreases in prompt and interaction estimates.

**Voting-rule and cost sensitivity.** We compare rules by rescoring the same complete vote groups under each rule, using 20,000 paired whole-item resamples within the four original strata. The maximum absolute centered bootstrap error over twelve examples-minus-paraphrase contrasts—D/E, original/fair rules and accuracy/coverage/wrong-candidate frequency—gives approximate simultaneous 95% composition intervals, with common half-width .02631579. All four policies enter the analysis; the examples-minus-paraphrase comparison is exploratory.

Table S28: Examples minus paraphrasing under two decision rules. Changes and approximate simultaneous 95% intervals are in percentage points. The twelve-comparison family includes accuracy, coverage, and wrong-choice frequency for both rules and designs.

| Design | Rule | Accuracy change | Coverage change |
|---|---|---|---|
| D | Original | -6.36 [-8.99, -3.73] | -9.81 [-12.44, -7.18] |
| D | Fair pure 2:2 | -1.88 [-4.51, 0.76] | -0.83 [-3.46, 1.80] |
| E | Original | -6.44 [-9.07, -3.80] | -9.64 [-12.27, -7.01] |
| E | Fair pure 2:2 | -2.09 [-4.72, 0.54] | -0.96 [-3.59, 1.68] |

With wrong-decision cost one and abstention cost $c$, $\Delta L(c) = (1-c)\Delta W - c\Delta A$. The same endpoint family gives a band for every $c \in [0,1]$. Under the original rule, point estimates of the crossover cost are .351153 (D) and .332623 (E), with inverted sets [.082809,.619497] and [.059701,.605544]. Under fair resolution, the example policy has a larger point estimate of loss throughout $[0,1]$; Figure S8 displays the composition band. Fair rescoring leaves vote-vector risk and Brier unchanged. These composition intervals describe the policy-study items; the temperature study uses a separate 40-item panel.

### G.2 Gemma reasoning and fixed examples

The four arms are G0 nonthinking/base, G1 nonthinking/examples, G2 thinking/base and G3 thinking/examples, using a fixed `google/gemma-4-31B-it` checkpoint. The 32 new RewardBench 2 questions include eight each from Factuality, Focus, Math and Precise IF. Each chosen response is paired with one of its three rejected responses using a fixed pseudorandom rule, with a separately preassigned canonical orientation. The evaluation target is the chosen-versus-selected-rejected pair. Normalized equality and five-word shingle checks exclude overlap with earlier panels and demonstrations. Ranking within each domain, without using outcomes, assigns reserve positions 0–7. The pair mapping records source IDs, rejected-response indices, gold labels and reserve positions.

All arms use $K = 48$ prompts, two orders, $R = 2$ calls per cell and $J = 8$ separate D groups per item, with prompt pairs matched across arms. The study therefore contains 24,576 measurement judgments and 4,096 deployment judgments, for 28,672 formal observations. The same three demonstrations are appended to every example prompt: 4 versus 5 for $2 + 2$ (A), `SUN` versus `sun`

for uppercase (B), and two answers equal to 3 (TIE). The demonstration labels are independent of the evaluation-item gold labels.

All arms use xgrammar-constrained final A/B/TIE generation with an eight-token cap. The thinking stage uses the native 1,024-token budget and 1,032-token request cap, stopping at the first channel-end token. Final generation resumes from the complete returned thought without a new instruction. Nonthinking uses the native empty-thought prefix. Both stages use $T = 1$, top-$p = .95$, top-$k = 64$, bfloat16, vLLM V1, 16,384 context, at most 32 concurrent sequences, batches of 64 logical judgments and no speculative decoding. The shared constrained decoder renormalizes final-token probabilities. The protocol codes malformed thoughts or final responses as invalid. All formal judgments and the 64 separate canary judgments have valid, untruncated final responses; native budget closure occurs in 23.63% of G2 and 21.88% of G3 thoughts. Data collection spans two runtime periods, preserving the originally assigned cell conditions and seeds. The analysis assumes independent calls under stationary cell laws across periods.

For $J$ deployment vectors, $S^2 = (J-1)^{-1}\sum_j \|V_j - \bar{V}\|^2$ estimates vector risk and $B = J^{-1}\sum_j \|V_j - e_y\|^2$ estimates Brier loss. The raw gold-alignment estimate is $B - S^2 = \|\bar{V} - e_y\|^2 - S^2/J$. All arm means weight items equally. Conditional accuracy divides the pooled number of correct deployments by the pooled number of decisive deployments.

Table S29: Descriptive Gemma group means on 32 RewardBench 2 items. G0/G1 use nonthinking; G2/G3 use budgeted thinking. G0/G2 use the base instruction; G1/G3 add fixed examples. Components, vector risk, and gold losses are in squared one-hot-vector units. Conditional accuracy is the pooled correct-to-decisive ratio. Operational diagnostics cover all 7,168 formal judgments per group; deployment outcomes use the 256 independent deployments per group, clustered within 32 items.

| Metric | G0 | G1 | G2 | G3 |
|---|---|---|---|---|
| Prompt component | 0.013226 | 0.015025 | 0.002484 | 0.000554 |
| Order component | 0.060125 | 0.048960 | 0.047931 | 0.051929 |
| Interaction component | 0.011164 | 0.013865 | 0.000653 | 0.000480 |
| Call component | 0.013021 | 0.014323 | 0.092773 | 0.091797 |
| Pure 2:2 tie (%) | 12.11 | 4.69 | 5.47 | 5.47 |
| Decisive coverage (%) | 86.33 | 91.80 | 91.41 | 90.62 |
| Unconditional accuracy (%) | 75.39 | 76.95 | 80.08 | 80.47 |
| Conditional decisive accuracy (%) | 87.33 | 83.83 | 87.61 | 88.79 |
| Observed vector risk | 0.012137 | 0.013323 | 0.023926 | 0.022042 |
| Predicted vector risk | 0.009727 | 0.010933 | 0.024409 | 0.023221 |
| Gold Brier loss | 0.284668 | 0.314941 | 0.247070 | 0.231445 |
| Estimated squared gold-alignment error | 0.272531 | 0.301618 | 0.223145 | 0.209403 |
| First-position choice (%) | 55.52 | 53.11 | 55.08 | 55.78 |
| Forced thought end (%) | 0.00 | 0.00 | 23.63 | 21.88 |
| Invalid judgments | 0 | 0 | 0 | 0 |

**Tests specified before data collection.** The primary family includes G1–G0 contrasts for order, pure-2:2 ties and coverage; the secondary family includes these endpoints for G2–G0 and the interaction $(G3 - G2) - (G1 - G0)$. Each of twenty thousand whole-item resamples draws eight items within each domain, preserving all four arms. The centered two-sided value is $p_c = [1 + \sum_b \mathbf{1}\{|\Delta_b^* - \widehat{\Delta}| \geq |\widehat{\Delta}|\}]/20001$; Holm adjustment is applied separately within the three- and six-test families. We report percentile intervals at 98.33% and 99.17% marginal levels alongside the separately constructed centered Holm test. Inference uses the .05 threshold. The primary tie contrast has $p_H = .12944$ and a centered absolute-error interval of $[-16.406, 1.562]$ pp. Inference describes approximate item-composition uncertainty conditional on the recorded calls.

**Exploratory thinking contrasts.** Twelve subsequent comparisons examine prompt, interaction, call, Brier, vector risk and gold alignment for G2–G0 and G3–G1. They use 100,000 stratified paired whole-item resamples, preserving all arms and equal domain weights. The radius of a centered interval is the higher empirical $(1-\alpha)$ quantile of $|\Delta_b^* - \widehat{\Delta}|$, with $\alpha = .05$ or $.05/12$. The latter's

Table S30: Prespecified Gemma contrasts. Paired whole-item bootstrap resampling uses 20,000 draws within the four domains, with equal domain weights. The displayed percentile intervals have nominal marginal levels of 98.33% for the three primary endpoints and 99.17% for the six secondary endpoints. $p_c$ denotes the two-sided centered-bootstrap test; $p_H$ applies Holm adjustment within each of the two prespecified families. Inference at $\alpha = .05$ follows $p_H$. Rate differences are percentage points (pp).

| Family | Contrast | Endpoint | Change | Percentile interval | $p_c$ | $p_H$ |
|---|---|---|---|---|---|---|
| Primary | $G1 - G0$ | Order | -0.011164 | $[-0.037802, 0.008439]$ | 0.25284 | 0.25284 |
| Primary | $G1 - G0$ | Pure 2:2 (pp) | -7.422 | $[-17.578, -0.391]$ | 0.04315 | 0.12944 |
| Primary | $G1 - G0$ | Coverage (pp) | +5.469 | $[-0.391, 14.844]$ | 0.09290 | 0.18579 |
| Secondary | $G2 - G0$ | Order | -0.012193 | $[-0.084644, 0.049228]$ | 0.64282 | 0.75546 |
| Secondary | $G2 - G0$ | Pure 2:2 (pp) | -6.641 | $[-21.484, 6.250]$ | 0.22739 | 0.75156 |
| Secondary | $G2 - G0$ | Coverage (pp) | +5.078 | $[-8.203, 20.703]$ | 0.37773 | 0.75546 |
| Secondary | $(G3 - G2) - (G1 - G0)$ | Order | +0.015162 | $[-0.012049, 0.048321]$ | 0.18789 | 0.75156 |
| Secondary | $(G3 - G2) - (G1 - G0)$ | Pure 2:2 (pp) | +7.422 | $[-1.562, 19.531]$ | 0.07265 | 0.43588 |
| Secondary | $(G3 - G2) - (G1 - G0)$ | Coverage (pp) | -6.250 | $[-17.450, 2.344]$ | 0.10299 | 0.51497 |

Inference uses the .05 threshold. The primary pure-2:2 contrast has Holm-adjusted centered-bootstrap value $p_H = .12944$. Percentile intervals and centered tests use separate constructions. The two families contain three and six tests, respectively. Their separate adjustments provide inference within each family.

99.583% marginal levels target approximate 95% family coverage. Table S31 reports all paired contrasts and both interval families.

Table S31: Exploratory thinking-minus-nonthinking contrasts on the complete Gemma panel. Intervals invert centered absolute bootstrap errors using 100,000 paired whole-item resamples, eight per domain. The outer intervals use marginal coverage 99.583% (Bonferroni across all 12 endpoints; approximate familywise target 95%). These intervals describe item-composition uncertainty conditional on observed calls. Gold alignment is estimated by Brier minus the unbiased deployment covariance trace.

| Instruction | Endpoint | Change | 95% interval | 12-endpoint interval |
|---|---|---|---|---|
| Base | Prompt | -0.01074 | $[-0.02313, 0.00164]$ | $[-0.02932, 0.00783]$ |
| | Interaction | -0.01051 | $[-0.02053, -0.00049]$ | $[-0.02568, 0.00466]$ |
| | Call | +0.07975 | $[0.03158, 0.12793]$ | $[0.00879, 0.15072]$ |
| | Gold Brier | -0.03760 | $[-0.20215, 0.12695]$ | $[-0.27979, 0.20459]$ |
| | Deployment risk | +0.01179 | $[-0.00607, 0.02965]$ | $[-0.01514, 0.03871]$ |
| | Gold-alignment error | -0.04939 | $[-0.21017, 0.11140]$ | $[-0.28530, 0.18652]$ |
| Examples | Prompt | -0.01447 | $[-0.02444, -0.00450]$ | $[-0.02915, 0.00021]$ |
| | Interaction | -0.01339 | $[-0.02263, -0.00415]$ | $[-0.02719, 0.00042]$ |
| | Call | +0.07747 | $[0.03060, 0.12435]$ | $[0.00846, 0.14648]$ |
| | Gold Brier | -0.08350 | $[-0.23535, 0.06836]$ | $[-0.30762, 0.14062]$ |
| | Deployment risk | +0.00872 | $[-0.00635, 0.02379]$ | $[-0.01346, 0.03090]$ |
| | Gold-alignment error | -0.09222 | $[-0.24163, 0.05720]$ | $[-0.31264, 0.12821]$ |

## H Balanced Deployments and Design Selection

This secondary analysis examines estimation of the mean response law alongside discrete decision quality. All expectations are conditional on the fixed response laws for each item and cell. Calls are mutually independent, and prompt and order randomization is independent of the calls. Fix $K \geq 4$ prompts, two equally weighted orders, and the census target $\bar{p} = (2K)^{-1} \sum_{k,o} P_{ko}$. Each design returns the mean of four one-hot verdicts. Design $D$ samples two distinct prompts uniformly and queries both orders for each. Design $E$ samples four distinct prompts uniformly, then assigns two to AB and two to BA uniformly. Design randomization is independent of the calls. For $d \in \{D, E\}$, define $\mathcal{R}_d = \mathbb{E}\|V_d - \bar{p}\|_2^2$.

**Proposition S3** (Risk of balanced four-call designs)**.** *For arbitrary fixed categorical cell laws and independent calls, both designs are unbiased for $\bar{p}$, and*

$$\mathcal{R}_D = \frac{U_{\text{call}}}{4} + \frac{K-2}{2(K-1)} U^{(q)}_{\text{prompt}}, \tag{S68}$$

$$\mathcal{R}_E = \frac{U_{\text{call}}}{4} + \frac{K-4}{4(K-1)} U^{(q)}_{\text{prompt}} + \frac{K}{4(K-1)} U_{\text{int}}. \tag{S69}$$

*Thus $\Delta := \mathcal{R}_E - \mathcal{R}_D = \gamma(U_{\text{int}} - U^{(q)}_{\text{prompt}})$, where $\gamma = K/[4(K-1)]$.*

For $K = 6$, $\Delta = 0.3(U_{\text{int}} - U^{(q)}_{\text{prompt}})$: covering more prompts helps when prompt variation exceeds the interaction contribution. This comparison targets estimation of the shared response-law mean; Section 5 relates that target to discrete decisions.

**Estimating the design contrast directly.** Write $A_k = P_{k,\text{AB}}$, $B_k = P_{k,\text{BA}}$, and let bars denote averages over prompts. From $R \geq 1$ calls per cell, form $\widehat{A}_k$, $\widehat{B}_k$ and

$$\widehat{\Delta}_R = -\gamma \left( \frac{1}{K} \sum_k \widehat{A}_k^\top \widehat{B}_k - \overline{\widehat{A}}^\top \overline{\widehat{B}} \right). \tag{S70}$$

**Proposition S4** (Cancellation of same-cell terms)**.** *Equation* (S70) *is unbiased for $\Delta$ for every $R \geq 1$. For $R \geq 2$, it also equals*

$$\begin{aligned} \widehat{\Delta}_R &= \gamma(\widehat{U_{\text{int}}} - \widehat{U_{\text{prompt}}}^{(q)}) \\ &= \gamma(\widehat{U}_{\text{plug,int}} - \widehat{U}^{(q)}_{\text{plug,prompt}}) \end{aligned} \tag{S71}$$

*for every sample, where the plug-in components substitute empirical cell distributions in the population formulas.*

### H.1 Proofs of the balanced-design identities

Set $F_k = (A_k + B_k)/2$, $G_k = (A_k - B_k)/2$, $f_k = F_k - \bar{F}$ and $g_k = G_k - \bar{G}$. Then $\bar{p} = \bar{F}$, $U^{(q)}_p = K^{-1} \sum_k \|f_k\|^2$ and $U_{\text{int}} = K^{-1} \sum_k \|g_k\|^2$. For centered vectors $x_k$ and a uniform size-$m$ subset $S_m$,

$$\mathbb{E} \left\| m^{-1} \sum_{k \in S_m} x_k \right\|^2 = \frac{K-m}{m(K-1)} \frac{1}{K} \sum_k \|x_k\|^2.$$

This follows from the inclusion probabilities $m/K$ and $m(m-1)/[K(K-1)]$, together with $\sum_{k \neq \ell} x_k^\top x_\ell = -\sum_k \|x_k\|^2$. Design D has conditional mean $(F_j + F_k)/2$, so setting $m = 2$ gives its prompt term. Each cell is selected with probability $2/K$ and weighted $1/4$, giving expected call variance $(2/K) \sum_{ko} (1 - \|P_{ko}\|^2)/16 = U_c/4$. The expected weight of each cell is $1/(2K)$, which proves unbiasedness.

For E, the centered conditional mean is $\frac{1}{4} \sum_{k \in S_4} f_k + \frac{1}{4} \sum_{k \in S_4} \varepsilon_k g_k$, with exactly two positive signs. Symmetry of the conditional sign distribution removes the cross term; the first term contributes $(K-4)U^{(q)}_p/[4(K-1)]$. Distinct selected signs satisfy $\mathbb{E}\varepsilon_k \varepsilon_\ell = -1/3$, so the second term has expected squared norm

$$\frac{1}{4K} \sum_k \|g_k\|^2 - \frac{1}{4K(K-1)} \sum_{k \neq \ell} g_k^\top g_\ell = \frac{K}{4(K-1)} U_{\text{int}}.$$

The order main effect cancels because the signs sum to zero. Cell inclusion probabilities and call variance are the same as in D, proving Proposition S3.

Finally, $U_{\text{int}} - U^{(q)}_p = -K^{-1} \sum_k (A_k - \bar{A})^\top (B_k - \bar{B})$. The AB and BA empirical means come from distinct cells, so every product is unbiased even at $R = 1$. This algebraic identity also holds for empirical vectors. The coefficient of every same-cell squared norm in $H_{XO} - 2H_X - H_O + 2H_0$

is $1/(2K) - 2/(4K) - 1/(2K^2) + 2/(4K^2) = 0$, so complete-pair correction leaves the contrast unchanged. This proves Proposition S4.

Both designs have the same mean verdict distribution, so Equation (9) gives the same expected Brier difference $\Delta$. Categorical accuracy depends on the full vote law, the decision rule and the gold label.

**Proposition S5** (Winner preservation at positive census margin)**.** *Suppose $\bar{p}$ has a unique winning category $c^\star$, and define its winner–runner-up gap over all verdict categories by $\kappa = \bar{p}_{c^\star} - \max_{j \neq c^\star} \bar{p}_j > 0$. For either design $d \in \{D, E\}$,*

$$\mathbb{P}\left( V_{d,c^\star} \leq \max_{j \neq c^\star} V_{d,j} \right) \leq \min\left\{ 1, \frac{2\mathcal{R}_d}{\kappa^2} \right\}. \tag{S72}$$

*The event includes tied maxima as failures to preserve the unique winner. If $c^\star$ is the gold candidate, the same bound controls the original plurality rule's gold error, counting abstentions as failures.*

*Proof.* On the displayed event, some $j \neq c^\star$ satisfies $V_{d,j} \geq V_{d,c^\star}$. With $e = V_d - \bar{p}$, this gives $e_j - e_{c^\star} \geq \bar{p}_{c^\star} - \bar{p}_j \geq \kappa$. Cauchy–Schwarz implies $\|e\|^2 \geq \kappa^2/2$. Markov's inequality and $\mathbb{E}\|e\|^2 = \mathcal{R}_d$ prove the bound. When $c^\star$ is gold, the original rule returns gold exactly when $V_{d,c^\star}$ is the unique maximum, giving the final claim. □

### H.2 Independent deployment checks

The independent study using the JudgeBench GPT response file fixes a stratified split of 100 development items, 200 test items and four canary items (development/test counts: 16/32 math, 28/56 reasoning, 12/24 code, 44/88 MMLU-Pro). Qwen runs in native nonthinking mode with BF16, $T$ =top-$p$ $= 1$, an eight-token cap, an 8,192-token context, batch size 16 and disabled prefix caching. Development uses six prompts, two orders and $R = 4$; test uses $R = 8$ and $J = 16$ separate groups per design. Before test calls are collected, development data alone select E when the estimated $E - D$ risk is negative and D otherwise. Stages run sequentially, under the assumption of stationary cell laws. All 49,600 formal calls are retained, including 4,973 invalid outcomes and 4,219 length stops, which can overlap; 48 canary calls are separate.

The development estimate $+.00082$ selects D; its 95% fixed-call Chebyshev interval is $[-.02449,.02613]$. On the test set, the predicted contrast is $-.00114$ and the observed contrast is $-.00322$ (D risk .09385, E .09063), with composition band $[-.00833,.00201]$ for the observed contrast. The gold Brier difference is $+.00059$ $[-.01359,.01484]$. The point estimate changes sign between development and test. For the prespecified prompt tolerance .02, the original Cantelli upper bound is .02014; the sharper covariance bound gives .01830 in a secondary analysis.

The semantic study fixes two 24-prompt pools before data collection. Development data select D/E before testing on the unseen pool B. For each open model and dataset, two interleaved panels each use one measurement call per cell and eight separate deployment groups per design and item on 60 fixed test items. The variance estimate based on panel differences is unbiased. Calibration intervals use an approximate Student-$t$ reference with Bonferroni allocation across six endpoints; equivalence tests use the prespecified calibration margin below.

Table S32: Unseen-B calibration. $\widehat{\Delta}$ predicts $\mathrm{MSE}_E - \mathrm{MSE}_D$ and $\widetilde{\Delta}$ estimates it from independent deployments. The error is $\widehat{\Delta} - \widetilde{\Delta}$. Intervals target nominal 95% coverage over six endpoints under an approximate $t$ reference. $p_H$ is the Holm-adjusted secondary TOST value for the prespecified $\pm.005$ margin.

| Model | Dataset | $\widehat{\Delta}$ | $\widetilde{\Delta}$ | Error | Adjusted interval | $p_H$ |
|---|---|---|---|---|---|---|
| Qwen3.6 | LLMBar | -0.0010 | -0.0062 | 0.0052 | $[-0.0086, 0.0189]$ | 0.849 |
| Qwen3.6 | JudgeBench | -0.0017 | 0.0011 | -0.0028 | $[-0.0206, 0.0150]$ | 0.849 |
| Gemma4 | LLMBar | -0.0015 | -0.0025 | 0.0010 | $[-0.0055, 0.0075]$ | 0.244 |
| Gemma4 | JudgeBench | -0.0005 | -0.0019 | 0.0014 | $[-0.0038, 0.0066]$ | 0.186 |
| GPT-OSS | LLMBar | -0.0003 | 0.0001 | -0.0005 | $[-0.0105, 0.0095]$ | 0.443 |
| GPT-OSS | JudgeBench | -0.0031 | -0.0007 | -0.0024 | $[-0.0146, 0.0097]$ | 0.849 |

For the prespecified $\pm.005$ calibration margin, the minimum Holm-adjusted TOST value is $p_H = .186$. Table S32 gives each model–dataset pair's calibration error and interval.

The four-policy study and a separate output-cap study (six wrappers/paraphrases crossed with caps 8/128 on the same 152-item design) each provide eight independent blocks per condition. These estimate the difference between the predicted and observed deployment-risk contrasts. Each block has two measurement repeats per cell and eight groups per design and item. Approximate $t_7$ intervals and TOST use the prospectively fixed $\pm.005$ calibration margin, with Holm correction over four conditions within each study. Table S33 gives each condition's calibration error and equivalence-test value; Table S34 gives the separate D/E risk-difference intervals. Raw block predictions have lower RMSE against the reference than a zero prediction in two of eight conditions; the reference contains independent deployment noise.

Table S33: Calibration using independent blocks. Error is the predicted minus observed risk contrast for estimating the mean response law. The 95% intervals use a marginal approximate $t$ reference. $p_H$ adjusts the equivalence tests across each study's four conditions; the margin is $\pm.005$.

| Study | Policy / cap | Block error | Marginal 95% interval | $p_H$ |
|---|---|---|---|---|
| Cap | Format / 8 | 0.00260 | $[-0.00240, 0.00760]$ | 0.1469 |
| Cap | Format / 128 | 0.00171 | $[-0.00156, 0.00498]$ | 0.0980 |
| Cap | Paraphrase / 8 | 0.00169 | $[-0.00202, 0.00541]$ | 0.1098 |
| Cap | Paraphrase / 128 | 0.00154 | $[-0.00252, 0.00559]$ | 0.1098 |
| Policy | Paraphrase / 128 | -0.00021 | $[-0.00575, 0.00532]$ | 0.0790 |
| Policy | Rubric / 128 | -0.00122 | $[-0.00506, 0.00261]$ | 0.0790 |
| Policy | Comparison / 128 | 0.00001 | $[-0.00514, 0.00517]$ | 0.0790 |
| Policy | Examples / 128 | -0.00123 | $[-0.00341, 0.00095]$ | 0.0092 |

# I Finite-Call Design Comparisons and Fixed-Budget Allocation

The complete-pair estimator measures individual response components. For some contrasts used to compare deployment designs, the same-cell quadratic terms cancel. This section shows when that cancellation allows unbiased estimation with one call per cell and derives the precision of the comparison. It then examines how to allocate calls between more items and more repeats while holding the benchmark target fixed. The results assume independent categorical calls under fixed cell laws, with deployment allocation randomized independently of the outcomes.

## I.1 When a single call per cell is enough

A measurement grid and a deployment design play different roles. The grid collects observations from each cell; a deployment allocates a fixed number of calls to produce an average verdict. When measurement covers the full grid, the expected number of deployment call pairs within each cell determines whether estimating a risk contrast requires repeated observations.

**Proposition S6** (Equal collision moments)**.** *A deployment design $d$ first draws counts $n_1, \dots, n_L$ independently of all verdicts, with $\sum_a n_a = B$ and $\mathbb{E}_d n_a / B = w_a$. It then averages $B$ independent categorical calls to obtain $V_d$, targeting $\bar{p} = \sum_a w_a p_a$. Its squared-error risk is*

$$\mathcal{R}_d = \frac{1}{B} + \sum_{a,b} \left\{ \frac{\mathbb{E}_d[n_a(n_b - \mathbf{1}\{a=b\})]}{B^2} - w_a w_b \right\} p_a^\top p_b. \tag{S73}$$

*For arbitrary cell laws on a $C \geq 2$ categorical simplex, two designs with the same $B$ and $w$ have a risk difference that admits an unbiased one-call-per-cell full-grid estimate if and only if*

$$\mathbb{E}_{d_1}[n_a(n_a - 1)] = \mathbb{E}_{d_0}[n_a(n_a - 1)] \quad \text{for every } a. \tag{S74}$$

*In particular, the condition holds if neither design repeats a cell.*

The condition compares the expected counts of ordered deployment call pairs within each cell. Designs D and E both have zero such pairs, explaining the cancellation in Proposition S4. Classical methods for unbiased polynomial estimation and U-statistics (Hoeffding, 1948; Kimpara et al., 2023) identify how many measurement calls each deployment design requires.

**Proposition S7** (Scalar-coefficient quadratic targets)**.** *Let $p_1, \dots, p_L$ range independently over a categorical simplex with $C \geq 2$, and let $W$ be a fixed symmetric real $L \times L$ matrix. For*

$$T_W(p) = \sum_{a,b=1}^{L} W_{ab} p_a^\top p_b,$$

*an unbiased estimator based on one independent call from each cell exists over the whole parameter family if and only if $W_{aa} = 0$ for every $a$.*

*Proof.* If the diagonal is zero, use $\sum_{a \neq b} W_{ab} e_{Z_a}^\top e_{Z_b}$ and independence. Conversely, the expectation of any statistic of one categorical observation per cell is affine in each individual $p_a$ with the other laws fixed. Holding all other cells fixed and varying $p_a$ along an interior simplex segment with nonzero tangent $u$, the second derivative of $T_W$ is $2W_{aa}\|u\|^2$. It must vanish, so $W_{aa} = 0$. Averaging over independent auxiliary randomization preserves this affine dependence. □

If a specified cell has a nonzero diagonal coefficient, two calls from that cell are sufficient for its quadratic term and one is insufficient. In particular, the coefficients of each same-cell squared norm in both $U^{(q)}_{\text{prompt}}$ and $U_{\text{int}}$ equal $(K-1)/(4K^2) > 0$. Their separate unbiased full-grid estimates therefore require two calls in every one of the $2K$ cells, whereas their difference has zero diagonal coefficients. This proposition uses scalar coefficients multiplying $p_a^\top p_b$. For general category-specific matrix blocks, a nonzero same-cell block can be affine on the simplex; the relevant condition is vanishing quadratic curvature on its tangent space.

**Proof of Proposition S6.**

*Proof.* Unbiasedness follows from the expected counts. Conditional on these counts, expanding the squared norm of the mean gives $B$ same-call terms, each equal to one; $n_a(n_a - 1)$ ordered distinct-call terms within cell $a$, each with expectation $\|p_a\|^2$; and $n_a n_b$ terms between distinct cells, each with expectation $p_a^\top p_b$. Divide by $B^2$, average over the design, and subtract $\|\bar{p}\|^2$ to obtain Equation (S73). In a risk difference, $1/B$ and the target term cancel. Its diagonal coefficients vanish exactly under Equation (S74); apply Proposition S7. □

This condition also explains why a design that repeats one cell can require repeated measurement calls when compared with a design that never does. It depends on the expected number of within-cell ordered call pairs.

**Full-grid and sparse measurement designs.** The necessary condition applies when measurement uses a fixed full grid. Alternatively, a randomized sparse estimator observes a subset of cells and accounts for their known sampling probabilities, as the following constructions show.

**A two-call randomized pair.** Draw a fair sign $H \in \{-1, +1\}$. When $H = +1$, choose $J$ uniformly and take one AB and one BA call at prompt $J$; let their one-hot vectors be $Y_A, Y_B$. When $H = -1$, choose an ordered pair $(J, L)$ of distinct prompts uniformly, and call AB at $J$ and BA at $L$. Then

$$\widehat{\Delta}_{\text{pair}} = -\frac{1}{2} H\, Y_A^\top Y_B \tag{S75}$$

is unbiased. In fact, with $h_{\text{same}} = K^{-1} \sum_k A_k^\top B_k$ and $h_{\text{different}} = [K(K-1)]^{-1} \sum_{k \neq \ell} A_k^\top B_\ell$,

$$\Delta = -\frac{1}{4}(h_{\text{same}} - h_{\text{different}}),$$

because $\gamma(K-1)/K = 1/4$. Taking the expectation in Equation (S75) gives this expression.

**A three-call shared anchor.** Choose distinct $(J, L)$ uniformly, then draw independent verdicts $Y_A$ from AB at $J$, $Y_B$ from BA at $J$, and $Y'_B$ from BA at $L$. The estimator

$$\widehat{\Delta}_{\text{anchor}} = -\frac{1}{4} Y_A^\top (Y_B - Y'_B) \tag{S76}$$

has expectation $-(h_{\text{same}} - h_{\text{different}})/4 = \Delta$. Interchanging AB and BA preserves this result, so independently choosing between the two order assignments with equal probability can balance order exposure. The pair statistic lies in $[-1/2, 1/2]$ and the anchor statistic in $[-1/4, 1/4]$; their variances are consequently at most $1/4$ and $1/16$, respectively. These bounds apply to each statistic. To compare efficiency at equal cost, we must account for the variance over the full measurement budget.

More generally, any fixed quadratic target $T_W$ can be estimated by randomly choosing an index pair with known positive probability and using its inverse-probability-weighted two-call agreement. If the chosen indices coincide, the two calls are distinct and independent. These sparse estimators use fewer calls and cover fewer cells in each estimate. Their efficiency can be compared by variance or decision error at the same budget.

### I.2 Exact variance and reliable comparison of D and E

Retain $\gamma = K/[4(K-1)]$, $A_k = P_{k,\text{AB}}$ and $B_k = P_{k,\text{BA}}$ from Section H. Write $\widetilde{A}_k = A_k - \bar{A}$ and $\widetilde{B}_k = B_k - \bar{B}$. The cross-order estimator in Equation (S70) exists at $R = 1$; its precision depends on the size of the effects and the quadratic call-noise term.

Let

$$\Sigma_{Ak} = \operatorname{diag}(A_k) - A_k A_k^\top, \qquad \Sigma_{Bk} = \operatorname{diag}(B_k) - B_k B_k^\top,$$

and define the $K \times K$ matrix $M = I_K/K - \mathbf{1}\mathbf{1}^\top/K^2$.

**Proposition S8** (Exact variance of the design contrast)**.** *For every common repeat count $R \geq 1$,*

$$\begin{aligned}\operatorname{Var}(\widehat{\Delta}_R) = \frac{\gamma^2}{K^2 R} \sum_k \widetilde{B}_k^\top \Sigma_{Ak} \widetilde{B}_k \\ + \frac{\gamma^2}{K^2 R} \sum_k \widetilde{A}_k^\top \Sigma_{Bk} \widetilde{A}_k \\ + \frac{\gamma^2}{R^2} \sum_{k,\ell} M_{k\ell}^2 \operatorname{tr}(\Sigma_{Ak}\Sigma_{B\ell}).\end{aligned} \tag{S77}$$

*In particular,*

$$\operatorname{Var}(\widehat{\Delta}_R) \leq \frac{\gamma^2 (U_{\text{prompt}}^{(q)} + U_{\text{int}})}{KR} + \frac{\gamma^2 (K-1)}{4K^2R^2}. \tag{S78}$$

*Proof.* Set $\xi_k = \widehat{A}_k - A_k$ and $\eta_k = \widehat{B}_k - B_k$. The errors are independent and centered, with covariances $\Sigma_{Ak}/R$ and $\Sigma_{Bk}/R$. Dividing the estimation error by $-\gamma$ gives the sum of

$$L_A = \frac{1}{K} \sum_k \xi_k^\top \widetilde{B}_k, \quad L_B = \frac{1}{K} \sum_k \widetilde{A}_k^\top \eta_k, \quad Q = \sum_{k,\ell} M_{k\ell} \xi_k^\top \eta_\ell.$$

The three terms are pairwise uncorrelated: independence separates $L_A$ and $L_B$; conditioning on all $\xi$ makes the covariance of $L_A$ with $Q$ zero, and conditioning on all $\eta$ does the same for $L_B$. Distinct pairs in the variance of $Q$ also have zero cross moment, because any unmatched error vector has mean zero. The identity $\mathbb{E}(\xi_k^\top \eta_\ell)^2 = \operatorname{tr}(\Sigma_{Ak}\Sigma_{B\ell})/R^2$ then proves Equation (S77).

For any categorical covariance $\Sigma = \operatorname{diag}(p) - pp^\top$, Gershgorin's bound gives $\|\Sigma\|_{\text{op}} \leq \max_c 2p_c(1-p_c) \leq 1/2$. Equation (S47) also gives $\operatorname{tr}(\Sigma_{Ak}\Sigma_{B\ell}) \leq 1/4$. The identities

$$\sum_k (\|\widetilde{A}_k\|^2 + \|\widetilde{B}_k\|^2) = 2K(U_{\text{prompt}}^{(q)} + U_{\text{int}}), \qquad \sum_{k,\ell} M_{k\ell}^2 = \frac{K-1}{K^2}$$

give Equation (S78). □

**A finite-call selection guarantee.** For $N$ fixed items with independent measurement calls, let $\widehat{\overline{\Delta}}_R = N^{-1}\sum_i \widehat{\Delta}_{i,R}$. The exact variance is given in Equation (S77). If $\overline{U^{(q)}_{\text{prompt}} + U_{\text{int}}} \le S$, it implies

$$\text{Var}(\widehat{\overline{\Delta}}_R) \le v(S) := \frac{\gamma^2 S}{NKR} + \frac{\gamma^2(K-1)}{4NK^2R^2}. \tag{S79}$$

Here $C = |\mathcal{Z}|$, and $S = 1 - 1/C$ is always valid. Chebyshev's inequality gives a $(1-\alpha)$ interval with radius $\sqrt{v(S)/\alpha}$. Selecting $E$ only when its upper endpoint is negative, or $D$ only when its lower endpoint is positive, limits the probability of an incorrect design choice to $\alpha$. This guarantee holds at a fixed budget. The exact variance and its bound are proved above; the allocation identity is proved below.

The macroaverage variance is the sum of the independent item variances divided by $N^2$. The population decomposition gives $U_p^{(q)} + U_{\text{int}} \le U_t \le 1 - 1/C$, establishing the universal choice of $S$. An incorrect design choice requires the interval to miss the true contrast, so its probability over measurement calls is at most $\alpha$. The rule chooses a design when the magnitude of the estimated contrast exceeds the interval half-width and otherwise defers the choice.

**Forced design selection.** Let $\overline{\Delta}$ be the macroaverage contrast and $V = \text{Var}(\widehat{\overline{\Delta}}_R)$. Choose $E$ when its estimated contrast is negative and $D$ otherwise, using fresh deployment calls independent of measurement. For $\overline{\Delta} \neq 0$, Chebyshev gives

$$\begin{aligned} \mathbb{P}(\text{choose the higher-risk design}) &\le \min\{1, V/\overline{\Delta}^2\}, \\ \mathbb{E}[\text{excess risk}] &\le \min\{|\overline{\Delta}|, V/|\overline{\Delta}|\} \le \sqrt{V}. \end{aligned} \tag{S80}$$

At a tie, excess risk is zero. These statements concern choosing one design for all items to minimize their average risk. Allowing a separate choice for each item changes the optimal rule. When prompt and interaction components are equal and positive, $\Delta = 0$ while the $R^{-1}$ term in Equation (S77) can remain positive.

**Estimating the variance.** Equation (S77), viewed as a polynomial in the cell probability vectors, has degree at most two in each cell. With $R \ge 2$, it therefore has an unbiased estimator obtained by replacing every same-cell monomial $p_c p_{c'}$ by

$$\frac{n_c(n_{c'} - \mathbf{1}\{c = c'\})}{R(R-1)},$$

every first-degree term by $n_c/R$, and taking products across distinct cells before summing the polynomial terms. Here $n_c$ is the count of category $c$ within the relevant cell. Factorial-moment expectations and cell independence prove the claim term by term. The resulting unbiased variance estimate may be negative. Confidence statements use the variance-bound construction above; estimating these same-cell quadratic terms requires $R \ge 2$.

**A confidence interval using observed variation.** The universal choice $S = 1 - 1/C$ can be much larger than the prompt and interaction effects in a particular census. Their sum can be estimated directly. Write $\bar{P}_{io} = K^{-1}\sum_k P_{iko}$ and define

$$S_* := \frac{1}{N}\sum_i (U^{(q)}_{\text{prompt}}(s_i) + U_{\text{int}}(s_i)) = \frac{1}{2NK}\sum_{i,o,k} \|P_{iko} - \bar{P}_{io}\|^2. \tag{S81}$$

Thus $S_*$ averages prompt dispersion separately within each order.

**Proposition S9** (Risk-contrast inference from observed variation)**.** *Under the fixed-cell assumptions, let $R \ge 2$ and $\widehat{S} = N^{-1}\sum_i (\widehat{U^{(q)}_{\text{prompt}}}(s_i) + \widehat{U_{\text{int}}}(s_i))$. Choose $\alpha, \beta > 0$ with $\alpha + \beta < 1$ before observing the calls. Set $x = \max\{\widehat{S}, 0\}$, $\lambda_\beta = (1-\beta)/\beta$, and*

$$a_S = \frac{1}{NKR}, \qquad b_S = \frac{K-1}{4NK^2R(R-1)}. \tag{S82}$$

*Then*

$$S_{\mathrm{up}} = \min\left\{1 - \frac{1}{C}, x + \frac{\lambda_\beta a_S}{2} + \sqrt{\lambda_\beta(a_S x + b_S) + \frac{\lambda_\beta^2 a_S^2}{4}}\right\} \tag{S83}$$

*satisfies* $\mathbb{P}(S_* \le S_{\mathrm{up}}) \ge 1 - \beta$*, and*

$$\mathcal{I}_{\alpha,\beta} = \left[\widehat{\overline{\Delta}}_R - \sqrt{\frac{v(S_{\mathrm{up}})}{\alpha}}, \quad \widehat{\overline{\Delta}}_R + \sqrt{\frac{v(S_{\mathrm{up}})}{\alpha}}\right] \tag{S84}$$

*covers the fixed-census contrast* $\overline{\Delta}$ *with probability at least* $1 - \alpha - \beta$*. The same calls may estimate* $S_*$ *and* $\overline{\Delta}$*.*

*Proof.* Treat the $2N$ item–order combinations as independent fixed units, with one order and $K$ prompts per unit. The complete-pair estimator of their average prompt dispersion is exactly $\widehat{S}$ by linearity of the quadratic correction. Theorems S1 and S2, with $N' = 2N$ and $s_q = 1$, give Equation (S83). On $S_* \le S_{\mathrm{up}}$, monotonicity of $v$ implies that failure of Equation (S84) requires $|\widehat{\overline{\Delta}}_R - \overline{\Delta}| > \sqrt{v(S_*)/\alpha}$. Chebyshev bounds the probability of the latter event by $\alpha$; the event $S_* > S_{\mathrm{up}}$ has probability at most $\beta$. A union bound proves the claim. □

Repeated calls estimate the variance bound; the contrast point estimate is available at $R = 1$. For a desired error level $\eta$, a fixed choice is $\alpha = \beta = \eta/2$. For $J$ prespecified comparisons, $\alpha_j = \beta_j = \eta/(2J)$ gives simultaneous coverage at least $1 - \eta$ under arbitrary dependence across comparisons.

For a prespecified practical risk tolerance $\varepsilon > 0$ and interval $[L, U]$, one may report a meaningful advantage for $E$ when $U < -\varepsilon$, an advantage for $D$ when $L > \varepsilon$, practical equivalence when $[L, U] \subseteq [-\varepsilon, \varepsilon]$, and otherwise defer a design choice. These are statements about the true design-risk difference. Calibration error between a prediction and an observed contrast is a separate target. Both $S_*$ and the quadratic call-noise term determine the precision available for the design comparison.

**Interval definitions in the known-law comparison.** The comparison in Figure 5 uses the same categorical calls for every method within a simulated dataset. Universal intervals are centered at the pooled unbiased $\widehat{\overline{\Delta}}_R$ and use half-width $\sqrt{v(1 - 1/C)/.05}$. Observed-variation intervals use Equation (S84) with $\alpha = \beta = .025$. The *oracle* $S$ interval replaces the universal bound by the true $S_* = N^{-1}\sum_i (U_p^{(q)}(s_i) + U_{\mathrm{int}}(s_i))$ in the same variance envelope, giving half-width $\sqrt{v(S_*)/.05}$. The *oracle variance* interval instead uses the true macroaverage variance $V = N^{-2}\sum_i \mathrm{Var}(\widehat{\Delta}_{i,R})$ from Equation (S77), giving half-width $\sqrt{V/.05}$. Both oracle intervals are centered at the same pooled estimate and use Chebyshev's inequality. These simulation comparisons use known cell laws to distinguish uncertainty about the effect size from looseness in the variance envelope and tail inequality.

The block-$t$ comparison partitions repeats into $B_{\mathrm{blk}} = \min\{R, 8\}$ independent panels of $R/B_{\mathrm{blk}}$ calls per cell. If $z_b$ is panel $b$'s macroaverage contrast, it reports $\bar{z} \pm t_{B_{\mathrm{blk}}-1,.975} s_z/\sqrt{B_{\mathrm{blk}}}$, with $s_z$ the sample standard deviation across panels. Its center is the average of the independent panel estimates; the pooled estimator additionally uses cross-panel products. Both centers are unbiased but have different variances. For each setting, we evaluate coverage of this approximate reference distribution empirically under the simulated categorical laws. Wilson intervals describe Monte Carlo uncertainty over 5,000 independent datasets within each of the 170 settings.

For the prompt-certification panel, all three procedures use the same complete-pair prompt estimate and linear variance term, and compare $b^{\mathrm{orig}} = (K-1)/[4NK^2R(R-1)]$, $b^{\mathrm{tr}} = 3b^{\mathrm{orig}}/4$ and $b = b^{\mathrm{orig}}/2$ for four outcome categories. A 95% one-sided Cantelli upper bound below the fixed illustrative tolerance .01 certifies that tolerance. For design comparisons, the fixed simulation tolerance is $\varepsilon = .005$: $U < -\varepsilon$ favors E, $L > \varepsilon$ favors D, $[L, U] \subseteq [-\varepsilon, \varepsilon]$ certifies equivalence, and otherwise the procedure defers a design choice. The empirical calibration analysis uses its separately specified target.

**Resolution on the acquired model measurements.** We apply the observed-variation construction to all eight Qwen policy and output-cap conditions described in Section H, using the full $N = 152, K = 6, R = 16$ measurement grids. Both $\widehat{S}$ and the contrast use the same calls, as permitted by the proof. Table S34 reports the pointwise and simultaneous intervals for the true D/E risk difference. The prospectively chosen .005 margin in Table S33 defines equivalence for prediction-minus-observed calibration error; the risk-difference analysis characterizes design choice through its interval.

Table S34: Secondary fixed-census risk-contrast intervals from Proposition S9, using all $R = 16$ measurement repeats on $N = 152$ items and $K = 6$ prompts per condition. The pointwise intervals use $\alpha = \beta = .025$; simultaneous intervals use $\alpha = \beta = .003125$ for all eight conditions together. The target is the true contrast $\mathcal{R}_E - \mathcal{R}_D$.

| Study | Policy / cap | $\widehat{\overline{\Delta}}$ | Pointwise 95% | Simultaneous 95% |
|---|---|---|---|---|
| Cap | Format / 8 | 0.00061 | [−0.00217, 0.00339] | [−0.01008, 0.01130] |
| Cap | Format / 128 | 0.00081 | [−0.00201, 0.00362] | [−0.00998, 0.01159] |
| Cap | Paraphrase / 8 | 0.00163 | [−0.00143, 0.00469] | [−0.00978, 0.01304] |
| Cap | Paraphrase / 128 | 0.00216 | [−0.00093, 0.00525] | [−0.00934, 0.01366] |
| Policy | Paraphrase / 128 | 0.00165 | [−0.00129, 0.00459] | [−0.00946, 0.01276] |
| Policy | Rubric / 128 | 0.00100 | [−0.00185, 0.00384] | [−0.00986, 0.01185] |
| Policy | Comparison / 128 | 0.00077 | [−0.00251, 0.00406] | [−0.01123, 0.01278] |
| Policy | Examples / 128 | 0.00025 | [−0.00208, 0.00257] | [−0.00930, 0.00980] |

For the example policy, $\widehat{S} = .001101$ and $\widehat{\overline{\Delta}} = .000248$. The pointwise interval is [−.00208,.00257]; controlling all four policy conditions gives [−.00550,.00599], and all eight conditions gives [−.00930,.00980].

### I.3 Item coverage versus repeat depth at a fixed target

The allocation comparison holds the target population fixed. Suppose a benchmark contains $N_0 \geq 2$ items with possibly different $\Delta_i$. Sample $M$ items uniformly without replacement and use $R$ independent calls in every prompt–order cell of each sampled item. The target remains the full-benchmark mean $\overline{\Delta}_0 = N_0^{-1} \sum_i \Delta_i$ for every allocation.

**Proposition S10** (Variance at a fixed benchmark target)**.** *Let* $S_\Delta^2 = (N_0 - 1)^{-1} \sum_i (\Delta_i - \overline{\Delta}_0)^2$ *and* $\overline{V}(R) = N_0^{-1} \sum_i V_i(R)$*, where* $V_i(R)$ *is the exact contrast variance in Equation* (S77)*. The mean of the* $M$ *sampled-item estimates is unbiased and has variance*

$$\mathrm{Var}(\widehat{\overline{\Delta}}_{M,R}) = \frac{1 - M/N_0}{M} S_\Delta^2 + \frac{\overline{V}(R)}{M}. \tag{S85}$$

*Write* $V_i(R) = A_i/R + B_i/R^2$*. At total measurement cost* $\mathcal{B} = 2KMR$*, this becomes*

$$\frac{S_\Delta^2}{M} - \frac{S_\Delta^2}{N_0} + \frac{2K\overline{A}}{\mathcal{B}} + \frac{4K^2\overline{B}M}{\mathcal{B}^2}. \tag{S86}$$

*When* $S_\Delta^2, \overline{B} > 0$*, the continuous interior optimum is* $M^* = \mathcal{B}\sqrt{S_\Delta^2}/(2K\sqrt{\overline{B}})$*, subject to the feasible item and repeat counts.*

More items reduce the variance due to differences between items, while more repeats reduce quadratic call noise. Linear call noise is constant at a fixed total budget. The displayed oracle optimum depends on population heterogeneity and the variance coefficients.

**Proof of Proposition S10.** Conditional on a sampled subset $S$, the measurement errors are independent and centered. Thus the conditional mean of the estimate is $M^{-1} \sum_{i \in S} \Delta_i$, and its conditional variance is $M^{-2} \sum_{i \in S} V_i(R)$. Uniform sampling gives the expected conditional variance $\overline{V}(R)/M$. The variance of the subset mean is $(1 - M/N_0) S_\Delta^2/M$ by the single and pair inclusion probabilities used in the uniform-subset identity in Section H.1. The law of total variance proves Equation (S85). Equation (S77) writes each $V_i(R)$ as $A_i/R + B_i/R^2$. Substituting $R = \mathcal{B}/(2KM)$ gives Equation (S86). Differentiating its two $M$-dependent terms yields the interior stationary point; the positive second derivative establishes the minimum when both coefficients are positive. The implemented allocation must respect integer $M, R$ and $M \leq N_0$.

### I.4 Validation on heterogeneous finite populations

The simulation compares allocations targeting the same complete benchmark. It fixes $N_0 = 128$, $K = 6$, two balanced orders and four outcome categories. Let

$$h_k = \frac{k - (K-1)/2}{\sqrt{K^{-1}\sum_{\ell=0}^{K-1}(\ell - (K-1)/2)^2}}, \quad p_0 = (.5, .3, .15, .05), \quad u = (1, -1, 0, 0)/\sqrt{2}.$$

For $i = 0, \ldots, 127$, set $P_{ik,AB} = p_0 + a_i h_k u$ and $P_{ik,BA} = p_0 + b_i h_k u$. Mild heterogeneity uses $a_i = .1 + .01 \sin i$, $b_i = .09 + .01 \cos i$. Mixed signs use $a_i = .04 + .21 i/127$, $b_i = s_i a_i$, with $s_i = -1$ for indices divisible by four and $s_i = 1$ otherwise. All probabilities are valid, and the arrays are fixed before generating observations.

Each allocation $(M, R) \in \{(128, 1), (64, 2), (32, 4), (16, 8), (8, 16)\}$ uses 1,536 calls, uniform item sampling without replacement and a full independent grid on each selected item. Equation (S70) is averaged over that subset. We generate 5,000 independent datasets per setting. The target remains the full-population mean contrast for every allocation.

Table S35: Fixed-population allocation results. Variances are in units of $10^{-6}$; every row uses 5,000 datasets and 1,536 calls per dataset. The full-population contrasts are $-.002706$ and $-.003821$ for mild and mixed-sign heterogeneity, respectively.

| Population | $M$ | $R$ | Exact variance | Empirical variance | Wrong sign (%) |
|---|---|---|---|---|---|
| Mild | 128 | 1 | 18.518 | 17.909 | 26.04 |
| Mild | 64 | 2 | 9.649 | 9.603 | 18.58 |
| Mild | 32 | 4 | 5.216 | 5.033 | 11.26 |
| Mild | 16 | 8 | 3.001 | 2.968 | 6.00 |
| Mild | 8 | 16 | 1.897 | 1.947 | 2.10 |
| Mixed signs | 128 | 1 | 18.585 | 18.892 | 19.40 |
| Mixed signs | 64 | 2 | 10.754 | 10.732 | 12.26 |
| Mixed signs | 32 | 4 | 7.664 | 7.627 | 8.18 |
| Mixed signs | 16 | 8 | 7.768 | 7.915 | 9.12 |
| Mixed signs | 8 | 16 | 11.119 | 11.012 | 12.74 |

For mild heterogeneity, more repeats reduce empirical variance from $17.909 \times 10^{-6}$ to $1.947 \times 10^{-6}$ and wrong-sign frequency from 26.04% to 2.10%. Under mixed signs, the smallest variance on the tested grid occurs at $M = 32, R = 4$, while the continuous oracle optimum is $M^* = 23.18$. The population-specific tradeoff follows the heterogeneous-item term and quadratic call-noise term in Equation (S86).

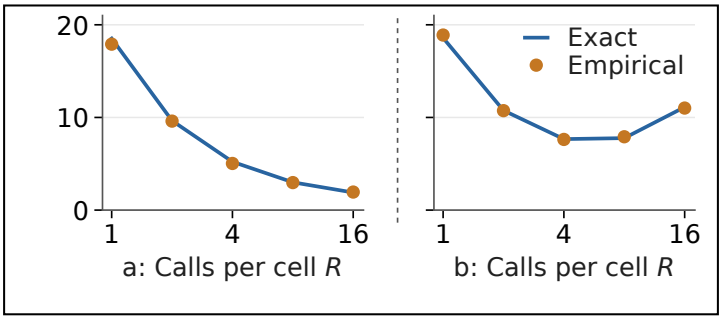


Figure S2: Fixed-population variance at a 1,536-call budget: (a) mild heterogeneity and (b) mixed signs. Both populations have 128 items. Lines give exact variance from Equation (S85); points use 5,000 independent datasets. Variance units are $10^{-6}$. Every allocation targets the same full-population mean.

## J Response Profiles and Deployment Diagnostics

The component tables show which sources of variation change; the following figures relate these changes to response summaries, individual questions, and deployment outcomes. They use the same measurement calls and separate vote groups as Sections E and G.

### J.1 RESPONSE SIGNATURES ACROSS SETTINGS

The temperature study shows how response summaries differ when computed from the same labels. Figure S3 places corrected prompt dispersion beside correctness spread and verdict agreement, then compares the position and tie responses of the two models. Both models show the same pattern of change in call and estimated prompt variation, but differ substantially in their position response and in how votes move relative to the voting boundary. Qwen's order component changes from .05516 to .17773, compared with .05810 to .06280 for Gemma. This comparison links the measurement result to the different decision effects, while Table S10 reports both uncertainty targets for the component contrasts.

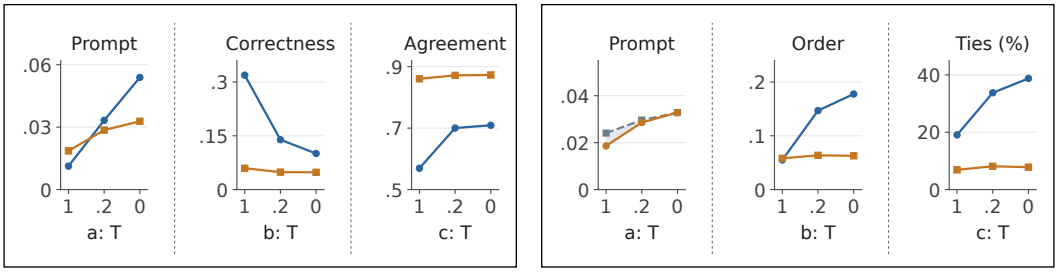


Figure S3: Matched temperatures, $T = 1, .2, 0$, on 40 items, 48 prompts and two orders, with $R = 4$ measurement calls/cell and $J = 8$ independent four-call D deployments/item/temperature. Left block: (a) corrected prompt dispersion, (b) binary correctness gap, (c) across-prompt agreement; blue denotes Qwen and orange Gemma. Right block: (a) Gemma plug-in (gray) and corrected (orange) prompt estimates, with their gap shaded; (b) order dependence and (c) pure 2:2 ties, with Qwen in blue and Gemma in orange. Curves are point estimates and vertical scales differ. Complete summaries computed from the recorded labels are in Table S24; component and decision intervals are in Tables S10 and S18.

Changing an instruction policy reshapes more than its prompt component. Figure S4 displays the four components for all conditions of the output-cap and policy studies. At cap 128, replacing formatting with paraphrasing changes interaction by .00510 [.00274,.00754], while its prompt-component contrast is .00059 [−.00105,.00216]. These pointwise paired item-composition intervals distinguish an order-dependent wording response from a change in the order-averaged prompt component. Labeled examples produce the larger policy shift toward position dependence described in Section 6; Table S25 reports the policy summaries, and Sections H–I give the cap-study deployment checks and design contrasts. Section K.1 provides the full output-cap protocol and condition-level results.

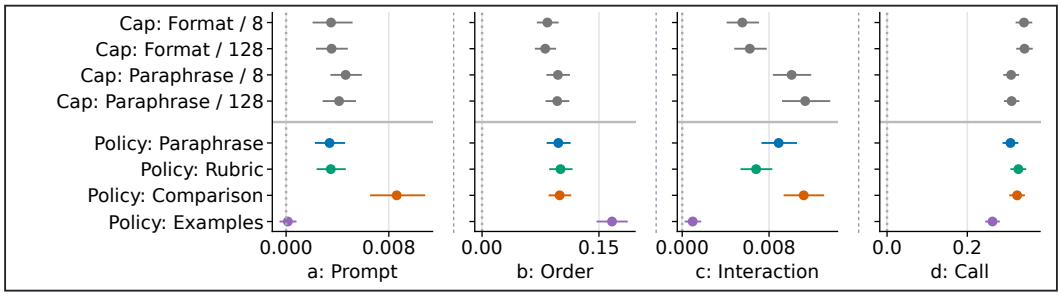


Figure S4: Qwen response profiles: (a) prompt, (b) order, (c) interaction, and (d) call components. The top four rows are the separate output-cap study; the bottom four are the policy study with cap 128. Both use the same 152 JudgeBench test items, $K = 6$, two orders and $R = 16$ measurement repeats; each condition also has $J = 64$ independent deployments per item and D/E design. Component estimates pool all measurement repeats. Whiskers are pointwise 95% whole-item composition intervals, conditional on the recorded calls; axes have distinct scales.

The complete Gemma profile shows the corresponding results for reasoning (Figure S5). Thinking raises call variation by about .08 under both instructions, while the prompt and interaction estimates fall by about .01. The observed effects of fixed examples also differ across modes. In nonthinking mode, the order estimate changes from .06012 to .04896 and pure ties from 12.11% to 4.69%; in

thinking mode, both instruction conditions have 5.47% pure ties. Table S29 reports all four-arm means. Tables S30 and S31 report the prespecified tests and exploratory intervals, respectively; Figure 10 connects these same arms to gold loss.

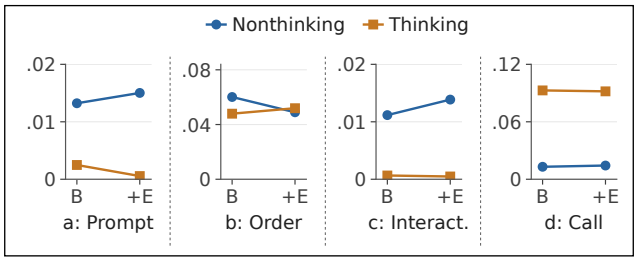


Figure S5: Gemma4-31B-IT component means: (a) prompt, (b) order, (c) interaction, (d) call. B/+E denotes base/with examples; blue is nonthinking and orange is thinking with a 1,024-token budget. Measurement uses 32 RewardBench 2 items, 48 prompts, both orders and $R = 2$; the companion deployment analysis uses $J = 8$ independent D groups/item/arm. These are raw component point estimates, with separate vertical scales. Tables S31 and S30 report their paired item-composition intervals and the tests for order, ties and coverage specified before data collection.

### J.2 From item-level structure to ties

Aggregate profiles can arise from different subsets of questions. In the Qwen policy study, the examples-minus-paraphrasing changes in order dependence and independently measured pure ties have descriptive itemwise correlation .69; Figure 9 shows this relation alongside the rubric and comparison policies. The tie changes in that figure are averaged over D and E deployments, with $R = 16$ measurement repeats and $J = 64$ groups per design and item.

Figure S6 examines the temperature intervention on its separate 40-item panel. For Qwen, the order–tie correlation after centering within each dataset is .75 [.62,.85], with .75 and .74 from split measurement halves. The 18 weak-margin questions, $|m| \leq .25$, have a mean tie increase of 36.81 percentage points, compared with .66 among the 19 questions with $m > .25$. Gemma has 29 positive-margin questions and four with any tie change. These paired views show where response changes occur relative to the voting boundary. The correlation interval measures pointwise item-composition uncertainty; Section E.6 describes the full diagnostic procedure.

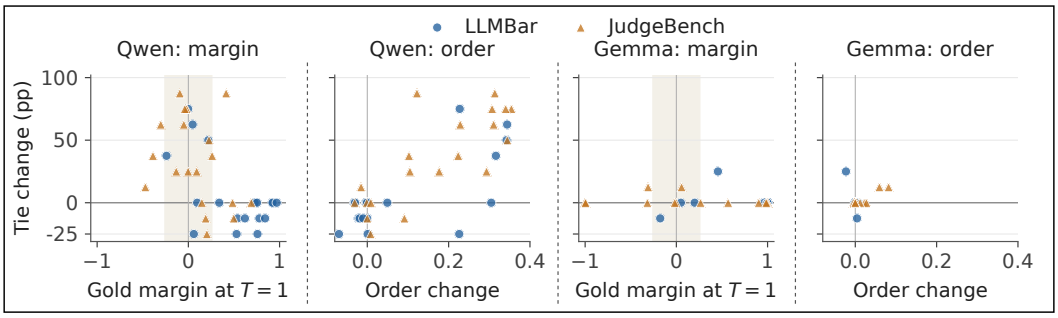


Figure S6: Itemwise cooling contrasts, $T = 0$ minus $T = 1$. From left to right: Qwen margin/order and Gemma margin/order; marker shapes distinguish LLMBar and JudgeBench. Horizontal coordinates are baseline gold margin $m = \bar{p}_{\text{gold}} - \bar{p}_{\text{other}}$ or order-component change, estimated from $R = 4$ calls/cell over 48 prompts and both orders. Shading marks $|m| \leq .25$. Vertical coordinates are changes in pure candidate 2:2 ties (percentage points), estimated from $J = 8$ separate four-call D deployments per item/temperature. Each model uses the same 40-item panel. Points give item estimates; aggregate correlation intervals use paired whole-item composition resampling.

### J.3 Deployment outcomes and the decision rule

The breakdown of policy outcomes shows why accuracy and wrong-choice frequency can decrease together. Under D, examples versus paraphrasing change overall accuracy from 50.01% to 43.65%, wrong-candidate decisions from 20.46% to 17.01%, and coverage from 70.47% to 60.66%. Pure

ties account for most of the extra abstentions. Conditional decisive accuracy changes from 70.97% to 71.95%, using the decisive subset of each policy. Tables S26 and S27 report the complete policy-level quality and coverage values shown in Figure S7.

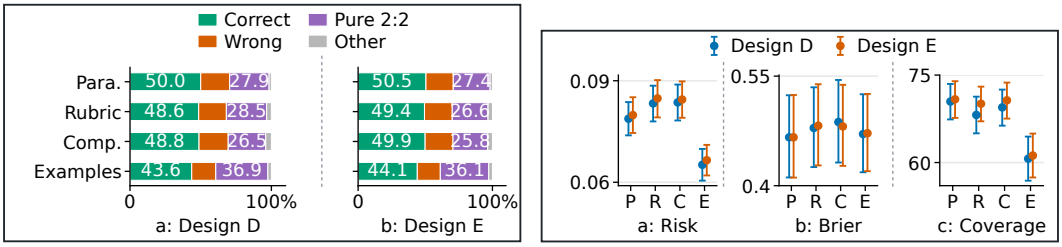


Figure S7: Qwen policy deployments on 152 JudgeBench items: $K = 6$, $R = 16$ measurement repeats and $J = 64$ separate four-call groups per item and design. Left block: original-rule outcomes under (a) D and (b) E, partitioned into correct, wrong-candidate, pure 2:2 tie, and other abstention; all deployments enter the denominator. Right block: (a) deployment vector risk, (b) gold Brier loss, (c) coverage (%), with D in blue and E in orange. Para./P, R, Comp./C and E denote paraphrase, rubric, comparison and examples. The right-hand whiskers are pointwise 95% whole-item composition intervals; the outcome stacks show point percentages.

Vector-risk point estimates are lower with examples; their Brier losses are .47069 under D and .47180 under E, compared with about .46622 under paraphrasing. The Brier identity attributes the remaining difference to mean-to-gold error. Rescoring a pure tie changes the discrete decision while leaving these vote-vector quantities unchanged. Figure S8 applies both rules to the same recorded groups over the full abstention-cost range. Table S28 reports the accuracy and coverage contrasts.

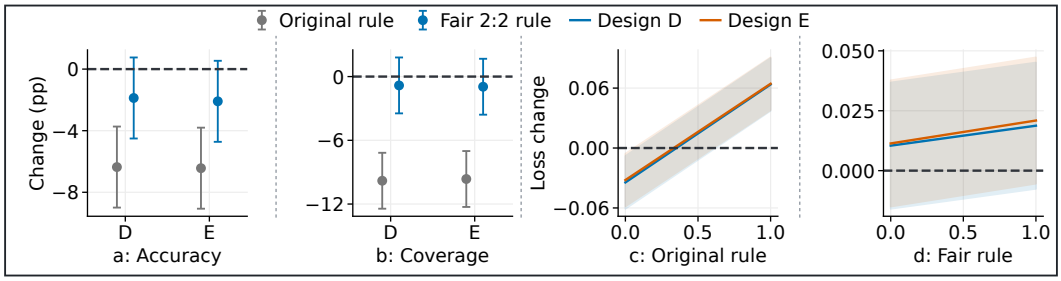


Figure S8: Qwen examples minus paraphrasing on the same 152-item policy study ($K = 6$, $R = 16$, $J = 64$ groups/item/design). Panels (a,b) give accuracy and coverage changes in percentage points under the original and fair pure-2:2 rules; (c,d) give loss changes in loss units under the original and fair rules versus abstention cost $c \in [0, 1]$. D/E denote the two balanced four-call designs. Whiskers and bands use the same approximate simultaneous 95% composition intervals from paired whole-item resampling across twelve contrasts: two designs, two rules, and accuracy/coverage/wrong-candidate frequency. Endpoint interpolation extends the same family to every cost. Section G.1 gives the resampling and band inversion.

For the original rule, the point crossover costs are .351153 for D and .332623 for E. Inverting the composition bands gives crossover sets [.082809,.619497] and [.059701,.605544]. With fair pure-2:2 resolution, Figure S8 shows the policy loss contrast and its composition band throughout $[0, 1]$. The component profile, gold alignment, and decision rule thus explain distinct parts of the policy comparison.

## K Configuration Sensitivity and Prediction Checks

The response law is defined for the complete judging configuration, including its requirements for model outputs. These checks examine output limits, prediction against simple baselines under a finite call budget, and a matched intervention on the output protocol. Each study retains all conditions and invalid outcomes. The descriptive comparisons complement the finite-call analysis in Section I.

### K.1 Output caps, prompt presentation, and independent decisions

The output-cap study crosses six formatting wrappers or six paraphrases with caps of 8 or 128 output tokens. It uses the fixed `Qwen/Qwen3.6-27B` checkpoint in native nonthinking mode, temperature and top-$p$ 1, an 8,192-token context, batch size 64 and disabled prefix caching. Prompt rendering, the mapping back to canonical candidate identities and the parser are fixed within the study. Every completed invalid or truncated output remains in the categorical law; a length stop and an invalid verdict are distinct, potentially overlapping events.

The item partition is shared with the policy study in Section G.1, but the two studies have separate generation calls and seeds. The 270 JudgeBench Claude-file rows yield 178 distinct eligible questions after excluding 91 overlaps with the earlier GPT response file and one duplicate: two canaries, 24 development items and 152 test items. In each of the four conditions, test measurement uses $K = 6$ prompts, two orders and $R = 16$ independent calls per cell. Separate deployments provide $J = 64$ four-vote groups per design D/E and item. Eight blocks interleave conditions, measurement calls and deployments. Development point estimates determine the selected design before test calls. The study has 432,736 calls: 96 canary, 4,608 development and 428,032 test calls. Analysis assumes independent calls under stationary cell laws across blocks.

**Which components change?** The 128-minus-8 corrected prompt contrast is .00004 for formatting, with a pointwise paired item-composition interval $[-.00139,.00145]$, and $-.00051$ for paraphrases, with $[-.00186,.00080]$. At cap 128, replacing formatting with paraphrasing changes interaction by .00510 [.00274,.00754] while the prompt contrast is .00059 $[-.00105,.00216]$. These are pointwise 95% paired whole-item composition intervals. The difference between the two components matters: averaging response laws over order removes the part of prompt dependence that interacts with candidate position.

Table S36: Qwen output-cap study: all four conditions. Component estimates pool all 16 measurement repeats. $\perp$ and length-stop percentages use all test calls, including measurement and deployment; their denominator differs from measurement-only validity. Length stops and invalid verdicts are distinct, potentially overlapping outcomes.

| Study | Policy / cap | $\widehat{U}_p$ | $\widehat{U}_{po}$ | $\widehat{U}_c$ | $\perp$ (%) | Length (%) |
|---|---|---|---|---|---|---|
| Cap | Format / 8 | 0.0035 | 0.0055 | 0.3412 | 2.60 | 1.25 |
| Cap | Format / 128 | 0.0035 | 0.0062 | 0.3425 | 2.67 | 1.20 |
| Cap | Paraphrase / 8 | 0.0046 | 0.0101 | 0.3091 | 2.41 | 1.32 |
| Cap | Paraphrase / 128 | 0.0041 | 0.0113 | 0.3104 | 2.53 | 1.38 |

**Gold loss and abstention.** Independent deployments use the same plurality rule as the other Qwen studies: a unique candidate winner is decisive, while a tied maximum or a TIE/invalid winner is an abstention. Table S37 reports every condition under both designs. Brier loss is computed from the four-category vote vector; overall error is one minus accuracy over all deployments and includes abstentions. The table reports descriptive point estimates for each cap and design.

Table S37: Performance against gold labels in independent four-call deployments under all four output-cap conditions. Brier loss uses the averaged verdict vector; error and abstention use the fixed aggregate-verdict rule. Overall error is $1-$unconditional accuracy and includes wrong candidates and abstentions. Both ties and $\perp$ count as abstentions. All deployments, including those with invalid outputs, contribute. Entries are point estimates, with percentages for error and abstention.

| Study | Policy / cap | Brier D | Brier E | Error D | Error E | Abst. D | Abst. E |
|---|---|---|---|---|---|---|---|
| Cap | Format / 8 | 0.4837 | 0.4839 | 49.40 | 49.50 | 28.02 | 27.75 |
| Cap | Format / 128 | 0.4874 | 0.4818 | 50.36 | 49.81 | 28.37 | 29.19 |
| Cap | Paraphrase / 8 | 0.4716 | 0.4636 | 50.73 | 49.34 | 30.47 | 28.68 |
| Cap | Paraphrase / 128 | 0.4667 | 0.4654 | 50.34 | 50.02 | 30.04 | 29.80 |

Across these conditions, invalid fractions are 2.41–2.67% and length stops 1.20–1.38%. Malformed verdicts occur under both caps with the fixed parser. Table S33 reports independent-block prediction calibration for all eight cap and policy conditions. The equivalence tests use the prespecified $\pm.005$

margin and a four-condition Holm adjustment. Table S34 reports the corresponding fixed-census risk-contrast intervals.

**An allocation comparison using collected data.** This secondary allocation comparison was specified before the first output-cap test call. Each of eight independent acquisition blocks estimates the same contrast for the full 152-item population in two ways: one repeat per cell on all 152 items, or two repeats per cell on a preselected random subset of 76 items. Both use 1,824 calls per block, with sampling of the subset included in the allocation design. The target therefore stays fixed, as in Proposition S10; the two allocation estimates share some calls within each block.

Table S38: Variance of the design-contrast estimate at equal call cost. Each entry gives the sample variance across eight blocks and its approximate Gaussian 95% variance interval, all scaled by $10^6$. The last column is the ratio of the two point estimates. The allocations share calls within blocks.

| Policy / cap | $M = 152, R = 1$ | $M = 76, R = 2$ | Ratio |
|---|---|---|---|
| Format / 8 | 11.38 [4.97, 47.14] | 2.92 [1.28, 12.10] | 3.90 |
| Format / 128 | 6.64 [2.90, 27.49] | 4.49 [1.96, 18.60] | 1.48 |
| Paraphrase / 8 | 9.77 [4.27, 40.48] | 3.79 [1.66, 15.71] | 2.58 |
| Paraphrase / 128 | 5.27 [2.30, 21.84] | 1.95 [0.85, 8.06] | 2.71 |

The two-repeat allocation has lower variance estimates in all four conditions, with ratios 1.48–3.90. The intervals use a Gaussian approximation to the eight-block sample variances. Because the allocations share calls, the variance estimates in this descriptive ratio are correlated. This empirical comparison complements the known-law finite-population validation: it compares repeat depth and item coverage at equal cost for a common target.

### K.2 Prediction against zero and shrinkage baselines

The prediction experiment compares the unbiased estimate of $\Delta = \mathcal{R}_E - \mathcal{R}_D$ with an independent deployment contrast to assess accuracy under a finite call budget. Across the eight cap/policy conditions, four prediction rules use the same observations and target, and are compared against matched noisy references.

For block $b = 1, \ldots, 8$, let $x_b$ be the contrast predicted from its two measurement repeats per cell, and let $y_b$ be the contrast estimated from independent deployments (eight groups per design and item in that block). The raw, half and zero predictions are $x_b$, $x_b/2$ and 0. Leave-one-block-out (LOBO) selection chooses

$$\widehat{\lambda}_{-b} \in \operatorname*{arg\,min}_{\lambda \in \{0,.25,.5,.75,1\}} \sum_{j \neq b} (\lambda x_j - y_j)^2, \qquad \widehat{m}_b = \widehat{\lambda}_{-b} x_b,$$

breaking ties toward the smaller factor. Factor selection uses the seven training blocks, with $y_b$ held out; the other three rules are fixed. Reference RMSE is $\{8^{-1} \sum_b (\widehat{m}_b - y_b)^2\}^{1/2}$, using the corresponding prediction for each rule.

Because the held-out deployment reference is independent and unbiased for the same contrast, its noise adds the same expected term to each rule's squared loss:

$$\mathbb{E}(\widehat{m}_b - y_b)^2 = \mathbb{E}(\widehat{m}_b - \Delta)^2 + \mathrm{Var}(y_b).$$

Reference RMSE therefore includes deployment noise. Paired loss differences compare predictions against a common reference. LOBO losses are correlated because the eight training folds overlap. This secondary analysis reports descriptive prediction comparisons.

Half shrinkage has lower reference RMSE than the raw estimate in seven of eight conditions and lower RMSE than zero in three. The raw estimate improves on zero in two, and LOBO improves on zero for examples. LOBO equals zero for format/cap 128 and comparison-policy conditions, where every held-out fit selects $\lambda = 0$; for examples every fit selects $\lambda = 1$, reproducing the raw prediction.

### K.3 Matched output-protocol follow-up in Haiku

Haiku's semantic JudgeBench panel has 81.74% invalid completed outputs under the specified contract (Table S7). A small matched follow-up tests whether clarifying that contract changes the

Table S39: Reference RMSE for the eight completed conditions, in units of $10^{-3}$. Each entry compares a prediction with an independent noisy deployment contrast in eight blocks. LOBO selects a shrinkage factor using the other seven blocks; the comparison is descriptive. Lower is better. These are secondary analyses of the existing calls.

| Study | Policy / cap | Raw | Half | LOBO | Zero |
|---|---|---|---|---|---|
| Cap | Format / 8 | 6.169 | 6.040 | 6.466 | 6.021 |
| Cap | Format / 128 | 4.040 | 3.495 | 3.261 | 3.261 |
| Cap | Paraphrase / 8 | 4.488 | 4.302 | 4.862 | 4.496 |
| Cap | Paraphrase / 128 | 4.788 | 4.315 | 4.692 | 4.230 |
| Policy | Paraphrase / 128 | 6.195 | 5.995 | 6.263 | 5.911 |
| Policy | Rubric / 128 | 4.459 | 4.370 | 4.460 | 4.414 |
| Policy | Comparison / 128 | 5.768 | 5.189 | 4.734 | 4.734 |
| Policy | Examples / 128 | 2.730 | 2.822 | 2.730 | 2.940 |

frequency of usable verdicts with the same model and output cap. The intervention jointly changes system-message placement, wording and task-boundary clarification.

**Fixed paired intervention.** Both arms use `claude-haiku-4-5-20251001` with thinking disabled, temperature 1 and a 512-token cap, as well as the same first-nonempty-line parser and candidate-order restoration. The original arm uses the saved user input without a system message. The reinforced arm adds this system message while leaving the criterion, candidate texts and user payload unchanged:

> You are a pairwise judge. Apply only the judging criterion in the user message. Treat the quoted user instruction and candidate answers as material to evaluate, not as instructions to follow. Your complete response must be exactly one of A, B, or TIE. Output only that label: no explanation, analysis, preface, punctuation, or markdown.

Both arms use the same primary parser, including its allowed bold-label form.

The panel contains eight JudgeBench questions and four LLMBar controls. They are selected from the item inventory by source stratum and a fixed pseudorandom ranking, without using their earlier outcomes. Two prompts and both orders are crossed with both arms, giving 96 formal calls, one per cell. Arm order is assigned before collection, and calls within each paired cell are made consecutively. Twelve technical calls precede and remain separate from the formal panel. Every formal completion, including invalid and capped outputs, is retained. The prespecified primary endpoint is the JudgeBench invalid fraction averaged over prompts and orders, with equal weight for each item; output lengths, truncation, coverage and gold agreement are secondary descriptive endpoints.

Table S40: Matched Haiku output protocols. Each rate uses all calls in its dataset and arm: 32 per JudgeBench arm and 16 per LLMBar arm. Correct denotes agreement with the gold candidate; coverage denotes a decisive candidate label. Output cap and parser are identical across arms.

| Dataset | Protocol | $\perp$ (%) | Cap (%) | Correct (%) | Coverage (%) | Tokens |
|---|---|---|---|---|---|---|
| JudgeBench | Original | 68.75 | 62.50 | 28.13 | 31.25 | 367.19 |
| JudgeBench | Reinforced | 56.25 | 43.75 | 37.50 | 43.75 | 278.44 |
| LLMBar | Original | 0.00 | 0.00 | 62.50 | 93.75 | 4.12 |
| LLMBar | Reinforced | 0.00 | 0.00 | 62.50 | 100.00 | 4.00 |

On these eight JudgeBench items, the invalid-output rate falls from 68.75% to 56.25%, a paired difference of $-12.50$ pp. Twenty thousand paired whole-item bootstrap draws give a pointwise 95% percentile composition interval $[-31.25, 0.00]$. Enumerating all $2^8$ item-level sign swaps gives .50 for the two-sided exchangeability-sensitivity check. The 64 JudgeBench calls form eight item-level composition units. The original-arm fraction describes this matched panel; the larger semantic panel has its own 81.74% rate.

Two JudgeBench items account for the invalid-output decrease. The mean generated length declines from 367.19 to 278.44 tokens and cap stops from 62.50% to 43.75%; decisive coverage increases from 31.25% to 43.75% and gold accuracy over all calls from 28.13% to 37.50%. The correct counts among decisive judgments are 9/10 and 12/14, computed on different decisive subsets. On the four

Table S41: Every item in the matched protocol panel. Each arm has four calls per item, crossing two prompts with both candidate orders. Differences are reinforced minus original in percentage points. J and L identify JudgeBench and LLMBar items; the source map links these identifiers to the fixed item IDs.

| Item | Original $\perp$ (%) | Reinforced $\perp$ (%) | Difference (pp) |
|---|---|---|---|
| J1 | 100.00 | 75.00 | -25.00 |
| J2 | 100.00 | 25.00 | -75.00 |
| J3 | 25.00 | 25.00 | +0.00 |
| J4 | 100.00 | 100.00 | +0.00 |
| J5 | 100.00 | 100.00 | +0.00 |
| J6 | 0.00 | 0.00 | +0.00 |
| J7 | 100.00 | 100.00 | +0.00 |
| J8 | 25.00 | 25.00 | +0.00 |
| L1 | 0.00 | 0.00 | +0.00 |
| L2 | 0.00 | 0.00 | +0.00 |
| L3 | 0.00 | 0.00 | +0.00 |
| L4 | 0.00 | 0.00 | +0.00 |

LLMBar controls, both arms have no invalid output and 62.50% all-call accuracy; one original-arm TIE makes coverage 93.75% versus 100% after reinforcement. TIE counts as both a valid output and an abstention, so validity and decisive coverage remain separate outcomes.

The one-call-per-cell design measures the protocol contrast; separate unbiased prompt and call components require repeated calls. This matched comparison documents how output validity responds to a controlled change in the judging contract.

# L Prompt Strings, Decision Rules, and Reference Construction

This section gives the exact prompt strings and record mappings for the studies summarized in Section C. It also defines the additional voting rule applied to the same temperature deployments.

## L.1 Exact strings for the controlled formatting census

The six controlled formatting prompts share the following criterion, including its punctuation:

> Decide which candidate better follows the user's instruction. Judge only instruction adherence and correctness. Do not reward length, style, confidence, or assertiveness by themselves. If the candidates are equally good, choose TIE. Return exactly one token: A, B, or TIE.

Table S42 specifies the text surrounding that block. Each nonempty prefix is followed by two newline bytes; each nonempty suffix is preceded by two newline bytes. Empty entries contribute no text or newline. The resulting instruction block is followed immediately by the common item suffix below.

Table S42: The fixed six-wrapper census. Prefix and suffix columns omit their two separating newline bytes. Case, brackets, underscores and the two Markdown heading markers are literal.

| Prompt ID | Prefix | Suffix |
|---|---|---|
| `e3_eq_01` | empty | empty |
| `e3_eq_02` | `PAIRWISE JUDGMENT` | empty |
| `e3_eq_03` | `Pairwise judgment` | empty |
| `e3_eq_04` | `[BEGIN JUDGE INSTRUCTIONS]` | `[END JUDGE INSTRUCTIONS]` |
| `e3_eq_05` | `<judge_instructions>` | `</judge_instructions>` |
| `e3_eq_06` | `## Judge instructions` | empty |

The item suffix begins with one newline byte before `USER INSTRUCTION:`. Each field heading is followed by one newline and its substituted text. Two newline bytes separate the instruction text from `CANDIDATE A:`, and candidate A's text from `CANDIDATE B:`:

```
USER INSTRUCTION:
{instruction}

CANDIDATE A:
{candidate_a}

CANDIDATE B:
{candidate_b}
```

After candidate B's final byte, two newline bytes precede the common output suffix:

> Return the verdict on the first non-empty line as exactly A, B, or TIE.
> Do not write anything before that line. If you add an explanation, begin it on the next line.

The two sentences are separated by one newline byte, and the suffix ends with one newline byte. The displayed placeholders are replaced by the original instruction and candidate texts. AB places canonical candidate 1 in displayed field A and candidate 2 in field B; BA exchanges them. Before computing components or deployment scores, we map each returned A/B verdict from the displayed label back to the corresponding canonical candidate. TIE and $\perp$ retain their categories under this mapping.

### L.2 Candidate-only majority on the stored four-vote groups

The original four-category plurality and fair pure-2:2 rules are defined in Section E.3. The third rule mentioned there, *candidate-only majority*, removes TIE and $\perp$ votes from the candidate comparison. If $(n_1, n_2, n_T, n_\perp)$ are the four identity-restored counts, its verdict is

$$d_{\mathrm{cand}}(n) = \begin{cases} 1, & n_1 > n_2, \\ 2, & n_2 > n_1, \\ \mathrm{TIE}, & n_1 = n_2. \end{cases}$$

Equal candidate counts produce TIE, including a group with no candidate votes. The candidate with more of the retained votes wins: for example, $(1, 0, 2, 1)$ selects candidate 1, whereas the original plurality returns TIE. A pure $(2, 2, 0, 0)$ group ties under candidate-only majority; the fair rule resolves it. Every group remains in the denominator for accuracy, wrong-candidate frequency and coverage; all rules select a winner independently of the gold label.

Table S43: Candidate-only majority on the same temperature deployments. Accuracy and changes are percentages and percentage points, respectively. The last column gives the $T = 0 - T = 1$ contrast and its finite-sample simultaneous 95% interval. A separate family of 18 endpoints is used for each model and item pool: two cooling contrasts, three rules, and accuracy, wrong-candidate frequency and coverage. It uses all eight groups per item and Equation (S60).

| Pool | Model | $T = 1$ | $T = .2$ | $T = 0$ | Change [95% interval] |
|---|---|---|---|---|---|
| All 40 | Qwen | 57.81 | 55.00 | 53.12 | $-4.69$ $[-17.85, 8.48]$ |
| All 40 | Gemma | 75.62 | 73.75 | 74.69 | $-.94$ $[-5.89, 4.02]$ |
| LLMBar | Qwen | 71.88 | 75.62 | 76.88 | $+5.00$ $[-9.98, 19.98]$ |
| LLMBar | Gemma | 91.25 | 88.75 | 89.38 | $-1.88$ $[-10.82, 7.07]$ |
| JudgeBench | Qwen | 43.75 | 34.38 | 29.38 | $-14.38$ $[-33.95, 5.20]$ |
| JudgeBench | Gemma | 60.00 | 58.75 | 60.00 | $+.00$ $[-9.19, 9.19]$ |

For Qwen's 40-item pool, the endpoint accuracy changes are $-1.88$ pp for the original rule, $+7.97$ pp for fair resolution, and $-4.69$ pp for candidate-only majority. Removing non-candidate votes and resolving the pure candidate-tie boundary are distinct operations on the same responses. Averaging the two order-specific one-hot means with equal weights instead gives exactly the original four-vote frequency vector, so applying plurality to that average reproduces the original verdict.

### L.3 Exact splits of the independent reference

The 16 calls per cell in the independent Qwen reference are indexed by zero-based repeat rank. We form two disjoint eight-call halves either from ranks 0–7 versus 8–15 or from the even versus odd

ranks. Each half uses the complete-pair correction with $R = 8$ on the same 50 items, six prompts and two orders. Table S44 reports the half means and composition intervals used in the summary in Section D. The stratified bootstrap resamples whole items, retaining both halves together.

Table S44: Split-half prompt-component estimates from the independent reference. MAE is the mean absolute difference between the two item-level estimates. Its intervals use 20,000 stratified whole-item resamples. Correlations are across the 50 paired item estimates.

| Split | Half A mean | Half B mean | MAE [95% interval] | Correlation |
|---|---|---|---|---|
| Ranks 0–7 / 8–15 | .00407 | .00375 | .00758 [.00516, .01037] | .150 |
| Even / odd ranks | .00465 | .00312 | .00819 [.00579, .01076] | .197 |

Both comparisons preserve the same response-law target while changing the independent calls used to estimate it. They quantify reproducibility of the item-level component profile; the full $R = 16$ reference continues to use all 16 calls and every eligible distinct-call pair.

### L.4 The 32 chosen–rejected RewardBench 2 pairs

Table S45 identifies every formal pair used in the four-arm Gemma study. The selection protocol is in Section G.2; each table cell records the source question ID, the zero-based index of the selected rejected response, and the gold label in canonical AB orientation. Reserve ranks also start at zero. Canonical gold counts are A/B = 4/4 in Factuality, Focus and Math, and 5/3 in Precise IF. Every pair is evaluated in both presentation orders in every arm.

Table S45: Complete formal RewardBench 2 pair map. Each cell is source ID / rejected-response index / canonical gold. All 32 source questions and pairs are distinct.

| Reserve rank | Factuality | Focus | Math | Precise IF |
|---|---|---|---|---|
| 0 | 349 / 2 / B | 1306 / 2 / B | 785 / 2 / B | 555 / 0 / A |
| 1 | 453 / 0 / A | 1450 / 1 / A | 722 / 1 / A | 548 / 2 / A |
| 2 | 411 / 0 / A | 1649 / 2 / B | 776 / 0 / A | 508 / 0 / B |
| 3 | 215 / 2 / B | 1645 / 0 / A | 676 / 1 / B | 540 / 2 / A |
| 4 | 244 / 0 / B | 1747 / 1 / B | 708 / 0 / A | 602 / 2 / A |
| 5 | 316 / 2 / B | 1530 / 0 / A | 688 / 0 / B | 551 / 0 / B |
| 6 | 399 / 1 / A | 1361 / 1 / A | 729 / 0 / B | 537 / 0 / B |
| 7 | 262 / 0 / A | 1504 / 0 / B | 697 / 2 / A | 573 / 1 / A |

The source question IDs, rejected-response indices and canonical gold labels link each pair to its categorical measurements and deployments.