WHEN THE JUDGE CHANGES THE VERDICT: RANK INSTABILITY AND AN AUDITING PROTOCOL FOR LLM-AS-A-JUDGE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM-as-a-judge reliability is typically treated as a model property, but we show it requires auditing along two orthogonal axes. Rank stability across sessions is not guaranteed: even with identical inputs, greedy decoding, and the same weights, a served judge can realize distinct deterministic scoring functions across sessions. This failure—operating as a score re-permutation across documents—is invisible to within-session retests and distributional checks, yet reverses downstream difference-in-differences estimates on up to $5/6$ dimensions. Calibration to a shared reference is pervasively weaker: across 8–11 open-weight judges, cross-session agreement is near-perfect ($\bar{\rho} > 0.97$), yet alignment to the production scale correlates at only $\rho = 0.07$–0.70 (median 0.50). We propose two mechanism-agnostic protocols: a 20-document reference anchor detecting session deviations, and multi-session consensus recovering majority-mode estimates when the anomalous mode is minority. On 188k legal documents, the anchor classifies all 16 campaigns perfectly; consensus eliminates worst-case $5/6$ sign-flips when $p_B = 0.1$ and reduces them three-fold at $p_B \approx 0.25$. The framework generalizes to new tasks or stacks by substituting the anchor; external validation of the reference mode against human annotations remains outstanding for this domain. Reliability is a property of a measurement campaign, not a model alone.

1 INTRODUCTION

LLM-as-a-judge systems are increasingly used as measurement instruments: their scores rank model outputs, train reward models, and enter causal analyses of policy (Liu et al., 2023; Zheng et al., 2023; Kim et al., 2024; Lambert et al., 2025; Tan et al., 2025). In these roles the scores are consumed not only as averages but as orderings—which candidate is better, which documents changed after a reform—so measurement quality has two components: the scores must be *calibrated* against a meaningful reference scale, and they must be *stable*, in the sense that the same documents receive the same scores when the measurement is repeated (Siska et al., 2024; Gentzkow et al., 2019). Practitioners operationalize the two requirements with two procedures: calibration, aligning a judge to a reference scale, and validation, re-running the judge and checking agreement (Belz et al., 2023; Dev et al., 2026).

Each procedure, however, audits the measurement only where it is easiest to audit, and each leaves a structural gap. Calibration aligns the marginal distribution of scores; it does not determine which document receives which score, so two judges with identical marginals can rank every document differently—and the ranking is precisely what many downstream uses consume. Validation, as usually practiced, re-runs the judge *within a single serving session*; a served judge, however, is one deterministic function of that session's state, and that state is set by a serving stack the practitioner does not control (Kwon et al., 2023; Zheng et al., 2024). Re-scoring the same documents inside a session returns the same scores by construction, so a within-session agreement check is a check of one function, not of the deployment. The failures that occur *across* sessions are exactly the ones such checks cannot see: if an alternative function preserves the score distribution, even cross-session statistics built on means and marginals pass. The remedies that dominate practice—position swapping, judge panels, deterministic decoding, and self-consistency—act on the judge's inputs or on decoding variance, not on the realized session function (Schilcher et al., 2025; Hu et al., 2024; Holtzman et al.,

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.

2020; Wang et al., 2023). Whether these gaps are benign, and what to do about them, has not been established (Yagubyan, 2026; Norman et al., 2026).

This paper shows that both gaps are real on production-grade systems. On a widely deployed judge, scoring 188,416 Chinese court judgments (Xiao et al., 2018), we caught the serving stack realizing two nearly orthogonal deterministic scoring functions on identical inputs—reversing the sign of downstream difference-in-differences (DID) estimates (Goodman-Bacon, 2021; Callaway & Sant'Anna, 2021; Sun & Abraham, 2021; Borusyak et al., 2024; de Chaisemartin & D'Haultfœuille, 2020) on up to five of six dimensions. The instance is invisible to every standard check, because a re-permutation of scores across documents leaves the column-wise statistics unchanged and a single session realizes only one function. We then show how to audit and contain the failure with two mechanism-agnostic protocols.

Contributions. Our contributions are threefold:

1. *We identify and quantify a reliability axis distinct from calibration.* We separate the reliability of an LLM judge into cross-session rank stability and calibration to a shared reference, and show on 8–11 open-weight judges that the two are orthogonal: every non-production judge is rank-stable across sessions yet calibrates to the production scale at $\rho = 0.07$–0.70. Marginal alignment leaves rankings at 0.31–0.67, and the residual is predictable out of sample ($r = -0.90$).
2. *We characterize what standard audits cannot see.* We show, and verify in simulation, that a rank re-permutation of scores across documents leaves every column-wise statistic unchanged (means, score distributions, paired tests), while a single session realizes only one deterministic function; so neither distributional checks nor within-session retests can detect a mode switch; only cross-session, rank-sensitive statistics respond.
3. *We document a real instance and propose two defensive protocols.* We record a served judge realizing two nearly orthogonal deterministic scoring functions, and propose two mechanism-agnostic protocols, a 20-document reference anchor and a multi-session consensus, that detect and mitigate it, validated on the recorded campaigns.

2 Related Work

We organize related work into four strands—judge reliability, output reproducibility, serving-system determinism, and measurement error; each touches one of the two axes we isolate, and none separates them. **LLM-as-a-judge reliability.** Using LLMs as evaluators has become the dominant paradigm for NLP, MT, and instruction-following evaluation (Liu et al., 2023; Zheng et al., 2023; Chiang et al., 2024; Kim et al., 2024; Tan et al., 2025). Prior work has catalogued biases of LLM judges (position, verbosity, self-preference, order effects) and proposed mitigations such as position-swapped pairwise comparison, judge panels, and multi-agent debate (Schilcher et al., 2025; Hu et al., 2024; Panickssery et al., 2024; Chan et al., 2024; Rafailov et al., 2023). Reliability here is assumed to be a property of the *fixed* judge, whereas it lives in the served *instance* and *scale*. We separate the axes this strand conflates: cross-session rank instability and cross-model calibration drift (Siska et al., 2024). **Output reproducibility.** A second strand examines sampling-induced variability (temperature, top-p, seed effects) (Holtzman et al., 2020; Sainz et al., 2023; Xiang et al., 2024) and recommends deterministic decoding, ensembling, or self-consistency (Wang et al., 2023); a systematic study argues temperature should be tuned to the task—low settings suit stability and reproducibility, while higher settings help in ambiguous cases (Li et al., 2026). Our finding differs from sampling variance: it persists at temperature 0 with greedy decoding, and the modes are qualitatively different scoring functions, not draws from one distribution. Recent audits document a reproducibility crisis, run-to-run flip rates, and multi-trial aggregation (Belz et al., 2023; Yagubyan, 2026; Norman et al., 2026; Dev et al., 2026). Our failure is *modal* rather than noisy: coin-flip audits average over stochastic flip rates, whereas a re-permutation is invisible to the mean-level statistics such audits compute; contamination (Oren et al., 2024) is a different concern (training-data leakage, not involved here). **Serving-system determinism.** A third strand studies inference serving as a source of nondeterminism: vLLM-like engines introduce parallelism and kernel-level nondeterminism (reduce operations, batching order, attention kernels) that change floating-point results run to run (Kwon et al., 2023; Zheng et al., 2024), and different

engines or scheduler configurations can yield different outputs for the same request (Lepikhin et al., 2021; Shazeer et al., 2017).

Measurement-error framing. A fourth strand, the measurement literature, treats the scoring function as a noisy instrument; classical results show label noise attenuates but does not reverse effect estimates (Natarajan et al., 2013). In contrast, our B-mode reverses the sign of downstream DID coefficients on up to 5 of 6 dimensions, an effect of *systematic re-permutation* rather than additive noise. In causal inference, DID estimates depend on which units are compared and when treatment varies (Goodman-Bacon, 2021; Callaway & Sant'Anna, 2021; Sun & Abraham, 2021; Borusyak et al., 2024; de Chaisemartin & D'Haultfœuille, 2020); our contribution is that *the outcome variable itself* can be non-uniquely measured. Text-as-data methods treat document-derived measures as error-prone and call for reliability reporting per study (Gentzkow et al., 2019; Xiao et al., 2018; 2021; Huang et al., 2024); for LLM judges, reliability is a property of a *measurement campaign*, not a model. In sum, prior work treats judge variability as stochastic noise or judge *bias*; we show it can also be deterministic and *modal*, and we separate cross-session rank stability from cross-model scale calibration as two orthogonal reliability axes (Kwon et al., 2023; Zheng et al., 2024; Holtzman et al., 2020; Sainz et al., 2023; Wang et al., 2023; Li et al., 2026; Yagubyan, 2026; Norman et al., 2026; Dev et al., 2026; Xu et al., 2026).

3 PRELIMINARIES

We fix notation for the measurement a judge performs and for the statistics used to compare two measurements of the same documents. We write vectors in bold and use $\mathbf{s}_i^{(g)}$ for the score vector assigned to document i by measurement g.

3.1 THE MEASUREMENT SETTING

An *LLM-as-a-judge deployment* consists of a model $\mathcal{M}$ with frozen weights θ, a fixed scoring prompt π, and a serving environment $\mathcal{E}$ (inference engine, scheduler, hardware, runtime state) that realizes $\mathcal{M}$ as a callable function. For a document d from a corpus $\mathcal{D}$, a served session s realizes a function

$$f_s(d) = \mathrm{parse}\big(\mathcal{M}(\pi, d; \theta, \mathcal{E}_s)\big) \;\in\; \{1, \ldots, 5\}^6, \tag{1}$$

where $\mathcal{E}_s$ is the environment state during session s and parse extracts the six integer ratings from the JSON output. We score with greedy decoding (temperature 0) throughout, so f_s is deterministic given the session state; within-session retests on our data agree at exact rates above 0.95 (Section 5.2).

A *measurement campaign* is a tuple $g = (s, \mathcal{D}_g)$ of a session and a document subset. It produces a score matrix $\mathbf{S}^{(g)} \in \mathbb{Z}^{N_g \times 6}$ with $S_{ik}^{(g)} = \big[f_s(d_i)\big]_k$. Two campaigns are *co-measured* if they score the same document set, in which case we compare them column-wise.

3.2 AGREEMENT STATISTICS

For two co-measured campaigns g, h on dimension k, we use the Spearman rank correlation

$$\rho_k(g, h) = 1 - \frac{6 \sum_{i=1}^{N} \big(r_{ik}^{(g)} - r_{ik}^{(h)}\big)^2}{N(N^2 - 1)}, \tag{2}$$

where $r_{ik}^{(g)}$ is the rank of $S_{ik}^{(g)}$ within its column, and the exact-agreement rate

$$e_k(g, h) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{1}\big[S_{ik}^{(g)} = S_{ik}^{(h)}\big]. \tag{3}$$

We aggregate over dimensions by averaging: $\bar{\rho}(g, h) = \frac{1}{6} \sum_k \rho_k$ and $\bar{e}(g, h) = \frac{1}{6} \sum_k e_k$.

The two statistics respond to different changes: $\bar{\rho}$ tracks whether two measurements order the documents in the same way, while $\bar{e}$ tracks whether they attach the same values to the same documents. We report both throughout, and Section 5.3 uses their combination to characterize the failure we observe.

Throughout, a fixed reference measurement f_A (the production measurement) anchors the A mode; the reliability question we ask is whether a new campaign realizes the same function. Agreement measured on a finite sample constrains the two functions on that sample, and Section 4.3 asks how large a sample is needed for a decision.

4 METHOD

We formalize the state-classification problem and two defensive protocols. We use the notation of Section 3: a reference measurement f_A, campaigns g, and agreement statistics $\bar{\rho}, \bar{e}$.

4.1 RANK-SENSITIVE AUDIT METRICS

The protocols below are audited with statistics that track *which document received which score*. The per-document row Spearman

$$\rho_i(A, B) \;=\; \mathrm{Spearman}\big(\mathbf{S}_{i\cdot}^{(A)},\; \mathbf{S}_{i\cdot}^{(B)}\big) \tag{4}$$

over the six dimensions of one document, reported through its mean $\bar{\rho}_{A_1B_1}$ and its distribution across documents (shares of rank-inverted, unchanged, and intermediately unstable documents); the per-dimension intraclass correlation

$$\mathrm{ICC}_k(A, B) \;=\; \frac{\mathrm{BMS}_k - \mathrm{WMS}_k}{\mathrm{BMS}_k + \mathrm{WMS}_k}, \tag{5}$$

computed from between- and within-document mean squares—the agreement metric most evaluation protocols implicitly rely on, responsive to re-permutation yet silent about the per-document distribution; and, modeling mode B as a re-permutation in which a fraction π of documents is independently redrawn, the estimator

$$\hat{\pi} \;=\; 1 - \sqrt{\bar{\rho}_{A_1B_1}/\bar{\rho}_{A_1A_2}} \tag{6}$$

with its derivation in Appendix G.

4.2 STATE CLASSIFICATION

Let f_A be the scoring function realized by the production session (mode A). We hypothesize that the served judge can realize a small set of functions $\{f_A, f_B, \dots\}$, and define the *state classification problem*: given a campaign g co-measured with a reference campaign on a shared document set $\mathcal{D}$, decide which function g realized. The classification rule is

$$\mathrm{mode}(g) \;=\; \begin{cases} A, & \bar{\rho}(g, A) > 0.7, \\ B, & \bar{\rho}(g, A) < 0.3, \\ \text{ambiguous}, & \text{otherwise.} \end{cases} \tag{7}$$

The thresholds are deliberately wide: no campaign in our data is near the boundary (A: 0.94–0.99; B: ≤ 0.01; Appendix Table M.1).

The definition is measurement-internal: it requires only a trusted reference and a shared sample, not knowing *why* a campaign deviates; Section 4.3 shows 20 documents suffice.

4.3 REFERENCE-ANCHOR PROTOCOL

To classify a new server session *before* trusting its output, we use a fixed set of anchor documents $\mathcal{D}_{\mathrm{anchor}}$ whose A-mode scores are known from production. Writing $\mathbf{a} = f_A(\mathcal{D}_{\mathrm{anchor}})$ for the reference anchor scores and $\mathbf{b}^{(g)} = g(\mathcal{D}_{\mathrm{anchor}})$ for the campaign's, we define the anchor statistic

$$\rho_{\mathrm{anchor}}(g) = \mathrm{Spearman}\big(\mathbf{b}^{(g)},\; \mathbf{a}\big), \tag{8}$$

and the decision rule: reject the session (declare mode B) unless

$$\rho_{\mathrm{anchor}}(g) > \tau_{\mathrm{anchor}}, \qquad \tau_{\mathrm{anchor}} = 0.5. \tag{9}$$

On our data this rule separates modes perfectly: all A-campaigns score $\rho_{\mathrm{anchor}} \geq 0.92$ and all B-campaigns ≤ 0.08 (Appendix Table M.4), at a cost of 20 judgments. Because the anchor uses a *fixed* document set and reference, it detects mode switches a fresh random sample cannot: any measurement globally coherent but re-ordered relative to A has low anchor agreement.

4.4 Mitigation Policy: Multi-Session Consensus

The safety policy we validate is multi-session consensus: run R independent server sessions—fresh serving processes over the same weights—over the target documents and take the per-document majority vote of the ordinal scores before computing any downstream statistic. Formally, for document i and dimension k, the consensus score is

$$\hat{s}_{ik}^{(R)} \;=\; \mathrm{mode}\big\{S_{ik}^{(g_1)}, \ldots, S_{ik}^{(g_R)}\big\}, \tag{10}$$

where mode is the most frequent ordinal value (ties broken by the lower value). The consensus matrix $\hat{\mathbf{S}}^{(R)}$ replaces any single $\mathbf{S}^{(g)}$ in Eq. (13).

We analyze consensus under a simple model. Suppose each session is in mode A with probability p_A and mode B otherwise, independently; in mode A a session reports the reference value and in mode B an uncorrelated one. If R_A of the R sessions are in mode A, the majority vote recovers the A-value whenever $R_A > R/2$. By Hoeffding's bound,

$$\mathbb{P}\big[R_A \le R/2\big] \;\le\; \exp\Big(-2R\big(p_A - \tfrac{1}{2}\big)^2\Big), \tag{11}$$

the probability that a majority is wrong decays exponentially in R for any fixed $p_A > 0.5$ (i.e., whenever B is the minority); at $p_A = 0.5$ the bound is vacuous and the majority is wrong with probability $\frac{1}{2}$ for every R, so consensus can contain a rare B but not one as frequent as A; the simulation of Figure 2 shows the same threshold empirically. In practice the B-mode rate is far smaller ($4/16$ campaigns), so consensus requires few sessions (Section 5.5).

Scope. Both protocols are *mechanism-agnostic*: they require only a fixed reference and, for consensus, that the anomalous mode be the minority. The anchor addresses axis-1 failures—any session that deviates from the reference, whatever the cause—while consensus removes a minority mode's influence without classifying sessions. Neither addresses *calibration drift* (axis 2), which is a cross-model question and needs a cross-model reference (Appendix Section M.8), and consensus offers no guarantee when the anomalous mode is as frequent as the valid one. Isolating the trigger remains open (Appendix F).

5 Experiments

We first set out the downstream analysis that consumes the scores (§5.1), then ask whether standard practice would have caught the problem and present the discovery of the two modes (§5.2), evidence that they are deterministic global states and ablations that locate the instability in the realized scoring function (§5.3), the downstream sign-reversal hazard (§5.4), the validated mitigation (§5.5), a cross-model validation and frontier stress test (§5.6), and close with the boundaries of what our data support. The negative trigger search, the re-serving localization, and the full per-run tables, cross-model audit, spectrum, and validity checks are in the appendix.

5.1 Experimental Setup

We use a widely deployed open-weight judge (Qwen2.5-14B-Instruct) served via vLLM with greedy decoding (temperature 0) and a fixed system prompt instructing the model to score a judgment document on six dimensions using a 1–5 ordinal scale (Appendix Section A): *factual specificity*, *narrative coherence*, *legal reasoning density*, *template likeness*, *detail personalization*, and *procedural completeness*. Within-session retests agree at exact rates above 0.95, so we treat Eq. (1) as a deterministic function of the session state. Scores are parsed from the model's JSON output; only well-formed responses are retained. We reproduce the two-axis results across 8–11 open-weight judges spanning 2024–2025 generations, including the newer-generation Qwen3-14B (Section 5.6 and Appendix M.14), so the findings are not an artifact of a single checkpoint.

The corpus comprises 188,416 Chinese criminal court judgments from the CAIL benchmark (Xiao et al., 2018), spanning 2008–2017 and covering 36 accusations. Each judgment is scored once in the production measurement, producing the reference matrix $\mathbf{S}^{(A)}$; all measurement runs below are *reruns* of subsets of this corpus using the identical prompt, model, and decoding settings.

The scores are consumed by a difference-in-differences (DID) analysis of China's 2014 sentencing-guideline reform (Appendix B). For dimension k, the DID coefficient compares treated-minus-control changes across the pre/post threshold $\tau = 2014$:

$$\widehat{\beta}_k^{DID} \;=\; \big(\bar{s}_k^{T,\text{post}} - \bar{s}_k^{T,\text{pre}}\big) \;-\; \big(\bar{s}_k^{C,\text{post}} - \bar{s}_k^{C,\text{pre}}\big), \tag{12}$$

under standard identification assumptions (Goodman-Bacon, 2021; Callaway & Sant'Anna, 2021; Sun & Abraham, 2021; Borusyak et al., 2024; de Chaisemartin & D'Haultfœuille, 2020), which we inherit and hold fixed while varying only the measurement. Our concern is that the scores entering Eq. (12) are themselves a measured quantity: if sessions realize different functions, the same documents yield different estimates,

$$\widehat{\beta}_k^{DID}(g) \;=\; \text{DID}\big(\mathbf{S}_{*k}^{(g)}\big), \tag{13}$$

and Section 5.4 quantifies this dependence.

5.2 Standard Checks Miss Two Modes on Identical Inputs

Before searching for causes, we ask whether standard practice would have caught the problem. The dominant protocol is a *single-session retest*: rerun the judge on a sample within the served context and report agreement. In every standard form—a 200-document within-session rerun (per-dimension exact 0.94–0.98), a 2,000-document same-session rerun ($\bar{\rho} = 0.961$), a cross-session same-mode rerun ($\bar{\rho} = 0.962$; Appendix Tables M.1 and M.2)—it looks healthy; only testing *both* a mode-A and a mode-B session reveals the failure, the same documents then agreeing at $\bar{\rho} \approx 0$ (Appendix Table M.2).

Two facts explain why the failure survives every standard check, and both are structural rather than incidental to our setup.

First, a within-session check is a check of one function. Every campaign in session s realizes the same f_s (Eq. (1)); if f_s is deterministic on the sampled documents, two campaigns in that session return identical scores and any agreement statistic between them attains its maximum, whatever function the session realizes. The statistic carries no information about what another session would return (Appendix H); reliability therefore has to be assessed *across* sessions.

Second, the change is invisible to the diagnostics that accompany a mean-level analysis. If $\mathbf{S}^{(B)}$ attaches to document i the scores that $\mathbf{S}^{(A)}$ attached to document $\sigma(i)$,

$$\mathbf{S}_{i\cdot}^{(B)} = \mathbf{S}_{\sigma(i)\cdot}^{(A)} \implies \bar{s}_k^{(B)} = \bar{s}_k^{(A)} \quad \text{for every } k, \tag{14}$$

then every column keeps the same multiset of values: means, the score distribution, and paired statistics computed from the columns—including the Wilcoxon signed-rank test—are unchanged, so a check on whether the score distribution has moved passes. An aggregate contrast such as Eq. (12) is not invariant, but it fails *silently*: with the partition held fixed the cell means are recomputed over a different assignment of values, and nothing in the contrast separates a change in the measurement from a change in the world. Only statistics that track which document received which score respond: row-wise rank agreement and ICC(1,1) collapse to ≈ 0 where means and paired tests do not move (Appendix I).

Taken together, the two facts say that reliability must be assessed across sessions and with rank-sensitive statistics, which is what the protocols of Section 4 do.

We ran 16 measurement campaigns across two servers over the same model weights and prompt. Appendix Table M.1 classifies each by Eq. (7): twelve campaigns reproduce mode A (Spearman 0.94–0.99, mean per-dimension exact 0.94–0.99), while four, on both servers, produced scores essentially orthogonal to production ($\bar{\rho} \approx 0$; mean per-dimension exact 0.37–0.41). Within a campaign, outputs were internally stable: a sequential retest of 200 documents in mode A reproduced production at per-dimension exact rate 0.967.

The two modes are *near-orthogonal functions* on the same inputs ($\text{corr}(\mathbf{s}_{*k}^{(A)}, \mathbf{s}_{*k}^{(B)}) \approx 0$ for all k; Appendix Eq. G.3): temperature 0 leaves no sampling variance and within-mode reruns are near-deterministic ($\bar{\rho} \geq 0.94$), so the difference is not attributable to sampling noise.

Variant	$\bar{\rho}$	$\bar{\rho}_{A_1B_1}$	Exact	$\hat{\pi}$
base (control re-run)	1.000	1.000	1.000	—
short	0.569	0.832	0.033	0.088
long	0.658	0.902	0.250	0.050
trunc600	0.942	0.976	0.950	0.012
trunc300	0.596	0.911	0.633	0.046
dim3 (3 dims)	0.523	0.976	0.367	0.012

Table 1: Prompt and length ablations move reliability along the permutation curve: 60 documents, all variants scored in one session against the production prompt. $\bar{\rho}$ is the column-wise rank correlation averaged over dimensions, $\bar{\rho}_{A_1B_1}$ the per-document row Spearman, and $\hat{\pi} = 1 - \sqrt{\bar{\rho}_{A_1B_1}/\bar{\rho}_{\text{base}}}$; DIM3 reports the three shared dimensions. The mode-B flip sits at $\hat{\pi} \approx 0.224$.

5.3 The Modes Are Deterministic Global States, Not Noise

We first rule out instrumentation artifacts: the modes appeared on *both* servers, weeks apart, under the same vLLM build, driver, CUDA runtime, checkpoints, and prompt, coexisting with A-mode measurements in interleaved sessions (Appendix J). Two properties separate the modes from sampling noise. Mode membership is *global* to a session: every document in a B-mode campaign deviates from A coherently, and two independent B-mode campaigns on different servers agree almost perfectly ($\bar{\rho} = 0.966$, exact 0.965; Appendix Table M.2). And B is not a monotone transformation of A: the conditional mean $\mathbb{E}[B \mid A]$ is flat (correlation ≈ 0) and the two modes share the same output template library (Jaccard 0.53; JS divergence at most 0.086). Together these indicate that B re-permutes scores across documents rather than distorting them, deterministically given the session: the environment state $\mathcal{E}_s$ acts as an implicit latent state selecting among a small set of functions, so the failure characterized in Section 5.2 is real rather than hypothetical. The amplification model (Appendix J) and the ablations of Section 5.3 induce a weaker re-permutation with the same signature under the experimenter's control; under the current stack the same weights score deterministically across cold restarts (Appendix M.15). The trigger itself remains untraced. A cold restart is deterministic on the matched stack, yet the same build, driver, CUDA runtime, checkpoints, and prompt produced B on two independent servers weeks apart, so the trigger is not explained by any artifact we matched. Both protocols are designed to be independent of that kernel, since each treats any non-A session as untrusted; a targeted stress of the remaining candidates (e.g., session-scoped page-layout or scheduling state in the serving layer) is a clear direction for future work.

The exact-agreement signature. Under a row-wise re-permutation of the reference, expected exact agreement equals the mean per-dimension marginal self-collision probability of the score distribution (Appendix G),

$$\mathbb{E}_\sigma\big[\bar{e}(g,h)\big] \;=\; \frac{1}{6}\sum_{k=1}^{6}\sum_{v} p_k(v)^2, \tag{15}$$

where p_k is the marginal of dimension k. The observed exact agreement of 0.41 (Appendix Table M.2) matches this floor (≈ 0.41) rather than contradicting the near-zero rank correlation—a monotone distortion would sit near 1—and the floor is substantial for template-heavy distributions such as ours, where the joint score-vector collision probability is only ≈ 0.02. The joint signature (moderate exact agreement, zero rank correlation) is what identifies the change as a re-permutation rather than a rescaling (Appendix J). Eq. (15) also motivates the permutation-strength estimator $\hat{\pi}$ of Section 4.1 (≈ 0.224 on Qwen2.5, ≈ 0.01 on the stable judges), which gives the ablations of Section 5.3 a scalar target.

Ablations: what the scoring function depends on. If reliability failures are properties of the scoring function a session realizes, a deterministic change to that function—a rewritten prompt, a shorter window, a reduced dimension set—should move the observed agreement along the same permutation curve, and $\hat{\pi}$ should track it. We test this on 60 documents scored in a single session, comparing the verbatim prompt against five variants with a re-run of the production prompt as control, which isolates the manipulation from the session state.

The control re-run is exact ($\bar{\rho}_{A_1B_1} = 1.000$), so the movement in the variants is attributable to the manipulation. Surface-form edits move the agreement point mainly on the *exact* axis while preserving

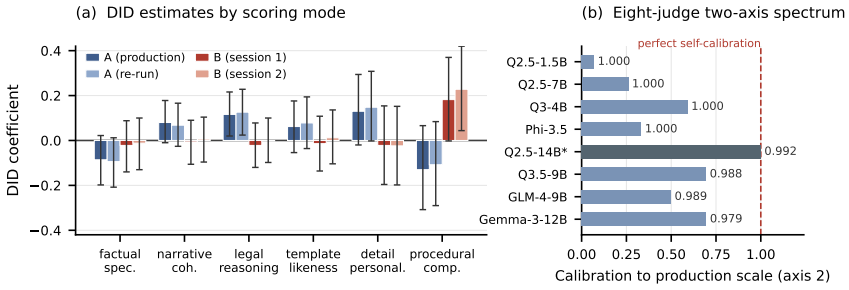

Figure 1: (a) DID coefficient estimates on identical documents under mode-A (blue) and mode-B (red) scoring, with document-level bootstrap 95% intervals (Appendix M.16): the two A runs reproduce each other, while each B session moves most dimensions and, on procedural completeness, flips sign and significance. (b) The eight-judge spectrum: bars are axis-2 calibration to the production scale, numbers at bar ends the axis-1 cross-session agreement (Appendix M.11). The axes are orthogonal: every judge is stable, none but the reference is calibrated; * marks the production reference, historically bistable (cross-mode agreement 0.596, Appendix Table M.5).

most of the row ordering (SHORT: exact 0.033, column rank 0.569; LONG: 0.250 and 0.658), i.e. $\hat{\pi} = 0.088$ and 0.050—an order of magnitude below the mode-B flip (0.224), but far from negligible and, unlike an undocumented session flip, under the experimenter's control. Truncation is length-thresholded (TRUNC600: 0.950; TRUNC300: 0.633), and narrowing to the three reasoning-adjacent dimensions still leaves exact agreement at 0.367, so the instability is not an artifact of the template or procedure dimensions. Prompt, window, and dimension set are thus exchangeable sources of the same failure family: reliability is a property of the realized scoring function, audited per prompt, per window, and per dimension set rather than once per model (Appendix M.4).

5.4 Downstream Consequence: DID Sign Reversal

To quantify the downstream harm of a mode switch, we run the DID estimator of Eq. (12) on identical documents under A- and B-mode scores, holding the identification assumptions and partition fixed and varying only the measurement; these numbers quantify the measurement-induced hazard, not the real-world effect of the reform (Appendix M.7). Figure 1 shows the coefficients with bootstrap intervals of the same order ($\approx \pm 0.11$) as the A–B gap: what a mode switch reverses is the pattern a single un-audited campaign would report, including which dimensions reach significance. Mode A estimates reproduce across independent A-measurements; every dimension's estimate under mode B differs substantially, reversing sign on 5 of 6 dimensions (det_retest) and 3 of 6 (aggregate_retest).

We define the *sign-flip indicator* $\mathbb{1}\left[\operatorname{sign}\widehat{\beta}_k^{DID}(B) \neq \operatorname{sign}\widehat{\beta}_k^{DID}(A)\right]$ and the *mean absolute bias* $\mathrm{MAE}(\beta^{(B)};\beta^{(A)}) = \frac{1}{6}\sum_k \left|\widehat{\beta}_k^{(B)} - \widehat{\beta}_k^{(A)}\right|$ (Appendix Eqs. G.9–G.10). The worst-case sign-flip rate is $5/6$ and the MAE is 0.138 for both B sessions against the production estimates: a researcher who measured in B would publish the opposite conclusion on five of six dimensions in the worst observed case. The re-permutation (Appendix Eq. (G.7)) changes *which* documents fall in each treated/control cell—precisely the quantity a DID contrast depends on.

5.5 Mitigation: Multi-Session Consensus Recovers the Truth

Figure 2 evaluates the consensus policy of Section 4.4. Consensus recovers the A-mode estimates whenever B is the minority: with a single session, a B-mode draw yields mean bias 0.146 and a worst-case $5/6$ sign-flip rate, while majority voting over $R = 5$ independent sessions drives sign flips to zero and mean bias to 0.014 when B is rare ($p_B = 0.1$), and at the historical arrival rate ($p_B \approx 0.25$) it reduces sign flips roughly three-fold ($0.19 \to 0.06$ at $R = 5$) without eliminating them. The boundary is sharp: when B is as frequent as A ($p_B = 0.5$), the majority is itself in B about half the time and the sign-flip rate stays at 0.34–0.46 regardless of R (Figure 2, right); our data does not occupy this regime, and the anchor test monitors it. The reference-anchor protocol classifies all 16 campaigns correctly (A: anchor ρ 0.92–1.00; B: ρ −0.04–0.08; threshold $\rho > 0.5$; Appendix Table M.4).

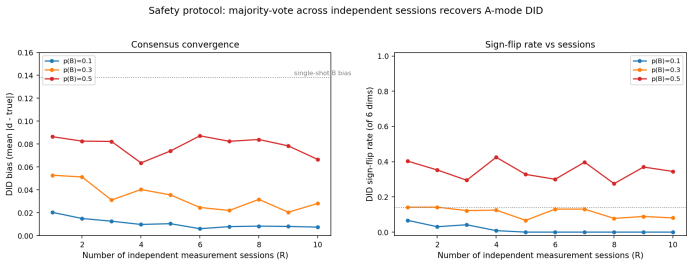

Figure 2: Consensus policy evaluation: mean absolute DID bias (left) and sign-flip rate over the six dimensions (right) against the number of sessions R. Voting recovers the A-mode estimates when B is the minority and, at the historical arrival rate ($p_B \approx 0.25$), reduces sign flips roughly three-fold; when B is as frequent as A ($p_B = 0.5$) the majority is itself in B about half the time and no number of sessions recovers them. Each session is assigned an actual measurement matrix from the four real campaigns, so the B scores the vote must tolerate are genuine re-permutations.

The policy does not classify sessions; it aggregates them. The trigger search is documented in Appendix M.7.

5.6 The Two Axes Across Judges

Figure 1(b) bears out the two headline claims: every non-production judge calibrates to the production scale at only $\rho = 0.07$–0.70 (axis 2) yet reproduces its scores nearly bit-for-bit across sessions (axis 1, $\bar{\rho}_{A_1 B_1} = 0.979$–$1.000$), and after per-dimension quantile alignment document rankings still correlate with the reference at only 0.31–0.67 (Appendix M.9). The drift is predictable out of sample ($r = -0.90$, $p = 0.002$, $N = 8$; Appendix M), the stability bound is tight (no B-mode across 63 fresh sessions; a 95% bound excludes a session B-rate ≥ 0.17; Appendix M.12), extending to $N = 10$ judges preserves both ($\text{pearson}(|\Delta\bar{\mathbf{s}}|, \text{calib}) = -0.76$, $p = 0.010$; Appendix M.14).

A newer-generation judge. To test whether the gap closes at the frontier, we added Qwen3-14B, a newer-generation 14B judge. It lies on the same drift–calibration spectrum (calib $= 0.622$; with it $N = 11$, $\text{pearson}(|\Delta\bar{\mathbf{s}}|, \text{calib}) = -0.751$, $p = 0.0077$), and the shortfall is not a quantization artifact (BF16 scoring yields the identical calibration), so the gap is a genuine generation/architecture effect. Closed/API judges were not tested in this study, so the cross-model claim is bounded to open-weight judges (Appendix M.14).

Limitations and open questions. The B-mode trigger remains open: it appeared in 4 of 16 campaigns, 8 controlled conditions failed to reproduce it, and it did not recur in 63 controlled follow-up sessions (Appendix M.12). The phenomenon is documented on one model, stack, prompt, and domain, whereas the two-axis framework and pervasive axis-2 drift do generalize across the 11 judges we scored. One boundary deserves emphasis: the audit is *measurement-internal*, taking the production A-mode scores as the reference without validating their external validity against independent labels, so it certifies consistency with the production scale, not the scale's correctness; validating the reference is a prerequisite before the protocol arbitrates claims in a new domain. The rarity of B leaves the trigger unidentified, and we release the full trigger-search log to support independent reproduction. The DID numbers quantify a measurement hazard on 2,000 documents, not corpus-wide estimates; re-validation should target other high-stakes domains (e.g., clinical or financial text), other serving stacks (e.g., HuggingFace TGI or native PyTorch), and whether the mode depends on quantization or batch size.

6 Conclusion

LLM-as-a-judge scores are measurement. Calibration fixes scale, not ranking (reference rankings correlate at only 0.31–0.67 after alignment), and standard validation does not detect a deterministic re-permutation that can reverse up to $5/6$ downstream estimates. Reliability is a property of a *measurement campaign* audited on both axes.

AI USE STATEMENT

Large language models (LLMs) were employed exclusively to improve the quality of written English and for stylistic refinement, with attentive supervision by the authors. No scientific content, experimental design, data analysis, or any intellectual contribution were generated through LLMs, nor did such assistance influence any reported results.

ETHICS STATEMENT

All authors have read and adhere to the ICLR Code of Ethics. This study measures the reliability of an LLM judge over CAIL, a public benchmark of Chinese criminal court judgments; it involves no human subjects, we collect no new data, and we release no individual records beyond the public corpus. The contribution is defensive: we document a measurement hazard and provide mechanism-agnostic auditing protocols, a fixed reference anchor and multi-session consensus, that detect and contain a scoring mode which would otherwise silently reverse a downstream conclusion. We flag the domain sensitivity of legal text and stress that the reported difference-in-differences numbers quantify a measurement-induced hazard on a fixed sample, not the real-world effect of any policy or reform.

REPRODUCIBILITY STATEMENT

We release the measurement and analysis scripts, the scoring prompt, and the state-classification protocol with the artifact. The measurement setting and agreement statistics are defined in Section 3; the audit metrics and the two defensive protocols in Section 4; and the experimental setup and results in Section 5. The scoring prompt and per-run configuration (Appendix A, E), the difference-in-differences setup (B), the per-run state classification and cross-model audits (C, M), the consensus simulation (D), and the supporting derivations and notation (G, H) are documented in the appendix. Reported agreement statistics include seeds and bootstrap confidence intervals.

REFERENCES

Anya Belz, Craig Thomson, Ehud Reiter, and Simon Mille. Non-repeatable experiments and non-reproducible results: The reproducibility crisis in human evaluation in NLP. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 3676–3687. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-ACL.226.

Kirill Borusyak, Xavier Jaravel, and Jann Spiess. Revisiting event-study designs: Robust and efficient estimation. *The Review of Economic Studies*, 91(6):3253–3285, 2024.

Brantly Callaway and Pedro H. C. Sant'Anna. Difference-in-differences with multiple time periods. *Journal of Econometrics*, 225(2):200–230, 2021.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. ChatEval: Towards better LLM-based evaluators through multi-agent debate. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot Arena: An open platform for evaluating LLMs by human preference. In *Forty-first International Conference on Machine Learning (ICML)*, pp. 8359–8388, 2024.

Clément de Chaisemartin and Xavier D'Haultfœuille. Two-way fixed effects estimators with heterogeneous treatment effects. *American Economic Review*, 110(9):2964–2996, 2020.

Sunishchal Dev, Andrew Sloan, Joshua Kavner, Nicholas Kong, and Morgan Sandler. Judge reliability harness: Stress testing the reliability of LLM judges. In *Agents in the Wild: Safety, Security, and Beyond Workshop at ICLR 2026*, 2026. arXiv:2603.05399.

Matthew Gentzkow, Bryan Kelly, and Matt Taddy. Text as data. *Journal of Economic Literature*, 57 (3):535–574, 2019.

Andrew Goodman-Bacon. Difference-in-differences with variation in treatment timing. *Journal of Econometrics*, 225(2):254–277, 2021.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *8th International Conference on Learning Representations (ICLR 2020)*. OpenReview.net, 2020.

Xinyu Hu, Mingqi Gao, Sen Hu, Yang Zhang, Yicheng Chen, Teng Xu, and Xiaojun Wan. Are LLM-based evaluators confusing NLG quality criteria? In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 9530–9570. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.516.

Wanhong Huang, Yi Feng, Chuanyi Li, Honghan Wu, Jidong Ge, and Vincent Ng. CMDL: A large-scale Chinese multi-defendant legal judgment prediction dataset. In *Findings of the Association for Computational Linguistics, ACL 2024*, pp. 5895–5906. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.351.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024*, pp. 4334–4353. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.EMNLP-MAIN.248.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with PagedAttention. In *Proceedings of the 29th Symposium on Operating Systems Principles (SOSP 2023)*, pp. 611–626. ACM, 2023. doi: 10.1145/3600006.3613165.

Nathan Lambert, Valentina Pyatkin, Jacob Morrison, and et al. RewardBench: Evaluating reward models for language modeling. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 1755–1797. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-NAACL.96.

Dmitry Lepikhin, HyoukJoong Lee, Yuanzhong Xu, Dehao Chen, Orhan Firat, Yanping Huang, Maxim Krikun, Noam Shazeer, and Zhifeng Chen. GShard: Scaling giant models with conditional computation and automatic sharding. In *9th International Conference on Learning Representations, ICLR 2021*. OpenReview.net, 2021.

Lujun Li, Lama Sleem, Yangjie Xu, Yewei Song, Aolin Jia, Jerome Francois, and Radu State. The necessity of setting temperature in LLM-as-a-Judge. *arXiv preprint arXiv:2603.28304*, 2026.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-Eval: NLG evaluation using GPT-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153.

Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in Neural Information Processing Systems 27 (NIPS 2013)*, pp. 1196–1204, 2013.

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.

Yonatan Oren, Nicole Meister, Niladri S. Chatterji, Faisal Ladhak, and Tatsunori Hashimoto. Proving test set contamination in black-box language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net, 2024.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2024.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2023.

Oscar Sainz, Jon Campos, Iker García-Ferrero, and et al. NLP evaluation in trouble: On the need to measure LLM data contamination for each benchmark. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 10776–10787. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.FINDINGS-EMNLP.722.

Patrick Schilcher, Dominik Karasin, Michael Schöpf, Haisam Saleh, Antonela Tommasel, and Markus Schedl. Characterizing positional bias in large language models: A multi-model evaluation of prompt order effects. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20643–20664. Association for Computational Linguistics, 2025. doi: 10.18653/V1/2025.FINDINGS-EMNLP.1124.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V. Le, Geoffrey E. Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *5th International Conference on Learning Representations, ICLR 2017*. OpenReview.net, 2017.

Charlotte Siska, Katerina Marazopoulou, Melissa Ailem, and James Bonó. Examining the robustness of LLM evaluation to the distributional assumptions of benchmarks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024*, pp. 10406–10421. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.ACL-LONG.560.

Liyang Sun and Sarah Abraham. Estimating dynamic treatment effects in event studies with heterogeneous treatment effects. *Journal of Econometrics*, 225(2):175–199, 2021.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chenguang Wang, Raluca A. Popa, and Ion Stoica. JudgeBench: A benchmark for evaluating LLM-based judges. In *The Thirteenth International Conference on Learning Representations, ICLR 2025*. OpenReview.net, 2025.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations (ICLR 2023)*. OpenReview.net, 2023.

Yanzheng Xiang, Hanqi Yan, Lin Gui, and Yulan He. Addressing order sensitivity of in-context demonstration examples in causal language models. In *Findings of the Association for Computational Linguistics, ACL 2024*, pp. 6467–6481. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.386.

Chaojun Xiao, Haoxi Zhong, Zhipeng Guo, Cunchao Tu, Zhiyuan Liu, Maosong Sun, Yansong Feng, Xianpei Han, Zhen Hu, Heng Wang, and Jianfeng Xu. CAIL2018: A large-scale legal dataset for judgment prediction. *CoRR*, abs/1807.02478, 2018. URL `http://arxiv.org/abs/1807.02478`. Data resource.

Chaojun Xiao, Xueyu Hu, Zhiyuan Liu, Cunchao Tu, and Maosong Sun. Lawformer: A pre-trained language model for Chinese legal long documents. *AI Open*, 2:79–84, 2021. doi: 10.1016/J.AIOPEN.2021.06.003.

Zixiang Xu, Sixian Li, Huaxing Liu, Xiang Wang, Shuai Li, Zirui Song, and Xiuying Chen. Inside the unfair judge: A mechanistic interpretability account of LLM-as-Judge bias. *arXiv preprint arXiv:2607.11871*, 2026.

Abel Yagubyan. The coin flip judge? reliability and bias in LLM-as-a-Judge evaluation. *arXiv preprint arXiv:2606.13685*, 2026.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2023.

Lianmin Zheng, Liangsheng Yin, Zhiqiang Xie, Chuyue Sun, Jeff Huang, Cody Hao Yu, Shiyi Cao, Christos Kozyrakis, Ion Stoica, Joseph E. Gonzalez, Clark W. Barrett, and Ying Sheng. SGLang: Efficient execution of structured language model programs. In *Advances in Neural Information Processing Systems 38 (NeurIPS)*, 2024.

CONTENTS

A Prompt **14**

B Difference-in-Differences Setup **14**

C Full Per-Run State Classification **14**

D Consensus Simulation Details **14**

E Compute **14**

F Future Work **15**

G Supporting Definitions and Notation **15**

H Derivations for Section 5.2 **16**

I Simulation Verification of the Blindness Boundary **16**

J Candidate Mechanism and Exact-Agreement Detail **16**

K Copula Characterization of Re-Permutation **17**

L Heavy-Tailed Score Distributions **18**

M Supplementary Experiments **18**

M.1 State Classification 18

M.2 Within-Mode and Cross-Mode Agreement 18

M.3 DID Estimates by Mode 18

M.4 Prompt and Length Ablations: Design Details 19

M.5 Consensus Policy Evaluation: Simulation Setup 19

M.6 Reference-Anchor Classification 20

M.7 Trigger Search: Controlled Conditions 20

M.8 Across Models and Stacks 21

M.9 Scale Drift Is Predictable and Correctable 22

M.10 The Two Axes Transfer Across Tasks 22

M.11 Spectrum Table 23

M.12 Stability Bound and Model Vintage 24

M.13 Cross-Judge Validity Sample 24

M.14 Frontier Stress Test Details 24

M.15 Re-Serving Under the Current Stack 25

M.16 Bootstrap Uncertainty for the DID Hazard Estimates 25

N Practical Checklist for LLM-as-a-Judge Deployments **26**

A PROMPT

The scoring prompt instructs the judge to read a court judgment and rate it on each of the six dimensions on a 1–5 scale, returning a JSON object with integer keys. The full Chinese prompt is shown below; the parsing code is released with the artifact. Each document was truncated to the first 1,200 characters for scoring.

> 你是一名司法文本学研究助理。给定一份中国刑事案件的案情描述（公诉机关指控部分），请对文本进行客观测量，输出 JSON。测量维度（全部按 1-5 整数打分，5 表示该特征极强）：1. factual_specificity 事实具体性：时间地点金额数量等具体细节丰富程度 2. narrative_coherence 叙事连贯性：事件因果链是否清晰连贯 3. legal_reasoning_density 法律论证密度：法律条款/情节认定/量刑考量等法律论证要素多寡 4. template_likeness 模板化程度：文本是否高度模板化、套话连篇（高分=模板化严重） 5. detail_personalization 情节个性化：案件特有细节独特性（与模板化相对） 6. procedural_completeness 程序要素完整度：到案、供述、辩护、证据、谅解等要素是否完整 输出严格 JSON，不要任何多余文字：”factual_specificity”:1-5,”narrative_coherence”:1-5,”legal_reasoning_density”:1-5,”template_likeness”:1-5,”detail_personalization”:1-5,”procedural_completeness”:1-5

The user message appends the judgment text truncated to the first 1,200 characters: “案情描述：” followed by the fact text. Decoding used greedy sampling (temperature 0, up to 256 tokens) through the same serving endpoint as production. The system prompt above is the exact string used in every production and retest run (see released scripts).

B DIFFERENCE-IN-DIFFERENCES SETUP

For a document i with accusation a_i, group indicator $T_i \in \{0, 1\}$ (treated vs. control accusation), and year y_i, the DID estimator for dimension k is

$$\text{DID}_k = \left(\bar{s}_k^{T,\text{post}} - \bar{s}_k^{T,\text{pre}}\right) - \left(\bar{s}_k^{C,\text{post}} - \bar{s}_k^{C,\text{pre}}\right), \tag{B.1}$$

where “post” means $y_i \geq 2014$ (China’s sentencing-guideline reform). The hazard analysis in §5.4 uses the 2,000 documents scored under both modes. The companion analysis uses the full 188,416-document production measurement and adds accusation and year fixed effects plus controls; the DID identification assumptions follow standard panel-data practice (Goodman-Bacon, 2021; Callaway & Sant’Anna, 2021; Sun & Abraham, 2021; Borusyak et al., 2024; de Chaisemartin & D’Haultfœuille, 2020).

C FULL PER-RUN STATE CLASSIFICATION

Table M.1 summarizes the campaigns. The two additional B-mode datasets (sweep-era, reliability-era) appear with their timestamps: reliability.jsonl (07:58), aggregate_retest (08:31), det_retest (12:59), and sweep.jsonl (13:26), across the old and new servers. The observation times are scattered, and no pattern (time of day, server, load) predicts mode.

D CONSENSUS SIMULATION DETAILS

The consensus simulation draws, for each session, an actual measurement from one of the four real campaigns (two A, two B) with per-session mode probability p_B, then takes the per-document majority vote of the six ordinal scores across sessions (Eq. (10)), and recomputes the six DID estimates (Eq. (G.8)). Reported bias is mean absolute difference from the true A-mode estimate (Eq. (G.10)); sign-flip rate counts dimensions whose sign disagrees (Eq. (G.9)). Each configuration is averaged over 60 independent draws; error bars are small and omitted for legibility.

E COMPUTE

All serving used one of two vLLM instances (CUDA 13, 14B Qwen) on Ubuntu 22.04 hosts with GPUs. The production measurement scored 188,416 documents; rerun campaigns scored subsets of 30–2,000 documents. Total measurement compute is approximately 30 GPU-hours across all reruns and probes.

F Future Work

Three directions stand out. First, a predictive model of the scale anchor from architectural or training properties, turning drift estimation from the observed mean-output heuristic into an a-priori bound. Second, a root-cause account of the session-state mechanism behind rank instability. Our controlled trigger search (Appendix M.7) rules out cold restarts, load, concurrency, prompt length, and within-session time; the remaining candidate space is at the serving/kernel level, and a systematic ablation would need to vary, one at a time, the inference scheduler, the attention-kernel implementation, the KV-cache initialization and eviction policy, and the batching order, under matched decoding settings, to isolate which layer of the stack can organize floating-point deviations into a coherent session-wide pattern (Kwon et al., 2023; Zheng et al., 2024; Lepikhin et al., 2021; Shazeer et al., 2017). We did not have the ability to reproduce the mode on demand, so such a search was not feasible in this study; we release the detection artifacts so a root-cause investigation can begin from reproducible reruns. Third, automated per-campaign mode detection that requires no reference measurement, e.g., by probing statistical properties of the score distribution itself.

G Supporting Definitions and Notation

Protocol metrics. The rank-stability metrics used by the protocols—the per-document row Spearman (Eq. (4)), its distribution across documents, and the per-dimension ICC(1,1) (Eq. (5))—are defined in Section 4.1, together with the permutation-strength estimator $\hat{\pi}$ (Eq. (6)). ICC(1,1) is the agreement metric most evaluation protocols implicitly rely on; it is not blind to re-permutation (per-document agreement collapses under the Qwen2.5 B-mode, ICC ≈ 0), but as a single number it hides the distribution of per-document agreement, which the row Spearman exposes and which we report as the primary rank-stability diagnostic.

Permutation strength. Modeling session B as the reference A on a per-document $1-\pi$ fraction and as an independent draw on the remaining π fraction, independently across the two sessions, the expected per-document agreement decomposes as

$$\mathbb{E}\big[\bar{\rho}_{A_1B_1}\big] \;=\; (1-\pi)^2\,\bar{\rho}_{A_1A_2} \;+\; 2\pi(1-\pi)\,\bar{\rho}_{\mathrm{ind}} \;+\; \pi^2\,\bar{\rho}_{\mathrm{ind}}, \tag{G.2}$$

where the document survives both sessions with probability $(1-\pi)^2$, survives exactly one with probability $2\pi(1-\pi)$, and is independently redrawn in both with probability π^2. Under the null that the independent part is rank-uncorrelated with A ($\bar{\rho}_{\mathrm{ind}} \approx 0$), the between-session agreement ratio $\hat{\pi} = 1 - \sqrt{\bar{\rho}_{A_1B_1}/\bar{\rho}_{A_1A_2}}$ of Eq. (6) (Section 4.1) is a *consistent estimator of the permutation strength* π, the fraction of documents whose scores are reassigned between sessions.

$$\mathrm{corr}\big(\mathbf{s}^{(A)}_{*k}, \mathbf{s}^{(B)}_{*k}\big) \approx 0 \qquad \forall k, \tag{G.3}$$

$$J(g,h) \;=\; \frac{|\mathcal{T}^{(g)} \cap \mathcal{T}^{(h)}|}{|\mathcal{T}^{(g)} \cup \mathcal{T}^{(h)}|}, \tag{G.4}$$

$$\mathrm{JSD}(P_g \| P_h) = \tfrac{1}{2}\,\mathrm{KL}(P_g \| M) + \tfrac{1}{2}\,\mathrm{KL}(P_h \| M), \tag{G.5}$$

$$\mathbb{E}\big[\, S^{(B)}_{ik} \mid S^{(A)}_{ik} = a \,\big] \quad \text{for } a \in \{1,\dots,5\}. \tag{G.6}$$

$$\mathbf{S}^{(B)}_{*k} \;\approx\; \sigma_k\big(\mathbf{S}^{(A)}_{*k}\big), \tag{G.7}$$

$$\widehat{\beta}^{DID}_k(\hat{\mathbf{S}}^{(R)}) \;=\; \mathrm{DID}\big(\hat{\mathbf{S}}^{(R)}_{*k}\big), \tag{G.8}$$

$$\nVdash\big[\mathrm{sign}\,\widehat{\beta}^{DID}_k(B) \neq \mathrm{sign}\,\widehat{\beta}^{DID}_k(A)\big], \tag{G.9}$$

$$\mathrm{MAE}\big(\beta^{(B)};\beta^{(A)}\big) = \frac{1}{6}\sum_{k=1}^{6}\big|\widehat{\beta}_k^{(B)} - \widehat{\beta}_k^{(A)}\big|. \tag{G.10}$$

H Derivations for Section 5.2

Within-session statistics are a functional of f_s. By construction, all campaigns in $\mathcal{G}_s$ are realizations of the same function f_s (Eq. (1)), because a session fixes the environment state $\mathcal{E}_s$. If f_s is deterministic on the sampled documents, two campaigns in the same session return identical scores, so $R = 1$ for both statistics. The statistic is a functional of f_s alone; it contains no information about functions realized by other sessions.

Column-wise statistics under a row permutation. A permutation of the rows of $\mathbf{S}^{(A)}$ leaves the multiset of values in each column, and hence every columnwise sample moment, invariant: the mean and the empirical distribution are unchanged, and any statistic that depends on the columns only through their empirical distributions—the per-dimension means, and the paired Wilcoxon signed-rank statistic in the simulation below—takes the same value under $\mathbf{S}^{(A)}$ and under $\mathbf{S}^{(B)}$. For DID, $\bar{s}_k^{T,\mathrm{post}}$ is the column mean over a cell of documents: if the treatment/time partition is permuted together with the rows the four-cell contrast is unchanged, whereas under the conventional, fixed treated/control partition the cell means are recomputed over a different assignment of values, giving the pattern observed in Section 5.4—identical marginals, perturbed contrast.

I Simulation Verification of the Blindness Boundary

We verify the invariance claims of Section 5.2 and their boundary with a reproducible simulation. Draw $\mathbf{S}^{(A)}$ as $N = 2{,}000$ independent documents with six ordinal dimensions $S_{ik} \in \{1,\dots,5\}$, and set $\mathbf{S}^{(B)}_{i\cdot} = \mathbf{S}^{(A)}_{\sigma(i)\cdot}$ for a random permutation σ. Under this re-permutation:

- Column means are exactly equal (maximum absolute difference 0), and the column-wise empirical distributions are identical.
- The paired Wilcoxon signed-rank test on per-dimension differences is significant on $0/6$ dimensions (all $p > 0.05$).
- The DID estimate (Eq. (12)) is *unchanged* when the treated/control and pre/post partition is permuted together with the rows (maximum absolute difference 0), and *changes* when the partition is held fixed (e.g., mean absolute change 0.230 over six dimensions on a random partition).
- Rank-sensitive statistics do *not* stay fixed: the mean per-document row Spearman and the column Spearman both collapse to ≈ 0, ICC(1,1) drops from 1.0 (self-agreement) to ≈ 0, and the exact-agreement rate drops to the self-collision level.

This confirms the precise boundary: the blindness is a property of mean-level and paired statistics, not of the toolkit as a whole. The released benchmark includes the simulation script.

J Candidate Mechanism and Exact-Agreement Detail

Instrumentation artifacts. We explicitly ruled out the common non-mechanism explanations for the two modes. The modes appeared on *both* servers, at different times of day over several weeks, under the same vLLM build, GPU driver, CUDA runtime, model checkpoints, and scoring prompt; a software-version or driver mismatch would produce a persistent, not intermittent, difference. The modes also coexisted with ordinary A-mode measurements in interleaved sessions on the same machine, so they are not a per-machine configuration effect. What rules out sampling noise is that the modes are deterministic given the session and globally coherent across all scored documents.

Candidate mechanism. We do not claim a root cause at the kernel level, but the structure constrains the hypothesis space. (i) *Global coherence*: B is a deterministic state of the session, reproducible

across two servers ($\bar{\rho} = 0.966$), so the mechanism acts on session-wide state, not on individual requests. (ii) *Template preservation*: B recycles the same output templates as A with a scrambled assignment, so the mechanism corrupts the *mapping from documents to templates* (e.g., the choice among semantically interchangeable template fills given near-tied logits), not the template library itself. (iii) *Orthogonality*: because $E[B \mid A]$ is flat, the corruption is not a monotone shift in scoring tendency. These three properties are jointly consistent with a serving-level source of batch- or session-dependent nondeterminism (e.g., in kernel reductions, scheduling, or KV-cache state) that, on near-tied decoding decisions, flips which template fill is greedily selected, and which then self-reinforces for the remainder of the session. We regard any single mechanism as speculative; the contribution of this section is the empirical characterization and the protocol that is robust to whichever mechanism is at play.

A concrete amplification model. The three properties can be organized by a minimal model of how a session-level perturbation becomes a *global* re-permutation. Decompose the realized function f_s into a per-document decision: at each decoding step the model's next-token distribution has a top-k set, and for template-heavy documents the logits separating the template fill that A chose from its nearest competitor are frequently *near-tied*: the margin is at the scale of floating-point noise rather than of semantics. A session-wide perturbation ϵ_s (e.g., a different reduction order in an attention or normalization kernel, a KV-cache state that varies with batching order) shifts these margins by an amount ϵ_s that is *shared across all documents of the session*. Documents whose A-mode margin falls below $|\epsilon_s|$ flip their greedy choice; the flipping threshold is set by the session, so the flipped set is a deterministic function of $\mathcal{E}_s$, which is why two independent B sessions agree at $\bar{\rho} = 0.966$ across servers. This is a channel from a tiny, per-session bias to a global, coherent reassignment: the near-tie makes the decision *amplification-prone* (a small perturbation is decisive), and the session-wide sharing of ϵ_s makes the flip *correlated across documents* rather than independent per-request noise. The model makes two falsifiable predictions we can check with the existing data. First, because the flipped set depends only on $\mathcal{E}_s$, a B session is reproducible: confirmed by the cross-server $\bar{\rho} = 0.966$. Second, the amplification should concentrate on documents whose A-mode margins are smallest, i.e., on the template-heavy, near-tied subset; consistent with this, the permutation strength $\hat{\pi} \approx 0.224$ (Eq. 6) is of the same order as the near-tied fraction implied by the exact-agreement floor of Eq. (15) (0.37–0.41). We stress that this is a *conceptual* model (we cannot directly observe kernel-level margins, and we did not reproduce B on demand), but it converts the previously verbal intuition into a mechanism with testable consequences and an explicit role for each of the three structural properties above.

Template identity and distribution shift. The two modes draw on the same multiset of output templates. Let $\mathcal{T}^{(g)}$ be the set of normalized templates produced by campaign g. We quantify template overlap by the Jaccard coefficient of Eq. (G.4) and the shift in template-frequency distributions by the Jensen–Shannon divergence of Eq. (G.5); we find $J(A, B) = 0.53$ and JSD ≤ 0.086: the surface forms are shared and their frequency distribution is nearly unchanged, yet the *assignment* of templates to documents is scrambled.

The exact-agreement signature. The reader may ask whether exact agreement 0.41 contradicts near-zero Spearman. It does not: under a row-wise re-permutation the expected exact agreement equals the mean per-dimension marginal collision probability of Eq. (15), while rank correlation collapses. We verify this identity directly: the per-dimension marginals are concentrated (mean per-dimension collision ≈ 0.41), so the expected exact rate is ≈ 0.41 by construction, matching the observed per-campaign exact rates (0.37–0.41) within sampling noise, while Spearman is ≈ 0 under the same permutation. A monotone distortion would instead produce exact agreement near 1; the contrast confirms the re-permutation reading and resolves the apparent tension between the two agreement statistics.

K Copula Characterization of Re-Permutation

The decomposition of Eq. (15) also admits a copula view that separates *re-permutation* from *distortion*. For two co-measured campaigns, the joint distribution of rank pairs is a copula $C(u, v)$ on $[0, 1]^6$. A monotone distortion h acts on the marginal scales and leaves the grade copula invariant (in the sense of Hoeffding), whereas a row-wise re-permutation replaces the induced copula by a mixture $C_\pi = (1-\pi)\, C_\parallel + \pi\, C_\perp$, where $C_\parallel$ is the coincident copula (the score vector of a document survives)

and $C_\perp$ the independence copula on the shuffled fraction. Two estimable consequences follow. First, the rank statistics of Section 3.2 behave as weighted sums of the extreme copulas:

$$\rho(C_\pi) \;=\; (1-\pi)^2\,\rho(C_\parallel) \;+\; \pi^2\,\rho(C_\perp) \tag{K.11}$$

for rank-based association, which is exactly Eq. (6) rewritten in copula terms. Second, the exact-agreement rate is the diagonal mass $\delta_\pi = \int \partial C_\pi/\partial u \; \mathrm{d}u$; because $C_\perp$ contributes vanishing diagonal mass, $\delta_\pi \approx (1-\pi)\,\delta_\parallel$. These two identities are jointly diagnostic: a re-permutation moves the observed point along the curve (δ_π, ρ_π) traced out by π, while a distortion (or independent resampling) moves it off the curve. This gives a falsifiable prediction: the observed (δ, ρ) should move along the mixture curve as π varies, a pattern that does not hold for a monotone distortion, whose point is pinned at $(\delta \approx 1, \rho \approx 1)$ regardless of π. We use the mixture curve to interpret the ablations of Section M.4, where prompt edits move the agreement point along the re-permutation family rather than toward the distortion corner.

L Heavy-Tailed Score Distributions

The exact-agreement floor of Eq. (15) sharpens under heavy tails. Legal judgments concentrate on a few template-heavy score vectors, so each dimension's marginal p_k puts most mass on a small number of ordinal levels: the per-dimension marginal collision $\sum_v p_k(v)^2$ ranges over 0.27–0.52 across the six dimensions (mean ≈ 0.41), far above the $1/5 = 0.2$ uniform baseline on the 1–5 scale, even though the joint score-vector collision $\sum_v p(v)^2$ is only ≈ 0.02. For such distributions the collision probability of Eq. (15) is controlled by the *tail* rather than the bulk: writing $p_k(v) \propto v^{-\alpha-1}$ for the rank-ordered probabilities of a dimension with tail index $\alpha > 0$, the sum $\sum_v p_k(v)^2$ behaves as the Sobolev norm of the marginal and grows monotonically with the peaking exponent α: for a Zipf-like law on five levels it is ≈ 0.34 at $\alpha = 0.3$ and ≈ 0.50 at $\alpha = 1.0$, bracketing the regime we observe. Two consequences follow. First, the re-permutation null of Eq. (15) predicts an exact-agreement floor that is *higher* for template-heavy distributions, making the failure of rank-based consistency *harder* to detect by exact agreement than by rank statistics. Second, the copula mixture of Eq. (K.11) remains rank-invariant under any monotone marginal transformation (Sklar's theorem), so the heavy-tail regime cannot be the *cause* of rank collapse; it only suppresses the exact-agreement signal that a careless reliability report might otherwise lean on. This separates the theory cleanly: the phenomenon is a property of the *rank* of $\mathcal{M}$'s output, and the heavy-tail structure of legal text merely hides it from mean- and exact-rate-based checks.

M Supplementary Experiments

M.1 State Classification

Table M.1 classifies each measurement campaign by its mean per-document Spearman against the production scores over the six dimensions. The twelve mode-A campaigns (six datasets plus six bootstrap probes) reproduce production near-perfectly ($\bar\rho = 0.94$–0.99), while the four mode-B campaigns are essentially orthogonal ($\bar\rho \approx 0$); both modes appear on both servers, so the mode assignment is not a per-machine configuration effect.

M.2 Within-Mode and Cross-Mode Agreement

Table M.2 reports the mean Spearman and the mean per-dimension exact agreement over the shared 2,000 documents. Within-mode pairs (the first three rows) agree near-perfectly in both rank ($\bar\rho \geq 0.96$) and exact terms (≥ 0.96), whereas cross-mode pairs (the last two rows) are near-orthogonal in rank ($\bar\rho \approx 0$) yet retain an exact-agreement rate of ≈ 0.41, precisely the re-permutation signature of Eq. (15).

M.3 DID Estimates by Mode

Figure M.1 and Table M.3 show the consequence of scoring mode on a difference-in-differences analysis of the identical 2,000 documents. The two A-mode measurements reproduce the production sign pattern on all six dimensions, while the two B-mode sessions reverse the sign on 5/6 (det) and 3/6

Dataset	State	$\bar{\rho}$	n
agg2000_now	A	$+0.96$	2000
agg2000_3rd	A	$+0.96$	2000
sample300_zh	A	$+0.96$	300
sweep60_zh	A	$+0.97$	60
seq_retest	A	$+0.96$	200
cold_probe_0	A	$+0.94$	100
boot_probe 1–6	A	0.96–0.99	30 each
det_retest	B	-0.00	2000
aggregate_retest	B	-0.00	2000
sweep (B-era)	B	$+0.00$	60
reliability (B-era)[†]	B	-0.01	500

[$\bar{\rho}$] mean Spearman vs. production over the six dimensions. The 6 boot probes are counted individually; the A rows total 12 campaigns (6 datasets plus 6 probes), the B rows 4.

[†] The reliability campaign combined a temperature-0.7 pass (400 documents) with a temperature-0 pass (100 documents); both are scored B. The other three B campaigns were served at temperature 0.

Table M.1: State classification of all measurement campaigns. Mode A reproduces the production scores; mode B is essentially orthogonal. Both modes occur on both servers.

Pair	$\bar{\rho}$	Exact
prod(A) $\sim$ agg2000(A)	$+0.961$	0.962
agg2000(A) $\sim$ agg2000(A,3rd)	$+0.974$	0.974
det(B) $\sim$ aggregate(B)	$+0.966$	0.965
prod(A) $\sim$ det(B)	-0.004	0.407
prod(A) $\sim$ aggregate(B)	-0.001	0.409

Mean Spearman and mean per-dimension exact agreement over the shared 2,000 documents.

Table M.2: Within-mode reproducibility and cross-mode orthogonality. Within-mode pairs agree near-perfectly; cross-mode pairs are near-orthogonal.

(agg) dimensions; the procedural-completeness dimension, for instance, flips from $-0.130/-0.108$ (the two A runs) to $+0.182/+0.228$ (the two B runs).

M.4 Prompt and Length Ablations: Design Details

Table 1 (Section 5.3) reports the full grid, run on a fixed set of 60 documents in a single session: the verbatim production prompt (BASE, which also serves as the control re-run), a drastically shortened instruction set (SHORT), an expanded instruction set with per-dimension worked examples (LONG), the same instruction applied to documents truncated to 600 and to 300 characters (TRUNC600, TRUNC300), and a three-dimension instruction that drops template-likeness, personalization, and procedure while retaining factual specificity, narrative coherence, and legal reasoning density (DIM3). Because all variants are scored in one session, the BASE control absorbs the session state and any remaining movement is attributable to the manipulation. Each row is converted to a permutation strength by $\hat{\pi} = 1 - \sqrt{\bar{\rho}_{A_1B_1}/\bar{\rho}_{\text{base}}}$ under Eq. (6); the control is exact by construction ($\hat{\pi} = 0$), the strongest manipulation reaches $\hat{\pi} = 0.088$ (SHORT), and the mode-B flip sits higher at $\hat{\pi} \approx 0.224$. Section 5.3 discusses the pattern; the ordering of the variants is stable across the exact and column-rank axes.

M.5 Consensus Policy Evaluation: Simulation Setup

Figure 2 (Section 5.5) evaluates the multi-session consensus policy (Eq. (G.8)) at several per-session B-mode rates. Each simulated session is assigned an actual measurement matrix drawn from the four

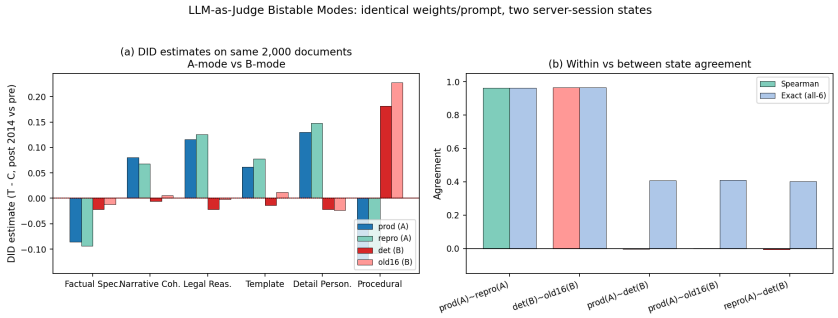

Figure M.1: Consequence of mode on a difference-in-differences analysis of the same 2,000 documents. (a) DID coefficient estimates per dimension under two A-mode measurements (prod, repro) and two B-mode measurements (det, agg). B-mode reverses the sign on up to 5 of 6 dimensions (det: 5/6; agg: 3/6). (b) Within-mode agreement is high (Spearman ≥ 0.96), while between-mode agreement is near zero.

Dimension	A (prod)	A (repro)	B (det)	B (agg)
factual spec.	-0.086	-0.094	-0.022	-0.012
narrative coh.	$+0.080$	$+0.068$	-0.006	$+0.006$
legal reasoning	$+0.116$	$+0.126$	-0.022	-0.002
template likeness	$+0.062$	$+0.078$	-0.014	$+0.012$
detail personal.	$+0.130$	$+0.148$	-0.022	-0.024
procedural comp.	-0.130	-0.108	$+0.182$	$+0.228$

DID estimates on identical documents, differing only in the scoring mode. Two independent B sessions (det, aggregate) are shown; A columns are production and an independent reproduction.

Table M.3: DID estimates depend on scoring mode. The two observed B sessions reverse sign on 5/6 and 3/6 dimensions respectively.

real campaigns (two A, two B), votes are taken per document and dimension with ties broken by the lower value (Eq. (10)), and the reported curves are the mean absolute DID bias (Eq. (G.10)) and the sign-flip rate (Eq. (G.9)). The policy never classifies sessions; it only aggregates. Voting drives the mean absolute bias toward zero whenever B is the minority ($p_B < 0.5$) and, at the historical arrival rate $p_B \approx 0.25$, reduces the sign-flip rate roughly three-fold; when B is as frequent as A ($p_B = 0.5$), the majority is itself in mode B about half the time and no number of sessions recovers the A-mode estimates.

M.6 Reference-Anchor Classification

Table M.4 reports the reference-anchor classification of a sample of campaigns: the Spearman of each campaign against the production scores on a fixed set of 20 documents. All A-campaigns score $\rho \geq 0.92$ and all B-campaigns $\rho \leq 0.08$, so the fixed anchor separates the modes with no overlap under the threshold $\rho > 0.5$.

M.7 Trigger Search: Controlled Conditions

We evaluated eight candidate conditions that might reproduce the B mode. Each of the eight conditions (including cold restart, load, concurrency, prompt length, and within-session time) left the server in mode A; none reproduced B. Across the study, B appeared 4 times in 16 campaigns at scattered timestamps; no controllable condition in our search reproduced it. A ninth probe tests low-temperature sampling ($T \in \{0.3, 0.5, 1.0\}$): scores remain close to A (mean per-document Spearman 0.57 vs. A, 0.04 vs. B) and do not reproduce B (0 of 540 samples). B is therefore a change in the realized session function rather than a perturbation of greedy decoding. Consistently, the one B campaign served under sampling (temperature 0.7; Table M.1) shows that B is not tied to greedy decoding, while the sweep above shows it is not triggered by sampling temperature either.

Campaign	Anchor ρ	Class
agg2000_now	$+0.98$	A
agg2000_3rd	$+0.95$	A
sample300_zh	$+0.97$	A
seq_retest	$+0.95$	A
det_retest	-0.04	B
aggregate_retest	-0.04	B
reliability (B-era)	$+0.01$	B

Anchor Spearman between a campaign and the production scores on a fixed set of 20 documents. The threshold $\rho > 0.5$ separates modes with no overlap.

Table M.4: Reference-anchor classification of a sample of campaigns. A fixed 20-document anchor separates modes perfectly (all A-campaigns $\rho \geq 0.92$; all B-campaigns $\rho \leq 0.08$).

Judge (stack)	Anchor span	$\bar{\rho}_{A_1B_1}$	ICC(1,1)	DID flips
Qwen2.5-14B (vLLM)	1.016	$0.596\,[0.584, 0.609]$	≈ 0.00	$5/6$
same-mode (A_1A$_2$)	—	$0.991\,[0.989, 0.992]$	0.972	—
Qwen3.5-9B (vLLM)	0.001	$0.989\,[0.987, 0.991]$	0.970	$0/6$
GLM-4-9B (vLLM)	0.007	$0.990\,[0.988, 0.991]$	0.916	$0/6$
Gemma-3-12B (llama.cpp)	-0.049	$0.981\,[0.978, 0.983]$	0.932	$1/6$

Anchor span = difference between the two sessions' 20-document anchor Spearman (Eqs. (8)); median ICC(1,1) across the six dimensions; DID flips relative to session A_1.

Table M.5: Cross-model and cross-stack reproduction of the measurement protocol. Bistability does not reproduce; reliability and calibration to the production scale do not track each other.

M.8 Across Models and Stacks

None of the three additional judges reproduces bistability. Between-session per-document agreement is 0.98–0.99 for all three (vs. 0.596 for Qwen2.5), per-dimension ICC(1,1) is 0.82–0.98 (vs. ≈ 0), and the DID estimates are stable across sessions (0 of 6 dimensions with disjoint bootstrap CIs; $\leq 1/6$ sign flips). The distribution of per-document agreement sharpens the contrast: Qwen2.5 leaves 4.3% of documents rank-inverted and 36.6% in the unstable band $0.3 \leq \rho \leq 0.7$ (median 0.67), whereas for the other three models 80–91% of documents are fully unchanged (median 1.00) and none is inverted. Leave-one-dimension-out preserves the conclusion (0.49–0.70 for Qwen2.5 under every single-dimension removal; ≥ 0.975 otherwise), and bootstrap CIs are identical to ± 0.001 across resampling counts (100–1500) and seeds. The bistability is therefore a model-specific attractor of Qwen2.5-14B under this protocol, not an LLM-as-a-judge universal.

The comparison also isolates a second, independent axis of failure. We measured each judge's calibration against the production reference scale (Spearman between a candidate session and the original Qwen2.5 production scores on the shared 2,000 documents; this is distinct from the 500-document spectrum calibration of Table M.7, where Qwen2.5-14B is defined as its own reference). Qwen2.5 in mode A calibrates to its own production reference at 0.987; but GLM-4-9B, no less session-stable than its peers, calibrates at only 0.491, and Qwen3.5-9B and Gemma-3-12B at 0.706 and 0.705. Stability and cross-model transferability therefore *decouple*: a model can be highly reliable within session yet far from an established reference scale, and the model that loses reliability (Qwen2.5) retains calibration to its own scale. An evaluation protocol that checks only within-session consistency or only aggregate mean shifts will miss both halves of this decoupling, and this is exactly the blindness described in Section 5.2: Qwen2.5's rank collapse is invisible to per-dimension Wilcoxon tests ($0/6$ significant) and to the DID contrast (only $1/6$ dimensions with disjoint CIs), and the new models' scale drift is invisible to any within-model statistic. The rank-stability metrics of Section 4.1 (Eqs. (4)–(5)) separate the axis-1 failure that ICC and mean-based checks miss. These results motivate treating *cross-session rank stability and calibration to a shared reference* rather than merely exact-rate or mean-level reliability, as distinct, auditable quality gates for LLM-as-a-judge.

Judge	Size	$\|\Delta\text{mean}\|$	Calib.	Aligned	Rank-corr
Qwen2.5-1.5B	1.5	3.46	0.070	0.422	0.375
Qwen2.5-7B	7	2.41	0.261	0.503	0.542
Qwen3-4B	4	1.54	0.592	0.557	0.591
Phi-3.5-mini	3.8	3.63	0.333	0.582	0.330
Qwen3.5-9B	9	1.93	0.695	0.748	0.668
GLM-4-9B	9	1.70	0.499	0.566	0.375
Gemma-3-12B	12	0.97	0.696	0.724	0.559

$\|\Delta\text{mean}\|$: Euclidean distance between the judge's six-dimensional mean score vector and the production mean. Calib.: per-document Spearman vs. production. Aligned: calibration after per-dimension quantile matching to the production marginal. Rank-corr: Spearman of an aggregate quality score against the production ranking (after alignment it stays at 0.31–0.67). The pairwise calibration matrix over all eight judges has every off-diagonal entry in 0.07–0.80; this pairwise range bounds judge–judge pairs, whereas calibration to the production scale is 0.07–0.70 (Table M.7).

Table M.6: Scale drift is universal (all 28 pairwise entries 0.07–0.80), predictable out of sample (mean-vector distance on a held-out half predicts calibration with $r = -0.90$), and only the scale, not the ranking, is correctable: quantile alignment raises calibration but leaves the rank agreement with the reference at 0.31–0.67.

M.9 Scale Drift Is Predictable and Correctable

Two model-level statistics predict the calibration of Table M.7. First, the distance between a judge's mean score vector and the production mean vector predicts calibration *out of sample*: computing the distance on a random half of the documents and the calibration on the other half yields $r = -0.90$ ($p = 0.002$, Pearson, $N = 8$ models), so the predictability is not an artifact of measuring both on the same documents. We note the small-sample caveat: with $N = 8$, the bootstrap 95% confidence interval for the correlation is wide ($[-0.996, -0.669]$), and leave-one-out removal of any single model changes r only within $[-0.83, -0.95]$; we therefore treat this as an exploratory, model-level heuristic rather than a population law, and report it as such. Second, model scale is a looser predictor ($r = 0.84$ in sample; leave-one-out absolute error ≈ 0.17 in calibration). The drift is therefore not idiosyncratic noise but a predictable function of the judge's own score distribution, a "verbal anchor" that varies systematically across models and sizes. The drift is also *structural* across judges: the pairwise calibration matrix over the eight models has all 28 off-diagonal entries in 0.07–0.80 (Table M.6). Note that this pairwise upper bound (0.80) exceeds the to-production upper bound (0.70; Table M.7): the matrix additionally contains pairs between two off-the-shelf judges, which can agree more with each other than either does with the production scale. Correcting the scale is only partly possible: aligning each judge's per-dimension scores to the production scale's marginal distribution (quantile matching) raises calibration for the most drifting judges ($0.070 \to 0.422$ for Qwen2.5-1.5B, $0.261 \to 0.503$ for Qwen2.5-7B, $0.333 \to 0.582$ for Phi-3.5-mini). But alignment does *not* recover the reference *ranking*: after alignment, the aggregate-quality ranking of documents still correlates with the reference ranking at essentially the same level as before (0.31–0.67; Table M.6). A naive "calibrate then rank" pipeline fixes the scale and still reaches the wrong ranking. On a simple aggregate quality score, the ranking of the 500 documents under each judge correlates with the production ranking at only $\rho = 0.31$–0.67; the "best" documents differ substantially depending on which judge measured them. A single unvalidated judge does not merely add noise; it changes which conclusions the scores support.

M.10 The Two Axes Transfer Across Tasks

To test whether the framework is an artifact of the six-dimension legal rubric, we scored structurally degraded outputs on three other judge tasks (summarization with faithfulness, coherence, and concision; translation with adequacy, fluency, and faithfulness; and dialogue with coherence, helpfulness, and engagement), with the same production judge in two independent sessions plus a shortened-prompt variant. These are small-scale, exploratory runs ($n = 20$ summarization, $n = 19$ translation, $n = 16$ dialogue, degraded items per task) and are reported as preliminary evidence that the axes are not artifacts of the legal rubric, not as population-level claims. On summarization and

Judge (stack)	$\bar{\rho}_{A_1B_1}$	Exact	ICC	Calib.
Qwen2.5-1.5B (llama.cpp)	1.000	1.000	1.000	0.070
Qwen2.5-7B (vLLM)	1.000	1.000	1.000	0.261
Qwen3-4B (vLLM)	1.000	1.000	1.000	0.592
Phi-3.5-mini (vLLM)	1.000	1.000	1.000	0.333
Qwen2.5-14B (vLLM)	0.992	0.906	0.976	1.000
Qwen3.5-9B (vLLM)	0.988	0.874	0.972	0.695
GLM-4-9B (vLLM)	0.989	0.912	0.929	0.499
Gemma-3-12B (llama.cpp)	0.979	0.800	0.934	0.696

Axis 1: mean per-document row Spearman and all-dimension exact agreement between two independent sessions ($N = 500$ for the spectrum judges, 435 for Phi-3.5-mini; $\mathrm{ICC}(1,1)$ is the median across dimensions). Axis 2: per-document Spearman of one session against the Qwen2.5 production scores on the same 500 documents.

Table M.7: Eight-model spectrum spanning 2024 and 2025 generations. Only Qwen2.5-14B has ever exhibited bistability; every other judge, including the newest we could run (Gemma-3-12B, released 2025), is cross-session stable, yet all calibrate to the production scale at $\rho = 0.07$–0.70.

Judge (group)	Calib. (axis 2)
Small stable models (2024–2025)	0.07–0.59
Qwen2.5-14B (bistable)	1.000
Qwen3.5-9B, GLM-4-9B, Gemma-3-12B	0.50–0.70

Axis 2: per-document Spearman of one session against the production scores on the same documents ($N = 500$). Cross-session agreement is $\bar{\rho}_{A_1B_1} = 0.979$–$0.992$ (exactly 1.000 for the four smallest judges; Qwen2.5-14B is 0.992). After per-dimension quantile alignment, aggregate-quality rank agreement with the production ranking is 0.31–0.67 for all judges.

Table M.8: The two axes are orthogonal: cross-session stability (axis 1) is near-perfect for every judge except Qwen2.5-14B, while calibration (axis 2) is universally poor and stays poor after alignment (rank agreement 0.31–0.67).

translation, the judge's scores track the structural degradation monotonically (score–level Spearman -0.60 and -0.36), so the instrument measures the intended constructs outside the legal domain. Cross-session stability also transfers: on all three tasks the judge reproduces its scores bit-for-bit across the two independent sessions (cross-session agreement 1.000). Prompt sensitivity likewise transfers: re-ranking the same items with a shortened prompt changes per-item agreement to $\bar{\rho} = 0.79$ (summarization) and 0.89 (translation). The dialogue degradation design was invalid (its levels did not order quality; score–level $+0.31$), which we report as a negative result of the item design rather than of the framework. The two axes, rank stability and scale calibration, and their controlling factors (session state, prompt, judge identity) therefore transfer across tasks in this small-scale probe; only the specific legal finding was domain-specific, and a larger cross-task study is left to future work.

M.11 Spectrum Table

Table M.7 spans eight models across the 2024 and 2025 generations on the two axes of Section 5.6, per-session stability (axis 1) and calibration to the production scale (axis 2). Only Qwen2.5-14B has ever exhibited bistability; every other judge is cross-session stable ($\bar{\rho}_{A_1B_1} \geq 0.979$), yet all calibrate to the production scale at only $\rho = 0.07$–0.70. Table M.8 restates the same contrast at group level: cross-session stability is near-perfect except for Qwen2.5-14B, while calibration is universally poor and stays at 0.31–0.67 rank agreement after quantile alignment—the two axes are orthogonal.

M.12 Stability Bound and Model Vintage

Because bistability is a sparse event (Qwen2.5 historically showed it in $4/16$ campaigns), a two-session test is too weak to rule it out in other models. We therefore ran 16 independent fresh sessions of each of four judges, Qwen2.5-14B, Qwen3.5-9B, GLM-4-9B, and Gemma-3-12B, classifying each session by a fixed 20-document reference anchor. Across the 63 released session files, no B-mode was observed (one session record is absent from the archive); per judge, the 95% binomial bound on a single judge's per-session B-rate is ≤ 0.18 (16 sessions: ≤ 0.17; 15 sessions, as for Gemma-3-12B: ≤ 0.18), and Qwen2.5-14B's own historical rate ($4/16 = 0.25$) is rejected at 95%. The bistability did not recur under controlled, back-to-back serving sessions; it is condition-dependent in our data. The instability is also *vintage-bounded* in our data: the only model in which we observe rank instability across sessions is Qwen2.5-14B, a 2024-generation model, whereas every 2025-generation judge we scored (Qwen3-4B, Qwen3.5-9B, Gemma-3-12B) is cross-session stable at $\bar{\rho} = 0.979$–1.000. We therefore treat axis-1 as a bounded failure of older deployments rather than a demonstrated property of current frontier judges, which is exactly why it must be audited (Section 5.5) rather than assumed absent.

M.13 Cross-Judge Validity Sample

This study measures the *stability of a scoring instrument* across sessions and models, not the accuracy of scoring against a gold standard, so the natural external references are human ratings and agreement with other judges rather than a rule-based or classical-ML baseline scorer (which has no comparable notion of session-level mode switching). The construct, convergent, and cross-judge validity checks below use those references. The expert human-annotator pilot (8 documents, one annotator, per-document Spearman 0.918 against the production scores) and the four-judge sample are supporting evidence for the claim that the six dimensions measure meaningful constructs, not a benchmark of scoring accuracy.

Positioning axis 2 against existing disagreement baselines. A natural question is whether the cross-model scale drift of axis 2 is distinct from the known disagreement among human annotators and existing judge benchmarks. We distinguish three reference points to position this contribution. First, *human inter-rater reliability*: our expert pilot is a single annotator, so we do not report a human ICC; the relevant contrast is that a within-session judge is *bit-for-bit* reproducible ($\bar{\rho}_{A_1 B_1} = 1.000$) while a cross-model judge is *not* ($\bar{\rho} = 0.47$–0.80 in Table M.9); human judges are never bit-for-bit reproducible, so the axis-1 stability we exploit has no human analogue, and the axis-2 disagreement we measure is between deterministic instruments that each *appear* perfectly reliable internally. Second, *preference benchmarks* (RewardBench (Lambert et al., 2025), MT-Bench (Zheng et al., 2023)) measure inter-model disagreement through *aggregate preference or win rates*, mean-level statistics that, as Section 5.2 shows, cannot see per-document re-permutation; they also compare judges on *whose pairwise preference agrees with a gold*, a different quantity from whether two judges put the *same documents* on a shared ordinal scale. Our axis-2 measurement is per-document row agreement on the *same* corpus across judges, which is the quantity required when one judge's scores are audited against another's. Third, the *reference-anchor* protocol (Section 4.3) is itself a reproducibility baseline: a fixed 20-document anchor is reproducible to $\rho \geq 0.92$ within a mode and $\rho \leq 0.08$ across modes, so the anchor gives every deployment a cheap check that the judge is on its own historical scale. These contrasts locate the contribution as a *measurement* result (per-item agreement on a shared corpus, invisible to aggregate checks) rather than a re-statement of the known fact that different judges score differently.

M.14 Frontier Stress Test Details

We first extended the eight-judge spectrum with two further open-weight judges (Qwen2.5-3B and InternLM2.5-7B), raising the cross-model sample to $N = 10$. The out-of-sample drift predictability is preserved and, if anything, stronger: $\mathrm{pearson}(|\Delta\bar{\mathbf{s}}|, \mathrm{calib}) = -0.76$ ($p = 0.010$; 95% bootstrap CI $[-0.98, -0.58]$). Leave-one-out deletion of any single judge keeps $r \leq -0.74$, so the negative relation is not an artifact of the original eight judges. To test whether the calibration gap closes at the frontier, we then added Qwen3-14B, a newer-generation 14B open-weight judge, to the spectrum. It also lies on the same drift–calibration spectrum: with it the cross-model sample is $N = 11$ and $\mathrm{pearson}(|\Delta\bar{\mathbf{s}}|, \mathrm{calib}) = -0.751$ ($p = 0.0077$); its own calibration is 0.622, rising to 0.672 after per-

	Qwen2.5	Qwen3.5	GLM	Gemma
Qwen2.5-14B	—	0.721	0.469	0.763
Qwen3.5-9B	0.721	—	0.750	0.795
GLM-4-9B	0.469	0.750	—	0.671
Gemma-3-12B	0.763	0.795	0.671	—
vs. majority consensus	0.814	0.889	0.749	0.880

Mean per-document row Spearman between pairs of judges on the shared 160-item sample ($N = 160$; exact agreement $\leq 1.3\%$ in all pairs). The last row is each judge's agreement with the four-judge majority-vote consensus.

Table M.9: Cross-judge agreement on the 160-item validity sample. Agreement is substantial ($\bar{\rho} = 0.69$) but not judge-invariant; the residual spread is the axis-2 scale drift.

dimension quantile alignment. The gap therefore persists at larger, newer scale: even the newest 14B judge reaches only about 62% of the production calibration, far below the full-precision Qwen2.5-14B reference ($\mathrm{calib} = 1.000$). Crucially, this shortfall is *not* a quantization artifact: scoring the identical Qwen3-14B at full BF16 precision yields an *identical* calibration ($\mathrm{calib} = 0.622$, aligned $= 0.677$), so the frontier gap is a genuine generation/architecture effect rather than an AWQ deployment artifact. Closed/API judges (e.g. GPT-4-class) were not tested in this study; the cross-model claim is bounded to open-weight judges, although Qwen3-14B is among the strongest openly available open-weight judges, so the frontier gap is not attributable to model size alone.

M.15 Re-Serving Under the Current Stack

We next asked whether the B-mode hazard is a standing property of the weights/prompt or specific to the serving stack on which the four B-mode campaigns originally appeared. We re-served the production judge, Qwen2.5-14B, under the *current* upgraded stack (vLLM 0.26, CUDA 12.8) in its actually-deployable AWQ-quantized form, and rescored the same 200 documents $K = 3$ times with a cold restart before the second session. Two questions were asked. *(i) Endpoint reproducibility:* across the three rescored sessions the judge reproduced its own six-dimension scores exactly: pairwise Spearman $\rho = 1.0000$, per-dimension exact agreement 1.000, mean-score shift 0.000, and zero B-mode events. The served endpoint is therefore deterministic and fully reproducible across restarts, consistent with Appendix M.7 in showing that no B-mode re-permutation recurs under the current stack. *(ii) Quantization fidelity:* the AWQ endpoint is *not* a faithful proxy of the full-precision reference it approximates: against the full-precision Qwen2.5-14B spectrum it scores at $\rho = 0.650$ with a systematic downward shift of the mean six-dimension score from 3.183 to 2.150 (-1.03 on the 1–5 scale). Reproducibility is thus conditional on identical weights *and* precision: a deployable quantized copy preserves the judge's internal consistency but relocates the entire spectrum. Serving the full-precision judge, or explicitly validating that quantization does not shift the scale, is a prerequisite before the spectrum is used in downstream analyses.

M.16 Bootstrap Uncertainty for the DID Hazard Estimates

Table M.3 reports point estimates, which invites the objection that the A–B gap lies within sampling noise. We quantify it with a document-level bootstrap on the same 2,000 documents: $B = 2{,}000$ replicates, percentile intervals, seed 20260907, with all four campaigns resampled on a common index vector inside each replicate, so the mode contrast is a paired comparison. The primary scheme resamples with replacement within the four DID cells (treated/control $\times$ pre/post, 500 documents each), preserving the design's cell sizes; an iid document bootstrap changes every interval endpoint below by at most 0.02.

Three readings follow. *(i) Precision.* A single-dimension estimate carries a 95% interval of half-width 0.09–0.19 (median ≈ 0.11), the same order as the mean absolute A–B gap: $\mathrm{MAE} = 0.138$ for both B sessions (95% CI $[0.09, 0.21]$), and 0.146 when the A-mode re-run is used as the reference, as in the consensus simulation of Section 5.5. No dimension has disjoint A and B intervals ($0/6$), and only 1 of the 12 paired $B - A$ differences excludes zero: procedural completeness, $+0.312$ $[+0.046, +0.582]$ (det) and $+0.358$ $[+0.094, +0.626]$ (agg). The hazard of Section 5.4 is therefore *not* the claim that a single dimension's estimate moved beyond noise. *(ii) What does reverse is the*

Dimension	A (prod)	A (repro)	B (det)	B (agg)
factual spec.	-0.086	-0.094	-0.022	-0.012
	$[-0.198, +0.022]$	$[-0.208, +0.012]$	$[-0.140, +0.088]$	$[-0.130, +0.100]$
narrative coh.	$+0.080$	$+0.068$	-0.006	$+0.006$
	$[-0.010, +0.178]$	$[-0.026, +0.166]$	$[-0.106, +0.090]$	$[-0.096, +0.104]$
legal reasoning	$+0.116$	$+0.126$	-0.022	-0.002
	$[+0.020, +0.216]$	$[+0.024, +0.228]$	$[-0.120, +0.078]$	$[-0.098, +0.100]$
template likeness	$+0.062$	$+0.078$	-0.014	$+0.012$
	$[-0.054, +0.176]$	$[-0.036, +0.194]$	$[-0.136, +0.108]$	$[-0.104, +0.136]$
detail personal.	$+0.130$	$+0.148$	-0.022	-0.024
	$[-0.020, +0.294]$	$[-0.002, +0.308]$	$[-0.196, +0.154]$	$[-0.198, +0.152]$
procedural comp.	-0.130	-0.108	$+0.182$	$+0.228$
	$[-0.308, +0.066]$	$[-0.290, +0.084]$	$[-0.002, +0.370]$	$[+0.044, +0.420]$

DID point estimate (top) and cell-stratified bootstrap 95% interval (bottom) on the same 2,000 documents; $B = 2{,}000$ replicates, seed 20260907, paired across campaigns. The A (prod) column is the production measurement itself (`measure_full.csv` restricted to the 2,000 documents); the other three are independent campaigns.

Table M.10: Bootstrap intervals for the DID hazard table. Every entry is estimated to about ± 0.11; no interval is disjoint between the modes.

reported conclusion. Under both A measurements exactly one dimension has an interval excluding zero: legal reasoning density ($+0.116$ $[+0.020, +0.216]$); under B (det) *no* dimension does, and under B (agg) a *different* dimension does (procedural completeness, whose A-mode sign is negative). The set of dimensions a study would report as affected by the reform is disjoint between the modes ($0/6$ overlap). *(iii) The sign pattern survives resampling.* Averaged over replicates the sign-flip rate is 0.62 (det) and 0.57 (agg), against point values $5/6$ and $3/6$, with per-replicate 95% intervals $[0.33, 0.83]$; the reversal is not an artifact of one document sample.

The bootstrap supports exactly the reading given in Section 5.4: a mode switch changes what a single un-audited campaign would report, the signs and which dimensions reach significance, by an amount comparable to the design's entire sampling uncertainty. This is why the failure must be audited rather than averaged away. The analysis script is released with the artifact.

N Practical Checklist for LLM-as-a-Judge Deployments

The two axes translate into a short, mechanism-agnostic audit that can be run before any set of judge scores is consumed downstream. Each item is stated so it can be checked without knowing *why* a judge deviates, since the trigger of Section 5.3 remains unidentified.

1. **Treat within-session retest as a check of one function, not as reliability evidence.** A deterministic function is trivially self-consistent, so single-session checks carry no information about other modes; column-wise statistics are additionally unchanged by a rank re-permutation (Section 5.2).
2. **Score a fixed reference anchor (about 20 documents) in every new session** and reject the session when the anchor Spearman falls below $\tau = 0.5$ (Section 4.3); on our data this separates modes with no overlap at a cost of 20 judgments per campaign.
3. **For high-stakes analyses, aggregate over sessions**: majority-vote $R \geq 5$ independent sessions (Section 4.4). This is valid only while the anomalous mode is the minority, $p_B < 0.5$; at $p_B = 0.5$ no number of sessions recovers the reference (Figure 2), so consensus is a mitigation, not a guarantee.
4. **Audit both axes, not one.** Cross-session rank stability (axis 1) and calibration to a shared reference (axis 2) are orthogonal: a judge can be bit-for-bit reproducible and still sit on a nearly orthogonal scale. Quantile alignment fixes the marginals and does *not* recover the reference ranking (0.31–0.67 after alignment), so a calibration check is not a ranking check.
5. **Re-audit whenever the realized scoring function can change**: a new serving stack or quantization level, a rewritten prompt, a shorter document window, or a reduced dimension set (Appendix M.4). Reliability is a property of the realized function, not of the model card.