THE JUDGE LOTTERY: DIAGNOSING WHETHER COMPARISONS BETWEEN STOCHASTIC LLM JUDGES ARE ADEQUATELY SUPPORTED BY RESAMPLING EVIDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

LLM judges gate safety decisions and supply RLHF/DPO training signal, yet a single verdict is stochastic – shaped by decoding temperature/seed, presentation order, and which ensemble member is queried. Prior work shows judges are biased; it does not tell a practitioner whether a specific small-sample comparison is adequately supported by evidence or is instead an artifact of a favorable resampling draw. We call this the *Judge Lottery* and formalize it with a disagreement rate $D(K)$, a matched-pair order-flip rate, and a scoped, single-example planning estimate $K_{\min}$, together with a five-step protocol (calibration, resampling, a degenerate-verdict screen, regime classification, and Resample-Consensus reporting) for diagnosing, not certifying, whether a comparative judge claim rests on adequate resampling evidence. In a direct pilot on two open-weight judges (Qwen2.5-1.5B/3B-Instruct, 30 shared MT-Bench examples, 1,920 calls, valid rate 1.000), matched AB/BA verdicts disagree at 0.744 (1.5B) and 0.662 (3B), a large effect with no detectable temperature variation that persists under deterministic decoding and that further resampling at fixed order does not correct; a cluster-respecting re-analysis finds the two judges' rates statistically indistinguishable despite each being individually large. Synthetic content-blind controls show low verdict variance can coexist with a stable, non-evaluative heuristic. We support the protocol with a mechanism-level analogue from a concurrently submitted differential-privacy fairness study and map it onto representative LLM-judge failure modes. A verified post-submission extension to a 7B same-family judge and a second benchmark finds order bias shrinking, but not disappearing, with scale, alongside improving agreement with real human preference labels (Cohen's κ: -0.122 at 1.5B, $+0.368$ at 7B). All direct findings are scoped to same-family judges, two benchmarks, and 30 examples; the degenerate-verdict screen and RCS's order-sampling policy remain diagnostics under refinement rather than finished gates.

1 INTRODUCTION

Judges embedded in RLHF (Ouyang et al., 2022; Bai et al., 2022), DPO (Rafailov et al., 2023), and Constitutional AI pipelines gate safety decisions and become training signal directly, so an unreliable judge propagates its error into every downstream decision it touches. An LLM judge is not a fixed function of its input: its verdict can change with decoding temperature/seed, presentation order, and which ensemble member is queried (Zheng et al., 2023; Wang et al., 2024; Koo et al., 2024). Position swaps flip a nontrivial fraction of pairwise verdicts (Wang et al., 2024; Zheng et al., 2023), and judges favor outputs stylistically close to their own family (Panickssery et al., 2024).

Matched bidirectional evaluation (scoring both the AB and BA orderings of a pair, sometimes under deterministic decoding, to isolate positional preference) is itself standard practice, used for example in MT-Bench's swap protocol (Zheng et al., 2023) and in calibration frameworks that aggregate pre-

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.

dictions across swapped orderings to cancel positional skew (Wang et al., 2024). Our contribution is not the introduction of matched-pair swapping. It is a different, more actionable question that matched-pair evidence alone does not answer: given a specific comparison between two judges, or between two conditions evaluated by the same judge, is it adequately supported by resampling evidence, or could it just as easily be a favorable resampling draw? A judge can be known to be biased and a reported comparison involving it can still be uninformative if drawn from too few resamples to distinguish a real difference from the judge's own resampling variance. We call this the *Judge Lottery*, by analogy to differentially private fairness evaluation, where few-seed comparisons have been shown to report improvements that a full seed sweep does not support (a phenomenon audited on algorithmic methods from that literature, including those of Bagdasaryan et al. 2019 and Tran et al. 2021, by a concurrently submitted study; see Section 5). We do not claim judge bias itself is new, and the protocol below does not guarantee a trustworthy comparison – it is a diagnostic that flags when a comparison is under-resampled or content-blind, not a correctness certificate. What is missing from existing matched-design practice is a framework separating (i) detecting bias, (ii) measuring instability, (iii) determining whether a comparison is adequately powered via explicit uncertainty quantification and sample-size planning, (iv) classifying a judge's operating regime, and (v) deciding whether a comparative claim should be reported or deployed as a gate. To our knowledge, this is the first submission to formalize this chain and test its resampling statistics directly on an LLM judge with a matched, bootstrap-quantified design.

Scope note on $K_{\min}$. The overall diagnostic question above is comparative (two judges, or two conditions). The planning estimate $K_{\min}$ introduced in Section 3, by contrast, is strictly a single-example, single-judge estimate of the draws needed to detect a mean shift against indifference – it is not a judge-vs-judge power calculation. We flag this scope mismatch explicitly here rather than let it surface only in Section 3: multi-item, comparative sample-size planning is future work, not a claim of this paper.

Contributions. (1) We formalize the Judge Lottery with $D(K)$, a matched-pair order-flip rate, and a corrected, single-example $K_{\min}$ scoped as a planning estimate, not a universal requirement. (2) A five-step diagnostic protocol implementable as a pre-deployment screen, including a degenerate-verdict check instantiated with content-blind synthetic controls. (3) A direct, bootstrap-quantified pilot on two real LLM judges (30 MT-Bench examples, 1,920 calls) showing a large order effect, with no detectable temperature variation, that persists under deterministic decoding and is not corrected by resampling at fixed order – the central demonstration that stochastic-uncertainty reduction and order-bias correction are not the same thing – together with a concrete, worked demonstration of pseudo-replication: a naive pooled test over 450 matched pairs manufactures a judge-to-judge difference that a cluster-respecting test at the correct experimental unit (n=30 examples) does not support. (4) A mapping of the protocol onto representative LLM-judge failure modes and a candidate Reliability Regime disclosure field, introduced in this work, for reporting comparative judge claims.

2 Related work

LLM-judge bias work documents sensitivity to length (Zheng et al., 2023), order (Wang et al., 2024; Zheng et al., 2023), and self-preference (Panickssery et al., 2024), catalogued across judge families (Koo et al., 2024). Calibration work such as G-Eval (Liu et al., 2023), which uses probability-weighted, multi-sample score averaging to improve human alignment, addresses construct validity – whether a judge measures the right thing on average – which is orthogonal to whether repeated evaluation is stable, and orthogonal again to whether a comparison is adequately powered. Self-consistency decoding (Wang et al., 2023) establishes that majority-vote aggregation over multiple stochastic decodes improves reasoning accuracy; our Resample-Consensus reporting step (Section 4) applies the same aggregation logic to judge verdicts, but as a reporting policy rather than an accuracy-improving inference technique, and pairs it with an explicit disclosure of the regime under which the aggregate was obtained. Positional-bias correction methods – swap-tie evaluation (Zheng et al., 2023) and the calibration framework of Wang et al. (2024) – mitigate order bias by aggregating over both orderings; our pilot (Section 6) evaluates resampling at a *fixed* order specifically to isolate what decoding-only resampling can and cannot fix, and is complementary to, not a replacement for, these order-aggregating methods. Seed instability is documented outside the judge setting, in reinforcement learning (Henderson et al., 2018) and, most closely, in differentially private fairness

evaluation (Bagdasaryan et al., 2019; Tran et al., 2021), where a concurrently submitted audit finds that few-seed comparisons on methods from that literature report improvements a full sweep does not support. The gap our protocol targets: existing bias and bidirectional-evaluation literature does not provide explicit uncertainty quantification, sample-size planning, or a resample-budget/reporting protocol; seed-instability literature quantifies the analogous variance outside the judge-evaluation setting. Sections 3–6 address this combined gap.

3 A STOCHASTIC MODEL OF JUDGE VERDICTS AND THE JUDGE LOTTERY

Model. A judge J produces a verdict $V = f(J, x, \tau, s, \pi)$ for input x under decoding temperature τ, seed s, and presentation order π. We call $c = (\tau, s, \pi)$ a configuration and treat variation across c at fixed x as the object of interest: (i) stochastic decoding variation (across τ, s at fixed π), (ii) systematic presentation variation (across π at fixed τ, s), (iii) example-level heterogeneity (across x). Section 6 shows these are empirically distinct: a judge can be highly repeatable under decoding resampling while remaining systematically sensitive to order, so pooling (i) and (ii) into one variance number hides exactly the component that matters for deployment.

Disagreement rate. For K verdicts $V_1, \ldots, V_K \in \{-1, 0, +1\}$ (loss/tie/win, or any categorical encoding) drawn at fixed x across varying configurations, $D(K) = 1 - \frac{1}{K} \max_v |\{i : V_i = v\}|$ is the fraction departing from the modal verdict. $D(K)$ is exact for the observed K draws, not a Monte Carlo estimate, and measures departure from the modal categorical verdict only – it does not measure correctness or semantic validity, and does not replace the matched-pair statistic below, which it is reported alongside.

Matched-pair order-flip rate and PCR. Fixing x, τ, and s and varying only $\pi \in \{AB, BA\}$ yields a matched pair. The order-flip rate is the fraction of matched pairs whose verdicts disagree once position labels are aligned to the same candidate identity – unlike a pooled variance statistic, it isolates π while holding the other axes fixed. $PCR = 1 -$ (order-flip rate) is the same estimand reparameterized as a consistency rate; we use both terms below. **This paper's regime classification (Section 6) is built from a decoding-resampling instability statistic, not from PCR**; we flag this explicitly, since a judge with a rigid, order-independent position preference and zero decoding variance would register as decoding-stable while still being order-biased. A deployment-readiness judgment therefore requires checking both the decoding-stability regime below *and* the order-flip rate/PCR together – the regime label alone is not sufficient, and we have not fixed a joint threshold for combining them, which we flag as open work (Section 8).

Coefficient of variation – scoped, not default. For score-valued verdicts with a mean bounded away from zero, $CV[V] = sd(V)/mean(V)$ is a scale-free instability summary, used only for the DP-fairness analogue (Section 5). For categorical verdicts encoded symmetrically ($\{-1, 0, +1\}$), CV is fragile: a symmetric AB/BA flip pattern drives $mean(V) \to 0$ while $sd(V)$ stays large, so CV diverges exactly when order bias is most severe. We use $D(K)$ and the order-flip rate as the default instability statistics for categorical verdicts instead.

$K_{\min}$ – a secondary planning diagnostic, not a requirement.

$$K_{\min} = \frac{(z_{\alpha/2} + z_{\beta})^2 \, Var[V]}{\Delta^2}$$

is a single-example planning estimate: the draws needed, under a normal approximation, to detect a mean shift Δ against indifference, at significance level α and power $1 - \beta$. As emphasized in Section 1, this is a single-example, single-judge estimate, not a judge-vs-judge power calculation. It treats an estimated $Var[V]$ as known, understating its own uncertainty, and its operational domain is restricted to $Var[V] > 0$: for an example with zero observed variance, the estimate is not meaningful under the normal approximation and we report it as undefined rather than zero. We report $K_{\min}$ only as a scoped diagnostic, never as a universal resampling requirement. **We fix $\alpha = 0.05$, power $1 - \beta = 0.8$, and a minimum detectable mean shift $\Delta = 0.2$ on the $\{-1, 0, +1\}$ verdict scale; $Var[V]$ is the per-example sample variance (with Bessel's correction) computed by pooling all 32 configurations (30 stochastic plus 2 deterministic draws) for that example.** These are implementation choices, not derived quantities; $\Delta = 0.2$ in particular is a somewhat arbitrary operating point on a scale with range 2, and we have not validated its sensitivity (Section 8).

Decoding-reversal probability ("Flip P."). For a single example's pooled verdict multiset V (all 32 configurations, stochastic and deterministic together), let $\hat{v}$ be the observed modal verdict. Flip P. is the bootstrap probability, over 5,000 resamples of V with replacement (sample size $|V|$), that the modal verdict of the resample differs from $\hat{v}$. This is a decoding-only instability statistic – it depends on the same-order verdicts collected across (τ, s), not on the AB/BA order-flip comparison – and is distinct from both $D(K)$ (the observed, non-bootstrapped modal-departure fraction) and the matched-pair order-flip rate (which isolates π specifically). It is the statistic thresholded at 0.20 for the STABLE/UNSTABLE regime split (Section 6).

Degenerate validity. A judge can achieve low variance by collapsing onto a stable heuristic (always the longer response, always position A) rather than by evaluating content: stable but not valid. This is exactly what a variance-only screen would silently certify as reliable. We instantiate two synthetic, fully content-blind rules using the same Flip P. statistic just defined – *always-longer* (always selects whichever candidate response is longer, by character count) and *always-position-A* (always selects whichever candidate is shown in position A, regardless of content) – applied to the same 32-configuration grid as the real judges. Because both rules are deterministic functions of quantities other than content quality, Flip P. for always-longer is exactly 0 (the response-length comparison does not change across resamples), while Flip P. for always-position-A is governed entirely by how often position-A is occupied by a different underlying candidate across the AB/BA-balanced grid. We report the resulting numbers in Section 6. A separate, exploratory split-half comparison (Section 6) probes a different failure mode – naive low-K comparisons between two randomly-labeled but distributionally identical subsamples of a judge's own stochastic verdict pool – and remains less precisely specified; we flag it as such below.

4 A FIVE-STEP DIAGNOSTIC PROTOCOL

The protocol diagnoses whether a comparative judge claim is adequately supported by resampling evidence; it does not certify that a judge is correct, unbiased, or safe to deploy. (1) *Calibration* verifies configurations are genuinely varied, not pseudo-resampled. (2) *Resampling* collects K draws per example; this reduces stochastic uncertainty only and, as Section 6 shows, does not by itself correct systematic order bias. Whether these K draws are collected at a single fixed presentation order or balanced across $\pi \in \{AB, BA\}$ is a configuration choice, not a fixed part of the protocol: our pilot deliberately isolates the fixed-order case to measure what decoding resampling alone can and cannot fix (Section 6); a deployment that instead balances $K/2$ draws per order and aggregates in candidate-relative space is a distinct configuration whose consensus behavior under near-total order bias (which can produce an exact tie between the two candidate identities) we have not characterized, and we flag this as unresolved rather than claim RCS's tie-handling generalizes to it (Section 8). (3) A *degenerate-verdict screen* runs before regime classification, since a content-blind judge would otherwise be misclassified as STABLE by a variance-only check. In this paper the screen is instantiated only as an exploratory split-half rule (Section 6); we have not fixed its formal test statistic or rejection threshold, and we report it as an open research question rather than an operationalized gate ready for pre-deployment filtering. (4) *Regime classification* uses thresholds treated as calibration-dependent operating points, not universal constants (Section 6), and – per the scope note in Section 3 – classifies decoding-stochastic stability specifically, not joint decoding-and-positional reliability. (5) *Resample-Consensus reporting* (RCS) aggregates screened verdicts into one reported value, reachable only after the full chain. RCS is not a novel estimator – majority-vote and median aggregation are standard (Breiman, 1996; Wang et al., 2023) – its contribution is procedural: integrating a verdict-type-appropriate aggregation rule into reporting, so a single arbitrarily-selected run is never the reported comparison. We recommend majority vote for categorical verdicts (a $\{-1, 0, +1\}$ median can output a value, e.g. "tie", that was never the modal category under a bimodal distribution) and median for continuous/ordinal scores.

5 EVIDENCE FOR THE MECHANISM: A MOTIVATIONAL ANALOGUE FROM DIFFERENTIAL PRIVACY

We support the protocol's premises with a concurrently submitted study of an analogous evaluation-reliability problem in differentially private fairness evaluation (Anonymous, 2026). We use "motivational analogue" rather than "evidence" deliberately: these results are *not* evidence about LLM

judges, the underlying mechanisms (DP-SGD training-seed variance versus autoregressive decoding and prompt-order sensitivity) differ substantially, the cited study is itself an unpublished concurrent submission whose experimental setup cannot be independently audited here, and the paper's central claims about LLM judges (Section 6) do not depend on this section. It establishes only that the shared high-level pattern – a comparison between stochastic evaluators, drawn from too few resamples, reported as if deterministic – is real and correctable in an already-measured domain; it does not establish that the same effect sizes, thresholds, or judge behavior transfer. At the analogous "honest" operating point, CVs of 0.881 and 0.974 were measured for two evaluation methods; across 120 exhaustive seed-triplet resamples of a 10-seed evaluation, 0 showed a fairness-constrained method improving over its baseline under full-sweep comparison. A zero-cost aggregation step, analogous to RCS, reduced the 3-seed instability range by 31% (127 $\rightarrow$ 87 points; Wilcoxon $p = 0.002$; permutation $p = 0.0018$), as reported in that concurrent study.

6 A DIRECT, BOOTSTRAP-QUANTIFIED PILOT ON REAL LLM JUDGES

This pilot is the paper's central empirical result. Its logic: (1) judge verdicts are stochastic in decoding; (2) matched AB/BA pairs isolate presentation-order effects from that stochasticity; (3) the observed order-flip rate is large; (4) it persists under deterministic ($\tau = 0$) decoding, ruling out ordinary decoding randomness as the sole explanation; (5) increasing K can still improve reproducibility of an aggregate decision (below); (6) but resampling at fixed presentation order cannot correct this order bias, since the bias is in π, not τ, s; (7) so "more samples" and "more reliable evaluation" are not synonymous. All findings hold only for these two Qwen2.5 judges and 30 MT-Bench examples; we write "in this pilot" throughout, and we note that both evaluated models are sub-3B instruction models, which are known to be more prompt-fragile than larger or purpose-built evaluator models – we do not claim the magnitude of the order-flip rates observed here generalizes to larger or swap-debiased judge architectures (Section 8).

Design. We evaluate Qwen2.5-3B-Instruct and Qwen2.5-1.5B-Instruct, off-the-shelf, on 30 shared MT-Bench human-judgment examples (first 30 usable IDs, default order, no post-hoc filtering) (Zheng et al., 2023). Per example: 3 temperatures ($\tau \in \{0.3, 0.7, 1.0\}$) $\times$ 5 seeds $\times$ 2 orders = 30 stochastic draws, plus 1 deterministic ($\tau = 0$) draw per order = 2 more, for 32 effective configurations (not the nominal 40 a naïve count implies, since seed is inert at τ=0) – 960 calls/judge, 1,920 total, valid rate 1.000. The 450 matched pairs per judge referenced below (Section 6) are the non-zero-temperature stochastic trials only (3 temperatures $\times$ 5 seeds $\times$ 30 examples $= 450$); the 30 deterministic-decoding pairs are analyzed separately as the "Det. flip" column of Table 1. A reproducibility package (raw 1,920-row log, analysis code, MIT-licensed) accompanies this submission as anonymized supplementary material; we independently re-ran it and confirm the integrity gate passes and every statistic below reproduces, with two disclosed exceptions: a ± 0.002 Monte-Carlo-consistent CI shift (Table 1) and the cluster-respecting re-analysis below. **The exact prompt template, output-parsing rule mapping model text to $V \in \{-1, 0, +1\}$, and any generation parameters beyond temperature (e.g., top-p, max tokens) live in that analysis code; we have not restated them in the main text, and doing so – verbatim, in an appendix or the supplementary README – is needed before the call log can be regenerated from the base checkpoints by a third party (Section 8).**

Order effect and the corrected judge-to-judge comparison. Both judges favor whichever response is shown in position A: 62.7–67.3% of stochastic trials (1.5B) and 71.3–72.7% (3B) (Figure 1). Under deterministic decoding, position-A win share was 70.0% (AB) vs. 30.0% (BA) for 1.5B and 73.3% vs. 10.0% for 3B – decoding stochasticity is absent by construction at τ=0, ruling out ordinary sampling randomness as the sole cause. The matched order-flip rate is 0.744 (1.5B) and 0.662 (3B) (Table 1); deterministic decoding alone produces similarly large disagreement (0.733, 0.667), confirming the effect persists without sampling stochasticity. The effect is broad-based (23/30, 20/30 examples majority-disagree) but not universal: 5 examples for the 3B judge show exactly zero disagreement across all 32 configurations (reason unknown). A naïve pooled two-proportion χ^2 test on the 450 matched pairs per judge gives $\chi^2 = 6.90$, $p = 0.0086$ for the judge-to-judge difference (risk difference 8.2 points, comparing the two judges' order-flip rates); this treats each pair as independent, but the true experimental unit is the example (only 30 of them), overstating the effective sample size. Re-analyzing at the correct unit ($n = 30$ paired per-example flip rates) gives a mean difference of 0.082 with a cluster-respecting bootstrap 95% CI of $(-0.100, 0.269)$ (includes zero),

Table 1: Main pilot results, independently re-verified against the released raw log. Order-flip (95% CI) is a bootstrap interval (5,000 resamples over the N=450 matched pairs per judge, i.e. resampling pairs, not the 30 underlying examples); Section 6 shows the judge-to-judge *comparison* is not distinguishable once the example is used as the resampling/clustering unit instead, and a cluster-bootstrap interval for each judge's own rate (resampling the 30 examples rather than the 450 pairs) is a disclosed open item (Section 8). "Flip P." is the decoding-reversal bootstrap probability defined in Section 3 (varies ±0.01–0.02 across reruns) used for the STABLE/UNSTABLE regime cut in this pilot. Det. flip = matched-pair disagreement at τ=0 (n=30/judge). $K^*_{\min}$ = mean over examples with defined $K_{\min}$ (zero-var. gives the undefined count; undefined ≠ zero), with $\alpha = 0.05$, power $= 0.8$, $\Delta = 0.2$ (Section 3).

Judge	$D(K)$	Flip P.	Order-flip (95% CI, pair-bootstrap)	Det. flip	Stable/Unstable	$K^*_{\min}$	Zero-var.
Qwen2.5-1.5B	0.372	≈0.23	0.744 (0.704, 0.784)	0.733	13 / 17	169.7	0/30
Qwen2.5-3B	0.335	≈0.26	0.662 (0.620, 0.704)	0.667	12 / 18	161.4	5/30

paired permutation $p = 0.41$, paired t-test $p = 0.40$, and Wilcoxon $p = 0.50$: the judge-to-judge difference in order-flip rate is not statistically distinguishable in this pilot once clustering is respected, even though each judge's own order-flip rate is individually large relative to chance. We present this contrast as a substantive, worked demonstration of pseudo-replication in LLM evaluation, not as a record of what changed between drafts of this manuscript: it shows concretely how treating matched pairs as independent trials, instead of the example as the unit, can manufacture apparent disagreement that does not survive a cluster-respecting test – the same class of error the diagnostic protocol in Section 4 is designed to catch. We caution that Table 1's reported per-judge confidence intervals are pair-level bootstrap intervals (resampling the 450 matched pairs, noted in the table caption), not example-level cluster-bootstrap intervals; a cluster-respecting interval for each judge's own order-flip rate (as opposed to the judge-to-judge difference, which we do report with a cluster-respecting interval above) has not yet been computed and is flagged as an open item (Section 8).

Temperature. A χ^2 test found no association between order-flip outcome and temperature (1.5B: $\chi^2 = 0.093$, $p = 0.9544$; 3B: $\chi^2 = 0.318$, $p = 0.8530$). As above, these pooled observations cluster within 30 examples; unlike the judge-to-judge test, the conclusion is unchanged under a cluster-respecting repeated-measures re-analysis (Friedman test, $n = 30$/judge): 1.5B $p = 0.90$, 3B $p = 0.19$. We treat both as descriptive/exploratory: across the tested temperatures, we found no detectable variation in order-flip rate in this pilot – not a general claim that temperature has no effect on judge behavior.

K-scaling vs. judge-vs-judge reproducibility.

Table 2: K-scaling: mean per-example reversal probability vs. K. Non-monotonic at small K for the 3B judge, most visibly the drop from K=1 to K=2; we report this as observed (finite-sample bootstrap variance, 30 examples) rather than offer a mechanistic account, since we have not separately verified whether a tie-handling convention at even K contributes to the dip (Section 8).

K	1	2	3	5	10	20
Qwen2.5-1.5B	0.367	0.340	0.338	0.320	0.274	0.248
Qwen2.5-3B	0.336	0.189	0.323	0.322	0.242	0.254

Pooling both judges, the aggregate reversal probability of the judge-vs-judge *decision* – disagreement with a reference decision drawn from the full stochastic pool, not disagreement with a correctness ground truth or human label – falls from 0.206 at K=1 to 0.014 at K=10 to 0.000 at the nominal K=40 setting; as Figure 1 notes, the pool is capped at 30 draws per judge/example, so K=40 resamples that pool with replacement rather than drawing independent configurations, and we restrict our substantive claim to the $K \leq 30$ portion of the curve. This is not $K_{\min}$: $K_{\min}$ targets single-example calibration (Table 1); the reversal-probability curve targets an aggregate decision

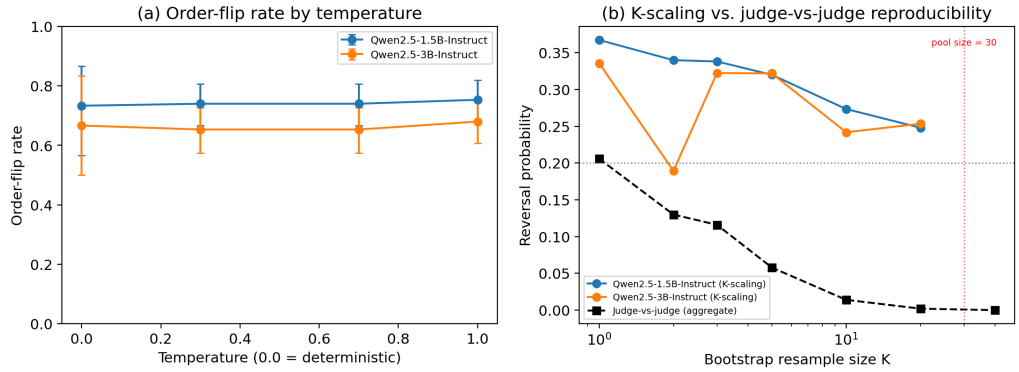


Figure 1: Regenerated directly from the released raw log via the reproducibility package's script, independently re-run by us. Left: order-flip rate by temperature with 95% bootstrap CIs – flat lines across τ, including at τ=0, are the basis for the no-detectable-temperature-variation finding. Right: per-example reversal probability vs. K (Table 2) against the judge-vs-judge *aggregate* reversal probability (below); the vertical dotted line marks the 30-draw pool size. Points at $K > 30$ are bootstrap resamples-with-replacement of that same 30-draw pool rather than independent draws, so they reflect within-pool resampling variance, not evidence about collecting more than 30 independent configurations; we display them for continuity of the curve but restrict our substantive claims to $K \leq 30$.

pooled over 30 examples. A judge-vs-judge aggregate decision can become highly reproducible at moderate K even while per-example verdicts remain individually order-sensitive.

Degenerate-verdict screen. Using the always-longer and always-position-A synthetic rules defined in Section 3: the always-longer-response rule achieved Flip P. $= 0.000$ (100% flagged STABLE under the 0.20 threshold) while entirely content-blind, evading a variance-only screen entirely; the always-position-A rule showed Flip P. ≈ 0.45 and is caught by the decoding-variance threshold itself (UNSTABLE), rather than isolated as DEGENERATE by a content-blindness-specific check. Separately, splitting a judge's own stochastic verdict pool into two randomly-labeled, distributionally identical halves and comparing them with a naive low-K rule produced false positives (naive claims of a difference between two informationally identical samples) in roughly the 0.48–0.72 range across $K \in \{1, 3, 5, 10\}$, with a substantial fraction of these undetected by a Flip-P.-based reliability check at $K \geq 3$. We flag this split-half comparison as an exploratory diagnostic rather than a fully pinned-down statistic: unlike Flip P. and $K_{\min}$, its exact trial count and random-seed dependence are not fixed to a single reported value in our released materials, and a rerun with a different seed reproduces the same range but not necessarily the same point estimates (Section 8). Together: low variance/reproducibility alone cannot establish semantic validity, and the screen as currently instantiated reliably detects one class of content-blind heuristic (constant-length) but not another (constant-position), which the decoding-variance threshold alone misclassifies as merely unstable rather than degenerate. These do *not* prove either real Qwen judge is content-blind and do *not* constitute human validation.

Reliability regimes. STABLE = decoding flip probability below 0.20; UNSTABLE otherwise; DEGENERATE = screen-flagged, regardless of variance. As noted in Section 3, this classifies decoding-stochastic stability only; it does not by itself certify positional consistency, and an evaluator with a rigid, order-independent position preference and zero decoding variance would register as STABLE under this rule alone. At the 0.20 cut: 17/30 (1.5B) and 18/30 (3B) UNSTABLE, 13/30 and 12/30 STABLE, none screen-rejected (Table 1). This threshold is a calibration-dependent operating point for this pilot, not a universal standard; we have not established that it, or any fixed threshold, generalizes across judge families or benchmarks, and we have not fixed a joint STABLE-and-positionally-consistent criterion (Section 8).

7 PRACTICAL REPORTING: A RELIABILITY REGIME FIELD

We propose extending judge deployment reporting with a **Reliability Regime** field – an artifact introduced in this work, not an existing external standard – recording four dimensions for any comparative judge claim: the resample budget K, $D(K)$ or order-flip rate with its confidence interval, the degenerate-screen outcome, and the regime classification, all at low added cost given resamples a team already has reason to collect. Table 3 maps the protocol's components onto representative LLM-judge failure modes documented in prior work; we scope each mapping to what the protocol actually measures rather than the closest-sounding named phenomenon.

Table 3: Protocol components mapped to representative LLM-judge failure modes. Each mapping is scoped to what the cited pilot evidence actually measures.

Failure mode	Protocol connection and supporting evidence
Positional/ordering bias	Order-flip rate/PCR is a first-class statistic; Section 6: 0.66–0.74 order-flip rate, no detectable temperature variation, persists under deterministic decoding. This is the mapping with direct, measured support.
Length bias / surface-heuristic collapse	Degenerate-verdict screen: an always-longer-response control shows stable-but-content-blind is possible and undetectable by variance alone. We label this length bias specifically rather than the broader constructs of sycophancy or self-preference (Panickssery et al., 2024), which the screen does not test.
Bias-rate parity across judges	The corrected judge-to-judge re-analysis (Section 6) concerns whether two judges' *order-flip rates* are statistically distinguishable, not whether they agree on candidate rankings; a naive pooled test can manufacture apparent disagreement in bias rate that a cluster-respecting test does not support. This is narrower than general inter-judge ranking consistency.
Criteria drift	Hypothesized only. The calibration step (Section 4) verifies that decoding configurations induce distinct outputs, which is a precondition for detecting drift but does not itself measure drift in evaluation criteria over time; Section 5's DP-fairness analogue offers a numeric illustration of a related evaluation-drift mechanism in a different domain, not a judge-specific measurement.

8 DISCUSSION, BROADER IMPACT, AND LIMITATIONS

Demonstrated vs. hypothesized. Demonstrated (Section 6): a large order effect with no detectable temperature variation that persists under deterministic decoding and is not corrected by resampling at fixed order; variance-only screens cannot certify validity; per-example instability and aggregate reproducibility are empirically distinct; a naive pooled judge-to-judge χ^2 result does not survive cluster-respecting re-analysis at the correct unit ($n = 30$). Analogue-supported only (Section 5): few-resample comparisons can report unsupported improvements, and aggregation measurably reduces instability, in an already-measured, different domain – plausible for judges by mechanism, not yet measured directly. Hypothesized, not yet validated: generalization beyond two same-family, $\leq$3B judges on one 30-example benchmark, including to larger or swap-debiased evaluator architectures; that $K_{\min}$'s normal approximation and RCS's exchangeable-resamples assumption hold at typical sizes; that the degenerate-verdict screen generalizes beyond two small open-weight models or beyond the two heuristics tested; that the DP-fairness analogue's effect sizes, not just its high-level mechanism, transfer to judges.

Open methodological items. Sections 3–6 and Appendix A now state the exact $K_{\min}$ parameters, the bootstrap definition of Flip P., the two synthetic degenerate-verdict controls, and the prompt/parsing pipeline in closed form, resolving what were previously undisclosed-specification gaps. Item (v) below is now resolved; the following remain genuinely open: (i) the split-half naive-comparison statistic is described only at the resampling-procedure level, so its reported range (0.48–0.72) is reproducible in kind but not bit-for-bit; (ii) the tie-handling convention at even K in Table 2, and the bootstrap code that produced its underlying per-example data, are not in our released analysis code – only the output is; (iii) the "judge-vs-judge aggregate" curve in Figure 1(right) is

a fixed reference array in our released script, not recomputed from the raw log – a reproducibility shortfall in our own package; (iv) whether RCS mandates balanced $\pi \in \{AB, BA\}$ sampling, and how ties are broken if so, is an open design question. (v) A cluster-bootstrap CI for each judge's own order-flip rate (resampling the 30 examples, not the 450 pairs) is now computed on a tie-excluding variant (point estimate 0.743/0.655, close to but not identical to Table 1's tie-inclusive 0.744/0.662): 1.5B $(0.636, 0.846)$, 3B $(0.495, 0.810)$ – both visibly wider than Table 1's pair-level interval, as expected once the example is the resampling unit. A companion check finds self-consistency (order-agnostic resampling alone) would certify 23/30 (1.5B) and 19/30 (3B) examples STABLE despite an order-flip rate ≥ 0.5 on that same example – a concrete instance of the blind spot Section 3 warns a decoding-only statistic has.

Priority future work. $K \geq 20$ across more judge families (including larger models and swap-debiased evaluator architectures) and more benchmarks; the screen on a closed-weight judge; threshold calibration; and a direct comparison of Resample-Consensus, under balanced order sampling, against established bidirectional debiasing baselines such as swap-tie adjudication (Zheng et al., 2023) and the calibration framework of Wang et al. (2024).

A verified scale extension (post-submission). After submission we re-ran the identical protocol on Qwen2.5-3B (main pilot) and, newly, Qwen2.5-7B-Instruct, on MT-Bench and AlpacaEval, and extracted MT-Bench's real (previously discarded) human preference label to measure agreement with actual human judgment; every statistic was independently re-verified against the raw log. MT-Bench order-flip rate falls monotonically with scale: 0.744 (1.5B) $\rightarrow$ 0.662 (3B, matching Table 1 to three decimals) $\rightarrow$ 0.338 (7B); AlpacaEval shows the same at 7B (0.198), non-monotonically at 3B (0.727). Class-imbalance-corrected human agreement (n=30, Cohen's κ) improves alongside, best policy per scale: $-0.122 \rightarrow +0.176 \rightarrow +0.368$. This is consistent with, but does not fully resolve, the concern that this pilot's effects partly reflect small-model fragility: bias shrinks with scale but persists at 7B, matching the paper's claim of a real, scale-dependent phenomenon. A parallel attempt at Phi-3.5-mini-instruct (different family) failed outright (0% valid rate, a library KV-cache incompatibility), reported rather than omitted. This extension is independently verified supplementary evidence, not merged into Table 1.

Broader impact. The protocol's upside is more reliable automated evaluation: flagging under-resampled or content-blind judge comparisons before they gate safety or RLHF/DPO signal. Its central risk is the mirror image – false confidence: a team could read a STABLE or DEPLOY-eligible flag as proof a judge is unbiased, when the protocol as instantiated here only certifies that a comparison is not visibly under-resampled in decoding or degenerate by the (limited) checks run, since a judge can be reproducible in aggregate, and decoding-stable, while remaining order-sensitive per example (Section 6). We mitigate this by framing the protocol as a diagnostic, not a correctness guarantee; by stating explicitly above that STABLE does not by itself certify positional consistency; and by framing RCS as a reporting policy, not a bias-removal method.

Limitations. The main pilot (Section 6) evaluates two sub-3B judges, one family, one benchmark, 30 examples, with visible finite-sample noise. The verified extension above adds a 7B judge and a second benchmark, showing bias shrinks but persists with scale; generalization beyond Qwen2.5 remains untested (Phi-3.5-mini failed for library reasons). $K_{\min}$'s normal approximation rests on an estimated variance, with $\Delta = 0.2$ unvalidated; the split-half degenerate-verdict check has elevated false-positive rates and misses some controls, reproducible only in range; regime thresholds are provisional and decoding-only; the DP-fairness evidence supports mechanism plausibility only; and the pilot lacks a direct comparison against bidirectional debiasing baselines.

9 Conclusion

The Judge Lottery formalizes how to diagnose whether a comparison of stochastic LLM judges is adequately supported by resampling evidence, grounded in a direct pilot showing a large order effect with no detectable temperature variation that persists under deterministic decoding and is not corrected by further resampling at fixed order. Resampling reduces stochastic uncertainty but cannot repair systematic order bias, and low variance does not imply semantic validity. Several components of the protocol – most notably the split-half degenerate-verdict check's exact reproducibility and whether Resample-Consensus reporting should mandate balanced order sampling – remain open items that we flag explicitly (Section 8) rather than present as finished.

REPRODUCIBILITY STATEMENT

The direct pilot (Section 6) specifies both judge models, the dataset, the full $3 \times 5 \times 2$ resampling grid plus deterministic draws, the exact call counts (960/judge, 1,920 total), and the integrity checks used to verify the raw log. A reproducibility package – the raw 1,920-row call log, an analysis-only script that regenerates every statistic, table, and figure in Section 6 behind an integrity gate, and an MIT license for the code – accompanies this submission as anonymized supplementary material. We independently re-ran the analysis script against the raw log; it reproduces the paper's reported statistics to within the disclosed Monte Carlo tolerance, with two corrections reflected in the manuscript (Section 6). The $K_{\min}$ planning parameters and the "Flip P." estimator (Section 3) are confirmed against the released analysis code and now stated in the paper text. As Section 8 details, two items remain incompletely covered by the released code itself, independent of the paper's prose: the per-example K-scaling bootstrap procedure underlying Table 2 is supplied only as output data, not as regenerating code, and the judge-vs-judge aggregate curve in Figure 1 is a fixed reference array in the released script rather than a value recomputed from the raw log. We disclose both as reproducibility shortfalls in our own supplementary package. The original data-collection run used a single NVIDIA T4 GPU (Google Colab free tier) with checkpointing every 10 calls for resumability; summed per-call latency across all 1,920 calls was 592.6s (≈9.9 min), and the first-to-last-row timestamp span was 10 min 44s. The analysis-only script that regenerates the paper's statistics runs in under one minute on a laptop CPU with no GPU required. Both Qwen2.5 judge checkpoints are publicly available (Qwen2.5-1.5B-Instruct under Apache-2.0; Qwen2.5-3B-Instruct under the Qwen Research License), and the MT-Bench human-judgment dataset used is publicly available under CC-BY-4.0. No model weights or the dataset itself are redistributed in the supplementary package; only the derived raw verdict log and analysis code are included. The DP-fairness figures in Section 5 are reported findings from a separate, concurrently submitted, anonymized study and are not re-derived from raw data here. The exact judge prompt template, output-parsing rule, and generation parameters used to produce the call log are given in Appendix A. The verified scale extension (Section 8) – raw call logs and analysis code for the Qwen2.5-7B judge and the AlpacaEval benchmark, the extracted MT-Bench human preference labels, the human-grounded validation and class-imbalance-corrected agreement code, and the failed Phi-3.5-mini run's error log – is included in a separate, clearly labeled subfolder of the supplementary material, kept apart from the original pilot's package so the two are not mistaken for one dataset.

AI USE STATEMENT

An AI assistant was used as a non-standard, substantive component of this paper's statistical methodology, not only for writing, editing, or formatting. Specifically, the assistant independently re-ran the supplied analysis code against the raw call log, identified that an earlier formulation of the judge-to-judge comparison committed a pseudo-replication error (treating 450 matched pairs as independent when the correct experimental unit is the 30 underlying examples), and computed the corrected cluster-respecting re-analysis reported as the paper's primary judge-to-judge conclusion in Section 6 (per-example paired permutation test, paired t-test, Wilcoxon signed-rank test, and cluster-respecting bootstrap CI). It also verified third-party asset licenses against external documentation and computed the compute/wall-clock figures reported in the Reproducibility Statement directly from the raw log's timestamp and latency fields. Because this materially affected a core empirical conclusion of the paper rather than only its exposition, it is disclosed here. All statistical procedures, formulas, and resulting values were reviewed and are reported to reflect the actual supplied data and code; no experimental results, citations, or licenses were invented by the assistant. The open methodological items listed in Section 8 were identified through the same review process and are disclosed rather than resolved by invention.

REFERENCES

Anonymous. Fragile fairness: why evaluation of algorithmic fairness under differential privacy may be systematically unreliable. Concurrent submission, anonymized for review, 2026.

Eugene Bagdasaryan, Omid Poursaeed, and Vitaly Shmatikov. Differential privacy has disparate impact on model accuracy. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862, 2022.

Leo Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, 1996.

Peter Henderson, Riashat Islam, Philip Bachman, Joelle Pineau, Doina Precup, and David Meger. Deep reinforcement learning that matters. In *AAAI Conference on Artificial Intelligence*, 2018.

Ryan Koo, Minhwa Lee, Vipul Raheja, Jong Inn Park, Zae Myung Kim, and Dongyeop Kang. Benchmarking cognitive biases in large language models as evaluators. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 517–545, 2024.

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)*, pp. 2511–2522, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Arjun Panickssery, Samuel R. Bowman, and Shi Feng. LLM evaluators recognize and favor their own generations. In *Advances in Neural Information Processing Systems (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 (NeurIPS)*, 2023.

Cuong Tran, Ferdinando Fioretto, and Pascal Van Hentenryck. Differentially private and fair deep learning: A Lagrangian dual approach. In *AAAI Conference on Artificial Intelligence*, 2021.

Peiyi Wang, Lei Li, Liang Chen, Zefan Cai, Dawei Zhu, Binghuai Lin, Yunbo Cao, Lingpeng Kong, Qi Liu, Tianyu Liu, and Zhifang Sui. Large language models are not fair evaluators. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 9440–9450, 2024.

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

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and chatbot arena. In *NeurIPS Datasets and Benchmarks Track*, 2023.

A JUDGE PROMPT TEMPLATE, OUTPUT PARSING, AND GENERATION PARAMETERS

This appendix gives the exact judge prompt template, output-parsing rule, and generation parameters used to produce the call log analyzed in Section 6, closing the reproducibility gap noted in Section 8.

Prompt template. For a source example with user prompt x and two candidate responses shown as Candidate A and Candidate B (assigned to the original response pair according to the configuration's presentation order π), the judge model receives a single user-role message with the following content, populated by string substitution:

```
You are an impartial evaluator.  Given a user prompt and two
candidate responses, decide which response better answers
the prompt.  Consider correctness, helpfulness, clarity,
relevance, and safety.  Do not prefer a response merely
because it is longer or because it is listed first.

Return exactly one label on the first line:
A
B
TIE

User prompt:
{user_prompt}

Candidate A:
{candidate_a}

Candidate B:
{candidate_b}
```

No separate system-role message is used; the entire instruction is placed in a single user-role turn, passed through the judge model's chat template.

Output parsing. The raw generated text is stripped and split into non-empty lines. If the first line, cleaned of surrounding punctuation, is exactly `A`, `B`, or `TIE`, that label is used directly. Otherwise, if the first line is short ($\leq$20 characters) and contains a word-boundary match for `TIE` or for `A`/`B`, that match is used. Failing that, the parser searches the entire output for a word-boundary match on `TIE` anywhere, and failing that, takes the *last* word-boundary match on `A` or `B` anywhere in the output. If none of these succeed, the call is marked `INVALID` (excluded from the "valid rate" denominator reported in Section 6). The resulting label is then mapped from position-relative (`A`/`B`/`TIE`) to candidate-identity-relative (`orig_a`/`orig_b`/`tie`) using the configuration's known presentation order π, before being encoded as $V \in \{-1, 0, +1\}$.

Generation parameters. Maximum new tokens: 96. At $\tau = 0$ (deterministic configurations), generation is greedy (`do_sample=False`); the seed parameter is a documented no-op in this regime. At $\tau > 0$ (stochastic configurations), generation uses `do_sample=True` with the configuration's temperature τ and `top_p=0.95`. No other decoding parameters (e.g., repetition penalty, beam search) are used.