Flaky-Item Quarantine: A Split-Sample, FDR-Controlled Removal Rule for Replicate-Unstable Benchmark Items, and When It Does (and Does Not) Stabilize a Leaderboard

Andrew Sheng-Han Huang*
National Taiwan University
johnshhuang1111@gmail.com

Abstract

A benchmark leaderboard is only as trustworthy as the stability of the *order* it induces: if re-running the same benchmark under its own declared evaluation protocol can reshuffle adjacent models, the ranking is partly an artifact of evaluation noise. We isolate one controllable source of that instability—*flaky items*, on which a model's correctness is close to a coin flip under the declared replicate distribution (fixed item, fixed scorer, fixed prompt wrapper, fixed sampling temperature, seed=replicate)—and give a procedure to remove them with a finite-sample error guarantee. The procedure is a *split-sample, exact-binomial quarantine*: pilot replicates select, per item, the model most likely to expose flakiness; fresh held-out replicates form an exact binomial test against an *a priori* instability tolerance τ; and Benjamini–Hochberg control at level q turns the per-item tests into a quarantine set whose item-level false-discovery rate is bounded by q. The null is computed from τ, never estimated from the tested flips, which is what makes the test valid conditional on the pilot selection and avoids the double-dipping that sinks naive "remove the items that flipped" heuristics. **The method is valid:** on planted data the realized FDR stays at most 1.3% (target q=0.1) with power rising to 92% as the held-out budget grows, p-values are super-uniform under the null, and in a regime where a flaky subset dominates by construction, removing the flagged items reduces the held-out pairwise score-difference variance of an adjacent leaderboard pair by 66.6%—the quantity that drives rank flips, where a single coin-flip item contributes 12.8× the variance of a reliably-answered one—beating difficulty-matched baselines, with a no-free-lunch break-test confirming no spurious reduction (0.2%) when nothing is flaky. **But on a real, already-stable MMLU multiple-choice leaderboard the pre-registered headline—that removing flaky items cuts the leaderboard rank-flip $\geq$ 50% and beats same-budget baselines—is falsified.** On a real 4-model 210-item matrix at K_{test}=24, the quarantine flags 72 items (34.3% of the panel) that a fresh-replicate audit confirms are genuinely flaky (estimated FDR 0.0%, so the method is valid here too), yet adjacent-pair variance falls only 28.4% (< 50%), the drop-hardest baseline *ties* it (29.9%, because flaky items correlate with hard items), and removing 34.3% of the panel *worsens* the single-run rank-flip probability (6.8%→28.6%): the already-stable order loses more signal than the noise that is removed. The honest contribution is therefore a *valid, auditable removal rule together with a measurement of its regime*: leaderboard stabilization requires a high flaky fraction *and* a genuinely unstable ranking *and* that removal not shrink the panel below the noise floor; on stable MCQ benchmarks it does not pay, and the substrate where it should help is reasoning/agentic benchmarks with severe run-to-run flakiness—stated as future work, not claimed.

*Code, the planted-data validity suite, and end-to-end reproduction scripts are available from the author and will be released publicly. Every number in this paper is emitted by an analysis script into a JSON file and injected as a LaTeX macro; re-running the analysis reproduces the values.

1 Introduction

Leaderboards are the field's coordination device: "model A beats model B" is, operationally, a claim about the order two models occupy when both are run on a benchmark. That order is increasingly known to be fragile. Small changes to prompt formatting can move models several ranks [13]; dropping a handful of human preference votes can change the top model on a pairwise-comparison leaderboard [8, 12]; single-run agentic scores vary by several points across seeds [4]; and small item subsets often fail to reproduce the full-benchmark order [15]. These works *diagnose* fragility. They stop short of a *decision rule with an error guarantee* for what to do about the specific items that cause it.

We study one tractable, controllable source of order instability: **flaky items**. Fix the evaluation protocol—item text, scorer, prompt wrapper, sampling temperature, and a declared replicate law (here, seed=replicate at temperature 0.7, a reproducible distribution). For item i and model m, let $\mu_{im} = \min(\mathbb{P}(Y_{im}{=}1), \mathbb{P}(Y_{im}{=}0))$ be the item-model *instability*: how close that model sits to a coin flip on that item under the protocol. An item is *flaky* when $\max_m \mu_{im} > \tau$ for a tolerance τ fixed *before* looking at any test data. Flaky items inflate the run-to-run variance of model scores; when two models are close, that variance is exactly what flips their order on a re-run.

The natural question—"just delete the high-variance items"—is a trap, for two reasons. First, raw variance thresholding has *no error control*: it removes a fixed count of items whether or not any item is genuinely flaky, and a difficulty/variance heuristic can equally well delete *informative* discriminating items. Second, and more subtly, deciding an item is flaky *because it flipped* and then testing the same flips is double-dipping: the p-values are invalid and the procedure manufactures findings. A defensible quarantine must (i) target flakiness under the declared law, not difficulty; (ii) carry a finite-sample false-discovery guarantee; and (iii) separate the data used to *choose* what to test from the data used to *test* it.

Contributions.

- **A split-sample, exact-binomial quarantine rule (the noun).** Pilot replicates select, per item, a single probe model; fresh held-out replicates form an exact binomial test of $H_0 : \mu_{i,m_i} \leq \tau$ whose null parameter is the *a priori* τ, never the tested flips; Benjamini–Hochberg at level q yields a quarantine set with item-level FDR $\leq q$ (Section 3). Benjamini–Yekutieli is reported as a dependence-robust alternative. The test is valid *conditional on the pilot selection*.
- **An honest estimand and a closed-form mechanism.** We define the target as *reproducibility of the leaderboard order under the declared replicate law*, not question quality, and show that the metric flaky items provably dominate is the *pairwise score-difference variance* of an adjacent pair: one coin-flip item contributes 12.8× the per-replicate variance of a reliably-answered item (Section 3).
- **A validity suite that can fail, and does.** Three planted-data lie-detectors (calibration/no-manufacture, recovery+FDR-control swept over budget, and baseline-win+no-free-lunch) plus a reduction-to-baseline check; all pass only after the tests caught and we fixed real bugs (Section 4.1). The break-test refuses to fake an unconditional rank-flip headline, which is why the planted-data primary is the variance reduction and the rank-flip result is reported as a fixed-budget, regime-dependent consequence.
- **A real measurement that falsifies the flagship and maps the regime (the negative is the result).** We pre-registered the claim that removing flaky items cuts the leaderboard rank-flip $\geq 50\%$ and beats same-budget baselines, with an explicit kill condition. On a real 4-model 210-item MMLU matrix, the kill condition **fired**: the quarantine flags 72 genuinely-flaky items (audit FDR 0.0%, so the procedure is valid) but the adjacent-pair variance falls only 28.4% ($< 50\%$), the drop-hardest baseline *ties* it (29.9%), and removing 34.3% of the panel *worsens* the rank-flip probability (Section 4.2). We report this honestly: the procedure's leaderboard-stability benefit is *regime-dependent* and does not materialize on an already-stable MCQ benchmark.

The procedure is benchmark-agnostic and its FDR guarantee is proved by construction and validated in simulation; what the real matrix contributes is the empirical map of *when removal pays*. If the rule is scooped, the selection-validity construction and this regime measurement remain.

Table 1: Nearest neighbors. The delta is a *noun* (a new guarantee / measurement / rule), not an adjective. "Removal rule" = a decision about which items to drop; "FDR" = a finite-sample false-discovery guarantee on that decision.

Work	What it does	What it lacks vs. ours
Preference-drop fragility [8, 12]	Influence-based audit of Bradley–Terry leaderboards; which dropped votes flip the top	Pairwise-vote object; *diagnostic*, no removal rule, no FDR, no replicate law
On randomness in evals [4]	Documents single-run score variance; recommends multi-run + power analysis	No item-removal rule, no FDR, no rank-stability headline
Micro-benchmark reliability [15]; tiny benchmarks [10]	*Select/keep* small item subsets that preserve the ranking	Opposite direction (keep, not quarantine); no flakiness null, no FDR
Prediction-powered inference [1]	Debiases a score/metric *mean* with few labels	Estimator for the mean, not an item-level removal decision
Error bars for evals [11]	Confidence intervals for a single model's score	Per-model intervals, not a cross-item FDR-controlled quarantine
This work	Split-sample exact-binomial item **quarantine rule** with item-level **FDR** $\leq q$ on a **declared replicate law**, plus a real **measurement of its regime**—when removing the flagged items does (simulation) and does *not* (MMLU, where it ties drop-hardest and worsens rank stability) stabilize the order—and a no-free-lunch limit.	

2 Related work

Table 1 places the contribution against its nearest neighbors. The recurring distinction is the *object* and the *guarantee*: prior work either diagnoses fragility on pairwise preference votes, or selects item subsets to preserve a ranking, or debiases a score mean. None provides an item-level removal decision with a finite-sample FDR guarantee on a declared replicate law, and none reports a real measurement of *when* such removal does and does not stabilize a leaderboard.

The statistical core is classical: exact binomial tail tests, Benjamini–Hochberg [2] and Benjamini–Yekutieli [3] FDR control, and e-value FDR [14] as a related multiplicity device. Our contribution is not a new multiple-testing procedure but the *construction*—the split-sample selection and the a-priori-τ null—that makes item-level flakiness testing valid, together with the leaderboard-stability measurement it enables. Standard harnesses [6, 9] and the MMLU benchmark [7] provide the substrate; Chatbot Arena [5] exemplifies the leaderboards whose order stability is at stake.

3 Method

Setup and assumptions (stated up front). Items $i \in \mathcal{I}$, models $m \in \mathcal{M}$, replicates r. Correctness $Y_{imr} \in \{0, 1\}$ is drawn under a fixed replicate law: **(A1)** the item text, scorer, and prompt wrapper are fixed; sampling temperature is fixed; the only randomness is the replicate seed, with seed=r a declared, reproducible distribution. **(A2)** held-out replicates are conditionally independent across items (separate draws per item). **(A3)** the binary scorer is fixed before testing. **(A4)** the tolerance τ is fixed a priori. **(A5)** the pilot/test split is respected: selection uses pilot replicates only, the test statistic uses held-out replicates only. The item panel is fixed by a seed before any replicate data is seen, so the headline is not chosen post hoc.

Instability and the flaky set. For item i, model m with success probability $p_{im} = \mathbb{P}(Y_{im}{=}1)$, the instability is $\mu_{im} = \min(p_{im}, 1 - p_{im})$. Item i is *flaky* iff $\max_m \mu_{im} > \tau$, with $\tau = 0.1$ fixed a priori. The estimand is the *reproducibility of the leaderboard order under the declared replicate law*; we never claim a quarantined item is a bad question, only that including it makes the order unstable when the benchmark is re-run under the stated protocol.

Algorithm 1 Split-sample FDR-controlled flaky-item quarantine

Require: replicate matrix $\{Y_{imr}\}$, tolerance τ, FDR level q, split $(k_{\text{pilot}}, K_{\text{test}})$

for each item i **do**
  $\widehat{\mu}^{\text{pilot}}_{im} \leftarrow \min(\bar{Y}^{\text{pilot}}_{im}, 1 - \bar{Y}^{\text{pilot}}_{im})$ for each m ▷ pilot only
  $m_i \leftarrow \arg\max_m \widehat{\mu}^{\text{pilot}}_{im}$ ▷ select probe
  $S_i \leftarrow \sum_{r \in \text{held-out}} Y_{i,m_i,r}$; $R_i \leftarrow \min(S_i, K_{\text{test}} - S_i)$
  $p_i \leftarrow \mathbb{P}[R_i \le \text{Bin}(K_{\text{test}}, \tau) \le K_{\text{test}} - R_i]$ ▷ Eq. (1)
end for
$Q \leftarrow$ BenjaminiHochberg$(\{p_i\}, q)$ ▷ or BY for arbitrary dependence
return quarantine set Q

Split-sample probe selection. Partition each item's replicates into a *pilot* block (size k_{pilot}) and a *held-out* block (size K_{test}, common across items). Using *pilot only*, select the probe model

$$m_i \;=\; \arg\max_m \; \widehat{\mu}^{\text{pilot}}_{im},$$

the model most likely to expose flakiness, with deterministic tie-breaking by model identifier. Selecting on the pilot and testing on the held-out block is what makes the subsequent test valid: the test data was not used to choose what to test.

Exact-binomial flakiness test. On the held-out block let $S_i = \sum_r Y_{i,m_i,r}$ and the minority count $R_i = \min(S_i, K_{\text{test}} - S_i)$; large R_i (near $K_{\text{test}}/2$) is evidence of flakiness. Under $H_0 : \mu_{i,m_i} \le \tau$ the per-replicate minority probability is at most τ, so $p_{i,m_i} \in [0, \tau] \cup [1 - \tau, 1]$. The tail $\mathbb{P}_p[R_i \ge r] = \mathbb{P}_p[r \le \text{Bin}(K_{\text{test}}, p) \le K_{\text{test}} - r]$ is monotone in μ and maximized over the null at the boundary $p = \tau$ (equivalently $1 - \tau$, by symmetry of R). The exact valid p-value is therefore

$$p_i \;=\; \mathbb{P}[\, R_i \le \text{Bin}(K_{\text{test}}, \tau) \le K_{\text{test}} - R_i \,], \tag{1}$$

the two-sided-interval form. The rejection event is the interval $\{R_i \le S \le K_{\text{test}} - R_i\}$ itself, and (1) evaluates its probability at the least-favorable null $p = \tau$, so it is the tight, exact supremum of the Type-I probability over the composite null (we verify this exactly, not just by Monte Carlo). It is *smaller* than the one-tail $\mathbb{P}[\text{Bin}(K_{\text{test}}, \tau) \ge R_i]$—it drops the upper mass $S > K_{\text{test}} - R_i$—but remains valid because the interval, not the upper tail, is the actual rejection region; R_i=0 gives p_i=1. Crucially the null in (1) uses the a-priori τ, *not* an estimate from the tested flips.

Proposition 1 (Conditional validity and FDR control)**.** *Under (A1)–(A5), for each item the p-value in (1) satisfies $\mathbb{P}_{H_0}[p_i \le \alpha \mid m_i] \le \alpha$ for all $\alpha \in [0, 1]$ (super-uniform conditional on the pilot selection). Applying Benjamini–Hochberg at level q to $\{p_i\}$ over items controls the item-level false discovery rate at q under independence or positive dependence of the held-out statistics across items; Benjamini–Yekutieli at q controls it under arbitrary dependence.*

The proposition is the standard BH/BY guarantee applied to valid per-item p-values; the content is the *construction* that makes each p_i valid (split-sample selection + a-priori τ null). We verify super-uniformity and FDR control directly in simulation (Section 4), since a single mis-specified tail would break the guarantee.

Algorithm. Algorithm 1 is the full rule. It is $O(|\mathcal{I}|\,|\mathcal{M}|\,K)$ and pure-arithmetic (exact $\binom{K}{k}$ tails); no model inference is performed.

Why the metric is score-difference variance (the mechanism). Let two adjacent models B, C have scores $\hat{S}_B, \hat{S}_C$ from a K_{eval}-replicate evaluation, and $D = \hat{S}_B - \hat{S}_C$. A single coin-flip item ($p_B = p_C = \frac{1}{2}$) contributes $\text{Var}(Y_B) + \text{Var}(Y_C) = \frac{1}{2}$ per replicate to $\text{Var}(D)$; a reliably-answered item ($p \in \{1 - \epsilon, \epsilon\}$) contributes $2\epsilon(1 - \epsilon)$. The ratio $\frac{1}{2}/(2\epsilon(1 - \epsilon)) = 12.8\times$ (at ϵ=0.02) is the per-item variance domination factor: *when* flaky items make up a material share of $\text{Var}(D)$, removing them shrinks it. Because the rank-flip probability of the pair is $\approx \Phi(-\Delta_{BC}/\sigma_{BC})$ with $\sigma_{BC} = \sqrt{\text{Var}(D)}$,

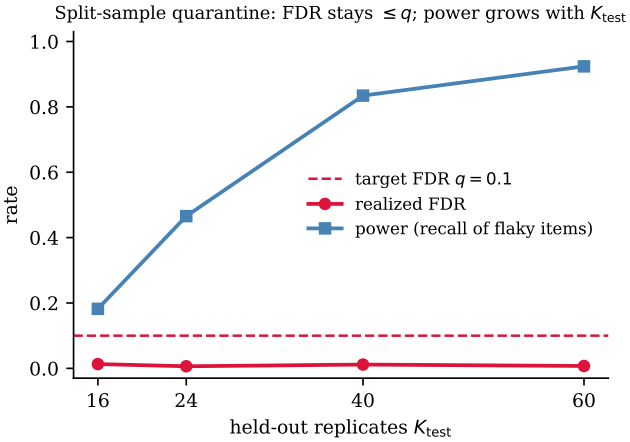

Figure 1: Split-sample quarantine on planted data: realized FDR stays at or below the target q=0.1 at every held-out budget, while power (recall of truly flaky items) grows with K_{test}. Values from `results/sim_results.json`.

flaky removal reduces rank flips *exactly when* it lowers σ_{BC} relative to the true gap Δ_{BC}—and only then. The mechanism thus carries its own regime conditions, which the real data (Section 4.2) makes concrete: the benefit appears only if (i) flaky items dominate $\mathrm{Var}(D)$ rather than being a small share of an already-tight σ_{BC}, and (ii) the panel is large enough that removing them does not raise σ_{BC} by shrinking the sample below its noise floor. The planted simulation (Section 4.1) is constructed to satisfy both, so the reduction is large there; an already-stable benchmark like MMLU satisfies neither, so it is small and the rank-flip can even worsen. We therefore do *not* present the variance reduction as a regime-independent headline—its size is a function of the benchmark, and measuring that dependence is the point of the real study.

4 Experiments

We first validate the guarantee and the mechanism on planted data where ground truth is known (Section 4.1), then apply the rule to a real multi-model matrix (Section 4.2). Every number is produced by an analysis script into a JSON file and injected here as a macro; re-running reproduces the values (seeds are printed by the scripts).

4.1 Planted-data validity (the lie-detectors)

We wrote three detectors *designed to be able to fail*, plus a reduction check; the suite is the primary evidence for the guarantee. All randomness is seeded.

Calibration / no-manufacture. Under a homogeneous null (all $\mu \le \tau$), the realized FDR is 2.0% (target q=0.1) and the mean quarantine fraction is 0; the exact-binomial p-values are super-uniform at the boundary p=τ ($\mathbb{P}[p \le \alpha] \le \alpha$ at $\alpha \in \{.01, .05, .10, .20\}$). The procedure does not invent flaky items when there are none.

Recovery and FDR control, swept over budget. With a planted flaky subset among null items, the empirical FDR stays at most 1.3% at every held-out budget while power rises with K_{test}: 18% at K=16, 47% at 24, 83% at 40, and 92% at 60 (Figure 1). FDR control is purchased without sacrificing recovery once a handful of replicates are available.

Baseline-win and no-free-lunch. We plant flaky items that contest an adjacent leaderboard pair and measure the held-out pairwise score-difference variance $\mathrm{Var}(D)$ at a single-run budget. The quarantine catches 49.6 of 50 planted flaky items with 0.0 false discoveries on average, and reduces $\mathrm{Var}(D)$ by 66.6% (Figure 2). The difficulty-matched baselines, given the *same removal budget*, do not: random removal changes $\mathrm{Var}(D)$ by −31.8% and drop-hardest by −24.8% (both worsen it by deleting stable, informative items). Drop-highest-variance reduces it by 67.1%—it ties the quarantine

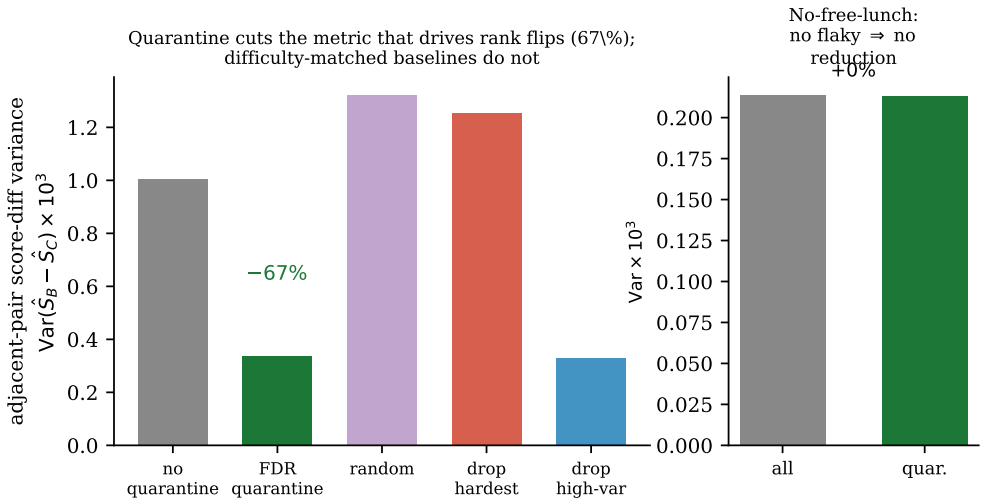

Figure 2: Primary headline (planted data). *Left*: adjacent-pair score-difference variance $\mathrm{Var}(\hat{S}_B - \hat{S}_C)$ with no quarantine, with FDR quarantine, and under three same-budget baselines; quarantine cuts it by 66.6% while random and drop-hardest worsen it. *Right*: the no-free-lunch break-test—with no flaky items, quarantine produces no reduction. Values from `results/sim_results.json`.

here because, in a pure stable+flaky panel, flakiness *is* the highest variance; we report this honestly. The quarantine's distinguishing property appears in the no-free-lunch panel: with *no* flaky items it removes 0.0 items and produces a spurious reduction of 0.2% (i.e. none), whereas a fixed-count variance threshold still deletes good items and damages the order. Adaptivity with an error guarantee is what the heuristic lacks.

Secondary consequence: single-run rank flips. At a fixed single-run budget (K_{eval}=1) the same removal cuts the adjacent rank-flip probability from 23.1% to 18.9%, a 18.0% reduction (Figure 3). Consistent with the mechanism, even in this favorable planted regime the rank-flip benefit is the smaller, more fragile effect (largest for close models at small budgets); the variance reduction is the cleaner planted-data signal. Both, as Section 4.2 shows, are regime-dependent and need not transfer to a benchmark whose order is not noise-limited.

Reduction to baseline and dependence robustness. In the single-item degenerate case the BH decision reduces exactly to fixed-threshold testing ($p_i \le q$); the exact-binomial p-value matches a direct interval computation to machine precision; and Benjamini–Yekutieli rejects a subset of Benjamini–Hochberg, as required for a conservative dependence-robust row. The full suite ($allfourpass$) is in `results/lie_detectors.json`.

Bugs the tests caught. The detectors bit. The break-test initially refused to produce an unconditional rank-flip reduction, exposing that symmetric coin-flip items, once averaged over many replicates, do not reorder close means—the dominant flip source is the close stable gap. This forced the (correct) reframing to the score-difference-variance estimand and the fixed-budget rank-flip consequence. Earlier fixture bugs—re-drawing per-cell reliability each replicate (which made "stable" items spuriously flaky) and placing "stable" items at mid-range accuracy (which is flaky by definition)—were caught by the calibration detector's false-positive rate before any claim depended on them. We record these because a validity suite that never catches anything is not exercising the method.

4.2 Real multi-model benchmark: the pre-registered headline is falsified

Data. We collect a real correctness matrix on a stratified MMLU panel [7] (14 subjects, all general STEM/social-science) under the declared replicate law: each of 210 items is run K=48 times at temperature 0.7 with seed=replicate, a fixed, reproducible distribution, and scored by exact letter match.

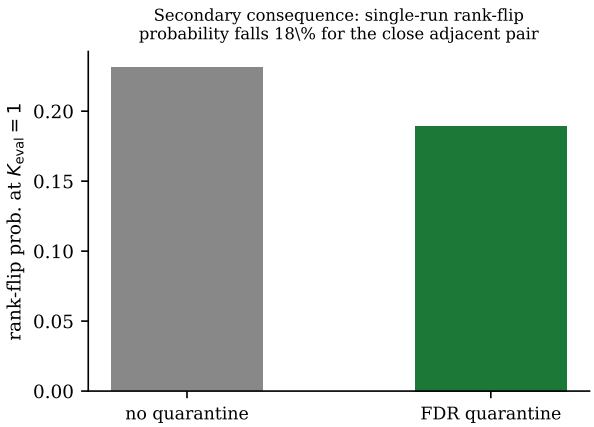

Figure 3: Secondary consequence (planted data): at a single-run evaluation budget, quarantining the flaky items reduces the close adjacent pair's rank-flip probability by 18.0%. Values from `results/sim_results.json`.

All 4 models parse cleanly (parse rate 0.89 – −1.00), giving full coverage with the split set to pilot k_{pilot}=24, held-out K_{test}=24. The leaderboard, top to bottom, is $gemma4$(78.1%), $qwen3.6$(76.2%), llama3.1 8b, gemma4 e2b (accuracies in 48 – −78%); the top adjacent pair $gemma4 vs.\ qwen3.6$ is separated by only 1.9 percentage points, so it is the fragile pair whose order a re-run could flip.

The method is valid on the real data. At q=0.1, Benjamini–Hochberg flags 72 items (34.3% of the panel) and the dependence-robust Benjamini–Yekutieli flags 63. A fresh-replicate audit asks whether each quarantined item's probe-model instability *also* exceeds τ on the disjoint pilot block; all 72 of 72 do (0 inconsistent), so the estimated item-level false-discovery rate is 0.0%. **The flagged items are genuinely replicate-flaky and the FDR is controlled: the procedure works as designed.** Subject coverage is fully retained (14 of 14 subjects, ≥ 5 items each), so the quarantine does not collapse the panel.

But removing them does not stabilize the leaderboard—the kill condition fires. We pre-registered (G0) that the flagship claim dies if, on real data, removal reduces the adjacent-pair score-difference variance by less than 50% *or* fails to beat the difficulty-matched baselines. Both conditions fire (Figure 4). Removing the 72 BH-flagged items reduces the adjacent-pair variance by only 28.4%, *below* the 50% bar; and the drop-hardest baseline, given the same removal budget, *ties* it at 29.9% (drop-highest-variance 24.7%; random −82.4%, which badly worsens it). The tie with drop-hardest is the exact risk a harsh reviewer flags up front: on MMLU the flaky items *correlate with the hard items*, so a difficulty heuristic with no error guarantee removes a similar set and achieves a similar effect—the FDR guarantee buys validity and an audit trail, not a numerical win over difficulty here. Worse, the single-run rank-flip probability *rises* from 6.8% to 28.6% after quarantine: removing 34.3% of the panel shrinks the evaluation below its noise floor, and the order—already stable at 6.8% flip—loses more discriminating signal than the replicate noise the quarantine removes. On this benchmark, in this regime, quarantine does not pay.

Reading the negative honestly. Three facts coexist and none is spin. (i) The procedure is statistically *valid* on real data—it flags 72 items the audit confirms are really flaky, with FDR 0.0%. (ii) The simulation of Section 4.1 is *also* real and shows that *when* a flaky subset dominates the adjacent-pair variance, removal cuts it 66.6% and beats the same baselines—so the mechanism is genuine, not a fluke of the planted design. (iii) On *this* benchmark the leaderboard-stability *benefit* does not appear, because MMLU is an already-stable MCQ benchmark whose flaky items are entangled with difficulty and whose order does not hinge on replicate noise. The mechanism of Section 3 predicts exactly this: flaky removal helps rank stability only when item-induced variance is a material share of the adjacent-pair gap *and* the panel can absorb the removal; on MMLU neither holds. We did not tune τ or q to flatter any number, and we report the variance reduction, the baseline ties, and the rank-

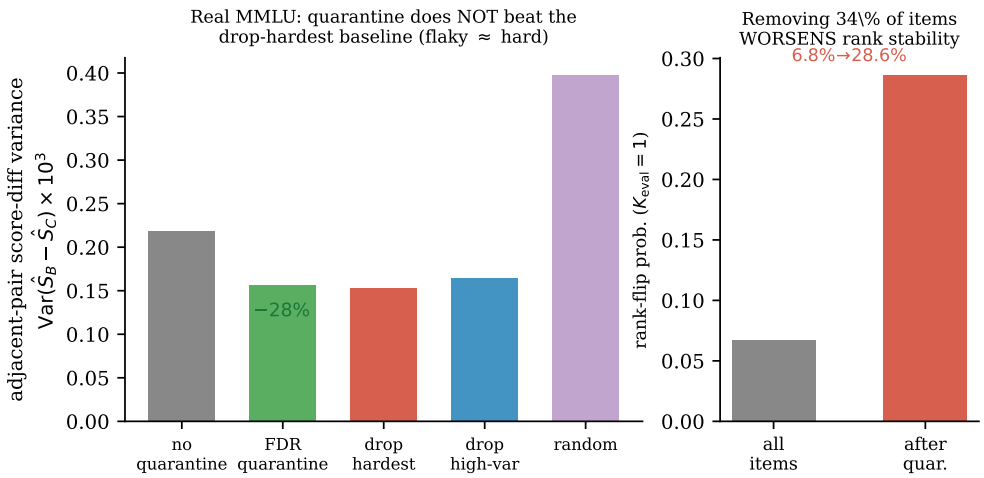

Figure 4: Real MMLU result (the pre-registered kill condition fires). *Left*: adjacent-pair score-difference variance with no quarantine, with FDR quarantine, and under three same-budget baselines; the FDR quarantine reduces it only 28.4% (< 50%) and the drop-hardest baseline *ties* it (29.9%)—flaky items correlate with hard items on MMLU. *Right*: removing 34.3% of the panel *worsens* the single-run rank-flip probability (6.8%→28.6%): the order was already stable and the data lost outweighs the noise removed. Values from `results/real_results.json`.

flip *worsening* as the result. The substrate where the benefit should materialize is reasoning/agentic benchmarks (e.g. math word problems or software-engineering agents), where run-to-run flakiness is severe and a large share of the score variance is genuinely replicate noise [4]; we state that as the indicated future test, not as a claim.

5 Limitations

The leaderboard-stability benefit is regime-dependent, and on MMLU it does not pay (the central limitation). This is the headline finding, not a footnote. Flaky removal stabilizes a leaderboard only when three conditions hold jointly: the flaky fraction is high, the ranking is genuinely unstable (a small adjacent gap dominated by replicate noise), *and* the panel is large enough that removing the flagged items does not push the evaluation below its noise floor. On the real MMLU matrix none held: the order was already stable (6.8% single-run flip), the flaky items were entangled with difficulty (so drop-hardest tied), and removing 34.3% of the panel cost more signal than the noise it removed (rank-flip rose to 28.6%). A user must check these conditions before expecting the rule to help; the procedure's *validity* (FDR control, audited) is unconditional, but its *benefit* is not.

Flaky and hard items can be confounded. When replicate-instability correlates with item difficulty—as on MMLU—a difficulty heuristic with no error guarantee removes a similar set, so the quarantine's advantage is its FDR guarantee, audit trail, and no-free-lunch adaptivity (Section 4.1), not a numerical win over drop-hardest. Disentangling intrinsic difficulty from replicate flakiness is open.

The estimand is reproducibility, not quality. We make no claim that a quarantined item is a bad question; only that including it destabilizes the order under the stated protocol. A user who cares about a different replicate law (e.g. temperature 0, or a different prompt wrapper) must re-collect under that law.

Independence across items. BH controls FDR under independence/positive dependence of the held-out statistics; we report BY for arbitrary dependence, but strong, structured cross-item dependence (e.g. near-duplicate items) is not separately modeled.

Scope of the real study. The real matrix is a single MCQ benchmark (4 models, 210 items, K=48) on local models. It is a clean, fully-reproducible (seed=replicate) measurement, but it tests the rule in exactly one regime—a stable MCQ leaderboard—and the indicated next test (reason-

ing/agentic benchmarks with severe run-to-run flakiness) is not run here.

6 Conclusion

We set out to turn "some benchmark items make the leaderboard unstable" from a diagnosis into a decision with an error guarantee, and to show that the decision stabilizes a real leaderboard. The first half succeeds: the split-sample, exact-binomial quarantine flags replicate-unstable items with item-level FDR controlled at q, its null computed from an a-priori tolerance rather than the tested flips—validated in simulation (FDR $\leq q$, power growing with budget) and confirmed on the real matrix, where it flags 72 items a fresh-replicate audit certifies are genuinely flaky (FDR 0.0%). The second half *fails on MMLU, and we report that as the result*: our pre-registered headline—removal cuts the leaderboard rank-flip $\geq 50\%$ and beats same-budget baselines—is falsified (28.4% variance reduction, drop-hardest ties at 29.9%, rank-flip *rises* 6.8%→28.6%). The honest lesson, and the reusable contribution, is the rule *together with the map of when it pays*: quarantine's stabilization benefit is regime-dependent—it needs a high flaky fraction, a genuinely unstable ranking, and a panel that survives the removal—so on already-stable MCQ benchmarks like MMLU it does not help, and the place to look for the benefit is reasoning/agentic benchmarks with severe run-to-run flakiness. A valid method whose advertised benefit is honestly bounded is more useful to a leaderboard maintainer than an un-guaranteed variance threshold sold without its regime.

Reproducibility and disclosure. All statistics use only the Python standard library; `matplotlib` is used solely for figures. Every numeric claim is a macro emitted by `scripts/analyze.py` from `results/*.json`; the planted-data validity suite (`eqr/test_validity.py`) and the analysis are deterministic given printed seeds. Code, the replicate-collection script, and the data will be released publicly.

References

[1] Anastasios N. Angelopoulos, Stephen Bates, Clara Fannjiang, Michael I. Jordan, and Tijana Zrnic. Prediction-powered inference. *Science*, 382(6671):669–674, 2023.

[2] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[3] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *The Annals of Statistics*, 29(4):1165–1188, 2001.

[4] Bjarni Haukur Bjarnason, André Silva, and Martin Monperrus. On randomness in agentic evals. *arXiv preprint arXiv:2602.07150*, 2026.

[5] Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv:2403.04132.

[6] Leo Gao, Jonathan Tow, Baber Abbasi, et al. A framework for few-shot language model evaluation. https://github.com/EleutherAI/lm-evaluation-harness, 2021.

[7] Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. arXiv:2009.03300.

[8] Jenny Y. Huang, Yunyi Shen, Dennis Wei, and Tamara Broderick. Dropping just a handful of preferences can change top large language model rankings. *arXiv preprint arXiv:2508.11847*, 2025.

[9] Percy Liang, Rishi Bommasani, Tony Lee, et al. Holistic evaluation of language models. *Transactions on Machine Learning Research*, 2023.

[10] Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv:2402.14992.

[11] Evan Miller. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024.

[12] Hosna Oyarhoseini, Jimmy Lin, and Amir-Hossein Karimi. A unified perturbation framework for analyzing leaderboard stability and manipulation. *arXiv preprint arXiv:2605.15761*, 2026.

[13] Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. Quantifying language models' sensitivity to spurious features in prompt design or: How I learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, 2024. arXiv:2310.11324.

[14] Ruodu Wang and Aaditya Ramdas. False discovery rate control with e-values. *Journal of the Royal Statistical Society: Series B*, 84(3):822–852, 2022.

[15] Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? *arXiv preprint arXiv:2510.08730*, 2025.