Resolution Diagnostics for Paired LLM Evaluation

Anany Kotawala [1]

Abstract

Across two public LLM leaderboards, many displayed pairwise rankings do not meet a conventional paired-test resolution target under the actual paired evaluation design: 11 of 40 Open LLM Leaderboard v1 pairwise comparisons and 4 of 9 MMLU-Pro top-10 adjacent-rank pairs are unresolved at $(\alpha, 1-\beta)=(0.05, 0.8)$. The MMLU-Pro count rises to $6/9$ under real subject-level clustering and stays at 5–6 out of 9 in 99.9% of category-bootstrap resamples. We frame paired LLM evaluation as a hypothesis-testing problem, invert level-α, power-$(1-\beta)$ tests, and report a per-pair resolution ratio $q=N/N^\star$ as the primary diagnostic. A sharp small-effect expansion with an explicit second-order constant shows that the widely-used unpaired Cohen-h-plus-$(1-\rho)$ shortcut deviates from the correct $N^\star$ by approximately a factor of two in the close-comparison regime, a deficit that three of five off-the-shelf calculators (Cohen 1988, G*Power, R `pwr`) silently inherit when the user post-multiplies their per-arm output by $(1-\rho)$. The unresolved-pair pattern remains under multiplicity correction and anytime-valid sequential testing.

1. Introduction

Modern LLM leaderboards rank models by percentage-point gaps on shared-prompt benchmarks. A leaderboard saying "model A scores 78.3%, model B scores 77.5%" converts a 0.8-point gap into headlines and product decisions; but the assertion that A is meaningfully better than B is a statistical claim about the gap, not the gap itself. On the four-task Open LLM Leaderboard v1, 11 of 40 displayed pairwise rankings do not meet the paired-test resolution target at $(\alpha, 1-\beta)=(0.05, 0.8)$; on the MMLU-Pro top-10, 4 of 9 adjacent-rank pairs remain unresolved at the actual benchmark size N=12,032. Concretely, the displayed gap between gemma-7B and Llama-3-8B on HellaSwag is $\hat{\delta}=+0.46$ pp at n=10,042: significant by asymptotic χ^2_1 (p=0.049), not by the exact conditional binomial (p=0.054), with a paired-bootstrap 95% CI on $\bar{D}$ containing zero. This is the kind of claim a resolution diagnostic surfaces. Whether a leaderboard has the resolution to support these claims depends on its size, on the within-prompt structure of the data, and on how many other adjacent gaps share the table.

The natural inference is a paired hypothesis test: for binary accuracy, McNemar's test on the discordant pairs with the corresponding required-N (McNemar, 1947; Connor, 1987); for graded scores, a paired-t or paired bootstrap. The machinery is classical. What is missing is a *resolution-reporting protocol* that says, given the size and structure of a benchmark, what gaps can be distinguished from sampling noise at conventional Type-I and Type-II error. Power calculators in common use focus on unpaired comparisons (Miller, 2024), and there is no standard for what to publish alongside a headline gap.

We treat shared-prompt LLM benchmarks as paired hypothesis-testing problems and derive a resolution-reporting framework by inverting level-α, power-$(1-\beta)$ tests. The framework yields three quantities: the minimum detectable effect (MDE) at the current N; the required paired sample size $N^\star$ at a target effect; and the resolution ratio $q = N/N^\star$.

The methodological building blocks are classical: Wald inversion, McNemar-Connor required-N, Bonferroni/Holm multiplicity, design-effect cluster correction, and anytime-valid e-processes. This paper contributes three things on top. First, a sharp characterisation of a specific misuse pattern (one new lemma with an explicit constant). Second, an empirical demonstration that the resulting diagnostic flags several displayed rankings as unresolved at the target $(\alpha, 1-\beta)$ resolution level. Third, a packaged reporting protocol that exposes each diagnostic entry as a one-line call. Concretely:

New theory.

- A **sharp small-effect expansion** (Lemma 1) for the un-

[1]Princeton University, Princeton, NJ, USA. Correspondence to: Anany Kotawala <akotawala@princeton.edu>.

Accepted to the *ICML 2026 Workshop on Hypothesis Testing*, Seoul, South Korea, 2026. Copyright 2026 by the author(s).

paired Cohen-h-plus-$(1-\rho)$ shortcut: the ratio $n_h/N^\star$ deviates from $\frac{1}{2}$ by at most $C(p,\rho)\delta^2 + O(\delta^4)$ with an explicit second-order constant $C(p,\rho)$ that lets practitioners compute the admissible $\delta^\star$ at which the shortcut is reliable (Corollary 1). The leading-order factor of two is not new; the explicit constant and the uniform convergence on compact admissible sets are.

New empirical findings.

- A **calculator-misuse characterisation** across five power-analysis tools at the worked example $(p_A, p_B, \rho)=(0.65, 0.60, 0.30)$: three of five (Cohen's 1988 textbook formula, G*Power 3.1, R `pwr`) silently underestimate required-N by a factor of two when their per-arm output is post-multiplied by $(1-\rho)$ (Table 1). The misuse pattern is easily reproduced on every widely-used calculator we tried; whether it occurs in published LLM-evaluation work is a separate empirical question that we do not claim to settle here.
- A **real-leaderboard reanalysis with prospective validation**: paired-McNemar required-N is median $2.15\times$ smaller than the unpaired Gaussian formula of Miller (2024) on the same data (IQR $[1.60, 2.75]$; Figure 1), and on three real OLL v1 pairs the framework's $N^\star$ prescription delivers empirical McNemar power 0.80 ± 0.03 over $M{=}1000$ bootstrap trials, with sub-prescription ($0.8\,N^\star$) and super-prescription ($1.2\,N^\star$) correctly under- and overshooting (Table 10). The diagnostic is calibrated on data, not just in asymptotic theory.
- **Real subject-level cluster sensitivity**: using MMLU-Pro's 14 subject categories as natural clusters, the unresolved adjacent-pair count rises from $4/9$ at IID to $6/9$ (Table 4); two pairs flip from comfortably resolved to $N^\star > 3N$. A category-bootstrap CI on the cluster-corrected unresolved count and LOSO sensitivity (Section F) both support the flip.

Methodological integration.

- Multiplicity, multi-arbiter, and anytime-valid sequential stress tests applied jointly to a single leaderboard family (§6–§6.5). Each component is classical; the joint pipeline and per-pair verdict table (Table 5) are, to our knowledge, the first end-to-end resolution characterisation of a public LLM leaderboard.
- A pip-installable package, `llm-power`, exposing the diagnostic for benchmark designers and leaderboard maintainers.

The empirical verdicts in §6 use Equation (6) directly and do not depend on the explicit constant in Lemma 1; the lemma's role is to sharply quantify the well-known factor-of-two pattern in a closed form practitioners can use to pre-screen their tools.

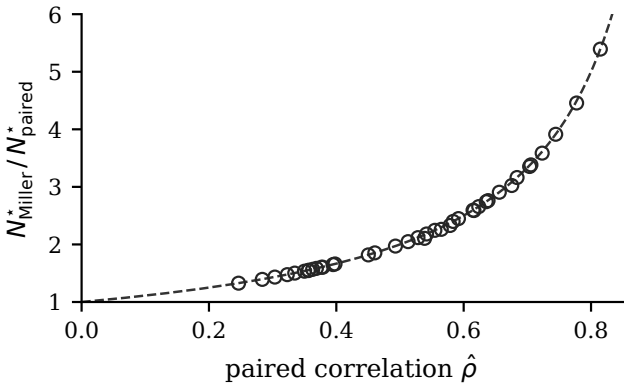

Figure 1. Efficiency gain of paired McNemar over the unpaired Gaussian formula of Miller (2024) across the 40 OLL v1 pairs. Open circles are empirical $N^\star_{\text{Miller}}/N^\star_{\text{paired}}$ at the pair's $\hat{\rho}$; dashed curve is the textbook prediction $1/(1-\rho)$ in the equal-marginal limit. Median empirical gain is $2.15\times$.

Paper roadmap. §3 defines the inversion and the resolution ratio q. §4 instantiates for paired-binary accuracy and proves the shortcut lemma; §4.1 traces it through five real calculators. §5 reports finite-sample calibration. §6 applies the diagnostic to OLL v1 and MMLU-Pro; §6.3–§6.5 stress-test the verdict under multiplicity, real subject clustering, and anytime-valid sequential testing. The headline empirical finding is Table 5; the headline theoretical finding is Lemma 1.

2. Related work

Paired-binary tests. McNemar's test (McNemar, 1947) conditions on the discordant-pair count $b{+}c$. The large-sample required-N formula is Connor (1987); mid-p (Liddell, 1983) and continuity-corrected variants (Agresti & Min, 2005) are the standard small-$b{+}c$ upgrades.

Power for NLP/LLM evaluation. Card et al. (2020) argued NLP comparisons are routinely underpowered; Dror et al. (2018) survey appropriate tests, including McNemar for paired binary outcomes. Methodologically closest is Miller (2024), who gave a closed-form required-N for the *unpaired* Gaussian-accuracy case. The unpaired formula is appropriate for independent samples; our comparison quantifies the efficiency loss when the evaluation design is actually paired. On the same 40 OLL v1 pairs we study (§6.1), the paired McNemar required-N is a median $2.15\times$ smaller than Miller's unpaired formula at the empirical $\hat{\rho}$ (IQR $[1.60, 2.75]$, range $[1.33, 5.39]$); the empirical efficiency gain matches the textbook prediction $1/(1-\rho)$ for paired vs. unpaired Wald tests (Figure 1, mean residual -0.009, max 0.062).

Madaan et al. (2024) measure benchmark variance across 13 tasks. Jo & Wilson (2025) give a clustered bootstrap for ability-estimation precision; that is an estimator-variance statement, distinct from test power. Polo et al. (2024) sub-

sample benchmarks for efficient point-estimate accuracy, also distinct from the required-N for paired hypothesis tests we address. Construct-validity critiques (Freiesleben & Zezulka, 2025; Bean et al., 2025; Alaa et al., 2025) are orthogonal: a benchmark can have high construct validity and still be too small for the test at hand.

To our knowledge, prior work treats these elements separately. The present paper integrates paired-difference variance with test inversion, leaderboard-scale multiplicity, and a per-pair resolution diagnostic on a real LLM benchmark, and packages the result as a reusable tool.

3. Methodology

Paired setup. Two models A, B are evaluated on the same N prompts, which we treat as an i.i.d. sample from a target prompt superpopulation; without this, the gap on exactly these items is a fixed quantity and no hypothesis test is needed. Per-prompt scores X_i^A, X_i^B may be binary, graded in $[0, 1]$, or real-valued. Define the paired difference $D_i = X_i^A - X_i^B$ and the paired-mean estimator $\hat{\delta} = N^{-1} \sum_i D_i$.

Test and inversion. Under a normal approximation,

$$T_N = \hat{\delta}/\widehat{\mathrm{SE}}(\hat{\delta}), \quad \widehat{\mathrm{SE}}(\hat{\delta}) = \sigma_D/\sqrt{N}, \tag{1}$$

where $\sigma_D^2 = \mathrm{Var}(D_i)$. For two-sided level-$\alpha$ tests, $H_0 : \mathbb{E}[D_i] = 0$ is rejected when $|T_N| \geq z_{1-\alpha/2}$. Under a fixed alternative $\mathbb{E}[D_i] = \delta$,

$$1 - \beta = \Phi\left(\frac{|\delta|}{\sigma_D}\sqrt{N} - z_{1-\alpha/2}\right) + O(N^{-1/2}), \tag{2}$$

where the $O(N^{-1/2})$ term is the standard finite-sample remainder of the normal approximation (Hall, 1992); Section 5 characterises it empirically.[1] Inverting at target power gives

$$N^\star(\delta; \alpha, \beta) = \left(\frac{(z_{1-\alpha/2}+z_{1-\beta})\,\sigma_D}{|\delta|}\right)^2, \tag{3}$$

$$\delta_{\mathrm{MDE}}(N; \alpha, \beta) = \frac{(z_{1-\alpha/2}+z_{1-\beta})\,\sigma_D}{\sqrt{N}}. \tag{4}$$

Definition 1 (Resolution ratio)**.** For a paired leaderboard pair with observed gap $\hat{\delta}$ on N shared prompts, the *resolution ratio* is $q := N/N^\star(\hat{\delta})$. The displayed gap is *statistically resolvable* at $(\alpha, 1-\beta)$ if $q \geq 1$.

For a single pair, q is a deterministic monotone transform of the squared Wald statistic: $q = T_N^2/(z_{1-\alpha/2} + z_{1-\beta})^2$, so $q \geq 1 \Leftrightarrow |T_N| \geq z_{1-\alpha/2} + z_{1-\beta} \approx 2.80$. Per-pair q therefore carries no information beyond the (Wald-statistic) p-value; its value-add over a p-value is interpretive, plus the aggregation layer (multiplicity, clustering, anytime-validity) where the q scale composes more naturally than p-values do.

$q < 1$ does not assert equality of the two models, nor does it overturn a fixed-N p-value: it says the benchmark of this size does not have the target resolution for a gap of the displayed magnitude under the stated $(\alpha, 1-\beta)$ operating point. The HellaSwag boundary pair in §1 illustrates the distinction: it rejects asymptotically at p=0.049 yet has $q \approx \frac{1}{2}$, so a nominally significant gap sits only halfway to the $(0.05, 0.8)$ resolution target. We use q as the paper's load-bearing reporting quantity throughout.

Prospective vs. diagnostic. Equation (3) has two uses. Plugging in a *pre-specified* target δ gives a prospective sample-size requirement for benchmark designers. Plugging in the *observed* $\hat{\delta}$ gives a resolution diagnostic. We do *not* compute power at the observed effect to argue that a particular ranking is true or false (the misuse criticised by Hoenig & Heisey (2001)); instead we compute $N^\star(\hat{\delta})$ as a benchmark-design diagnostic: the required N to detect a gap of the observed magnitude, irrespective of the significance verdict on the current sample.

Paired variance and multiplicity. On shared prompts, $\sigma_D^2 = \mathrm{Var}(X^A) + \mathrm{Var}(X^B) - 2\,\mathrm{Cov}(X^A, X^B)$. State-of-the-art LLMs solve overlapping subsets of items, so $\mathrm{Cov}(X^A, X^B)$ is large and independent-proportion variances inflate MDE and $N^\star$. A K-model leaderboard simultaneously displays up to $\binom{K}{2}$ pairwise tests; replacing α with a Bonferroni-, Holm- (Holm, 1979) or BH-adjusted α' (Benjamini & Hochberg, 1995) in Equation (3) directly inflates $N^\star$. Family-level resolution is the relevant object whenever a leaderboard summarises many pairs at once.

Sequential / anytime-valid extension. Public leaderboards update continuously, so the relevant N is a stopping time chosen post-hoc. The same σ_D^2 plugs into a confidence sequence (Howard et al., 2021; Ramdas et al., 2023) or a paired-Bernoulli mixture e-process, replacing $z_{1-\alpha/2}$ with a time-uniform boundary $u(n)$ satisfying $\Pr(\sup_n |T_n| \geq u(n)) \leq \alpha$ under H_0; Section 6.5 reports an $\approx 2\times$ threshold inflation and one additional unresolved MMLU-Pro pair.

4. Paired-binary instantiation and the shortcut lemma

For binary accuracy the paired-difference variance is

$$\sigma_D^2 = p_A q_A + p_B q_B - 2\rho\sqrt{p_A q_A\, p_B q_B}, \;\; q_\cdot = 1 - p_\cdot, \tag{5}$$

[1] The exact two-sided power is $\pi_N(\delta) = \Phi(-z_{1-\alpha/2} - \mu) + 1 - \Phi(z_{1-\alpha/2} - \mu)$ with $\mu = |\delta|\sqrt{N}/\sigma_D$. The second term dominates at conventional target powers (the first contributes $< 10^{-3}$ at $1-\beta = 0.8$); Equation (3) inverts the dominant term.

with ρ the within-pair Bernoulli correlation. Substitution into Equation (3) yields the McNemar-Connor required paired count

$$N^\star = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2\, \sigma_D^2}{(p_A - p_B)^2}, \tag{6}$$

which agrees asymptotically with the χ^2_1 McNemar test on the discordant counts $(b, c) \sim \mathrm{Binomial}(b{+}c, \frac{1}{2})$ under H_0 (McNemar, 1947; Connor, 1987).

Admissible correlations. Not every ρ is achievable for given marginals (p_A, p_B): the Hoeffding bound on Bernoulli correlation pins ρ to an interval $[\rho_{\min}, \rho_{\max}]$ with

$$\rho_{\max} = \sqrt{\frac{\min\{p_A(1{-}p_B), (1{-}p_A)p_B\}}{\max\{p_A(1{-}p_B), (1{-}p_A)p_B\}}}, \tag{7}$$

$$\rho_{\min} = -\sqrt{\frac{\min\{p_A p_B, (1{-}p_A)(1{-}p_B)\}}{\max\{p_A p_B, (1{-}p_A)(1{-}p_B)\}}}. \tag{8}$$

Equation (5) requires ρ to lie in this admissible interval; we restrict all numerical claims accordingly.

The unpaired-to-paired shortcut. A natural temptation, when only an unpaired Cohen-h calculator is on hand, is to read off the per-arm $n_{\mathrm{unp}} = (z_{1-\alpha/2} + z_{1-\beta})^2/h^2$ (Cohen, 1988, Eq. 6.3.2) with $h = 2\arcsin\sqrt{p_A} - 2\arcsin\sqrt{p_B}$, then apply Cohen's generic $(1{-}\rho)$ paired adjustment (Cohen, 1988, Ch. 2):

$$n_h = (1-\rho)\,(z_{1-\alpha/2} + z_{1-\beta})^2/h^2. \tag{9}$$

This is a natural shortcut when only an unpaired Cohen-h calculator is to hand (see the comparison of §4.1). Lemma 1 is a sharp small-effect characterisation of this failure mode: not a foundational theorem about paired tests, but a tight second-order expansion for a misuse that these tools make easy. The shortcut deviates from $N^\star$ by approximately a factor of two in the close-comparison limit; the deviation of $n_h/N^\star$ from $\frac{1}{2}$ is $O(\delta^2)$, not $O(\delta)$, and the ρ-dependent term in C vanishes at $p = \frac{1}{2}$ (relevant since many benchmark accuracies cluster near $\frac{1}{2}$). The technical contribution is the explicit constant $C(p, \rho)$ in Equation (11).

Lemma 1 (Sharp small-effect bound)**.** *Fix $p \in (0, 1)$ and δ small enough that $(p{+}\delta/2,\ p{-}\delta/2) \in (0, 1)^2$, and ρ in the Hoeffding-admissible region for this pair, with ρ bounded away from 1. Then for every sufficiently small $|\delta|$,*

$$\left|\frac{n_h(p{+}\frac{\delta}{2},\ p{-}\frac{\delta}{2},\ \rho)}{N^\star(p{+}\frac{\delta}{2},\ p{-}\frac{\delta}{2},\ \rho)} - \frac{1}{2}\right| \;\le\; C(p,\rho)\,\delta^2 \;+\; O(\delta^4), \tag{10}$$

with explicit constant

$$C(p,\rho) = \frac{1}{2}\left|\frac{(1{+}\rho)(1{-}2p)^2}{16(1{-}\rho)\,p^2(1{-}p)^2} - \frac{1}{6\,p(1{-}p)}\right|, \tag{11}$$

and convergence uniform on compact subsets of the admissible region. In particular $\lim_{\delta\to 0} n_h/N^\star = \frac{1}{2}$, and at $p = \frac{1}{2}$ the $(1{-}2p)^2$ term vanishes a fortiori, giving $C(\frac{1}{2}, \rho) = 1/3$ independent of ρ.

Corollary 1 (Underestimation in the close-comparison regime)**.** *For any $\epsilon \in (0, \frac{1}{2})$ there exists $\delta^\star(p, \rho, \epsilon)$ such that whenever $|\delta| \le \delta^\star$, $n_h \le (\frac{1}{2} + \epsilon)\,N^\star$, i.e. the shortcut underestimates $N^\star$ by at least $(\frac{1}{2} - \epsilon)\,N^\star$. Concretely $\delta^\star(p, \rho, \epsilon) = \sqrt{\epsilon/C(p, \rho)}$ to leading order; e.g. $\delta^\star(0.65, 0.3, 0.05) \approx 0.43$.*

Interpretation. The shortcut n_h is approximately one half of $N^\star$ at small δ, with the deviation above or below $\frac{1}{2}$ depending on (p, ρ). The leading-order factor of two follows from $\mathrm{Var}(X^A - X^B)$ algebra and is not new; the explicit constant and the uniform convergence are what Corollary 1 relies on to deliver a usable admissible $\delta^\star$.

When the shortcut matters operationally. In the close-comparison regime, the shortcut produces approximately $N^\star/2$ with a deviation that is $O(\delta^2)$. What changes with $|\delta|$ is the operational consequence. A verdict flips (resolved vs. unresolved at benchmark size N) only when $N^\star$ and n_h straddle N. For cross-tier comparisons where $\max(N^\star, n_h) \ll N$, both return "resolved" and the underestimate is benign. The verdict-changing regime is precisely the close-comparison regime that dominates leaderboard adjacencies: 17/40 OLL v1 pairs have $|\hat{\delta}| \le 5$ pp (Table 3) and all 9 MMLU-Pro top-10 adjacent pairs have $|\hat{\delta}| \le 7$ pp (Table 7). The shortcut is material exactly where adjacent-rank claims rest, and benign where cross-tier comparisons would resolve either way.

Proof sketch (full proof in Section A). The strategy is to Taylor-expand both h^2 and σ_D^2 around the midpoint p, exploiting the symmetry of $\arcsin\sqrt{\cdot}$ to kill the linear-in-δ terms. With $u = p(1{-}p)$, the result is $h^2 = (\delta^2/u)\cdot[1 + \delta^2\cdot H_2 + O(\delta^4)]$ where $H_2 = 1/(12u) + (1{-}2p)^2/(16u^2)$, and $\sigma_D^2 = 2u(1{-}\rho) + (\delta^2/4)\cdot[\rho(1{-}2p)^2/u - 2(1{-}\rho)] + O(\delta^4)$. Substituting and combining the two $O(\delta^2)$ corrections, $n_h/N^\star = \frac{1}{2} - \frac{\delta^2}{2}\big[(1{+}\rho)(1{-}2p)^2/(16(1{-}\rho)u^2) - 1/(6u)\big] + O(\delta^4)$, from which Equation (11) follows. □

Numerical check. Per-cell verification on a (p, ρ, δ) grid (`e6_ratio_heatmap.csv`; full details in Section A) matches $C(p, \rho)$ to four significant figures at small δ. At $\delta{=}0.05$, $|n_h/N^\star - \frac{1}{2}| \le 0.0008$; at $\delta{=}0.20$, ≤ 0.014. On the 40 OLL v1 pairs (§6.1), the empirical ratio $n_h/N^\star$ has median 0.5002, IQR $[0.4999, 0.5035]$, range $[0.487, 0.562]$; the 7 close-comparison pairs ($|\hat{\delta}| \le 2\%$) hit $\frac{1}{2}$ to four decimals.

The structural cause is that $(1{-}\rho)$ is correctly applied to the single-arm variance $p(1{-}p)$, but the paired-difference

Table 1. Calculator-misuse comparison at $(p_A, p_B, \rho) = (0.65, 0.60, 0.30)$, $(\alpha, 1-\beta) = (0.05, 0.8)$. "Per-arm" is the formula or calculator's native return value; "$\times(1-\rho)$" is the naive paired-sample-size readout. The correct $N^\star = 1{,}028$ from Equation (6). Three of the five formulas/tools return a per-arm K/h^2 that, when multiplied by $(1-\rho)$, gives the shortcut n_h: half the correct $N^\star$. `statsmodels` and `llm_power` return values that do not suffer this misuse pattern. Tool versions and invocation commands in artifact under `appendix_table1/`.

Calculator	convention	per-arm	$\times(1-\rho)$
Cohen 1988, Eq. 6.3.2	K/h^2	736	515
G*Power 3.1 (2-prop. z)	K/h^2	736	515
R `pwr::pwr.2p.test`	K/h^2	736	515
`statsmodels.NormalIndPower`	$2K/h^2$	1,471	1,030
`llm_power` (paired)	$\mathrm{Var}(\Delta)/\delta^2$	–	1,028

variance Equation (5) carries an additional factor of two from $\mathrm{Var}(X^A)+\mathrm{Var}(X^B)$. The fix is to use the paired-difference variance directly. We show next that this misuse pattern occurs in widely-used power tools.

4.1. Calculator-misuse comparison

Table 1 traces the comparison. Three of the five tools (Cohen's textbook, G*Power 3.1, R's `pwr`) reproduce the shortcut's factor-of-two underestimate when $(1-\rho)$ is applied to a per-arm K/h^2 output, because the paired adjustment is being applied to a single-arm variance rather than to $\mathrm{Var}(\Delta)$. The remaining two recover the correct $N^\star \approx 1{,}028$: `statsmodels.NormalIndPower` uses a $2K/h^2$ per-arm convention that already accounts for both arms, and `llm_power` computes the paired-difference variance directly. This is not a calculator-correctness issue: every tool returns a defensible quantity. It is a user-error pattern that the tools make easy, and Lemma 1 is its quantitative consequence at the leaderboard adjacency regime.

Field practice. The same pattern surfaces in the published literature. The unpaired Gaussian approximation of Miller (2024) is the prevalent closed-form required-N treatment for LLM accuracy comparisons; paired-variance corrections via McNemar–Connor are rarely surfaced, even on data with shared prompts. The NLP power advocacy of Card et al. (2020) and the test-selection survey of Dror et al. (2018) flag McNemar's availability but do not contrast paired and unpaired required-N on the same data. Lemma 1 and Table 1 together explain why this shortcut is quantitatively material at leaderboard adjacencies.

5. Finite-sample calibration

Equation (3) treats the rejection threshold as exact at any N, but the underlying normal approximation has a finite-N remainder. We check empirically how big that remainder is. We calibrate five paired-binary test variants on simulated paired-Bernoulli populations: McNemar χ^2_1, exact conditional binomial, mid-p, continuity-corrected McNemar, and the paired bootstrap of Definition 2. The grid covers marginal accuracies $p \in \{0.5, 0.7, 0.9\}$, latent-Gaussian correlations $\rho_z \in \{0, 0.4, 0.8\}$ used to generate the Bernoulli pair, and sample size n=500. The per-cell δ under H_1 is tuned so the McNemar-Connor asymptotic target is exactly 0.80 at each cell (Monte Carlo standard error ≈0.6 pp on Type-I, ≈1.0 pp on power). Table 2 reports cell-level Type-I deviations and empirical power; the bootstrap is the variant we recommend when no closed-form σ_D is available.

Definition 2 (Paired bootstrap test)**.** Given a paired score-matrix $\{(X_i^A, X_i^B)\}_{i=1}^N$, the two-sided percentile-bootstrap test of $H_0 : \mathbb{E}[D_i] = 0$ at level α resamples prompt indices with replacement, computes $\bar{D}^{(b)}$ on each resample, and rejects when $0 \notin \mathrm{CI}_{1-\alpha}$ from the percentile distribution of $\{\bar{D}^{(b)}\}_{b=1}^B$.[2]

Empirical Type-I and power. All five variants are calibrated within 1.1 pp of $\alpha = 0.05$. Under H_1 tuned to a 0.80 asymptotic target, the asymptotic trio (McNemar χ^2_1, mid-p, paired bootstrap) achieves median empirical power 0.79, while the exact and continuity-corrected variants are ≈3 pp conservative; the remaining gap is the $O(N^{-1/2})$ remainder of the normal approximation in Equation (2). On synthetic Bernoulli and Beta$(4, 2)$ graded marginals, the bootstrap tracks the parametric required-N to within 4–6% (Section B). *Recommendation:* use Equation (6) with empirical $\hat{\rho}$ for binary accuracy; use the paired bootstrap of Definition 2 for graded metrics.

6. Results

We apply the inversion to Open LLM Leaderboard v1 (§6.1) and the OLL v2 MMLU-Pro top-10 (§6.2). All required-N figures are at $(\alpha, 1-\beta) = (0.05, 0.8)$.

6.1. OLL v1: 40 unique pairwise comparisons

We pulled per-question 0/1 scores for five 7–8B-parameter open-weights models (Llama-3-8B±Instruct, Mistral-7B-Instruct-v0.2, Gemma-7B±it) on four tasks (ARC-Challenge, HellaSwag, Winogrande, GSM8K) from the EleutherAI `lm-evaluation-harness` dumps released via the OLL `details_*` repos. Each task contributes $\binom{5}{2} = 10$ unique pairwise comparisons (not adjacent-rank only), 40 comparisons in total.

Figure 2 plots $N^\star$ from Equation (6) at the observed $(\hat{p}_A, \hat{p}_B, \hat{\rho})$ against the actual benchmark size; markers

[2] The percentile variant is chosen for simplicity; the studentized bootstrap (Hall, 1992, Ch. 3) is asymptotically more accurate. Empirically (Table 2), the percentile bootstrap's Type-I deviation on our 5-cell grid is within 0.9 pp of nominal, so the simpler variant suffices at the N regime of interest.

Table 2. Empirical Type-I and power of five paired-binary test variants on a 5-cell (p, ρ, n) grid (M=1500 trials per cell; setup in §5).

Variant	Null	Data	$\lvert\hat{\alpha}-\alpha\rvert_{\max}$ (pp)	Power med. (H_1 at 0.80)	max dev. from 0.80 (pp)
McNemar χ^2_1	marginal	binary	0.9	0.79	5.3
Exact conditional	sharp	binary	1.1	0.76	10.2
Mid-p	sharp	binary	0.8	0.79	5.9
Continuity-corrected	marginal	binary	1.1	0.76	10.6
Paired bootstrap (Definition 2)	marginal	binary/graded	0.9	0.79	5.2

Randomisation on discordant signs coincides with the exact conditional binomial in the binary case. The paired-t / paired-bootstrap row applies to graded data only (Efron & Tibshirani, 1993; Hall, 1992).

Table 3. Fraction of OLL v1 pairs with resolution ratio $q < 1$, binned by $|\delta|$. The two ratio columns are $r = N^\star/N = 1/q$ summarised across pairs in each bucket; $r > 1$ means unresolved.

$\lvert\delta\rvert$ bucket	pairs	unresolved	r med.	r worst
$\leq 1\%$	3	3 (100%)	94	1,892
1%–2%	4	4 (100%)	4.2	6.8
2%–5%	10	4 (40%)	0.75	2.8
5%–15%	17	0 (0%)	0.15	0.65
$>15\%$	6	0 (0%)	0.03	0.07
all	40	11 (28%)	0.16	1,892

above $y = x$ are unresolved. The $|\delta|$-binned breakdown (Table 3) shows the resolution boundary near $|\delta| \approx 5\%$: every pair with $|\delta| \leq 2\%$ is unresolved and every pair with $|\delta| > 5\%$ is resolved; the 2–5% band is mixed.

Bootstrap uncertainty on $N^\star$. Plug-in $N^\star$ inherits the sampling uncertainty of $(\hat{p}_A, \hat{p}_B, \hat{\rho})$. We resample prompts with replacement ($B = 500$), recompute, and report the 5th–95th-percentile interval. Among the 7 OLL v1 close pairs ($|\delta| \leq 2\%$), 4 have a 5th-percentile $N^\star$ exceeding N (robustly unresolved); the remaining 3 have intervals that span the diagonal because $\hat{\delta}$ is poorly estimated near zero.

Prospective design validation. To verify that the framework's $N^\star$ prescription actually achieves the target power on real data, we picked three OLL v1 pairs spanning $|\delta| \in [6.3, 10.1]$ pp, computed $N^\star$ via Equation (6) at $(\hat{p}_A, \hat{p}_B, \hat{\rho})$, bootstrap-subsampled to $N^\star$ prompts, and ran McNemar at α=0.05 over M=1000 trials each. Empirical power at $N^\star$ lands at 0.796–0.827, within $\pm$2.7pp of the 0.80 target; at $0.8\,N^\star$ and $1.2\,N^\star$ it correctly under- and over-shoots. The framework's prescription is calibrated on the data, not only in asymptotic theory. Per-pair table in Section H.

Multi-arbiter agreement on close pairs. On every $|\delta| \leq 2\%$ pair, four arbiters (asymptotic McNemar χ^2_1, exact two-sided conditional binomial, mid-p, and the paired bootstrap of Definition 2) agree on six of seven close pairs (all unanimous fail-to-reject); only the leaderboard-displayed-significant HellaSwag pair flagged in §1 produces a split verdict ($\chi^2_1 : p$=.049 rejects; exact: p=.054 does not). Per-pair p-values in Section H.

6.2. MMLU-Pro paired item-level tightening (OLL v2)

For the largest OLL v2 task with accessible per-item correctness, we pull the gated `open-llm-leaderboard/*_details` parquets for the top-10 MMLU-Pro models (N = 12,032 items) and compute $N^\star$ for the 9 adjacent-rank pairs. The paired-Bernoulli formula (Equation (6)) and the discordance-form McNemar required-N agree to within 1% on every pair (Figure 3). Four of nine adjacent-rank pairs are unresolved at N = 12,032 under the adjacent-rank multiplicity convention $m = 9$. The all-pairs convention $m = \binom{10}{2} = 45$ inflates $N^\star$ by $\approx$2.14, but every unresolved pair already has $q < 1$, so all four verdicts stand.

Lemma 1 holds tightly on this data: the shortcut n_h is below $N^\star$ by the predicted factor of two on every pair (median ratio 0.500, range $[0.496, 0.500]$, despite ρ ranging over $[0.45, 0.99]$).

External replication. Section D reports a closed-source frontier panel (4 models, MMLU-Pro N=1,350 all-four-scored): one adjacent pair (Llama-4-Maverick vs. DeepSeek-V3.2, gap 2.8 pp) below resolution, shortcut/$N^\star$ in $[0.496, 0.500]$ replicating Lemma 1 on a different model class (illustrative; deployments rotate).

Sensitivity to $\hat{\rho}$ misspecification. The diagnostic uses the empirical paired-correlation $\hat{\rho}$, which is itself estimated. Perturbing $\hat{\rho}$ by $\pm$0.10 on every pair (clamped to the admissible interval) moves the OLL v1 unresolved count to $[9, 12]/40$ and the MMLU-Pro count to $[2, 4]/9$; the qualitative leaderboard message survives the perturbation. Three of 40 OLL v1 pairs and 2 of 9 MMLU-Pro pairs flip verdict; in every case a boundary pair with $|\hat{\delta}| \leq 4$ pp, including the HellaSwag pair from §1 ($\hat{\delta}$=0.46 pp, $\hat{\rho}$=0.81) which becomes resolved at $\hat{\rho}$+0.10=0.91. Boundary-case sensitivity is expected.

Stress-test sequence. The next three subsections stress the verdicts of §6 against family-level multiplicity (§6.3), real subject-level clustering (§6.4), and anytime-valid sequential testing (§6.5). Each adjustment strictly tightens the verdict (Table 5); the qualitative leaderboard message, that displayed adjacent gaps in this regime are often under-supported, survives all three.

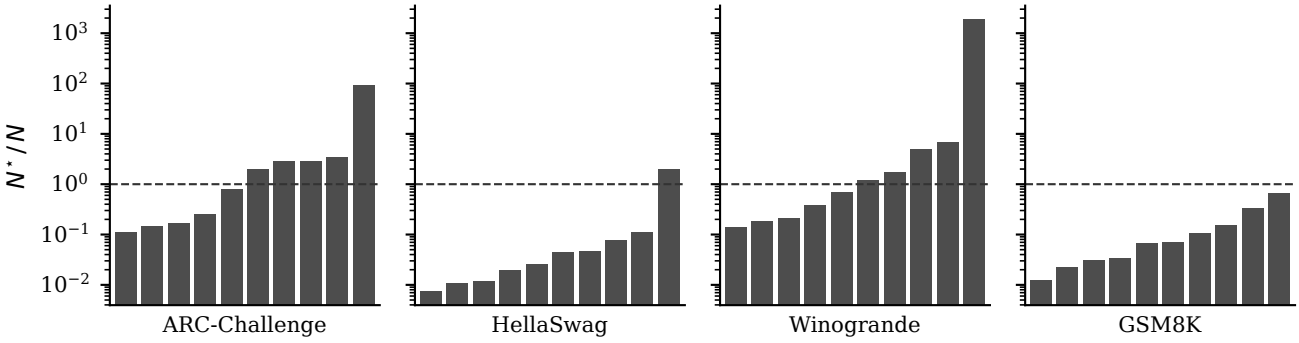

Figure 2. Resolution diagnostic on 40 OLL v1 pairwise comparisons, faceted by task. Each bar gives $r = N^\star/N = 1/q$ for one pair, sorted ascending within task; $N^\star$ at observed $(\hat{p}_A, \hat{p}_B, \hat{\rho})$ via Equation (6). Bars above the dashed line ($r > 1$) are unresolved at $(\alpha, 1-\beta) = (0.05, 0.8)$ (11/40 across all panels).

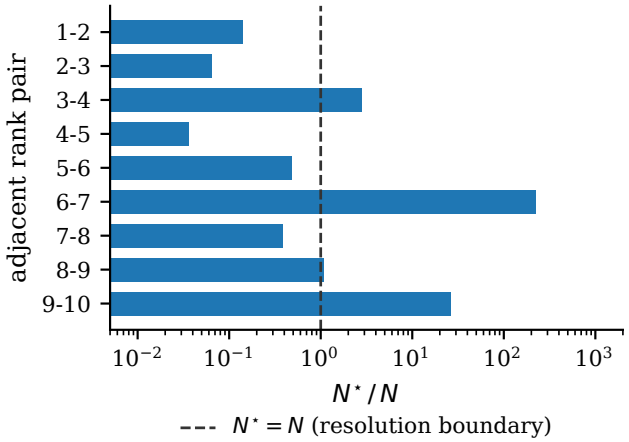

Figure 3. MMLU-Pro top-10 adjacent pairs (OLL v2, N = 12,032). Bars give $r = N^\star/N = 1/q$ for each adjacent-rank pair (log scale); the dashed line marks $r = 1$. Pairs whose bar extends past the dashed line ($r > 1$, unresolved) are unresolved at $(\alpha, 1-\beta) = (0.05, 0.8)$: four of nine.

6.3. Leaderboard-scale multiplicity

We report multiplicity under two pre-declared families: *adjacent-rank claims* (family size $K-1$) and *all-pairs leaderboard claims* (family size $\binom{K}{2}$). Families are declared before observing p-values.

Bonferroni / Holm. Running the 40 OLL v1 pairs at nominal $\alpha = 0.05$ inflates the family-wise error rate. Bonferroni control replaces $z_{1-\alpha/2} = 1.960$ with $z_{1-\alpha/(2\cdot 40)} \approx 3.227$, multiplying $N^\star$ by $\approx$2.11. Šidák yields a near-identical $2.10\times$ inflation; Holm (Holm, 1979) is bounded above by Bonferroni at every position, so $2.11\times$ is an upper bound on the Holm-adjusted required-N. Applying $2.11\times$ to Table 3, every $|\delta| \leq 2\%$ pair remains unresolved, the 2–5% unresolved fraction rises from 4/10 to 6/10, and 16/17 of the 5–15% pairs stay resolved (the worst-case pair at r=0.65 flips since $0.65\times2.11>1$); the $> 15\%$ bucket stays fully resolved.

Benjamini–Hochberg. Controlling the false-discovery rate at 0.05 via Benjamini–Hochberg (Benjamini & Hochberg, 1995) is less conservative. On these data it effectively coincides with the $|\delta| \approx 5\%$ resolution boundary (close pairs produce p-values far from significance, large-gap pairs essentially at zero), so no $|\delta| > 5\%$ rejection is retracted under FDR control.

6.4. Cluster-aware sensitivity

Real benchmarks contain topic clusters, near-duplicates, and templated subdomains (Madaan et al., 2024; Jo & Wilson, 2025) that induce intra-cluster correlation in the per-item paired difference D_i. The cluster-aware required-N scales the IID estimate by the design effect $\mathrm{DE} = 1 + (\bar{m} - 1)\,\mathrm{ICC}(D)$, where $\bar{m}$ is mean cluster size and $\mathrm{ICC}(D)$ is the intra-cluster correlation of D. We compute $\mathrm{ICC}(D)$ empirically on MMLU-Pro using the dataset's 14 subject categories (Madaan et al., 2024) as natural clusters ($\bar{m} \approx$ 859 items per category); Table 4 reports per-pair ICC, DE, and cluster-adjusted $N^\star$ for the 9 adjacent-rank top-10 pairs.

Real-data verdict. Median empirical $\mathrm{ICC}(D) = 0.0010$ across the 9 pairs; the distribution is heavy-tailed (range $[-3\times10^{-4},\ 3.6\times10^{-2}]$). Median DE = 1.88 but two pairs (rank 4 vs 5, rank 5 vs 6) show large clustering (DE = 31.5 and 6.7): paired-difference homogeneity is high within categories for these pairs because the model gap is dominated by specific subject domains. Under real cluster correction, the unresolved count rises from 4/9 at IID to 6/9: rank 4 vs 5 flips from $N^\star_{\mathrm{IID}} = 432$ (well-resolved) to $N^\star_{\mathrm{cluster}} = 13{,}621$ ($> N$); rank 5 vs 6 flips from 5786 to 39,009.

Bootstrap stability of the verdict. With only K=14 clusters, individual ICC point estimates carry nontrivial uncertainty. We address this by cluster-bootstrapping the verdict: resample the 14 categories with replacement (B=1000, seed 42) and recompute the unresolved-pair count. We hold $(\hat{p}_A, \hat{p}_B, \hat{\rho})$ at the full-data estimates so the resulting CI isolates cluster-structure uncertainty. The three pairs driving the IID-to-cluster flip have 5th-percentile ICC bounds that are all strictly positive (rank 4 vs 5 at $[0.021, 0.044]$, rank 5 vs 6 at $[0.003, 0.010]$, rank 6 vs 7 at $[0.004, 0.020]$),

Table 4. Real-data cluster sensitivity on MMLU-Pro top-10 adjacent-rank pairs, using the dataset's 14 subject categories as clusters ($K = 14$, $\bar{m} \approx 859$). $\mathrm{ICC}(D)$ is the empirical intra-cluster correlation of the paired difference D_i (one-way ANOVA estimator); $\mathrm{DE} = 1+(\bar{m}-1)\mathrm{ICC}^+$ with $\mathrm{ICC}^+ = \max(\mathrm{ICC}, 0)$. Bold $N^\star$ exceeds the actual N=12,032.

Pair	$\lvert\hat{\delta}\rvert$ pp	$\hat{\rho}$	ICC(D)	DE	Cluster $N^\star$
1 vs 2	1.18	0.92	0.0004	1.37	2,327
2 vs 3	1.73	0.93	−0.0003	1.00	778
3 vs 4	0.10	0.99	0.0002	1.19	**40,660**
4 vs 5	6.61	0.46	0.036	31.5	**13,621**
5 vs 6	1.88	0.45	0.0067	6.74	**39,009**
6 vs 7	0.08	0.49	0.012	11.5	$\geq 10^7$
7 vs 8	0.91	0.90	−0.0002	1.00	4,628
8 vs 9	0.86	0.75	0.0010	1.88	**24,632**
9 vs 10	0.22	0.58	0.0029	3.52	$\geq 10^6$

so the cluster signal that flips them is unlikely to be a K=14 artefact. Across the full bootstrap, the unresolved-pair count is $5/9$ in 45% of resamples and $6/9$ in 55%; $\Pr(\text{unresolved} \geq 5) = 99.9\%$, and only 1 of 1000 resamples returns to the IID $4/9$ count. Full per-pair CIs are in Section G.

Sensitivity, not certification. §6.4 should be read as a cluster-sensitivity analysis. Three robustness checks support the conclusion despite K=14: (i) the cluster-bootstrap above (verdict is at 5–6 unresolved out of 9 in 99.9% of resamples), (ii) a leave-one-subject-out (LOSO) recomputation that drops each MMLU-Pro category in turn (Section F), which holds the unresolved count at $6/9$ on 11 of 14 drops and $5/9$ on the remaining 3, and (iii) the cluster-definition sensitivity below. None of these collapses the result below $5/9$ except under random clusters (the null check), so the headline $4/9 \rightarrow 6/9$ flip is not an artifact of either category sampling or any single high-ICC subject.

Cluster-definition sensitivity. Subject categories are one of many possible clusterings. We rerun the §6.4 pipeline under three alternatives: random clusters of K=14 (null check); difficulty quartiles (K=4, binned by per-item mean accuracy across the top-10); and subject sub-clusters (K=28, each subject split in half by item parity). Random clusters give ICC≈0 and revert to the IID $4/9$ verdict (the null check). Difficulty quartiles produce a stronger cluster signal (median ICC=0.019, max 0.19) and $9/9$ unresolved; subject sub-clusters give $5/9$. The unresolved count thus spans $[5, 9]/9$ across non-null definitions, with the headline $6/9$ at the midpoint.

6.5. Anytime-valid leaderboard testing

Public leaderboards update continuously: each new model triggers up to $K-1$ new pairwise tests against the existing entries. A fixed-n test loses Type-I control when stopping can be decided after seeing the data. We replace the fixed McNemar-Connor threshold $z_{1-\alpha/2}$ by an *anytime-valid* threshold (valid simultaneously at every n) derived from a paired-Bernoulli mixture e-process (Howard et al., 2021; Ramdas et al., 2023; Vovk & Wang, 2021). Construction, mixture choice, and Type-I/stopping calibration are in Section C.

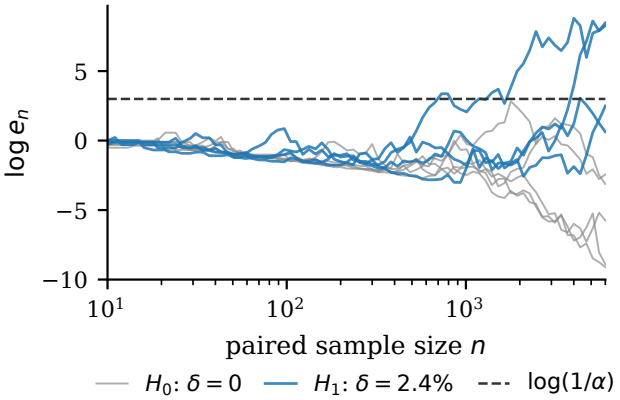

Figure 4. Mixture e-process trajectories for paired Bernoulli under H_0 (δ=0, grey) and H_1 (δ=2.4%, blue), calibrated to an ARC pair ($\hat{\rho}$=0.64). The horizontal dashed line is the rejection threshold $\log(1/\alpha)$ at α=0.05. H_0 trajectories stay below; H_1 trajectories cross between $n{\sim}10^3$ and $n{\sim}5\times10^3$ (fixed-n McNemar-Connor $N^\star$=2362 on this pair).

Applied to MMLU-Pro adjacencies. At N=12,032 the time-uniform threshold inflates $N^\star$ by 2.15× vs. fixed-n. The numerical proximity to the m=45 Bonferroni inflation (2.14×, §6.2) and the m=40 figure (2.11×, §6.3) is coincidence: anytime-validity gives a time-uniform boundary; Bonferroni is a family-size correction.

Anytime-valid testing flips one extra verdict relative to fixed-n. Five of nine adjacent-rank pairs are unresolved under anytime-validity, vs. $4/9$ under fixed-n and Bonferroni-9. The extra pair is rank 5 vs. 6 ($\hat{\delta}$=1.88pp, $\hat{\rho}$=0.45): the fixed-n exact McNemar p-value is 5.8×10^{-5}, comfortably rejecting. The anytime-valid threshold, however, is stricter than a one-shot test at the realised discordance, so the same data fails to cross it. Figure 4 traces sample $\log e_n$ trajectories under H_0 and H_1.

Verdict across procedures. Table 5 pulls the unresolved counts together. On the larger OLL v1 family (m=40) the anytime-valid threshold-inflation at typical N coincides with Bonferroni's; on the smaller MMLU-Pro family (m=9) anytime-validity dominates, flagging one additional pair as sequentially-not-resolvable. The columns of Table 5 answer subtly different questions and are not directly comparable as more-vs-less-correct: the anytime-valid column scores whether a verdict is resolvable *under continuous monitoring of an infinitely-extending stream*, which is a stricter criterion than fixed-n McNemar applied once.

Table 5. Unresolved-pair count across testing paradigms.

Procedure	OLL v1	MMLU-Pro
Fixed-n	$11/40$	$4/9$
Bonferroni / Holm	$14/40$	$4/9$
Anytime-valid	$14/40$	$5/9$
Real-ICC clustered	n/a	$6/9$

7. Conclusion

LLM leaderboard claims should report not only a gap and a p-value, but the benchmark resolution needed to make that gap detectable under the actual paired design.[2] Across OLL v1 and MMLU-Pro, many displayed adjacent gaps fall below this resolution once pairing, multiplicity, clustering, or anytime-validity are made explicit. Clustering is verified by a category-bootstrap and LOSO (Sections F and G).

Scope. The load-bearing claims (the shortcut lemma and the unresolved-pair verdicts) apply to the close-comparison regime where leaderboard adjacencies live. For cross-tier comparisons with large gaps, both paired and unpaired calculators return "resolved" at typical benchmark sizes, and the shortcut's factor-of-two underestimate is benign. The lemma is sharpest near $p{=}\frac{1}{2}$, which is where many benchmark accuracies cluster.

Limitations. Our empirics are restricted to binary accuracy, the displayed metric on the headline benchmarks; graded metrics and pairwise-preference leaderboards are handled only methodologically via Definition 2 (validated to 4–6% on Beta$(4, 2)$ marginals, Section B). The MMLU-Pro cluster analysis rests on $K{=}14$ subject categories; LOSO and a category bootstrap (Sections F and G) support the headline $4/9 \rightarrow 6/9$ flip, but a benchmark with finer-grained natural clusters would tighten the design-effect estimate. The closed-source frontier panel (Section D) is a snapshot under rotating deployments and replicates Lemma 1 only as illustration, not as a durable head-to-head comparison.

Future work. Four directions extend the framework. (i) Generalise Lemma 1 away from the equal-marginal midpoint and to multi-arm power. (ii) Derive a PRDS-aware required-N for leaderboard families with overlapping models, building on Benjamini & Yekutieli (2001). (iii) Plug in a fully clustered-prompt bootstrap with empirically-estimated ICC (Jo & Wilson, 2025) as the principled refinement of the design-effect estimate. (iv) Instantiate the resolution diagnostic on judge-scored and pairwise-preference (Chatbot-Arena-style) leaderboards via Bradley–Terry e-processes. Construct validity (Freiesleben & Zezulka, 2025; Bean et al., 2025; Alaa et al., 2025) composes with resolution: a benchmark can be statistically resolvable and still fail to measure what it claims to.

[2]Code and raw responses: https://github.com/akotawala10/llm-power.

References

Agresti, A. and Min, Y. Simple improved confidence intervals for comparing matched proportions. *Statistics in Medicine*, 24(5):729–740, 2005.

Alaa, A., Hartvigsen, T., Golchini, N., Dutta, S., Dean, F., Raji, I. D., and Zack, T. Position: Medical large language model benchmarks should prioritize construct validity. *arXiv preprint arXiv:2503.10694*, 2025. URL https://arxiv.org/abs/2503.10694. ICML 2025 (Position track).

Bean, A. M., Kearns, R. O., Romanou, A., Hafner, F. S., Mayne, H., et al. Measuring what matters: Construct validity in large language model benchmarks. *arXiv preprint arXiv:2511.04703*, 2025. URL https://arxiv.org/abs/2511.04703. NeurIPS 2025 Datasets and Benchmarks Track.

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

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

Card, D., Henderson, P., Khandelwal, U., Jia, R., Mahowald, K., and Jurafsky, D. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274, 2020.

Cohen, J. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2nd edition, 1988.

Connor, R. J. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43(1): 207–211, 1987.

Dror, R., Baumer, G., Shlomov, S., and Reichart, R. The hitchhiker's guide to testing statistical significance in natural language processing. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1383–1392, 2018.

Efron, B. and Tibshirani, R. J. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

Freiesleben, T. and Zezulka, S. The benchmarking epistemology: Construct validity for evaluating machine learning models. *arXiv preprint arXiv:2510.23191*, 2025. URL https://arxiv.org/abs/2510.23191.

Grünwald, P., de Heide, R., and Koolen, W. Safe testing. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2024. Read paper, with discussion.

Hall, P. *The Bootstrap and Edgeworth Expansion*. Springer, 1992.

Hoenig, J. M. and Heisey, D. M. The abuse of power: The pervasive fallacy of power calculations for data analysis. *The American Statistician*, 55(1):19–24, 2001.

Holm, S. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Howard, S. R., Ramdas, A., McAuliffe, J., and Sekhon, J. Time-uniform, nonparametric, nonasymptotic confidence sequences. *Annals of Statistics*, 49(2):1055–1080, 2021.

Jo, N. and Wilson, A. What does your benchmark really measure? a framework for robust inference of AI capabilities. *arXiv preprint arXiv:2509.19590*, 2025. URL https://arxiv.org/abs/2509.19590.

Liddell, F. D. K. Simplified exact analysis of case-referent studies: Matched pairs; dichotomous exposure. *Journal of Epidemiology and Community Health*, 37(1):82–84, 1983.

Madaan, L., Singh, A. K., Schaeffer, R., Poulton, A., Koyejo, S., Stenetorp, P., Narang, S., and Hupkes, D. Quantifying variance in evaluation benchmarks. *arXiv preprint arXiv:2406.10229*, 2024. URL https://arxiv.org/abs/2406.10229.

McNemar, Q. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947.

Miller, E. Adding error bars to evals: A statistical approach to language model evaluations. *arXiv preprint arXiv:2411.00640*, 2024. URL https://arxiv.org/abs/2411.00640.

Polo, F. M., Weber, L., Choshen, L., Sun, Y., Xu, G., and Yurochkin, M. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Ramdas, A., Grünwald, P., Vovk, V., and Shafer, G. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Vovk, V. and Wang, R. E-values: Calibration, combination and applications. *Annals of Statistics*, 49(3):1736–1754, 2021.

Waudby-Smith, I. and Ramdas, A. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 86(1):1–27, 2024.

Appendix

Supplementary material to "Resolution Diagnostics for Paired LLM Evaluation." Sections A–J cover, respectively, the proof of Lemma 1, synthetic Bernoulli/graded validation, the mixture e-process construction and calibration, the illustrative closed-source replication, raw discordance and pair details, leave-one-subject-out cluster robustness, the cluster-bootstrap CIs on MMLU-Pro ICC, prospective validation and multi-arbiter agreement on OLL v1 close pairs, a reporting checklist for paired-leaderboard claims, and the `llm-power` *API.*

A. Proof of Lemma 1

Let $\phi(t) = \arcsin\sqrt{t}$, so $h(p_A, p_B) = 2[\phi(p_A) - \phi(p_B)]$ with $\phi'(t) = 1/[2\sqrt{t(1-t)}]$ and $u(t) := t(1-t)$. Under the midpoint parameterization $p_A = p + \delta/2$, $p_B = p - \delta/2$, expand symmetrically around p:

$$\phi(p_A) - \phi(p_B) = \delta\,\phi'(p) + \frac{\delta^3}{24}\phi'''(p) + O(\delta^5), \tag{12}$$

the even-order (δ^2, δ^4, ...) Taylor terms cancelling by symmetry of ϕ around p. Squaring,

$$h^2 = 4\delta^2\phi'(p)^2\left[1 + \tfrac{\delta^2}{12}\,\phi'''(p)/\phi'(p) + O(\delta^4)\right]. \tag{13}$$

Direct computation gives $\phi'(p)^2 = 1/(4u)$ and $\phi'''(p)/\phi'(p) = 1/u + 3(1-2p)^2/(4u^2)$, so

$$h^2 = \frac{\delta^2}{u}\left[1 + \delta^2\left(\tfrac{1}{12u} + \tfrac{(1-2p)^2}{16u^2}\right) + O(\delta^4)\right]. \tag{14}$$

The $O(\delta^2)$ correction inside the bracket contributes at the same order as the σ_D^2 correction below, so it cannot be dropped.

For the paired-difference variance Equation (5), expand symmetrically. The sum $u(p_A) + u(p_B) = 2u(p) + (\delta^2/4)u''(p) + O(\delta^4)$, and since $u''(t) = -2$,

$$u(p_A) + u(p_B) = 2u(p) - \tfrac{\delta^2}{2} + O(\delta^4). \tag{15}$$

For the product, write $u(p_A) \approx u(p) + (\delta/2)u'(p) + (\delta^2/8)u''(p)$ and similarly for $u(p_B)$, then

$$\begin{aligned} u(p_A)u(p_B) &= \left[u(p) - \tfrac{\delta^2}{4}\right]^2 - \left[\tfrac{\delta}{2}u'(p)\right]^2 + O(\delta^4) \\ &= u(p)^2 - \tfrac{\delta^2}{2}u(p) - \tfrac{\delta^2}{4}u'(p)^2 + O(\delta^4), \end{aligned} \tag{16}$$

so

$$\sqrt{u(p_A)u(p_B)} = u(p) - \delta^2\left(\tfrac{1}{4} + \tfrac{u'(p)^2}{8u(p)}\right) + O(\delta^4). \tag{17}$$

With $u'(p) = 1-2p$, substitution into Equation (5) gives

$$\begin{aligned} \sigma_D^2 &= 2u(p)(1-\rho) \\ &\quad + \tfrac{\delta^2}{4}\left[\rho\,\tfrac{(1-2p)^2}{u(p)} - 2(1-\rho)\right] + O(\delta^4). \end{aligned} \tag{18}$$

The shortcut-to-paired ratio is then

$$\begin{aligned} \frac{n_h}{N^\star} &= \frac{(1-\rho)\,\delta^2}{h^2\,\sigma_D^2} = \frac{1}{2\,(1+\delta^2 H_2)(1+\delta^2 V_2)} \\ &= \tfrac{1}{2}\left[1 - \delta^2(H_2 + V_2) + O(\delta^4)\right], \end{aligned} \tag{19}$$

where the corrections are

$$H_2 = \tfrac{1}{12u} + \tfrac{(1-2p)^2}{16u^2}, \tag{20}$$

$$V_2 = \tfrac{\rho(1-2p)^2}{8u^2(1-\rho)} - \tfrac{1}{4u}. \tag{21}$$

Combining,

$$H_2 + V_2 = \frac{(1+\rho)(1-2p)^2}{16(1-\rho)u^2} - \frac{1}{6u}. \tag{22}$$

Taking absolute values,

$$\left|\frac{n_h}{N^\star} - \frac{1}{2}\right| \le \frac{1}{2}\left|\frac{(1+\rho)(1-2p)^2}{16(1-\rho)u^2} - \frac{1}{6u}\right| \delta^2 + O(\delta^4), \tag{23}$$

giving $C(p,\rho)$ as in Equation (11). Convergence is uniform on any compact (p,ρ) set in which $u(p) \ge u_0 > 0$ and $1-\rho \ge \eta_0 > 0$, because $C(p,\rho)$ is then bounded above and the $O(\delta^4)$ remainder uniform. Corollary 1 follows by inverting $C(p,\rho)\delta^2 \le \epsilon$. □

Numerical cross-check. Direct per-cell evaluation on the heatmap grid of `e6_ratio_heatmap.csv` confirms the leading-order coefficient: the empirical $|n_h/N^\star - \frac{1}{2}|/\delta^2$ matches the closed-form $C(p,\rho)$ to four significant figures at small δ on every cell ($p \in \{0.5, 0.65, 0.8\}$, $\rho \in [0, 0.7]$, $\delta \in [0.005, 0.20]$; median relative error 0.08% at $\delta \le 0.05$). The full inequality $|n_h/N^\star - \frac{1}{2}| \le C(p,\rho)\delta^2 + O(\delta^4)$ from Lemma 1 holds with the explicit $O(\delta^4)$ remainder: at the upper end of the grid (e.g. $(p,\rho,\delta) = (0.8, 0.5, 0.20)$), the $O(\delta^4)$ correction adds up to $\approx 17\%$ on top of $C\delta^2$, which is the expected behaviour of a small-δ Taylor bound. At $p = 0.5$ the $(1-2p)^2$ factor vanishes a fortiori, giving the clean sanity check $C(0.5,\rho) = 1/(12u) = 1/3$ for every admissible ρ (independent of ρ). Off-midpoint the constant grows with both $|1-2p|$ and ρ: empirically $C(0.65, 0.0) \approx 0.31$, $C(0.8, 0.5) \approx 0.80$, $C(0.65, 0.9) \approx 0.67$.

B. Synthetic Bernoulli and graded validation

Figure 5 reports bootstrap power against n on synthetic paired Bernoulli data with $p_A = 0.65$, $\delta \in \{0.02, 0.04, 0.08\}$, $\rho \approx 0.3$ via a latent Gaussian copula, $N = 30{,}000$. The bootstrap crosses the 0.8 target within 5% of $N^\star$ from Equation (6) for all three δ. Figure 6 repeats the exercise on Beta$(4, 2)$ marginals (graded stress test); the bootstrap tracks the paired-t required-N to within 4–6%.

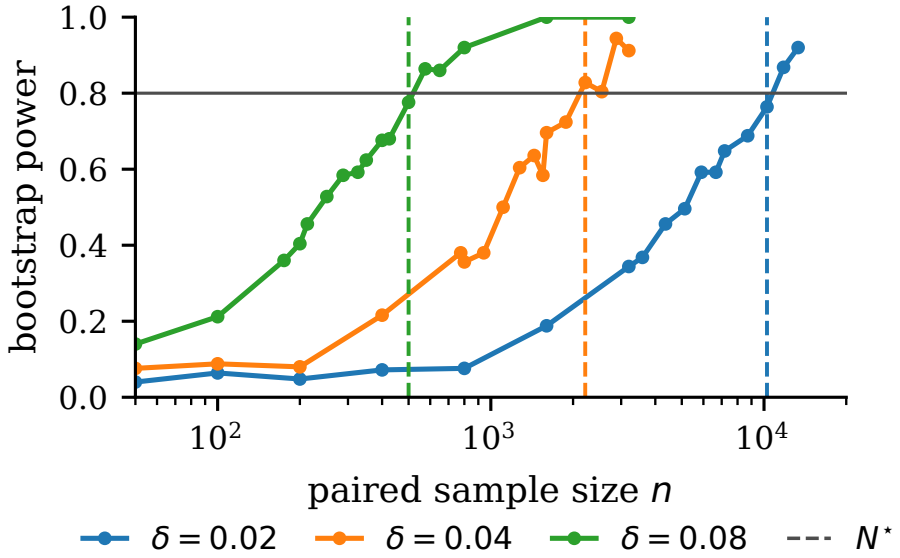

Figure 5. Bootstrap power against paired sample size n on synthetic paired Bernoulli data, three δ. Dashed verticals mark $N^\star$ from Equation (6).

C. Mixture e-process: construction and calibration

Construction. Conditioned on $b+c$ discordant pairs, the sign sequence is i.i.d. Bernoulli$(\frac{1}{2})$ under H_0. With a discrete mixture prior ν over the alternative discordance probability $\theta \in (0, \frac{1}{2}) \cup (\frac{1}{2}, 1)$, the e-process is

$$e_n = \int \frac{\theta^{b_n}(1-\theta)^{c_n}}{(1/2)^{b_n+c_n}} \, \mathrm{d}\nu(\theta), \tag{24}$$

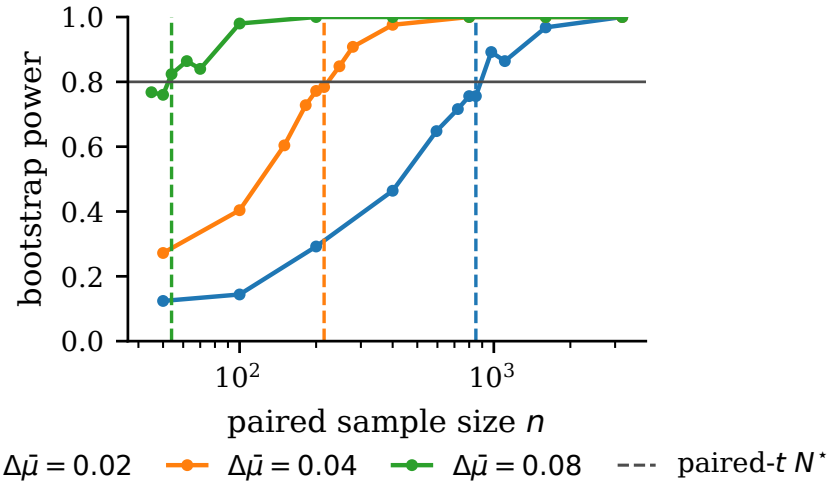

Figure 6. Non-Bernoulli stress test: paired graded scores from Beta$(4, 2)$ marginals. Bootstrap tracks the paired-t required-N within $\pm 6\%$.

By Ville's inequality (a martingale maximal inequality), e_n has $\Pr(\sup_n e_n \geq 1/\alpha) \leq \alpha$ at any stopping time. The rejection rule "reject the first n at which $e_n \geq 1/\alpha$" is therefore anytime-valid.

Mixture choice. We use a discrete uniform mixture over $\theta \in \{0.01, \ldots, 0.49, 0.51, \ldots, 0.99\}$ with equal weights, for two reasons. First, it is conjugate to the discordant-binomial likelihood, so the integral in Equation (24) is a closed-form sum. Second, it is the natural default in the betting-based formulation of safe testing (Waudby-Smith & Ramdas, 2024), where the prior plays the role of a uniform prior over wagers. Alternatives that grow faster against a specific point alternative (e.g. GROW, Grünwald et al., 2024) require knowing the alternative density and are sensitive to misspecification. Empirically, a Beta$(2, 2)$ mixture and a discrete two-point mixture at $\theta \in \{0.4, 0.6\}$ both produce stopping times within $\approx 8\%$ of the uniform on our calibration pairs.

Calibration. We calibrate the mixture e-process on simulated paired Bernoulli calibrated to two ARC pairs ($\hat{\delta} = 2.4\%, 7.8\%$; $\hat{\rho} = 0.64, 0.54$; M=600 trials each). The empirical Type-I rates are 0.035 and 0.043 (Monte Carlo SE $\approx$0.9 pp, both within nominal α=0.05); the H_1 rejection rate is 97–100%. The expected stopping time is 1.84–$2.32\times$ the fixed-n McNemar-Connor $N^\star$ on resolved pairs: the time-uniform cost of dropping the pre-specified-n assumption.

D. Closed-source frontier panel (illustrative)

We replicate Lemma 1 on a closed-source frontier panel as a robustness check on a different model class and evaluation route; we include it as illustrative rather than load-bearing because deployments rotate. Four pay-walled API models (GPT-5.5 and GPT-5.4 on Azure OpenAI; DeepSeek-V3.2 and Llama-4-Maverick-17B-128E-Instruct-FP8 on Azure AI Foundry) were evaluated on a 3,000-item random subsample of MMLU-Pro on 2026-05-08 under a 5-shot CoT prompt; dropping items where any model errored leaves N=1,350 all-four-scored items. Across the 3 adjacent-rank pairs the shortcut/$N^\star$ ratio sits in $[0.496, 0.500]$ (median 0.497), replicating Lemma 1's factor-of-two. The Llama-4-Maverick vs. DeepSeek-V3.2 pair (gap 2.8 pp, $\rho = 0.48$) is unresolved at N=1,350; the other two are resolved (Figure 7). Exact API identifiers, call timestamps, the verbatim 5-shot CoT prompt, and the 3,000-item index list (seed 42 over the MMLU-Pro test split) are in the artifact under `appendix_d/`.

E. Raw discordance and pair details

Tables 6 and 7 list the per-pair raw discordance counts (b, c), marginals, correlation, and both p-value variants for the seven OLL v1 close pairs (§6.1) and the nine MMLU-Pro adjacent-rank pairs (§6.2). These let a reader verify the headline HellaSwag p_{χ^2}=0.049 vs. p_{exact}=0.054 split (§1, row 4 of Table 6) without re-running the code, and trace each rank pair to its concrete model identifiers.

Table 6. Raw discordance and pair details for the seven OLL v1 close pairs ($|\hat{\delta}| \leq 2$pp). b=n_{AB} and c=n_{BA} are the discordant counts; $\hat{\rho}$ is the empirical Bernoulli correlation; $N^\star$ is the McNemar-Connor required-N from Equation (6). The HellaSwag row reconciles the §1

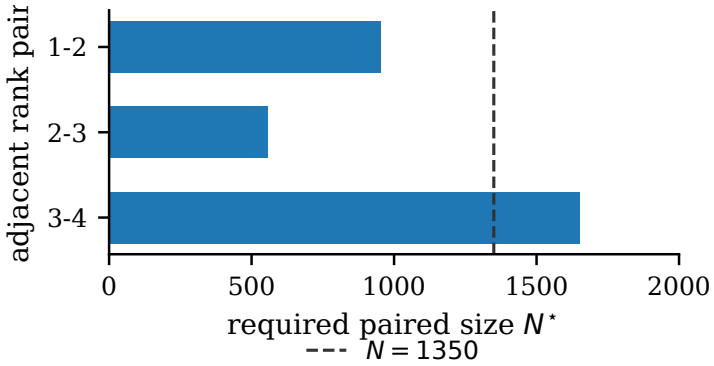

Figure 7. Frontier panel on a 3,000-item MMLU-Pro subsample (N=1,350). Bars give $N^\star$ for each adjacent-rank pair; the dashed line marks the actual N. Llama-4-Maverick vs. DeepSeek-V3.2 (rank 3 vs. 4) is unresolved at $(0.05, 0.8)$.

example: $(b-c)/N = 46/10{,}042 = 0.46$ pp. The $\approx 2.4\times10^6$ Winogrande Mistral-I/Llama-3-8B figure reflects $\hat{\delta} \approx 0$ ($b - c = 1$ out of 241 discordant pairs) and should be read as "far beyond resolution" rather than a precise budget.

Task	Model A	Model B	N	$\hat{p}_A$	$\hat{p}_B$	b	c	$\hat{\rho}$	p_{χ^2}	p_{exact}	$N^\star$
ARC-C	gemma-7b	Llama-3-8B-Instruct	1,172	0.6109	0.6075	98	94	0.66	0.773	0.829	110,379
ARC-C	Llama-3-8B-Instruct	Llama-3-8B	1,172	0.6075	0.5922	81	63	0.74	0.134	0.156	4,081
ARC-C	gemma-7b	Llama-3-8B	1,172	0.6109	0.5922	100	78	0.68	0.099	0.115	3,375
HS	gemma-7b	Llama-3-8B	10,042	0.8247	0.8202	295	249	0.81	**0.049**	**0.054**	20,255
Wino	Mistral-7B-I-v0.2	Llama-3-8B	1,267	0.7719	0.7711	121	120	0.46	0.949	1.000	2,396,624
Wino	gemma-7b	Mistral-7B-I-v0.2	1,267	0.7845	0.7719	119	103	0.49	0.283	0.314	8,616
Wino	gemma-7b	Llama-3-8B	1,267	0.7845	0.7711	98	81	0.59	0.204	0.232	6,152

Table 7. Raw discordance and model identifiers for the nine MMLU-Pro top-10 adjacent-rank pairs (N=12,032). p_{χ^2} and p_{exact} are the asymptotic McNemar and exact conditional-binomial variants. $N^\star$ here is the IID McNemar–Connor value from Equation (6); cluster-adjusted values are in Table 4. Model names are the Open LLM Leaderboard v2 display names, lightly abbreviated for layout; full Hugging Face identifiers appear in the artifact at `experiments/a3_mmlupro_mcnemar.csv`.

Pair	Model A	Model B	$\hat{p}_A$	$\hat{p}_B$	b	c	$\hat{\rho}$	p_{χ^2}	p_{exact}	$N^\star$
1 vs 2	calme-3.2-78b	calme-3.1-78b	0.7303	0.7185	253	111	0.92	9.9×10^{-14}	7.1×10^{-14}	1,697
2 vs 3	calme-3.1-78b	CalmeRys-78B-Orpo	0.7185	0.7012	284	76	0.93	$<10^{-15}$	$<10^{-15}$	778
3 vs 4	CalmeRys-78B-Orpo	calme-2.4-rys-78b	0.7012	0.7002	32	20	0.99	0.096	0.126	34,092
4 vs 5	calme-2.4-rys-78b	Reflection-70B	0.7002	0.6341	1871	1076	0.46	$<10^{-15}$	$<10^{-15}$	433
5 vs 6	Reflection-70B	Arcee-Blitz	0.6341	0.6153	1680	1454	0.45	5.4×10^{-5}	5.8×10^{-5}	5,787
6 vs 7	Arcee-Blitz	Homer-Qwen2.5-72B	0.6153	0.6145	1449	1439	0.49	0.852	0.867	2,727,127
7 vs 8	Homer-Qwen2.5-72B	ultiima-72B-v1.5	0.6145	0.6054	352	242	0.90	6.4×10^{-6}	7.3×10^{-6}	4,628
8 vs 9	ultiima-72B-v1.5	Qwen2.5-72B	0.6054	0.5968	787	684	0.75	7.2×10^{-3}	7.8×10^{-3}	13,086
9 vs 10	Qwen2.5-72B	QwentileSwap	0.5968	0.5945	1227	1200	0.58	0.584	0.598	314,370

F. Leave-one-subject-out cluster robustness

Section 6.4 reports a $4/9 \rightarrow 6/9$ unresolved-count flip on MMLU-Pro top-10 adjacent pairs under real subject-level clustering (K=14 categories). To confirm the flip is not driven by any single category, we recompute the cluster-corrected unresolved count after dropping each MMLU-Pro category in turn and recomputing the per-pair ICC, design effect, and $N^\star_{\text{cluster}}$ on the remaining items.

Table 8 lists the LOSO results: across the 14 drops, the unresolved count stays at $6/9$ for 11 drops and at $5/9$ for three drops (health, law, psychology). No drop collapses the result below $5/9$, so the cluster-induced flip is not an artifact of a single high-ICC category.

Table 8. Leave-one-subject-out (LOSO) recomputation of the cluster-corrected unresolved-pair count on MMLU-Pro top-10 adjacent-rank pairs. Each row drops one of the 14 subject categories and reruns the §6.4 pipeline on the remaining items.

Dropped category	Unresolved
biology, business, chemistry, computer science, economics, engineering, history, math, other, philosophy, physics (11 drops)	6/9
health, law, psychology (3 drops)	5/9
Base (no drop, Table 4)	6/9

G. Cluster-bootstrap CIs for MMLU-Pro ICC

Section 6.4 reports a category-bootstrap stability check on the MMLU-Pro cluster-induced verdict flip. Table 9 lists the per-pair 5–95% bootstrap intervals on $\mathrm{ICC}(D)$, design effect, and cluster-corrected $N^\star$. Resampling protocol: at each of $B{=}1000$ iterations (seed 42), draw $K{=}14$ subject categories with replacement, recompute the one-way ANOVA ICC on the paired-difference series D_i using the resampled cluster structure, derive $\mathrm{DE} = 1 + (\bar{m} - 1)\,\mathrm{ICC}^+$, and report $N^\star_{\mathrm{cluster}} = N^\star_{\mathrm{IID}} \cdot \mathrm{DE}$. The IID inputs $(\hat{p}_A, \hat{p}_B, \hat{\rho}, N^\star_{\mathrm{IID}})$ are held fixed at their full-data estimates so the CIs isolate cluster-structure uncertainty (full-bootstrap CIs would only be wider). The "$\Pr(\text{unres.})$" column reports the fraction of B bootstraps in which $N^\star_{\mathrm{cluster}} > N{=}12{,}032$ for that pair.

Table 9. Cluster-bootstrap CIs on MMLU-Pro top-10 adjacent-rank pairs ($B{=}1000$, $K{=}14$ categories resampled with replacement). Brackets are 5–95% percentile bounds on the bootstrap distribution; "pt" is the full-data point estimate from Table 4. "$\Pr(\text{unr.})$" is the fraction of bootstrap iterations for which the pair is unresolved at $N{=}12{,}032$. Three pairs (rank 4 vs 5, 5 vs 6, 6 vs 7) drive the cluster-induced verdict tightening and have ICC bounds (bold) well above zero. The bootstrap distribution of the count-of-unresolved-pairs (out of 9) puts 44.9% mass at 5, 55.0% at 6, 0.1% at 4, and 0% above 6.

Pair	$\mathrm{ICC}(D)$ pt $[5, 95\%]$	DE pt $[5, 95\%]$	Cluster $N^\star$ pt $[5, 95\%]$	$\Pr(\text{unr.})$
1 vs 2	0.0004 [−0.0003, 0.0012]	1.37 [1.00, 1.95]	2,327 [1,697, 3,308]	0.000
2 vs 3	−0.0003 [−0.0008, 0.0001]	1.00 [1.00, 1.12]	778 [778, 868]	0.000
3 vs 4	0.0002 [−0.0005, 0.0008]	1.19 [1.00, 1.64]	40,660 [34,092, 55,765]	1.000
4 vs 5	**0.036 [0.021, 0.044]**	31.5 [17.4, 40.9]	13,621 [7,546, 17,672]	0.552
5 vs 6	**0.0067 [0.0026, 0.0104]**	6.74 [3.25, 9.78]	39,009 [18,800, 56,617]	0.997
6 vs 7	**0.012 [0.004, 0.020]**	11.5 [4.12, 17.7]	31.4M [11.2, 48.3]M	1.000
7 vs 8	−0.0002 [−0.0007, 0.0002]	1.00 [1.00, 1.15]	4,628 [4,628, 5,314]	0.000
8 vs 9	0.0010 [0.0002, 0.0016]	1.88 [1.14, 2.36]	24,632 [14,873, 30,865]	1.000
9 vs 10	0.0029 [0.0015, 0.0037]	3.52 [2.33, 4.08]	1.11M [0.73, 1.28]M	1.000

H. Prospective design validation and multi-arbiter agreement

Table 10 reports the prospective validation referenced in §6.1; Table 11 reports the multi-arbiter agreement check on every $|\hat{\delta}| \leq 2$pp pair.

Table 10. Prospective validation: empirical McNemar power at framework-prescribed $N^\star$ on three OLL v1 pairs ($M{=}1000$ bootstrap trials per cell; Monte Carlo SE ≈1.3 pp at the 0.8 target); columns probe sub-, on-, and super-prescription.

				Empirical power at		
Pair	$\lvert\delta\rvert$	$\hat{\rho}$	$N^\star$	$0.8\,N^\star$	$N^\star$	$1.2\,N^\star$
HS, Mistral-I / Llama-I	6.3pp	0.68	193	0.71	**0.83**	0.89
ARC, Llama-3-8B / gemma	7.8pp	0.54	294	0.69	**0.80**	0.88
HS, Llama-3-8B / gemma	10.1pp	0.57	120	0.72	**0.81**	0.88

Table 11. Four-arbiter agreement on the seven OLL v1 close pairs ($|\hat{\delta}| \leq 2$pp). R/F: reject/fail-to-reject at $\alpha{=}0.05$. The HellaSwag boundary pair (bold) is the only split verdict.

Pair ($\hat{\delta}$pp)	χ^2_1	exact	mid-p	CI$\ni 0$	verdict
ARC, gem/L-I (0.34)	.77	.83	.77	yes	FFFF
ARC, L-I/L-8B (1.54)	.13	.16	.13	yes	FFFF
ARC, gem/L-8B (1.88)	.10	.12	.10	yes	FFFF
HS, gem/L-8B (0.46)	**.049**	**.054**	**.049**	**yes**	**RFRF**
Wi, Mi/L-8B (0.08)	.95	1.00	.95	yes	FFFF
Wi, gem/Mi (1.26)	.28	.31	.28	yes	FFFF
Wi, gem/L-8B (1.34)	.20	.23	.20	yes	FFFF

I. Reporting checklist

Table 12. Reporting checklist for any paired LLM leaderboard claim. Each row is a direct consequence of an inversion in §3 or its multiplicity adjustment in §6.3.

Quantity	Definition	Why
$\hat{\delta}$	observed gap	headline effect
Paired test	McNemar / boot. / t	matches metric
N	paired prompts	budget
δ_{MDE}	Equation (4)	current-N resolution
$q = N/N^\star$	Definition 1	gap supported when $q \geq 1$
$N^\star$ CI	B=500 bootstrap	reveals whether $N^\star$ CI straddles N
Multiplicity	Bonf. / BH / none	required for family-level claims only
Per-item raw	$0/1$ matrix	third-party check

J. `llm-power` API

`cohens_h(p1, p2)` – Cohen's h for two proportions.
`paired_bootstrap_delta(scores_a, scores_b, ...)` – prompt-bootstrap CI on $\hat{\delta}$.
`bootstrap_power(...)` – empirical power against a reference score-matrix.
`parametric_required_n_proportions(p1, p2, paired, rho)` – shortcut n_h (Cohen convention).
`parametric_required_n_paired_binary(p1, p2, rho)` – $N^\star$ from Equation (6).
`required_n_mcnemar(n_ab, n_ba, n_observed)` – discordance-form McNemar required-N.
`parametric_required_n_paired(mean_diff, sd_diff)` – paired-t required-N for graded data.