# LEADERBOARDS RANK FINER THAN THEY RESOLVE

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

## ABSTRACT

Machine learning judges progress by ranking methods on shared benchmarks, but a benchmark is an instrument, with a resolution below which it returns noise. That resolution is almost never reported, so a reader cannot tell which leaderboard gaps are evidence and which are chance. Power analysis names it the minimum detectable difference: the smallest true accuracy gap a benchmark's own significance test detects at least eight times in ten. We ask three questions of it. First, how much of what the field publishes does its data support? Across 202 leaderboards in eight families, the median board ranks dozens of systems from first to last, yet its data can separate them into only two to five groups, depending on how strictly a group is drawn. Second, does a benchmark's resolution predict which of its published orderings hold when measured again? We label each accuracy ordering resolvable or not, then rescore it on MMLU subjects its label never saw. Orderings the benchmark cannot resolve reverse in 21.5% of cases, against 0.27% of those it can. With the items split at random instead, the same contrast holds on five further boards, one in software engineering. Reversals concentrate below a benchmark's resolution, which is computable in accuracy points before any replication exists. Third, is resolution a property of the benchmark alone, so that its authors could compute it once and print it on the benchmark's card? It is not. It also depends on how often the two compared systems disagree, a rate that belongs to the comparison rather than to the benchmark and cannot be reliably borrowed from a neighbouring board. Resolution has to be measured on the board itself, from the per-item outcomes its leaderboard already holds. This needs no new experiment, and leaderboards should report it.

## 1 INTRODUCTION

Machine-learning leaderboards compare systems by ordering their scores on a shared test set. The resulting ranking does not indicate whether the evaluation can reliably detect the differences between systems. Statistical power provides the relevant design criterion: the probability of rejecting a null hypothesis for a specified true effect. Power analysis therefore complements significance testing by quantifying the performance differences an evaluation can detect (Card et al., 2020).

The minimum detectable difference (MDD) expresses this criterion on the scale of accuracy. Under the sampling assumptions of a paired binary test, the threshold depends on the number of test items $n$ and the disagreement rate $d$: the probability that exactly one of the two systems is correct. Disagreement here concerns correctness, rather than differences between raw predictions. An observed gap below MDD may still be statistically significant; the threshold describes detection power rather than the significance of a particular observed comparison.

We apply established power analysis to examine benchmark resolution across 202 leaderboards. Card et al. (2020) previously evaluated minimum detectable effects on nine natural-language-processing benchmarks. Our contribution is an empirical study of resolution across evaluation families, its relationship to ranking stability, and the transferability of the disagreement rates needed to estimate it. The exact paired-test power is evaluated numerically; a normal approximation explains its dependence on sample size and disagreement.

Where per-item outcomes are available, we estimate disagreement directly. Elsewhere, we report resolution conditional on an assumed rate and examine sensitivity to that assumption. We then evaluate whether orderings below the threshold reverse more often on held-out items and whether



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

rates estimated on one benchmark transfer to another. Together, these analyses assess what resolution estimates contribute to the interpretation and design of benchmark comparisons.

**Contributions.**

1. We characterize resolution across 202 leaderboards from eight families. At a common reference disagreement rate, the median board has two single-linkage groups or five greedy tiers. Across boards, the median approximate item requirement for the observed top-two gap is 49 times the current size (Sections 3 and 7).
2. We evaluate ranking stability on items excluded from label assignment. In the subject-disjoint evaluation, orderings below MDD reverse in 21.5% of cases, compared with 0.27% above it. We examine this association using a second evaluation suite and split-half comparisons on five additional boards (Section 4).
3. We assess disagreement-rate transfer across seven benchmarks. Of forty-two ordered transfers, eleven retain 0.25 or less of the target-fitted regression's $R^2$. The rate also changes 1.21× between system populations on fixed items. These results support estimating disagreement for the benchmark and systems of interest (Section 5).

We adopt the power-analysis framework of Card et al. (2020) and the paired-sample sizing approach of Connor (1987). The contribution is the survey and empirical evaluation of ranking stability and rate transfer. Human-preference and non-binary generation-quality metrics require different statistical models and are outside the paired-binary analysis.

## 2 The minimum detectable difference

The **minimum detectable difference** (MDD) of a benchmark-and-protocol pair (Fig. 1a) is the smallest true accuracy gap its own significance test declares significant with at least 80% power at $\alpha = 0.05$. We examine this threshold through controlled perturbations and paired-test power calculations.

The perturbation analysis introduces defects of known size into a fitted model and measures the rejection rate. On four small classification datasets McNemar's exact test crosses 80% power at 1.44 accuracy points on digits ($n$ = 1,797), 2.49 accuracy points on breast-cancer ($n$ = 569) and 5.48 accuracy points on ABIDE ($n$ = 758), all lower bounds (the pooled outcomes are dependent). On ADNI ($n$ = 124) no gap in range reaches that power (Appendix Fig. 10).

For independently sampled paired binary outcomes, power depends on the true gap, test-set size $n$ and disagreement rate $d$, the share of items on which exactly one of two systems is correct. The normal approximation at 80% power and $\alpha = 0.05$ gives the sizing criterion $n \gtrsim (z_{0.975} + z_{0.80})^2\, d/\Delta^2$, based on Connor's paired-sample formula. Inverting this approximation gives an MDD proportional to $\sqrt{d/n}$. We instead obtain the exact-test threshold by numerically inverting the binomial-mixture power function. Both calculations require a measured or assumed disagreement rate as well as the item count (Appendix Fig. 9, Appendix Tables 3 and 5; Appendix A.1).

## 3 Power-based resolution of published leaderboards

We evaluated 202 leaderboards using the reference disagreement rate $d = 0.2549$, measured on HELM's MMLU items (Fig. 2). The median test-set size is 303 items and the median MDD is 8.35 accuracy points. These are conditional resolution estimates at a common rate. The 80%-power threshold differs from the significance boundary: under the normal approximation, a gap 1.43 times smaller can already reach significance.

We summarize each ranking using two grouping rules at this threshold. Greedy grouping starts a new tier when a score falls a full MDD below the current tier leader. Single linkage instead splits the ranking only at adjacent gaps of at least one MDD. The median leaderboard has five greedy tiers and two single-linkage groups; 98.9% of adjacent gaps fall below MDD. These groups describe separation at the specified power threshold, not equivalence within groups or simultaneous significance between them.

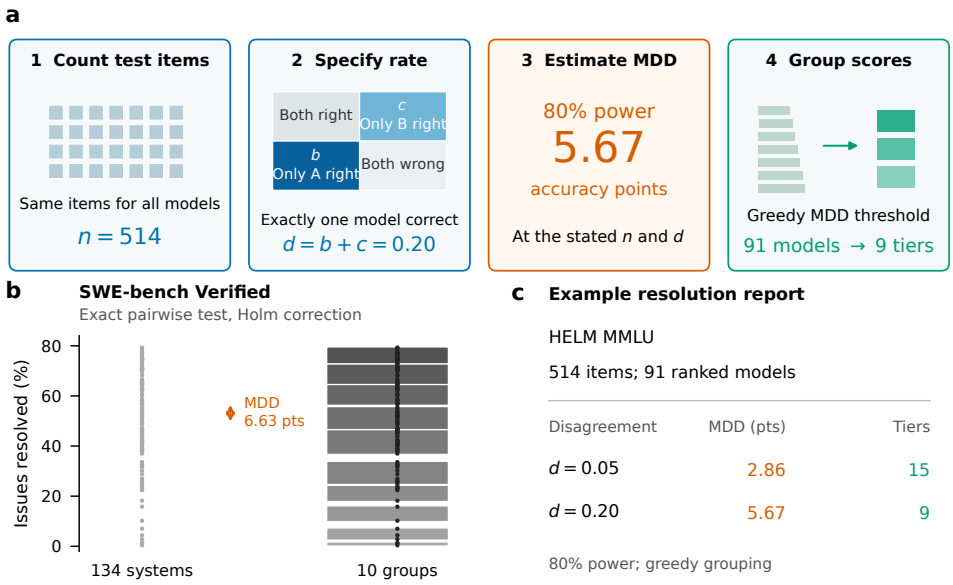



Figure 1: **Estimating and reporting benchmark resolution.** **a**, Worked example for HELM's MMLU evaluation. The test set contains $n = 514$ items. Correctness disagreement is $d = b + c$, where exactly one system is correct. At the illustrative rate $d = 0.20$, MDD is 5.67 accuracy points at 80% power. Greedy grouping starts a new tier when a score falls one MDD below the current tier leader, producing 9 tiers among 91 models. The measured rate 0.2549 instead gives 6.38 accuracy points and 8 tiers. The illustrative rate is not used in the survey; ten benchmark calculations are reported in Appendix Fig. 9. **b**, SWE-bench Verified: 134 systems and ten groups from exact pairwise testing with Holm correction over 8,911 pairs. These groups use a different criterion from the power-based tiers in **a**. The orange reference shows the per-comparison MDD of 6.63 accuracy points at the measured rate. **c**, Example resolution report. At $d = 0.05$, the limit is 2.86 accuracy points with 15 greedy tiers; at $d = 0.20$, it is 5.67 with 9 tiers. Reporting the rate and grouping criterion is necessary to interpret either count.

The survey contains 267 published tables across eight families. Removing 65 duplicate tables leaves 202 distinct boards: 182 from HELM, including 143 subsets of parent boards, and 20 outside HELM. We therefore report both populations separately. The median non-HELM test set contains 1,655 items, compared with 236 for HELM; the corresponding median MDDs are 3.58 and 9.44 accuracy points. Non-HELM boards have a median of 19.5 greedy tiers compared with 5 for HELM, and their score ranges span 22.8 detection limits compared with 5.27. Nevertheless, both populations have a median of 2 single-linkage groups, and the non-HELM population has a higher share of adjacent gaps below MDD. The prevalence of small adjacent gaps therefore extends beyond HELM.

Grouping and disagreement assumptions affect the estimated resolution. Across the 182 HELM boards, 78.5% of greedy tier boundaries join adjacent systems whose gap is below MDD. On the 80 HELM boards with measured disagreement, using each board's own rate changes the median greedy count from 5 to 4. Sensitivity intervals for the counts span approximately a factor of two. The median measured rate is 0.35; over the measured rate spread, the median MDD ranges from 7.1 to 11.6 accuracy points. The reference rate is below this median. Where outcomes are unavailable, a sensitivity range is therefore more informative than a single estimate based on a borrowed rate (Section 5).

Appendices B.2–B.5 report the family-specific analyses. Increasing the MMLU item count twenty-seven-fold increases its greedy tier count by a factor of 2.9. Vision has the finest estimated resolution, but still has fewer tiers than ranked systems. Interpretation of the embedding results is limited by incomplete metric provenance in the published archive. Three of the eight families release per-item outcomes, permitting exact pairwise tests and direct evaluation of ranking stability.

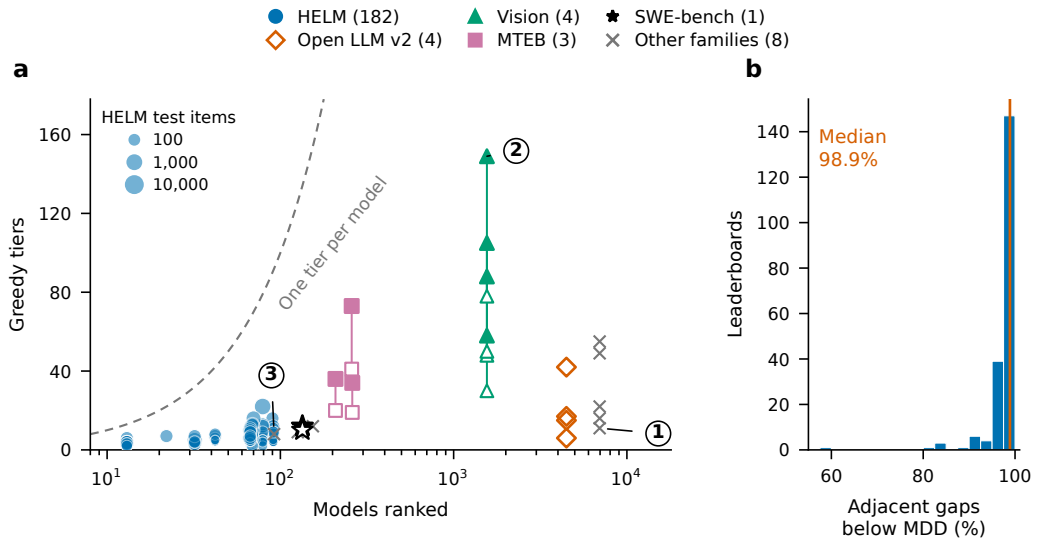



Figure 2: **Power-based resolution across published leaderboards. a**, Number of ranked systems (logarithmic horizontal axis) and greedy tier count (linear vertical axis). Each point is a leaderboard; colour and shape indicate the family. The dashed reference curve marks one tier per system. HELM marker area represents item count, with the size key shown in the panel; other families use fixed marker sizes. Most estimates use the reference disagreement rate $d = 0.2549$. For vision and MTEB, filled markers use $d = 0.05$ and linked open markers show the reference-rate sensitivity calculation, which is infeasible for some pairs at their accuracies (Methods). The filled SWE-bench star shows 12 power-based tiers at the reference rate; its linked open star shows ten groups from exact pairwise tests with Holm correction. These grouping criteria are distinct. Grey crosses represent the remaining surveyed boards. Callouts identify Winogrande (1; largest model count), ImageNet-R (2; largest tier count), and GAIA validation (3; smallest non-HELM test set). **b**, Fraction of adjacent score gaps below MDD, using the reference rate for all boards; the median is 98.9%. Greedy tiers use the gap from the current tier leader. Single-linkage grouping merges adjacent scores separated by less than MDD and cannot yield more groups at the same threshold. Estimated tiers are not confidence sets for ranks. Per-board values and grouping rules are reported in Appendix A.8.

## 4 Ranking reversals on held-out items

We evaluated whether the detection threshold distinguishes orderings that remain stable on held-out items. Each pair was labelled according to whether its observed accuracy gap exceeded MDD, then evaluated on items excluded from label assignment. We use "resolvable" and "unresolvable" for these power-based labels; neither label is a significance-test result.

HELM evaluates MMLU on 514 items from five of its 57 subjects. We matched 59 models evaluated on both this subset and the full 14,042-item benchmark. Labels were assigned on the 514 items and outcomes measured on the 52 subjects HELM does not use, comprising 13,528 items. The labelling items, 3.7% of the benchmark, were excluded from outcome evaluation. Among 1,711 model pairs, 15 were tied on the labelling items and excluded. Of the remaining 1,696, the 596 that the limit cannot separate reverse in 21.5% of cases; the 1,100 it can separate, in 0.27%. The two 95% bootstrap intervals are 15.72–27.75% and 0.00–1.06% (Fig. 3). We resample models to account for dependence among pairs: each model contributes to up to 58 comparisons. Labelling by significance instead gives 29.2% of 411 non-significant orderings (22.02–37.33%) against 0.86% of 1,285 (0.00–2.11%); the limit moves 185 of the latter to the unresolvable side and widens the separation.

Random item splits show the same association without a systematic subject shift. We split the 514 items into halves, assigned labels on one half and evaluated the ordering on the other. Reversal rates were 18.2% below MDD and 0.25% above it. Scored in sample, on HELM's own items, the two rates are 21.0% and 0.18%. Across 300 random panels of whole subjects drawn to the same size, unresolvable orderings reverse more often in every panel.

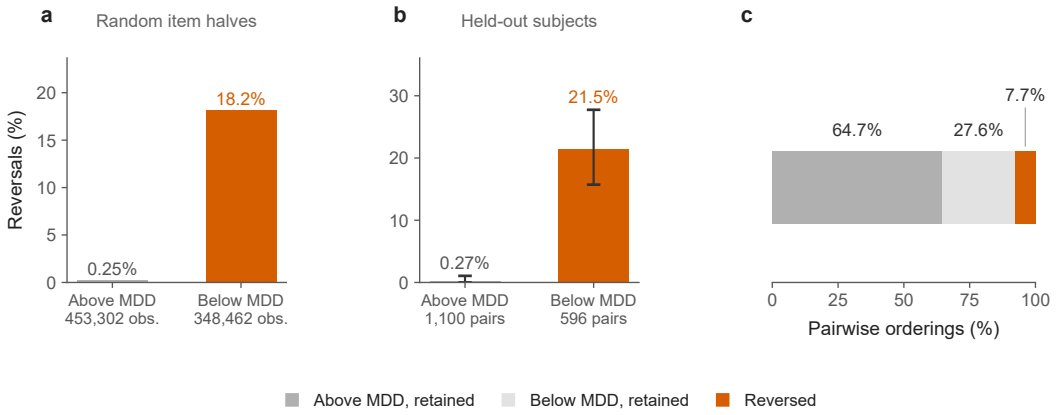



Figure 3: **Ranking reversals are more frequent below MDD. a**, HELM's 514 MMLU items are randomly divided into halves for 200 splits. Pairs are labelled using one half and evaluated on the other. Reversal rates are 0.25% above MDD and 18.2% below it; counts are pair-by-split observations. **b**, Orderings labelled on HELM's subset are evaluated on the 52 remaining MMLU subjects (13,528 items). Of 1,100 orderings above MDD, 0.27% reverse, compared with 21.5% of the 596 below it. Error bars show 95% model-clustered bootstrap intervals. **c**, Outcomes for the same 1,696 orderings; 7.7% reverse overall. Power-based labels describe detection at a specified true gap and are distinct from significance-test decisions. Details of the splits and bootstrap procedure are in Appendices A.10 and A.1.

We repeated the analysis using Open LLM v1, which evaluates MMLU over 111 models and 13,937 items with per-item outcomes. At that board's own measured rate, unresolvable orderings reverse in 20.6% of cases against 0.10% of resolvable ones, and more often in 300 of the 300 panels drawn. Scoring each panel's ordering on the subjects it did not draw, so that the label and the outcome share no item, gives 21.9% against the same 0.10%, on model-clustered intervals that do not overlap.

We extended the split-half evaluation to five additional boards with per-item outcomes. Labels use the sample size and pair-specific disagreement measured on the first half; reversal is evaluated on the second half. SWE-bench Verified ranks 134 agentic systems on 500 software engineering issues. There, unresolvable orderings reverse in 17.75% of cases against 0.09% of resolvable ones, on 95% model-clustered intervals of 15.99–19.83% and 0.07–0.11%. Four further language boards show the same contrast. HellaSwag gives 18.56% against 0.04%, on intervals of 14.91–23.20% and 0.03–0.06%; GSM8K 17.88% against 0.10%, on 5.15–35.15% and 0.02–0.25%; ARC 17.31% against 0.10%, on 15.41–19.87% and 0.08–0.13%; and Winogrande 18.94% against 0.19%, on 17.16–21.45% and 0.14–0.25% (`src/reversal_second.py`). These five split each board's items at random and so extend the test to further boards; only the two subject-disjoint tests above evaluate orderings on subjects excluded from label assignment.

MDD supplies a threshold in accuracy points from $n$ and $d$ without calibration on observed reversals. This interpretation does not establish superior predictive performance: a $\sqrt{n}$-scaled gap rule performs similarly, and slightly better on the 3.5% of cases where the rules disagree. The association also extends to item-response ability (Rodriguez et al., 2021): a board leaves fewer pairs unresolvable, yet those below that score's own limit still reverse in 17.5–20.0% of cases against 0.04–0.21% above it (Appendix A.12). Labelled instead at one assumed rate for every pair, anywhere from $d = 0.1855$ to 0.4947 (a limit of 5.46 to 8.84 accuracy points at $n = 514$), orderings below the limit reverse in 14.7–22.8% of cases against 0.00–0.61% above it on HELM, and in 19.4–25.8% against 0.03–0.21% on Open LLM v1, on intervals that do not touch; HELM's unresolvable count moves from 543 to 891 of the 1,696 orderings (`src/reversal_assumed_rate.py`). Both harnesses score MMLU, so the five assumed-rate families without per-item outputs are outside this test.

## 5 Transferability of disagreement-rate estimates

Disagreement depends jointly on the benchmark and the systems compared (Fig. 4). Aggregate accuracy scores do not determine it, although it can be measured in the three survey families with per-item outcomes. Within one evaluation suite, median rates range from 0.131 to 0.352 across boards: a factor of 2.68, corresponding to 1.64 in the approximate detection limit. On the same 514 MMLU items, the rates for 91-model and 111-model populations are 0.2549 and 0.3093, respectively, a 1.21× difference. Resolution estimates must therefore identify the system population as well as the test set.

We evaluated whether a regression fitted on one board could estimate rates on another. We extend the regression approach of Card et al. (2020), modelling disagreement using mean accuracy, the absolute accuracy gap, the product of the two accuracies, and the accuracy-implied upper bound on disagreement. We define *retained fit* as the transferred regression's target $R^2$ divided by the $R^2$ of a regression fitted on the target itself. A ratio of 1 matches target fitting; 0 corresponds to predicting the target mean. Across all 42 ordered transfers between seven boards, retained fit ranges from −24.664 to 0.990, with a median of 0.704; 19 transfers retain at least 0.75 and 11 retain at most 0.25 (Appendix Table 26; Appendix A.9).

Nine of the eleven low-retention transfers involve Winogrande, whose items offer two answer options. All four transfers with Pearson $r < 0.85$ also involve Winogrande; the minimum is 0.473. Among leave-one-board-out evaluations, excluding Winogrande yields correlations of 0.905–0.990 and median retained fit of 0.774 over the remaining 30 transfers. Controls for chance accuracy and extrapolation beyond the source accuracy range do not explain this pattern under the pre-specified thresholds. Resampling systems and refitting the regressions confirms that at least one transfer retains at most 0.25 in every resample (Appendix Table 30). The resampling analysis does not resolve the stronger claims that fewer than half retain 0.75 or that some transfers lose correlation.

Target-board recalibration improves typical fit but leaves substantial variation. With one target pair, retained fit is negative in 73% of draws; with 100 pairs, its 5th percentile is 0.537 (Appendix Table 32). Applying Card et al.'s published regression unchanged preserves correlation on all seven boards, but retained fit ranges from −0.757 to 0.900 (Appendix Table 27). When direct estimation is unavailable, resolution should therefore be reported across a plausible range of rates, subject to the accuracy-implied feasibility bounds.

The additional regression terms do not measurably improve transfer. Card et al.'s three-parameter form retains a median 0.734, compared with 0.704 for the full model; the difference is not distinguishable from zero. A two-parameter gap-only baseline retains a median 0.707 without target refitting, and each method performs better in 21 of the 42 transfers. The full regression outperforms this baseline in 19 of 30 transfers to multiple-choice target boards, compared with 2 of 12 transfers to free-response target boards. This difference between evaluation types persists under resampling. For generated answers, disagreement is already close to the accuracy gap, limiting the benefit of the additional terms (Appendix A.9).

Per-item outcomes also permit direct significance testing without an assumed rate. On HELM's MMLU subset, 46.9% of the 4,095 pairs among 91 models are not significant after correction for multiple comparisons. SWE-bench Verified evaluates 134 agentic systems on 500 issues and releases per-item outcomes (Appendix Fig. 13). Here, 25.8% of pairwise orderings are not significant under the same procedure, yielding ten groups; 129 of the 133 adjacent comparisons are not significant even before correction.

## 6 Related work

**Power analysis and sample size.** Card et al. (2020) estimate the minimum detectable effect of McNemar's test from test-set size and agreement, the complement of the disagreement rate used here, and report estimates for nine benchmarks. Our analysis adopts this framework. Bowman & Dahl (2021) discuss the implications for benchmark size, and Miller (2024) derive a related inversion. The paired-sample sizing approximation follows Connor (1987); exact sign-test constraints follow Dixon & Mood (1946). Minimum detectable differences also inform evaluation design in information retrieval (Sakai, 2016b). The paired-binary setting additionally constrains disagreement by the

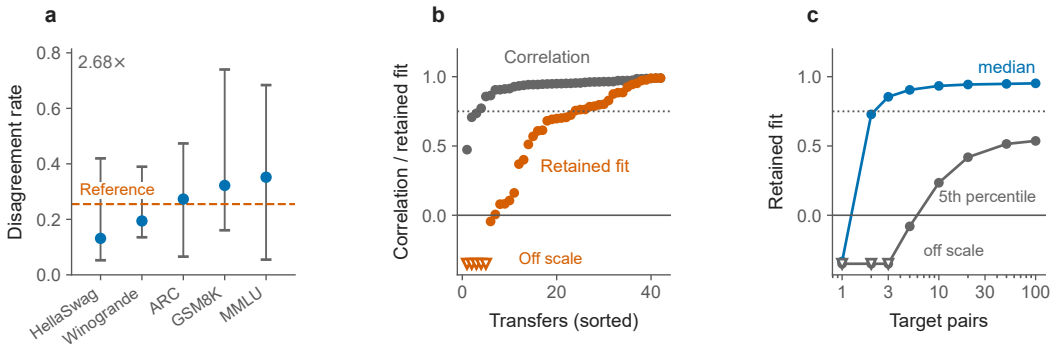



Figure 4: **Disagreement rates and the limits of transfer.** **a**, Median pairwise correctness disagreement on five Open LLM v1 boards evaluated with the same harness; bars span the 5th–95th percentiles across pairs. Medians differ by a factor of 2.68. The dashed line marks the reference rate $d = 0.2549$. **b**, All 42 ordered transfers between seven boards, sorted independently for each curve. Pearson correlation ranges from 0.473 to 0.990. Retained fit is transferred $R^2$ divided by target-fitted $R^2$; it ranges from −24.664 to 0.990 (median 0.704), with eleven transfers at or below 0.25. Open triangles mark the five values below the plotting range (−2.3 to −24.7), all from Winogrande. The dotted line marks 0.75. **c**, Median and 5th-percentile retained fit after recalibration using $k$ target pairs, over 200 draws per transfer. The median exceeds 0.75 at three pairs; the 5th percentile reaches 0.537 at 100 pairs. Open triangles mark off-scale percentiles of −2.5, −30.3 and −1.7 at one, two and three pairs, respectively.

accuracy gap, yielding a rate-independent lower bound on resolution. Appendix A.11 gives the methodological context, including the test-selection guidance of Dror et al. (2018).

**Uncertainty in rankings.** Critical-difference diagrams, WMT evaluations, Chatbot Arena and rank intervals summarize uncertainty in system comparisons (Demšar, 2006; Bojar et al., 2013; Chiang et al., 2024; Neuhof & Benjamini, 2026b). These approaches concern observed comparisons or plausible rankings. MDD instead specifies the true accuracy difference needed to attain a target power for a given test and sampling model. The distinction matters when choosing evaluation size. We report family-wise and false-discovery-rate corrections where multiple comparisons are interpreted together (Appendix A.11).

**Resolution, efficiency and stability.** Kotawala (2026) invert paired-test power to estimate per-pair sample requirements on two boards. Yauney et al. (2026) estimate detectable ability differences and assess whether smaller evaluations reproduce full-benchmark judgments; Perlitz et al. (2024) also study evaluation efficiency. Rank-uncertainty methods quantify uncertainty from per-observation scores (Neuhof & Benjamini, 2026a;b), and Chandrahas (2026) provide tools for corrected comparisons. Buckley & Voorhees (2000) examine how conclusion reversals depend on threshold and sample size. We assign power-based labels without fitting to reversal outcomes and evaluate their association with reversals on held-out items. Data-removal robustness (Huang et al., 2025) is a related but different criterion (Appendix D.7). Variation in prompts, decoding, seeds and hardware (Hochlehnert et al., 2025) lies outside the item-sampling model considered here.

**Disagreement estimation and transfer.** Previous work measures or models disagreement within evaluation settings (Card et al., 2020; Pembury Smith & Ruxton, 2020; Singh, 2026); Yang & Wang (2026) examine its implications for annotation. Our transfer analysis evaluates whether a relationship estimated on one benchmark retains predictive fit on another. The contributions are therefore the survey across 202 leaderboards, the held-out reversal evaluations, and the empirical assessment of disagreement-rate transfer, rather than a new power statistic.

## 7 Discussion

The survey and held-out evaluations identify complementary limitations of leaderboard rankings. Many adjacent gaps are below the specified power threshold, and these orderings reverse more often on held-out items. Adding systems to a leaderboard does not increase the number of evaluated items. Under the normal approximation, MDD decreases with the square root of item count and increases with the square root of disagreement: a threefold change in disagreement changes MDD by approximately 1.7-fold. Resolution therefore depends on the evaluation design as well as the number of systems ranked.

The per-comparison thresholds do not control error across a complete ranking, which contains a median of 2,346 comparisons per board. Using a Bonferroni threshold of $\alpha = 0.05/\binom{M}{2}$ for $M$ systems, 24 of the 182 distinct HELM boards cannot reach 80% power at any feasible effect size under the specified disagreement rate. None of those 24 holds more than 65 items, so dropping them leaves a median limit of 8.25 accuracy points against the 9.44 Section 3 reports over all 182. Across the 158 that remain, the correction raises that median to 13.67 accuracy points, cuts the median board from 5 tiers to 4, and takes the boards supporting no distinction from 0 to 4. A stricter threshold can only widen a limit and shrink a tier count, so this bounds the optimism of the per-comparison limits.

Power analysis also specifies the sample size required for a target gap. At 80% power, exact evaluation for a gap of two accuracy points requires 3,424 items at $d = 0.17$, the median rate between successive SWE-bench record-holders. The corresponding normal approximation scales quadratically with the inverse gap. Of the 202 boards, 11 meet the exact item requirement for two accuracy points and 5 for one accuracy point. Resolving a board's own top two systems is harder still: treating observed gaps as design targets, the all-board median approximate requirement is 49-fold the current item count. On 34 of the 202 boards, the top two observed scores are tied. A zero target gap has no finite required sample size; this does not establish that the systems' population accuracies are equal.

The estimates have four scope limitations. First, five of the eight families lack per-item outcomes and require assumed disagreement rates. Predicted power can also have substantial per-pair error despite a small average signed bias. Across 200 random splits, estimating the rate and gap on half the items gives absolute power errors of 9.8–15.4 on the 0–100 power scale; using the full item set reduces these to 5.0–7.9. The latter evaluation uses information from the items on which rejection is measured. Second, tier counts are summaries of an estimated threshold, rather than confidence sets for ranks. Third, the sampling model does not capture run-to-run variation in training or inference. Fourth, the reversal results do not establish consequences for model adoption: the tested proxies—likes, thirty-day downloads and declared base models—did not pass their controls (Appendix A.10).

We recommend reporting resolution estimates with the test, target power, disagreement assumptions and grouping rule. Per-item outcomes enable direct estimation and pairwise tests; multiplicity correction is required when interpreting those tests simultaneously. These additions make the scope of a ranking explicit without requiring new model training.

## 8 Conclusion

Published leaderboards frequently report adjacent score differences below their estimated minimum detectable difference. In evaluations with per-item outcomes, orderings below this threshold reverse substantially more often on held-out items than orderings above it. Power-based resolution therefore provides information about ranking stability that a score alone does not convey, while remaining distinct from a test of an observed difference.

Resolution depends on the test-set size, evaluation protocol and systems being compared. Disagreement rates do not transfer consistently across benchmarks, so estimates should state their rate assumptions and use measured per-item outcomes wherever available. Reporting these estimates alongside scores, uncertainty intervals and appropriately corrected pairwise tests would clarify the precision of benchmark comparisons and inform the sample sizes required for future evaluations.

### AI USE STATEMENT

In this work, we used generative AI tools to implement methods (the analysis code: the leaderboard harvest, the exact limit evaluator, the reversal and transport tests, and the figure scripts); to design and give feedback on methodology and experiments, including robustness tests that were then run; to clean and reformat the harvested leaderboard data; to help formulate the mathematical claims of the theory appendix and draft their proofs; to generate the synthetic data of the simulation grids and implanted defects, through code they wrote; and to interpret results. Translation and qualitative or thematic data analysis are not applicable to this work. Additionally, we used generative AI tools to draft and edit the text, to prepare figures, to search and summarise the literature, and to format references. An automated checker compared numerical claims with stored artifacts, with coverage evaluated using planted defects (Reproducibility statement). Independent AI reviewers also checked the claims and proofs against the stored results. We take responsibility for the final content of this work, including text, claims and artifacts produced with the aid of generative AI.

### ETHICS STATEMENT

The survey, reversal and transport analyses use only public leaderboard outputs (HELM, the Open LLM Leaderboard, SWE-bench Verified, BigCodeBench, GAIA, MTEB and the image-classification suites) under their published terms; they contain no personal data. The implanted-defect experiments of Section 2 use two de-identified human-subject neuroimaging datasets, ADNI and ABIDE, obtained under their data use agreements. We made no attempt to re-identify participants and report only aggregate statistics. Generative AI tools had no access to the ADNI or ABIDE data: the code for those datasets was written against a documented schema and run on institutional compute, and only aggregate results were returned.

### REPRODUCIBILITY STATEMENT

The supplementary material contains the analysis code and the stored result artifacts from which the reported numbers are read, with a README that maps scripts to artifacts. Before submission, an automated checker compared the numerical claims and regenerated table bodies of the manuscript source with these artifacts, and the table values and decimal figures of this text were confirmed identical to that source; the checker reads the manuscript source and is not included. The artifacts support inspection of the reported results independently of model training or leaderboard collection. The appendix specifies the exact paired-test power calculation, its normal approximation, disagreement assumptions, grouping rules and evaluation splits. Per-board results, pairwise comparisons and transfer estimates are reported in the supplementary tables; Appendix A.8 provides supporting analyses.

**Checker validation.** Artifact perturbation tests detected 45 of 47 predefined defects. Prose perturbation tests detected 25 of 28 numeric changes in the evaluated front-matter sample, excluding comments, citation years, reference targets and graphics options. A detection required the build to fail; artifact tests additionally required a diagnostic identifying the corrupted file. The three undetected prose changes concerned the target power, significance level and illustrative disagreement rate. These detection rates apply to the tested defect classes and do not establish complete coverage.

**Reproduction limits.** Twelve analysis scripts require per-item caches that are not distributed, and four retrieve data from changing leaderboard APIs. Their stored outputs permit verification of reported results, but do not ensure that a new data collection will reproduce the same leaderboard snapshot.

### ACKNOWLEDGMENTS

Data collection and sharing for the Alzheimer's Disease Neuroimaging Initiative (ADNI) is funded by the National Institute on Aging (National Institutes of Health Grant U19AG024904). The grantee organization is the Northern California Institute for Research and Education. In the past, ADNI has also received funding from the National Institute of Biomedical Imaging and Bioengineering, the Canadian Institutes of Health Research, and private sector contributions through the Foundation for the National Institutes of Health (FNIH) including generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech;

BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics.

## REFERENCES

T. W. Anderson and S. M. Samuels. Some inequalities among binomial and Poisson probabilities. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pp. 1–12. University of California Press, 1967.

B. M. Bennett and R. E. Underwood. On McNemar's test for the $2 \times 2$ table and its power function. *Biometrics*, 26(2):339–343, 1970. doi: 10.2307/2529083.

Ondřej Bojar, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. Findings of the 2013 workshop on statistical machine translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation (WMT)*. Association for Computational Linguistics, 2013.

Samuel R. Bowman and George E. Dahl. What will it take to fix benchmarking in natural language understanding? In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pp. 4843–4855. Association for Computational Linguistics, 2021. URL https://aclanthology.org/2021.naacl-main.385/.

Tamara Broderick, Ryan Giordano, and Rachael Meager. An automatic finite-sample robustness metric: when can dropping a little data make a big difference? arXiv preprint, 2020.

Chris Buckley and Ellen M. Voorhees. Evaluating evaluation measure stability. In *Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 33–40, 2000. doi: 10.1145/345508.345543.

Dallas Card, Peter Henderson, Urvashi Khandelwal, Robin Jia, Kyle Mahowald, and Dan Jurafsky. With little power comes great responsibility. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 9263–9274. Association for Computational Linguistics, 2020. URL https://aclanthology.org/2020.emnlp-main.745/.

Ben Carterette and Mark D. Smucker. Hypothesis testing with incomplete relevance judgments. In *Proceedings of the 16th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 643–652, 2007. doi: 10.1145/1321440.1321530.

Shreyas K Chandrahas. evalci: A python library for statistically rigorous comparison of language model evaluations, 2026.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios N. Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael I. Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024.

Pierre Colombo, Nathan Noiry, Ekhine Irurozki, and Stephan Clémençon. What are the best systems? new perspectives on NLP benchmarking. In *Advances in Neural Information Processing Systems 35 (NeurIPS)*, 2022.

Robert J. Connor. Sample size for testing differences in proportions for the paired-sample design. *Biometrics*, 43(1):207–211, 1987. doi: 10.2307/2531961.

Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. MS MARCO: Benchmarking ranking models in the large-data regime. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 1566–1576, 2021. doi: 10.1145/3404835.3462804.

Amir Dembo and Ofer Zeitouni. *Large Deviations Techniques and Applications*. Springer, 2 edition, 2010. doi: 10.1007/978-3-642-03311-7. Sanov's theorem and the Cramér rate function.

Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

W. J. Dixon and A. M. Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946. doi: 10.1080/01621459.1946.10501898.

Rotem Dror, Gili Baumer, Marina Bogomolov, and Roi Reichart. Replicability analysis for natural language processing: Testing significance with multiple datasets. *Transactions of the Association for Computational Linguistics*, 5:471–486, 2017. doi: 10.1162/tacl_a_00074. URL https://aclanthology.org/Q17-1033/.

Rotem Dror, Gili Baumer, Segev Shlomov, and Roi Reichart. 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), Volume 1: Long Papers*, pp. 1383–1392. Association for Computational Linguistics, 2018. URL https://aclanthology.org/P18-1128/.

Rotem Dror, Segev Shlomov, and Roi Reichart. Deep dominance — how to properly compare deep neural models. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 2773–2785. Association for Computational Linguistics, 2019. doi: 10.18653/v1/P19-1266.

Stephen W. Duffy. Asymptotic and exact power for the McNemar test and its analogue with R controls per case. *Biometrics*, 40(4):1005–1015, 1984. doi: 10.2307/2531151.

Morten W. Fagerland, Stian Lydersen, and Petter Laake. The McNemar test for binary matched-pairs data: mid-p and asymptotic are better than exact conditional. *BMC Medical Research Methodology*, 13:91, 2013. doi: 10.1186/1471-2288-13-91.

Andreas Hochlehnert, Hardik Bhatnagar, Vishaal Udandarao, Samuel Albanie, Ameya Prabhu, and Matthias Bethge. A sober look at progress in language model reasoning: Pitfalls and paths to reproducibility. In *Conference on Language Modeling (COLM)*, 2025.

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

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, 2025.

Bahram Jafrasteh, Ehsan Adeli, Kilian M. Pohl, Amy Kuceyeski, Mert R. Sabuncu, and Qingyu Zhao. Statistical variability in comparing accuracy of neuroimaging based classification models via cross validation. *Scientific Reports*, 15(1), 2025. doi: 10.1038/s41598-025-12026-2.

Tom Kocmi, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Anton Dvorkovich, Christian Federmann, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Benjamin Marie, Christof Monz, Kenton Murray, Masaaki Nagata, Martin Popel, Maja Popović, Mariya Shmatova, Steinthór Steingrímsson, and Vilém Zouhar. Findings of the WMT24 general machine translation shared task: The LLM era is here but MT is not solved yet. In *Proceedings of the Ninth Conference on Machine Translation (WMT)*. Association for Computational Linguistics, 2024.

Anany Kotawala. Resolution diagnostics for paired LLM evaluation, 2026. ICML 2026 Workshop on Hypothesis Testing.

Peter A. Lachenbruch. On the sample size for studies based upon McNemar's test. *Statistics in Medicine*, 11(11):1521–1525, 1992. doi: 10.1002/sim.4780111110.

John M. Lachin. Power and sample size evaluation for the McNemar test with application to matched case-control studies. *Statistics in Medicine*, 11(9):1239–1251, 1992. doi: 10.1002/sim.4780110909.

Mitchell Levine and Mary H. H. Ensom. Post hoc power analysis: An idea whose time has passed? *Pharmacotherapy*, 21(4):405–409, 2001.

Quinn McNemar. Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, 12(2):153–157, 1947. doi: 10.1007/BF02295996.

Olli S. Miettinen. The matched pairs design in the case of all-or-none responses. *Biometrics*, 24(2): 339–352, 1968. doi: 10.2307/2528039.

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

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks, 2026a.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation, 2026b.

David Otero, Javier Parapar, and Álvaro Barreiro. Towards reliable testing for multiple information retrieval system comparisons. In *Advances in Information Retrieval (ECIR 2025), Part II*, Lecture Notes in Computer Science, pp. 424–439, 2025. doi: 10.1007/978-3-031-88711-6_27.

Art B. Owen. Empirical likelihood ratio confidence intervals for a single functional. *Biometrika*, 75 (2):237–249, 1988. doi: 10.1093/biomet/75.2.237.

Matilda Q. R. Pembury Smith and Graeme D. Ruxton. Effective use of the McNemar test. *Behavioral Ecology and Sociobiology*, 74(11):133, 2020. doi: 10.1007/s00265-020-02916-y.

Yotam Perlitz, Elron Bandel, Ariel Gera, Ofir Arviv, Liat Ein-Dor, Eyal Shnarch, Noam Slonim, Michal Shmueli-Scheuer, and Leshem Choshen. Efficient benchmarking of language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), Volume 1: Long Papers*, pp. 2519–2536. Association for Computational Linguistics, 2024.

Nils Reimers and Iryna Gurevych. Why comparing single performance scores does not allow to draw conclusions about machine learning approaches. *arXiv preprint arXiv:1803.09578*, 2018.

Pedro Rodriguez, Joe Barrow, Alexander Miserlis Hoyle, John P. Lalor, Robin Jia, and Jordan Boyd-Graber. Evaluation examples are not equally informative: How should that change NLP leaderboards? In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pp. 4486–4503, 2021.

Patrick Royston. Exact conditional and unconditional sample size for pair-matched studies with binary outcome: a practical guide. *Statistics in Medicine*, 12(7):699–712, 1993. doi: 10.1002/sim.4780120709.

Tetsuya Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 5–14, 2016a. doi: 10.1145/2911451.2911492.

Tetsuya Sakai. Topic set size design. *Information Retrieval Journal*, 19(3):256–283, 2016b. doi: 10.1007/s10791-015-9273-z.

Tetsuya Sakai. *Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power*. The Information Retrieval Series. Springer Singapore, 2018a. ISBN 978-981-13-1198-7. doi: 10.1007/978-981-13-1199-4.

Tetsuya Sakai. Topic set size design for paired and unpaired data. In *Proceedings of the 2018 ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR)*, pp. 199–202, 2018b. doi: 10.1145/3234944.3234971.

Mark Sanderson and Justin Zobel. Information retrieval system evaluation: effort, sensitivity, and reliability. In *Proceedings of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 162–169, 2005. doi: 10.1145/1076034.1076064.

M. Anthony Schork and George W. Williams. Number of observations required for the comparison of two correlated proportions. *Communications in Statistics - Simulation and Computation*, 9(4): 349–357, 1980. doi: 10.1080/03610918008812161.

Amogh Singh. Certifying compressed language models: An audit and a statistical toolkit. *arXiv preprint arXiv:2608.15046*, 2026.

Julián Urbano. Test collection reliability: a study of bias and robustness to statistical assumptions via stochastic simulation. *Information Retrieval Journal*, 19(3):313–350, 2016. doi: 10.1007/s10791-015-9274-y.

Julián Urbano, Harlley Lima, and Alan Hanjalic. Statistical significance testing in information retrieval: An empirical analysis of Type I, Type II and Type III errors. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 505–514, 2019. doi: 10.1145/3331184.3331259.

Ellen M. Voorhees and Chris Buckley. The effect of topic set size on retrieval experiment error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 316–323, 2002. doi: 10.1145/564376.564432.

William Webber, Alistair Moffat, and Justin Zobel. Statistical power in retrieval experimentation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM)*, pp. 571–580, 2008. doi: 10.1145/1458082.1458158.

Eddie Yang and Dashun Wang. Benchmark illusion: Disagreement among LLMs and its scientific consequences. *arXiv preprint arXiv:2602.11898*, 2026.

Gregory Yauney, Shahzaib Saqib Warraich, and Swabha Swayamdipta. How reliable is language model micro-benchmarking? In *International Conference on Learning Representations (ICLR)*, 2026.

# A Methods

## A.1 Minimum detectable difference

For a specified benchmark, evaluation protocol and model pair, the minimum detectable difference (MDD) is the smallest true difference in the reported metric for which the protocol's significance test rejects the null hypothesis with at least 80% probability at $\alpha = 0.05$. We estimated MDD using controlled model perturbations and an exact power calculation.

*Controlled model perturbations.* The reference estimator was $\ell_2$-penalised logistic regression at $C = 1$, with features standardised within each training fold. We perturbed the fitted estimator by a controlled amount and measured the resulting performance gap. To assess sensitivity to the reference estimator, we repeated the label-corruption procedure with a linear support-vector machine, a random forest and a multilayer perceptron; the estimated limits changed little across these estimators (Appendix Table 9b). Boundary perturbations were restricted to the two linear estimators, for which a separating weight vector is available. We used two distinct perturbations: displacement of the fitted weight vector by a fraction $\delta$ of its norm along a fixed random direction, without refitting; and corruption of a fraction $\delta$ of training labels followed by refitting. The first modifies the construction of Jafrasteh et al. (2025), who perturb both models symmetrically to retain a zero true difference and assess false rejections. Perturbing only one model instead produces a non-zero difference for power estimation. Urbano et al. (2019) similarly control the underlying difference using simulated information-retrieval data fitted to TREC runs, then estimate Type I, Type II and Type III error rates. Our perturbations provide controlled differences without simulating the evaluation data. For each value of $\delta$, we recorded rejection frequency over 100 independent stratified 10-fold splits. MDD was the interpolated crossing of 80% power. The two perturbation mechanisms differed in power by at most 11.6 percentage points over gaps of 2 to 8 accuracy points, supporting approximate robustness to the perturbation mechanism over this range.

*Exact power evaluation.* McNemar's test depends on the paired predictions through the test-set size $n$ and two rates, $b$ (first model right, second wrong) and $c$ (the reverse), with gap $b - c$ and disagreement $d = b + c$. The power is then a finite mixture over the discordant count, $\pi(n, \Delta, d) = \sum_m \binom{n}{m} d^m (1-d)^{n-m} \beta_m(\theta)$ with $\theta = \frac{1}{2} + \Delta/2d$. We obtain $\mathrm{MDD}(n, d)$ by bisection on this finite sum, without simulation or model training. The corresponding normal approximation gives the planning requirement $n \geq (z_{0.975} + z_{0.80})^2 d/\Delta^2$; the finite-sample calculation uses the exact mixture instead. Monte Carlo estimates of the same quantity differ by at most 0.07 accuracy points for the values regenerated by the survey scripts and do not change the qualitative conclusions. The released exact evaluator supplies the exact-test quantities reported here. The exact power of McNemar's test is classical (Schork & Williams, 1980; Duffy, 1984)—Schork & Williams evaluate the unconditional mixture above by multinomial enumeration over all outcomes—and Carterette & Smucker (2007) inverted the conditional power to a required trial count; we use this inversion. Solving for the $d$ at which the closed form reproduces an experimentally measured limit fits one parameter to one observation and therefore provides no independent goodness-of-fit test (Remark 6): in accuracy points, digits gives 1.44 measured against 1.42 at an implied $d = 0.045$, breast-cancer 2.49 against 2.50 at $d = 0.043$, and ABIDE 5.48 against 5.47 at $d = 0.278$. The implied rates can instead be compared with rates measured from per-item outcomes as a plausibility check. The implied rate lies within the measured range for digits and just above that range for breast-cancer. No directly measured rate is available for ABIDE.

The controlled-perturbation design and power curves are shown in Appendix Fig. 10. Exact-test survey calculations are reported in Appendix Fig. 9 and Appendix Table 3; implied-rate comparisons are reported in Appendix Table 5.

## A.2 Leaderboard survey

We surveyed eight leaderboard families. Inclusion required published per-model scores and an evaluation size verifiable against the benchmark's source data. Leaderboards with scores but no verifiable evaluation size were excluded.

*HELM* (247 tables, 182 distinct boards, six suites). Every group table was examined and retained only if it reports a per-item binary outcome and a single instance count shared by all models. Thirteen group

files report a mean over their sub-scenarios in that column rather than a total — HELM's MMLU summary shows 102.8 where the benchmark comprises 514 items, and 246.4 where it comprises 14,042. Taken as sizes without summing over their sub-scenarios, those means understate $n$ and overstate the limit on those thirteen boards by up to 7.6×. We therefore summed the sub-scenario counts. The retained boards span 30 to 14,042 items. The median tier count (5) and median share of adjacent positions below the detection limit (98.7%) are unchanged whether these thirteen tables use the published counts, corrected counts or are excluded. Two additional suites contained no eligible tables.

*HELM boards with a measured rate.* A board is said to release enough per-item outcomes to estimate its disagreement rate when it satisfies the screening criteria in `src/helm_measure_d.py`. The release must publish at least three run directories for the scenario. At least three of the twelve runs sampled from them must expose a `per_instance_stats.json` with a statistic taking values in $\{0, 1\}$ over more than twenty instances. The statistic's name must not be on a fixed exclusion list of dataset and runtime properties: `num_references`, for instance, counts an item's gold answers, is identical for every model, and so yields a rate of zero by construction rather than an outcome. At least one model pair must share twenty or more instance identifiers. The median disagreement rate must also be non-zero; the screening procedure treats a zero median as evidence of a column that is constant across models. We recorded the failed criterion for every excluded board. Eighty of the 182 distinct boards passed this screening and form the population used for Appendix Table 17 and the reported HELM per-board rates.

*Open LLM Leaderboard* v2 (4,476 of 4,576 models retrieved) and v1 (6,960 of 7,260), counted separately because they share no benchmarks and use different harnesses. Sizes were verified from each source dataset's evaluated split. Scores are taken from the raw columns; the headline columns are rescaled for chance, which breaks the binomial model.

*Vision*: 1,556 classifiers from the timm ImageNet results, with sizes verified against mirrors of the evaluation sets because the published tables carry accuracies but no instance count.

*Scores that are not unweighted item means.* We checked whether each score lay on the $1/n$ lattice implied by the reported item count: a mean over $n$ binary outcomes must be an integer multiple of $1/n$. Scores from ARC, HellaSwag, Winogrande, GSM8K, the vision boards and SWE-bench are 100% consistent, and MATH Lvl 5 and MMLU-PRO 99.9%. The three MTEB boards are 3.5%, 5.8% and 5.0% consistent, Open LLM v1's MMLU 13.2%, and Open LLM v2's IFEval and MuSR 3.7% and 2.4%: those six are not unweighted means over items — MTEB's are means over repeated evaluation runs, v1's MMLU is a 57-subject macro-average, IFEval mixes prompt- and instruction-level accuracies, and MuSR is a macro-average over three subsets of 250, 256 and 250 items — so McNemar's decomposition into discordant pairs does not strictly apply to them. For the macro-averages the size and the sign of the error can be measured. A macro-average gives an item in a subject of $n_s$ items the weight $1/(Sn_s)$, so the sampling error of its gap exceeds that of the unweighted mean the closed form assumes; under a common per-item variance the ratio of the two standard errors is $\sqrt{A/H}$, the arithmetic over the harmonic mean of the subject sizes. On Open LLM v1's MMLU, where per-item outcomes let it be measured pair by pair with each subject's own rates, the macro gap's standard error is 1.19 times the unweighted one at the median pair (1.16–1.23 over the central 90% of pairs, above one on every pair), compared with the equal-variance reference of 1.21. For HELM's five similarly sized subjects, the corresponding measured ratio is 0.98. The closed form therefore *understates* that board's limit, which is 1.43 accuracy points and not 1.20, and the board resolves 47 tiers and not 55 (`src/mx36_macro_average_bound.py`); Appendix Table 11 reports the uniform closed-form estimate, whereas the text uses the corrected estimate. Agreement between the measured and equal-variance ratios supports using the latter as an approximation when outcomes are unavailable. For MuSR's three similarly sized subsets the ratio is 1.0001, with negligible effect on the results. For MTEB's run-averages and IFEval's mixed accuracies no bound is available, since the closed form omits run-to-run variance and MTEB is read at an assumed rate; bounding them would take per-run scores and per-prompt outcomes, which neither publishes, so those limits remain approximate. No alternative denominator up to 15,000 makes IFEval or MuSR consistent with unweighted item means (14.5% and 13.0% at best). Dropping the three MTEB boards leaves the survey's medians unchanged at 5 tiers and 98.7%; it removes the family that supplies the top of the reported range of supported shares, so the range is reported with MTEB included. HELM's MMLU is a fourth case and a benign one: its published score is a macro-average, but its per-item outcomes are released and every

micro-average lies exactly on the $1/514$ lattice. The exact tests use those outcomes rather than the published macro-average.

*MTEB*: computed from a stored parquet snapshot rather than the live query interface, which returns different row sets between identical calls—two runs gave one task 200 and 208 models. The snapshot is released with the artifacts.

*SWE-bench Verified, BigCodeBench, GAIA*: 134 systems, 154 and 126 on BigCodeBench's two splits, and 92. GAIA's size is confirmed from the published scores themselves, every one being an exact multiple of $1/165$.

We excluded tables with incompatible metrics or unverifiable evaluation sizes. These comprised seven HELM tables whose models were scored on differing instance counts—the LegalBench parent table and its `abercrombie` and `corporate_lobbying` subsets (50 against 95 and 490 items), `wikifact`, `pubmed_qa`, and the contrast-set variants of `boolq` and `imdb`—while the three LegalBench subsets scored on a single count throughout are retained; F1 and BLEU tables (not per-item binary); TruthfulQA in v1 (MC2 scores probability mass, not per-item correctness); ToxicConversationsClassification in MTEB (originally excluded on an unsupported metric inference, discussed below); GPQA (gated); BBH (leaderboard evaluates an unverified subset); ImageNet-Sketch (mirrored size disagrees with the cited figure); and ImageNet-ReaL (re-labels the same images as ImageNet val). The ToxicConversationsClassification split contains 3,954 positive examples among 50,000, giving majority-class accuracy of 92.1%, whereas the lowest archived score is 12.6%. Scores below that baseline do not identify the metric: a classifier can perform worse than the majority rule. Current task metadata specify accuracy as the main score, so the original inference that the archived values represented average precision does not justify exclusion. The task remains outside the reported quantitative survey, limiting its MTEB coverage; no conclusion about that task's resolution follows from this exclusion.

We formed tiers by traversing the ranking from the highest score and starting a new tier when a model fell at least one detection limit below the current tier's leader. At a uniform threshold, this greedy partition has the fewest tiers subject to every within-tier pair lying below the threshold; its count also equals the largest set of systems separated pairwise by at least that threshold. This interpretation assumes a common limit for all pairs. In practice, between the 5th and 95th percentiles the disagreement rate spans 0.169–0.543 on HELM's MMLU and 0.140–0.561 on SWE-bench, so the pair-specific limit runs roughly 0.24–1.70 times the one used. The upper-bound reading requires the limit used to lie below the *smallest* pair-specific limit on the board; on SWE-bench it does not, and the count at the median rate (11) understates the largest pairwise-distinguishable set (15). Under heterogeneous rates the tier count is neither an upper nor a lower bound on pairwise resolution. Single linkage provides a complementary partition: at the same threshold it never produces more tiers and gives fewer on all but three boards at the reference rate. All between-tier pairs then exceed the threshold, a property the greedy partition has only when the below-threshold relation is transitive. Where per-item outcomes are released, we also apply the greedy procedure using the Holm-corrected exact test rather than a single limit, so on SWE-bench Verified and HELM's MMLU the reported counts are the stricter of the two. Power-based and significance-based tiers have different interpretations. The normal approximation places the 80% power boundary 1.43 times above the significance boundary. A pair within a power-based tier may therefore still be separated by the exact test, as 5 of the 34 Open LLM record advances (each a new leader against the one it displaced) are. Single-comparison and Holm thresholds also give different tier counts on the same board: SWE-bench Verified resolves nineteen tiers at the single-comparison threshold and ten under Holm correction, and an exact-test tier count quoted without qualification is the corrected one. Families without per-item outputs use the closed form at the reference rate $d = 0.2549$, measured on HELM's MMLU items; vision and MTEB are also read at the optimistic 0.05.

Family-specific data and results are detailed in Appendices B.2, B.3, B.4 and B.5.

### A.3 Exact pairwise testing

For HELM Lite's MMLU, per-item outcomes were retrieved for every run and filtered to the test split; the released files also contain validation instances (111 rows where the leaderboard reports 100), and mixing them would inflate $n$. Runs were resolved through each table row's recorded run identifier, never by matching model display names. Each model's reconstructed accuracy was compared against

the leaderboard's own published value and the model discarded on any disagreement; all 91 matched to within $10^{-6}$. We applied McNemar's exact conditional binomial test to all 4,095 pairs. Because a ranking asserts all pairwise orderings simultaneously, Holm-corrected results are reported alongside uncorrected ones. The uncorrected analysis consequently reports more distinguishable pairs.

HELM's published MMLU score is the mean of the five subject accuracies rather than the mean over all 514 items: reconstructed the first way it reproduces every published value exactly, and the second way it differs by up to $5.7 \times 10^{-3}$. The per-item outcomes are binary regardless, and every micro-average lies on the $1/514$ lattice, so McNemar's decomposition applies to the pairwise test even though the headline score is a macro-average. Across the 4,095 pairs, the share not reaching $\alpha = 0.05$ is 22.34% uncorrected, 23.42% under Benjamini–Hochberg and 46.94% under Holm. Pairs whose accuracies are exactly tied—29 on this board and 36 on SWE-bench—have no observed score difference and are counted as unsupported, the convention throughout. Because the reported statistic counts *non*-rejections, it is largest under Holm, the strictest correction, and smallest uncorrected; all three are reported. The exact conditional test is conservative (Fagerland et al., 2013), which on its own would inflate the non-rejection share; replacing it with the mid-$p$ variant moves the uncorrected share by $-0.93$ percentage points and the Holm share by $-0.54$. The effect of test conservativeness on these shares is therefore smaller than the effect of multiplicity correction.

*Comparison with published sign-test calculations.* We checked the exact evaluator against the independently computed table of Dixon & Mood (1946), which reports the number of sign observations needed to detect a given departure from an even split 95% of the time (their Table 2, p. 562). Under McNemar the discordant count is $m \sim \mathrm{Bin}(n, d)$ and, conditional on it, the sign test operates at $\theta = \frac{1}{2} + \Delta/2d$, so their table is this specification in the corresponding sign-test parameterisation: $n = m/d$, and as $\Delta/d \to 0$ their requirement becomes $(z_{1-\alpha/2} + z_\tau)^2\, d/\Delta^2$, the algebraic limit of Definition 2 solved for $n$. Two comparisons are possible and they answer different questions. Our *algebraic* expression is the $\sigma_1 = \frac{1}{2}$ limit of Dixon and Mood's own normal approximation, so comparing the two on their large-$n$ cells — which they computed by that approximation, and which are italicised in the original — provides an algebraic consistency check. The median ratio is 1.009 over the twelve cells nearest an even split; this agreement is expected from the shared approximation. The *exact* evaluator provides a separate numerical check, because their Table 2 corresponds to the conditional sign test at a fixed discordant count. Inverting it at their $\alpha$ and at 95% power reproduces 26 of their 36 cells to the integer, including all 21 cells with $n \le 66$. The first discrepancy occurs at $n = 106$: all 24 cells below this value agree, as do two of the twelve cells with $n \ge 106$. This partition by the first discrepancy is a retrospective summary. The agreeing cells have a printed-to-exact ratio of 1.000; the ten discrepant cells have ratios of 0.981–0.997. Of the ten misses, 6 fall in the two rows Dixon and Mood's own footnote marks as computed by approximation ($p_0 = 0.40$ and 0.45); the remaining 4, at $p_0 = 0.30$ and 0.35, do not, and their discrepancies remain unexplained. Approximation beyond the rows identified in the footnote is one possible explanation. The published table and our exact evaluator therefore agree on every cell below $n = 106$, and six of their ten disagreements lie in rows the table itself marks as approximated. The same paper also supplies the six-disagreement minimum (p. 561) below which the exact test cannot reject at any outcome, which our floor adopts.

We also applied exact testing to SWE-bench Verified. Each submission's results file lists the instance identifiers it resolved; correctness is membership in that list, scored over the benchmark's canonical 500 instances taken from the dataset rather than from any submission, since SWE-bench reports resolution as a fraction of all 500 regardless of how many a system attempts. All 134 submissions were usable and all 8,911 pairs were tested exactly.

On SWE-bench Verified, 25.8% of all pairs are non-significant after correction. Before correction, 129 of 133 adjacent pairs (97.0%) are non-significant, with a median adjacent gap of 0.400 accuracy points. After correction all 133 adjacent pairs are non-significant, and the 134 systems form ten greedy tiers.

### A.4 Benchmarks measured directly

The controlled perturbation experiments used ADNI (124 participants, balanced to 62 per class, one session each, 19,701 features), ABIDE (758 participants, 16,110 features), breast-cancer (569 samples, 30 features) and digits (1,797 samples, 64 features). The ADNI and ABIDE connectome pipeline is shared with a companion manuscript under review. We also used twenty synthetic benchmarks

in which sample size and class separation varied independently. Because ADNI is longitudinal, we retained one session per participant before cross-validation. Cohort balancing used a seed independent of the repeat seed, ensuring the same cohort across job shards.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (`adni.loni.usc.edu`). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The original goal of ADNI was to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). The current goals include validating biomarkers for clinical trials, improving the generalizability of ADNI data by increasing diversity in the participant cohort, and providing data concerning the diagnosis and progression of Alzheimer's disease to the scientific community. For up-to-date information, see `adni.loni.usc.edu`.

### A.5 Statistics and reproducibility

All tests were two-sided at $\alpha = 0.05$. The paired $t$-test used 10 per-fold accuracies (9 d.f.). McNemar's test was the exact binomial test on discordant pairs pooled across the out-of-fold predictions from all ten folds. These tests therefore differ in their unit of analysis as well as in their form, ten fold means against up to 1,797 items, and the gap between their limits reflects that choice of unit as well as the statistic. Pooling can make the McNemar limits optimistic: the ten models have overlapping training data and items in one test fold share a fitted model, so the paired outcomes are not independent. The measured limits should therefore be interpreted as optimistic estimates (Remark 1). Power is the proportion of 100 independent repeats in which the test rejected; every power estimate rests on exactly 100 repeats, run as sharded jobs and pooled from raw per-repeat indicators rather than shard means. Uncertainty on MDD is 10,000 bootstrap resamples of the repeats, resampled as rows of a repeats-by-magnitude matrix so the pairing from a shared split survives. Where power never reaches 80% no value is reported, the largest swept gap is quoted as a bound, and the share of resamples that also fail to reach 80% power is given.

**Power target and sampling error.** Unless stated otherwise, all limits use the conventional 80% power target. At 50% power a limit of two accuracy points needs 1,633 items and a limit of one accuracy point needs 6,530; at 90% power they rise to 4,466 and 17,863. At 50%, 80% and 90% power the median board has 98.5%, 98.9% and 99.9% of its adjacent positions below the limit. This share is algebraically determined by the number of systems and separated blocks, so its small change does not independently validate a choice of power target. The tier count and the item requirement do respond to the power target, and they move by a factor of 2.7 across this range (Appendix Table 14). Sampling error in the disagreement rate is measurable where per-item outcomes exist: bootstrapping SWE-bench's 134 *systems* rather than its 8,911 pairs — the pairs are not independent, and resampling them understates the dominant component, since one system enters up to 133 pairs at once — gives $\hat{d} = 0.268$ with a standard error of 0.0070 (2,000 resamples), and SWE-bench's own tier count moves between 11 and 12 across $\hat{d} \pm 2$ standard errors. Transport is the larger uncertainty: the rate is measured in the three families that release per-item outcomes and borrowed for the other five.

Predictions for every adjacent pair on the twelve sizing boards are recorded in the pre-registration for evaluation against future entrants.

### A.6 Data and code availability

**Data.** All leaderboard sources are public: HELM results and per-item outputs, the Open LLM Leaderboard contents datasets, the timm ImageNet result tables, the MTEB results archive, the SWE-bench experiments repository, and the BigCodeBench and GAIA result sets. Retrieval commands are in the released code, and every derived quantity is released as an artifact, including the MTEB snapshot. ABIDE is shared through the Preprocessed Connectomes Project; breast-cancer and digits are included in scikit-learn; synthetic benchmarks are generated by seeded code. ADNI data are available from `adni.loni.usc.edu` under a data use agreement prohibiting redistribution of participant-level data; none appears here or in the artifacts, and derived participant-level data can be

released only through LONI-IDA. All aggregate quantities behind every figure and table are released as artifacts.

**Code.** The anonymised supplementary material contains the analysis scripts and stored result artifacts; the numerical consistency checker reads the manuscript source and is not included. The planned public release includes the controlled-perturbation procedure, the closed-form limit, the leaderboard collection and exact tests, and a checker that compares covered numerical claims with stored artifacts and exits non-zero on a mismatch. Numerical literals not covered by a stored artifact are reported separately without causing the checker to fail.

### A.7 Power-based labels and the gap-rule control

The initial within-benchmark and disjoint-subject analyses label pairs by whether their comparison reaches $\alpha = 0.05$. The main paper instead uses power-based labels: "cannot resolve" denotes an observed gap below the limit at 80% power, which under the normal approximation is 1.4294 times the gap at which the test rejects; the exact boundary is a step function of $n$ and $d$ and the ratio varies with both. The two criteria disagree on 185 of the 1,695 pairs, 10.9%, in one direction only, and relabelling by the limit itself *widens* the separation: 0.18% reversal among the 1,100 pairs it calls resolvable against 21.0% among the 595 it does not, a ratio of 115.6 against 41.1 under the significance label; on the disjoint-subject analysis, 0.27% against 21.5% (0.00–1.06% and 15.72–27.75%, the dyadic model-clustered intervals of the main text). We report both labels.

To assess whether the power-based rule adds predictive information beyond gap size, we compared it with a threshold on the observed accuracy gap. Ranking the same 1,695 pairs by accuracy gap alone and cutting at the same bin size reproduces the same split — 28.78% against 0.70% — because $n$ is fixed within a benchmark and, at a common disagreement rate, the limit is itself a gap threshold. We also compared the rules across sample sizes. Pooling draws at $n = 64$, 128, 256 and 514, a gap rule *scaled by* $\sqrt{n}$ disagrees with the limit on 3.5% of observations and, where they differ, is the better of the two, separating reversals 36.3-fold against the limit's 33.0 and admitting 0.91 times as many reversals among pairs each calls resolved (0.88 on the random split). Since $\mathrm{MDD}(n, d) \approx z\sqrt{d/n}$, the two rules differ only by $\sqrt{d}$, which provides no improvement in this comparison. The scaled rule uses a common threshold across all draws and sample sizes. It declares a pair unresolvable when $|\widehat{\Delta}|\sqrt{n}$ falls below a single cut, and the cut is fixed by rank rather than chosen: it is the smallest value at which the rule flags as many of the pooled observations as the limit flags, the matched-selectivity convention already used for the unscaled gap rule above. Equivalently the threshold is $c/\sqrt{n}$ accuracy points with the same $c$ at every size, and $c$ is determined by the limit's own flag count, so nothing in the comparison is tuned to the reversal outcome (`src/reversal_audit.py`).

### A.8 Survey accounting and composition

This section specifies the survey populations and the sensitivity of its summary statistics to disagreement rates, partition rules and board composition.

**Board-specific disagreement rates.** Per-item outcomes support a board-specific disagreement estimate for 80 of the 182 distinct HELM boards. Their median rate is 0.35, compared with the reference rate of 0.2549. Across the measured distribution (5th–95th percentile, 0.19–0.49), the median limit ranges from 7.1 to 11.6 accuracy points, the median tier count from 6 to 4, and the median share of adjacent positions below the limit from 98.5 to 100.0%. These comparisons use the same 165 boards, for which all rates yield defined limits (Appendix Table 17). Because the reference rate is below the median measured rate, it assigns at least as much resolution as the median rate on every board. If each rate instead uses all boards with a defined limit, the upper end of the median limit is 13.1 accuracy points. This larger value reflects the inclusion of 17 small boards that are excluded at the lowest rate.

Under the normal approximation, a threefold change in disagreement changes the limit by approximately 1.7, reflecting its $\sqrt{d}$ dependence.

**Accounting for 267 tables and 202 boards.** The survey comprises 267 published leaderboard tables from eight families, representing 202 distinct boards after removing 65 duplicated HELM

boards. The families cover language, vision, embeddings, code and agentic software engineering. Test-set sizes were obtained from benchmark documentation and scores from the published tables. HELM accounts for 182 of the 202 boards, including 143 sub-scenario slices whose parent tables are often also included. We therefore report survey statistics both with and without HELM. The number of distinguishable tiers varies tenfold across families, but the supported fraction of the published ordering remains between 0.3% and 9.0% under a common standard: single comparisons at the reference rate. In every family, the median board has at least 94% of adjacent positions below the detection limit. Combining results computed under different standards instead gives an upper endpoint of 14%: this mixes an optimistic disagreement rate for one family with Holm-corrected exact tests for another and should not be interpreted as a like-for-like comparison.

**Disagreement-rate ranges in three model populations.** Three Open LLM v1 populations are used in the analyses. The four boards with cached per-item outputs have a 2.57× disagreement-rate range. The record-holder populations on five boards have a 2.68× range (Appendix Table 13). The 111-model population used in the cross-board and subject-panel experiments has a 2.60× range across two boards. These values describe different model populations; the main text reports the range among record holders.

**Sensitivity of tier counts to pair-specific rates.** A tier count based on one disagreement rate is an estimate of resolution, not a general bound. Pair-specific rates span 0.169–0.543 and 0.140–0.561 between the 5th and 95th percentiles on HELM's MMLU and SWE-bench Verified, respectively. Using each pair's own rate, exact maximum clique calculations identify 8 mutually distinguishable systems on HELM's MMLU, compared with a uniform-rate count of 9. The corresponding counts on SWE-bench Verified are 15 and 11. Under uncorrected exact tests, the maximum clique sizes are 11 and 19; under Holm correction they are 6 and 10. A uniform-rate count can therefore overstate or understate resolution, depending on the comparison standard. Across the 80 HELM boards with measured rates, the median count is 5 at the reference rate and 4 at each board's own rate, with a range of 4 to 6 when each board's rate varies between its 5th and 95th percentiles (Appendix Table 17). The counts vary within roughly a factor of two across these analyses, without changing the broad between-board comparisons.

**Model-population and item-panel effects on disagreement.** Disagreement can differ even when model populations are evaluated on the same items. A direct comparison of HELM's 514 MMLU items with Open LLM v1's 13,937 items confounds item coverage with model population. Restricting Open LLM v1's 111 models to HELM's five subjects holds the item panel fixed. On the same 514 items, HELM's 91 models have a disagreement rate of 0.255 and Open LLM v1's models have a rate of 0.309, a population-related factor of 1.21 and a 1.10-fold difference in the limit. The Open LLM v1 rate changes little between this panel and the full item set (0.309 versus 0.316 on all 13,937 items), indicating that the model population accounts for most of the difference. Among successive record holders, the rate is 0.352, 1.38× the HELM reference.

**Relationships among resolution summaries.** The share of adjacent positions below the detection limit is algebraically determined by the single-linkage block count. By Prop. 57, $a = 1 - (k^{\mathrm{SL}} - 1)/(M - 1)$, where $k^{\mathrm{SL}}$ is the number of separated blocks. The identity holds to 0.0000 percentage points on every board. For example, $M = 69$ and $k^{\mathrm{SL}} = 2$ imply $a = 98.53\%$, compared with the observed median of 98.7%. The adjacent-position share therefore does not provide independent evidence beyond the block count. On the median board, the score range $R$ spans approximately five detection limits $L$: the median $R/L$ is 5.3, with a range of 1.1–48.2. Single linkage partitions the median board into two separated blocks. The greedy tier count is bounded by the same score range, $k \leq 1 + \lfloor R/L \rfloor$, with a median ratio to that bound of 0.83; we therefore report $R/L$ alongside the count. Across the 182 distinct HELM boards, 78.5% of greedy tier boundaries (658 of 838) separate two adjacent systems whose gap is below the threshold. Single linkage has no such boundaries and satisfies $k^{\mathrm{SL}} = M - a(M - 1)$; its median count is 2 compared with 5 for the greedy partition, and 80 of 182 boards form one block. Because single linkage can connect systems through chains of small gaps, the two partitions answer different questions and are both reported.

**Effects of survey composition.** Of the 202 boards, 182 are HELM tables and 143 are sub-scenario slices whose parent tables are often also counted. The implications of this composition differ across

summary statistics (Appendix Table 19). The median share of adjacent positions below the limit increases from 98.9% across all boards to 99.94% across the 20 non-HELM boards, although the least resolvable individual board is from HELM. In contrast, sample-size adequacy depends strongly on composition. Across all 202 boards, 12 meet the algebraic two-accuracy-point target (6%); the counts are 4 of 182 HELM boards and 8 of 20 non-HELM boards (40%). The all-board median of 303 items largely reflects HELM sub-scenarios: the HELM median is 236, compared with 1,655 for non-HELM boards. Removing every slice leaves 59 boards, with the same 12 and 5 meeting the two- and one-accuracy-point targets and a median size of 1,000 items. The median factor of additional items required for leader certification decreases from 49 across all boards to 17 across non-HELM boards and to 25 after removing only the slices. Thus, the severity of the sizing shortfall is influenced by the many small sub-scenarios, whereas the high prevalence of unresolved adjacent positions persists outside HELM. Resolving every adjacent pair is a substantially more demanding target, requiring a median 167-fold increase in items (89–480-fold).

Restricting HELM to its 39 scenario-level boards gives a median of 565 instances and 32 models, five greedy tiers and 97.1% of adjacent positions below the limit (Appendix Table 19). The prevalence of unresolved adjacent positions therefore persists after excluding the smaller sub-scenario tables.

### A.9 Controls, null models and procedural cautions

The following analyses compare the reported effects with explicit controls and null models, and specify the scope of the pre-registered transfer criterion.

**Recalibration and the target-fit denominator.** Recalibration improves some transfers but does not eliminate the largest failures. Across the seven boards, a two-parameter affine refit raises the minimum retained ratio (recalibrated fit divided by the target's own fit) to 0.231, whereas a slope-only refit gives −0.511. In the three-board analysis, the corresponding minima are 0.886 for an affine refit and 0.769 for an intercept refit. Two of the four quantities in Appendix Table 25 are mathematically identical: for a single predictor, the $R^2$ of an OLS affine refit equals the squared Pearson correlation. This equality holds to machine precision for all 42 transfers, so the table contains three distinct summaries. Under MDD $\propto \sqrt{d}$, borrowing a rate produces median discrepancies of 12.6% in the limit and 23.8% in the item requirement. Each retained ratio uses the target's own *in-sample* fit as its denominator; an optimistic target fit would therefore lower the ratio. Replacing that denominator with a 5-fold cross-validated fit changes no ratio by more than 0.1% (Appendix Table 26). Of the 42 transfers, 19 attain at least 0.75 and 11 are at or below 0.25. The latter 11 determine the pre-registered non-transferability classification.

**Comparison with a gap-only predictor.** To test whether the fitted disagreement map improves on a simpler predictor, we fitted a two-parameter rule using only $|a_i - a_j|$ and an intercept on the source board and transferred it in the same way. The gap-only rule achieves a median retained ratio of 0.707, compared with 0.704 for the full map, and performs better on 21 of the 42 transfers (interval 17–24). The difference between the medians is −0.003, with an interval from −0.105 to +0.077, providing no evidence of an overall advantage for the full map. Relative performance differs by target response format, and this difference persists under resampling. On multiple-choice targets, the full map performs better on 19 of 30 transfers (16–22), with a median of 0.700 compared with 0.677 for the gap-only rule. On free-response targets—SWE-bench patches and GSM8K arithmetic answers—it performs better on 2 of 12 transfers (1–5), with medians of 0.734 and 0.779, respectively. The full map's performance is similar across formats (0.700 versus 0.734); the larger difference is in the gap-only rule (0.677 versus 0.779). Free-response tasks lack a fixed guessing floor and shared multiple-choice distractors, which is consistent with the observed closer relationship between disagreement and the accuracy gap. The results support the gap-only rule as a competitive predictor on the evaluated free-response boards, rather than a universal replacement for the fitted map.

**Exact tests on four Open LLM v1 boards.** Cached per-item outputs support exact tests for every model pair on four Open LLM v1 boards: GSM8K (14 models, 1,319 items), ARC (6, 1,170), HellaSwag (111, 10,019) and MMLU (110, 13,937). Their disagreement rates range from 0.1243 to 0.3194, a 2.57-fold variation within one leaderboard family. Estimates based on only five models per board give a larger range of 3.4×, showing that the estimated range also depends on the model population.

**Resolution at different model-population sizes.** For the five-model HellaSwag comparison, all 10 pairs and all four adjacent positions are distinguishable; three of four adjacent positions are distinguishable on MMLU. In the 111-model comparisons, the exact test separates 21 of 110 adjacent positions on HellaSwag and 20 of 110 on MMLU. After Holm correction over the 6,105 comparisons on each board, these counts decrease to 12 and 9, with 34 and 25 greedy tiers, respectively. Thus, ten-thousand-item evaluations can support an ordering of five systems while leaving many adjacent comparisons unresolved in a larger population.

**Scope of the pre-registered transfer criterion.** The pre-registered criterion uses the minimum retained ratio across the six ordered transfers among three boards with sufficient per-item outcomes (Appendix Table 25). A minimum of at least 0.75 is classified as transportable and a minimum of at most 0.25 as non-transportable. This minimum cannot increase when additional boards are included. The clustered interval for the transfer that determines the three-board minimum, $[-0.195, 0.315]$, crosses the 0.25 boundary. Without recalibration, that transfer is the only one of the six at or below 0.25, and 2 reach 0.75. All six reach 0.75 after either refit, with minima of 0.886 for the affine refit and 0.769 for the intercept refit. Because rates vary across boards and model populations, a borrowed rate is an assumption about the comparison, rather than a fixed property of a benchmark. Limits computed at borrowed rates should therefore be accompanied by a range across plausible rates. We use $d = 0.2549$ as a common reference within the measured range and provide per-board estimates in Appendix Table 13 for recalculation.

**Successive records on SWE-bench Verified.** SWE-bench Verified archives submissions by arrival date. Ordering all 134 submissions chronologically and retaining each running maximum gives 33 records and 32 successive score increases. Exact tests against the preceding record fail to distinguish 24 of 32 increases; after multiplicity correction, the count is 29 of 32 (Fig. 12).

**Arrival-order permutation as a test of aggregate progress.** Permuting submission order tests whether the observed record sequence is consistent with random arrival. Across random orders, 134 systems produce 4.9 records on average, compared with the observed 33 ($P < 0.0005$). The observed record increments have a median of 1.70 accuracy points, a mean of 2.34 and a maximum of 8.80. The unusually long record sequence supports aggregate progress even though most individual increments are not statistically distinguishable. These results address different scales of change: the benchmark records substantial cumulative improvement, while the 75% non-rejection fraction describes its ability to distinguish successive increments on 500 items.

**Held-out power estimation and cross-board rate substitution.** Resampling a board's items using its own estimated rates reproduces predicted power to within 0.8 percentage points. This agreement is a computational consistency check: the resampling distribution is the same empirical law used by the power calculation. Estimating the disagreement rate and gap on one half of the items, then measuring rejection frequency on the other half, provides a separate test of parameter estimation. Across all pairs, without selection by predicted power, the mean signed errors are −0.3, −0.1, −1.1 and −2.8 percentage points of power in the four board-and-size conditions. Mean per-pair |error| ranges from 9.0 to 14.5 percentage points (Appendix Table 18). Thus, small average bias coexists with substantial pair-level error. Restricting evaluation to a band selected by predicted power would preferentially select overestimates and induce regression to the mean; we therefore report the unselected results.

Split variation is substantial. Across 200 splits, mean held-out per-pair error ranges from 9.8 to 15.4 percentage points across the four conditions, with a 95% split range of 8.2–18.3. Estimating both the rate and the gap from the full item set reduces the error to 5.0–7.9 percentage points, approximately half the held-out value in every condition. Using only the full-set *gap* recovers essentially all of this reduction (6.96 versus 7.09 at HELM $n = 100$); using only the full-set *rate* provides no reduction and is slightly worse in all four conditions.

The reduction reflects overlap with the tested items, rather than sample size alone. A sample-size-only explanation based on $1/\sqrt{m}$ predicts a ratio of 0.707. For equal halves, however, $\hat{d}_{\text{full}} = (\hat{d}_{\text{half}} + \hat{d}_{\text{tested}})/2$ exactly. An error proportional to the weight assigned to items outside the tested half predicts ratios of 1, $2/3$, $1/2$ and 0 for estimates based on the untested half, that half plus half the tested items ($3n/4$), the full set, and the tested half alone. Across 200 splits, the measured ratios are 0.713–0.723, 0.514–0.516 and 0.087–0.123 for the last three conditions, compared with predictions

of $2/3$, $1/2$ and 0 (Appendix Table 31). All four full-set intervals contain 0.500 and exclude 0.707. The two conditions using $n/2$ items differ only in item identity, separating item overlap from sample size. The $3n/4$ condition remains 7–8% above its prediction in all four cases, while the tested-half condition has a non-zero residual consistent with Monte Carlo noise.

The tested-half calculation cannot assess model misspecification. Its parameters describe exactly the empirical population from which the draws are sampled, so the calculated power is the expectation under that design. With 200 draws per pair, signed residuals are +0.013, +0.003, +0.007 and −0.008 percentage points, compared with absolute errors of 0.9–1.7. These nearly zero-mean residuals quantify simulation noise within this construction. The held-out error of 9.8–15.4 percentage points is consequently attributable to parameter estimation within the specified resampling model; the experiment does not establish the accuracy of that model for other sampling processes. Likewise, estimates from the full item set are unsuitable for evaluating calibration on unseen data, because they include the tested half. They are appropriate for the distinct task of predicting what a subsample of an already observed benchmark will resolve. For that task, the results favour estimating both $d$ and $\Delta$ from all available benchmark items, with the main improvement attributable to the gap estimate.

One split shows residual bias on SWE-bench at $n = 200$: $-2.8$ percentage points, decreasing in magnitude as the item count increases ($-4.2$, $-2.8$, $-1.3$). The closer relationship between disagreement and gap on free-response boards offers a possible explanation, because errors in the two plug-in estimates may then fail to offset one another. This explanation is not established: across splits, the largest bias sometimes occurs on the multiple-choice HELM board at $n = 200$. A test of the response-format hypothesis would require a pre-registered comparison across boards.

Five of the eight surveyed families require a disagreement rate borrowed from another board. We evaluated this substitution where rates were measured on both source and target. At a 2.6× mismatch between Open LLM v1's HellaSwag and MMLU, borrowing the higher rate increases false-unresolvable classifications 2.4×. Borrowing the lower rate produces the only false-resolvable classifications in these tests: 8 of 6,105 pairs. For boards with closely matched rates, either direction changes the error by at most a tenth of the within-board error (Appendix Table 22). Under MDD $\propto \sqrt{d}$, overestimating disagreement raises the limit, whereas underestimating it can assign resolution that the measured rate does not support.

Across twenty synthetic benchmarks with independently varied size and class separation, MDD/SE has a through-origin slope of 4.05 and ranges from 2.58 to 9.82. The ratio increases with class separation, as expected from $\text{MDD/SE} = Z\sqrt{d/p(1-p)}$. The slope uses the McNemar column of the synthetic results, not the paired-$t$ column, for which this identity does not apply (Appendix Table 4). We report the relationship rather than a through-origin $R^2$, which would be inflated by the shared $1/\sqrt{n}$ scaling.

**Target-board measurements for recalibration.** A transferred disagreement map requires target-board measurements to calibrate its scale. We refitted the map using $k$ anchor pairs and evaluated it on pairs excluded from recalibration. A single anchor produced a fit worse than predicting the target mean in 73% of draws, with a median retained ratio of $-0.33$. The median exceeds 0.75 at $k = 3$, but the 5th percentile remains below this threshold throughout the evaluated range: it is 0.235 at ten anchors and 0.537 at a hundred. Thus, no evaluated anchor count provides consistently high retained fit across the sampled boards and draws (Appendix Table 32).

**Reversal rates across random subject panels.** We assessed sensitivity to HELM's choice of five MMLU subjects using 300 random subject panels of approximately the same ∼514 items. Each model pair was labelled using the closed-form limit at the panel's own size. Below-limit orderings reversed in 18.5% of cases, compared with 1.3% for above-limit orderings (Wilson intervals 18.3–18.6% and 1.30–1.37%, respectively). The below-limit reversal rate was higher in 300 of the 300 panels. Applying the same procedure to HELM's five subjects gave 18.5% and 0.0%, indicating that the qualitative separation is not specific to HELM's subject selection. This separation is also not unique to MDD: a $\sqrt{n}$-scaled gap rule performs similarly and slightly better on the 3.5% of observations where the labels disagree. The limit supplies a threshold with a power interpretation, rather than a better ranking of reversal risk. A concurrent method based on agreement with a second measurement (Yauney et al., 2026) estimates its threshold from that agreement, treating the full benchmark as ground truth. Here the threshold is specified before the second measurement is used to evaluate it.

**Within-benchmark ordering and clustered uncertainty.** HELM reports MMLU results on 514 items, whereas the full benchmark contains 14,042; 59 models have results under both evaluations. We compare micro-accuracy on the 514 items with micro-accuracy on the full benchmark. HELM's published five-subject macro-average differs in ordering from its micro-average on one of 4,066 model pairs. The within-benchmark analysis contains 1,695 orderings: the 1,696 used in the disjoint-subject analysis, excluding one additional pair tied on the full benchmark. Of these, 127 (7.5%) reverse on the full benchmark. Among the 1,285 pairs significant on HELM's items, 0.70% reverse; among the 410 non-significant pairs, 28.8% reverse. Power-based labels give 0.18% and 21.0% across 1,100 above-limit and 595 below-limit pairs, respectively. The reversal-rate ratio is then 115.6, compared with 41.1 for the significance labels. The power-based threshold is stricter and gives a larger separation in these data, but the gap-rule control shows that it does not improve the ordering of reversal risk.

The 1,695 pairs are formed from 59 models, so an individual model contributes to as many as 58 comparisons. We therefore bootstrap models rather than pairs. The dyadic bootstrap retains each sampled model's multiplicity when forming pairs. A `set`-based alternative discards multiplicities and acts as a random ∼63% subsample rather than an $n$-out-of-$n$ bootstrap, yielding narrower intervals. Unless otherwise specified, reported intervals are 95% percentile intervals from 2,000 model bootstrap resamples. For significance-based labels in the within-benchmark analysis, the dyadic intervals are 0.00–1.69% and 21.7–36.6%. For the disjoint-subject analysis shown in Fig. 3, they are 0.00–1.06% and 15.72–27.75% under power-based labels, and 0.00–2.11% and 22.02–37.33% under significance labels. The corresponding `set`-based intervals for the latter are 0.17–1.72% and 23.60–34.71%. All methods retain a clear separation between the two categories, although their uncertainty estimates differ.

**Dependence, multiplicity and record chronology.** Three qualifications apply to successive SWE-bench record comparisons. First, successive records recur as reference systems, so the 32 comparisons are dependent and a Wilson interval based on independent claims is inappropriate. A bootstrap over the 134 systems gives a non-rejection share of 75.0%, with a 95% interval of 61.5–81.5%. Because this statistic uses a running maximum and bootstrap resamples duplicate systems, the percentile interval is an indication of variability rather than a calibrated coverage guarantee. Changing the classification of one comparison changes the estimate by approximately three percentage points. Second, 24 is the uncorrected count. Five of the eight significant comparisons have $P = 0.024$ to 0.045; under Benjamini–Hochberg correction at $q = 0.05$, only three remain significant. The non-rejection count becomes 29 of 32, and the last significant record comparison moves from April 2025 to 2024–06–12. The date is therefore particularly sensitive to the threshold, shifting by ten months under multiplicity correction. Selection of record increases and multiplicity both favour treating the uncorrected non-rejection share as an optimistic account of resolution. Third, the construction treats every new maximum as a successive advance. Three records share the date 2024–04–02 and are ordered alphabetically within that date; they represent a single baseline deposit rather than an observed sequence of publications. The comparison setting the uncorrected date, 30 April 2025, is also one of five cases where the closed-form power label differs from the exact significance test: its 4.2-point gain lies below the 5.12-accuracy-point limit.

**Cumulative improvement beyond successive increments.** Failure to distinguish successive records does not imply an absence of cumulative improvement. The final record exceeds the record at the last significant comparison by 9.2 accuracy points ($P = 1.5e-07$). Six of the 8 comparisons between earlier records and the final record are significant, although none of the eight individual intervening increments is. The benchmark therefore supports improvement over the full interval without identifying a statistically distinguishable contribution at each step.

**Null distributions for unresolved record increments.** The fraction of non-significant record increments depends on the reference process. A running maximum over a growing pool tends to have smaller increments as the record sequence lengthens, even without temporal progress. A length-matched null, formed by drawing and sorting 33 of the 134 systems, gives a non-rejection share of 83.6%, compared with the observed 75.0%, but its median gap is smaller (1.40 versus 1.70 accuracy points). A process-matched null based on permuted arrivals produces about five records and a larger median gap of 3.20 accuracy points. To control both sequence length and gap scale, we used rejection sampling to retain draws with a median increment within 15% of the observed value. This spacing-matched null gives 83.1% non-rejection (95% range 71.9–93.8), compared with

75.0% in the observed sequence. Thus, the observed sequence has fewer unresolved increments than a random sequence of comparable size and spacing. Together with 33 observed records versus 4.9 under random arrival, this supports aggregate progress, while exact tests still leave most individual increments unresolved. The spacing-matched one-sided test gives $P = 0.030$ when ties count against the null comparison and $P = 0.105$ when ties count in its favour. Both conventions are reported because the statistic lies on the discrete lattice $k/32$. The length-matched null gives $P = 0.0175$; its test cannot be combined with the spacing-matched null's mean, because they describe different reference distributions. These comparisons qualify the interpretation of the 24 unresolved increases: the record sequence's length, rather than its unresolved fraction, is the evidence for progress.

**Scope of the reversal evidence.** The reversal analyses distinguish changes in subject coverage from sampling noise. HELM's 514 MMLU items comprise five complete subjects from the 14,042-item benchmark. The within-benchmark and disjoint-subject analyses therefore assess sensitivity to subject coverage, whereas random split halves assess item-sampling variability. All three analyses use per-item outcomes; they establish reversal separation on boards with such outcomes and do not directly validate the five surveyed families without them. The limit provides a common effect-size threshold from $n$ and measured $d$, allowing comparisons across benchmark sizes. The qualitative result also replicates on Open LLM v1's MMLU: with a different harness, 111 models and subject panels of ∼514 items labelled at the board's own rate, below-limit and above-limit pairs reverse at 20.6% and 0.10%, respectively. The model-clustered intervals are disjoint, the separation holds in 300 of 300 panels, and it persists at every evaluated size up to 4,000 (Appendix Table 23). These are pair-level rates under each pair's majority label, matching the estimand of the clustered intervals. Pooling observations instead gives 19.8% and 0.64%. Although the separation replicates, the reversal rate itself is not determined by $(n, d)$; it also depends on the distribution of gaps on each board.

**Sensitivity to the sampled systems and organisations.** The transfer results depend on the systems included. The verified populations contain 111 systems on four Open LLM boards and 21 on GSM8K, representing 100 and 16 uploading organisations, respectively. We assessed this dependence with an organisation-cluster bootstrap, refitting every map within each resample, and with 200 draws retaining one system per organisation (Appendix Table 30). The pre-specified rule confirms a statement if it holds in at least 95% of resamples under both schemes. Three statements satisfy this rule. In every resample, at least one transfer has a retained ratio of 0.25 or less, so high retained fit is not consistent across all transfers. The full map's advantage over the gap-only rule is larger for multiple-choice than free-response targets in every resample, as is the reported interpretation of Card et al.'s published relationship. Two broader statements are unsupported at the point estimates and under resampling. A claim that most transfers retain at least 0.75 requires a strict majority (16 of 30 in the original rule, or 22 of 42 here); only 19 do so, with an interval of 12–23, and the criterion holds in 7.9% of resamples. A claim that every transfer preserves correlation with the target relationship (minimum $r$ at least 0.85) holds in 3.5%. The median $r$ is 0.951 (0.938–0.959), but the minimum is 0.473 (0.206–0.859). All four transfers with $r$ below 0.85 originate from Winogrande. The median retained ratio is 0.704 (0.563–0.773). We therefore report the counts and intervals, and restrict the conclusions to the three statements supported under both resampling schemes.

### A.10 Limitations and implications for benchmark design

This section separates the assumptions underlying the resolution estimates from the empirical implications for benchmark size and reporting.

**Sample-size requirements.** At $d = 0.17$, the algebraic sizing expression requires 3,336 items for a two-accuracy-point gap and 13,344 for a one-accuracy-point gap. Exact test power gives requirements of 3,424 and 13,534, respectively, increases of 2.6% and 1.4%. We report both to distinguish the convenient approximation from the exact requirement. The selected rate approximates the observed median disagreement between successive SWE-bench record holders (0.172); it is not a universal default or a lower bound. Record holders need not be the most similar model pairs. On HellaSwag and MMLU, their rates exceed the full-population rates (0.1213 versus 0.1314 and 0.3156 versus 0.3518, respectively, with the full-population rate listed first). The median record-holder rate is 0.273. Since 0.17 is below the survey reference rate, the resulting requirements are correspondingly smaller.

Across the 88 measured board-specific rates, 0.0910 to 0.4889, the two-accuracy-point requirement ranges from 1,786 to 9,594 items.

Adequacy depends on which boards are counted. Among the twelve largest, 7 of 12 meet the two-accuracy-point target. Across all 202 distinct boards, 12 meet the algebraic requirement of 3,336 items at $d = 0.17$, whereas 11 meet the exact requirement of 3,424. The corresponding counts are 14 at 0.10 and 9 at 0.2549. Five boards meet the one-accuracy-point target. The median board contains 303 items, approximately an eleventh of the two-accuracy-point requirement, and 127 of 202 contain fewer than 500. On 34 of the 202 boards (17%), including SWE-bench Verified and GAIA, the top two observed scores are exactly tied. Substituting a zero observed gap into the sizing calculation yields no finite requirement; this does not establish that the population performances are equal. Across all 202 boards, the median requirement is 49 times the current item count (26–141 across the rate range); across the 168 boards with positive top-two gaps, it is 33 times (18–96). Eleven of the 202 boards meet their leader-comparison requirement.

**Reversals under random item splitting.** To isolate item-sampling variability, we randomly split HELM's 514 MMLU items into halves of 257, repeated 200 times. Each pair's label uses only the first half, and its ordering is evaluated on the second. The gap reverses sign in 0.25% of 453,302 above-limit observations (0.18–0.36%) and in 18.2% of 348,462 below-limit observations (16.4–20.4%). Each observation is one model pair in one split, so pairs recur across splits. The intervals therefore use the dyadic model bootstrap rather than resampling individual observations. Both halves cover the same benchmark distribution; only the sampled items differ.

**Multiplicity-adjusted survey limits.** The primary survey uses per-comparison $\alpha = 0.05$, while a full ranking implies many pairwise comparisons: a median of 2,346 per board. For a board of $M$ systems, using $\alpha = 0.05/\binom{M}{2}$ across the 182 distinct HELM boards increases the median limit from 8.25 to 13.67 accuracy points. The median tier count decreases from 5 to 4, and the mean from 5.91 to 4.06. Boards with no supported distinction increase from 0 to 4, and those with at most two tiers from 8 to 20. On 24 of the 182 boards, no feasible effect size attains 80% power at the corrected level. A stricter significance threshold necessarily raises the limit or leaves it unchanged and cannot increase the tier count. The comparison therefore quantifies the optimism of per-comparison resolution summaries for a simultaneous ranking claim. We use Bonferroni because the survey tables lack the pairwise $p$-values required by a step-down procedure. Holm correction would be no stricter, so these Bonferroni limits and counts provide, respectively, upper and lower bounds relative to that correction.

**Model-population growth and the two Open LLM versions.** Adding systems to a fixed test set lengthens the ranking without increasing the number of test items available to distinguish them. Additional evaluations require inference, whereas enlarging a benchmark also requires new suitable items and often annotation. The two Open LLM Leaderboard versions have the largest model populations and the smallest supported fractions in this survey, 0.4% and 0.3%. The earlier version supports the smaller fraction despite its larger suite—five benchmarks and 27,842 items, compared with four and 14,653—because it ranks more models. This comparison illustrates how ranking density can offset improvements in absolute test-set resolution.

**Assumptions and limits of applicability.** Exact testing requires per-item outputs. We test three leaderboard families exhaustively—HELM, SWE-bench Verified and Open LLM v1—and use assumed rates for the other five. BigCodeBench, GAIA and Open LLM v2 use 0.2549; vision and MTEB use 0.05, the lowest rate considered. Direct validation would require per-item outcomes that were not obtained for these families. The coverage of detailed case studies is also uneven: SWE-bench, vision, MTEB and Open LLM v1 receive separate analyses, whereas GAIA and BigCodeBench appear in the survey tables without equivalent case studies.

A separate lower-limit calculation does not require an assumed disagreement rate. For non-negative discordant rates, $b - c = \Delta$ and $b + c = d$ with $b, c \geq 0$ imply $d \geq |\Delta|$. The most favourable boundary is therefore $d = |\Delta|$, giving $\mathrm{MDD}(n, |\Delta|)$. A gap below this value fails to attain the target power at that boundary by theorem; exhaustive numerical evaluation also verifies failure at every admissible $d$ for the 119 surveyed board sizes (Open question 26). Among the 59 boards that are not sub-scenario slices, the median fraction of pairs below this floor is 5.13%. Across all 202 boards it is 20.99% (range

0.16–95.88%). The composition-specific medians are 25.12% across the 182 HELM boards, 9.05% across HELM's 39 scenario-level boards, and 2.95% across the 20 non-HELM boards. Non-HELM family medians range from 0.28% for vision to 9.34% for GAIA (Appendix Table 19).

The floor has a count interpretation: $n\,\Delta_{\mathrm{fl}}(n) \to 7.906$ (Corollary 21). Approximately eight additional correct items are therefore needed under the most favourable disagreement structure. The floor scales as $1/n$, whereas the usual fixed-rate limit scales as $1/\sqrt{n}$. Eight items correspond to 3.33 accuracy points at HELM's median size of 236 and 0.48 at the non-HELM median of 1,655. The proportion of pairs below the floor also depends on the gap distribution: GAIA's 165 items yield 9.34%, below the HELM median despite its smaller size. Pooling pairs across boards gives 3.27%, dominated by the two Open LLM leaderboards, which contain 96.8% of all pairs. We report board-level medians to match the other survey summaries. The count identity $b - c = n\Delta$ applies only to an unweighted mean of binary item outcomes; restricting the analysis to the 181 boards with that metric gives a median of 24.88%. HELM's three-decimal reporting introduces additional uncertainty: the all-board median lies between 19.91 and 21.73%, and the HELM median between 23.62 and 25.76%.

For every surveyed board, $\mathrm{MDD}(n, 0.2549) \geq \Delta_{\mathrm{fl}}(n)$. Consequently, all pairs below the floor are also below the reference-rate limit, and the corresponding median fraction is no larger than the reference-rate value of 53.51%. The floor can still be substantial on small boards, reaching 95.88% on a 65-item HELM table. Removing the assumed rate does not remove the other assumptions: item independence, unconditional treatment of the discordant count, 80% power, and substitution of the observed gap for the unknown true gap. The last step is an observed-power construction, whose relationship to significance testing is well established (Hoenig & Heisey, 2001; Levine & Ensom, 2001). Consistent with this relationship, a $\sqrt{n}$-scaled gap rule orders reversal risk as well as the MDD rule. The power-based tier threshold corresponds to exact-test per-comparison $\alpha$ values of 0.0038–0.0047 across boards with 165 to 50,000 items. Its contribution is an effect-size interpretation for comparing benchmarks, not additional statistical power beyond the underlying test. Computing the floor by interpolation on a $d$ grid instead of solving for it overestimates the threshold when interpolation clamps below the first feasible knot. Across all 202 boards, this occurs on every board, with a median factor of 1.06× and a maximum of 3.64× at $n = 50{,}000$. We instead solve for the threshold using the exact power calculation, including the requirement that the design can produce the six disagreements needed for rejection.

An assumed rate can also be incompatible with the reported accuracies: $d$ cannot exceed $\min(a_i + a_j,\, 2 - a_i - a_j)$. The reference rate 0.2549 exceeds this bound on 45.1% of adjacent IMDB pairs, 32.4% of ImageNet-A pairs and 24.3% of Banking77 pairs. The rate 0.05 is feasible for every adjacent pair on all seven boards examined. An infeasibly high rate overstates the limit and undercounts tiers. Capping each pair at its feasible maximum increases the count on 43 of the 202 boards, by a median of 1 tier and a maximum of 5 (`src/mx43_rate_feasibility.py`). Using the reference rate of 0.2549 instead of 0.05 also increases the apparent resolution shortfall in vision and MTEB: vision's median adjacent-below-limit share rises from 98.9% to 99.5%, while its median tier count falls from 96.5 to 49; the MTEB values change from 94.2% to 98.5% and from 36 to 20. We report the more favourable 0.05 case for these families. Two independently measured family rates agree with the reference 0.2549 to within 0.0131, but another domain varies by 2.68×, so this agreement does not justify a universal rate.

Tier interpretation depends on the partition rule. The greedy partition minimises the number of blocks subject to no distinguishable within-block pair (Theorem 30). Single linkage never yields more blocks and yields fewer on all but three boards at the reference rate, because it allows within-block pairs connected through intermediate systems to exceed the threshold. The analysis also assumes a paired test on a shared test set; it describes the distinctions available under such testing even when published rankings do not report a test. Sample-size targets concern item-resampling uncertainty and do not account for run-to-run or seed-level variability. The appendix's cross-run analysis shows that the item-based power calculation does not transfer directly to that source of variation.

We have not established that unresolved rankings change model adoption. A pre-registered analysis considered likes, thirty-day downloads, derivative counts, and derivative counts after excluding merge-like entries. Each proxy was to be evaluated on unresolved pairs only if it passed a control on resolved pairs. Likes and downloads failed that control. Among resolved pairs, the higher-scoring model had more derivatives in 38.8% of 48,189 comparisons and in 41.5% of 50,741 after excluding merges. These pair sets are not nested: each proxy samples at most 260 entries per board and excludes

ties in derivative count. Both derivative-count controls failed, with the same direction of association on every eligible board. Under the pre-specified rule, the analysis on unresolved pairs was therefore not interpreted. The inverse association between score and derivative count is consistent with users building on base models rather than leaderboard winners, but it does not establish that explanation. Downstream adoption remains a motivation rather than a demonstrated result. Finally, the statistical analysis concerns binary per-item outcomes. It does not establish analogous limits for generation quality or human-preference metrics, and tables with unverifiable metrics or sizes were excluded.

**Implications for benchmark reporting.** Three reporting changes follow from the analyses without requiring new test items. First, report a detection limit or power curve with the benchmark, including the assumed or measured disagreement rate. Second, report grouped resolution summaries alongside ranks and specify the grouping rule. WMT has published significance clusters since Bojar et al. (2013), and Chatbot Arena groups models with overlapping uncertainty intervals (Chiang et al., 2024). For paired binary accuracy, the corresponding power threshold can be stated from $n$ and $d$. Greedy tiers should not be interpreted as groups separated on every pairwise comparison. Across the survey, the median board has 2 single-linkage blocks and 5 greedy tiers; the median ratio of greedy to single-linkage counts is 3.0, and 43% of boards form a single block under single linkage. These counts describe complementary partitions, not a confidence interval. Both can be reported, with $k^{\mathrm{SL}} = M - a(M-1)$ obtained directly from the adjacent-below-limit share (Appendix Table 10). Third, identify the metric of every score column. The archived score tables did not always provide enough metric information to verify comparability with binary accuracy or exact match. Distinguishing accuracy from metrics such as F1 and BLEU is necessary for both readers and automated audits. Increasing the number of evaluated systems alone does not improve the item-based resolution of an existing benchmark.

**Dependence on estimator, metric and attainable effect size.** The controlled experiments assess how the estimated limit depends on the reference estimator and metric. Across four estimators, the limit changes little under the evaluated perturbations. Across metrics, however, both its magnitude and ordering change: AUC is the most sensitive metric on one benchmark and the least sensitive on another, where the reference AUC of 0.995 leaves little room for improvement (Appendix Table 9). This evidence is restricted to four small datasets and the specified perturbations; it does not make MDD independent of model-pair disagreement. On HELM's NarrativeQA, with 91 models and 355 items, three published metrics yield limits from 6.06 to 6.78 accuracy points. Binarising its continuous $F_1$ at thresholds from 0.3 to 0.7 changes the limit from 5.89 to 7.39 accuracy points, a 25% range (Appendix Table 21).

The target power can also be unattainable over the available effect-size range. On the balanced 124-participant clinical cohort, training-label corruption reduced the competing model to approximately chance performance, creating a 14.72-accuracy-point gap against an attainable range of 15.0 accuracy points set by the reference accuracy. Power did not reach 80% anywhere along this perturbation sequence. This establishes insufficient power for the evaluated reference, competitor sequence and protocol; it does not establish that every possible model pair on the cohort is indistinguishable.

**Sensitivity of sizing and record comparisons to disagreement.** The broad separation of reversal rates changes little under nearby rate assumptions. Pooling observations rather than using the pair-level summary in Appendix Table 23, Open LLM v1 MMLU gives 19.8% versus 0.6% at its own rate of $d = 0.316$, and 21.3% versus 0.84% at HELM's 0.255. The qualitative separation persists across a 2.8× change in the limit induced by panel size. Because reversal probability varies smoothly with the gap relative to the limit, nearby thresholds select similar mixtures of pairs. This threshold insensitivity does not imply that disagreement is irrelevant to sizing. A two-accuracy-point gap requires 3,424 items at $d = 0.17$ and 6,962 at $d = 0.35$. Excluding HELM does not eliminate the rate variation: the range across the other exact-test boards is 0.121–0.352, 2.90×, compared with the reported 2.68× record-holder range.

The analysis of 66 published record increases separates the effects of the power criterion, rate aggregation and rate transfer (Appendix Table 24). With pair-specific rates and matched power, the formula reproduces the exact classification on all 65 evaluable cases. This identity defines the baseline rather than independently validating the method. Requiring 80% power changes 6 classifications; using board-level rates changes 10. Substituting HELM's rate for every board changes no additional

classifications in aggregate: it gives the same number of errors as the board-specific medians. The individual classifications nevertheless differ. Across the measured rate range, 54 claims are classified correctly throughout, 7 incorrectly throughout, and 5 depend on the rate; borrowing excludes 2 of those 5 rate-sensitive claims. No rate within this range produces a false-resolvable classification, whereas 0.05 produces five. The formula-based calculation uses published score differences, while exact testing uses per-item outcomes. Repeating the analysis with the matched per-item gain $(b - c)/n$ reproduces every cell of Appendix Table 24 and the 54/7/5 partition. The one claim with a genuine discrepancy between these quantities has zero discordance on 10,042 common items despite a published gain of 0.060 accuracy points; neither calculation yields an exact-test rejection.

The unchanged aggregate error count is specific to the evaluated rate range $[0.121, 0.352]$. The 0.05 assumption used for vision and MTEB lies outside it. Rate substitution can change the direction of individual errors even when their total count is constant, and the direction depends on whether the assumed rate is above or below the pair's actual rate. When disagreement is unknown, a larger plausible rate gives the more conservative requirement. The 3,424-item headline target uses $d = 0.17$, motivated by successive record comparisons; at the largest measured rate, the same gap requires more than twice as many items. Planning should therefore use a rate relevant to the intended model pair, or a conservative range when no such measurement is available.

### A.11 Related methods and additional benchmark comparisons

The following comparisons relate the analysis to prior statistical methods and examine benchmark populations that differ in evaluation size or model count.

**Relation to significance clusters and rank intervals.** Grouped rankings are established in statistical model comparison. Demšar (2006) introduced critical-difference diagrams to connect classifiers not separated by a Nemenyi test on mean ranks across datasets. WMT has clustered systems by significance since Bojar et al. (2013), using a stricter rule than the greedy partition considered here, and continues to do so (Kocmi et al., 2024). Chatbot Arena groups models with overlapping intervals (Chiang et al., 2024). Rank intervals (Neuhof & Benjamini, 2026b) bound each model's rank using its Holm-corrected wins and losses on per-observation scores, then aggregate across tasks conformally. These methods summarise the evidence in observed comparisons. Non-rejection alone cannot distinguish equal population performance from insufficient test-set resolution. MDD complements these summaries by quantifying the true gap needed to attain a specified power. Under the normal approximation, the 80% power boundary is 1.43 times the significance boundary (Proposition 37), so power-based tiers are no finer than corresponding significance-based cliques. MDD is expressed in accuracy units for one benchmark, whereas a critical difference is expressed in mean ranks across datasets. The latter depends on the rank design's known null variance; the former additionally requires disagreement through the variance $d/n$. This nuisance parameter motivates the transfer analyses. Neither the below-threshold relation nor non-significance is transitive, so greedy and single-linkage partitions provide different grouping summaries. For multiplicity control, Otero et al. (2025) recommend Benjamini–Hochberg rather than the family-wise correction used here; we report both. Colombo et al. (2022) instead change the aggregation rule while retaining a total order, addressing a different aspect of ranking.

**Prior power analysis in NLP evaluation.** Card et al. (2020) define the minimum detectable effect for McNemar's test at $\alpha = 0.05$ and 80% power using test-set size and pairwise agreement, with our $d$ corresponding to their $1 - P_a$. They compare the resulting limits with published gaps on nine benchmarks; we use the same power concept. Drawing on that analysis, Bowman & Dahl (2021) describe requirements of a few thousand examples for a one-point effect and two orders of magnitude more near the frontier, consistent with the sizing analysis in Appendix A.10. Miller (2024) derive the same inversion for evaluation design, while Dror et al. (2018) discuss McNemar testing for paired binary outcomes. Dror et al. (2017); Reimers & Gurevych (2018); Dror et al. (2019) also explain why individual scores are insufficient for model comparison. The latter two address training-seed variability, which the item-based limit here does not model. Our extension is empirical: the survey covers 202 distinct boards rather than nine datasets and measures disagreement from per-item outcomes where those outcomes are available.

**Code and agent benchmarks.** BigCodeBench and GAIA also have coarse resolution under the assumed-rate calculation (Appendix Table 12). GAIA contains 165 items, the smallest evaluation set outside HELM's sub-scenarios, and has a detection limit of 11.31 accuracy points. At that assumed rate, gaps smaller than this value among its 92 systems fall below the target-power threshold.

**Prior sample-size and collection-reliability methods.** McNemar's test (McNemar, 1947) and its power calculation have a long-established literature. The power function has both exact and asymptotic treatments (Bennett & Underwood, 1970; Schork & Williams, 1980; Duffy, 1984), as do sample-size requirements (Miettinen, 1968; Lachin, 1992; Royston, 1993; Lachenbruch, 1992). Our algebraic sizing expression is the approximation of Connor (Connor, 1987) with the alternative-variance term omitted. The six-discordance rejection floor is the sign-test minimum described by Dixon & Mood (1946). In the relevant large-sample regime, the algebraic expression reproduces their Table 2 with a median printed-value ratio of 1.009 and a maximum of 1.026; these cells also use a normal approximation, so the agreement is expected rather than an independent validation.

Information-retrieval research has likewise used power to design evaluation collections. Webber et al. (2008) tabulate the smallest detectable average-precision difference at 80% power. Sakai (2016b) use the term *minimum detectable difference* with the same convention and invert it to a topic requirement, developed further by Sakai (2018a). Carterette & Smucker (2007) compute exact sign-test power after removing ties and invert it to a trial count. For paired binary outcomes, this is McNemar's conditional test, with an expected $nd$ discordant trials; the unconditional calculation averages over the discordant count. Buckley & Voorhees (2000) examine reversal frequency as a joint function of threshold and sample size. The broader collection-reliability literature addresses related questions (Voorhees & Buckley, 2002; Sanderson & Zobel, 2005; Urbano, 2016; Urbano et al., 2019; Sakai, 2018b; 2016a; Craswell et al., 2021). A useful feature of the binary setting is that the disagreement rate is observable and bounded below by the absolute gap, enabling the rate-free floor. The nuisance variance for continuous retrieval metrics has no corresponding bound in this analysis. Finally, Fagerland et al. (2013) document conservativeness of the exact test; we assess its practical effect through the reported mid-$p$ comparison.

**Evaluation on all 57 MMLU subjects.** To assess the effect of HELM's restricted subject coverage, we repeated the resolution calculation on the complete 57-subject MMLU. HELM uses 514 of 14,042 items, comprising all items from five of the fifty-seven subjects. At the same disagreement rate measured from HELM's per-item outcomes, increasing the item count twenty-seven-fold reduces the limit by a factor of 5.3, from 6.40 to 1.20 accuracy points, consistent with $\sqrt{27.3}$ being 5.2. The tier count increases by a factor of 2.9, from 8 to 23. These tier counts use different model populations—91 HELM models and 65 models evaluated on every subject— and therefore do not isolate item count alone. The limits, which depend only on $n$ and $d$, do provide that controlled comparison. At the full size, 84.4% of adjacent positions remain below the limit (Appendix Table 2). The $\sqrt{n}$ dependence means that reducing a detectable gap fourfold requires sixteen times as many items, explaining why substantial ranking uncertainty can persist even with the larger evaluation set.

**Absolute resolution and the supported ranking fraction.** The family with the highest absolute resolution supports 6.2% of its ordering and leaves 98.9% of adjacent positions below the limit, despite an evaluation set approximately two orders of magnitude larger than the language boards. A larger tier count therefore need not imply a larger supported fraction. All eight family shares use the closed-form limit at 80% power. SWE-bench has a supported fraction of 9.0% at the reference rate (12 tiers for 134 systems). Its significance-based exact test gives 7.5% under Holm correction (ten tiers) and 14.2% without correction (19 tiers). The uncorrected rejection boundary lies approximately 1.43 times below the power boundary, so these counts use different criteria. Vision has the highest absolute tier count, but five families have a larger supported fraction (Fig. 7d).

### A.12 Item-response scores and score-specific resolution

We assess whether the accuracy-based results extend to rankings based on item-response scores.

**Ranking with item-response scores.** Accuracy-based resolution need not describe every scoring rule. Rodriguez et al. (2021) report more stable rankings across disjoint samples using item-response ability than using accuracy. We therefore repeated random item splitting on five boards, retaining

the accuracy-based power labels for each pair while ranking each half by accuracy, Rasch ability or two-parameter logistic (2PL) ability fitted to that half alone (`src/mx34_irt_followup.py`; 30 splits per board; five pre-specified interpretation rules; Appendix Table 28 and Appendix Table 29). The Rasch analysis is a control: the number-correct score is sufficient for Rasch ability, so it should reproduce accuracy ordering. It does so exactly on every board. With 2PL ordering, accuracy-labelled below-limit pairs reverse less often, with a rate ratio $\rho$ ranging from 0.33 on Open LLM v1's MMLU to 0.96 on GSM8K. The pre-specified rule classifies this evidence as *ambiguous*; the classification depends on the discrimination prior, with a tighter prior giving the category *carries*. The improvement is consistent with item-level discrimination rather than subject weighting: macro-averaging by subject gives $\rho$ of 1.00 and 1.14 on the two MMLU boards, while allowing one discrimination parameter per subject gives 1.05 and 0.91. Estimating item weights on half the systems and applying them unchanged to the other half changes $\rho$ by at most 0.07, indicating limited sensitivity to which systems are used to estimate those weights.

**Score-specific detection limits.** The preceding analysis labels 2PL orderings with an accuracy-based limit. We also evaluated each score against its own limit, defined as 2.80 paired standard errors of its gap, estimated by an item bootstrap that refits the model in every resample. Under this comparison, the pre-specified interpretation is *generalises*. For 2PL, pairs below the score-specific limit reverse in 17.5–20.0% of cases, compared with 0.04–0.21% above it; 96.9–98.6% of all 2PL reversals occur below the limit. For accuracy under the same construction, which agrees with the exact limit on 95.8–99.2% of pairs, the corresponding share is 97.3–98.1%. Thus, 2PL can reduce the detectable gap and the number of below-limit comparisons while retaining the relationship between score-specific resolution and reversal. The below-limit count is 0.57 times the accuracy count on Open LLM v1's MMLU and 0.83–0.95 times elsewhere; the interval excludes one on every board except GSM8K. The remaining below-limit pairs reverse at rates comparable to the accuracy-based analysis.

The same subject-transfer analysis cannot be performed for HELM's 59 matched models, for which the required per-item outcomes are available only on the 514-item panel. On Open LLM v1's MMLU, labelling a subject panel and evaluating on the remaining subjects gives $\rho$ of 0.68 (0.56–0.83) across 30 random panels and 0.76 on HELM's five subjects. Under the pre-specified rule, this is the category originally named "gains the word *accuracy*": the claim should be restricted to accuracy orderings rather than all possible scoring rules. Using eight splits on the two large boards gives the same category but does not separate the alternative interpretations.

# B Additional empirical results

## B.1 Fixed thresholds and prediction of ranking stability

We compared benchmark-specific detection limits with fixed accuracy-gap thresholds and evaluated whether the limits predict ranking stability. Across the 10 benchmarks in Appendix Table 3, disagreement rates from 0.05 to 0.40 yield limits from 0.54 to 16.28 accuracy points. This factor of 30 largely reflects test-set sizes that differ by two orders of magnitude. A fixed threshold of one accuracy point is lower than the benchmark-specific limit on 9 of the 10 benchmarks. A threshold of two accuracy points is too low on seven and too high on one. Even a threshold of five accuracy points misclassifies 3 of 10 benchmarks, with errors in both directions. The closed-form approximation instead assigns a threshold conditional on each benchmark's size and assumed disagreement rate.

The detection limit does not classify stable orderings more accurately than a $\sqrt{n}$-scaled gap rule (Methods). Its additional role is calibration: it specifies a threshold in accuracy points from $n$ and $d$ for an individual benchmark. In contrast, a gap-ranking rule requires a separate choice of cutoff, typically calibrated across a collection of comparisons.

We also evaluated 32 successive record claims on SWE-bench Verified, distinguishing the ranking of claims from the calibration of a decision threshold. Among claims made after each of three cutoff dates, the accuracy gap perfectly ranks the exact-test outcomes: the area under the curve is 1.000 at every cutoff. This is an internal-consistency result rather than an out-of-sample forecast. The predictor is the observed gap $100(b - c)/n$ and the label is the exact binomial test applied to the same $(b, c)$ table; neither is estimated using pre-cutoff data. Perfect ranking is not an algebraic identity. The exact test is monotone in the gap only when the number of discordant items is fixed; with different

discordant counts, a smaller gap can reject when a larger gap does not (Proposition 41). In the 25 claims after the first cutoff, the three significant advances happen to have the largest increments.

The power threshold and the significance boundary answer different questions. The limit at 80% power is 5.28 accuracy points, whereas the observed $\alpha = 0.05$ boundary is 4.0 accuracy points, a difference of 1.28 accuracy points. A true gap below an 80%-power threshold can still produce a significant comparison, although less reliably. When used to predict single-comparison significance, the power threshold labels every post-cutoff claim below the limit and has a balanced accuracy of 50%. The perfect ranking therefore does not imply accurate threshold calibration, and the positive class contains only 3 of 25 claims.

### B.2 Resolution of language-model leaderboards

We surveyed eight leaderboard families comprising 267 published tables and 202 distinct boards after counting 65 duplicated HELM boards once (Appendix Fig. 11, Appendix Table 1). Six families are discussed in five groups, with the two Open LLM Leaderboard generations considered together; BigCodeBench and GAIA are reported in Appendix Table 12. HELM contributes 182 of the 202 boards, corresponding to 247 tables with per-item binary outcomes across six evaluation suites (Appendix Table 10). Among these, 143 are sub-scenario slices, often accompanied by their parent scenario. We report survey summaries both with and without these slices (Appendix Table 19). The eligible population depends on the analysis: sizing requires published score vectors, whereas exact tests require per-item outcomes. Appendix Table 15 specifies these populations.

The median HELM board ranks 69 models on 236 test instances. It contains five tiers under the greedy construction and two under single linkage (Methods; Fig. 5). The greedy count is the more permissive summary. At the median, 98.7% of adjacent score gaps fall below the detection limit (Appendix Table 1). Thus, the total order reported by a typical board is substantially finer than either threshold-based partition. These partitions describe power at the assumed disagreement rate; they do not establish equality between models within a tier.

The result persists when sub-scenario slices are excluded. The 39 scenario-level boards have a median of five tiers, with 97.1% of adjacent gaps below the limit (Appendix A.8). Increasing the item count also leaves many adjacent gaps below the limit. The complete 57-subject MMLU has 27 times as many items as the HELM sample, reducing the limit by a factor of 5.3 but increasing the tier count by only a factor of 2.9. Even on the full benchmark, 84.4% of adjacent gaps remain below the limit (Appendix Table 2; Appendix A.11).

### B.3 Resolution as the model population increases

The Open LLM Leaderboard uses a different evaluation framework and contains a model population two orders of magnitude larger than HELM's. We analyzed 4,476 models on four benchmarks whose evaluated sizes were verified from the source datasets (Fig. 6, Appendix Table 6). MuSR yields six tiers, although its score is a macro-average over three subsets and its limit is therefore approximate. The same pattern holds for MATH Lvl 5 and MMLU-PRO, whose scores are item means: they yield 17 and 42 tiers among the same 4,476 models. On all four benchmarks, at least 99.9% of adjacent gaps fall below the limit; on two, every adjacent gap does so.

Open LLM v1 contains 6,960 models and shares no benchmarks with v2. At least 99.9% of adjacent gaps fall below the limit on each of its five eligible benchmarks. Its median tier count divided by its median model count is 0.3%, the lowest family-level value in the survey, although individual boards have lower values (Appendix Table 11, Appendix Fig. 2). Both generations therefore contain many more ranked positions than threshold-separated tiers.

For a fixed comparison, power depends on the number of test items and the systems' disagreement rate, rather than the number of models on the board. Adding models within an existing score range can increase the density of scores without improving pairwise resolution. In these two families, the ratio of median tiers to median model count is 0.4% for v2 and 0.3% for v1, the two lowest family-level values. These observations concern the surveyed populations; adding a model outside the existing score range could increase the tier count.

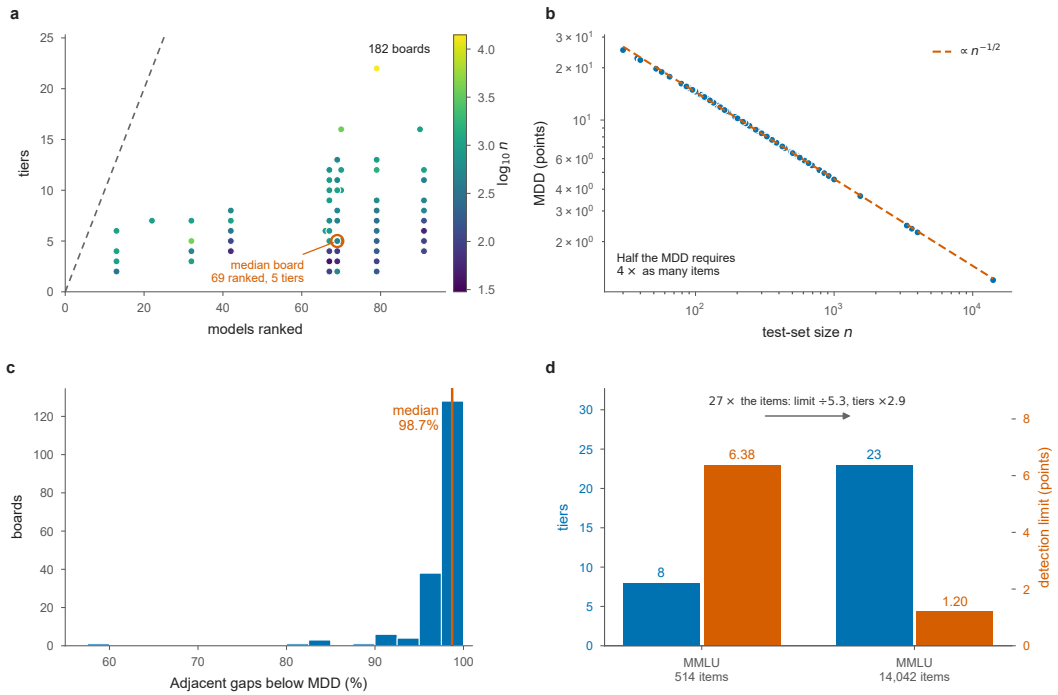



Figure 5: **Resolution of 182 distinct HELM leaderboards.** **a**, Tier counts versus model counts, colored by test-set size. The dashed diagonal denotes one tier per model. The circled median board contains 69 models, five greedy tiers, and two single-linkage groups. **b**, Detection limit versus test-set size on logarithmic axes, with an $n^{-1/2}$ reference. The disagreement rate is fixed, so the curve shows the mathematical size dependence rather than an empirical fit: halving the limit requires four times as many items. **c**, Distribution of the fraction of adjacent gaps below the detection limit across 182 boards; the median is 98.7%. **d**, HELM's 514-item MMLU sample versus the full 14,042-item benchmark, with tier counts on the left axis and detection limits on the right. Both use the closed-form approximation at the disagreement rate measured from HELM's per-item outcomes. Increasing the number of items 27-fold reduces the limit by a factor of 5.3 and increases the tier count by a factor of 2.9. An exact-test analysis of the sample, which does not assume a common disagreement rate, gives six tiers.

### B.4 Resolution of image-classification benchmarks

The vision family comprises 1,556 image classifiers evaluated on ImageNet and related distribution-shift benchmarks (Fig. 7). At the reference rate $d = 0.2549$, its median board has ten times as many tiers as the median HELM board. We also report an optimistic scenario at $d = 0.05$, which reduces the estimated limits for these high-accuracy models. Under this assumption, ImageNet contains 88 tiers and ImageNet-R contains 149 (Appendix Table 7). The latter is approximately three times the 49 tiers of HellaSwag, the language benchmark with the most tiers at the reference rate after accounting for Open LLM v1 MMLU's macro-averaged score (Methods). The larger vision tier counts are consistent with test sets of 30,000 and 50,000 instances and score ranges of seventy and fifty accuracy points, compared with two to five tiers on the median survey board.

At the optimistic rate, vision's median tier count is 6.2% of its median model count, and 98.9% of adjacent gaps remain below the limit. Absolute tier counts and their fractions of the model population therefore give different summaries: vision has the most tiers, but five families have a larger tier-to-model ratio under the common reference-rate comparison (Fig. 7d; Appendix A.11).

### B.5 Resolution of embedding leaderboards

The embedding family ranks approximately 250 text-embedding models using downstream classification scores (Fig. 8). At the common reference rate $d = 0.2549$ (Fig. 7d), it has a median of 20 tiers

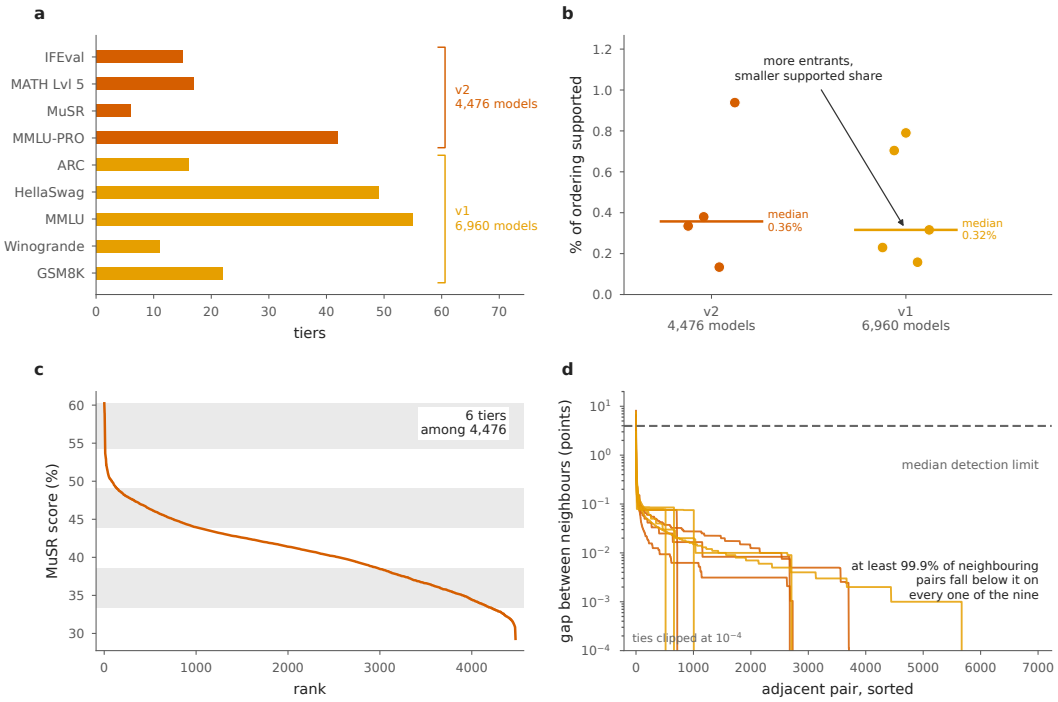



Figure 6: **Resolution of two Open LLM Leaderboard generations.** The v2 and v1 populations contain 4,476 and 6,960 models, respectively, and share no benchmarks. **a**, Tier counts by benchmark, with v2 in vermillion and v1 in gold; labels give model counts. **b**, Tier count as a fraction of the number of ranked models. The larger population has a smaller fraction. **c**, Ranked MuSR scores for all 4,476 models, with boundaries marking six tiers. **d**, Sorted adjacent score gaps for all nine benchmarks on a logarithmic scale, compared with the median detection limit. At least 99.93% of adjacent pairs on every benchmark fall below the limit; on four benchmarks, every pair does so. Tied scores have zero gaps and are displayed at $10^{-4}$.

among 258 models, a tier-to-model ratio of 7.8%. At the optimistic rate $d = 0.05$, it has 36 tiers and a ratio of 14%, while 94.2% of adjacent gaps remain below the limit (Appendix Table 8). This is the largest ratio obtained under the family-specific assumptions, which differ from the common-rate comparison. MTEB scores average ten evaluation runs rather than one set of binary item outcomes. McNemar's decomposition therefore does not apply directly, and these limits are approximate.

The embedding analysis covers only these three tasks. ToxicConversationsClassification was excluded during curation on the basis of an unsupported inference about its metric. We retain the original three-task analysis and make no resolution claim about the excluded task.

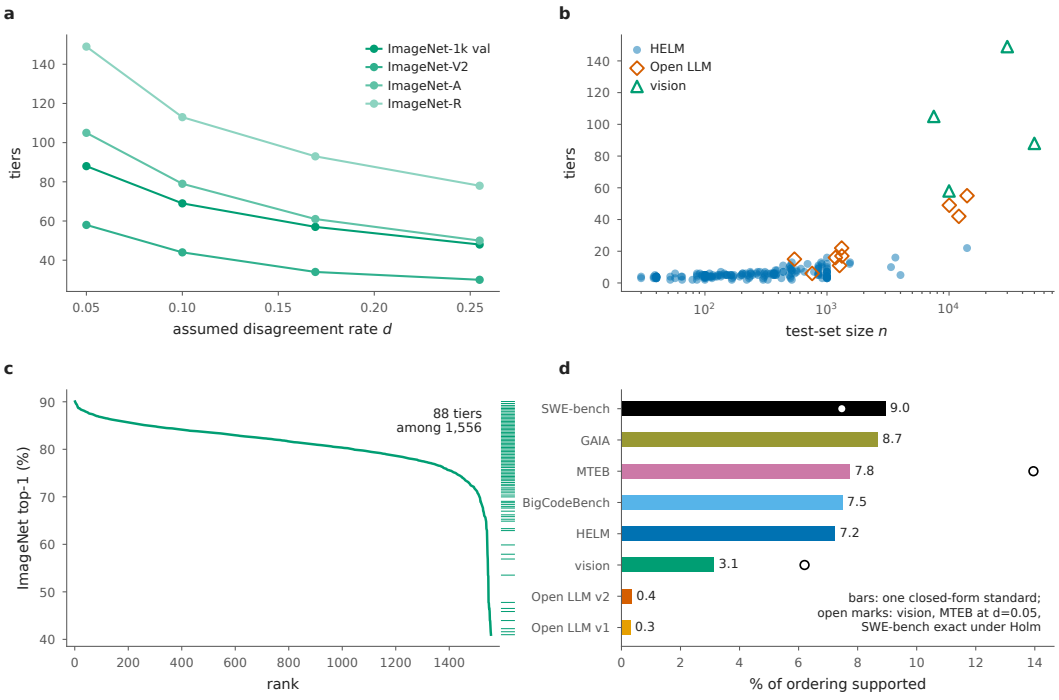



Figure 7: **Resolution of 1,556 image classifiers on ImageNet and related benchmarks.** **a**, Tier counts versus assumed disagreement rate. The family-specific analysis uses the optimistic rate $d = 0.05$, yielding the largest tier counts among the assumptions considered. **b**, Tier counts versus test-set size for vision and language families. Vision contains approximately an order of magnitude more tiers, consistent with its larger test sets, including 50,000-item ImageNet. **c**, Ranked ImageNet top-1 accuracy. The 88 tier boundaries are marked outside the right axis to indicate their density without obscuring the score curve. **d**, Ratio of median tier count to median model count for all eight families. Bars use the common closed-form limit at the reference disagreement rate. Vision's 3.1% ranks sixth, below SWE-bench, GAIA, MTEB, BigCodeBench, and HELM. Open markers show alternative criteria discussed in the text: vision at 6.2% and MTEB at 14.0% under $d = 0.05$, and SWE-bench at 7.5% under Holm-corrected exact tests.

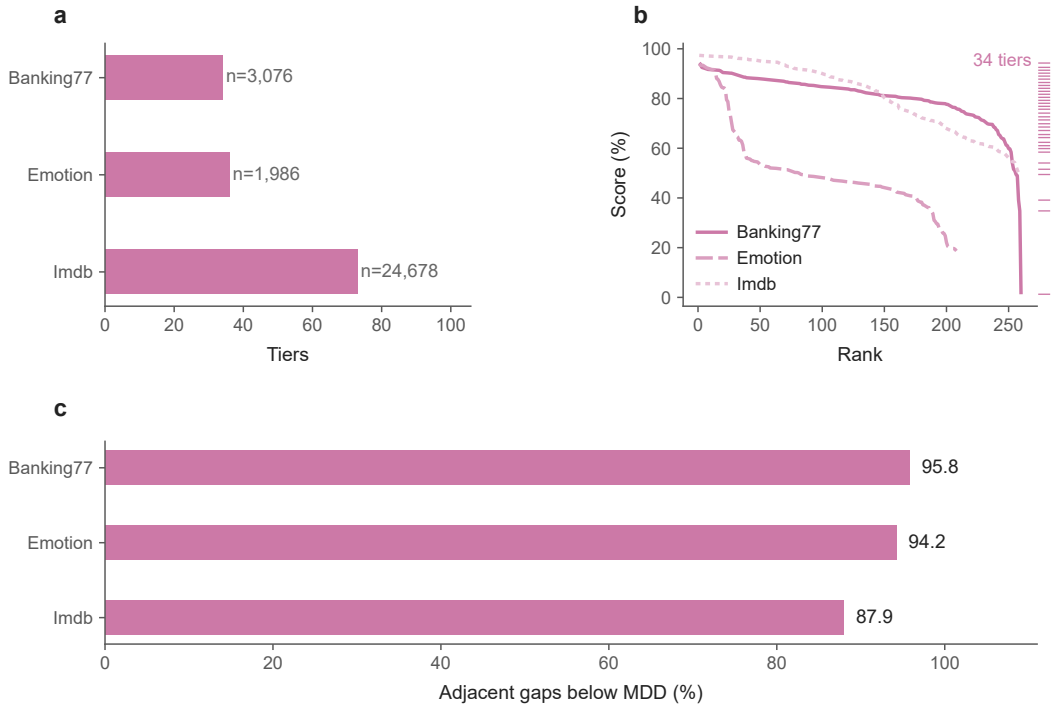



Figure 8: **Approximate resolution of three MTEB classification tasks. a**, Tier counts and evaluated sample sizes by task. **b**, Ranked scores for three tasks. Banking77's 34 tier boundaries are marked outside the right axis. **c**, Fractions of adjacent gaps below the assumed detection limit, with a minimum of 87.9% across these tasks. Results use the optimistic disagreement rate $d = 0.05$; limits are approximate because scores average multiple evaluation runs.

## C Supplementary figures and tables

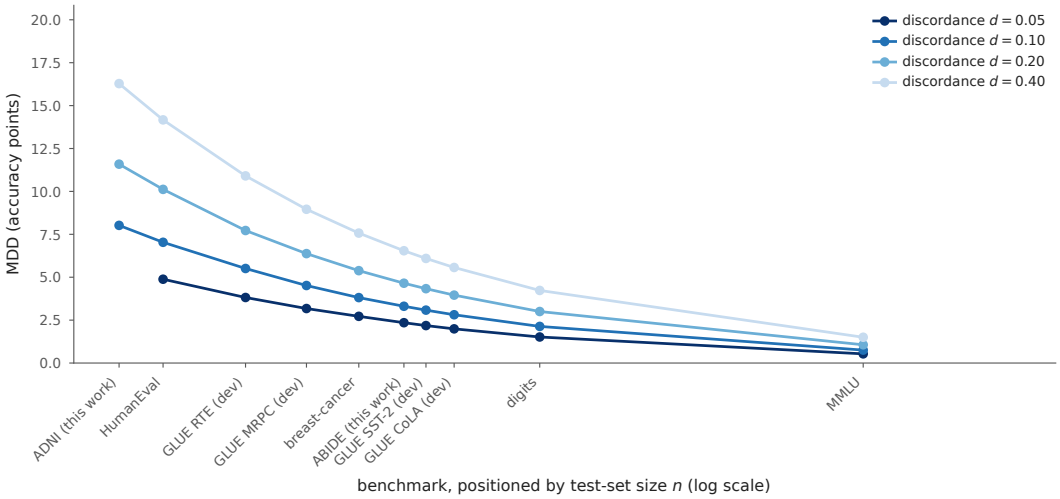



Figure 9: **Detection limits conditional on test-set size and disagreement.** Minimum accuracy gaps with 80% power under McNemar's exact test for ten benchmarks. Horizontal positions indicate published test-set sizes; each curve corresponds to an assumed disagreement rate $d$. The calculation requires no model training or reproduction of published scores. At the lowest disagreement rate considered, HumanEval's 164 items yield a limit of 4.88 accuracy points and GLUE RTE's 277 items a limit of 3.82 accuracy points. MMLU is the only benchmark shown with a limit below one accuracy point, attained at low disagreement rates. Combinations for which 80% power is unattainable are marked `n/a`: when $\Pr(\mathrm{Bin}(n, d) \geq 6) < 0.8$, the design cannot produce, with sufficient probability, the minimum six disagreements required for a two-sided rejection. Test-set sizes were checked against published sources; benchmarks with unverified sizes were excluded. These limits omit additional uncertainty from stochastic generation and prompt selection.

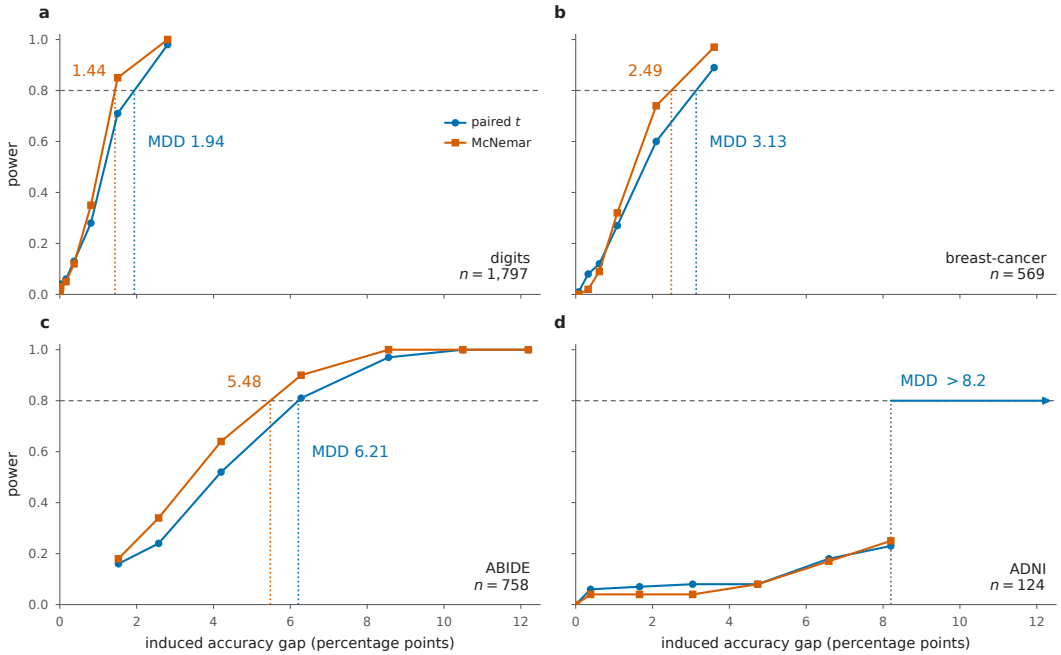



Figure 10: **Empirical power under controlled prediction perturbations.** Power is plotted against the induced accuracy gap on a common horizontal scale. **a**, digits, $n$ = 1,797. **b**, breast-cancer, $n$ = 569. **c**, ABIDE, $n$ = 758. **d**, ADNI, $n$ = 124. Blue denotes paired $t$-tests across cross-validation folds; vermillion denotes McNemar's exact tests on pooled predictions. Each point summarizes 100 independent stratified 10-fold repeats. Dotted vertical lines show the interpolated crossing of 80% power: 1.94, 3.13, and 6.21 accuracy points for the paired $t$-test on digits, breast-cancer, and ABIDE, respectively, and 1.44, 2.49, and 5.48 for McNemar's test. ADNI does not reach 80% power over the evaluated range, and 100% of bootstrap resamples retain this outcome. Its limit is therefore shown as a lower-bound arrow along the 80% power line. Inverting the closed-form approximation gives implied disagreement rates that can be compared with rates measured from per-item outcomes. The digits estimate, 0.045, lies within the measured range 0.0073–0.0963; the breast-cancer estimate, 0.0425, exceeds the measured maximum of 0.0414. No directly measured rate is available for ABIDE. Each implied rate fits one parameter to one observed crossing, so this inversion alone provides no residual goodness-of-fit check. Of the two available comparisons with measured ranges, one falls outside its range.

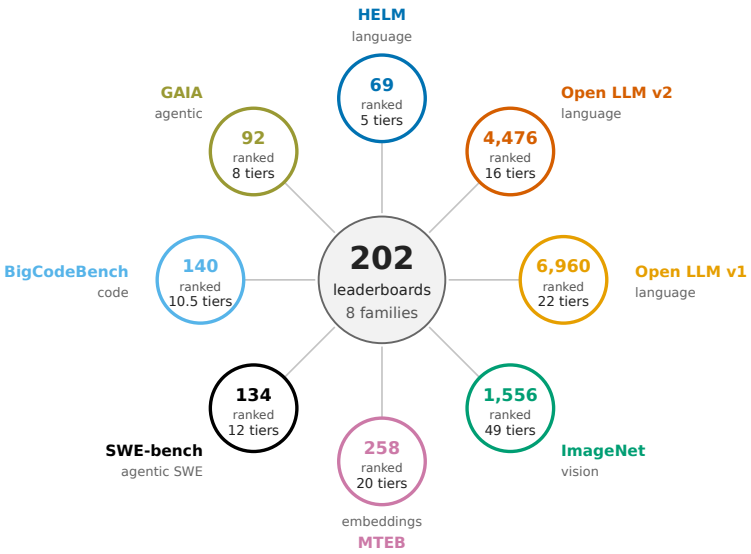



Figure 11: **Coverage and resolution of the leaderboard survey.** Eight leaderboard families comprise 202 distinct boards spanning language, vision, embeddings, code, and agentic software engineering. Each spoke reports the median model count and median tier count under the survey's common closed-form criterion. Model counts range from 69 to 6,960 and tier counts from five to 49. In every family, the median tier count is substantially smaller than the median number of ranked models.

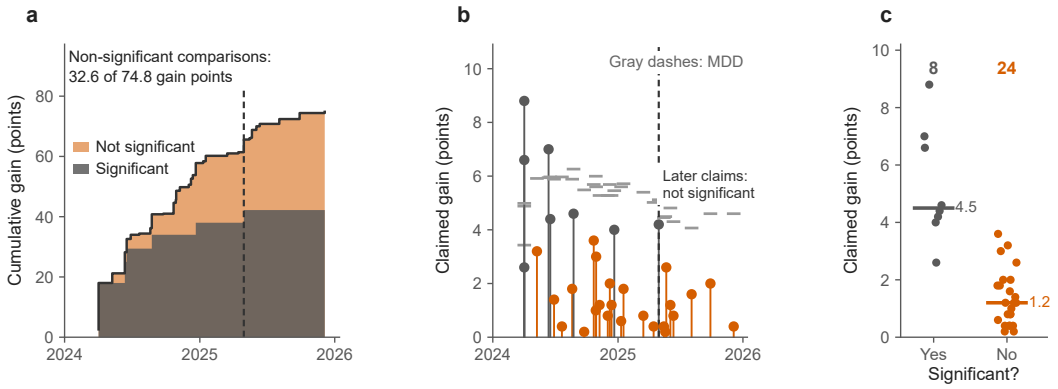



Figure 12: **Exact-test outcomes for successive SWE-bench Verified records.** Chronologically ordering 134 submissions and retaining each new maximum identifies 33 records and 32 successive improvements. Each improvement is tested against the record it replaces. **a**, Cumulative gains: of 74.8 accuracy points reported since April 2024, 32.6 arise from comparisons that do not reject under the stated exact-test criterion. **b**, Each observed gain compared with MDD$(500, d)$ at its pair-specific disagreement rate (gray dash). **c**, Median gains are 4.5 accuracy points for the 8 significant claims and 1.2 for the 24 non-significant claims. The dashed line marks 30 April 2025, the most recent significant claim under the base criterion; under Benjamini–Hochberg correction this date is 12 June 2024. Exact-test and closed-form classifications agree on 27 of 32 claims. The five disagreements reflect the distinction between an observed significance test and a target-power threshold.

The following tables report the survey populations, family-level results, exact pairwise tests, disagreement-rate transfer analyses, power validation, and sample-size requirements. "Unresolvable" and "indistinguishable" are retained in table headings as shorthand for the criterion specified in each caption. For power-based analyses they mean that the observed gap is below the stated detection limit, not that an exact test necessarily fails to reject or that the systems are equivalent. Tables using exact tests identify their significance criterion separately.

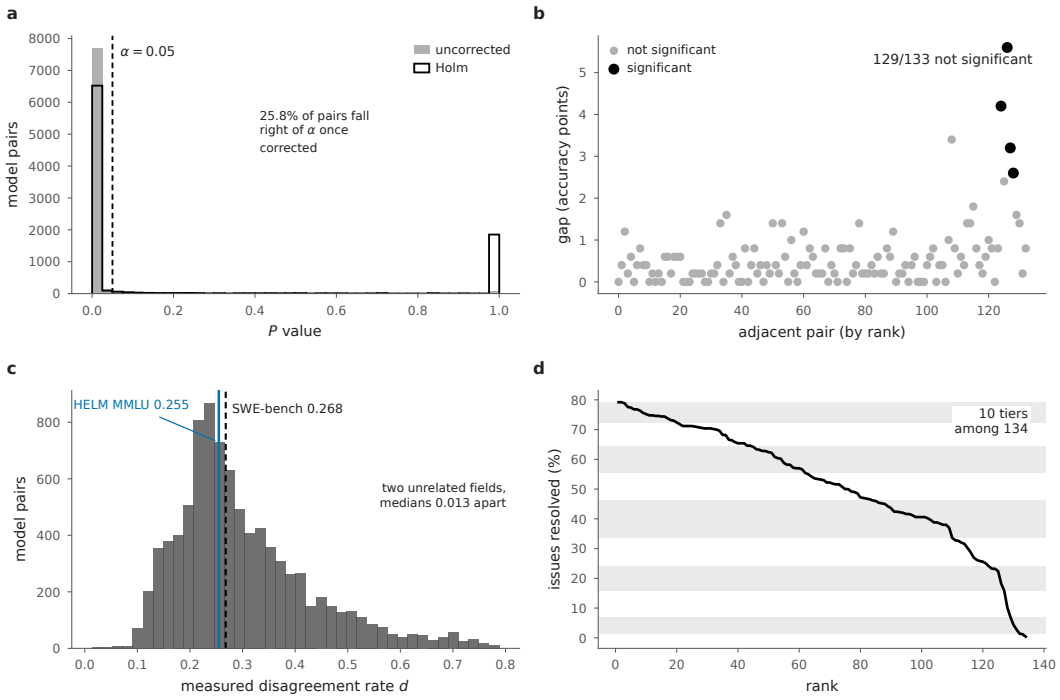



Figure 13: **Exact pairwise comparisons of 134 SWE-bench Verified systems.** All 8,911 comparisons use released per-item outcomes, without assuming a common disagreement rate. **a**, Uncorrected and Holm-corrected $P$ values; under Holm correction, 25.8% of pairs do not reject at $\alpha = 0.05$. **b**, Adjacent pairs ordered by accuracy gap and colored by significance. Four of 133 pairs reject; the median adjacent gap is 0.400 accuracy points. **c**, Distribution of measured pairwise disagreement rates, marking its median of 0.268 and HELM MMLU's median of 0.255, a difference of 0.013. **d**, Ranked scores with exact-test tier boundaries; the 134 systems form ten tiers.

| family | boards | models | tiers | % of order | adj. indist. | test |
|---|---|---|---|---|---|---|
| HELM (6 suites) | 182 | 69 | 5 | 7.2 | 98.7 | closed form |
| Open LLM v2 | 4 | 4,476 | 16 | 0.4 | 100.0 | closed form |
| vision (ImageNet) | 4 | 1,556 | 49 | 3.1 | 98.9 | closed form |
| SWE-bench Verified | 1 | 134 | 12 | 9.0 | 97.0 | exact |
| MTEB (embeddings) | 3 | 258 | 20 | 7.8 | 94.2 | closed form |
| Open LLM v1 | 5 | 6,960 | 22 | 0.3 | 100.0 | exact (4) |
| BigCodeBench | 2 | 140 | 10.5 | 7.5 | 99.3 | closed form |
| GAIA | 1 | 92 | 8 | 8.7 | 98.9 | closed form |

Table 1: **Summary of eight leaderboard families and 202 distinct boards.** The collection contains 267 published tables, including 65 duplicated HELM boards counted once. Values are family-level medians. Tier counts and fractions of the ordering use the common reference rate $d = 0.2549$ for single-comparison limits. Under the alternative optimistic rate $d = 0.05$, vision has 96.5 tiers and 6.2% of the ordering, and MTEB has 36 tiers and 14.0%. Holm-corrected exact tests give SWE-bench 10 tiers and 7.5%. The adjacent-indistinguishable column uses different criteria: vision and MTEB retain $d = 0.05$, yielding 98.9% and 94.2%, rather than the 99.5% and 98.5% obtained at the reference rate. Applying the common rate to those rows would increase the column minimum from MTEB's 94.2% to SWE-bench's 97.0%. The tier-to-model ratios under the common criterion span 0.3–9.0%. The two Open LLM generations are separate families because they use different benchmarks and evaluation frameworks. Median tier counts range from HELM's 5 to vision's 49. The table's adjacent-pair fractions are at least 94.2%, and tier-to-model ratios are at most 9.0% at the reference rate. The test column identifies families with exact analyses: HELM's MMLU panel, four Open LLM v1 boards, and SWE-bench Verified provide per-item outcomes. "Exact (4)" denotes four such boards. Other analyses use the closed-form limit at $d = 0.2549$, measured on HELM's MMLU items, except for the explicitly identified vision and MTEB cells. Exact-test non-rejection and a gap below a power threshold are distinct criteria.

| disagreement rate | MDD (pp) | tiers | adjacent indistinguishable |
|---|---|---|---|
| measured 5th pct ($d = 0.169$) | 0.979 | 27 | 53/64 (82.8%) |
| measured median ($d = 0.255$) | 1.200 | 23 | 54/64 (84.4%) |
| measured 95th pct ($d = 0.543$) | 1.749 | 18 | 58/64 (90.6%) |

Table 2: **Resolution of the complete 57-subject MMLU benchmark.** The analysis includes $n = 14{,}042$ items and 65 models evaluated on every subject. The assumed disagreement rates are the measured percentiles from the same benchmark's 514-instance sample. MDD is in accuracy points (pp); adjacent-indistinguishable counts denote gaps below that limit.

| benchmark | $n$ | $d = 0.05$ | $d = 0.10$ | $d = 0.20$ | $d = 0.40$ |
|---|---|---|---|---|---|
| ADNI (this work) | 124 | n/a | 8.02 | 11.59 | 16.28 |
| HumanEval | 164 | 4.88 | 7.04 | 10.12 | 14.17 |
| GLUE RTE (dev) | 277 | 3.82 | 5.51 | 7.72 | 10.90 |
| GLUE MRPC (dev) | 408 | 3.18 | 4.52 | 6.38 | 8.96 |
| breast-cancer | 569 | 2.72 | 3.81 | 5.38 | 7.57 |
| ABIDE (this work) | 758 | 2.35 | 3.31 | 4.65 | 6.54 |
| GLUE SST-2 (dev) | 872 | 2.18 | 3.08 | 4.33 | 6.09 |
| GLUE CoLA (dev) | 1,043 | 1.99 | 2.81 | 3.96 | 5.57 |
| digits | 1,797 | 1.52 | 2.13 | 3.00 | 4.23 |
| MMLU | 14,042 | 0.54 | 0.75 | 1.06 | 1.50 |

Table 3: **Exact-power detection limits for published benchmarks and two neuroimaging tasks.** Entries give accuracy gaps in accuracy points at 80% power. The symbol n/a indicates that no feasible gap reaches the target power at the specified sample size and disagreement rate.

| $n$ | separation | ref. acc. (%) | SE (pp) | MDD | MDD/SE |
|---|---|---|---|---|---|
| 200 | 0.6 | 84.7 | 2.547 | 10.26 | 4.03 |
| 200 | 1.0 | 88.6 | 2.243 | 10.28 | 4.58 |
| 200 | 1.6 | 96.9 | 1.218 | 5.70 | 4.68 |
| 200 | 2.4 | 99.3 | 0.600 | 5.89 | 9.82 |
| 400 | 0.6 | 78.4 | 2.056 | 7.44 | 3.62 |
| 400 | 1.0 | 91.7 | 1.382 | 4.43 | 3.21 |
| 400 | 1.6 | 94.1 | 1.178 | 4.75 | 4.03 |
| 400 | 2.4 | 98.0 | 0.702 | 3.46 | 4.93 |
| 800 | 0.6 | 88.4 | 1.133 | 3.75 | 3.31 |
| 800 | 1.0 | 94.0 | 0.841 | 3.08 | 3.66 |
| 800 | 1.6 | 98.1 | 0.478 | 2.15 | 4.50 |
| 800 | 2.4 | 99.0 | 0.360 | 2.61 | 7.26 |
| 1600 | 0.6 | 80.5 | 0.991 | 2.92 | 2.95 |
| 1600 | 1.0 | 96.7 | 0.445 | 1.15 | 2.58 |
| 1600 | 1.6 | 98.1 | 0.340 | 1.47 | 4.33 |
| 1600 | 2.4 | 99.5 | 0.170 | 0.81 | 4.79 |
| 3200 | 0.6 | 81.9 | 0.680 | 2.04 | 3.00 |
| 3200 | 1.0 | 95.0 | 0.386 | 1.25 | 3.25 |
| 3200 | 1.6 | 98.9 | 0.181 | 0.63 | 3.47 |
| 3200 | 2.4 | 99.3 | 0.146 | 0.79 | 5.44 |

Table 4: **Detection limits across a synthetic sample-size and class-separation grid.** Sample size and class separation vary independently. MDD is estimated using McNemar's test. Both MDD and the binomial standard error (SE), $\sqrt{p(1-p)/n}$ for reference accuracy $p$, are in accuracy points (pp). With $Z = z_{0.975} + z_{0.80}$, the predicted ratio is $Z\sqrt{d/p(1-p)}$. The fitted through-origin slope is MDD/SE = 4.05; individual ratios range from 2.58 to 9.82 and increase with class separation. We omit a through-origin $R^2$ because the shared $1/\sqrt{n}$ scaling can inflate that measure of agreement.

| | boundary defect | | label defect | |
|---|---|---|---|---|
| benchmark | paired $t$ | McNemar | paired $t$ | McNemar |
| ADNI | > 8.2 | > 8.2 | > 11.3 | > 11.3 |
| ABIDE | 6.21 [5.78, 6.88] | 5.48 [5.10, 5.83] | 6.43 [5.93, 6.86] | 5.90 [5.31, 6.28] |
| breast-cancer | 3.13 [2.84, 3.44] | 2.49 [2.11, 2.75] | 2.49 [2.33, 2.61] | 2.34 [2.15, 2.46] |
| digits | 1.94 [1.58, 2.14] | 1.44 [1.36, 1.58] | > 1.6 | 1.53 [1.45, 1.61] |

Table 5: **Empirically estimated detection limits under controlled defects.** Values are accuracy gaps in accuracy points at 80% power, with 95% bootstrap intervals over 100 repeats. A boundary defect displaces the fitted weight vector by a specified fraction of its norm; a label defect corrupts a fraction of training labels before refitting (Methods). Entries prefixed > are lower bounds because the evaluated range does not reach 80% power. The target remains unattained in 100% of bootstrap resamples for both ADNI defects and in 79.2% for the digits label-defect $t$-test. Each lower bound equals the largest gap reached by that perturbation sweep and therefore depends on the evaluated range. The wider sweep in Appendix Table 9 yields larger ADNI lower bounds for the same defect types. These empirical crossings are used to assess the closed-form approximation.

| benchmark | $n$ | models | spread | MDD | tiers | 5th–95th | adj. indist. |
|---|---|---|---|---|---|---|---|
| IFEval | 541 | 4,476 | 90.0 | 6.21 | 15 | 10–18 | 100.0 |
| MATH Lvl 5 | 1,324 | 4,476 | 71.5 | 3.95 | 17 | 13–21 | 100.0 |
| MuSR | 756 | 4,476 | 31.0 | 5.25 | 6 | 5–8 | 100.0 |
| MMLU-PRO | 12,032 | 4,476 | 63.0 | 1.30 | 42 | 30–51 | 99.9 |

Table 6: **Resolution of the Open LLM Leaderboard v2.** The snapshot contains 4,476 of 4,576 models (97.8%). One source page could not be retrieved; its identifier is recorded in the artifacts. GPQA is excluded because access is gated, and BBH because the evaluated subset of tasks could not be verified. Spread and MDD are in accuracy points. Adjacent-indistinguishable fractions denote gaps below the specified power threshold.

| evaluation set | $n$ | models | top-1 range | MDD | tiers |
|---|---|---|---|---|---|
| ImageNet-1k val | 50,000 | 1,556 | 41.0–90.1 | 0.28 | 88 |
| ImageNet-V2 | 10,000 | 1,556 | 31.6–82.8 | 0.64 | 58 |
| ImageNet-A | 7,500 | 1,556 | 0.5–88.6 | 0.73 | 105 |
| ImageNet-R | 30,000 | 1,556 | 13.4–90.8 | 0.36 | 149 |

Table 7: **Resolution of timm ImageNet evaluations under $d = 0.05$.** Published result tables provide accuracies without sample counts, so sizes were checked against evaluation-set mirrors. ImageNet-Sketch is excluded because the mirrored size of 30,100 differs from the commonly reported size. ImageNet-ReaL is excluded because it relabels the same 50,000 images as ImageNet validation. Top-1 ranges are percentages; MDD is in accuracy points.

| task | $n$ | models | score range | MDD | tiers | adj. indist. |
|---|---|---|---|---|---|---|
| Banking77Classification | 3,076 | 260 | 1.3–94.3 | 1.15 | 34 | 95.8 |
| EmotionClassification | 1,986 | 208 | 18.7–93.9 | 1.44 | 36 | 94.2 |
| ImdbClassification | 24,678 | 258 | 50.1–97.4 | 0.40 | 73 | 87.9 |

Table 8: **Approximate resolution of MTEB classification under $d = 0.05$.** The three task scores average ten evaluation runs rather than a single item-level binary outcome vector; only 3.5–5.8% lie on the $1/n$ lattice. McNemar's decomposition therefore does not apply directly. ToxicConversationsClassification was excluded during curation on the basis of an unsupported metric inference; the reported scope remains restricted to these three tasks. Calculations use a stored snapshot because repeated identical live queries returned different row sets. Score ranges are percentages; MDD is in accuracy points.

| **a** by metric (boundary defect, logistic regression) | | | |
|---|---|---|---|
| benchmark | accuracy | AUC | $F_1$ |
| ADNI | > 12.7 | > 18.6 | > 12.5 |
| ABIDE | 6.21 | 4.57 | 7.75 |
| breast-cancer | 3.13 | 9.77 | 3.00 |
| digits | 1.94 | 2.82 | 4.28 |

| **b** by estimator (label defect, accuracy) | | | | |
|---|---|---|---|---|
| benchmark | logistic | linear SVM | random forest | MLP |
| ADNI | > 14.7 | > 12.4 | > 10.4 | — |
| ABIDE | 6.43 | 6.39 | 6.50 | — |
| breast-cancer | 2.49 | 3.06 | 3.40 | 2.74 |
| digits | > 1.6 | 1.90 | 2.06 | 1.89 |

Table 9: **Detection-limit sensitivity to metric and estimator.** **a**, Metric-specific MDDs under boundary perturbations of logistic regression. **b**, Accuracy MDDs under label corruption for different estimators. Label corruption applies to every estimator; the boundary perturbation requires an exposed linear decision boundary. In these analyses, changing the estimator produces smaller changes than changing the metric, and the benchmark ordering is not preserved across metrics. Each value uses units of its associated metric: accuracy points for panel b and the accuracy column of panel a.

| suite | boards | $R/L$ | $k$ | $k^{\mathrm{SL}}$ | $k/k^{\mathrm{SL}}$ | amb. bnd. (%) | 1 blk |
|---|---|---|---|---|---|---|---|
| classic | 78 | 5.9 | 5 | 2 | 3.0 | 77.9 | 34 |
| mmlu | 58 | 5.8 | 5 | 2 | 2.5 | 77.0 | 18 |
| lite | 20 | 5.7 | 6 | 1 | 4.0 | 87.7 | 12 |
| medhelm | 19 | 3.4 | 3 | 1 | 2.0 | 64.3 | 11 |
| thaiexam | 6 | 4.7 | 5 | 1 | 5.0 | 96.2 | 5 |
| finance | 1 | 10.5 | 7 | 5 | 1.4 | 33.3 | 0 |
| all | 182 | 5.3 | 5 | 2 | 3.0 | 78.5 | 80 |

Table 10: **Resolution and tier structure across six HELM suites.** Columns $R/L$, $k$, $k^{\mathrm{SL}}$, and $k/k^{\mathrm{SL}}$ are medians of board-level values within each suite. The ratio $R/L$ compares the observed score range with the detection limit and bounds the greedy tier count by $k \leq 1 + \lfloor R/L \rfloor$. Small tier counts can thus arise from a large limit or a narrow score range. $k^{\mathrm{SL}}$ is the single-linkage count at the same threshold, the coarsest partition without an adjacent gap below the threshold at a group boundary. For $M$ entrants and adjacent-below-limit fraction $a$, it equals $M - a(M - 1)$. "Amb. bnd." is the fraction of greedy tier boundaries whose adjacent systems have gaps below the limit; "1 blk" counts boards with one single-linkage group. Eligibility requires a per-item binary metric and a single instance count. Two additional suites were inspected but contained no eligible tables.

| benchmark | $n$ | models | score range | MDD | tiers | adj. indist. |
|---|---|---|---|---|---|---|
| ARC | 1,172 | 6,960 | 19.7–87.5 | 4.20 | 16 | 100.0 |
| HellaSwag | 10,042 | 6,960 | 20.7–92.8 | 1.42 | 49 | 100.0 |
| MMLU | 14,042 | 6,960 | 17.8–89.4 | 1.20 | 55 | 100.0 |
| Winogrande | 1,267 | 6,960 | 47.2–91.5 | 4.04 | 11 | 100.0 |
| GSM8K | 1,319 | 6,960 | 0.0–88.2 | 3.96 | 22 | 100.0 |

Table 11: **Resolution of the archived Open LLM Leaderboard v1.** We retrieved 6,960 of 7,260 models (95.9%). Sample sizes were checked against each source dataset's evaluated split. TruthfulQA is excluded because v1 reports MC2, the probability mass assigned to true statements, rather than a binary correctness outcome. The same binary-outcome requirement excludes HELM F1 and BLEU tables. The MMLU row uses the uniform closed-form approximation. Accounting for its 57-subject macro-average gives a limit of 1.43 accuracy points and 47 tiers (Methods). Score ranges are percentages; MDD is in accuracy points.

| benchmark | $n$ | models | score range | MDD | tiers | adj. indist. |
|---|---|---|---|---|---|---|
| BigCodeBench-Complete | 1,140 | 154 | 7.9–62.4 | 4.26 | 12 | 99.3 |
| BigCodeBench-Instruct | 1,140 | 126 | 7.0–51.1 | 4.26 | 9 | 99.2 |
| GAIA (validation) | 165 | 92 | 0.0–99.4 | 11.31 | 8 | 98.9 |

Table 12: **Resolution of BigCodeBench and GAIA.** BigCodeBench evaluates code generation; its sample size is identical across all released dataset versions. GAIA evaluates agentic assistance. Its reported scores support the stated sample size: all 92 values are multiples of $1/165$, and level-1 scores are multiples of $1/53$, consistent with the 53/86/26 split. With 165 items, GAIA has a detection limit of 11.31 accuracy points, the largest among the individually named main-text boards. Some HELM sub-scenarios are smaller, reaching 25.26 accuracy points at $n = 30$. Score ranges are percentages; MDD is in accuracy points.

| board | models | pairs | median $d$ | 5–95% |
|---|---|---|---|---|
| HellaSwag | 14 | 46 | 0.1314 | 0.053–0.420 |
| Winogrande | 7 | 9 | 0.1942 | 0.135–0.390 |
| ARC | 13 | 36 | 0.2730 | 0.066–0.474 |
| GSM8K | 27 | 259 | 0.3222 | 0.161–0.740 |
| MMLU | 15 | 55 | 0.3518 | 0.055–0.684 |
| HELM MMLU | 91 | 4,095 | 0.2549 | — |
| SWE-bench | 134 | 8,911 | 0.2680 | — |

Table 13: **Variation in measured disagreement rates across boards and model populations.** Open LLM v1 rows report all-pairs median rates from per-item outcomes for five boards, alongside the HELM and SWE-bench reference rates. Within the Open LLM family, rates differ by a factor of 2.68, implying a factor of 1.64 in the approximate limit. The MMLU rates of 0.3518 and 0.2549 use different item panels: 13,937 Open LLM v1 items and 514 HELM items. Their 1.23-fold limit ratio therefore combines item-panel and model-population differences. Restricting Open LLM v1 outcomes to HELM's 514 items gives rates of 0.3093 and 0.2549, a 1.21× rate ratio and a 1.10-fold limit ratio attributable to the different model populations on this fixed panel. Open LLM rows include every cached run of the collected record-holding models. A run pair is eligible when it shares at least 100 scored items, and disagreement is computed only on those shared items. This admits 46 of 91 possible pairs among 14 HellaSwag runs; no other pair-level exclusion is applied. Counts are small, including nine Winogrande pairs, so the reported spread should be interpreted within these populations. Rates remain unmeasured for the other five families, whose closed-form results depend on the assumed rate.

| | | median MDD (pp) | | | median tiers | | | adj. indist. (%) | | |
|---|---|---|---|---|---|---|---|---|---|---|
| family | boards | $\tau$=.5 | .8 | .9 | $\tau$=.5 | .8 | .9 | $\tau$=.5 | .8 | .9 |
| BigCodeBench | 2 | 3.01 | 4.26 | 4.91 | 14.5 | 10.5 | 9 | 98.6 | 99.3 | 99.6 |
| GAIA | 1 | 8.21 | 11.31 | 12.86 | 11 | 8 | 8 | 98.9 | 98.9 | 100.0 |
| HELM | 182 | 6.80 | 9.44 | 10.75 | 6 | 5 | 4.5 | 98.5 | 98.7 | 98.9 |
| MTEB | 3 | 1.82 | 2.58 | 2.98 | 26 | 20 | 17 | 96.1 | 98.5 | 98.5 |
| Open LLM v1 | 5 | 2.80 | 3.96 | 4.56 | 31 | 22 | 20 | 100.0 | 100.0 | 100.0 |
| Open LLM v2 | 4 | 3.26 | 4.60 | 5.29 | 22.5 | 16 | 14 | 100.0 | 100.0 | 100.0 |
| SWE-bench | 1 | 4.61 | 6.47 | 7.43 | 15 | 12 | 10 | 99.2 | 100.0 | 100.0 |
| vision | 4 | 0.79 | 1.12 | 1.30 | 67 | 49 | 43.5 | 99.3 | 99.5 | 99.6 |
| all | 202 | — | — | — | 6 | 5 | 5 | 98.5 | 98.9 | 99.9 |

Table 14: **Sensitivity of survey summaries to the target power.** Column headings specify target power $\tau$; MDD is in accuracy points (pp). The primary analysis uses 80% power. This table also evaluates 50% and 90% for all 202 boards, grouped by family. Two HELM boards have no finite limit at 90% power and are excluded from that column. For a gap of two accuracy points, the required sample size increases by a factor of 2.7, from 1,633 items at 50% power to 4,466 at 90%. Across the survey, the median fraction of adjacent gaps below the limit remains high: 98.5%, 98.9%, and 99.9% at the three targets. This fraction is determined by $M$ and $k^{\mathrm{SL}}$ at each threshold, so it is not independent evidence beyond those quantities. Tier counts and sample-size requirements change with the target power even where adjacent-pair fractions remain near their upper bound.

| population | $N$ | statistics reported on it |
|---|---|---|
| published leaderboard tables | 267 | the collection |
| HELM tables | 247 | |
| distinct boards | 202 | *every survey statistic*: limits, tiers, adjacency, sizing, the rate-free floor |
| distinct HELM boards | 182 | $R/L$, $k^{\mathrm{SL}}$, ambiguous boundaries, per-board $d$ |
| boards with untied leaders | 168 | the top-1/top-2 requirement |
| boards releasing per-item outcomes | 7 | the exact tests |

Table 15: **Analysis populations and eligibility criteria.** A survey board requires both a published sample size and a score table; an exact-test analysis additionally requires per-item outcomes. Counting the 65 duplicated HELM entries once reduces 267 published tables to 202 distinct boards. Among these, 34 have tied leading scores and therefore no finite sample-size requirement for detecting the observed zero gap. The top-two sizing analysis consequently has 168 boards with finite requirements. Each reported statistic uses the population indicated here.

| suite | boards | med. $k$ | ceiling | $k$/ceiling | med. $k^{\mathrm{SL}}$ | $a$ (%) |
|---|---|---|---|---|---|---|
| classic | 78 | 5 | 6 | 0.83 | 2 | 98.5 |
| mmlu | 58 | 5 | 6 | 0.83 | 2 | 98.7 |
| lite | 20 | 6 | 6 | 0.85 | 1 | 100.0 |
| medhelm | 19 | 3 | 4 | 0.75 | 1 | 100.0 |
| thaiexam | 6 | 5 | 5 | 1.00 | 1 | 100.0 |
| finance | 1 | 7 | 11 | 0.64 | 5 | 81.0 |

Table 16: **Relation between adjacent gaps and tier counts.** The ceiling $1 + \lfloor R/L \rfloor$ bounds the greedy tier count for the observed score range and limit; $k^{\mathrm{SL}}$ denotes the single-linkage count. The adjacent-below-limit fraction satisfies $a = 1 - (k^{\mathrm{SL}} - 1)/(M - 1)$, with discrepancy 0.0000 in percent units on every board. Because medians are calculated separately, the identity need not hold after substituting marginal medians. For example, $M = 69$ and $k^{\mathrm{SL}} = 2$ give $a = 98.53\%$, whereas the measured median is 98.7%, corresponding algebraically to $k^{\mathrm{SL}} = 1.88$ at that model count. The median single-linkage count is 2. The ratio $R/L$ provides complementary information by comparing the population's score spread with the benchmark's detection limit.

| rate | $d$ | MDD (pp) | tiers | adj. (%) | $R/L$ | $k^{\mathrm{SL}}$ |
|---|---|---|---|---|---|---|
| 5th percentile | 0.1855 | 7.15 | 6 | 98.5 | 6.7 | 2 |
| borrowed | 0.2549 | 9.44 | 5 | 98.7 | 5.3 | 2 |
| median | 0.3488 | 11.06 | 4 | 98.9 | 4.5 | 2 |
| 95th percentile | 0.4947 | 13.13 | 4 | 100.0 | 3.8 | 1 |

Table 17: **Sensitivity to the empirical disagreement-rate distribution.** Rates are measured on 80 of 182 distinct HELM boards. The 5th, 50th, and 95th percentiles bracket the reference rate used in the survey. Each row recomputes the summaries at one common rate on every board with a finite limit: 182 boards at three rates and 165 at the lowest rate. The main text uses the 165-board intersection to keep the population fixed; this table uses all eligible boards at each rate. The reference rate is below the measured median and therefore yields more permissive resolution estimates. MDD is in accuracy points (pp).

| board | $n$ | pairs | power band | signed error | mean \|error\| |
|---|---|---|---|---|---|
| HELM MMLU | 100 | 286 | all | −0.3 | 13.9 |
| | 100 | 89 | 50–95% | −4.8 | 18.7 |
| | 200 | 290 | all | −0.1 | 14.5 |
| | 200 | 74 | 50–95% | −7.3 | 18.7 |
| SWE-bench Verified | 100 | 293 | all | −1.1 | 10.2 |
| | 100 | 73 | 50–95% | −7.2 | 18.3 |
| | 200 | 292 | all | −2.8 | 9.0 |
| | 200 | 40 | 50–95% | −9.6 | 21.9 |

Table 18: **Predicted power versus rejection rates on held-out items.** Disagreement and accuracy gap are estimated on half of each board's items. Power is evaluated using 200 draws with replacement from the other half for each of 300 randomly selected model pairs. Replacement sampling matches the independent-item model underlying the power calculation. Errors are in power units on a percent scale; banded rows retain pairs with predicted power between 50% and 95%. Unbanded rows evaluate all eligible pairs, whereas predictor-selected bands condition on estimation noise and require separate interpretation. Across all pairs, mean signed errors are −0.3, −0.1, −1.1, and −2.8, while mean absolute errors are 13.9, 14.5, 10.2, and 9.0. Estimating both quantities on all items reduces the absolute errors to 7.1, 7.5, 5.3, and 4.8. This comparison changes whether the evaluated items contribute to estimation, as well as the number of fitted items. For equal halves, a full-item empirical estimate averages the two half-item estimates, giving a $1/2$ error ratio under a linear approximation; all four intervals contain that ratio. The gap estimate accounts for most of the reduction: in the first cell, estimating only $\Delta$ on all items gives 6.9, compared with 7.1 when both quantities use all items; estimating only $d$ there gives 14.2, compared with 13.9 when neither does. The same qualitative pattern holds in all four cells (`src/mx28_validate_fullrate.py`). Signed error alone does not measure this variance reduction. Selecting a band using the held-out-design predictor makes full-item estimation appear better by 5 to 17 in power units, whereas selecting with the full-item predictor reverses that apparent comparison. A selector estimated on a third split, disjoint from estimates fitted on a quarter of the items, is consistent with the unbanded comparison: errors are −4.8, −7.3, −7.2, and −9.6 for half-item estimation versus −2.7, −4.0, −3.5, and −5.4 for full-item estimation. A selector cannot be disjoint from all estimation data when one design uses every item, so the primary comparison is unbanded. For a practitioner selecting a power band from a pilot estimate, the conditional behavior remains relevant: within each design's own predicted band, held-out-design optimism is 8.8, 8.2, 6.2, and 11.3, versus 2.5, 2.2, 2.3, and 6.3 for full-item estimation. Pilot estimates therefore overstate power by approximately six to eleven in these percent units, with smaller errors when the evaluated items enter the estimate. Over 200 splits, held-out mean signed errors are +0.33, +0.21, +0.09, and +0.28, with every interval containing zero. The single-split value of −2.8 is consequently not evidence of a persistent board-specific bias. The bound $|\text{signed}| \leq 0.33$ summarizes calibration over splits; it does not identify a mechanism. Each observed rejection rate also has approximately ±3.5 in percent units of binomial sampling noise from 200 draws. When predictions and resampling both use a board's full empirical outcome table, agreement is within 0.8 in percent units, as expected when estimation and evaluation refer to the same empirical distribution.

| population | boards | tiers | adj. (%) | $R/L$ | med. $n$ | 2 pp | 1 pp | floor (%) |
|---|---|---|---|---|---|---|---|---|
| all boards | 202 | 5 | 98.89 | 5.8 | 303 | 12 | 5 | 20.99 |
| HELM only | 182 | 5 | 98.72 | 5.3 | 236 | 4 | 1 | 25.12 |
| non-HELM only | 20 | 19.5 | 99.94 | 22.8 | 1,655 | 8 | 4 | 2.95 |
| non-HELM, MTEB dropped | 17 | 17 | 99.97 | 18.1 | 1,319 | 7 | 3 | 3.40 |
| HELM parents only | 39 | 5 | 97.06 | 7.0 | 565 | 4 | 1 | 9.05 |
| slices dropped, all families | 59 | 8 | 98.90 | 10.5 | 1,000 | 12 | 5 | 5.13 |
| item-mean boards only | 181 | 5 | 99.16 | 5.3 | 235 | 6 | 2 | 24.88 |

Table 19: **Survey summaries under alternative board populations.** Resolution summaries use $d = 0.2549$. Adequacy counts for gaps of two and one accuracy points (columns 2 pp and 1 pp) use $d = 0.17$, the median disagreement between successive SWE-bench record holders, and the algebraic sample-size requirement. For two accuracy points, this requires 3,336 items, whereas exact power inversion requires 3,424. The latter includes 11 of the 12 boards in the first row; the main text reports 11. High adjacent-below-limit fractions and small tier-to-model ratios persist without HELM. Sample-size adequacy is more composition-sensitive: 12 of 202 boards (6%) meet the algebraic two-point requirement, compared with 8 of 20 non-HELM boards (40%). The full-survey median of 303 items contrasts with 1,655 among non-HELM boards. Resolving each board's observed top-two gap requires a median 49-fold increase in items in the full survey, versus 17-fold outside HELM and 25-fold after removing sub-scenario slices. These sizing summaries are therefore strongly influenced by the small tables published within HELM. The final column gives the median rate-free fraction of pairs that cannot reject at any feasible disagreement rate. Its threshold scales as $1/n$, making it particularly relevant to small test sets. The last row retains 181 boards with unweighted item-mean scores, for which $b - c = n\Delta$ holds exactly.

| gap to resolve | $d = 0.05$ | $d = 0.10$ | $d = 0.17$ | $d = 0.25$ | $d = 0.35$ |
|---|---|---|---|---|---|
| 0.5 accuracy points | 16,064 | 31,778 | 53,760 | 78,880 | 110,277 |
| 1.0 accuracy points | 4,086 | 8,032 | 13,534 | 19,813 | 27,665 |
| 2.0 accuracy points | 1,035 | 2,043 | 3,424 | 4,998 | 6,962 |
| 5.0 accuracy points | 157 | 337 | 562 | 817 | 1,133 |

Table 20: **Sample sizes required for specified detection limits.** Each entry gives the smallest test-set size found by bisection on the exact-power limit for the row's gap, the column's disagreement rate, 80% power, and $\alpha = 0.05$. Values use numerical exact power inversion rather than the algebraic approximation. Measured rates within one evaluation framework range from 0.131 to 0.352; when per-item outcomes are available, the rate can be estimated for the intended model population. The survey's median board has 303 items, fewer than the 562 needed for a five-point gap at $d = 0.17$.

| metric | score range | median $d$ | MDD (pp) | 5–95% band |
|---|---|---|---|---|
| exact_match | 1.7–35.5 | 0.158 | 6.06 | 4.76–8.44 |
| quasi_exact_match | 2.3–57.2 | 0.177 | 6.43 | 5.14–9.87 |
| $F_1 \geq 0.5$ | 5.1–87.6 | 0.197 | 6.78 | 5.32–10.80 |

| $F_1$ binarized at | median $d$ | MDD (pp) |
|---|---|---|
| $F_1 \geq 0.3$ | 0.149 | 5.89 |
| $F_1 \geq 0.5$ | 0.197 | 6.78 |
| $F_1 \geq 0.7$ | 0.234 | 7.39 |

Table 21: **Sensitivity to metric choice and binarization threshold on NarrativeQA.** HELM's lite suite provides `exact_match`, `quasi_exact_match`, and continuous $F_1$ outcomes for the same 91 models and 355 items. Every reconstructed mean $F_1$ matches the published score to $10^{-4}$; no model is excluded. Upper: the two binary metrics differ on 17.6% of model–item cells. Using threshold 0.5 to binarize $F_1$, limits range from 6.06 to 6.78 accuracy points across metrics, a 12% range. Lower: continuous $F_1$ requires a threshold before McNemar's test can be applied, and HELM does not specify such a threshold. Changing it from 0.3 to 0.7 changes the limit from 5.89 to 7.39 accuracy points, a 25% range, approximately twice the metric-choice effect. In this analysis, increasing the threshold lowers scores and increases disagreement; this monotone pattern is empirical, rather than a general property of thresholding. Both ranges are smaller than the changes of up to 3× in Appendix Table 9. The present analysis uses natural differences among metrics on ranked model outputs, whereas Appendix Table 9 uses controlled boundary and label defects, so the analyses are not pooled. MDD is in accuracy points (pp).

| comparison | $d$ used | MDD (pp) | agree-ment | false-res. | false-unres. | vs. within |
|---|---|---|---|---|---|---|
| HELM, within | 0.255 | 6.38 | 88.8% | 0.00% | 11.2% | — |
| SWE-bench, within | 0.268 | 6.63 | 91.4% | 0.00% | 8.6% | — |
| HELM→SWE, borrowed | 0.255 | 6.47 | 92.1% | 0.00% | 7.9% | 0.92× |
| SWE→HELM, borrowed | 0.268 | 6.54 | 87.6% | 0.00% | 12.4% | 1.10× |

| comparison | $d$ used | MDD (pp) | agree-ment | false-res. | false-unres. | vs. within |
|---|---|---|---|---|---|---|
| HellaSwag, within | 0.121 | 0.98 | 96.2% | 0.00% | 3.8% | — |
| MMLU, within | 0.316 | 1.34 | 95.2% | 0.00% | 4.8% | — |
| HellaSwag→MMLU, borrowed | 0.121 | 0.83 | 98.8% | 0.13% | 1.1% | 0.25× |
| MMLU→HellaSwag, borrowed | 0.316 | 1.58 | 90.9% | 0.00% | 9.1% | 2.41× |

Table 22: **Effect of transferring a disagreement rate between boards.** "Within" applies a board's measured median $d$ to its own pairs; "borrowed" applies the source board's median to the target's pairs. Both threshold classifications are compared with exact tests on the target outcomes. "False-res." and "false-unres." denote the two directions of disagreement with those tests, rather than type-I and type-II error probabilities of the power threshold. Agreement and the two disagreement fractions sum to one; "vs. within" is the false-unresolvable fraction relative to the within-board value. Upper: HELM MMLU and SWE-bench have median rates 0.013 apart (0.255 and 0.268). Transferring the rate changes the error by approximately a tenth of its within-board value in either direction. No pair is classified as above the limit while failing the exact test in any of these four rows, including the within-board comparisons. This near-matched pair of rates provides only a limited test of transfer across substantial mismatch. Lower: Open LLM v1 HellaSwag and MMLU have rates 2.6× apart (0.121 and 0.316), with 111 jointly verified models and 6,105 pairs per board. Transferring the higher rate to HellaSwag raises the threshold and increases false-unresolvable calls 2.4×, from 3.8% to 9.1%. Transferring the lower rate to MMLU lowers the threshold and yields 8 of 6,105 pairs above the limit that do not reject in the exact test. Overall agreement nonetheless increases to 98.8%, because fewer significant comparisons fall below the threshold. The relation MDD $\propto \sqrt{d}$ explains the direction of threshold movement, but does not make a power threshold equivalent to a significance test. Open LLM v1 MMLU publishes a macro-average over 57 subjects. A micro-average differs by up to 1.99 accuracy points and would exclude 74 of the 111 verified models under the 0.6-accuracy-point reconstruction tolerance. MDD is in accuracy points (pp).

| $d$ label | panel $n$ | limit (pp) | resolvable % [95%] | unresolvable % [95%] | separating |
|---|---|---|---|---|---|
| 0.316 | 514 | 7.09 | 0.10 [0.04–0.20] | 20.6 [17.9–23.6] | 300/300 |
| 0.316 | 1,000 | 5.06 | 0.16 [0.06–0.33] | 20.6 [18.0–23.6] | 300/300 |
| 0.316 | 2,000 | 3.56 | 0.04 [0.00–0.12] | 20.3 [17.7–23.0] | 300/300 |
| 0.316 | 4,000 | 2.51 | 0.01 [0.00–0.05] | 20.7 [17.7–23.9] | 300/300 |
| 0.255 | 514 | 6.38 | 0.17 [0.07–0.32] | 22.1 [19.5–25.0] | 300/300 |
| 0.255 | 1,000 | 4.55 | 0.18 [0.07–0.34] | 21.9 [19.2–24.9] | 300/300 |
| 0.255 | 2,000 | 3.21 | 0.09 [0.02–0.23] | 22.3 [19.6–25.0] | 300/300 |
| 0.255 | 4,000 | 2.26 | 0.04 [0.00–0.12] | 21.8 [18.7–24.8] | 300/300 |

Table 23: **Replication of reversal-rate separation using a second evaluation framework.** The main-text 300-panel analysis is repeated on Open LLM v1 MMLU: 111 verified models, 13,937 shared items, and 57 subject shards, using the published macro-average score. Random panels of whole subjects are sampled to the target size. Each pair is labeled using the closed-form limit at the realized panel size and the rate in the first column, then its ordering is compared with the full board. Rows use either the board's measured rate (0.316) or HELM's (0.255); limits are in accuracy points (pp). Point estimates and intervals use the same pair-level estimand: each pair's mean reversal rate is calculated on panels assigning its majority label, and these rates are averaged over pairs. Pooling all observations instead gives reversal rates of 0.40–0.84% above the limit and 19.7–21.4% below it. The pooled above-limit rate falls outside the pair-level interval in all eight rows. This difference reflects selection of majority-label observations: pairs usually below the limit do not contribute their less frequent above-limit observations, among which reversals are concentrated. Intervals use a dyadic model bootstrap and retain resampling multiplicities when weighting pairs. Omitting those multiplicities produces an approximately 63% subset instead of an $n$-out-of-$n$ resample and yields narrower intervals in all sixteen cells. "Separating" counts panels with more reversals below than above the limit. Because the threshold is constant within each panel, selecting $|g| \geq L$ is equivalent to selecting the largest gaps at the same selectivity. We therefore also compare with a constant threshold of 6.04 accuracy points, matched to total selectivity across panel sizes. That rule yields a 37.2-fold reversal-rate ratio versus 31.1-fold for MDD at $n{\sim}514$, and matches or exceeds MDD in all eight rows. The replication supports an association between gap size and ranking stability, without establishing an advantage of MDD's scaling as $\sqrt{d/n}$. MDD additionally specifies a threshold location from the assumed rate and sample size. The analysis script prespecifies, without external registration, a failure criterion: overlap between the clustered reversal-rate intervals at $n{\sim}514$ under the board's own rate. Neither bootstrap meets that criterion; intervals are disjoint and all 300 of 300 panels separate at every size. This criterion is permissive given the approximately two-order separation observed in the initial analysis: overlapping intervals would require closing a 90-fold ratio between their nearest bounds. A stricter post-hoc criterion, a below-limit reversal rate at least fivefold the above-limit rate with disjoint intervals, is also satisfied (observed bound ratio: 90-fold). Below-limit rates remain near 20% across panel sizes, compared with approximately 18.5% on HELM. These rates depend on the mixture of gaps within each board and are not determined by $(n, d)$ alone. SWE-bench Verified has no larger scored counterpart, so its random-split analysis assesses within-board resampling rather than transfer to a larger item population.

| rate used | power | agree | false-res. | false-unres. |
|---|---|---|---|---|
| each pair's own $d$ | 50% | 65 | 0 | 0 |
| each pair's own $d$ | 80% | 59 | 0 | 6 |
| each board's own median $d$ | 80% | 56 | 0 | 10 |
| borrowed 0.255 (HELM MMLU) | 80% | 56 | 0 | 10 |
| borrowed 0.121 (lowest measured) | 80% | 59 | 0 | 7 |
| borrowed 0.352 (highest measured) | 80% | 54 | 0 | 12 |
| borrowed 0.05 (vision, MTEB) | 80% | 59 | 5 | 2 |
| borrowed 0.05 (vision, MTEB) | 50% | 58 | 8 | 0 |

| board | advance | gain | exact $P$ | own limit | cause |
|---|---|---|---|---|---|
| SWE-bench | sweagent_gpt4 | 6.60 | $1.8 \times 10^{-4}$ | 6.63 | own rate |
| SWE-bench | MASAI_gpt4o | 7.00 | $9.2 \times 10^{-4}$ | 6.63 | borrowing |
| SWE-bench | gru | 4.60 | $4.3 \times 10^{-2}$ | 6.63 | own rate |
| OLv1 GSM8K | Meta-Llama-3-70B-Instruct | 3.41 | $3.1 \times 10^{-3}$ | 4.44 | own rate |
| OLv1 HellaSwag | luxia-21.4b-alignment-v1.0 | 1.24 | $3.2 \times 10^{-6}$ | 1.02 | borrowing |

Table 24: **Sources of disagreement between power thresholds and exact tests on published record claims.** The analysis includes 66 advances with exact-test outcomes: 32 SWE-bench records and all eligible Open LLM v1 records on five boards. The exact test rejects for 25 claims. Upper: with pair-specific disagreement rates and a 50% power target, the threshold agrees with all 65 evaluable exact-test outcomes in this collection. This is an observed agreement, not a general identity between the approximate power boundary and a discrete exact test. Raising the target to 80% places 6 significant advances below the limit; the normal approximation increases the threshold by 1.43× relative to the 50%-power boundary. Replacing pair-specific rates with board medians increases this count from 6 to 10. Agreement falls by three rather than four because the denominator also changes: one claim lacks enough discordant items for a feasible pair-rate limit, so the pair-rate rows contain 65 claims and board-rate rows contain all 66. Applying HELM's 0.255 to every board gives the same total disagreement count as using board-specific medians, although individual classifications change. Across borrowed rates in $[0.121, 0.352]$, no pair classified above the limit fails the exact test in this collection. At 0.05, the optimistic rate used for vision and MTEB, five do; at matched 50% power, eight do. Since MDD $\propto \sqrt{d}$, higher assumed rates can only raise the threshold and lower assumed rates can only reduce it. Lower: the 5 advances whose classification changes with the borrowed rate. The 2 labeled *borrowing* exceed their own board's limit but fall below a transferred limit. The 3 labeled *own rate* already fall below their own board's limit and are included because another assumed rate changes their label. Of the 66 claims, 54 agree with the exact-test outcome throughout the stated range, 5 depend on the rate, and 7 disagree throughout. Gain and own limit are in accuracy points.

| fitted on | applied to | within | cross | ratio 95% | $r$ | recal. | disjoint | MDD err |
|---|---|---|---|---|---|---|---|---|
| HellaSwag | MMLU | 0.952 | 0.665 | [0.46, 0.86] | 0.971 | 0.944 | 0.644 | 12.5 |
| HellaSwag | SWE-bench | 0.972 | 0.686 | [0.48, 0.85] | 0.934 | 0.872 | 0.664 | 5.3 |
| MMLU | HellaSwag | 0.983 | 0.928 | [0.87, 0.98] | 0.990 | 0.979 | 0.918 | 13.3 |
| MMLU | SWE-bench | 0.972 | 0.103 | [-0.19, 0.32] | 0.928 | 0.861 | 0.054 | 15.0 |
| SWE-bench | HellaSwag | 0.983 | 0.936 | [0.91, 0.97] | 0.987 | 0.974 | 0.927 | 11.8 |
| SWE-bench | MMLU | 0.952 | 0.487 | [0.37, 0.62] | 0.948 | 0.899 | 0.458 | 13.3 |

Table 25: **Transfer of disagreement-rate regression across three boards.** Pairwise disagreement $d$ is regressed on published accuracies, extending the agreement regression of Card et al. (2020) with two terms derived from accuracy-based Fréchet bounds. Source coefficients are then applied to target-board pairs. "Within" is the target's fitted $R^2$; "cross" is the transferred $R^2$. "Ratio 95%" gives a dyadic model-bootstrap interval for cross divided by within, with coefficients refitted in every replicate. Pearson $r$ measures association between predicted and measured disagreement; "recal." is $R^2$ after an affine recalibration, and "disjoint" also excludes model overlap between source and target. "MDD err" is the median relative limit error, $|\sqrt{d'/d} - 1|$, expressed as a percentage, using MDD $\propto \sqrt{d}$. For a single predictor, affine OLS recalibration gives $R^2 = r^2$; this identity holds to $10^{-12}$ across all 42 transfers in the seven-board extension. In the three-board analysis, the worst raw transferred $R^2$ is 0.103, producing the prespecified not-transportable label. After affine recalibration, the smallest retained ratio is $0.861/0.972 = 0.886$; after an intercept-only adjustment it is 0.769. Both exceed the transportability threshold. The table's "recal." column reports raw refitted $R^2$, whose minimum is 0.861, rather than its ratio to the target fit. In the seven-board extension used in the main text, the smallest retained ratios after affine and slope-only recalibration are 0.231 and −0.511, respectively. Neither adjustment brings every transfer above the threshold.

*Interpretation of transfer summaries.* The prespecified statistic is the minimum across directed transfers and cannot increase when more boards are added. For the worst three-board transfer, the clustered interval $[-0.195, 0.315]$ crosses the 0.25 failure threshold. Among the six transfers, one point estimate is below 0.25 and 2 retain at least 0.75. Raw transferred $R^2$ has minimum 0.103 and median 0.6755; the smallest retained ratio is 0.106. The same transfer retains 0.106 in the seven-board analysis because both use the same 111 verified systems. Across all 42 transfers, the smallest retained ratio is −24.664 and the median is 0.704 (Appendix Table 26). These ratios divide by the target's fitted $R^2$ and should not be compared directly with raw $R^2$. The original design specifies three boards with sufficient per-item data; the seven-board analysis is an extension. On the original boards, association transfers more consistently than absolute scale. The extension also contains weakly associated transfers, particularly from Winogrande. The evaluated anchor-based recalibrations do not achieve reliable performance throughout the seven-board collection when scored on held-out pairs (Appendix Table 32). *Pooled and transferred regression.* Card et al. (2020) report $R^2 = 0.966$ for agreement regression fitted to pooled datasets. Pooling our three boards yields $R^2 = 0.900$ over 21,121 pairs, with board-specific values of 0.756–0.977. In contrast, the worst single-board transfer has a retained ratio of 0.106. The pooled fit uses observations from every target distribution during coefficient estimation; the transferred fit does not. Appendix Table 27 separately evaluates Card et al.'s published coefficients on all seven boards.

| source | target | within | $R^2_{\text{raw}}$ | ratio | Pearson $r$ | ratio (gap) | MDD err % |
|---|---|---|---|---|---|---|---|
| HELM MMLU | OLv1 ARC | 0.908 | 0.657 | 0.723 | 0.946 | 0.826 | 12.6 |
| HELM MMLU | OLv1 GSM8K | 0.979 | 0.811 | 0.828 | 0.964 | 0.912 | 9.4 |
| HELM MMLU | OLv1 HellaSwag | 0.983 | 0.956 | 0.972 | 0.987 | 0.634 | 7.0 |
| HELM MMLU | OLv1 MMLU | 0.952 | 0.942 | 0.989 | 0.972 | 0.829 | 2.2 |
| HELM MMLU | OLv1 Winogrande | 0.947 | 0.581 | 0.614 | 0.952 | 0.700 | 12.8 |
| HELM MMLU | SWE-bench | 0.972 | 0.078 | 0.080 | 0.914 | 0.669 | 14.1 |
| OLv1 ARC | HELM MMLU | 0.935 | 0.647 | 0.692 | 0.962 | 0.714 | 10.7 |
| OLv1 ARC | OLv1 GSM8K | 0.979 | 0.747 | 0.762 | 0.974 | 0.929 | 18.2 |
| OLv1 ARC | OLv1 HellaSwag | 0.983 | 0.960 | 0.977 | 0.990 | 0.845 | 10.3 |
| OLv1 ARC | OLv1 MMLU | 0.952 | 0.728 | 0.764 | 0.972 | 0.604 | 10.6 |
| OLv1 ARC | OLv1 Winogrande | 0.947 | 0.005 | 0.006 | 0.938 | 0.550 | 24.9 |
| OLv1 ARC | SWE-bench | 0.972 | 0.775 | 0.797 | 0.959 | 0.874 | 5.7 |
| OLv1 GSM8K | HELM MMLU | 0.935 | 0.706 | 0.755 | 0.942 | 0.821 | 9.3 |
| OLv1 GSM8K | OLv1 ARC | 0.908 | 0.335 | 0.369 | 0.911 | 0.901 | 20.1 |
| OLv1 GSM8K | OLv1 HellaSwag | 0.983 | 0.870 | 0.886 | 0.985 | 0.744 | 22.1 |
| OLv1 GSM8K | OLv1 MMLU | 0.952 | 0.763 | 0.801 | 0.905 | 0.693 | 6.7 |
| OLv1 GSM8K | OLv1 Winogrande | 0.947 | 0.838 | 0.885 | 0.963 | 0.630 | 5.1 |
| OLv1 GSM8K | SWE-bench | 0.972 | 0.389 | 0.401 | 0.950 | 0.877 | 16.7 |
| OLv1 HellaSwag | HELM MMLU | 0.935 | 0.532 | 0.569 | 0.955 | -0.247 | 12.6 |
| OLv1 HellaSwag | OLv1 ARC | 0.908 | 0.898 | 0.989 | 0.952 | 0.641 | 4.7 |
| OLv1 HellaSwag | OLv1 GSM8K | 0.979 | 0.767 | 0.784 | 0.961 | 0.761 | 14.8 |
| OLv1 HellaSwag | OLv1 MMLU | 0.952 | 0.665 | 0.698 | 0.971 | -0.112 | 12.5 |
| OLv1 HellaSwag | OLv1 Winogrande | 0.947 | 0.078 | 0.083 | 0.906 | -0.236 | 20.7 |
| OLv1 HellaSwag | SWE-bench | 0.972 | 0.686 | 0.706 | 0.934 | 0.777 | 5.3 |
| OLv1 MMLU | HELM MMLU | 0.935 | 0.925 | 0.990 | 0.965 | 0.760 | 3.4 |
| OLv1 MMLU | OLv1 ARC | 0.908 | 0.637 | 0.702 | 0.950 | 0.548 | 13.1 |
| OLv1 MMLU | OLv1 GSM8K | 0.979 | 0.857 | 0.875 | 0.964 | 0.782 | 7.8 |
| OLv1 MMLU | OLv1 HellaSwag | 0.983 | 0.928 | 0.944 | 0.990 | 0.372 | 13.3 |
| OLv1 MMLU | OLv1 Winogrande | 0.947 | 0.646 | 0.682 | 0.943 | 0.660 | 10.7 |
| OLv1 MMLU | SWE-bench | 0.972 | 0.103 | 0.106 | 0.928 | 0.240 | 15.0 |
| OLv1 Winogrande | HELM MMLU | 0.935 | -2.128 | -2.276 | 0.857 | 0.822 | 20.5 |
| OLv1 Winogrande | OLv1 ARC | 0.908 | -7.538 | -8.298 | 0.707 | 0.696 | 39.9 |
| OLv1 Winogrande | OLv1 GSM8K | 0.979 | -2.939 | -3.001 | 0.772 | 0.729 | 7.4 |
| OLv1 Winogrande | OLv1 HellaSwag | 0.983 | 0.159 | 0.161 | 0.864 | 0.522 | 21.6 |
| OLv1 Winogrande | OLv1 MMLU | 0.952 | -6.379 | -6.697 | 0.736 | 0.857 | 26.1 |
| OLv1 Winogrande | SWE-bench | 0.972 | -23.978 | -24.664 | 0.473 | 0.264 | 41.1 |
| SWE-bench | HELM MMLU | 0.935 | 0.569 | 0.609 | 0.954 | 0.604 | 10.7 |
| SWE-bench | OLv1 ARC | 0.908 | 0.840 | 0.925 | 0.943 | 0.890 | 4.2 |
| SWE-bench | OLv1 GSM8K | 0.979 | 0.771 | 0.787 | 0.985 | 0.914 | 17.3 |
| SWE-bench | OLv1 HellaSwag | 0.983 | 0.936 | 0.952 | 0.987 | 0.852 | 11.8 |
| SWE-bench | OLv1 MMLU | 0.952 | 0.487 | 0.512 | 0.948 | 0.430 | 13.3 |
| SWE-bench | OLv1 Winogrande | 0.947 | -0.042 | -0.044 | 0.948 | 0.444 | 24.3 |

Table 26: **All 42 directed disagreement-regression transfers among seven boards.** This analysis extends the original three-board design, whose six transfers appear in Appendix Table 25. "Within" is the target board's in-sample fitted $R^2$; "ratio" is $R^2_{\text{raw}}$ divided by within. Inflation of the denominator would lower the retained ratio. Under 5-fold cross-validation, the corresponding target fits are 0.9348, 0.9083, 0.9785, 0.9827, 0.9524, 0.9468, and 0.9722. The largest in-sample increment is +0.0009 on GSM8K's 210 pairs, and no retained ratio changes by more than 0.1%. A negative ratio indicates predictions worse than the target's mean. "Ratio (gap)" applies the same transfer procedure to a two-parameter regression on $|a_i - a_j|$. Pearson $r$ ranges from 0.473 to 0.990; retained ratios range from −24.664 to 0.990, with median 0.704. Of 42 transfers, 19 retain at least 0.75 and 11 retain at most 0.25. The latter satisfy the failure threshold used by the minimum-based rule. Nine of these eleven involve Winogrande, as do all four transfers with $r < 0.85$; the other two, 0.080 and 0.106, target SWE-bench. Entries are point estimates; Appendix Table 30 evaluates system-level uncertainty. The minimum −24.664 is a retained ratio across 42 transfers. In the original three-board analysis (Appendix Table 25), the minimum raw $R^2$ is 0.103 and the minimum ratio is 0.106.

**(a) All pairs**

| board | $R^2$ | retained | $r$ | err. $d$ (%) | err. limit (%) |
|---|---|---|---|---|---|
| Open LLM v1 MMLU | 0.857 | 0.900 | 0.974 | +10.2 | +5.0 |
| Open LLM v1 HellaSwag | 0.850 | 0.865 | 0.988 | +39.5 | +18.1 |
| Open LLM v1 Winogrande | 0.815 | 0.861 | 0.952 | -10.9 | -5.6 |
| HELM MMLU | 0.804 | 0.860 | 0.966 | +13.0 | +6.3 |
| Open LLM v1 GSM8K | 0.756 | 0.772 | 0.961 | +1.2 | +0.6 |
| Open LLM v1 ARC | 0.208 | 0.229 | 0.946 | +43.3 | +19.7 |
| SWE-bench Verified | -0.736 | -0.757 | 0.908 | +47.0 | +21.2 |

**(b) Card's band and feasibility**

| board | in-band share (%) | in-band err. $d$ (%) | infeasible (%) | reading |
|---|---|---|---|---|
| Open LLM v1 MMLU | 1.3 | +5.5 | 0.0 | carries |
| Open LLM v1 HellaSwag | 67.1 | +47.0 | 0.0 | carries |
| Open LLM v1 Winogrande | 59.9 | -11.3 | 0.0 | carries |
| HELM MMLU | 19.0 | +10.4 | 0.0 | carries |
| Open LLM v1 GSM8K | 43.3 | -11.8 | 13.8 | carries |
| Open LLM v1 ARC | 12.8 | +45.8 | 0.0 | fails |
| SWE-bench Verified | 9.7 | +41.8 | 10.7 | fails |

Table 27: **External application of Card et al.'s published agreement regression.** Card et al. (2020) model agreement, corresponding to one minus a disagreement rate, as $0.4142 + 0.5819\,\text{min_acc} - 0.4662\,\text{acc_diff}$. They fit this expression to ten high-performing models pooled across GLUE tasks (their Table 5, $R^2 = 0.966$). We apply the published coefficients without modification. "Retained" is the resulting board-level $R^2$ divided by the $R^2$ of a regression fitted on that board, as in Appendix Table 26. "Err. $d$" is median signed relative rate error; "err. limit" is the corresponding relative limit error under the $\sqrt{d}$ scaling. The prespecified thresholds are $\geq 0.75$ for transfer and $\leq 0.25$ for failure. Five boards exceed the transfer threshold, two meet the failure threshold, and none are intermediate. Pearson $r$ remains 0.908–0.988, whereas retained fit ranges from 0.900 to −0.757, indicating that association can persist without accurate calibration. Every system resample retains at least one success and one failure (Appendix Table 30). Card et al. fit GLUE and SQuAD separately and restrict their analysis to high-performing systems; they do not claim universal transfer. *Accuracy range.* Their fitted data begin near 65% accuracy. We therefore label pairs with lower-system accuracy at least 65% as in band. Within this band, signed rate errors range from −11.8% to +47.0%, compared with −10.9% to +47.0% overall. Restricting to the fitted accuracy range does not eliminate the observed failures. "Infeasible" denotes predicted rates outside the Fréchet interval $[|a_i - a_j|, \min(a_i + a_j, 2 - a_i - a_j)]$. *Different agreement targets.* Card et al. lack GLUE gold test labels and measure agreement between predictions; we measure agreement in correctness. With more than two classes, two incorrect predictions can differ, so prediction disagreement can exceed correctness disagreement. This distinction can contribute to upward rate error when their regression is interpreted as a correctness model. All but one reported limit errors are positive. The exception is binary Winogrande, where this particular discrepancy cannot arise and the error is −5.6%. These observations do not isolate the target-definition difference from other transfer effects. *Additional predictors.* Card's three-parameter form is nested within our five-parameter model. Refitting both on each source yields median retained fits of 0.734 and 0.704, respectively; our model performs better on 12 of 42 transfers (interval 10–24). The difference in medians is −0.030, with interval −0.114 to +0.034. The additional accuracy-product and feasibility-bound terms therefore do not show a detectable aggregate transfer gain. Feasibility bounds also enter Card et al.'s analysis through Lachenbruch's midpoint prior. Card's form exceeds the gap rule on 23 of 42 transfers (interval 17–26), including 20 of 30 with option-answer targets.

**(a) Reversal rates for unresolvable pairs and ratios**

| design | systems | accuracy reverse (%) | 2PL reverse (%) | $\rho$ [95%] | all-pair ratio [95%] |
|---|---|---|---|---|---|
| HELM MMLU | 91 | 17.82 | 14.13 | 0.79 [0.71, 0.88] | 0.79 [0.70, 0.88] |
| Open LLM v1 GSM8K | 21 | 18.55 | 17.79 | 0.96 [0.54, 1.53] | 0.96 [0.54, 1.51] |
| Open LLM v1 HellaSwag | 111 | 18.34 | 13.39 | 0.73 [0.54, 0.97] | 0.84 [0.63, 1.10] |
| Open LLM v1 MMLU | 111 | 23.26 | 7.63 | 0.33 [0.22, 0.48] | 0.48 [0.33, 0.69] |
| SWE-bench Verified | 134 | 17.87 | 16.71 | 0.94 [0.86, 1.02] | 0.94 [0.86, 1.02] |
| MMLU panel → other subjects (30 random) | 111 | 24.77 | 16.95 | 0.68 [0.56, 0.83] | 0.76 [0.62, 0.92] |
| HELM's five subjects → other 52 | 111 | 25.41 | 19.38 | 0.76 [0.56, 1.00] | 0.76 [0.56, 1.00] |

**(b) Alternative ordering and system-pool checks**

| design | $\rho$ under macro | $\rho$ under subj. 2PL | Kendall $\tau_K$ accuracy | Kendall $\tau_K$ 2PL | pool shift |
|---|---|---|---|---|---|
| HELM MMLU | 1.00 | 1.05 | 0.835 | 0.866 | +0.039 |
| Open LLM v1 GSM8K | — | — | 0.925 | 0.926 | -0.069 |
| Open LLM v1 HellaSwag | — | — | 0.967 | 0.973 | +0.046 |
| Open LLM v1 MMLU | 1.14 | 0.91 | 0.933 | 0.968 | +0.046 |
| SWE-bench Verified | — | — | 0.910 | 0.913 | +0.003 |
| MMLU panel → other subjects (30 random) | 1.04 | — | 0.843 | 0.881 | — |
| HELM's five subjects → other 52 | 0.96 | — | 0.802 | 0.849 | — |

Table 28: **Ranking by item-response ability while retaining accuracy-based power labels.** Each pair retains the label assigned to its accuracy gap on the labeling items; only the score used to order the systems changes. The ratio $\rho$ compares the below-limit reversal rate under a two-parameter logistic (2PL) item-response ordering with that under accuracy. $\tau_K$ is Kendall correlation between rankings from the two item sets and is distinct from target power $\tau$. The all-pair ratio uses every pair, including those above the accuracy threshold; its reduction is smaller because some above-limit pairs also reverse under 2PL. The first five rows use 30 split-half comparisons per board, stratified by subject when applicable. The last two label Open LLM v1 MMLU pairs on a whole-subject panel and evaluate rankings on the remaining subjects. Intervals jointly resample splits and systems. "Macro" ranks by mean subject accuracy; "subj. 2PL" uses one discrimination parameter per subject. Neither reproduces the reduction observed with item-specific discrimination, supporting the importance of that modeling choice within these analyses. "Pool shift" is the change in $\rho$ when item parameters are fitted on a random half of systems and then held fixed for the remainder. The Rasch control reproduces the accuracy rankings in every design. All fits converge. Joint maximum a posteriori estimation uses a prior standard deviation of 0.5 for log-discrimination; values of 0.25 and 1.0 yield the rule labels *carries* and *ambiguous*, respectively. This estimator is simpler than the hierarchical model of Rodriguez et al. (2021), so the results do not evaluate that model. Two independent runs initialized from scratch on different machines agree on every label and to within 0.004 on all values in this table and the next.

**(a) Unresolvable fractions and label agreement**

| design | accuracy unresolvable (%) | 2PL unresolvable (%) | $\phi$ [95%] | agrees (%) |
|---|---|---|---|---|
| HELM MMLU | 40.7 | 33.8 | 0.83 [0.80, 0.87] | 95.8 |
| Open LLM v1 GSM8K | 18.4 | 16.8 | 0.91 [0.72, 1.05] | 98.3 |
| Open LLM v1 HellaSwag | 8.4 | 7.5 | 0.88 [0.81, 0.96] | 99.2 |
| Open LLM v1 MMLU | 13.7 | 7.8 | 0.57 [0.47, 0.71] | 99.0 |
| SWE-bench Verified | 22.0 | 21.0 | 0.95 [0.93, 0.98] | 97.7 |
| MMLU panel → other subjects (30 random) | 28.5 | 22.4 | 0.79 [0.70, 0.88] | 98.0 |
| HELM's five subjects → other 52 | 36.9 | 27.1 | 0.73 [0.66, 0.82] | 97.5 |

**(b) Reversals relative to each score's own limit**

| design | accuracy reverse below / above (%) | 2PL reverse below / above (%) | accuracy reversals below limit (%) | 2PL reversals below limit (%) |
|---|---|---|---|---|
| HELM MMLU | 18.7 / 0.34 | 17.9 / 0.21 | 97.4 | 97.8 |
| Open LLM v1 GSM8K | 18.8 / 0.12 | 20.0 / 0.06 | 97.3 | 98.6 |
| Open LLM v1 HellaSwag | 18.6 / 0.03 | 17.5 / 0.04 | 98.1 | 97.1 |
| Open LLM v1 MMLU | 23.6 / 0.10 | 19.8 / 0.05 | 97.5 | 96.9 |
| SWE-bench Verified | 18.6 / 0.14 | 18.3 / 0.12 | 97.4 | 97.6 |
| MMLU panel → other subjects (30 random) | 25.3 / 0.73 | 22.6 / 1.02 | 93.2 | 86.5 |
| HELM's five subjects → other 52 | 26.2 / 0.08 | 26.4 / 0.36 | 99.5 | 96.4 |

Table 29: **Detection thresholds matched to the ranking score.** Appendix Table 28 changes the ordering score while retaining accuracy labels. Here each score has its own threshold: a pair is above the limit when its gap is at least $z_{0.975} + z_{0.80} = 2.80$ paired standard errors. Standard errors are estimated from 40 item resamples of the labeling set, refitting 2PL each time. "Agrees" is agreement between this construction for accuracy and the exact-power label. The ratio $\phi$ compares the below-limit fraction under 2PL with that under accuracy. The prespecified criterion requires at least 95% of 2PL reversals to occur below its own limit and at most 0.5% of above-limit pairs to reverse. All five split-half boards satisfy both conditions. The two panel designs compare different subjects, introducing a coverage shift beyond the sampling uncertainty represented by these standard errors; the thresholds account for a smaller fraction of reversals under that shift.

| quantity or statement | point | all verified systems | one per organisation | reading |
|---|---|---|---|---|
| median retained fit | 0.704 | 0.563 to 0.773 | 0.696 to 0.733 | |
| smallest retained fit | -24.664 | -701.612 to -0.962 | -229.386 to -18.417 | |
| median Pearson $r$ | 0.951 | 0.938 to 0.959 | 0.950 to 0.954 | |
| smallest Pearson $r$ | 0.473 | 0.206 to 0.859 | 0.309 to 0.513 | |
| transfers keeping $\geq 0.75$ (of 42) | 19 | 12 to 23 | 18 to 20 | |
| transfers keeping $\leq 0.25$ (of 42) | 11 | 7 to 16 | 10 to 11 | |
| transfers the gap rule wins (of 42) | 21 | 17 to 24 | 20 to 22 | |
| map beats gap rule, option-answer targets (of 30) | 19 | 16 to 22 | 19 to 20 | |
| map beats gap rule, generated-answer targets (of 12) | 2 | 1 to 5 | 1 to 2 | |
| median cost of borrowing, % of the limit | 12.6 | 11.0 to 13.6 | 12.1 to 12.8 | |
| median cost of borrowing, % of the item requirement | 23.8 | 21.3 to 27.0 | 23.2 to 24.4 | |
| median retained fit, full map minus gap rule | -0.003 | -0.105 to +0.077 | -0.015 to +0.021 | not separable |
| median retained fit, full map minus Card et al.'s form | -0.030 | -0.114 to +0.034 | -0.055 to -0.009 | not separable |
| scale is not reliably kept: smallest retained fit $\leq 0.25$ | holds | 100.0% | 100.0% | confirmed |
| shape is kept on every transfer: smallest $r \geq 0.85$ | fails | 3.5% | 0.0% | unresolved |
| most transfers carry: a majority, 22 of 42, keep $\geq 0.75$ | fails | 7.9% | 0.0% | unresolved |
| the map's advantage over the gap rule is larger on option-answer targets | holds | 100.0% | 100.0% | confirmed |
| Card et al.'s published line carries on one board and fails on another | holds | 100.0% | 100.0% | confirmed |

Table 30: **Uncertainty in transport summaries under system resampling.** The point estimates summarize the 42 transfers in Appendix Table 26. "All verified systems" gives 95% intervals from 1,000 organization-cluster resamples. Within each board, organizations are sampled with replacement and all their systems enter jointly. Pair weights are products of system multiplicities, and every source and target regression is refitted in each replicate. HELM and SWE-bench lack organization identifiers, so each system forms its own cluster. All boards are resampled; the smallest is GSM8K with 21 systems in 16 organizations. "One per organization" gives the 2.5–97.5 percentile range across 200 draws retaining one random system per uploading organization on Open LLM v1 boards. For statements, the entries give the proportion of resamples satisfying the statement. The table labels a statement confirmed when at least 95% satisfy it under both procedures, and unresolved otherwise. A difference is separable when both ranges exclude zero. Statements and decision thresholds are specified in `src/mx35_transport_robustness.py`; unit weights reproduce the tabulated ratios to $10^{-14}$. Both unresolved statements also fail at their point estimates. Four transfers, all from Winogrande, have $r < 0.85$, although the median $r$ varies little. A strict majority requires 16 of 30 transfers in the originally stated rule, or 22 of 42 here; the observed count is 19. These resampling labels describe robustness within the sampled system populations, rather than guarantees for unobserved boards.

| design | items fitted | tested half seen | ratio to held-out | predicted |
|---|---|---|---|---|
| held-out (A) | $n/2$ | none | 1 | 1 |
| A + half of B | $3n/4$ | half | 0.713–0.723 | 0.667 |
| all items | $n$ | all | 0.514–0.516 | 0.500 |
| tested half alone | $n/2$ | all | 0.087–0.123 | 0 |

Table 31: **Effect of overlap between estimation and evaluation items.** Entries are mean absolute pairwise errors relative to the held-out design, over 200 item splits and four combinations of board and sample size. Rows vary how much of the evaluated half contributes to estimation. The first and last use the same number of items, $n/2$, while changing their identity, which separates sample size from evaluation overlap. For equal halves, $\hat{d}_{\text{full}} = (\hat{d}_{\text{half}} + \hat{d}_{\text{tested}})/2$ holds exactly. A linear approximation to error in the remaining contribution from unseen items gives the predicted ratios. "Ratio to held-out" reports the range of point estimates across four cells. For the full-item row, the corresponding 95% intervals over 200 splits are 0.487–0.544, 0.473–0.556, 0.478–0.552, and 0.475–0.557. Each contains 0.500 and excludes 0.707, the ratio expected from sample-size scaling alone. The $3n/4$ condition exceeds its prediction by 7–8% in all four cells. The tested-half-only condition estimates the same empirical distribution from which evaluation draws are sampled, so its low error does not provide an independent validation of transfer. Its residual includes Monte Carlo uncertainty from 200 draws per pair. The comparison supports evaluation overlap as the principal explanation for the approximately halved full-item error.

| anchors $k$ | median | 25th pct | 5th pct | % worse than mean | % $\geq 0.75$ |
|---|---|---|---|---|---|
| 1 | -0.334 | -0.844 | -2.501 | 73.1% | 0.0% |
| 2 | 0.728 | -0.221 | -30.329 | 27.8% | 48.2% |
| 3 | 0.855 | 0.579 | -1.708 | 12.6% | 64.6% |
| 5 | 0.905 | 0.770 | -0.080 | 5.6% | 76.7% |
| 10 | 0.934 | 0.852 | 0.235 | 2.4% | 84.4% |
| 20 | 0.944 | 0.879 | 0.419 | 0.9% | 86.6% |
| 50 | 0.948 | 0.891 | 0.515 | 0.1% | 87.6% |
| 100 | 0.952 | 0.894 | 0.537 | 0.0% | 88.7% |

Table 32: **Recalibration using a limited number of target-board pairs.** For each of 42 directed transfers, an affine recalibration of source predictions is fitted to $k$ randomly selected target pairs. Retained fit is evaluated on pairs excluded from recalibration, using 200 draws per transfer. At $k = 1$, the two-parameter fit is underdetermined; the table reports NumPy's minimum-norm least-squares solution. At $k = 2$, the fit interpolates the anchors. Evaluating on those same anchors would reward interpolation rather than measure transfer, so all scores use held-out pairs. "% worse than mean" is the fraction of draws with negative $R^2$, indicating worse predictions than the target mean. One anchor gives a negative median retained ratio and no draw reaches 0.75. The median exceeds 0.75 at $k = 3$, but the 5th percentile remains below that threshold throughout the evaluated range: approximately $-30$ at $k = 2$, 0.235 at ten anchors, and 0.537 at a hundred. Thus, no evaluated anchor count gives uniformly high retained fit across the transfer and anchor-sampling distribution.

# D Theory of the detection limit

## D.1 The probability model and the estimand

Two systems $A$ and $B$ are scored on the same $n$ items. For item $i$ let $X_i, Y_i \in \{0, 1\}$ record whether $A$, respectively $B$, is correct. Write

$$B_n = \#\{i : X_i = 1, Y_i = 0\}, \qquad C_n = \#\{i : X_i = 0, Y_i = 1\}, \qquad M_n = B_n + C_n, \tag{1}$$

so the observed accuracy gap is $\widehat{\Delta} = (B_n - C_n)/n$ and the observed disagreement (discordance) rate is $\widehat{d} = M_n/n$. Throughout, $\mathrm{Bin}(m, \theta)$ has c.d.f. $F(\cdot\,; m, \theta)$; $\alpha \in (0, 1)$ is the level and $\tau \in (0, 1)$ the target power, with $\alpha = 0.05$ and 80% power ($\tau = 0.8$) in every number reported here. Write $z_p = \Phi^{-1}(p)$, $z_\alpha := z_{1-\alpha/2} = 1.959964$, $z_\tau := z_{0.8} = 0.841621$ and

$$Z := z_\alpha + z_\tau = 2.801585, \qquad Z^2 = 7.848880. \tag{2}$$

**Assumption 1** (Item superpopulation)**.** There is a probability distribution $Q$ on a space $\mathcal{Z}$ of items, and the $n$ benchmark items $Z_1, \ldots, Z_n$ are drawn i.i.d. from $Q$.

**Assumption 2** (Fixed, item-measurable systems)**.** Each system is a deterministic map $f : \mathcal{Z} \to \{0, 1\}$ giving its scored per-item outcome, fixed before the items are drawn and not a function of $Z_1, \ldots, Z_n$; thus $X_i = f_A(Z_i)$, $Y_i = f_B(Z_i)$.

**Assumption 3** (Independence across items)**.** The pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ are mutually independent.

**Assumption 4** (Common cell distribution)**.** There are $\beta, \gamma \geq 0$, $\beta + \gamma \leq 1$, with $\Pr(X_i{=}1, Y_i{=}0) = \beta$ and $\Pr(X_i{=}0, Y_i{=}1) = \gamma$ for every $i$.

**Assumption 5** (Binary, unweighted score)**.** The reported metric is the unweighted mean over items of a per-item $\{0, 1\}$ outcome. Metrics that are not of this form (F1, BLEU, MC2, average precision, AUC, macro-averages over unequal strata, and scores averaged over repeated evaluation runs) are outside the model. The survey's exclusion rules enforce this for metrics that are not per-item binary outcomes. Six surveyed boards nonetheless publish a macro- or run-average of binary outcomes; they are named in Methods, which measures the size and the sign of the resulting error on the one board whose per-item outcomes allow it.

**Assumption 6** ($d$ is a parameter)**.** Where a disagreement rate $d$ enters a detection limit it is the population rate of the pair under test, carried without sampling error. No uncertainty in $d$ is propagated anywhere in this appendix.

**Assumption 7** (Transported $d$)**.** Where one $d$ is applied to every pair on a board, or imported from a different benchmark, the pairs are assumed *degenerate* in $d$: $d_{xy} \equiv d$ almost surely. Exchangeability in $d$ is not enough, because $d \mapsto \mathrm{MDD}(n, d)$ is not affine.

**Assumption 8** (Prespecified, per-comparison design)**.** $n$, the item set, the scoring rule, $\alpha$ and $\tau$ are fixed before the outcomes are seen; there is no optional stopping and no outcome-dependent filtering of items. Every MDD is a *per-pair* quantity at level $\alpha$ and carries no multiplicity correction; simultaneous inference is §D.6.

Assumptions 1–2 imply 3–4, providing a sufficient sampling model for the analysis. Under 3–4 the vector $(B_n, C_n, n - M_n)$ is trinomial,

$$(B_n, C_n, n - M_n) \sim \mathrm{Multinomial}\Big(n; \tfrac{d+\Delta}{2}, \tfrac{d-\Delta}{2}, 1 - d\Big), \qquad \Delta := \beta - \gamma, \quad d := \beta + \gamma, \tag{3}$$

which is exactly the law sampled by `power()` in `src/testset_mdd.py`. The parameter space is

$$\mathcal{P} = \big\{(\Delta, d) : 0 \leq d \leq 1,\ |\Delta| \leq d\big\}. \tag{4}$$

For a fixed benchmark of $n$ items and deterministic systems evaluated by greedy decoding, $(X_i, Y_i)_{i \leq n}$ is a fixed table. The observed gap $\widehat{\Delta}$ is known exactly and has no sampling uncertainty conditional on that table. A detection limit therefore requires an explicit source of randomness. We distinguish three interpretations.

*(E1) Item resampling.* Assumptions 1–2. The estimand is the domain-level gap $\Delta = \mathrm{acc}_A(Q) - \mathrm{acc}_B(Q)$, and (3) is exact. This interpretation assumes i.i.d. sampling, whereas curated benchmark

items are deduplicated, difficulty-screened, and in cases such as SWE-bench Verified human-filtered from a larger pool, so i.i.d. sampling from $Q$ is an idealisation. In addition, $Q$ must be fixed independently of the systems, which pretraining contamination violates.

*(E2) Re-run randomness.* The items are fixed and the systems stochastic, $X_i \sim \text{Bernoulli}(\pi_i^A)$ independently across $i$ with item-specific rates; the estimand is $\Delta = n^{-1}\sum_i(\pi_i^A - \pi_i^B)$, the expected gap *on these items*. Assumption 3 is plausible here; Assumption 4 fails by construction, so $B_n$ is Poisson-binomial. By Lemma 5 the test of the sharp null remains exact; by Proposition 8 the power calculation does not transfer.

*(E3) Design-based randomisation.* No superpopulation is posited; the only randomness is the assignment of outcomes to systems within discordant pairs. The conditional test (5) is then a legitimate randomisation test of the sharp null "the two systems' outcomes are exchangeable within each discordant item", valid without 1 or 4. E3 admits no alternative indexed by a true gap, so power, and with it MDD, is undefined.

We adopt E1 to define generalisation from benchmark items to a specified item population. Where the evaluated items themselves constitute the population of interest, each reported MDD instead describes an exchangeable idealisation of the benchmark. Without E1, the exact $p$-values retain their E3 interpretation; Lemma 5 does not justify that interpretation (see Remark 2).

*Remark* 1 (Violations of 3 in both analyses). In the direct measurement analysis, predictions are pooled across cross-validation folds, so items in a common test fold share a fitted model: their outcomes are dependent, and the effective number of independent paired observations is below $n$, so the McNemar limits reported for digits, breast-cancer, ABIDE and ADNI are optimistic. This analysis also violates 2, since each item's prediction comes from a model fitted on $Z_{-i}$. In the leaderboard analysis, items cluster by MMLU subject, SWE-bench repository and MTEB document; 3 fails there too. The direction of the resulting error requires a dependence model: the conditional construction of Lemma 1 can fail, beyond any change in variance. Remark 9 quantifies the effect on the floor.

### D.2 The exact test

McNemar's exact conditional test rejects when

$$p(k;m) = \min\left\{2\min\left(F(k;m,\tfrac{1}{2}),\ 1-F(k-1;m,\tfrac{1}{2})\right),1\right\} < \alpha, \tag{5}$$

evaluated at $k = B_n$, $m = M_n$. Equation (5) is the rule implemented in `src/testset_mdd.py`, including the truncation at 1 and the strict inequality.

**Lemma 1** ($M_n$ is a cut for $\theta$)**.** *Under 3–4 with $(\Delta, d) \in \mathcal{P}$, $d > 0$,*

$$M_n \sim \text{Bin}(n,d), \qquad B_n \mid M_n{=}m \sim \text{Bin}(m,\theta), \qquad \theta := \frac{d+\Delta}{2d} \in [0,1], \tag{6}$$

*and the law of $M_n$ does not depend on $\Delta$.*

*Proof.* $M_n = \sum_i \mathbf{1}\{X_i \neq Y_i\}$ is a sum of $n$ i.i.d. Bernoulli$(d)$ variables, and $d$ does not involve $\Delta$. For the conditional law, with $\beta = d\theta$ and $\gamma = d(1-\theta)$, for $0 \le k \le m \le n$,

$$\Pr(B_n{=}k, M_n{=}m) = \frac{n!}{k!(m-k)!(n-m)!}\beta^k\gamma^{m-k}(1-d)^{n-m} = \binom{n}{m}d^m(1-d)^{n-m}\binom{m}{k}\theta^k(1-\theta)^{m-k};$$

divide by $\Pr(M_n{=}m) = \binom{n}{m}d^m(1-d)^{n-m}$. Since $\theta$ ranges over $[0,1]$ for every $d \in (0,1]$, the reparameterisation $(\Delta, d) \mapsto (\theta, d)$ is variation independent, which permits conditioning on $M_n$. □

**Definition 1** (Conditional and unconditional power)**.** For $m \ge 0$ let $R_m(\alpha) = \{k \in \{0,\dots,m\} : p(k;m) < \alpha\}$. The *conditional power* is $g_m(\theta) = \Pr_{K\sim\text{Bin}(m,\theta)}(K \in R_m(\alpha))$ and the *unconditional power* is $\pi(n,\Delta,d) = \Pr(B_n \in R_{M_n}(\alpha))$.

**Theorem 2** (Exact mixture representation)**.** *Under 3–4, for $(\Delta, d) \in \mathcal{P}$ with $d > 0$,*

$$\pi(n,\Delta,d) = \sum_{m=0}^{n}\binom{n}{m}d^m(1-d)^{n-m}\,g_m\left(\frac{d+\Delta}{2d}\right). \tag{7}$$

*The mixing weights are free of $\Delta$.*

*Proof.* Condition on $M_n$ and apply the tower property; the inner probability is $g_{M_n}(\theta)$ by Lemma 1, and $M_n \sim \mathrm{Bin}(n, d)$. □

**Lemma 3** (Structure and emptiness of the rejection set)**.** *Fix $m \geq 0$, $\alpha \in (0, 1)$.*

*(i) $p(\cdot; m)$ is symmetric, $p(k; m) = p(m - k; m)$, nondecreasing on $\{0, \ldots, \lfloor m/2 \rfloor\}$ and nonincreasing on $\{\lceil m/2 \rceil, \ldots, m\}$.*

*(ii) $R_m(\alpha) = \{0, \ldots, a_m\} \cup \{m - a_m, \ldots, m\}$ with $a_m = \max\{k < m/2 : 2F(k; m, \frac{1}{2}) < \alpha\}$, and $R_m(\alpha) = \emptyset$ (convention $a_m = -1$) when no such $k$ exists. The two tails are disjoint.*

*(iii) $R_m(\alpha) = \emptyset \iff 2^{1-m} \geq \alpha$. With $m_\alpha := \min\{m : 2^{1-m} < \alpha\} = \lfloor \log_2(2/\alpha) \rfloor + 1$, the test can reject only when $m \geq m_\alpha$; at $\alpha = 0.05$, $m_\alpha = 6$.*

*(iv) Whenever $R_m(\alpha) \neq \emptyset$, $m - 2a_m - 1 \geq 1$.*

*(v) $g_m(\frac{1}{2}) = 2F(a_m; m, \frac{1}{2}) < \alpha$ when $R_m \neq \emptyset$ and $= 0$ otherwise.*

*Proof.* (i) Symmetry of $\mathrm{Bin}(m, \frac{1}{2})$ gives $F(k) = 1 - F(m - k - 1)$, whence $p(k; m) = p(m - k; m)$; and $F(k) + F(k - 1) \leq 1 \iff F(k - 1) \leq F(m - k - 1) \iff k \leq m/2$, so for $k \leq m/2$ the inner minimum is $F(k)$, nondecreasing, and for $k \geq m/2$ it is $\Pr(K \geq k)$, nonincreasing. Truncation at 1 preserves both. (ii) A strict sublevel set of a symmetric unimodal lattice function is a symmetric union of two tails; $a_m < m/2$ gives $m - a_m > a_m$, so they are disjoint and $g_m = F(a_m; m, \theta) + \Pr_\theta(K \geq m - a_m)$ contains no overlapping events. (iii) The minimum of $p(\cdot; m)$ is at $k \in \{0, m\}$ with $p(0; m) = 2^{1-m}$; at $\alpha = 0.05$, $2^{1-m} < 0.05 \iff m > 1 + \log_2 20 = 5.3219$. (iv) $2F(a_m; m, \frac{1}{2}) < \alpha < 1$ forces $F(a_m; m, \frac{1}{2}) < \frac{1}{2}$. For odd $m$, $F(\frac{m-1}{2}; m, \frac{1}{2}) = \frac{1}{2}$, so $a_m \leq \frac{m-1}{2} - 1$ and $m - 2a_m - 1 \geq 2$; for even $m$, $a_m \leq m/2 - 1$ and $m - 2a_m - 1 \geq 1$. (v) Immediate from (ii) and symmetry. □

**Corollary 4** (Actual size)**.** *Under 3–4, with $n \geq m_\alpha$ and $d > 0$,*

$$\pi(n, 0, d) = \sum_{m \geq m_\alpha} \binom{n}{m} d^m (1 - d)^{n-m} 2F(a_m; m, \tfrac{1}{2}) < \alpha \Pr(M_n \geq m_\alpha) \leq \alpha.$$

*The nominal 5% test has size strictly below 5%: exactly 0.021989 at $(n, d) = (100, 0.10)$, 0.029784 at $(500, 0.05)$, 0.038153 at $(1797, 0.045)$ and 0.042431 at $(758, 0.2775)$.*

*Proof.* Immediate from Theorem 2 and Lemma 3(iii),(v); strictness uses $\Pr(M_n \geq m_\alpha) > 0$, which needs $n \geq m_\alpha$, $d > 0$. □

**Lemma 5** (Exactness under the sharp per-item null)**.** *Suppose 3 holds and item $i$ has cell probabilities $(\beta_i, \gamma_i, 1 - \beta_i - \gamma_i)$, not required common across items. If $\beta_i = \gamma_i$ for* every *$i$, then $B_n \mid M_n{=}m \sim \mathrm{Bin}(m, \frac{1}{2})$ exactly, and* (5) *has size at most $\alpha$.*

*Proof.* Let $D_i = \mathbf{1}\{X_i \neq Y_i\}$ and $S = \{i : D_i = 1\}$. Items with $\beta_i = \gamma_i = 0$ have $D_i = 0$ a.s. and never enter $S$. For the remainder, $\Pr(X_i{=}1, Y_i{=}0 \mid D_i{=}1) = \beta_i/(\beta_i + \gamma_i) = \frac{1}{2}$. By 3 items contribute independently, so conditionally on the realised set $S$ with $|S| = m$, $B_n$ counts successes in $m$ independent fair Bernoulli trials. This law depends on $S$ only through $|S|$, so conditioning on $M_n = m$ alone gives the same distribution. Size follows from Lemma 3(v). □

*Remark* 2 (The scope of Lemma 5). Its hypothesis is the *sharp* null $\beta_i = \gamma_i$ for every $i$. The null the reported $p$-values are read against is the composite $\Delta = n^{-1} \sum_i (\beta_i - \gamma_i) = 0$, a strictly larger null family. Lemma 5 does not address this larger set. Rejection concerns item-level asymmetry, rather than the conclusion "$A$ is more accurate than $B$ on the domain". Moreover, the lemma is formulated within E2, and under greedy decoding $\beta_i, \gamma_i \in \{0, 1\}$, so the sharp null becomes "the systems agree on every item", under which $M_n = 0$ a.s. and $p = 1$ a.s. Lemma 5 is therefore *vacuous* on the two boards whose $p$-values are reported. E3 provides their design-based interpretation.

**Lemma 6** (Under-dispersion of $B_n - C_n$ under heterogeneity)**.** *Under 3 with item-specific $(\beta_i, \gamma_i)$,*

$$\mathrm{Var}(B_n - C_n) = \mathbb{E}[M_n] - \sum_{i=1}^{n} (\beta_i - \gamma_i)^2 \;\leq\; \mathbb{E}[M_n],$$

*with equality iff $\beta_i = \gamma_i$ for every $i$.*

*Proof.* Write $B_n - C_n = \sum_i U_i$ with $U_i \in \{-1, 0, 1\}$, independent, $\mathbb{E}U_i = \beta_i - \gamma_i$ and $\mathbb{E}U_i^2 = \beta_i + \gamma_i$. Then $\mathrm{Var}(B_n - C_n) = \sum_i(\beta_i + \gamma_i) - \sum_i(\beta_i - \gamma_i)^2$ and $\mathbb{E}[M_n] = \sum_i(\beta_i + \gamma_i)$. □

**Open question 7** (Validity on the composite null)**.** *Under 3 with arbitrary item-specific $(\beta_i, \gamma_i)$ satisfying $\sum_i \beta_i = \sum_i \gamma_i$, the test* (5) *has size at most $\alpha$.*

*Evidence and the missing step.* An exhaustive search over 4,346 heterogeneous independent-item designs satisfying $\sum \beta_i = \sum \gamma_i$ (pure-$b$, pure-$c$ and symmetric-filler items; exact $(B, C)$ joint law by dynamic programming; exact conditional $p$-values) found maximum size $0.0296 < \alpha$. Lemma 6 gives a related variance result: heterogeneity strictly *under*-disperses $B_n - C_n$ relative to the $\mathrm{Bin}(m, \frac{1}{2})$ reference. The missing step is that under-dispersion of the variance does not by itself dominate the two *tails* of the exact conditional test: conditionally on $M_n = m$ the statistic is a sum of $m$ independent Bernoulli$(\theta_i)$ variables with $\sum_i(\theta_i - \frac{1}{2})$ not necessarily 0 on the realised $S$, and no majorisation argument extending the result from $\mathrm{Var} \le \mathbb{E}[M_n]$ to $\Pr(B_n \in R_{M_n}) \le \alpha$ has been supplied. The size bound therefore is supported by this finite search and remains unproved.

**Proposition 8** (Heterogeneity is not conservative for *power*)**.** *Assumption 4 cannot be dropped from the power calculation, and the trinomial power is neither an upper nor a lower bound on the power of a heterogeneous E2 design with the same $(n, \Delta, d)$. At $\alpha = 0.05$:*

*(i) $n = 120$, $d = \Delta = 0.05$: trinomial power $\Pr(\mathrm{Bin}(120, 0.05) \ge 6) = 0.558452$ (55.85%), while the design in which $nd = 6$ items have $\beta_i = 1$ and the rest $\beta_i = \gamma_i = 0$ has power 100%.*

*(ii) $n = 100$, $d = \Delta = 0.05$: trinomial power $\Pr(\mathrm{Bin}(100, 0.05) \ge 6) = 0.384001$ (38.40%), while the analogous design with $nd = 5$ deterministic discordances has power 0%.*

*Proof.* Both designs have $\bar\beta = d$, $\bar\gamma = 0$, hence the stated $\Delta, d$, and both have $M_n = nd$ a.s. with $B_n = M_n$; apply Lemma 3(iii). The trinomial values are Theorem 2 evaluated exactly. □

*Remark* 3 (Scope and two effects of heterogeneity). Proposition 8's designs are inadmissible under E1, where item difficulty is integrated out by the draw of $Z_i$ and 4 is a theorem rather than an assumption. The proposition concerns E2. Heterogeneity affects power in two ways: (a) $\mathrm{Var}(M_n) = \sum_i d_i(1 - d_i) \le n\bar d(1 - \bar d)$, so heterogeneity always *concentrates* $M_n$, which raises power when the concentration point is at or above $m_\alpha$ and lowers it otherwise; and (b) $B_n \mid M_n$ becomes Poisson-binomial when the direction ratios $\theta_i = \beta_i/(\beta_i + \gamma_i)$ are heterogeneous, independently altering power. The two deterministic examples above vary only (a); they establish that either direction is possible, without quantifying typical effects. To estimate the effect under a fitted model, we obtain per-item $(\beta_i, \gamma_i)$ from a two-parameter item-response model (Rodriguez et al., 2021) fitted to each board small enough to simulate item by item, and compare the exact test's power with the homogeneous design's at the same $n$, $d$ and $\Delta$. For pairs whose power lies between 50% and 95%, the mean difference is +0.06, −0.03 and +0.12 percentage points on HELM MMLU, GSM8K and SWE-bench Verified (68, 12 and 80 pairs; Monte-Carlo standard errors 0.02, 0.05 and 0.02). The homogeneous design run twice with independent seeds differs by 0.00, −0.11 and +0.03. At a five-times-coarser resolution the difference is indistinguishable from zero. At the reported precision it is detectable in the two larger samples, but its magnitude is about a tenth of a percentage point of power, relative to the 80% power at which the limit is defined (`src/mx31_item_heterogeneity.py`; all twelve fits converged). That simulation imposes local independence given the item parameters, so it bounds effects (a) and (b) under that model and does not address dependence *between* items. $\mathrm{MDD}(n, d)$ is a *characteristic* value of the exchangeable idealisation with the same size and disagreement rate, not a bound on the limit of any particular benchmark.

### D.3 The closed form and its error

**Definition 2** (MDD)**.** For $n \in \mathbb{N}$, $d \in (0, 1]$, level $\alpha$, target power $\tau$,

$$\mathrm{MDD}_{\alpha,\tau}(n, d) := \inf\{\Delta \in [0, d] : \pi(n, \Delta, d) \ge \tau\}, \qquad \inf \emptyset := +\infty. \tag{8}$$

The *algebraic* limit is $\mathrm{MDD}_{\mathrm{alg}}(n, d) := Z\sqrt{d/n}$ and the relative error is $\mathcal{E} := (\mathrm{MDD}_{\mathrm{alg}} - \mathrm{MDD})/\mathrm{MDD}$.

*Remark* 4 (Simulation, algebraic approximation and exact evaluation). The implementation contains two estimators of MDD. The simulation estimator, `mdd()` in `src/testset_mdd.py`, bisects the Monte Carlo `power()` over $S$ trinomial draws for $J = 18$ steps. Randomness is fresh at every step *within* a call, but each call re-seeds `default_rng(0)`, so successive grid evaluations share a random stream; the `exact_upto=2000` argument of `power()` is unused. The algebraic one, $Z\sqrt{d/n}$, appears only in `src/sizing.py` and one main-text sentence. The error analysis of this subsection applies to the algebraic estimator and to the $\sqrt{d}$ ratios of Corollary 11. It does not apply to the reported tier counts, which are evaluated exactly; Assumptions 6–7 remain relevant to those counts. Theorem 2 and Lemma 3(ii) give a deterministic alternative to simulation:

$$\pi(n, \Delta, d) = \sum_{m=m_\alpha}^{n} \binom{n}{m} d^m (1-d)^{n-m} \big[F(a_m; m, \theta) + 1 - F(m - a_m - 1; m, \theta)\big], \quad \theta = \tfrac{d+\Delta}{2d}, \quad (9)$$

which costs $O(n)$ binomial c.d.f. evaluations plus $O(n)$ inversions for $\{a_m\}$, and makes MDD a deterministic function of $(n, d, \alpha, \tau)$. Brent's method evaluates the detection limit to machine precision.

**Open question 9** (Two-term expansion of the relative error)**.** *As* $nd \to \infty$*,*

$$\mathcal{E} = -\frac{1}{Z\sqrt{nd}} + \frac{1}{nd}\left(\frac{z_\tau Z}{2} + \frac{1}{Z^2}\right) + o\left(\frac{1}{nd}\right) = -\frac{0.356941}{\sqrt{nd}} + \frac{1.306344}{nd} + o\left(\frac{1}{nd}\right), \quad (10)$$

*so* $\mathcal{E}$ *changes sign near* $nd^\dagger = (1.306344/0.356941)^2 = 13.4$.

*Unresolved derivation.* Three steps remain unproved. (a) The size deficit is averaged before the power is taken, $\bar\alpha := \mathbb{E}[\alpha_M]$ and then power at $\bar\alpha$; power is not linear in level, and the Jensen defect scales with $\mathrm{Var}(\alpha_M) = O(1/(nd))$, the same order as the second term. (b) Replacing the mixture (7) by its value at $m = nd$ is proved only for the *smooth* non-centrality factor $h(m) = \Phi\big((\sqrt{m}\,\Delta/d - z_\alpha)/\sqrt{1 - \Delta^2/d^2}\big)$; $g_m$ itself is not smooth in $m$ because $a_m$ is a step function, and the discarded lattice part is $\Theta((nd)^{-1/2})$, one order worse than the $O((nd)^{-1})$ claimed, since $\alpha - \alpha_m$ oscillates in $m$ with amplitude $4\phi(z_\alpha)/\sqrt{m}$ while the averaged deficit tends to $2\phi(z_\alpha)/\sqrt{nd}$. (c) A size deficit $\delta$ is converted into a shift of the *continuous* critical multiplier $z_{1-(\alpha-\delta)/2}$, which assumes a lattice test of realised size $\bar\alpha$ has the power of a continuous test at that level, an equivalence that does not hold for a discrete test. *Evidence.* At $nd = 4245$ the exact $\mathcal{E} = -0.00518108$ against $-0.00517095$ from (10) and $-0.00520097$ from the uncorrected coefficient. Residuals in units of $(nd)^{-3/2}$ are +2.79, +1.60, +2.44, +2.80 at $nd = 86, 131, 199, 4245$ under (10), versus +1.61, +0.14, +0.65, −5.50 under the uncorrected form.

**Proposition 10** (The algebraic limit is not uniformly optimistic)**.** $\mathrm{MDD}_{\mathrm{alg}}$ *is not below* MDD *at every* $(n, d)$*: the sign of* $\mathcal{E}$ *reverses, and the largest nd at which* $\mathrm{MDD}_{\mathrm{alg}} \geq$ MDD *found by exact enumeration is at least* 18.5 *(at* $d = 0.74$*,* $n = 25$*:* 48.2003 *vs* 48.0714 *accuracy points) within the range of disagreement rates this paper measures, and at least* 66.33 *(at* $d = 0.99$*,* $n = 67$*) over the full parameter space. The boundary is increasing in d; any quoted value is a property of the sweep grid, not of* $\mathcal{P}$*.*

**Corollary 11** (The $\sqrt{d}$ law)**.** $\mathrm{MDD}(n, d_1)/\mathrm{MDD}(n, d_2) = \sqrt{d_1/d_2}\,(1 + \mathcal{E}_1)^{-1}(1 + \mathcal{E}_2)$*. The paper's three* $\sqrt{d}$ *ratios are exact to* 0.10–0.81%*. The n-scaling is not equally safe: "27 times the instances divides the limit by* 5.3*" is* 5.3157 *exactly but* 5.2268 *under the closed form.*

**Theorem 12** (Monotone sandwich)**.** *Let* $I = [z_\alpha\sqrt{d/n},\ (z_\alpha + z_{0.99})\sqrt{d/n}]$*, let* $\pi_N$ *be the normal-approximation power and suppose*

$$\varepsilon := \sup_{\Delta \in I} |\pi_{\mathrm{ex}}(n, \Delta, d) - \pi_N(n, \Delta, d)| < 0.19. \quad (11)$$

*Then* $\mathrm{MDD}(n, d)$ *lies between the* $\pi_N$*-roots at levels* $\tau - \varepsilon$ *and* $\tau + \varepsilon$*, and consequently* $|\mathcal{E}| \leq \max\big\{|Z/u_{\tau-\varepsilon} - 1|,\ |Z/u_{\tau+\varepsilon} - 1|\big\}$ *with* $u_p$ *the corresponding normal multiplier.*

*Proof.* By Theorem 17, $\pi_{\mathrm{ex}}(n, \cdot, d)$ is continuous and strictly increasing, so $\{\Delta : \pi_{\mathrm{ex}} \geq \tau\}$ is an interval $[\Delta^*, d]$. On $I$, $\pi_N - \varepsilon \leq \pi_{\mathrm{ex}} \leq \pi_N + \varepsilon$, so $\pi_N(\Delta^*) \in [\tau - \varepsilon, \tau + \varepsilon]$ provided $\Delta^* \in I$; (11) with $\tau + \varepsilon < 0.99$ places both $\pi_N$-roots inside $I$ and hence $\Delta^*$ with them, since $\pi_N$ is strictly increasing on $I$. Inverting $\pi_N$ gives the bracket, and dividing by MDD gives the bound on $|\mathcal{E}|$. □

**Open question 13** (Explicit non-asymptotic bound)**.** *There is an explicit $K_0(nd, d)$ with $\varepsilon \le K_0(nd, d)/\sqrt{nd}$, assembled from a Berry–Esseen term, a lattice-cutoff term, an $M_n$-averaging term and a wrong-tail term.*

*Uncontrolled error terms.* The decomposition has three unresolved terms. (a) The current decomposition bounds the Berry–Esseen error, the substitution $s_{M_n} \to z_\alpha$, the substitution $\sqrt{M_n} \to \sqrt{nd}$ and the wrong tail, but does not bound the error from replacing the alternative standard deviation by the null one, i.e. the division by $\sqrt{1-\Delta^2/d^2}$. This factor enters the fixed point $u^* = z_\alpha + z_\tau\sqrt{1-u^{*2}/(nd)}$. Its contribution is $O(1/(nd))$, preserving the proposed rate, but the decomposition does not establish the bound at every $(n, d)$. (b) The averaging term bounds $\Delta/d \le Z/\sqrt{nd}$; since $\varepsilon$ is a supremum over $I$, the correct bound is $\sup_I \Delta/d = (z_\alpha + z_{0.99})/\sqrt{nd} = 4.286312/\sqrt{nd}$, larger by a factor 1.530. (c) $K_0$ is not a function of $nd$ alone: the averaging term carries $\|\sqrt{M_n} - \sqrt{nd}\|_2 \to \sqrt{1-d}/2$, so any limiting constant is $d$-dependent and 6.409 is at best $\sup_d \lim$, approached only as $d \to 0$; measured values at $nd = 2 \times 10^5$ are already *below* it and continue to decrease. Two additional terms require explicit treatment: $\mathbb{E}[M_n^{-1/2}] = +\infty$ because $\Pr(M_n = 0) > 0$, so the term must be written $\mathbb{E}[M_n^{-1/2}\mathbf{1}\{M_n \ge m_\alpha\}]$; and the released code sums a fifth term $\Pr(M_n < m_\alpha)$ that the display omits. No numerical value for the corrected $K_0$ is reported here. In comparisons with the exact mixture at every point where (11) holds, the proposed bound has slack 17–29× and is informative only for $nd \gtrsim 1.3 \times 10^3$, so it does *not* cover the operating points $nd = 86, 131, 199$.

**Proposition 14** (Monte Carlo error of the shipped estimator)**.** *Let $\widehat{\pi}_S$ be the Monte Carlo power from $S$ replicates and $\widehat{\Delta}_S$ the output of* `mdd()` *after $J$ bisection steps on $[0, d]$. Assume $\mathrm{MDD}(n, d) < \infty$. For $\varepsilon > 0$ let $\mathcal{B}_\varepsilon = \{\Delta \in [0, d] : |\pi(n, \Delta, d) - \tau| \le \varepsilon\} = [\Delta_\varepsilon^-, \Delta_\varepsilon^+]$, an interval containing $\Delta^* = \mathrm{MDD}(n, d)$ by Theorem 17. Then with probability at least $1 - (J+1)e^{-2S\varepsilon^2}$,* either `mdd()` *returns* `None`*, which can happen only if $\pi(n, d, d) \le \tau + \varepsilon$,* or $\widehat{\Delta}_S \in [\Delta_\varepsilon^-,\ \Delta_\varepsilon^+ + d2^{-J}]$. *If in addition $\pi(n, d, d) > \tau + \varepsilon$, the first branch is excluded.*

*Proof.* Fix $j \le J$. Conditionally on the history the evaluation point $\Delta_j$ is deterministic and $S\widehat{\pi}_S(\Delta_j) \sim \mathrm{Bin}(S, \pi(\Delta_j))$, so Hoeffding gives $\Pr(\widehat{\pi}_S(\Delta_j) - \pi(\Delta_j) > \varepsilon) \le e^{-2S\varepsilon^2}$ one-sidedly. Union over the $J+1$ evaluations (the feasibility check included) gives the event $G$ on which every comparison at a point outside $\mathcal{B}_\varepsilon$ agrees with the noiseless comparison. On $G$, if the feasibility check fails then $\widehat{\pi}_S(d) < \tau$, which outside $\mathcal{B}_\varepsilon$ would force $\pi(n, d, d) < \tau$, contradicting $\mathrm{MDD} < \infty$; hence $d \in \mathcal{B}_\varepsilon$, i.e. $\pi(n, d, d) \le \tau + \varepsilon$. Otherwise the loop is entered and, on $G$, `lo` is only assigned points where the outcome was "power $< \tau$", which outside $\mathcal{B}_\varepsilon$ forces $\Delta_j < \Delta_\varepsilon^+$, so $\texttt{lo} \le \Delta_\varepsilon^+$ throughout; symmetrically $\texttt{hi} \ge \Delta_\varepsilon^-$. At termination $\texttt{hi} - \texttt{lo} = d2^{-J}$ and $\widehat{\Delta}_S = \texttt{hi} \le \texttt{lo} + d2^{-J}$. □

**Corollary 15** (Magnitude)**.** *With $J = 18$ the term $d2^{-J}$ is negligible. At failure probability 0.05: $\varepsilon = 0.012186$ at $S = 20{,}000$ and $0.027249$ at $S = 4{,}000$ (used inside* `validate()`*). Through the normal approximation the quantity $2\varepsilon/[\phi(z_\tau)Z] = 2.5499\varepsilon$ is the* full *relative width of the bracket $\mathcal{B}_\varepsilon$ of Proposition 14, not a half-width:* 3.11% *and* 6.95% *of* MDD*, i.e. brackets* 0.045 *and* 0.100 *accuracy points wide on digits (*MDD $\approx$ 1.44 *accuracy points), equivalently* ±0.022 *and* ±0.050 *about the limit. A one-sided Bonferroni-normal band over the $J + 1 = 19$ evaluations gives* 2.01% *and* 4.50%*, i.e. widths* 0.029 *and* 0.065 *accuracy points, equivalently* ±0.014 *and* ±0.032 *accuracy points.* Directly measured*, the shipped estimator's s.d. over* 40 *seeds at $S = 20{,}000$ is* 0.015*,* 0.013 *and* 0.002 *accuracy points on the three calibration cells, whose limits are* 1.44*,* 2.49 *and* 5.48 *accuracy points and whose half-widths at $S = 20{,}000$ are therefore* ±0.022*,* ±0.039 *and* ±0.085 *accuracy points. Every measured s.d. lies below the smallest of those three half-widths under any pairing of cell to cell, so the Hoeffding bracket is conservative by a factor of about* 1.5 *at the least favourable pairing rather than by the factor of three the full width suggested. The two quantities are not the same functional — one is a spread, the other a* 95% *containment half-width — so the comparison fixes an order of magnitude and no more.*

*Remark* 5 (Algebraic bias relative to Monte Carlo error). The normal-approximation closed form has an error relative to the exact mixture ranging from −0.5% to −3.5% across all 40 cells of the benchmark table, and it is *optimistic in* 39 *of* 40: this systematic bias exceeds the symmetric Monte Carlo term of Corollary 15, so the table uses exact evaluation (Theorem 2) and has no Monte Carlo error.

*Remark* 6 (The calibration is exactly identified). `validate()` searches a grid of $d$ for the value minimising $|\mathrm{MDD}(n,d) - \widehat{\mathrm{MDD}}_{\text{measured}}|$: one free parameter fitted to one number, zero residual degrees of freedom. Three error terms exceed the reported agreement of 0.02 accuracy points, in decreasing order: (i) the *target* has a 95% bootstrap interval of ±0.11 to ±0.37 accuracy points (digits 1.44 [1.36, 1.58]; breast-cancer 2.49 [2.11, 2.75]; ABIDE 5.48 [5.10, 5.83]), 5–18× the claimed agreement; (ii) the minimum is taken over 236 grid values of a quantity subject to Monte Carlo noise, with grid spacing equivalent to 0.039 accuracy points, and the minimum of many noisy deviations is $\approx 0$ by construction, independently of resolution; (iii) the Monte Carlo term itself. Agreement after fitting therefore provides no independent validation. The comparison instead assesses plausibility within each benchmark: digits' implied $d = 0.045$ lies inside the 0.0073–0.0963 measured on digits; breast-cancer's implied $d = 0.0425$ lies *just above* the largest rate measured there, 0.04139; ABIDE's implied $d = 0.2775$ has no independent measurement, since `src/discordance.py` accepts only `{bc, digits}`. Every pair in `discordance.py` joins a model to a perturbation of itself, so none of these rates compares two distinct models.

### D.4 Monotonicity and the floor

**Lemma 16** (Derivative of the conditional power). *For $m$ with $R_m(\alpha) \neq \emptyset$ and $a = a_m$,*

$$g'_m(\theta) = m\binom{m-1}{a}\theta^a(1-\theta)^a\left[\theta^{m-2a-1} - (1-\theta)^{m-2a-1}\right], \tag{12}$$

*which is $> 0$ for $\theta \in (\frac{1}{2}, 1)$ and $= 0$ at $\theta = \frac{1}{2}$; it also vanishes at $\theta = 1$ whenever $a \geq 1$. If $R_m(\alpha) = \emptyset$ then $g_m \equiv 0$.*

*Proof.* By Lemma 3(ii), $g_m(\theta) = F(a; m, \theta) + \Pr_\theta(K \geq m - a)$. Using $\frac{d}{d\theta}\Pr_\theta(K \geq j) = m\binom{m-1}{j-1}\theta^{j-1}(1-\theta)^{m-j}$ and $\frac{d}{d\theta}F(a; m, \theta) = -m\binom{m-1}{a}\theta^a(1-\theta)^{m-1-a}$ with $\binom{m-1}{m-a-1} = \binom{m-1}{a}$ gives (12). By Lemma 3(iv) the exponent $m - 2a - 1 \geq 1$, so the bracket is strictly positive on $(\frac{1}{2}, 1]$; the prefactor $[\theta(1-\theta)]^a$ vanishes at $\theta = 1$ when $a \geq 1$. □

**Theorem 17** (Power is strictly increasing in the gap). *Fix $n \geq m_\alpha$ and $d \in (0, 1)$. Then $\Delta \mapsto \pi(n, \Delta, d)$ is a polynomial in $\Delta$, hence continuous, and is strictly increasing on $[0, d]$.*

*Proof.* Continuity is clear from (7). The map $\theta = (d + \Delta)/(2d)$ is a strictly increasing bijection $[0, d] \to [\frac{1}{2}, 1]$, and (7) is a nonnegative mixture of the nondecreasing $g_m$. For strictness on $(0, d)$ it suffices that some $m$ with $g'_m > 0$ carries positive weight: since $d \in (0, 1)$ and $n \geq m_\alpha$, the weight on $m = m_\alpha$ is positive and $a_{m_\alpha} = 0$, so (12) gives $g'_{m_\alpha}(\theta) = m_\alpha\left[\theta^{m_\alpha - 1} - (1-\theta)^{m_\alpha - 1}\right]$, which is $> 0$ throughout $(\frac{1}{2}, 1]$ because $\theta > 1 - \theta \geq 0$ there and $m_\alpha - 1 \geq 1$ by Lemma 3(iv). Strict increase on the closed interval follows by continuity. □

*Remark* 7. At $d = 1$ only $m = n$ carries weight, and by Lemma 16 $\partial\pi/\partial\Delta = 0$ at $\Delta = d = 1$ for every $n \geq 9$. This boundary is relevant to the counterexample in Proposition 25.

**Theorem 18** (Existence, attainment, uniqueness, feasibility). *Fix $n \geq m_\alpha$, $d \in (0, 1)$, $\alpha < \tau$. Then*

*(i)* $\sup_{\Delta \in [0,d]} \pi(n, \Delta, d) = \pi(n, d, d) = \Pr(\mathrm{Bin}(n, d) \geq m_\alpha)$;

*(ii)* $\mathrm{MDD}(n, d) < \infty \iff \Pr(\mathrm{Bin}(n, d) \geq m_\alpha) \geq \tau$;

*(iii) when finite the infimum is attained, lies in $(0, d]$, and is the unique solution of $\pi(n, \Delta, d) = \tau$.*

*Proof.* (i) The supremum is at $\Delta = d$ by Theorem 17; there $\gamma = 0$, $\theta = 1$, and $g_m(1) = \mathbf{1}\{m \in R_m\} = \mathbf{1}\{m \geq m_\alpha\}$ by Lemma 3(ii),(iii). Substitute into (7). (ii) Immediate. (iii) $\pi(n, \cdot, d)$ is continuous and strictly increasing on $[0, d]$ with $\pi(n, 0, d) < \alpha < \tau$ (Corollary 4) and $\pi(n, d, d) \geq \tau$; apply the intermediate value theorem, and note $\{\Delta : \pi \geq \tau\} = [\Delta^*, d]$ is closed. □

**Corollary 19** ($\mathrm{MDD} \leq d$). *Whenever $\mathrm{MDD}(n, d)$ is finite, $\mathrm{MDD}(n, d) \leq d$.*

**Definition 3** (Admissible disagreement rates). For a pair with accuracies $a_X \geq a_Y$ and gap $g = a_X - a_Y$, the Fréchet bounds on the joint table give

$$d \in \mathcal{D}(a_X, a_Y) := \left[\, g,\ \min(a_X, 1-a_Y) + \min(a_Y, 1-a_X) \,\right] = \left[\, g,\ \min(a_X + a_Y,\ 2 - a_X - a_Y) \,\right]. \tag{13}$$

*Proof of the identity and of attainability.* $\beta \le \min(a_X, 1-a_Y)$ and $\gamma \le \min(a_Y, 1-a_X)$ since $\beta \le \Pr(X{=}1)$, $\beta \le \Pr(Y{=}0)$ and symmetrically. If $a_X + a_Y \le 1$ then $1-a_Y \ge a_X$ and $1-a_X \ge a_Y$, so the two minima are $a_X$ and $a_Y$; otherwise they are $1-a_Y$ and $1-a_X$. In both branches the two minima differ by exactly $g$, so the upper bound is attained by a table with $\beta$ and $\gamma$ at their maxima, which is consistent with the margins. □

The admissible interval can be computed from the two reported accuracies. For two systems ($a_X = 0.90$, $a_Y = 0.85$) the admissible set is $[0.05, 0.25]$, not $[0.05, 1]$: four times narrower. Restricting the domain of a supremum cannot increase it, so a floor computed over $\mathcal{D}$ declares *more* pairs unresolvable than one computed over $[g, 1]$.

**Theorem 20** (The floor is a six-disagreement criterion)**.** *Fix $n \ge 1$ and $g \in (0,1)$. The following are equivalent:*

*(i)* $g < \mathrm{MDD}(n, g)$*;*

*(ii)* $\mathrm{MDD}(n, g) = +\infty$*;*

*(iii)* $\Pr(\mathrm{Bin}(n, g) \ge m_\alpha) < \tau$*;*

*(iv)* $g < \Delta_{\mathrm{fl}}(n) := \min\{\gamma \in (0,1) : \Pr(\mathrm{Bin}(n, \gamma) \ge m_\alpha) \ge \tau\}$.

*No monotonicity of* $\mathrm{MDD}(n, \cdot)$ *in $d$ is used.*

*Proof.* (ii)⇒(i) follows directly. (i)⇒(ii): if $\mathrm{MDD}(n, g)$ were finite then $\mathrm{MDD}(n, g) \le g$ by Corollary 19, contradicting (i). (ii) ⟺ (iii) is Theorem 18(ii) with $d = g$, together with the convention $\mathrm{MDD} = +\infty$ when $n < m_\alpha$, where the tail probability is $0 < \tau$. (iii) ⟺ (iv): $\gamma \mapsto \Pr(\mathrm{Bin}(n, \gamma) \ge m_\alpha)$ is continuous and strictly increasing on $(0,1)$ (stochastic ordering in the success probability), so its super-level set at $\tau$ is a closed interval $[\Delta_{\mathrm{fl}}(n), 1)$ and the minimum is attained. □

**Corollary 21** (The fixed point, and about eight items)**.** $\Delta = \mathrm{MDD}(n, \Delta)$ *has exactly one root, namely $\Delta_{\mathrm{fl}}(n)$, and $n\Delta_{\mathrm{fl}}(n) \to \lambda^* = 7.905993$, the root of $\Pr(\mathrm{Poisson}(\lambda) \ge 6) = \tau$. Numerically, $\Delta_{\mathrm{fl}}(n) = \lambda^*/n - 11.487/n^2 + O(n^{-3})$, accurate to five significant figures for $n \ge 80$.*

*Proof.* Uniqueness and the identification with $\Delta_{\mathrm{fl}}$ are Theorem 20. For the limit, $G_n(\lambda) := \Pr(\mathrm{Bin}(n, \lambda/n) \ge 6)$ converges pointwise to the continuous, strictly increasing $G(\lambda) = \Pr(\mathrm{Poisson}(\lambda) \ge 6)$, and each $G_n$ is nondecreasing, so by Pólya's theorem the convergence is uniform on compacts; uniform convergence of monotone functions to a strictly increasing continuous limit implies convergence of the roots, so $n\Delta_{\mathrm{fl}}(n) \to \lambda^*$. The $-11.487/n^2$ term is numerical only: $n(\lambda^* - n\Delta_{\mathrm{fl}}(n))$ runs over $[11.4873, 11.5361]$ for $n \in [20, 2{\times}10^5]$, with $n\Delta_{\mathrm{fl}}(2{\times}10^5) = 7.905936$. □

**Proposition 22** (Sufficiency of $\lambda^*/d$)**.** *Let $m_\alpha \le \lambda^*$, which at $\alpha = 0.05$ is the condition $\tau \ge \Pr(\mathrm{Poisson}(6) \ge 6) = 0.5543$ and holds at the paper's 80% power, where $\lambda^* = 7.906$. Then $n\Delta_{\mathrm{fl}}(n) < \lambda^*$ at every finite $n \ge m_\alpha$, so $n \ge \lambda^*/d$ is* sufficient *for* $\mathrm{MDD}(n, d) < \infty$.

*Proof.* Anderson & Samuels (1967, Corollary 2.1, eq. (2.3)) prove, for $n \ge \lambda$ and every integer $k \le \lambda - 1$, the strict inequality $P(k; \lambda) > B(k; n, \lambda/n)$ between the Poisson and binomial distribution functions. With $k = m_\alpha - 1$ this is $\Pr(\mathrm{Bin}(n, \lambda/n) \ge m_\alpha) > \Pr(\mathrm{Poisson}(\lambda) \ge m_\alpha)$ whenever $m_\alpha \le \lambda$. At $\lambda = \lambda^*$, where the Poisson power equals $\tau$ by definition, the binomial power at $\gamma = \lambda^*/n$ therefore exceeds $\tau$; since $\gamma \mapsto \Pr(\mathrm{Bin}(n, \gamma) \ge m_\alpha)$ is continuous and strictly increasing, the smallest $\gamma$ reaching $\tau$ is strictly below $\lambda^*/n$, which is the claim for $n \ge \lambda^*$. For $m_\alpha \le n < \lambda^*$ it is immediate, since $\Delta_{\mathrm{fl}}(n) \le 1$ gives $n\Delta_{\mathrm{fl}}(n) \le n < \lambda^*$. The hypothesis is needed: at 50% power the Poisson root is $\lambda^* \approx 5.67 < m_\alpha$, the cited inequality does not apply at $k = 5$, so the statement is not asserted there. □

*Illustration.* Exact evaluation gives

$$n\Delta_{\mathrm{fl}}(n) = 7.5224,\ 7.676,\ 7.791,\ 7.849,\ 7.883,\ 7.895,\ 7.904,\ 7.906$$

at $n = 30, 50, 100, 200, 500, 1000, 5000, 10^5$ — the threshold is approached *from below*. Consequently, at every finite $n$, a benchmark with fewer than about $7.91/d$ items can still have a detection limit: $n = 100$, $d = 0.079$ has $nd = 7.900 < \lambda^*$ yet $\Pr(\mathrm{Bin}(100, 0.079) \geq 6) = 0.81069 \geq \tau$; likewise $n = 157$, $d = 0.05$. The exact criterion is $\Pr(\mathrm{Bin}(n, d) \geq 6) \geq 0.8$, and $\lambda^*/d$ is its $n \to \infty$ envelope. The proximity of the normal sizing constant $Z^2 = 7.8489$, within 0.7% of $\lambda^*$, does not establish equivalence: the normal rule has no $m_\alpha$, and its rejection boundary at $\Delta = d$ is $M_n \geq z_\alpha^2 = 3.84$, whereas the exact test cannot reject below $M_n = 6$ — and the exact threshold is a step function of $(n, d)$, not a constant over $d$. The value of $\lambda^*$ also depends on the target power: it moves from $\approx 5.7$ at 50% power to 7.906 at 80% to $\approx 9.3$ at 90%, a factor 1.6 in the item threshold, so the assumed 80% target materially affects the result.

**Proposition 23** (MDD is not monotone in $d$)**.** *(i) $\pi(n, \Delta, \cdot)$ is not nonincreasing: at $n = 6$, $\Delta = 5/6$, $\pi(6, \frac{5}{6}, \frac{5}{6}) = (5/6)^6 = 0.3348980$ while $\pi(6, \frac{5}{6}, 1) = (11/12)^6 + (1/12)^6 = 0.5932925$.*

*(ii)* $\mathrm{MDD}(n, \cdot)$ *is not nondecreasing across the feasibility boundary:* $\Pr(\mathrm{Bin}(100, 0.06) \geq 6) = 0.559307 < \tau$ *so* $\mathrm{MDD}(100, 0.06) = +\infty$*, while* $\Pr(\mathrm{Bin}(100, 0.10) \geq 6) = 0.942423 \geq \tau$ *so* $\mathrm{MDD}(100, 0.10)$ *is finite.*

*(iii) Within the feasible region violations exist but are confined to large $d$. Sweeping every distinct board size in this survey (119 of them) on a* 0.001 *lattice gives* 407 *violations in* 114,204 *consecutive-$d$ comparisons, the lowest at $d = 0.966$ and the largest a relative decrease of* 0.4946% *at $n = 545$ ($d : 0.999 \to 1.0$,* 12.0771 → 12.0173 *accuracy points). The bound is a property of the grid and not of the function: a coarser lattice gives* 25 *violations in* 14,242 *comparisons and places the bound at* 0.97*. We therefore quote $d < 0.96$, which both lattices support, and store the sweep as an artifact.*

**Corollary 24** (Direction of a borrowed rate)**.** *Fix $n$ among this survey's board sizes and a true pairwise rate $d < 0.96$ in the feasible region. If the limit is computed at a borrowed rate $d'$ with $d \leq d' < 0.96$, then $\mathrm{MDD}(n, d') \geq \mathrm{MDD}(n, d)$, so every gap the borrowed threshold calls resolvable is called resolvable at the true rate: borrowing high cannot manufacture a resolvable call.* This step is grid-supported and not proved. *It rests on Proposition 23(iii), which is an exhaustive evaluation on a* 0.001 *lattice over the* 119 *board sizes in this survey, not a theorem about real-valued $d$; the largest violation it finds is a relative decrease of* 0.4946% *and could in principle lie between lattice points. Monotonicity is not available to be proved in general—Proposition 23(iii) exhibits genuine violations, so any proof must carry the restriction rather than remove it—and establishing it on* $[0, 0.96]$ *would require a monotonicity argument for the inverted binomial tail $\pi(n, \Delta, \cdot)$ that we do not have. Readers who need an unconditional statement should use the closed form, where the direction follows from* $\mathrm{MDD} \propto \sqrt{d}$ *with no restriction at all. The converse fails: $d' < d$ lowers the threshold, and any gap between* $\mathrm{MDD}(n, d')$ *and* $\mathrm{MDD}(n, d)$ *is called resolvable that the true rate does not resolve. Two restrictions matter: above $d \approx 0.96$ the exact limit is not monotone (Proposition 23(iii)), and across the feasibility boundary (Proposition 23(ii)) a borrowed rate can move a pair from an infinite limit to a finite one, a change of kind and not of direction. Neither binds in practice: the largest board-level rate measured anywhere in this survey is* 0.4889*, roughly half the bound. Appendix Table 22 is the empirical check at a* 2.6× *mismatch: borrowing high produced no false-resolvable calls; borrowing low produced 8 of 6,105 while pairwise agreement* rose*, so agreement, the diagnostic an author would reach for, ranks the unsafe direction first.*

*Proof.* (i) At $d = \Delta = 5/6$, $\gamma = 0$ and Theorem 18(i) gives $\Pr(M_6 \geq 6) = (5/6)^6$. At $d = 1$, $\beta = 11/12$, $\gamma = 1/12$, $M_6 = 6$ a.s. and $B_6 \sim \mathrm{Bin}(6, 11/12)$; $R_6(0.05) = \{0, 6\}$ since $p(6; 6) = 2^{-5} = 0.03125 < 0.05$ and $p(5; 6) = 2 \cdot \frac{7}{64} = 0.21875 > 0.05$. (ii) Theorem 18(ii) and the two tail probabilities. (iii) Exhaustive evaluation of (9). □

**Proposition 25** ($n$-monotonicity: true for the randomised test, false for the exact one)**.** *Let $\varphi^{\mathrm{rand}}$ be the symmetric randomised conditional test of exact size $\alpha$ at every $m \geq 1$, with $\varphi_0^{\mathrm{rand}} :\equiv \alpha$. Then*

*(a) $\varphi^{\mathrm{rand}}$ is UMPU for testing $\theta = \frac{1}{2}$ against the two-sided alternative $\theta \neq \frac{1}{2}$; against the one-sided alternative $\theta > \frac{1}{2}$ it is not optimal, the one-sided test being UMP there and strictly more powerful at every $\theta > \frac{1}{2}$. Nonetheless $\mathrm{MDD}_{\mathrm{rand}}(\cdot, d)$ is non-increasing in $n$;*

*(b)* $\mathrm{MDD}(\cdot, d)$ *for the exact test is* not *non-increasing: at* $d = 1$*,* MDD = 62.5432*,* 64.6901*,* 66.6037 *accuracy points at* $n = 17, 18, 19$ *— one extra item raises the limit by* 3.43%*. Violations are rare and confined to the extreme:* 4 *in* 48,500 *consecutive-n comparisons, all at* $d = 0.95$*, none at any* $d \le 0.94$*. The grid stops at* 0.95*; an independent sweep that carries it further finds violations at* 0.96 *and above, so this grid establishes exhaustively only that none occur at or below* 0.94*.*

*Proof.* (a) Symmetry of the region gives $\mathbb{E}_{1/2}[B\varphi^{\mathrm{rand}}] = \alpha\,\mathbb{E}_{1/2}[B]$, which with the one-parameter exponential family structure of $\mathrm{Bin}(m, \theta)$ is the unbiasedness side condition, so $\varphi^{\mathrm{rand}}$ is UMPU (Lehmann–Romano 4.4.1). To compare $n$ and $n+1$, let $\psi$ discard the last pair and apply the size-$\alpha$ test to the first $n$; $\psi$ is a level-$\alpha$ test based on the full data, and $\mathbb{E}[\psi \mid B_{n+1}]$ has the same power function and is a function of the sufficient, complete statistic; it is unbiased and has the power function of the UMPU test at $n$, so it lies in the class over which $\varphi^{\mathrm{rand}}$ is optimal at $n+1$, and optimality within that class gives power at $n+1$ at least that at $n$ at every $\theta \neq \frac{1}{2}$. Combined with $\mathrm{Bin}(n, d) \preceq_{\mathrm{st}} \mathrm{Bin}(n+1, d)$ and Theorem 17, this implies a nonincreasing root. (b) Exact evaluation; the mechanism is that $a_m$ is constant at 4 for $m = 17, 18, 19$ while $g_m$ *falls* within that plateau: at $\theta = 0.8127$, the direction ratio at which $n = 17$ has power exactly 80%, $g_{17}, g_{18}, g_{19}$ are 80.0%, 76.3% and 72.4%, recovering to 84.4% at $m = 20$, where $a_m$ steps to 5. Restoring the target power requires a larger root and hence a larger limit. □

*Remark* 8. The exact test is unbiased at its *attained* size $\alpha_m$, not at level $\alpha$: $g_m(\theta) < 0.05$ for $\theta \in (0.5, 0.598]$ at $m = 6$, $(0.5, 0.687]$ at $m = 8$, $(0.5, 0.527]$ at $m = 40$. That gap is precisely the mechanism of (b).

**Open question 26** (The floor holds at every admissible $d$)**.** *For* $n > m_\alpha$ *and* $g < \Delta_{\mathrm{fl}}(n)$*,* $\sup_{d \in \mathcal{D}} \pi(n, g, d) < \tau$*.*

*Unproved extension and boundary counterexample.* Theorem 20 proves the statement *at* $d = g$. Extending this result to "at every admissible $d$" requires that $d \mapsto \pi(n, g, d)$ be maximised at $d = g$ on the relevant region, and that is false in general: at $n = 50$, $\Delta = 0.02$, power is 0.048% at $d = 0.02$ but 3.872% at $d = 0.98$, an 80-fold increase, because at negligible signal the power *is* the size $\bar{\alpha}(n, d)$, which rises with $d$ (0.0298 → 0.0450 at $n = 500$). The extension fails at $n = m_\alpha = 6$: there rejection requires $M_n = n$, so $d = 1$ strictly dominates $d = g$, and for $g \in [0.926985, 0.963492)$ the floor declares the pair unresolvable at every $d$ while $\pi(6, g, 1) = \left(\frac{1+g}{2}\right)^6 + \left(\frac{1-g}{2}\right)^6 \ge 0.8$ (e.g. $g = 0.95$: $\pi(6, g, g) = 0.73509$ but $\pi(6, g, 1) = 0.85907$). Hence the hypothesis $n > m_\alpha$. *Evidence for* $n > m_\alpha$*:* an exhaustive scan found no further counterexample at $n \ge 7$; over 22 board sizes $n = 20 \ldots 5 \times 10^4$, $\sup_{g < \Delta_{\mathrm{fl}}} \sup_d \pi$ is 80%, approached only as $g \uparrow \Delta_{\mathrm{fl}}$ at $d = g$ and never attained; and the excess $\sup_d \pi - \pi(g, g)$ is *identically zero* for every $g/\Delta_{\mathrm{fl}} \ge 0.7$ at every $n \ge 12$, so the evaluated boundary cases have zero excess, rather than a small positive margin. The same calculation is implemented in `src/floor_every_d.py` on the sizes the floor is reported on: over the 119 distinct board sizes of this survey ($n = 30$ to 50,000; 27 gaps up to $0.999999\,\Delta_{\mathrm{fl}}$, each against at least 60 rates on $[g, 1]$) the supremum of the exact power is 79.9999%, below 80% at every one; over every $n$ from 7 to 200 it is 79.9999%, with no failing size; and the largest excess of $\sup_d \pi$ over $\pi(g, g)$ for $g \ge 0.7\Delta_{\mathrm{fl}}$ at $n \ge 12$ is $0.0e+00$.

*Remark* 9 (Sensitivity of the floor to item dependence). The variance interpretation of $\mathrm{MDD}(n, d)$ uses 3, whereas $\Delta_{\mathrm{fl}}(n)$ depends on an extreme *lower tail*: six successes against a mean of 7.9. Tail probabilities can be more sensitive to dependence than variances. Thus, although the floor requires no assumed disagreement rate, it depends more strongly on 3 than $\mathrm{MDD}(n, d)$ does. Under a beta-binomial cluster model with $n/k$ independent clusters and intra-cluster correlation $\rho$, on Open LLM MMLU ($n = 14{,}042$, $k = 246$ subject clusters) $\Delta_{\mathrm{fl}}$ moves from 0.0563 accuracy points (i.i.d.) to 0.1306 at $\rho = 0.05$ and 0.1874 at $\rho = 0.10$; on ImageNet-R ($n = 30{,}000$, $k = 300$) from 0.0264 to 0.0669 and 0.0979. The exact floor replaces grid-read thresholds of 0.1096 accuracy points on MMLU and 0.0743 on ImageNet-R (Remark 10); an intra-cluster correlation of 0.05 on MMLU or 0.10 on ImageNet-R lifts the floor above those grid values, reversing that correction. The correction holds under 3. Moreover, $\rho$ can be estimated on a board with known clusters: HELM's 514 MMLU items are five whole subjects, so the design effect is estimable from the stored $91 \times 514$ matrix. Over 1,000 randomly chosen model pairs the subject-cluster design effect on $\widehat{\Delta}$ has median 2.24 (interquartile 1.02–4.76, 95th percentile 9.05, above one for 75.3% of pairs), a median standard-error inflation of 1.50×, and an implied $\rho \approx 0.0122$. That is 4.1 times smaller than the $\rho = 0.05$ at

which the correction reverses, so the estimated correlation is below the reversal threshold on this board. The same measurement gives an effective $n$ of about 229 rather than 514, so under this cluster interpretation the limits for this board are optimistic by roughly 1.5× and the exact tests on it are anti-conservative. Both effects increase the apparent number of resolved pairs and understate the unresolved share. With only five clusters the per-pair figure is noisy; the median across pairs is more stable than the tail estimates. Under E1 (items drawn from the mixture these five subjects define) the items are exchangeable and carry no cluster design effect; the 2.24 applies only to generalisation across *subjects*, which §D.8 sets aside, and no quoted limit depends on it. The correction has been estimated only on this one board; SWE-bench Verified's repository-level clustering and MTEB's document-level clustering may produce analogous effects. Their direction remains unknown and is not assumed to be favourable.

*Remark* 10 (Anti-conservative error from a tabulated floor). Estimating the floor from a grid of $\mathrm{MDD}(n, \cdot)$ tabulated from $d = 0.002$, with infeasible knots dropped and linear interpolation clamped below the first surviving knot (`floor_threshold()` in `src/bestcase.py` solves the floor exactly), sets the threshold at $\mathrm{MDD}(n, d_0)$ at the first *feasible* grid point $d_0$ — not $\mathrm{MDD}(n, 0.002)$, which for 192 of the survey's 202 boards does not exist — and $\mathrm{MDD}(n, d_0) > \Delta_{\mathrm{fl}}(n)$ whenever $\Delta_{\mathrm{fl}}(n)$ is not itself a knot. Clamping therefore declares resolvable pairs unresolvable, on boards of *every* size: $n = 500$ uses 1.5877 accuracy points against $\Delta_{\mathrm{fl}} = 1.5766$; $n = 1172$ uses 0.7337 against 0.6737 (+8.9%); $n = 10{,}000$ uses 0.1284 against 0.0790 (+63%); $n = 50{,}000$ uses 0.0575 against 0.0158 (3.64×). Under the exact floor of Theorem 20, the unresolvable share falls on 170 of 202 boards, is identical on 32 and rises on none; the median threshold inflation is 1.06×.

### D.5 The tier operator

A leaderboard is a finite multiset of scores $S = \{s_1, \ldots, s_M\} \subset [0, 1]$, $M \geq 2$, sorted $s_1 \geq \cdots \geq s_M$, together with one resolution $L > 0$. Write $R = s_1 - s_M$ for the range and $w = \max_{1 \leq t < M}(s_t - s_{t+1})$ for the largest adjacent gap. Both are local to this subsection: $R$ here is a score range, not the rejection set $R_m(\alpha)$ of §D.2 or the rejection sets $\widehat{R}_\bullet$ of §D.6, and $w$ is neither the pair gap $g$ of §D.4 nor the conditional power $g_m$ of Definition 1.

**Assumption 9** (Detectability threshold)**.** A pair is *declared indistinguishable* when its observed gap is below $L$. By Theorem 17 this is the threshold at which the exact test attains power $\tau$ against a true gap of that size; it is not the threshold at which the test rejects (Proposition 37). Where $L = \mathrm{MDD}(n, d)$ is used, 6–7 are in force and $d$-monotonicity is invoked nowhere.

**Definition 4** (Indifference and dominance)**.** Index the entrants by distinct identifiers, so that two systems may share a score. Then $x \sim_L y \iff x \neq y$ and $|s_x - s_y| < L$, and $x \succ_L y \iff s_x - s_y \geq L$. $G_L = (S, \sim_L)$ is the indifference graph. Tied entrants ($s_x = s_y$, $x \neq y$) are therefore indifferent, as they must be for $G_L$ to be the incomparability graph of $\succ_L$; irreflexivity comes from $x \neq y$ rather than from excluding a zero gap.

**Lemma 27** (Semiorder)**.** *$\succ_L$ is a strict partial order and $G_L$ is its incomparability graph. On a finite $S$, $G_L$ is a unit interval graph and $(\succ_L, \sim_L)$ is a semiorder.*

*Proof.* Irreflexivity of $\sim_L$ is part of Definition 4 and that of $\succ_L$ follows from $L > 0$; if $s_x - s_y \geq L$ and $s_y - s_z \geq L$ then $s_x - s_z \geq 2L \geq L$. Incomparability of $x \neq y$ says $|s_x - s_y| < L$. For the classification, Scott–Suppes and Roberts are stated for a *strict* threshold and closed intervals, whereas Definition 4 uses $\geq L$; on a finite $S$ set $\varepsilon > 0$ below $\min\{\big||s_x - s_y| - L\big| : \text{the value is} > 0\}$ and replace $L$ by $L - \varepsilon$, which induces the same relation and satisfies the classical hypotheses. Then $I_x = [s_x - L/2, s_x + L/2)$ represents $\sim_L$. □

**Definition 5** (Tier operator)**.** $\mathcal{T}_L(S)$: set $\ell \leftarrow s_1$, open $B_1 = \{s_1\}$; for $t = 2, \ldots, M$, if $\ell - s_t \geq L$ open a new block with leader $\ell \leftarrow s_t$, else append $s_t$. Write $\mathcal{T}_L(S) = (B_1, \ldots, B_k)$, $k = k_L(S)$, leaders $\ell_1 > \cdots > \ell_k$.

The released implementation follows this rule: the eight copies of `tiers(scores, L)` in `src/` and `tiers(scores, limit_pp)` in `tool/resolution_card.py` all implement `if lead - x >= L: k += 1; lead = x`. Each score is compared with the tier *leader*.

**Lemma 28** (Well-definedness)**.** *$\mathcal{T}_L(S)$ depends only on the multiset $S$ and $L$; equal scores always share a block.*

*Proof.* The update reads only $\ell - s_t$. If $s_t = s_{t+1}$ then either $\ell - s_t < L$ and both join the current block, or $\ell - s_t \geq L$, $s_t$ becomes leader, and $\ell - s_{t+1} = 0 < L$. □

**Lemma 29** (Structure)**.** *Each $B_j$ is a clique of $G_L$ of diameter $< L$ contained in $(\ell_j - L, \ell_j]$; the leaders form a chain of $\succ_L$.*

*Proof.* Every $x \in B_j$ has $x \leq \ell_j$ and $\ell_j - x < L$ by the membership rule. For leaders, $\ell_j - \ell_{j+1} \geq L$ by construction, and for $i < j$, $\ell_i - \ell_j = \sum_{t=i}^{j-1} (\ell_t - \ell_{t+1}) \geq (j - i)L \geq L$. □

**Theorem 30** (Duality)**.** *Let $\omega(G_L)$ be the size of the largest pairwise-distinguishable set (a chain of $\succ_L$; an independent set of the indifference graph $G_L$, equivalently a clique of its complement, the significance graph, hence the symbol $\omega$) and $\kappa(G_L)$ the least number of blocks containing no distinguishable pair (a clique cover of $G_L$). Then $k_L(S) = \omega(G_L) = \kappa(G_L)$, and $k_L(S)$ is also the least number of half-open intervals of length $L$ covering $S$. $\mathcal{T}_L(S)$ attains the minimum and its leaders attain the maximum.*

*Proof.* If $C$ is a chain and $Q$ a partition into antichains, $|B \cap C| \leq 1$ for every $B \in Q$, so $|C| \leq |Q|$ and $\omega \leq \kappa$. By Lemma 29, $\mathcal{T}_L(S)$ is such a partition of size $k$ whose leaders form a chain of size $k$, so $\omega \geq k \geq \kappa$. For the cover, the $k$ intervals $(\ell_j - L, \ell_j]$ cover $S$; conversely a half-open interval of length $L$ contains at most one member of any chain. □

*Remark* 11. This construction realises Mirsky's theorem for $\succ_L$. The leader algorithm returns a cover and a chain of equal size, which certify each other's optimality. Reporting both makes the optimality of the count directly verifiable.

**Corollary 31** (Monotone in $L$)**.** *$L \mapsto k_L(S)$ is nonincreasing, so $L \in [L^-, L^+]$ gives $k_{L^+} \leq k_L \leq k_{L^-}$.*

*Proof.* A set pairwise $\geq L'$ apart with $L' \geq L$ is pairwise $\geq L$ apart, so $\omega(G_{L'}) \leq \omega(G_L)$; apply Theorem 30. □

**Definition 6.** A partition is *internally indistinguishable* (II) if no block contains a pair with gap $\geq L$, and *separated* (SEP) if every cross-block pair has gap $\geq L$.

**Proposition 32** (II, and SEP only under transitivity)**.** *$\mathcal{T}_L(S)$ is always II. It is SEP iff every connected component of $G_L$ is a clique, i.e. iff $\sim_L$ is transitive on $S$. For $S = \{0.80, 0.79, 0.78\}$, $L = 0.015$: $\mathcal{T}_L(S) = (\{0.80, 0.79\}, \{0.78\})$ and the pair $(0.79, 0.78)$ straddles a boundary while differing by $0.01 < L$.*

*Proof.* II is Lemma 29. If SEP then no edge crosses a boundary, so blocks are unions of components, and being cliques, so is each component. Conversely a clique component has diameter $< L$, so the algorithm places all its members in the block started at its maximum, and the next element is $\geq L$ below. □

**Proposition 33** (The "coarsest" question)**.** *$\mathcal{T}_L(S)$ is not the coarsest SEP partition, for two reasons: unless $\sim_L$ is transitive it is not SEP at all, and the coarsest SEP partition is the trivial one block. Among II partitions, however, $\mathcal{T}_L(S)$ attains the minimum cardinality $k_L(S)$, and it is maximal in the refinement order — no two of its blocks may be merged without violating II.*

*Proof.* A proper coarsening merges $B_i, B_j$ with $i < j$; the merged block contains $\ell_i, \ell_j$ with $\ell_i - \ell_j \geq L$, so II fails. □

**Proposition 34** (Canonical quantities)**.** *$k_L(S)$ is canonical; the minimising partition is not. For $S = \{0.80, 0.79, 0.78\}$, $L = 0.015$, both $(\{0.80, 0.79\}, \{0.78\})$ and $(\{0.80\}, \{0.79, 0.78\})$ are II of size 2 and neither refines the other. Two things are canonical: the top block $B_1 = \{x : s_1 - s_x < L\}$, the closed neighbourhood of the top-ranked system; and among II partitions into consecutive runs, $\mathcal{T}_L(S)$ is the unique one whose block starts $1 = t_1 < \cdots < t_k$ are simultaneously maximal.*

*Proof.* Let $P$ be an II run partition with starts $t_j$ and $\mathcal{T}_L$ have starts $v_j$; $t_1 = v_1 = 1$. Suppose $t_j \leq v_j$. If $t_{j+1} - 1 < v_j$ then $t_{j+1} \leq v_j \leq v_{j+1}$ which establishes the induction step. Otherwise $t_{j+1} - 1 \geq v_j$, and II gives $s_{t_j} - s_{t_{j+1}-1} < L$; since $s_{v_j} \leq s_{t_j}$ this yields $s_{v_j} - s_{t_{j+1}-1} < L$, so index $t_{j+1} - 1$ lies inside the greedy block opened at $v_j$, whence $t_{j+1} \leq v_{j+1}$. Induction gives $t_j \leq v_j$ for every $j \leq \min(|P|, k)$.

If $|P| < k$ then $t_{|P|} \le v_{|P|}$ and $P$'s last block would have to reach $s_M$ from $s_{t_{|P|}} \ge s_{v_{|P|}} > \ell_k \ge s_M$ with $k - |P| \ge 1$ further greedy boundaries strictly inside it, contradicting II. So $|P| \ge k$, with equality of the start vectors only for $\mathcal{T}_L$. □

**Proposition 35** (Single linkage is the SEP-optimal partition)**.** *Let $\mathcal{S}_L(S)$ cut at every adjacent gap $s_t - s_{t+1} \ge L$; its blocks are the connected components of $G_L$ (single-linkage clustering at threshold $L$). Then*

*(i) every SEP partition is coarser than or equal to $\mathcal{S}_L(S)$, so $\mathcal{S}_L$ is the unique finest SEP partition and* $\max\{|P| : P \text{ SEP}\} = k_L^{\mathrm{SL}}(S)$*;*

*(ii) $\mathcal{T}_L(S)$ refines $\mathcal{S}_L(S)$, so $k_L^{\mathrm{SL}}(S) \le k_L(S)$;*

*(iii)* $\max_{P \text{ SEP}} |P| = k_L^{\mathrm{SL}} \le k_L = \min_{P \text{ II}} |P|$*, with equality throughout iff $\sim_L$ is transitive, equivalently iff a partition that is simultaneously II and SEP exists.*

*Proof.* (i) SEP means every edge lies inside a block, so blocks are unions of components. (ii) At an SL cut $s_t - s_{t+1} \ge L$ the current leader $\ell \ge s_t$, so $\ell - s_{t+1} \ge L$ and $\mathcal{T}_L$ opens a block there; both partitions are into runs. (iii) Combine (i), (ii), Theorem 30 and Proposition 32. □

**Corollary 36** (Upper bound, under the $\ge L$ relation)**.** *Any partition in which systems in different tiers are guaranteed to have gap $\ge L$ has at most $k_L(S)$ blocks.*

*Proof.* One representative per block is pairwise $\ge L$ apart, i.e. a chain, so there are at most $\omega(G_L) = k_L(S)$ of them. □

**Proposition 37** (The threshold is power, not evidence)**.** *Under the normal approximation the exact test rejects at observed gap $\ge L_{\mathrm{sig}} := z_\alpha\sqrt{d/n} = L/1.42940$, whereas $L = \mathrm{MDD} = Z\sqrt{d/n}$. Every pair with observed gap in $[0.6996\,L,\ L)$ is significant at $\alpha$ and is nevertheless placed inside a tier. Consequently, by Corollary 31,*

$$k_L(S) \ \le\ k_{L_{\mathrm{sig}}}(S) = \omega(G_{L_{\mathrm{sig}}}),$$

*so under the reading "how many systems are pairwise* statistically separated*" the reported tier count is a lower bound, not an upper one. The upper-bound property of Corollary 36 holds only for the relation "gap $\ge L$".*

*Proof.* Immediate from Corollary 31 and $L_{\mathrm{sig}} < L$. □

*Empirical comparison.* On the Open LLM record-advance set, `artifacts/exact_vs_closed.json` records `closed_unres_exact_res = 5` and `exact_unres_closed_res = 0` out of 34: five advances (14.7%) are called indistinguishable by the tier threshold while the paired exact test rejects them, with no discrepancy in the reverse direction, consistent with the 1.43× factor. On HELM MMLU and SWE-bench Verified, whose tier constructions use exact $p$-values rather than $L$, few pairs show this discrepancy: no within-tier pair is significant uncorrected, and one per board is significant under Holm (8.17 accuracy points, $p_{\mathrm{Holm}} = 0.0172$ on HELM MMLU; 8.40 accuracy points, $p_{\mathrm{Holm}} = 0.049410$ on SWE-bench Verified).

**Theorem 38** (Range bounds)**.** $\dfrac{R+w}{L+w} < k_L(S) \le 1 + \left\lfloor \dfrac{R}{L} \right\rfloor$*, the upper bound attained e.g. by* $S = \{0, L, \ldots, (k-1)L\}$*. Exactly,* $R = \sum_{j=1}^{k} \mathrm{span}(B_j) + \Gamma$ *with* $\mathrm{span}(B_j) < L$ *and* $\Gamma$ *the sum of the* $k-1$ *boundary-straddling adjacent gaps.*

*Proof.* Upper: $\ell_1 - \ell_k \ge (k-1)L$ and $\ell_k \ge s_M$, so $R \ge (k-1)L$ and $k$ is an integer. Lower: if $\ell_{j+1} = s_{t+1}$ then $s_t \in B_j$, so $\ell_j - s_t < L$ and $s_t - s_{t+1} \le w$, giving $\ell_j - \ell_{j+1} < L + w$; also $\ell_k - s_M < L$, so $R < (k-1)(L+w) + L$. □

The reversed inequality $k \geq 1 + \lfloor R/L \rfloor$ is false. On the 30 measured board-by-$d$ cells (one board read at one value of $d$) the upper bound holds in every cell and is attained exactly once (Winogrande: $R = 44.278$ and $L = 4.0393$ accuracy points, $k = 11 = 1 + \lfloor 10.962 \rfloor$); the median ratio $k/(1 + \lfloor R/L \rfloor)$ is 0.757, and the lower bound is weak because $w$ is large relative to $L$ (median $w/L = 3.141$).

**Proposition 39** (Stability)**.** *For $0 < L \leq L'$ with $m = \lceil L'/L \rceil$, $\lceil k_L(S)/m \rceil \leq k_{L'}(S) \leq k_L(S)$, and the left inequality is attained. In particular an arbitrarily small relative increase in L can halve the count, so no Lipschitz bound in L exists and uncertainty must be propagated by re-evaluating at both endpoints (Corollary 31), never by linearising.*

*Proof.* Right: Corollary 31. Left: keep every $m$-th leader; consecutive kept leaders differ by $\geq mL \geq L'$, forming a $\succ_{L'}$-chain of size $\lceil k/m \rceil$; apply Theorem 30. Attainment: for $S = \{0, L, \ldots, (M-1)L\}$, $k_L = M$ and $k_{(1+\varepsilon)L} = \lceil M/2 \rceil$ for $0 < \varepsilon < 1$. □

*Propagation of numerical and rate uncertainty.* The released tool (`tool/resolution_card.py`) and the evaluator behind every limit in this paper (`src/mdd_exact.py`) evaluate the same exact sum; the Monte Carlo solver of Corollary 15 evaluates it by simulation. At all 220 stored board-by-$d$ cells, which contain the 202 boards at the reference rate and the 30 cells above, the tool and the evaluator agree to within $2.7 \times 10^{-13}$ accuracy points, and the tool's greedy tier construction reproduces every stored tier count; re-evaluated at both ends of that difference, no count changes. Uncertainty in the disagreement rate is propagated the same way, by re-evaluating each board at both ends of its measured rate distribution: the median HELM tier count ranges from 6 to 4 over the 165 boards on which every row of Appendix Table 17 is defined, and the SWE-bench Verified count ranges between 11 and 12 across $\hat{d} \pm 2$ standard errors. A tiers-per-model share divides the count by a fixed $M$, so it has the same relative uncertainty as the count.

**Proposition 40** (Heterogeneous limits)**.** *Let each pair carry $L_{xy} = \mathrm{MDD}(n, d_{xy})$ and let $\omega^*$ be the largest set pairwise distinguishable under those limits. If $L \leq \min_{x,y} L_{xy}$ then $k_L(S) \geq \omega^*$. If $L > \min_{x,y} L_{xy}$ then neither $k_L \geq \omega^*$ nor $k_L \leq \omega^*$ holds in general.*

*Proof.* If $C$ is pairwise distinguishable under the true limits then any $x, y \in C$ have $|s_x - s_y| \geq L_{xy} \geq L$, so $C$ is a $\succ_L$-chain and $|C| \leq \omega(G_L) = k_L$. For failure in one direction, place two systems a distance $\delta$ apart with $L_{xy} < \delta < L$: distinguishable, but one tier, so $k_L < \omega^*$. For the other, three systems at 0.70, 0.67 and 0.64 with $L = 0.025$ and pair limits $L_{12} = 0.022$, $L_{23} = 0.060$, $L_{13} = 0.063$: the minimum limit is below $L$, as the premise requires; the operator opens a tier at each step, $k_L = 3$; but only the first pair is distinguishable under the pairwise limits, so $\omega^* = 2 < k_L$. Those limits are realisable: at $n = 1{,}000$ they are $\mathrm{MDD}(n, d_{xy})$ for pairwise rates 0.06, 0.44 and 0.49, and one joint table of outcomes produces all three together with the three accuracies. Writing each item's outcomes for systems 1, 2, 3 as a triple, the counts $(1,1,1)$:420, $(1,1,0)$:235, $(1,0,0)$:40, $(1,0,1)$:5, $(0,1,1)$:15, $(0,1,0)$:0, $(0,0,1)$:200, $(0,0,0)$:85 give accuracies 700, 670, 640 and disagreement counts 60, 440, 490 in a thousand. □

The hypothesis concerns the minimum pairwise rate, not the 5th percentile of $d$, and it fails on the one board where every pair is checked. On SWE-bench Verified ($n = 500$, $M = 134$) the minimum pairwise $d$ recoverable from `artifacts/swebench.json` is 0.0140, below the 0.140 fifth percentile. The per-pair limit spans 0.239–1.701× the median-based $L$, with one pair outside the feasible region ($L_{xy} = \infty$); and at the median-based $L = 6.6309$ accuracy points, $k_L = 11$ while a greedy chain under the per-pair limits gives $\omega^* \geq 15$, and an exact maximum clique confirms $\omega^* = 15$. Three hundred and twenty-eight pairs are distinguishable per-pair but not at the uniform $L$, and none show the reverse discrepancy. Across the record advances of the Open LLM v1 boards the smallest *non-zero* pairwise $d$ is also 0.0140, on MMLU, and 0.0349 on GSM8K. Both lie below the $d = 0.05$ assumed for vision and MTEB, so $L \leq \min_{x,y} L_{xy}$ is not demonstrated there either.

**Proposition 41** (The exact-test walk has none of these guarantees)**.** *Let $x \succ_p y$ mean the paired exact test rejects. Then $\succ_p$ need not be monotone in the accuracy gap, because the conditional p-value depends on $(b, c)$ and not on $b - c$ alone; blocks produced by the walk need not be II, since membership is tested against the leader only; and the count is not in general an upper bound on $\omega(G_p)$.*

*Proof.* Take $n = 1000$ with joint counts $(A, B, C)$: $(1,1,1)$:178, $(1,1,0)$:9, $(1,0,1)$:2, $(1,0,0)$:382, $(0,1,1)$:357, $(0,1,0)$:19, $(0,0,1)$:3, $(0,0,0)$:50, so accuracies are $0.571 > 0.563 > 0.540$. Then $p_{AB} = 0.79958$, $p_{AC} = 0.273631$ and $p_{BC} = 6.61877 \times 10^{-5}$. The walk starts at $A$, absorbs $B$ and then $C$, reporting one tier, while $\{B, C\}$ is pairwise distinguishable, so $\omega(G_p) \geq 2 > 1 = k$. The same table gives non-monotonicity: the larger gap $A$ vs $C$ (3.1 accuracy points) is not significant while the smaller $B$ vs $C$ (2.3) is. □

*Remark* 12 (Verified optimality on the two boards with exact tests). The exact maximum clique size on the significance graph equals the reported tier count on *both* boards, under both readings: SWE-bench Verified 19 uncorrected and 10 Holm; HELM MMLU 11 uncorrected and 6 Holm. The HELM certificate can be verified from the released artifacts: `artifacts/helm_instances.json` contains the full $91 \times 514$ per-instance binary matrix, and recomputing all 4,095 $p$-values from it reproduces every stored summary exactly ($n_{\text{pairs}} = 4095$; discordance median 0.254864, p5 0.169261, p95 0.542802). Non-monotonicity occurs on both boards: largest non-significant gap 5.64 against smallest significant 3.31 accuracy points on HELM (11.67 vs 7.20 under Holm), and 4.60 vs 2.40 accuracy points on SWE-bench (10.40 vs 4.00 under Holm). These counts are *verified optimal* on the two boards, not guaranteed in general; Corollary 36 holds only for the relation "gap $\geq L$", which the closed-form families use.

**Theorem 42** (The floor as a tier construction)**.** *Assume 1–4 (so that, unconditionally on the discordant count, $M_n \sim \text{Bin}(n, d)$). At $d = |\Delta|$ we have $c = 0$, the two-sided exact conditional $p$-value is $p = \min\{1, 2^{1-m}\}$, the test rejects iff $m \geq m_\alpha = 6$, and the power is exactly $\Pr[\text{Bin}(n, d) \geq 6]$. Hence the floor at this boundary configuration is $L^*(n) = \Delta_{\text{fl}}(n)$ of Theorem 20: at $d = |\Delta|$ a pair is resolvable only if $|\Delta| \geq L^*(n)$, and this holds unconditionally. The stronger reading, that a pair with $|\Delta| < L^*(n)$ is resolvable at* no *admissible $d$, is Open question 26: it is assumed only for $n > m_\alpha$, where an exhaustive scan finds no exception, and it is false at $n = m_\alpha = 6$. The rate-free partition is $\mathcal{T}_{L^*(n)}(S)$, to which Theorem 30 and everything after it applies verbatim; calling it a floor over every admissible rate, rather than at $d = |\Delta|$, is conditional on that open question.*

*Proof.* With $c = 0$ the conditional law of $b$ given $m$ is degenerate at $m$, and $2\min\{F(m; m, \frac{1}{2}), 1 - F(m-1; m, \frac{1}{2})\} = 2^{1-m}$ before truncation; $2^{1-m} < 0.05 \iff m \geq 6$. Unconditionally $m \sim \text{Bin}(n, d)$, so the power is $\Pr[m \geq 6]$; the rest is Theorem 20. The necessity at $d = g$ is Theorem 20(i)⇔(iv). Its extension to *every* admissible $d$ is not proved here: it is Open question 26, which requires $d \mapsto \pi(n, g, d)$ to be maximised at $d = g$ below the floor, and the counterexample given there shows the hypothesis $n > m_\alpha$ cannot be dropped. □

*Remark* 13 (Dependence of the floor constant on the sampling model). A random $m$ corresponds to the unconditional trinomial model that `power()` samples. Under the conditional convention in which "at disagreement rate $d$" means $m$ fixed at $dn$, rejection at $c = 0$ is deterministic once $m \geq 6$ and the floor is exactly $6/n$, not $\lambda^*/n$ — approximately six rather than eight items, a 32% difference. Assumption 5 does not determine this choice; Assumptions 1–3 specify the sampling model and hence the constant. The floor is independent of the assumed value of $d$, but still depends on these sampling assumptions. The constant is also a function of the operating point: at 50% power, $\lambda^* \approx 5.67$; under a multiplicity correction $m_\alpha$ rises (to 18 at $\alpha/4095$ and 19 at $\alpha/8911$) and the floor rises with it.

### D.6 Multiplicity

Two symbols are local to this subsection. $\theta_k$ is system $k$'s accuracy, always subscripted by a system; the unsubscripted $\theta$ of §D.2 remains the direction ratio $(d + \Delta)/(2d)$ of Definition 1. The symbol $M$ denotes the number of systems on the board, not the discordant count $M_n$ of §D.1.

A board of $M$ systems generates $N = \binom{M}{2}$ pairwise statements. Ties are not excluded by the construction: HELM MMLU has 29 tied pairs and SWE-bench Verified 36, all with $\widehat{\Delta} = 0$, hence $p = 1$, hence counted in the "unsupported" numerator — 3.2% and 3.3% of it respectively. At those pairs the board asserts no strict ordering. The shares reported here count them as unsupported; excluding them gives $886/4066 = 21.79\%$ and $1040/8875 = 11.72\%$.

**Theorem 43** (Holm is valid under arbitrary dependence)**.** *Let $H_1, \ldots, H_N$ have $p$-values that are marginally super-uniform under their own nulls. The step-down procedure rejecting $H_{(i)}$ while $p_{(j)} \leq \alpha/(N - j + 1)$ for all $j \leq i$ controls FWER at $\alpha$, whatever the joint law.*

*Proof.* Let $T$ be the set of true nulls, $t = |T|$, and suppose a false rejection occurs. Let $i^*$ be the first step at which a member of $T$ is rejected; at that step at most $i^* - 1$ of the already-rejected hypotheses are false, so $N - i^* + 1 \geq t$, and the rejected true null has $p \leq \alpha/(N - i^* + 1) \leq \alpha/t$. Hence $\Pr(\text{any false rejection}) \leq \Pr(\min_{h \in T} p_h \leq \alpha/t) \leq t \cdot \alpha/t = \alpha$ by the union bound. □

**Theorem 44** (Shaffer's logical constraint gives a bounded relaxation)**.** *Equality of accuracies is an equivalence relation, so the set of true nulls is determined by a partition of the $M$ systems and the achievable true-null counts are $\mathcal{A}(M) = \{\sum_g \binom{m_g}{2}\}$. The step-$i$ divisor may be replaced by $t_i = \max\{t \in \mathcal{A}(M) : t \leq N - i + 1\}$, and for $2 \leq i \leq M$ the resulting threshold relaxation is at most $(M + 1)/(M - 1)$, attained at $i = 2$ where $t_2 = \binom{M-1}{2}$. For $i > M$ the same maximum was verified by exhaustive enumeration of $\mathcal{A}(M)$ on the two boards tested ($M = 91$ and $134$); it is not proved in general. Here $M \geq 3$ (at $M = 2$, $t_2 = 0$), and at $i = 1$ the divisor is $t_1 = N$, so that step is not relaxed at all.*

*Proof.* Validity is Theorem 43's proof with $t_i$ in place of $N - i + 1$, since the true-null count is always in $\mathcal{A}(M)$. For the bound, merging two groups of sizes $a, b$ into one of size $a + b - 1$ plus a singleton changes the count by $\binom{a+b-1}{2} - \binom{a}{2} - \binom{b}{2} = (a-1)(b-1) \geq 0$, so iterating from any $r \geq 2$ groups reaches $(M - r + 1, 1^{r-1})$, whose count $\binom{M-r+1}{2} \leq \binom{M-1}{2}$; hence $\binom{M-1}{2}$ is the largest achievable count strictly below $N$. Since $N - \binom{M-1}{2} = M - 1$, the value $\binom{M-1}{2}$ is admissible exactly for $2 \leq i \leq M$, and the ratio is maximised at $i = 2$: $(N - 1)/\binom{M-1}{2} = (M^2 - M - 2)/((M - 1)(M - 2)) = (M + 1)/(M - 1)$. □

*Empirical comparison.* The threshold relaxation is 1.0222 at $M = 91$ and 1.0150 at $M = 134$, with zero additional rejections on either board. Shaffer S1 and Holm give identical counts (1,922 and 2,295). An exhaustive dynamic program over $\mathcal{A}(M)$ confirms on both boards that no step $i > M$ relaxes the threshold by more than the ratio at $i = 2$. This conclusion remains unproved in general. The theorem bounds the threshold relaxation; the change in the count also depends on the atomic $p$-values.

**Lemma 45** (Equicorrelation under exchangeability)**.** *Suppose the per-item outcome vectors satisfy $\mathrm{Var}(X_{k1}) = v$ and $\mathrm{Cov}(X_{k1}, X_{l1}) = c$ for all $k \neq l$ (compound symmetry; $c = 0$ is not required). Then for the gap statistics $\widehat{\Delta}_{kl} = \overline{X_k} - \overline{X_l}$,*

$$\mathrm{Corr}(\widehat{\Delta}_{kl}, \widehat{\Delta}_{kj}) = +\tfrac{1}{2}, \qquad \mathrm{Corr}(\widehat{\Delta}_{kl}, \widehat{\Delta}_{lj}) = -\tfrac{1}{2}, \qquad \mathrm{Corr}(\widehat{\Delta}_{kl}, \widehat{\Delta}_{uv}) = 0$$

*for distinct indices, i.e. classical all-pairwise equicorrelation, with the sign set by the orientation of the shared system.*

*Proof.* $\mathrm{Cov}(X_k - X_l, X_k - X_j) = v - c - c + c = v - c$ and $\mathrm{Var}(X_k - X_l) = 2(v - c)$, giving $+\frac{1}{2}$; reversing the role of the shared system negates the numerator; and for disjoint pairs the four covariances cancel, $c - c - c + c = 0$. □

*Empirical correlations.* On the $91 \times 514$ HELM matrix, the values are +0.4944 (shared system, same position), −0.4956 (shared system, opposite position), +0.0036 (disjoint), each with s.d. $\approx 0.11$. The residual departure from $\pm\frac{1}{2}$ is explained by non-common variances (per-model Bernoulli variance 0.1543–0.2500, ratio 1.62), not by non-zero $c$: the between-model item-level correlation has mean +0.4201 and is positive for 98.6% of pairs, so $c = 0$ does not describe these data. The lemma allows nonzero covariance.

**Proposition 46** (The uncorrected count is a floor, for non-randomised procedures)**.** *Among procedures that reject $H_h$ only when $p_h \leq c_h(p)$ with critical constants $c_h \leq \alpha$, applied to these $p$-values, the number of non-rejections is at least the uncorrected count:* 22.34% *and* 12.07%.

*Proof.* Any such rejection set is contained in $\{h : p_h \leq \alpha\}$. □

The bound does not extend to every valid procedure. The exact conditional test has size $\alpha_m < \alpha$ strictly at every $m$ (Corollary 4), and the randomised symmetric test of exact size $\alpha$ — the same object that is UMPU in Proposition 25(a) — has an exactly uniform null $p$-value, so Theorem 43 applies unchanged, while its rejection set strictly contains $\{p \leq \alpha\}$. Empirically, randomisation reduces the per-comparison non-rejection rates to 21.70% and 11.73% (66 HELM pairs with $p > \alpha$ rejected with

mean probability 0.401); we report the non-randomised test, whose rejections do not depend on an auxiliary random draw.

**Proposition 47** (BH is bracketed; BY is bracketed only on one side)**.** *$\widehat{R}_{\mathrm{Holm}} \subseteq \widehat{R}_{\mathrm{BH}} \subseteq \{p \le \alpha\}$, so the BH non-rejection share lies between the uncorrected and Holm shares. For BY, $\widehat{R}_{\mathrm{BY}} \subseteq \{p \le \alpha\}$ holds, but $\widehat{R}_{\mathrm{Holm}} \subseteq \widehat{R}_{\mathrm{BY}}$ does not.*

*Proof.* For BH, $i(N-i+1) \ge N$ on $1 \le i \le N$, so $\alpha/(N-i+1) \le i\alpha/N$; if Holm rejects through step $i^*$ then $p_{(i^*)} \le i^*\alpha/N$ and BH's step-up index is at least $i^*$. For BY, $\max_i iq/(NH_N) = q/H_N < \alpha$ gives the containment in $\{p \le \alpha\}$. Counterexample for the other side: $N = 2$, $p = (0.02, 0.9)$, $q = \alpha = 0.05$, $H_2 = 1.5$. Holm rejects $H_{(1)}$ since $0.02 \le 0.025$; BY's constants are $(0.0166\ldots, 0.0333\ldots)$ and it rejects nothing. □

**Proposition 48** (The equality family is the wrong family)**.** *At any parameter with $\theta_k \ne \theta_l$ for every pair — true of any two real systems — every null in the equality family is false, so* FWER = FDR = PCER = 0 *at that parameter and the correction controls an error that does not occur there. The error that* can *occur is directional. Restate the family as $\{H^{\le}_{kl} : \theta_k \le \theta_l\} \cup \{H^{\ge}_{kl} : \theta_k \ge \theta_l\}$; exactly one member of each pair is true.*

The three rates are defined by suprema over the parameter space. The proposition is therefore a pointwise statement and does not imply that the rates vanish identically. Equality remains the least-favourable boundary of the directional null and determines the directional error bound.

**Lemma 49** (Per-comparison directional level)**.** *For each pair, the exact conditional test read directionally (reject $\theta_k \le \theta_l$ when the two-sided $p < \alpha$ and $b > c$) has directional error at most $\alpha/2$, conditionally on $m$ and hence unconditionally.*

*Proof.* On $\{b > c\}$ the two-sided $p$ equals $2\Pr_{1/2}(B \ge b)$, so the event $\{p < \alpha,\ b > c\}$ is $\{B \ge b_\alpha\}$ for a threshold with $\Pr_{1/2}(B \ge b_\alpha) < \alpha/2$. Here $\theta$ is the direction ratio of Definition 1 and not a system accuracy. Under $\theta \le \frac{1}{2}$ the $\mathrm{Bin}(m, \theta)$ upper tail is nonincreasing in decreasing $\theta$, so the probability is at most $\alpha/2$ for every $\theta \le \frac{1}{2}$; average over $m$. □

**Proposition 50** (Dependence-free directional FWER)**.** *Fix a configuration at which $\theta_k \ne \theta_l$ for every compared pair — the restriction of Proposition 48, and true of any two real systems. Rejecting directionally whenever the one-sided exact $p \le \alpha/N$ — i.e. the two-sided $p \le 2\alpha/N$ — controls the directional FWER at $\alpha$ at every such configuration, under any joint dependence among the $N$ comparisons.*

*Proof.* At a configuration with $\theta_k \ne \theta_l$ for every pair, exactly one of $\{H^{\le}_{kl}, H^{\ge}_{kl}\}$ is true for each pair, so the true-null count is $N$. By Lemma 49 each comparison's directional error at threshold $\alpha/N$ is at most $\alpha/N$; the union bound over $N$ comparisons gives $\alpha$ without any assumption on their joint law. □

*Scope of the untied-configuration restriction.* At $\theta_k = \theta_l$ both $H^{\le}_{kl}$ and $H^{\ge}_{kl}$ hold, so that pair contributes two true nulls rather than one; the procedure rejects in one direction or the other with probability at most $2\alpha/N$, and a union bound over $N$ tied pairs bounds the directional FWER by $2\alpha$ rather than $\alpha$. The result controls the probability of asserting an ordering opposite to the true one on the untied set. At $\theta_k = \theta_l$ no such ordering exists. Equality remains the least-favourable boundary of each one-sided null, the boundary at which Lemma 49 attains $\alpha/2$; it enters the bound through the parameter-space boundary. The stated error criterion concerns reversals of a strict population ordering. *Empirical comparison.* The non-rejection shares are $1{,}923/4095 = 46.96\%$ and $2{,}377/8911 = 26.67\%$, differing by one pair and 82 pairs from Holm's 46.94% and 25.75%. The directional formulation increases the non-rejection share by 0.02 and 0.92 percentage points.

*Remark* 14 (Observed non-rejection rates). Share of pairs not significant (%), HELM MMLU / SWE-bench Verified: PCER 22.344 / 12.075; randomised PCER 21.70 / 11.73; BH $q$=.05 23.419 / 12.378; BY $q$=.05 32.918 / 17.394; Holm 46.935 / 25.755; Shaffer S1 46.935 / 25.755; Bonferroni 48.596 / 27.427; dependence-free directional 46.96 / 26.67. The rejection set is already transitively closed — 0 intransitive triples of 121,485 and 392,084, uncorrected and under Holm — so applying transitive closure does not change the rejection sets here. We use BY for FDR control because it is valid under arbitrary dependence.

**Proposition 51** (Threshold relaxation from dependence and discreteness)**.** *Against Bonferroni, the achievable single-step threshold relaxation on HELM MMLU is* 1.568× *from discreteness alone (exact-atom min-P with dependence removed),* 1.663× *from Gaussian equicorrelated dependence alone, and* 2.181× *from both, the last measured by an item-wise permutation Westfall–Young min-P reference (*$B = 4{,}000$*) that assumes neither normality, nor homoscedasticity across pairs, nor continuity. At the permutation threshold the single-step non-rejection count is* 1,921 = 46.911%*, which* matches or improves on *Holm's* 1,922 = 46.935% *(Monte Carlo bracket* $[1912, 1924]$*), and the step-down permutation lower bound is* 45.52%*.*

These comparisons have three implications. (a) The results do not establish an advantage for step-down adjustment over accounting for dependence: single-step permutation max-$T$ is indistinguishable from Holm. (b) A Gaussian studentized-range reference gives 47.40% but requires three approximations: it uses a common standardiser across pairs, but $\widehat{d}$ spans 0.000–0.646 on HELM and 0.014–0.788 on SWE-bench, a $\sim 50\times$ spread in $\mathrm{Var}(\widehat{\Delta})$; it needs a common per-model variance (ratio 1.62); and it applies a continuous threshold $2.03 \times 10^{-5}$ to atomic $p$-values. The permutation min-$P$ reference of Proposition 51 needs none of the three, so it is the reference used here. (c) Tarone's discrete sharpening reduces the family to $K = 4094$ and $K = 8905$ but, as a discrete-*Bonferroni* procedure, gives 0 and 0 additional rejections; Tarone *within* Holm, a distinct procedure, yields 0 and 4. Thus, this Tarone procedure gives no additional rejections from discreteness, whereas the min-$P$ reference does.

**Lemma 52** (Bystander invariance)**.** *The per-comparison share is invariant, in expectation, to the number of systems on the board: for a uniformly drawn sub-board of* $M'$ *systems, exchangeability of the draw gives* $\mathbb{E}[\text{PCER share}] = 22.3443\%$ *exactly for every* $M'$*. The Holm share is not: it rises monotonically with* $M'$*, measured at* 27.15%*,* 33.46%*,* 38.85%*,* 43.12%*,* 45.38% *at* $M' = 5, 10, 20, 40, 60$ *against* 46.94% *at* $M' = 91$ *(*20,000 *draws; s.e.m.* $\leq 0.13$*).*

**Proposition 53** (The plug-in bias has no determined sign)**.** *Let* $\widehat{\eta}$ *be the share of pairs declared resolvable by comparing* $|\widehat{\Delta}_{kl}|$ *to* $\mathrm{MDD}(n, \widehat{d}_{kl}, \alpha)$*. Then* $\mathbb{E}|\widehat{\Delta}| \geq |\Delta|$ *(Jensen) and* $n \operatorname{Var} \widehat{\Delta} = d - \Delta^2$*; and* $\mathbb{E}[\mathrm{MDD}_{\mathrm{alg}}(n, \widehat{d})] \leq \mathrm{MDD}_{\mathrm{alg}}(n, d)$ *by Jensen, since the algebraic surrogate* $Z\sqrt{d/n}$ *is concave in* $d$*. For the exact limit the same inequality is not merely unproved but false:* $\mathrm{MDD}(n, d) = +\infty$ *whenever* $\Pr(\mathrm{Bin}(n, d) \geq m_\alpha) < \tau$ *(Theorem 20, Proposition 23(ii)), and* $\widehat{d} = M_n/n$ *falls in that region with positive probability, so* $\mathbb{E}[\mathrm{MDD}(n, \widehat{d})] = +\infty$*. For the exact test the sign of the plug-in term therefore rests on the bootstrap below. Neither yields a sign for the bias of* $\widehat{\eta}$*—so the sign is a measurement. A parametric bootstrap at each pair's fitted* $(\widehat{\Delta}, \widehat{d})$*, on both boards that publish per-item outcomes, settles it. On the statistic this proposition defines—the share with* $|\widehat{\Delta}| \geq \mathrm{MDD}(n, \widehat{d})$*—the bias is* $-0.31$ *percentage points on HELM MMLU (*31.50% *observed against* 31.19% *resampled) and* $-0.11$ *on SWE-bench Verified (*17.73% *against* 17.62%*). On the share of pairs the exact test does not reject, which is a different statistic and the one the survey tables report, it is* $+0.02$ *and* $-0.13$*. All four magnitudes are an order of magnitude below the effects the manuscript reports. A negative bias has opposite consequences for the two statistics. For* $\widehat{\eta}$ *it understates the supported share and so overstates the complementary unsupported share, which the manuscript claims is large; the overstatement is anti-conservative for that claim. Both measurements,* $-0.31$ *and* $-0.11$ *percentage points, carry that sign. The share of pairs the exact test does not reject is reported directly, so a negative bias understates it and is conservative for that claim;* $-0.13$ *on SWE-bench Verified carries that sign and* $+0.02$ *on HELM MMLU the opposite one. No single direction is claimed for the pair, since the two statistics do not move together.*

*Proof.* The direction of bias in a mean does not determine the direction of bias in an indicator. Writing the bias as $N^{-1} \sum_{k<l} [\Pr(|\widehat{\Delta}| \geq L_{kl}) - \mathbf{1}\{|\Delta| \geq L_{kl}\}]$, the summands have both signs: pairs just below the threshold contribute $\approx +\frac{1}{2}$, pairs just above $\approx -\frac{1}{2}$, and pairs far above contribute negatively, where $L_{kl} := \mathrm{MDD}(n, d_{kl}, \alpha)$ is the pair's limit. If every pair has $|\Delta| = 1.5 L_{kl}$ then $\eta = 1$ but $\Pr(|\widehat{\Delta}| \geq L_{kl}) < 1$, so $\mathbb{E}\widehat{\eta} < \eta$. □

*Remark* 15 (The simultaneous detection limit). Applying multiplicity correction to the limit changes the required gap: $\mathrm{MDD}(514, 0.2549)$ moves from 6.38 to 11.52 accuracy points at $\alpha/4095$, and $\mathrm{MDD}(500, 0.268)$ from 6.63 to 12.37 at $\alpha/8911$ — a factor of 1.81 and 1.87. Recomputing the closed-form families at their own $\alpha/\binom{M}{2}$ does not provide a common basis for comparing all eight families, because MDD ceases to exist where $\Pr(\mathrm{Bin}(n, d) \geq m_\alpha) < \tau$ with $m_\alpha = \lfloor \log_2(2/\alpha') \rfloor + 1$:

at $\alpha/\binom{92}{2}$, GAIA ($n = 165$) has $m_\alpha = 18$ and Pr = 0.0016, so no finite limit exists; and 24 of the 182 distinct HELM boards (13.2%) have no finite limit at $d = 0.2549$, rising to 116 of 182 (63.7%) at the $d = 0.05$ assumed for vision and MTEB. A single common corrected level, or an explicit attrition table, is required.

### D.7 Ranking robustness to item reweighting

A leaderboard based on the unweighted mean reports the ordering under uniform item weights. Let $w$ range over the simplex on the $n$ items and let $c_i = x_i - y_i \in \{-1, 0, 1\}$ be the per-item difference between two systems, with class masses $(d + \Delta)/2$, $(d - \Delta)/2$ and $1 - d$ for observed gap $\Delta > 0$ and disagreement rate $d$. The pair reverses under $w$ when $\sum_i w_i c_i \le 0$.

**Proposition 54** (Breakdown distance)**.** *For $0 < \Delta < d$ the smallest* $\mathrm{KL}(w \,\|\, \mathrm{unif})$ *over weightings that reverse the pair is*

$$\varepsilon^* = -\ln\bigl(1 - d + \sqrt{d^2 - \Delta^2}\bigr), \qquad \textit{and} \qquad \varepsilon^* = \frac{\Delta^2}{2d} + O\bigl((\Delta/d)^4\bigr).$$

*Proof.* The minimiser of KL subject to a linear constraint on a mean is the exponential tilt $w_i \propto e^{-\lambda c_i}$, with normaliser $\zeta(\lambda) = 1 - d + \frac{d+\Delta}{2} e^{-\lambda} + \frac{d-\Delta}{2} e^{\lambda}$. Setting $\zeta'(\lambda) = 0$ gives $e^{2\lambda} = (d + \Delta)/(d - \Delta)$ and $\zeta = 1 - d + \sqrt{d^2 - \Delta^2}$. The constraint binds at the optimum, so the attained divergence is $-\ln \zeta$. Expanding $\sqrt{d^2 - \Delta^2} = d - \Delta^2/(2d) + O(\Delta^4/d^3)$ gives the second form. □

The construction is the $I$-projection of the uniform weights onto a mean constraint: $w^*$ is the implied probability of empirical likelihood (Owen, 1988) and $\varepsilon^*$ the Cramér rate for $c$, so under multinomial resampling of the items Pr(reversal) $= e^{-m\varepsilon^*(1+o(1))}$ at $m$ draws (Dembo & Zeitouni, 2010). The following results connect this divergence to sample-size requirements and item-removal robustness.

**Corollary 55** (Reciprocal of the item requirement)**.** *Under the closed form, $n = Z^2(d - \Delta^2)/\Delta^2$ items are needed to resolve the pair, so $\varepsilon^* n \to Z^2/2 = 3.9244$ as $\Delta/d \to 0$, with $Z$ the constant of §D.3. Measured on the pairs with $\Delta/d < 0.1$, the median of $\varepsilon^* n$ is* 3.9248–3.9263 *on all seven boards that release per-item outcomes.*

**Proposition 56** (Dropping items returns the gap)**.** *Reversing the pair by* removing *$k$ items the leader answered and the trailer missed leaves the gap at $(n\Delta - k)/(n - k)$, which vanishes at $k = n\Delta$. The smallest removable fraction is therefore exactly $\Delta$, whatever $d$ is.*

A robustness metric of that form (Broderick et al., 2020), which on preference data locates the comparisons responsible for a reversal (Huang et al., 2025), therefore returns the accuracy gap on an accuracy benchmark, providing no additional information. Over the 37,292 pairs on the seven boards the smallest removable fraction equals the gap at machine precision, with largest absolute difference 0. Continuous reweighting distinguishes pairs with similar gaps: among the 82 HellaSwag pairs whose gap is $2.0 \pm 0.1$ accuracy points, the disagreement rate runs 0.034 to 0.339 and $\varepsilon^*$ spans a factor of 13.0, so pairs with similar item-removal thresholds differ in their robustness to reweighting. This quantity is independent of the item count and is computed from $(d, \Delta)$ alone (`src/mx39_breakdown.py`).

### D.8 Unproved results and limits of the assumptions

*Relevance to the reported results.* The four open questions below are not required for the main-text figures. The principal results for the three families with per-item outcomes—disagreement rates and their variation, transfer, ranking reversals, and tier counts—use exhaustive exact testing rather than the approximations addressed by the conjectures. Failure of the conjectures would affect the accuracy of the *closed form* for the five families whose disagreement rates are assumed. Corresponding sample-size targets and survey-wide counts would then require exact evaluation. Where reported together, the exact and approximate requirements are 3,424 and 3,336 items for a gap of two accuracy points, a 2.6% difference. Assumptions 3–4 underlie every closed-form result, so we consider their implications first.

Item independence and exchangeability (3–4) cannot be established from the observed data alone. Under E1 they follow from i.i.d. sampling (1–2), which curated benchmarks do not satisfy exactly.

Items are deduplicated, screened for difficulty, and, in SWE-bench Verified, filtered by human review. These procedures do not constitute i.i.d. sampling from $Q$. HELM's MMLU also fixes the allocation at 100/100/100/114/100 items across five subjects, whose mean difficulty lies in 0.4381–0.8600; this is a stratified design. Simulating the fixed-allocation design at the measured per-stratum $(n_s, d_s)$ under four global nulls, including the extreme feasible cross-stratum reversal, gives type-I error 0.0381–0.0412, compared with 0.0406 for the i.i.d. design. For two strata, validity follows because $\mathrm{Var}(m_1 \mid m) \le m/4$ makes the conditional law under-dispersed relative to $\mathrm{Bin}(m, \frac{1}{2})$. The estimand nevertheless changes: $\theta_k$ is population accuracy for these five subjects at the stated weights, rather than accuracy on MMLU as a whole or the fixed 514 items. Removing one subject reverses 50–167 pairwise orderings (1.2–4.1%). Item clustering also violates 3 in both analyses (Remark 1). The direction of its effect requires a dependence model and is especially relevant to the floor, which depends on an extreme lower tail. At $\rho = 0.05$ on Open LLM MMLU, clustering reverses the floor correction discussed in Remark 9.

Uncertainty and heterogeneity in disagreement rates, excluded by Assumptions 6–7, can exceed the approximation errors bounded in this appendix. Because MDD $\propto \sqrt{d}$, a relative error $\epsilon_d$ in $d$ is $\epsilon_d/2$ in MDD. At the reference point $(n, d) = (514, 0.2549)$ the *relative* sampling s.e. of $\widehat{d} = M_n/n$ on the same $n$ items is $\sqrt{(1-d)/(nd)} = 7.54\%$, hence 3.77% in MDD — larger than $|\mathcal{E}| = 2.23\%$ there, before any heterogeneity. The rate $d$ also varies across pairs: the 5th–95th range is 0.169–0.543 on HELM MMLU and 0.140–0.561 on SWE-bench, with a minimum of 0.0140; $d$ also depends on *which* models are compared (MMLU gives 0.38 on one population and 0.25 on another), so the constant-rate assumption is contradicted by the measured variation.

Four results remain open, with their missing steps identified explicitly. The sufficiency of $\lambda^*/d$ is proved in Proposition 22 using an inequality of Anderson & Samuels (1967). The remaining questions are: Open question 7 (validity of the exact test under a heterogeneous composite null, for which the variance identity does not control the tails); Open question 9 (the two-term expansion — the $\alpha$-averaging Jensen defect, the lattice term, and the size-to-power equivalence are all uncontrolled at the retained order); Open question 13 (the explicit non-asymptotic bound — one substitution error is not bounded, the averaging constant is understated by 1.530, and $K_0$ is $d$-dependent); and Open question 26 (the floor holds at every admissible $d$ — false at $n = m_\alpha$, verified but unproved for $n > m_\alpha$).

Two proved properties limit the interpretation of tiers. Their threshold specifies target power rather than observed evidence (Proposition 37). In addition, $\mathcal{T}_L$ does not satisfy SEP: a boundary guarantees separation of the two tier leaders, not of every pair across the boundary. Among the 280 board-by-$d$ cells with an adjacent-indistinguishable share, 80.72% of boundaries (2,186 of 2,708) separate adjacent systems whose gap is below the threshold. The median ratio $k_L/k_L^{\mathrm{SL}}$ is 3.0, and 112 of 280 cells form a single separated block. Proposition 57 recovers the separated count from the quantities already reported in the tables.

**Proposition 57** (The separated count is already reported)**.** *Let $a = \#\{t : s_t - s_{t+1} < L\}/(M-1)$ be the reported share of adjacent positions that are indistinguishable. Then $k_L^{\mathrm{SL}}(S) = M - a(M-1)$.*

*Proof.* $k^{\mathrm{SL}} = 1 + \#\{t : s_t - s_{t+1} \ge L\} = 1 + (M-1)(1-a)$. □