PRACTICAL CERTIFICATION OF BINARY-SCORE LEADERBOARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Leaderboards report orderings between models, rarely with a guarantee that the reported directions are right. We study certification of the population ranking of a fixed set of models scored correct or incorrect on shared, independently sampled items. We derive instance-specific lower bounds on the number of items any procedure needs to certify the ranking: one for a single comparison, given by an exact finite sum, and a mixture bound that prices several comparisons at once without assumptions on their dependence. We then show that testing all K pairs of M models with McNemar's test at $\delta/(M-1)$ in place of δ/K, and releasing the order only when all pass, certifies a complete ranking with error at most δ. On eighteen configurations from five benchmark families, resampling each benchmark's own items, this certificate needs 1.61 to 1.81 times the larger lower bound to certify the complete ranking correctly 95% of the time, and McNemar's test on the hardest single comparison alone is within 2% of its bound.

1 INTRODUCTION

A leaderboard prints an order (Liang et al., 2022; Srivastava et al., 2022), readers treat that order as settled, and the models near each other in it are the ones people argue about. The ordering between adjacent models is usually the claim readers act on, and it is rarely the claim a leaderboard has tested.

The gap between being right and proving it is large. Under the empirical law of the MMLU bank, with twenty models, the order of the sample means is entirely correct in 95% of runs at roughly $17{,}600$ items per model, yet the lower bound of §3 shows that no procedure can certify that order 95% of the time with fewer than $42{,}607$ items per model.

The gap comes either from a wasteful procedure or from the comparison itself, and telling these apart needs a number that belongs to the benchmark, not to a procedure.

What this paper does. We derive that number and measure how close a certificate comes to it (Figure 1). The number is an instance-specific lower bound (Theorem 1), built from a second item law that reverses one comparison while keeping the conditional law of the whole score vector, so it survives the dependence between comparisons and an exact finite sum evaluates it at every sample size. Mixing such alternatives for several comparisons (Theorem 2) raises it by up to 41.0%.

The certificate applies McNemar's exact test to the items where two models disagree (Theorem 3); empirical Bernstein needs 1.36 to 1.57 times as many items. A certified order with an inversion must reverse one of the $M-1$ pairs adjacent in the true order, so testing every pair at $\delta/(M-1)$ in place of δ/K suffices under any dependence (Theorem 4), which is worth 18.3% to 23.8% of the items. Applying Holm's step-down procedure directly to these data-selected one-sided p-values does not control directional errors, and we give an i.i.d. instance where that error exceeds 0.05 (Proposition 5).

On eighteen configurations across five benchmark families, the certificate needs 1.61 to 1.81 times the stronger lower bound to certify the complete ranking correctly 95% of the time, under each bank's own empirical law. On each configuration's hardest pair the per-pair test is within 1.9% of its bound, so the remaining gap concerns simultaneous certification; a proposal–validation variant recovers 3.3% to 9.9% of the items on every twenty-model configuration. With the models chosen on 30% of each benchmark, no complete ranking closes on the other 70%, but the uniform-budget certificate certifies 40% to 98% of the relations there.

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.

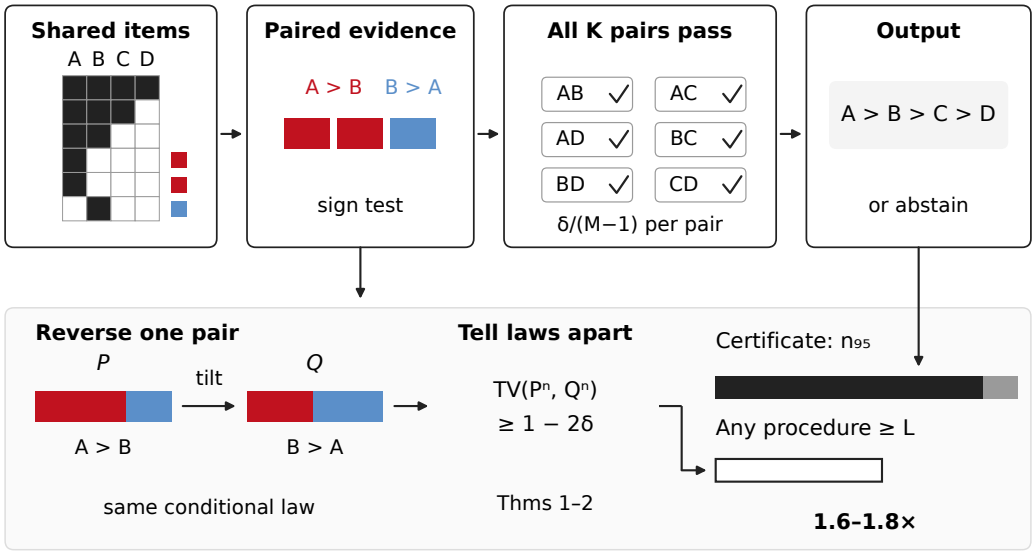

Figure 1: **Overview** (schematic). Top: disagreements between two models give an exact sign test (Theorem 3), and the complete order is released only if all K pairs pass at $\delta/(M-1)$ (Theorem 4). Bottom: tilting the law of one pair's disagreements (P) to a nearby law that reverses the pair (Q), with every other conditional law kept, bounds the items any procedure needs (Theorem 1; Theorem 2 mixes several tilts). The certificate needs 1.6 to 1.8 times the larger bound (L), the grey end showing the range over configurations. Partial orders use a separate rule (§2.4).

Contributions.

- Instance-specific lower bounds on the items any certification procedure needs (§3), for one comparison and for several at once, with no assumption on how comparisons depend, and a worst-case theorem showing the multiplicity cost cannot be removed in general.
- A ranking certificate built on McNemar's exact test that pays $\delta/(M-1)$ in place of δ/K (§4), and a counterexample to Holm applied to data-selected one-sided p-values.
- The measurement joining the two (§5): 1.61 to 1.81 times the stronger bound on eighteen replay configurations, and certificates on held-out items.

2 What a leaderboard certificate is

This section fixes what a certificate guarantees, which ranking it concerns and how its cost is counted; Appendix A collects the notation. We use one example throughout: twenty open language models ranked on MMLU by exact-match accuracy, whose accuracies run from 0.46 to 0.66, so that neighbouring models in that list differ by about one accuracy point.

2.1 What the ranking is a ranking of

A benchmark holds a *bank* of items, and MMLU's bank holds 14,042 questions. We treat that bank as a sample from a larger law G over items, so the ranking is a ranking of how models do under G. Writing $Y_{m,i} \in \{0,1\}$ for the score of model m on item i, as under exact-match or pass/fail scoring (Chen et al., 2021), models are ranked by

$$\theta_m \;=\; \mathbb{E}_{i\sim G}[\,Y_{m,i}\,]\,. \tag{1}$$

Anyone who reports a confidence interval on a benchmark score is already working with (1).

We write $Y = (Y_1, \ldots, Y_M)$ for the vector of scores on one item drawn from G, and $\mathbb{P}$ for its law. In the replay experiments G draws items uniformly with replacement from a published bank, which we call the *replay law*; the truth is then known, and so is every quantity computed from it.

A ranking places j above k, and that placement is an *inversion* when $\theta_k > \theta_j$. Ties are not inversions, so two models with equal θ can go in either order. Throughout, cost is counted in independent item draws per model, not in model calls or tokens (Appendix I).

2.2 Shared items and paired differences

Every opened model sees the same drawn items, so for a pair $e = (j, k)$ this gives a per-item difference $D_{e,i}$, and we write Δ_e for its mean, the gap of the pair:

$$D_{e,i} \;=\; Y_{j,i} - Y_{k,i} \;\in\; \{-1, 0, +1\}, \qquad \Delta_e \;=\; \mathbb{E}[D_{e,i}] \;=\; \theta_j - \theta_k. \tag{2}$$

Items differ in difficulty and a shared item cancels most of that difficulty from the comparison. That is worth a good deal: on the MMLU example the neighbouring pair with the smallest gap has paired variance 0.19, while the same two models scored on separate items give 0.50.

The $D_{e,i}$ are independent across items, but they are heavily dependent across pairs because one model appears in many pairs, and nothing below assumes otherwise.

2.3 Three probabilities, kept apart

A certificate involves three numbers, and running them together is a common source of confusion.

Inferential validity. The chance that some comparison we certify points the wrong way. This is what a guarantee controls, and §4 shows it costs nothing beyond bounded scores and independent items.

Certification power. The chance that every comparison clears its threshold, so that a complete ranking comes out. This depends on the true gaps and on the budget, and we measure it.

Ranking risk. The chance that a forced complete ranking contains an inversion.

The three come apart. A procedure that certifies nothing has perfect validity and no power, and a procedure whose empirical order is usually right can still prove almost none of it, which is the case on MMLU in §1. Validity is a statement over the draw of the items: the probability of outputting a wrong certificate is at most δ. Appendix X reports the rate among output certificates for one low-power synthetic cell.

2.4 The output rule

A certificate reads a sample and outputs a set of *certified edges* $j \to k$, and, when those edges determine one, a total order. Its guarantee is that the probability that any output relation is an inversion is at most δ. A pair with no edge is *unresolved*. A procedure can also output a *forced* total order that completes the certified edges arbitrarily. That order carries no guarantee, and its chance of containing an inversion is the ranking risk above.

§4 gives two certificates that meet this guarantee in different ways. The uniform-budget certificate tests each of the K pairs at δ/K and outputs its certified edges. The ranking certificate tests every pair at the larger level $\delta/(M-1)$ and outputs the whole order or nothing, and its guarantee rests on a different argument.

3 How many items the problem needs

A procedure that gets the ranking right under two laws that order it differently must tell those laws apart (Figure 1, bottom). The work is in choosing a second law that is close, because a close law gives a strong bound.

3.1 The nearest alternative

Fix one neighbouring pair and write $p_+ = \mathbb{P}(D_e = +1)$, $p_- = \mathbb{P}(D_e = -1)$, and $s = p_+ + p_-$. The discordant items are the ones the two models answer differently and s is their share, which on the hardest MMLU pair is $s = 0.32$, so about two thirds of items say nothing about that comparison.

An obvious alternative swaps p_+ and p_-, but that law is far away and gives a weak bound. The nearest alternative sits on the boundary between the two orderings:

$$Q_0(D_e = +1) = Q_0(D_e = -1) = s/2, \qquad Q_0(Y \mid D_e) = \mathbb{P}(Y \mid D_e). \tag{3}$$

The second half of (3) is what makes this usable, because the whole score vector keeps its conditional law given D_e, so every other model's behaviour given D_e is unchanged and the construction survives the dependence between pairs. Reweighting items is enough to produce it, and the weights are positive.

Under Q_0 the two models tie, and a tie is not yet a wrong ranking, so we lean past the boundary by $\eta > 0$ and let η go to zero. The bound below uses total variation, but the Kullback–Leibler divergence shows most simply why this alternative is the right one. From $\mathbb{P}$ to Q_0 it is

$$J \;=\; p_+ \log \frac{2p_+}{s} \;+\; p_- \log \frac{2p_-}{s} \;=\; \frac{\Delta_e^2}{2s} + O\big(\Delta_e^4/s^3\big)\,. \tag{4}$$

The swapped law has $2\Delta_e^2/s$ to leading order, so the nearest alternative is about four times closer, and a lower bound built on it is about four times larger.

3.2 The bound

Let A be the event that a procedure certifies the true order. Validity makes A unlikely under every reversing law Q_η, and letting η go to zero carries this to Q_0. A procedure that makes A likely under $\mathbb{P}$ and unlikely under Q_0 separates the two laws, and total variation measures exactly how well any event can separate them.

Theorem 1 (Item requirement)**.** *Draw items independently from G, and fix a neighbouring pair e with $\Delta_e > 0$ and $p_- > 0$. Let a procedure certify a correct complete ranking under $\mathbb{P}$ with probability at least $1 - \delta$, and let it certify an incorrect one with probability at most δ under every law Q_η of* (3) *that reverses the pair. Then*

$$\mathrm{TV}(\mathbb{P}^n, Q_0^n) \;\geq\; 1 - 2\delta, \qquad \mathrm{TV}(\mathbb{P}^n, Q_0^n) = \sum_m \mathbb{P}\{\mathrm{Bin}(n, s) = m\}\, \mathrm{TV}\big(\mathrm{Bin}(m, a), \mathrm{Bin}(m, \tfrac{1}{2})\big)\,, \tag{5}$$

with $a = p_+/s$. The smallest n satisfying (5) *bounds from below the item requirement of every procedure whose output is a function of at most n items drawn from G. That includes procedures that stop early or choose which models to score from the items seen so far, and no property of the other $M - 2$ models is used. That includes sequential tests (Wald, 1945) and confidence sequences (Waudby-Smith et al., 2021) whose item count is bounded by n.*

The same argument with validity α and success probability $1 - \beta$ gives $1 - \alpha - \beta$; we use $\alpha = \beta = \delta$. The second equality is exact and holds for every n, because the tilt in (3) leaves $\mathbb{P}(D_e = 0)$ alone, so the number of discordant items has the same law under $\mathbb{P}$ and Q_0. Given that number, the count of $+1$s is all that the likelihood ratio depends on, so those counts are sufficient and the sum in (5) is over their law. Appendix B has the details, along with two weaker closed forms that we use as sanity checks.

3.3 Paying for more than one comparison

Theorem 1 looks at one adjacent comparison, and a leaderboard has $M - 1$ of them. A procedure has to be safe against any one of them being reversed, and the same argument extends to cover that.

Theorem 2 (Mixture bound)**.** *Let $e_1, \ldots, e_k$ be adjacent pairs of the true order, each with a positive gap and $\mathbb{P}(D_{e_r} = -1) > 0$, let Q_r tilt pair e_r just past the tie as in* (3)*, and let $\pi_r \geq 0$ be weights summing to one. A procedure satisfying the hypotheses of Theorem 1 for every Q_r has*

$$\mathrm{TV}(\mathbb{P}^n,\ \textstyle\sum_r \pi_r Q_r^n) \;\geq\; 1 - 2\delta. \tag{6}$$

With A as before, $\mathbb{P}^n(A) \geq 1 - \delta$, while $Q_r^n(A) \leq \delta$ for each r because that order is wrong under Q_r, so the mixture also gives A probability at most δ. With $k = 1$ this is Theorem 1. When the components pull in different directions, the mixture can be harder to tell from $\mathbb{P}$ than any one of them, and that is where the price of several comparisons enters a lower bound.

Every component is the construction of (3), so nothing is assumed about how pairs depend. The total variation equals $\mathbb{E}_{\mathbb{P}}[(1 - \sum_r \pi_r L_r)_+]$ with $L_r = dQ_r^n / d\mathbb{P}^n$ a function of pair e_r's discordant counts, using the k hardest adjacent pairs with equal weights. We select a candidate n_0 on search draws and estimate the total variation at that fixed value on fresh draws. If its empirical Bernstein upper confidence bound is below $1 - 2\delta$, Theorem 2 rules out n_0 items, and since the total variation is non-decreasing in n it rules out every smaller sample too. All eighteen reported bounds hold simultaneously with probability at least 0.95 over these Monte Carlo draws (Appendix N).

How much the mixture adds depends on how many pairs are nearly as hard as the hardest, which is a property of the leaderboard. Consider a family built so that m comparisons are equally hard and independent. There any procedure that outputs the correct ranking with probability $1 - \delta$ everywhere in the family, and so any certificate that does, needs the single-pair requirement at error $1-(1-\delta)^{1/m}$, and some procedure attains it (Appendix D). That shows the multiplicity cost cannot be removed in general.

3.4 Reading the bound

Figure 2 evaluates both bounds on all eighteen configurations, at the parameters of the replay law. A statement about the unknown population needs parameters that hold simultaneously over every neighbouring pair of a comparison set fixed on independent items, which gives a much smaller bound (Appendix C). Either way the bound concerns probability, and a single run on a smaller sample can still succeed, and §5 reports one benchmark where that happens.

4 A certificate that gets close to it

The per-pair test is classical: it is the conditional binomial test behind McNemar (1947). What is new is how it connects to ranking-level error control and to the bounds of §3. We show that certifying a *ranking* costs $M - 1$ levels instead of K, and that a step-down shortcut does not keep the guarantee.

4.1 The sign of a paired difference needs no concentration bound

A paired difference on binary scores takes three values, so its mean has a special form, and from (2) we get

$$\Delta_e \;=\; p_+ - p_-, \qquad \text{so} \qquad \Delta_e > 0 \iff p_+ > p_-. \tag{7}$$

Deciding which model is better is therefore a question about two probabilities, and the discordant items answer it. Concentration inequalities are built to handle magnitudes, whereas (7) asks only for a sign, which is why nothing in this section estimates a variance.

Theorem 3 (Exact per-pair certificate)**.** *Let n_+ and n_- count the discordant items in each direction, let $m = n_+ + n_-$, and certify $j \to k$ when $n_+ > n_-$ and $\mathbb{P}\{\mathrm{Bin}(m, \frac{1}{2}) \geq n_+\} \leq t$ for a level t fixed before the items are drawn. Then the certificate of pair e is an inversion with probability at most t, for every n and every item law.*

Appendix E gives the proof. Conditional on m the count n_+ follows $\mathrm{Bin}(m, q)$ with $q = p_+/(p_+ + p_-)$, and an inversion needs $q < \frac{1}{2}$, so $\mathrm{Bin}(m, q)$ is dominated by $\mathrm{Bin}(m, \frac{1}{2})$ and nothing is approximated. The bound holds at every n with no concentration inequality, and the discreteness of the binomial can leave it slack at small m.

With both under the uniform-budget certificate and counting complete rankings, empirical Bernstein needs 1.36 to 1.57 times as many items as Theorem 3, and across 90 cells the exact test certifies at least as often in every cell and strictly more often in 56 (Appendix T).

4.2 A ranking costs $M - 1$ levels, not K

There are $K = M(M-1)/2$ comparisons and a certificate has to be right about all of them at once, so the obvious price is a level of δ/K each. That price is higher than it needs to be, because a ranking is not K separate claims. Its errors are witnessed by $M - 1$ comparisons, and $\delta/(M-1)$ suffices.

Theorem 4 (Ranking certificate)**.** *Compute the K exact p-values p_{jk} of Theorem 3, each in the direction of the sample means $\hat{\theta}$. Certify the whole empirical order when every one of them is at most*

$\delta/(M-1)$*, and otherwise certify no total order. Then the probability of certifying an order that contains an inversion is at most* δ*. Nothing is assumed about the dependence between pairs.*

Proof. Fix one order π^* that is consistent with θ, breaking any population ties by a rule chosen before the items are drawn, so that $\theta_{\pi_1^*} \geq \cdots \geq \theta_{\pi_M^*}$. Suppose the certified order contains an inversion. It then differs from π^*, which contains none, so by transitivity of the empirical means some π^*-adjacent pair (a, b) has $\hat{\theta}_b > \hat{\theta}_a$. The rule certified that pair at level $\delta/(M-1)$, so the event lies in $E_{ab} = \{\hat{\theta}_b > \hat{\theta}_a,\ p_{ab} \leq \delta/(M-1)\}$. Since $\theta_a \geq \theta_b$, the argument of Theorem 3 bounds $\mathbb{P}(E_{ab})$ by $\delta/(M-1)$, the tied case included, because a binomial is stochastically increasing in its success probability. The $M-1$ pairs are unknown to us but fixed, so a union bound over them gives δ. □

Fixing π^* in advance and bounding the event $\hat{\theta}_b > \hat{\theta}_a$ with ties included is what covers population ties, and Appendix E gives an order that an argument ignoring them would miss.

The threshold is larger by a factor $M/2$, ten-fold at $M = 20$, and the union bound runs over the true neighbours without our having to identify them.

4.3 Holm with data-selected directions

Holm's procedure raises each threshold as rejections accumulate (Holm, 1979), and most leaderboard comparisons are easy, so it is tempting to feed it the K p-values of Theorem 4. Those p-values choose their direction from the data, and Theorem 3 bounds their wrong-direction rate only at a level fixed in advance. Holm over the $2K$ pre-specified one-sided nulls $\theta_j \leq \theta_k$ with valid p-values does control directional errors, but at these levels at most one direction per pair rejects, so there are at most K rejections, each faces at most $\delta/(K+1)$, and every edge it certifies is also certified by the uniform-budget certificate.

Proposition 5 (Holm on data-selected p-values)**.** *There is an instance with $M = 4$ binary-scored models and $n = 10^6$ i.i.d. items on which Holm applied to the p-values of Theorem 4 certifies an inverted edge with probability at least $0.0502 > \delta = 0.05$.*

Two independent blocks hold two nearly tied models each, far apart across blocks. The four cross-block comparisons reject at once, so the two tied comparisons face $\delta/2$ and δ. Appendix F computes the error exactly over the random discordant counts. Sending the within-block gap to zero at fixed discordant counts and then letting the counts grow gives the limit $\delta + \delta^2/4$; at a fixed positive gap the error tends to zero as n grows, which is why the instance fixes n.

The excess occurs on runs that reject only the first hard comparison, where no complete order is output; the complete-order error on this instance is 0.0056. Theorem 4 provides the complete-order certificate used in this paper.

4.4 When the ranking does not close

Theorem 4 is all or nothing. For a partial order, the uniform-budget certificate certifies each pair at δ/K, valid by Theorem 3 and a union bound under arbitrary dependence. The certified edges form a partial order, and its transitive closure costs no further multiplicity: whenever the certified edges are all correct, so is every edge they imply, as $a \to c$ follows from $a \to b$ and $b \to c$. The two are separate certificates, and we evaluate them separately. An evaluator who reports the complete order when it closes and the partial order otherwise runs each at $\delta/2$, which keeps the combined output within δ.

4.5 Choosing the witnesses first

Theorem 4 pays $\delta/(M-1)$ on every pair because it does not know which pairs witness an error. A *proposal–validation* certificate buys that knowledge: it orders the models on a proposal part of the items, then tests only the $M-1$ comparisons adjacent in that order on the remaining items, each at the full level δ, and certifies the order if all pass. Given the proposal, an incorrect proposed order has a fixed adjacent inversion that must pass its own level-δ test, so validity needs no union bound (Appendix J); the price is the proposal items, all of which count toward n.

Table 1: Certificates on held-out items. Models are chosen by R1 and R3 on 30% of the bank and certified on the other 70%. *Closed*: configurations whose complete ranking Theorem 4 certifies. The uniform-budget certificate at δ/K is a separate rule: the share of relations it certifies, and the proved prefix, the number of top models each certified above every model ranked below it. The last column is descriptive. Per-configuration values are in Appendix Q.

	held-out items, models chosen on the rest				whole bank
Benchmark	Items	Closed	Relations at δ/K	Prefix	closed
GSM8K	923	0 of 2	64–78%	0	0 of 2
MMLU	9,829	0 of 4	80–96%	0–3	0 of 4
ARC-C	820	0 of 4	44–73%	0	0 of 4
HellaSwag	7,029	0 of 4	89–98%	1–8	4 of 4
WinoGrande	887	0 of 4	40–68%	0–1	0 of 4

5 What this costs on real leaderboards

We run the certificate on eighteen leaderboard configurations across five benchmark families. Model sets overlap within a family, so these are eighteen configurations and not eighteen independent replications.

5.1 Instances and protocol

The scores come from the published correctness matrix of Polo et al. (2024) over 395 open models, and five binary-scored benchmarks enter: GSM8K (Cobbe et al., 2021), MMLU (Hendrycks et al., 2020), ARC-Challenge (Clark et al., 2018), HellaSwag (Zellers et al., 2019) and WinoGrande (Sakaguchi et al., 2020). The last three were fixed by a written selection rule before any outcome on them was computed.

Each configuration takes $M \in \{10, 20\}$ models under one of two rules fixed in advance. Both require adjacent accuracies to differ by at least 0.01. R1 chains the models with the smallest accuracy span, and R3, allowing up to twice that span, the chain whose comparisons differ most in difficulty (Appendix O). Both are chosen to make the configurations hard. Sampling uses the replay law of §2, and $\delta = 0.05$. The quantity n_{95} is the number of items at which the complete ranking is certified and correct in 95% of draws, read off a completion curve with $8{,}000$ draws per grid point, with a bootstrap interval of about $\pm 1.5\%$.

5.2 What a benchmark's own items prove

Both rules read the bank, so Table 1 uses a selection/validation split: models are chosen on a random 30% of each bank and certified on the other 70%. No complete ranking closes on the held-out items, but the uniform-budget certificate certifies 40% to 98% of the relations. Four HellaSwag rankings close on the whole bank, which also chose their models, and we report that only as a description.

5.3 How close the certificate runs to the limit

Figure 2 puts every configuration on one axis. Against the larger of the two lower bounds, the certificate needs 1.61 to 1.81 times the items, and the bootstrap interval on each n_{95} moves the ratio by at most 0.04. The denominators are certified values, Theorem 1 exactly and Theorem 2 by fresh draws, all eighteen at once at computational confidence 0.95.

The three benchmarks fixed by the written selection rule give 1.61 to 1.80 and the two original ones give 1.62 to 1.81. Rebuilt on MMLU and HellaSwag with gap floors 0.005 to 0.02 and M from 5 to 20, the bounds move by two orders of magnitude, while the ratio is stable across floors and rises with M from 1.43 to at most 1.83 (Appendix R).

5.4 What the remaining factor is

On each configuration's hardest pair, the exact test at level δ reaches 95% power with 0.6% to 1.9% more items than Theorem 1 requires, so for one comparison neither the test nor the bound can move much. The factor in Figure 2 concerns simultaneous certification.

Some of that cost is unavoidable in general (Appendix D), and Theorem 2 captures part of it on these leaderboards. Some is recoverable, since the proposal–validation certificate of Appendix J saves 3.3% to 9.9% on every twenty-model configuration. Theorem 4 beats the uniform-budget certificate by 18.3% to 23.8% on every configuration.

Figure 3 shows what n_{95} measures on two twenty-model configurations, one whose bank closes and one whose bank does not, next to the probability that the sorted order is simply right. The four HellaSwag banks that close do so on one realisation: same-size bootstrap samples from them close in only 0.3% to 40.6% of draws (Appendix H).

5.5 Empirical checks of error control

Validity was tested on synthetic near-ties, where the truth is set exactly, with two gap shapes, gaps down to 0.0005, $M \in \{4, 10\}$, two sample sizes, and scores independent or correlated across models (Appendix T). Over 48 cells of $2{,}000$ runs, Theorem 4 certified $11{,}588$ complete rankings, and its worst cell put a wrong direction in 1.5% of runs against the allowed 5%, as the worst of the tested cells, with the rate counted over all runs.

6 Related work

Ranking and selection. A single-sample procedure for ranking the means of normal populations with known variances goes back to Bechhofer (1954). Our setting replaces that model with binary scores on shared items, where an exact test and a lower bound given by an exact finite sum are available and nothing need be assumed about how the comparisons depend on each other. The exact test is the conditional binomial test behind McNemar (1947), the continuous certificate of Appendix K

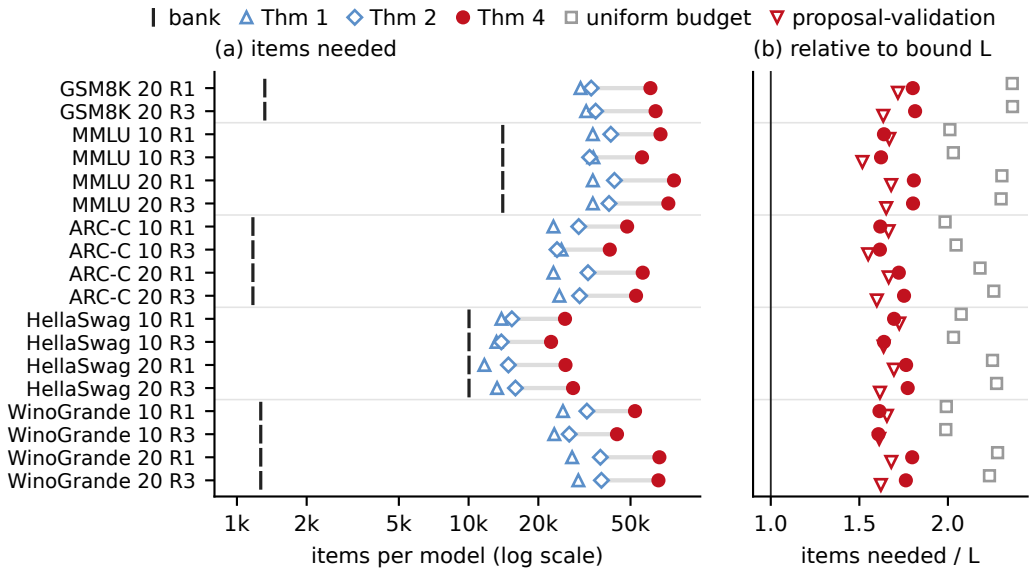

Figure 2: All eighteen configurations. **(a)** Items per model; the whisker on n_{95} is a bootstrap interval for estimating it under the bank's empirical law, and the bounds are certified values. **(b)** n_{95} over the larger bound L; the proposal–validation certificate uses a fixed 30% proposal and is charged for all its items.

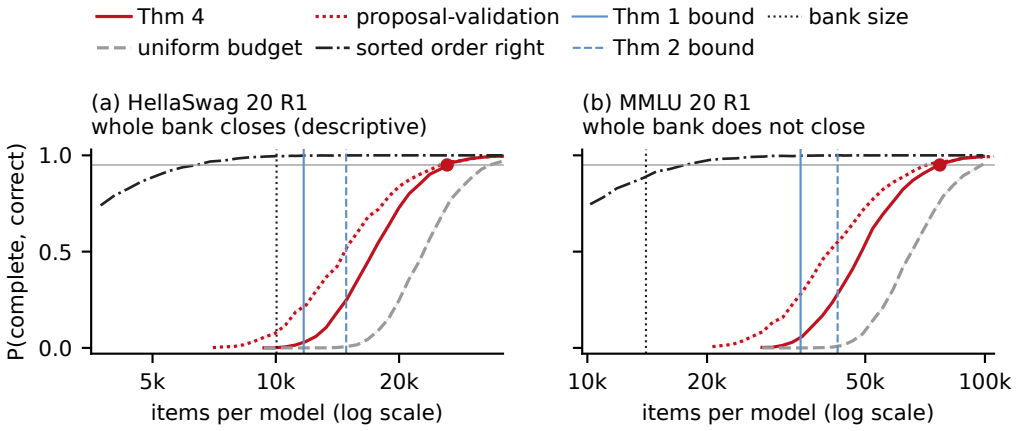

Figure 3: Probability of certifying the complete ranking correctly under the bank's empirical law, with 95% Monte Carlo bands; the dot is n_{95}. The black dash-dot curve is the probability that the sorted order has no inversion, which no certificate controls. Panel labels describe the one run on the whole bank.

is the empirical Bernstein bound of Maurer & Pontil (2009), and Holm (1979) controls the familywise error for valid p-values of pre-specified hypotheses (§4.3).

Inference for ranks. Mogstad et al. (2023) build simultaneous confidence sets for ranks from estimated means with a general covariance, the statistical antecedent for this paper. Binary scores on shared items make an exact test available, and we ask the design question their framework leaves open, namely how many items a certificate needs.

Certified evaluation of language models. Zhou et al. (2026) certify evaluation with e-processes, valid under optional stopping. Neuhof & Benjamini (2026) give rank intervals from pairwise directional hypotheses tested with Holm, a different target from a complete-order certificate, and Wu et al. (2026) select queries adaptively from historical structure. Each delivers a valid statement at a chosen budget, and this paper asks how much of that budget the statement inherently costs.

Adaptive and smaller evaluation. Balkır et al. (2026) use item-response theory to choose informative items, and reach a confident ranking from fewer of them, with continuous scores. For binary scores the ratios in Figure 2 bound what any procedure drawing items independently from G can recover. Polo et al. (2024) reproduce benchmark scores from a few hundred items, and their matrices supply every instance in §5, but a score recovered to within a point can leave open the comparison it decides.

Fixed-budget ranking in bandits. Karpov & Zhang (2020) give a fixed-budget guarantee for coarse ranking under batched sampling, Saad et al. (2023) rank experts actively across tasks, and Lalitha et al. (2023) adapt a fixed budget to heterogeneous variances. Our lower bound follows the change-of-measure tradition of Kaufmann et al. (2014) and Chernoff (1959), but shared items make every model see the same draw, so the alternative in (3) leaves the conditional law of the score vector untouched.

7 Conclusion

We give computable instance-specific lower bounds and a complete-ranking certificate built on McNemar's exact test at $\delta/(M-1)$. Measured under each bank's empirical law on the eighteen replay configurations, it needs 1.61 to 1.81 times the larger bound, in independent item draws. On each configuration's hardest pair, the per-pair test is within 1.9% of its bound.

REPRODUCIBILITY STATEMENT

Every number in this paper is produced by a released script from the public correctness matrix of Polo et al. (2024), checked against a recorded SHA-256 digest; Appendix V maps each claim to its script and output file, and Appendices B to J give the proofs. Replay draws items with replacement from a bank, so every sample size is a size under the bank's empirical law. Random streams are seeded deterministically, from a SHA-256 digest of the configuration and purpose or a fixed integer. Completion curves use $8{,}000$ draws per grid point, the proposal–validation curves $3{,}000$, the Bernstein comparison and the sensitivity to δ $4{,}000$, the baselines 600 paired replicates, and the synthetic validity sweep $2{,}000$ runs per cell; the mixture bounds use $6{,}000$ draws for the search and $400{,}000$ for certification (Appendix O). Compared methods share draws, which makes their comparisons paired.

AI USE STATEMENT

Generative AI tools were used for language polishing, for writing code, and for feedback on technical content; all AI-generated content was rigorously reviewed by the authors.

REFERENCES

Esma Balkır, Alice Pernthaller, Marco Basaldella, José Hernández-Orallo, and Nigel Collier. Confident Rankings with Fewer Items: Adaptive LLM Evaluation with Continuous Scores. *arXiv preprint arXiv:2601.13885*, 2026. doi: 10.48550/arXiv.2601.13885.

Robert E. Bechhofer. A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with known Variances. *The Annals of Mathematical Statistics*, 1954. doi: 10.1214/aoms/1177728845.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé de Oliveira Pinto, and Jared Kaplan. Evaluating Large Language Models Trained on Code. *arXiv preprint arXiv:2107.03374*, 2021. doi: 10.48550/arXiv.2107.03374.

Herman Chernoff. Sequential Design of Experiments. *The Annals of Mathematical Statistics*, 1959. doi: 10.1214/aoms/1177706205.

Peter E. Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, and Carissa Schoenick. Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge. *arXiv preprint arXiv:1803.05457*, 2018. doi: 10.48550/arXiv.1803.05457.

C. J. Clopper and E. S. Pearson. THE USE OF CONFIDENCE OR FIDUCIAL LIMITS ILLUSTRATED IN THE CASE OF THE BINOMIAL. *Biometrika*, 1934. doi: 10.1093/biomet/26.4.404.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, and Lukasz Kaiser. Training Verifiers to Solve Math Word Problems. *arXiv preprint arXiv:2110.14168*, 2021. doi: 10.48550/arXiv.2110.14168.

Fabian Dickmann and Nikolaus Schweizer. Faster comparison of stopping times by nested conditional Monte Carlo. *The Journal of Computational Finance*, 2016. doi: 10.21314/jcf.2016.221.

B. Efron. Bootstrap Methods: Another Look at the Jackknife. *The Annals of Statistics*, 1979. doi: 10.1214/aos/1176344552.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, and Dawn Song. Measuring Massive Multitask Language Understanding. *arXiv preprint arXiv:2009.03300*, 2020. doi: 10.48550/arXiv.2009.03300.

Sture Holm. A Simple Sequentially Rejective Multiple Test Procedure. *Scandinavian Journal of Statistics*, 1979. doi: 10.2307/4615733.

Nikolai Karpov and Qin Zhang. Batched Coarse Ranking in Multi-Armed Bandits. In *Advances in Neural Information Processing Systems*, 2020.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the Complexity of Best Arm Identification in Multi-Armed Bandit Models. *arXiv preprint arXiv:1407.4443*, 2014. doi: 10.48550/arXiv.1407.4443.

Anusha Lalitha, Kousha Kalantari, Yifei Ma, Anoop Deoras, and Branislav Kveton. Fixed-Budget Best-Arm Identification with Heterogeneous Reward Variances. *arXiv preprint arXiv:2306.07549*, 2023. doi: 10.48550/arXiv.2306.07549.

Percy Liang, Rishi Bommasani, Tong Lee, Dimitris Tsipras, Dilara Soylu, and Michihiro Yasunaga. Holistic Evaluation of Language Models. *arXiv preprint arXiv:2211.09110*, 2022. doi: 10.48550/arXiv.2211.09110.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 1976. doi: 10.1093/biomet/63.3.655.

Andreas Maurer and Massimiliano Pontil. Empirical Bernstein Bounds and Sample Variance Penalization. *arXiv preprint arXiv:0907.3740*, 2009. doi: 10.48550/arXiv.0907.3740.

Quinn McNemar. Note on the Sampling Error of the Difference Between Correlated Proportions or Percentages. *Psychometrika*, 1947. doi: 10.1007/bf02295996.

Magne Mogstad, Joseph P. Romano, Azeem M. Shaikh, and Daniel Wilhelm. Inference for Ranks with Applications to Mobility across Neighbourhoods and Academic Achievement across Countries. *The Review of Economic Studies*, 2023. doi: 10.1093/restud/rdad006.

Bitya Neuhof and Yuval Benjamini. Rank Intervals for Leaderboards: A Hierarchical Framework for Model Evaluation. *arXiv preprint arXiv:2606.08679*, 2026. doi: 10.48550/arXiv.2606.08679.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: evaluating LLMs with fewer examples. *arXiv preprint arXiv:2402.14992*, 2024. doi: 10.48550/arXiv.2402.14992.

El Mehdi Saad, Nicolas Verzélen, and Alexandra Carpentier. Active Ranking of Experts Based on their Performances in Many Tasks. *arXiv preprint arXiv:2306.02628*, 2023. doi: 10.48550/arXiv.2306.02628.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An Adversarial Winograd Schema Challenge at Scale. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020. doi: 10.1609/aaai.v34i05.6399.

Aarohi Srivastava, Abhinav Rastogi, Abhishek S. Rao, Abu Awal Shoeb, Abubakar Abid, and Adam Fisch. Beyond the Imitation Game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022. doi: 10.48550/arXiv.2206.04615.

Alexandre B. Tsybakov. Introduction to Nonparametric Estimation. *Springer series in statistics*, 2008. doi: 10.1007/b13794.

A. Wald. Sequential Tests of Statistical Hypotheses. *The Annals of Mathematical Statistics*, 1945. doi: 10.1214/aoms/1177731118.

Ian Waudby-Smith, David Arbour, Ritwik Sinha, Edward H. Kennedy, and Aaditya Ramdas. Time-uniform central limit theory and asymptotic confidence sequences. *arXiv preprint arXiv:2103.06476*, 2021. doi: 10.48550/arXiv.2103.06476.

Skyler Wu, Yash Nair, and Emmanuel J. Candés. Efficient Evaluation of LLM Performance with Statistical Guarantees. *arXiv preprint arXiv:2601.20251*, 2026. doi: 10.48550/arXiv.2601.20251.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a Machine Really Finish Your Sentence? 2019. doi: 10.18653/v1/p19-1472.

Zhijian Zhou, Z Zhenyu Ye, Zhaorun Chen, Bo Li, and Feng Liu. CELEUS: Certifiable and Efficient LLM Evaluation via E-Processes. *arXiv preprint arXiv:2606.20820*, 2026. doi: 10.48550/arXiv.2606.20820.

A NOTATION

Symbol	Meaning
M, $K = M(M-1)/2$	number of models, and of pairs of models
G	the law of one item; in replay, the empirical law of a bank
$Y = (Y_1, \dots, Y_M) \in \{0,1\}^M$	the scores of all models on one item drawn from G
$\mathbb{P}$	the law of Y
$\theta_m = \mathbb{E}[Y_m]$	accuracy of model m under G, the quantity models are ranked by
π^*	a true order: one refinement of the weak order of θ, fixed in advance
$e = (j,k)$, $D_e = Y_j - Y_k$	a pair and its paired difference on one item
$\Delta_e = \theta_j - \theta_k$	the gap of pair e
p_+, p_-	$\mathbb{P}(D_e = +1)$ and $\mathbb{P}(D_e = -1)$
$s = p_+ + p_-$, $a = p_+/s$	the discordant share, and the share of discordant items favouring j
n	items drawn, per model; every model sees the same items
$n_+, n_-, m = n_+ + n_-$	discordant counts of a pair in a sample
p_{jk}	exact one-sided p-value of Theorem 3 in the direction of the sample means
δ	the allowed probability of certifying any inversion
δ_e	the level a pre-specified budget assigns to pair e, with $\sum_e \delta_e \le \delta$
Q_η, Q_0	the tilted alternative of (3), leaning past the tie by η, and its limit
J	the per-item divergence from $\mathbb{P}$ to Q_0, (4)
$L_r = dQ_r^n / d\mathbb{P}^n$	the likelihood ratio of the tilt of pair e_r over n items
n_{95}	items at which a certificate closes the ranking correctly with probability 0.95
$n_2(t)$	items at which the level-t exact test on one pair reaches power 0.95
m (in App. D)	number of independent hard comparisons in the worst-case family

Three uses of "exact". The word is used in three senses, which should not be confused. An *exact test* (Theorem 3) has its error bounded at every n with no approximation. An *exact expression* (the finite sum in Theorem 1) evaluates a total variation without asymptotics, and an *exact computation* (Appendix F) evaluates a probability without Monte Carlo. The mixture bound is certified numerically, and the smallest achievable sample size lies between the larger bound and the certificate's requirement.

Throughout, an *inversion* places j above k when $\theta_k > \theta_j$. Ties are not inversions, so a certificate is correct whenever everything it asserts agrees with θ weakly.

B PROOF OF THEOREM 1

Fix the neighbouring pair $e = (j,k)$ with $\Delta_e > 0$, and write $p_+ = \mathbb{P}(D_e = +1)$, $p_- = \mathbb{P}(D_e = -1)$, $s = p_+ + p_-$ and $a = p_+/s$.

The alternative. Let Q_η reweight the item law so that

$$Q_\eta(D_e = +1) = s(\tfrac{1}{2} - \eta), \qquad Q_\eta(D_e = -1) = s(\tfrac{1}{2} + \eta), \qquad Q_\eta(D_e = 0) = 1 - s, \tag{8}$$

and $Q_\eta(Y \mid D_e) = \mathbb{P}(Y \mid D_e)$ for the whole score vector Y. The weights are $w(+1) = (\frac{1}{2} - \eta)s/p_+$, $w(-1) = (\frac{1}{2} + \eta)s/p_-$ and $w(0) = 1$, which is why the theorem assumes $p_- > 0$. They are positive and average to one, so Q_η is an item law. Under Q_η we have $\Delta_e = -2\eta s < 0$, so the ranking of j and k is reversed.

Only the marginal of D_e changes, and every conditional law given D_e is carried over. This is what makes the construction usable under dependence, and it is weaker than leaving the other models alone: with $Y_1 \sim \mathrm{Bern}(0.6)$, $Y_2 = 1 - Y_1$ and $Y_3 = Y_1$, the tilt moves $\mathbb{E}[Y_3]$ from 0.6 to 0.5. What the argument needs is that the conditional kernel is shared, which it is, and no independence between pairs is used anywhere below.

The two-point argument. Let Ψ be any procedure that draws n items from the item law and outputs a ranking, possibly adaptively and possibly stopping early. Let A be the event that its output ranks j above k. By hypothesis $\mathbb{P}^n(A) \geq 1-\delta$ and $Q_\eta^n(A) \leq \delta$, because under Q_η that output contains an inversion. Hence

$$\mathrm{TV}(\mathbb{P}^n, Q_\eta^n) \;\geq\; \mathbb{P}^n(A) - Q_\eta^n(A) \;\geq\; 1-2\delta. \tag{9}$$

Total variation over the n drawn items dominates total variation over any function of them, and an adaptive transcript is such a function. So (9) constrains sequential procedures too.

The exact identity. Let $\eta \to 0$ and write Q_0 for the limit. The bound (9) survives the limit. Q_η and Q_0 differ only in the probabilities of $D_e = \pm 1$, each by $s\eta$, so $\mathrm{TV}(Q_\eta, Q_0) = s\eta$ and $\mathrm{TV}(Q_\eta^n, Q_0^n) \leq ns\eta$. The triangle inequality then gives $\mathrm{TV}(\mathbb{P}^n, Q_0^n) \geq 1 - 2\delta - ns\eta$ for every $\eta > 0$, hence $\mathrm{TV}(\mathbb{P}^n, Q_0^n) \geq 1-2\delta$. The limit is needed because Q_0 itself has $\Delta_e = 0$, a tie, under which no output is an inversion. The tilt leaves $\mathbb{P}(D_e = 0)$ unchanged, so the number of discordant items m has law $\mathrm{Bin}(n, s)$ under both $\mathbb{P}$ and Q_0. The conditional law of Y given D_e is shared by the two laws. So conditional on m, the likelihood ratio between $\mathbb{P}^n$ and Q_0^n depends on the sample only through the count n_+ of $+1$s. So (m, n_+) is sufficient, and $n_+ \mid m$ is $\mathrm{Bin}(m, a)$ under $\mathbb{P}$ and $\mathrm{Bin}(m, \frac{1}{2})$ under Q_0. Total variation is preserved under sufficiency, which gives

$$\mathrm{TV}(\mathbb{P}^n, Q_0^n) = \sum_{m=0}^{n} \mathbb{P}\{\mathrm{Bin}(n, s) = m\}\, \mathrm{TV}\big(\mathrm{Bin}(m, a), \mathrm{Bin}(m, \tfrac{1}{2})\big)\,. \tag{10}$$

The right-hand side is a finite sum of exactly computable terms, so no asymptotic approximation enters. It is non-decreasing in n. The smallest n at which it reaches $1-2\delta$ is therefore the bound reported in Table 2 and Table 5. □

Two weaker closed forms, used as checks. Write $J = p_+ \log(2p_+/s) + p_- \log(2p_-/s)$ for the divergence from $\mathbb{P}$ to Q_0 at one item, and $\mathrm{kl}(u, v)$ for the binary relative entropy. The boundary form of the change-of-measure argument (Tsybakov, 2008) gives $n \geq \mathrm{kl}(1-\delta, \delta)/J$. The Hellinger affinity $A = 1 - s + \sqrt{s/2}\,(\sqrt{p_+} + \sqrt{p_-})$ gives $n \geq \log\Big(2\sqrt{\delta(1-\delta)}\Big) / \log A$. On the six GSM8K and MMLU configurations these fall below (10) by factors of 2.04 to 2.05 and 1.63 to 1.64. We report (10) and use the other two to check it.

Why the nearest alternative matters. The alternative that swaps p_+ and p_- has divergence $2\Delta_e^2/s$ to leading order, against $\Delta_e^2/(2s)$ for (8). The bound scales inversely with the divergence, so the swapped law loses a factor of about four.

C THE BOUND UNDER PARAMETER UNCERTAINTY

Theorem 1 is a statement about a law whose parameters are known. The replay experiments use the bank's own law, where they are. A statement about the unknown item population must estimate them, and it must do so for a set of comparisons that does not itself depend on the data used for the estimate.

We therefore use the selection/validation split of Appendix Q. The models and their order come from the 30% selection part, which fixes the $M-1$ neighbouring pairs before the validation part is seen. On the 70% validation part, each pair gets Clopper–Pearson (Clopper & Pearson, 1934) limits at level $\eta/(2(M-1))$ per side, $\eta = 0.05$, on its discordance s and on its direction a, and the bound of Theorem 1 is evaluated at the corner that makes it smallest, s at its upper limit and $|a - \frac{1}{2}|$ at its largest. The bound falls as s grows and as $|a - \frac{1}{2}|$ grows. Flipping each discordant item's direction with a fixed probability moves a towards $\frac{1}{2}$ and leaves $\frac{1}{2}$ fixed, so a smaller $|a - \frac{1}{2}|$ is a garbling of a larger one, and more discordant items cannot lower the total variation. The corner needs $|a - \frac{1}{2}|$ bounded from above, not a two-sided interval for a. If the true $a \geq \frac{1}{2}$ the corner can only fail when the upper limit falls below a, and if $a < \frac{1}{2}$ only when the lower limit rises above it, so for each fixed pair only one tail of a, fixed by the true parameter, is ever charged, together with the upper tail of s. Hence with probability at least $1-\eta$ over the validation items every pair's corner value is below its true value, and the maximum over pairs is a lower confidence bound on the item requirement for that fixed model set. Nothing is chosen after the validation items are seen.

Table 2: Theorem 1 under parameter uncertainty. Whole bank: plug-in value for the configurations of Table 5, descriptive of the bank. Models chosen on the selection part: plug-in value on the validation items, and the simultaneous lower confidence bound at $\eta = 0.05$, which holds for the item population because the comparison set was fixed before the validation items were used.

	whole bank	models chosen on the selection part		
Configuration	plug-in	validation items	plug-in	simultaneous
GSM8K, M=20, R1	30,432	923	77,098	710
GSM8K, M=20, R3	32,180	923	519,406	844
MMLU, M=10, R1	34,378	9,829	3,050,605	11,403
MMLU, M=10, R3	34,552	9,829	773,881	9,911
MMLU, M=20, R1	34,378	9,829	2,052,065	10,717
MMLU, M=20, R3	34,378	9,829	127,918	6,580
ARC-C, M=10, R1	23,238	820	337,201	812
ARC-C, M=10, R3	25,177	820	75,351	636
ARC-C, M=20, R1	23,238	820	505,781	690
ARC-C, M=20, R3	24,648	820	50,234	497
HellaSwag, M=10, R1	13,867	7,029	333,129	6,192
HellaSwag, M=10, R3	13,194	7,029	7,074,407	8,488
HellaSwag, M=20, R1	11,696	7,029	7,074,407	7,228
HellaSwag, M=20, R3	13,272	7,029	22,727	2,880
WinoGrande, M=10, R1	25,544	887	155,976	885
WinoGrande, M=10, R3	23,433	887	547,145	941
WinoGrande, M=20, R1	27,979	887	311,953	756
WinoGrande, M=20, R3	29,764	887	355,164	819

The plug-in value exceeds the simultaneous bound by a factor of 7.9 to 978.8 across the eighteen configurations, because the bound moves as $1/(a-\frac{1}{2})^2$ while a is estimated from the discordant items alone, and because some pairs that were at least 0.01 apart on the selection part are nearly tied on the validation part, which makes their plug-in values very large. Statements about the population use the right-hand column; the plug-in columns describe the observed items. On six configurations the lower confidence bound exceeds the number of validation items, so for each of them, with confidence 0.95, no procedure certifies its ranking with probability 0.95 from that many items, which is consistent with none of the held-out rankings closing. The confidence is simultaneous over the fixed comparisons within a configuration, not across the eighteen configurations.

Truncation of the finite sum. The sum (10) is evaluated over m within ten standard deviations of ns, skipping terms of weight below 10^{-17}. Each omitted term contributes at most its binomial weight, so the computed value T_{trunc} satisfies $T_{\mathrm{trunc}}(n) \le T(n) \le T_{\mathrm{trunc}}(n) + \tau(n)$ with τ the omitted mass. For each reported bound L we checked $T_{\mathrm{trunc}}(L-1) + \tau(L-1) < 1 - 2\delta$ on the pair that attains it: τ is at most 4×10^{-15}, and the margin below $1 - 2\delta$ is at least 5.9×10^{-7} (Table 4). So every reported integer is certified exactly, not approximately.

D THE MOST THE MULTIPLICITY CAN COST

Theorem 2 measures the multiplicity cost on a given leaderboard. This appendix solves one family of leaderboards exactly, which shows that cost cannot be removed in general.

The family. Take m blocks, each holding two models whose scores are drawn independently of every other block. In block r the discordance D_r takes the value $+1$ with probability $(s + v_r h)/2$, the value -1 with probability $(s - v_r h)/2$, and 0 otherwise, where $v_r \in \{-1, +1\}$ is the unknown orientation and $0 < h < s$. When $D_r = 0$ the pair scores $(1, 1)$ with a probability b_r that does not depend on v_r, and $(0, 0)$ otherwise, so D_r carries all the information about v_r. Block means are spaced so that every comparison across blocks is far wider than h, so the correct ranking is determined by v and distinct v give distinct rankings. All 2^m orientations belong to the family.

Let e_n be the smallest achievable error, at n items, of deciding a single orientation from n draws of one block. The sign of $\sum_i D_{r,i}$, with a fair coin on ties, attains it. The single-block requirement at error ε is defined as the smallest n with $e_n \le \varepsilon$.

Theorem 6 (Worst-case multiplicity cost)**.** *On this family, among procedures that always output a complete ranking, the smallest achievable worst-case probability of outputting a wrong one is exactly* $1-(1-e_n)^m$*. Consequently the number of items needed for error at most* δ *is the single-block requirement at error* $1-(1-\delta)^{1/m}$*, which lies between* δ/m *and* $\delta/(m(1-\delta))$*.*

Proof. Put the uniform prior on v. The blocks are independent, so the posterior factors across r, and the largest posterior probability of the whole vector is the product of the largest posterior probabilities of its coordinates. Those coordinate posteriors are independent under the prior predictive law, so the best achievable probability of a fully correct answer is $(1-e_n)^m$, and the worst-case error is at least the Bayes error $1-(1-e_n)^m$. Deciding each block separately with its optimal test, using independent tie-breaking coins, attains that value at every v. □

The item requirement also binds procedures that may abstain, for example the certificates of §4: one that outputs the correct complete ranking with probability at least $1-\delta$ becomes a procedure of the theorem's class, with error at most δ, once abstention is replaced by any fixed ranking.

The theorem is sharp, so it also shows what an easier argument misses. A marginal argument would ask each coordinate to have error about δ/m and conclude the same thing, but that step is false. Flipping all m decisions together with probability $\delta/2$ keeps the chance of a fully correct answer above $1-\delta$, while every marginal error is near $\delta/2$. The product-prior argument is what rules such correlated procedures out. A Fano bound on the full hypercube loses the $\log m$ entirely, because the entropy and the information both grow linearly in m; restricting it to the $m+1$ orientations that differ from one baseline in a single block recovers $nJ \ge (1-\delta)\log(m+1) - \log 2$.

At $M = 20$ and $\delta = 0.05$ the family has $m = 9$, and the information lower bound at error $1-(1-\delta)^{1/9}$ is 1.93 times the one at error δ. Both quantities are information lower bounds, and their ratio indicates the scale of the multiplicity cost.

E Proof of Theorem 3, and ties in Theorem 4

Why Theorem 4 fixes the tie-breaking in advance. With $\theta_A = \theta_B > \theta_C$ and $\pi^* = (A, B, C)$, the empirical order B, C, A puts C above A, which is an inversion, yet neither adjacent pair of π^* is inverted: A and B are tied, and B above C is correct. The pair (A, B) is nonetheless empirically reversed, and bounding the event $\hat\theta_b > \hat\theta_a$ for every π^*-adjacent pair, ties included, is what charges this run to one of the $M-1$ witnesses.

Proof. Suppose the certified direction $j \to k$ is an inversion, so $\theta_j < \theta_k$ and $p_+ < p_-$. Let

$$c(m) \;=\; \min\left\{c : \mathbb{P}\{\mathrm{Bin}(m, \tfrac{1}{2}) \ge c\} \le \delta_e\right\},$$

so the certification event is $\{n_+ \ge c(m)\}$. Conditional on m, the count n_+ is $\mathrm{Bin}(m, q)$ with $q = p_+/(p_+ + p_-) < \frac{1}{2}$. A binomial is stochastically increasing in its success probability, so

$$\mathbb{P}(n_+ \ge c(m) \,|\, m) \;\le\; \mathbb{P}\big\{\mathrm{Bin}(m, \tfrac{1}{2}) \ge c(m)\big\} \;\le\; \delta_e.$$

Averaging over m gives the claim. The case $\theta_j > \theta_k$ is symmetric. When $\theta_j = \theta_k$ neither direction is an inversion, so the event is empty. Nothing in the argument uses n, the item law, or the behaviour of the other models. □

The certificate for scores outside $\{0, 1\}$**.** Scores bounded in $[0, 1]$ without being binary use the same output rule on a one-sided empirical Bernstein bound (Maurer & Pontil, 2009). Certify $j \to k$ when $|\bar D_e| > \sqrt{2\hat V_e L/n} + 14L/(3(n-1))$ with $L = \log(2/\delta_e)$, writing $\hat V_e$ for the sample variance of the paired differences. Validity follows from a union bound over the K pairs and holds under arbitrary dependence between them. Theorem 3 replaces this whenever scores are binary, and §5 measures what the replacement is worth.

F HOLM'S PROCEDURE ON DATA-SELECTED p-VALUES

Holm's procedure is a natural choice for the K p-values of Theorem 4, each computed in the direction the data favour. This appendix records why that is attractive, why that particular use is invalid, and which use of Holm is valid. Nothing here concerns Holm's procedure applied to pre-specified hypotheses with valid p-values, which controls the familywise error under any dependence.

Why it looks right. Write $A_e(q)$ for the event that pair e rejects in the correct direction at threshold q, and let H be any set of comparisons. Holm sorts the p-values $P_{(1)} \le \cdots \le P_{(K)}$ and rejects $P_{(i)}$ when $P_{(l)} \le \delta/(K-l+1)$ for every $l \le i$. Suppose each pair outside H has $P_e \le \delta/K$ and each pair in H has $P_e \le \delta/|H|$. Consider position i. At most $|H|$ of the K values exceed δ/K, so if $P_{(i)} > \delta/K$ then at least $K-i+1 \le |H|$ values remain, giving $\delta/(K-i+1) \ge \delta/|H| \ge P_{(i)}$; and if $P_{(i)} \le \delta/K$ the comparison passes outright. Every position passes, so

$$\mathbb{P}(\text{complete}) \;\ge\; 1 - \sum_{e \in H} \mathbb{P}(A_e(\delta/|H|)^c) - \sum_{e \notin H} \mathbb{P}(A_e(\delta/K)^c)\,. \tag{11}$$

On a leaderboard almost every comparison clears δ/K at once, so $|H|$ is small and the hard comparisons are effectively tested near δ. That is a large power gain over δ/K. Among prefixes of the p-value ranking with size at most forty, the $|H|$ maximising (11) is 1, 2, 4, 6, 9 and 15 on the six GSM8K and MMLU configurations.

Why this use is invalid. The p-value $\mathbb{P}\{\mathrm{Bin}(m, \frac{1}{2}) \ge \max(n_+, n_-)\}$ picks its direction from the data. Theorem 3 bounds its wrong-direction rate at a level fixed before the data, and under a tie it falls below t with probability close to $2t$, so it is not a valid p-value for either one-sided null. Holm's guarantee needs valid p-values, and the step-down threshold is random, so the comparison tested last faces δ by itself, so any contribution from the others pushes the total past δ.

Take $M = 4$ in two independent blocks. Block one is (Y_1, Y_2) with outcomes 11, 10 and 01 of probability 0.8, $0.2q$ and $0.2(1-q)$; block two is (Y_3, Y_4) with 00, 10 and 01 of the same probabilities. Put $q = \frac{1}{2} + \epsilon$, so all four means are strictly ordered while the two within-block comparisons are nearly tied. Here $K = 6$, the four cross-block comparisons reject at thresholds $\delta/6$ through $\delta/3$, and the two tied comparisons face $\delta/2$ and δ.

We first fix the number m of discordant items in each tied pair and let $\epsilon \to 0$. Let $F(t)$ be the probability that a tied pair produces a p-value at most t, and note that its direction is wrong with probability $\frac{1}{2}$ independently of that. Writing $u = \delta/2$ and $v = \delta$, the pair with the smaller p-value is tested at u and the other at v, so

$$\mathbb{P}(\text{no inversion}) = (1 - F(u))^2 + \tfrac{1}{2}\,\mathbb{P}(s \le u,\, l > v) + \tfrac{1}{4}\,\mathbb{P}(s \le u,\, l \le v), \tag{12}$$

with s and l the smaller and larger of the two p-values. Evaluating (12) exactly, at m discordant items per tied pair and $\epsilon \to 0$:

m	$\mathbb{P}$(certified inversion)
2,000	0.04710
20,000	0.04916
200,000	**0.05049**
2,000,000	**0.05053**

Writing $A = F(u)$ and $B = F(v)$, (12) gives an error of $A - \frac{3}{4}A^2 + \frac{1}{2}AB$, and as m grows $A \to \delta$ and $B \to 2\delta$, so the error tends to $\delta + \delta^2/4 = 0.050625$. The excess grows with m, so it is not an artefact of discreteness. The same construction under the uniform budget gives 0.0165, and under Theorem 4 it gives 0.00082.

The i.i.d. instance of Proposition 5. The table fixes m and lets $\epsilon \to 0$. Proposition 5 is about the experiment itself, with n i.i.d. items, $q = \frac{1}{2} + \epsilon$ fixed and positive, and random discordant counts $m_1, m_2 \sim \mathrm{Bin}(n, 0.2)$, independent across the two blocks. Given m, the count favouring the better model is $\mathrm{Bin}(m, q)$, so every term of (12) is an exact binomial sum, and averaging over m_1 and

m_2 is exact. The only other event is that some cross-block comparison fails to clear $\delta/6$; on its complement Holm's decisions on the tied pairs are exactly the rule above, because a tied p-value sorted ahead of a cross-block one lies below $\delta/6$ and passes wherever it sits. A cross-block pair has $\mathbb{P}(D = +1) - \mathbb{P}(D = -1) \geq 0.8 - 0.4\epsilon$, and Hoeffding's inequality bounds the chance that its p-value exceeds $\delta/6$ by $\exp(-3 \times 10^5)$ at $n = 10^6$. With $n = 10^6$ and $\epsilon = 10^{-6}$ the probability of a certified inversion is at least 0.05025, and at $n = 2 \times 10^6$ at least 0.05028. The coefficients $\frac{1}{2}$ and $\frac{1}{4}$ of (12) assume an exact tie; at $\epsilon > 0$ the computation splits each event into its right and wrong directions and uses the exact probabilities of each. For fixed ϵ and $n \to \infty$ the error tends to zero, which is why the instance fixes n.

The valid use of Holm. Pre-specify both one-sided nulls for every pair, $\theta_j \leq \theta_k$ and $\theta_k \leq \theta_j$, with the valid p-values $\mathbb{P}\{\mathrm{Bin}(m, \frac{1}{2}) \geq n_{j \succ k}\}$, and run Holm over all $2K$ of them, reporting j above k when $\theta_j \leq \theta_k$ is rejected. This controls the probability of any certified inversion at δ under any dependence. It is never better than the uniform-budget certificate: at most one null per pair can be rejected at these levels, so every rejection comes among the first K and faces at most $\delta/(K+1) < \delta/K$, and every edge it certifies is also a uniform-budget edge. On held-out items it certifies as many relations as the uniform-budget certificate in ten of the eighteen configurations and up to four fewer in the rest (Table 9).

Directional errors and complete-order errors. The term in (12) that pushes the total past δ is the one where only the smaller p-value rejects, and a complete order needs both. The complete-order error probabilities are 0.00490, 0.00547, 0.00557 and 0.00561 at the four values of m above. The counterexample concerns the certified-edge output of §2.4. Closed testing (Marcus et al., 1976) controls the familywise error under arbitrary dependence when its local tests are valid.

The evidence for Proposition 5 is the exact computation.

G THE WEIGHTED SCHEME AND THE INTERSECTION–UNION TEST

Error budgets are free, so a prior ordering of the models can be spent on the comparisons it expects to be hard. This appendix reports what that buys for the edge certificate.

The scheme. Given an ordering σ from any source fixed before the validation items, put

$$\delta_e = \frac{\alpha\delta}{M-1} \text{ if } e \text{ is } \sigma\text{-adjacent}, \qquad \delta_e = \frac{(1-\alpha)\delta}{K-M+1} \text{ otherwise}, \tag{13}$$

with $\alpha \in (0, 1)$. Validity is immediate from Theorem 3 and a union bound, for any σ whatsoever, because the weights are independent of the validation sample. A pair's required sample size scales as $\log(1/\delta_e)$. Against the uniform split a correct prior gains a factor $\log(K/\delta)/\log((M-1)/(\alpha\delta))$. An arbitrarily wrong one loses at most $\log((K-M+1)/((1-\alpha)\delta))/\log(K/\delta)$. At $\alpha = 0.9$ and $\delta = 0.05$ those factors are 1.28 and 1.31 at $M = 10$, and 1.36 and 1.27 at $M = 20$. Setting $\alpha = 1$ removes the floor and makes the loss unbounded, since a pair given zero budget never certifies.

What it is worth. We ran (13) at $\alpha = 0.9$ against the equal split δ/K on the same items with the same test, on four configurations at two budgets, 600 paired replicates per cell, judged by a paired two-standard-error rule. With the order estimated from a pilot of 5% of the items, charged only to the weighted scheme, it is better in six of the eight cells and inconclusive in two. The largest gain is GSM8K at $M = 20$ on set R3 at $40{,}000$ items, from 0.327 to 0.603. With the true order it is better in all eight. With one adjacent transposition it is better in six and inconclusive in two, and with three it is better in four, inconclusive in three and worse in one, which is the bounded loss the floor guarantees. The equal split is itself dominated by Theorem 4 for complete rankings, so the weights matter for the certified partial order and not for the headline comparison.

The intersection–union test. A chain test on a frozen order tests each σ-adjacent pair at the full level δ and takes the intersection. Under a correct σ its completion rate equals Holm's in seven of the 8 cells we ran and exceeds it by 0.006 in the eighth. Under a single wrong adjacent transposition its completion rate is 0 in all 8 cells, and the same under three. With a pilot-estimated order it ranges

from 0.108 to 0.853, while Holm on the full sample ranges from 0.930 to 1.000. Holm's figures are descriptive completion rates; Proposition 5 establishes a directional-error counterexample for this use of Holm.

The two results bracket what a prior is worth. A chain test takes all of it and loses everything when the order is wrong, while the floored scheme of (13) takes most of it and bounds the loss.

H A BANK THAT CERTIFIED, AND A BANK SIZE THAT CERTIFIES

HellaSwag at $M = 20$ closes on its $10{,}042$ items, while Theorem 1 puts the requirement at $11{,}696$. Both statements hold, because the bound constrains the probability of success and one realised bank can succeed below it.

All four HellaSwag configurations close on the observed bank under Theorem 4. Resampling each bank at its own size with $8{,}000$ paired bootstrap (Efron, 1979) draws gives the probability that a bank of that size would close, under that bank's own empirical law.

Configuration	Bank	Theorem 4	uniform budget	step-down
HellaSwag, M=10, R1	10,042	0.060	0.001	0.598
HellaSwag, M=10, R3	10,042	0.406	0.141	0.827
HellaSwag, M=20, R1	10,042	0.003	0.000	0.487
HellaSwag, M=20, R3	10,042	0.022	0.000	0.610

The right-hand column reports descriptive completion rates for the step-down procedure of Appendix F. Under Theorem 4 the figures are 0.3% to 40.6%, so a bank that proved its leaderboard is far from a bank size that reliably would.

I ITEMS, TOKENS AND REPEATED ANSWERS

Every result in the main text is stated in items per model. This appendix relates that unit to what an evaluation run actually spends.

Evaluating one model on one item is a *unit*, and opening a unit costs a fixed amount κ before any answer is produced, because the prompt must be read and the harness must run. Asking the same question again costs one more, so a unit opened with r repeats costs

$$\kappa + r, \qquad \text{so a run that opens } n \text{ items on each of } M \text{ models costs } Mn(\kappa + r). \tag{14}$$

We measure κ in tokens on real runs, and it is about 64 for short-answer tasks, falling to about 0.55 once the model writes out a chain of thought, because a long answer dwarfs the prompt. Setting $\kappa = 0$ recovers the usual accounting in which every sample costs the same.

When a unit is opened with $r > 1$ answers, its score is the mean of the r binary answers, and the within-unit repeats reduce only the answer-level noise. At a fixed target variance, the continuous cost-optimal number of repeats follows the square-root rule of Dickmann & Schweizer (2016). Over the integers, $r = 2$ is cheaper than $r = 1$ if and only if $\kappa\,\sigma_w^2 > 2\,\sigma_b^2$ (Proposition 12). All experiments in the main text use $r = 1$.

J A PROPOSAL–VALIDATION RANKING CERTIFICATE

Theorem 4 spends $\delta/(M-1)$ on every pair because it does not know which pairs witness an error. Buying the candidate order first makes the witnesses known.

Proposition 7 (Proposal–validation certificate)**.** *Split the items into a proposal part and a validation part, independently. Order the models by their proposal-part means to obtain $\hat{\pi}$. On the validation part, test each of the $M-1$ comparisons adjacent in $\hat{\pi}$, one-sided in the direction $\hat{\pi}$ claims, and certify $\hat{\pi}$ when every p-value is at most δ. Then the probability of certifying an order that contains an inversion is at most δ.*

Proof. Condition on the proposal part, which fixes $\hat{\pi}$. If $\hat{\pi}$ contains an inversion, some pair (a, b) adjacent in $\hat{\pi}$ has $\theta_a < \theta_b$, since otherwise the population means are non-increasing along $\hat{\pi}$. Certifying $\hat{\pi}$ requires that pair's validation p-value to be at most δ in the direction a above b. The validation items are independent of $\hat{\pi}$, so that direction is fixed in advance and Theorem 3 bounds the event by δ. Averaging over the proposal completes the proof. One witness suffices and it is fixed once $\hat{\pi}$ is, so no union bound is needed. □

The threshold is δ in place of $\delta/(M-1)$, and the price is the items spent on the proposal. With 30% of the items used to propose, which was the best of 10, 20, 30 and 40% on every configuration, the requirement against Theorem 4 is:

Configuration	Theorem 4	Proposal–validation	Items saved
GSM8K, M=20, R1	2.00	1.91	+4.7%
GSM8K, M=20, R3	1.99	1.80	+9.9%
MMLU, M=10, R1	1.96	2.00	−1.8%
MMLU, M=10, R3	1.62	1.52	+6.5%
MMLU, M=20, R1	2.24	2.08	+7.1%
MMLU, M=20, R3	2.12	1.94	+8.4%
ARC-C, M=10, R1	2.08	2.14	−2.9%
ARC-C, M=10, R3	1.62	1.55	+4.1%
ARC-C, M=20, R1	2.43	2.35	+3.3%
ARC-C, M=20, R3	2.14	1.96	+8.8%
HellaSwag, M=10, R1	1.88	1.91	−1.7%
HellaSwag, M=10, R3	1.72	1.72	+0.1%
HellaSwag, M=20, R1	2.24	2.15	+3.9%
HellaSwag, M=20, R3	2.13	1.94	+8.8%
WinoGrande, M=10, R1	2.05	2.10	−2.6%
WinoGrande, M=10, R3	1.87	1.87	−0.3%
WinoGrande, M=20, R1	2.38	2.22	+6.6%
WinoGrande, M=20, R3	2.22	2.04	+8.0%

Ratios are to the single-pair bound, and n_{95} is read off an interpolated curve of complete and correct certification with 3,000 draws per point. At $M = 20$ the split certificate is better on all ten configurations, by 3.3% to 9.9%. At $M = 10$ it is better on three of eight and worse by up to 2.9% on the rest, because the threshold gains a factor of nine there, against nineteen at $M = 20$, and an order proposed from 30% of the items is wrong often enough to cost more than that. We report Theorem 4 in the main text because it needs no split and wins where the split does not, and we note that for large leaderboards the split certificate is the better choice.

K Scores that are not binary

The main text assumes binary scores, since exact match and multiple choice are how most leaderboards score an answer, and Theorem 3 uses that assumption in an essential way. This appendix states what survives when a score is any number in $[0, 1]$, for example a mean over several sampled answers, a partial-credit grade, or a judge's rating rescaled to the unit interval.

Setting. Model m receives a score $Y_m \in [0, 1]$ on each item, the items are drawn independently from G and every model sees the same n items. For a pair $e = (j, k)$ the paired differences $D_{e,1}, \ldots, D_{e,n}$ lie in $[-1, 1]$, are independent and identically distributed with mean $\Delta_e = \theta_j - \theta_k$, and are dependent across pairs in an arbitrary way. Write $\bar{D}_e$ for their mean and $\hat{V}_e$ for their unbiased sample variance.

The certificate. Fix levels $\delta_e \geq 0$ with $\sum_e \delta_e \leq \delta$ before the sample is drawn, and put

$$\mathrm{rad}_e \;=\; \sqrt{\frac{2\hat{V}_e \log(2/\delta_e)}{n}} \;+\; \frac{14 \log(2/\delta_e)}{3(n-1)}. \tag{15}$$

Certify $j \to k$ when $\bar{D}_e > \mathrm{rad}_e$ and $k \to j$ when $\bar{D}_e < -\mathrm{rad}_e$, and leave the pair unresolved otherwise.

Theorem 8 (Certificate for bounded scores)**.** *For any levels fixed as above, the probability that some certified direction is an inversion is at most δ. Nothing is assumed about the item law, the variances, the gaps or the dependence between pairs.*

Proof. Fix a pair e and let B_e be the event that it is certified in a direction that is an inversion. The gap Δ_e is a fixed number, so exactly one of three cases holds. If $\Delta_e > 0$, an inversion at e means certifying $k \to j$, which requires $\bar{D}_e < -\mathrm{rad}_e$ and hence $\bar{D}_e - \Delta_e < -\mathrm{rad}_e$. If $\Delta_e < 0$, then symmetrically B_e implies $\bar{D}_e - \Delta_e > \mathrm{rad}_e$. If $\Delta_e = 0$, neither direction is an inversion and B_e is empty.

In each case B_e lies inside a one-sided deviation event, and the side is fixed by Δ_e and not by the data. Put $X_i = (D_{e,i} + 1)/2 \in [0, 1]$. These are independent with mean $(\Delta_e + 1)/2$ and unbiased sample variance $\hat{V}_e/4$. The empirical Bernstein inequality of Maurer & Pontil (2009) gives, with probability at least $1 - \delta_e$,

$$\mathbb{E}X - \bar{X} \;\le\; \sqrt{\frac{2(\hat{V}_e/4)\log(2/\delta_e)}{n}} + \frac{7\log(2/\delta_e)}{3(n-1)},$$

and applying the same inequality to $1 - X_i$, whose sample variance is also $\hat{V}_e/4$, gives the bound for $\bar{X} - \mathbb{E}X$. Multiplying by two and using $\bar{X} - \mathbb{E}X = (\bar{D}_e - \Delta_e)/2$ turns either statement into the one-sided bound $\pm(\bar{D}_e - \Delta_e) \le \mathrm{rad}_e$ that B_e violates. Hence $\mathbb{P}(B_e) \le \delta_e$ in every case, and a union bound over pairs, which needs no independence, gives $\sum_e \delta_e \le \delta$. □

Why one tail is enough. The data choose which direction to certify, and it is tempting to pay for both tails. Only one tail is ever dangerous, because the truth is not random: for a given pair only one direction can be an inversion, and the bad event is defined by that fixed direction and not by the observed sign. The one-sided form uses $\log(2/\delta_e)$ where a two-sided one would use $\log(4/\delta_e)$. Since the leading term of the sample size is linear in that logarithm, the two-sided version would need about 19% more items for a single pair at $\delta = 0.05$, and 7.8% to 9.2% more at the split δ/K for K between 45 and 190.

What the binary case adds. For binary scores $D_{e,i} \in \{-1, 0, +1\}$, and the sign of Δ_e is the sign of $p_+ - p_-$ (7). Theorem 3 tests that sign exactly, with no variance estimate and no concentration slack, and Appendix T measures what this saves. Theorem 4 also has a bounded-score analogue. Its proof uses only that each π^*-adjacent event E_{ab} has probability at most $\delta/(M-1)$, so it holds verbatim when the exact p-value is replaced by a one-sided test that is valid at level $\delta/(M-1)$ for $\Delta_{ab} \ge 0$, for example (15) with $\delta_e = \delta/(M-1)$.

L Cycles, abstention and forced rankings

A certificate returns a set of certified directions. This appendix states how to turn that set into an output, and what can be guaranteed when an evaluator insists on a complete ranking whether or not the data support one.

The output rule. Fix per-pair tests that are valid at levels δ_e with $\sum_e \delta_e \le \delta$: the exact test of Theorem 3 for binary scores, or (15) otherwise. Draw a directed edge $j \to k$ for every certified direction. If the resulting digraph has a directed cycle, abstain. In *certified mode*, output a ranking only if every pair is certified and the digraph is acyclic, in which case the ranking is its unique topological order, and otherwise abstain. In *forced mode*, always output the lexicographically first topological order of the certified digraph, which completes the certified edges to a total order without adding any claim of its own.

Lemma 9 (Cycles on the good event)**.** *On the event that no certified direction is an inversion, every certified edge $j \to k$ has $\theta_j \ge \theta_k$. Any directed cycle then consists of models with exactly equal θ, and ordering those models in any way creates no inversion.*

Proof. A certified $j \to k$ with $\theta_k > \theta_j$ would be an inversion, so $\theta_j \ge \theta_k$. Around a cycle $j_1 \to j_2 \to \cdots \to j_L \to j_1$ this gives $\theta_{j_1} \ge \theta_{j_2} \ge \cdots \ge \theta_{j_L} \ge \theta_{j_1}$, which forces equality. An inversion needs a strict inequality, so no order of a tied set contains one. □

On the good event a cycle carries no information beyond a tie, and off it the certificate has already failed, so abstaining on a cycle costs nothing. A cycle among models whose scores are visibly different is evidence that the good event did not hold.

The shared-item design has no cycles at all. When every model is scored on the same items with equal weights, $\bar{D}_{jk} = \bar{Y}_j - \bar{Y}_k$ holds identically, and a certified $j \to k$ requires $\bar{Y}_j > \bar{Y}_k$, both for (15) and for Theorem 3, where $n_+ > n_-$ is the same condition. Every certified edge therefore points down the order of a single vector of sample means, so the certified digraph is acyclic on every sample, whether or not the good event holds. Lemma 9 is empty for the design this paper uses.

It stops being empty as soon as models are scored on different numbers of items, which any procedure that allocates the budget unevenly will do. Table 3 measures this on the eighteen configurations. Each model j receives N_j items drawn uniformly from $[(1-c)n, n]$, a pair is compared on the first $\min(N_j, N_k)$ shared items, and we record how often the digraph of empirical comparisons, with no certification threshold at all, contains a directed cycle. At $c = 0$ the rate is zero by the identity above. At $n = 2{,}000$ and $c = 0.75$ it reaches 0.286, and it falls to at most 0.019 at $n = 10{,}000$. When the relation has a cycle, "compare the pairs, then sort" has no answer until a tie-breaking convention supplies one, and the convention then decides the output. Cycles among *certified* edges are rarer, since they need an error or an exact tie, but empirical comparisons are what a forced ranking falls back on.

Table 3: How often the empirical comparison digraph contains a directed cycle when models are scored on unequal numbers of items, over $2{,}000$ replicates per cell. The column c is the allocation spread: each model's item count is uniform on $[(1-c)n, n]$. The column $c = 0$ is the shared-item design, where the rate is zero by an identity.

	$n = 2{,}000$				$n = 10{,}000$			
Configuration	$c = 0.0$	$c = 0.25$	$c = 0.5$	$c = 0.75$	$c = 0.0$	$c = 0.25$	$c = 0.5$	$c = 0.75$
GSM8K, M=20, R1	0.000	0.006	0.036	0.144	0.000	0.000	0.001	0.004
GSM8K, M=20, R3	0.000	0.000	0.003	0.022	0.000	0.000	0.000	0.000
MMLU, M=10, R1	0.000	0.013	0.040	0.158	0.000	0.000	0.002	0.009
MMLU, M=10, R3	0.000	0.000	0.002	0.016	0.000	0.000	0.000	0.000
MMLU, M=20, R1	0.000	0.021	0.100	0.286	0.000	0.000	0.001	0.019
MMLU, M=20, R3	0.000	0.002	0.015	0.088	0.000	0.000	0.001	0.002
ARC-C, M=10, R1	0.000	0.008	0.025	0.097	0.000	0.000	0.000	0.005
ARC-C, M=10, R3	0.000	0.001	0.002	0.015	0.000	0.000	0.000	0.000
ARC-C, M=20, R1	0.000	0.009	0.053	0.171	0.000	0.000	0.001	0.006
ARC-C, M=20, R3	0.000	0.004	0.015	0.067	0.000	0.000	0.001	0.001
HellaSwag, M=10, R1	0.000	0.001	0.004	0.030	0.000	0.000	0.000	0.000
HellaSwag, M=10, R3	0.000	0.001	0.001	0.005	0.000	0.000	0.000	0.000
HellaSwag, M=20, R1	0.000	0.001	0.006	0.051	0.000	0.000	0.000	0.001
HellaSwag, M=20, R3	0.000	0.001	0.004	0.021	0.000	0.000	0.000	0.000
WinoGrande, M=10, R1	0.000	0.006	0.035	0.103	0.000	0.000	0.000	0.005
WinoGrande, M=10, R3	0.000	0.002	0.005	0.028	0.000	0.000	0.000	0.003
WinoGrande, M=20, R1	0.000	0.015	0.061	0.226	0.000	0.000	0.001	0.004
WinoGrande, M=20, R3	0.000	0.006	0.041	0.140	0.000	0.000	0.000	0.005

Theorem 10 (Certified mode)**.** *In certified mode, the probability of outputting a ranking that contains an inversion is at most δ, for every n, every item law and every configuration of gaps.*

Proof. A ranking is output only if every pair is certified and the digraph is acyclic. On the good event of Theorem 3 or Theorem 8, which has probability at least $1 - \delta$, every certified edge agrees with θ weakly by Lemma 9. The topological order of a complete acyclic tournament then places j above k only when $\theta_j \geq \theta_k$, which is not an inversion. □

This is the uniform-budget certificate of the main text. Theorem 4 improves on it while still testing all K pairs, by a union bound over $M - 1$ fixed witness events.

Theorem 11 (Forced mode)**.** *In forced mode, fix $\varepsilon \geq 0$ and call an output wrong by more than ε if it places some j above some k with $\theta_k - \theta_j > \varepsilon$. Then*

$$\mathbb{P}\{\textit{the output is wrong by more than } \varepsilon\} \;\leq\; \delta \;+\; \mathbb{P}\{\textit{some pair with } |\Delta_e| > \varepsilon \textit{ is unresolved}\}.$$

Proof. Suppose the output places j above k with $\theta_k - \theta_j > \varepsilon \geq 0$, so $|\Delta_{jk}| > \varepsilon$. If the pair was certified, it was certified as $j \to k$, because a certified $k \to j$ is an edge of the digraph and a topological order respects it. That is an inversion, and all such events together have probability at most δ. Otherwise the pair was unresolved. A union bound over the two cases gives the claim. □

The theorem separates the two things a forced ranking can get wrong. The first term is validity, and it is paid for by the choice of levels. The second is power, and it depends on the unknown gaps and variances and on n. On our banks the gap between the two is large. The order of the sample means has no inversion with probability 0.95 at 0.26 to 0.57 times the Theorem 1 bound (the dash-dotted curves of Figures 4 to 6), so an evaluator who sorts is usually right at a budget where Theorem 1 shows that no procedure can yet certify that it is. That difference is exactly the part of a leaderboard's correctness that nothing certifies.

M How many repeats to buy

Appendix I prices a unit opened with r answers at $\kappa + r$. If an evaluator can ask the same question more than once, the choice of r trades items against repeats. This appendix states the optimum. The square root is not ours, since it is Proposition 3.1 of Dickmann & Schweizer (2016), who solve the same problem for nested Monte Carlo. We state it in the evaluation setting because the integer version decides whether a second answer is worth buying, and the continuous version gets that decision wrong by a factor of two.

Proposition 12 (Repeats under a fixed overhead)**.** *Let the score of a unit opened with r answers have variance $\sigma_b^2 + \sigma_w^2/r$, where $\sigma_b^2 = \mathrm{Var}_i(p_{m,i})$ is the variance of the item-level success probability across items and $\sigma_w^2 = \mathbb{E}_i \,\mathrm{Var}(Y \mid i)$ is the answer-level variance within an item. At a fixed target variance, the cost is proportional to $f(r) = (\kappa + r)(\sigma_b^2 + \sigma_w^2/r)$, which is minimised over real $r > 0$ at*

$$r^* = \sqrt{\kappa\, \sigma_w^2/\sigma_b^2}.$$

Over the integers, $r = 2$ is cheaper than $r = 1$ if and only if $\kappa\, \sigma_w^2 > 2\, \sigma_b^2$.

Proof. With n items the variance of the mean is $(\sigma_b^2 + \sigma_w^2/r)/n$, so reaching a target v needs $n = (\sigma_b^2 + \sigma_w^2/r)/v$ items at cost $n(\kappa + r)$ each per model, which is $f(r)/v$. Expanding, $f(r) = \kappa\sigma_b^2 + \sigma_w^2 + r\sigma_b^2 + \kappa\sigma_w^2/r$, so $f'(r) = \sigma_b^2 - \kappa\sigma_w^2/r^2$, which vanishes only at r^*, and $f''(r) = 2\kappa\sigma_w^2/r^3 > 0$. For the integer comparison, $f(1) - f(2) = (\kappa\sigma_w^2 + \sigma_b^2) - (\kappa\sigma_w^2/2 + 2\sigma_b^2) = \kappa\sigma_w^2/2 - \sigma_b^2$, which is positive exactly when $\kappa\sigma_w^2 > 2\sigma_b^2$. □

The continuous optimum exceeds one when $\kappa\sigma_w^2 > \sigma_b^2$, and rounding it up would buy a second answer in the band $\sigma_b^2 < \kappa\sigma_w^2 \leq 2\sigma_b^2$, where it does not pay. With $\kappa \approx 64$ on short-answer tasks, a second answer pays as soon as the within-item share of the variance exceeds about one part in 33. With $\kappa \approx 0.55$ on chain-of-thought tasks it pays only if σ_w^2 is more than 3.6 times σ_b^2. Under greedy decoding $\sigma_w^2 = 0$ and repeats buy nothing.

Theorem 3 certifies binary scores; for scores formed by averaging $r > 1$ answers, use Theorem 8. The proposition gives a planning rule for specified variance components, including σ_w^2. The leaderboards of Section 5 contain one answer per model and item.

N The procedures, step by step

This appendix writes out every procedure in the paper at the level of detail needed to reimplement it. The certificates are what an evaluator runs on their own data. The bounds are what we compute on a bank to say how far the certificates are from the best possible. Each procedure names the script in the released code that implements it.

N.1 CERTIFICATES

Procedure 1: the exact p-value of one pair (Theorem 3).

Input: the binary scores of models j and k on the same n items.

1. Count n_+, the items where j is right and k is wrong, and n_-, the items where k is right and j is wrong. Put $m = n_+ + n_-$. Items where both are right or both are wrong are ignored.
2. Put $p_{jk} = \mathbb{P}\{\mathrm{Bin}(m, \frac{1}{2}) \geq \max(n_+, n_-)\}$, the upper tail of a fair binomial.
3. The direction is $j \to k$ if $n_+ > n_-$ and $k \to j$ if $n_- > n_+$. If $n_+ = n_-$, then $p_{jk} \geq \frac{1}{2}$ and the pair is never certified at any level below one half.

Output: p_{jk} and its direction. Certifying at level t means $p_{jk} \leq t$. *Code:* `w4_provable.pvals`.

Procedure 2: the ranking certificate (Theorem 4).

Input: an $n \times M$ binary score matrix on shared items, and δ.

1. Run Procedure 1 on all $K = M(M-1)/2$ pairs.
2. If every $p_{jk} \leq \delta/(M-1)$, output the order of the sample means as certified.
3. Otherwise output no total order.

Guarantee: the probability of outputting an order with an inversion is at most δ, under any dependence between models. *Code:* `w4_provable.reject_rank`.

The rule tests all K pairs, including the $K - M + 1$ that are not adjacent in the sample order. That is what the proof needs: the witnesses are adjacent in the *true* order, which is unknown, and every true-adjacent pair must have cleared the threshold.

Procedure 3: the certified partial order (uniform budget).

Input: as Procedure 2, and a level t fixed in advance, δ/K alone or $\delta/(2K)$ when Procedure 2 is run on the same sample at $\delta/2$.

1. Run Procedure 1 on all K pairs.
2. Draw an edge $j \to k$ for every pair with $p_{jk} \leq t$, in its direction.
3. Output the transitive closure of these edges. On shared items it has no cycles (Appendix L), so it is a partial order.

Guarantee: with probability at least $1 - Kt$, no edge in the output is an inversion. The transitive closure adds none, because on that event every edge agrees with θ weakly and weak agreement is transitive. *Code:* `w4_provable.reject_bonf`.

The closure is where the partial order earns its keep. On a bank where only the wide gaps clear δ/K, a chain of certified edges $a \to b \to c$ certifies a above c with no test of its own.

Procedure 4: the proposal–validation certificate (Appendix J).

Input: as Procedure 2, and a proposal fraction η, which is 0.3 in all our experiments.

1. Split the items at random into a proposal part of $\lfloor \eta n \rceil$ items and a validation part of the rest. The split is fixed before any score is looked at.
2. Order the models by their proposal-part means to get $\hat{\pi}$.
3. On the validation part, run Procedure 1 on the $M - 1$ pairs that are adjacent in $\hat{\pi}$, each in the direction $\hat{\pi}$ claims: with u validation items favouring the claimed winner and v favouring the other model, $p = \mathbb{P}\{\mathrm{Bin}(u + v, \frac{1}{2}) \geq u\}$, and $p = 1$ when $u + v = 0$. A pair whose counts point the other way has $p > \frac{1}{2}$.
4. If all $M - 1$ p-values are at most δ, output $\hat{\pi}$. Otherwise output no total order.

Code: `w11_split.split_rates`.

N.2 BOUNDS

Procedure 5: the single-pair bound (Theorem 1).

Input: a score matrix that defines $\mathbb{P}$, and δ.

1. Order the models by their means under $\mathbb{P}$. For each of the $M-1$ adjacent pairs with $p_+ > p_- > 0$, compute p_+, p_-, s, a and the per-item divergence J of (4).
2. Take the three pairs with the smallest J. For each, find by bisection the smallest n at which the sum (10) reaches $1-2\delta$. The sum has $n+1$ terms, each a binomial probability times the total variation between two binomials, which is a finite sum of its own.
3. Report the largest of the three. The sum is truncated to terms of non-negligible weight, and the omitted mass is checked to leave the reported integer unchanged (Appendix C).

Code: `m11_r3.lower_bound`.

Every one of the three is a valid bound, so their maximum is too. The divergence J only picks which pairs to evaluate. It never enters the reported number, which is the exact sum. Three candidates are enough on every configuration here, since the pair with the smallest J also gives the largest exact bound in all eighteen.

Procedure 6: the mixture bound (Theorem 2).

Input: as Procedure 5, a number of draws R, and a set of candidate sizes k.

1. Sort the adjacent pairs by J, hardest first. For a set of k of them, the likelihood ratio of tilt r at the tie is $\log L_r = n_{r,+} \log \frac{s_r}{2p_{r,+}} + n_{r,-} \log \frac{s_r}{2p_{r,-}}$, where $n_{r,\pm}$ are pair r's discordant counts in the sample.
2. Estimate $\mathrm{TV} = \mathbb{E}_{\mathbb{P}}\big[(1 - \frac{1}{k}\sum_r L_r)_+\big]$ by drawing n items from $\mathbb{P}$ R times, and record the mean and its standard error.
3. Call n excluded if the upper end of a 95% interval on TV is below $1-2\delta$. Walk a grid of step 3% upward from 0.85 times the single-pair bound to bracket the first non-excluded n, and bisect inside the bracket for the last excluded n. The bound is that n plus one.
4. Choose k on one set of $R/3$ draws, and recompute the bound for the chosen k on a fresh set of R draws.
5. *Certification.* Freeze k and n_0, the bound of step 4 minus one. Draw $400{,}000$ fresh samples of n_0 items, and bound $\mathbb{E}Z$, $Z = (1 - \frac{1}{k}\sum_r L_r)_+ \in [0,1]$, from above by the empirical Bernstein inequality of Maurer & Pontil (2009) at level $\gamma/36$, $\gamma = 0.05$. If the upper bound is below $1-2\delta$, report $n_0 + 1$. Otherwise try once more at $n_0 = \lfloor 0.97 \times \text{bound} \rfloor$, and if that fails too, report the Theorem 1 bound.

Code: `w13_mixture.smallest_admissible`, `w24_mixture_verify.py`.

Steps 1 to 4 are a search, and a search does not by itself give its output a guarantee. Step 5 does, because the total variation is non-decreasing in n, since the first n of $n+1$ draws are a function of all of them under both laws, and the n-draw marginal of a mixture of $(n{+}1)$-fold products is the same mixture of n-fold products. One excluded n_0 therefore excludes every smaller n, and $n_0 + 1$ is a bound. With two pre-specified attempts on each of eighteen configurations, a union bound makes all reported bounds hold together with probability at least $1-\gamma$ over the Monte Carlo draws. All sixteen mixture bounds that exceed Theorem 1 were certified at the first attempt, so every reported number is unchanged. Sampling is fast because only the k pairs' discordance matters, so items with the same pattern of ± 1 over those pairs are pooled, and a sample is a multinomial draw over at most a few thousand patterns.

Table 4 gives, for each test, the frozen k and n_0, the mean and variance of Z over the $400{,}000$ fresh draws, the upper bound U, and its margin below $1-2\delta = 0.9$. The smallest margin is 0.0022. The last column is the omitted mass τ of the Theorem 1 sum at $L-1$ (Appendix C).

Table 4: The certification test of every Theorem 2 bound, and the truncation of every Theorem 1 sum. Rows with dashes are the two configurations where Theorem 2 does not exceed Theorem 1, which is then the reported bound.

Configuration	k	n_0	$\bar{Z}$	$\widehat{\mathrm{Var}}(Z)$	U	$0.9 - U$	Thm 1 tail τ
GSM8K, M=20, R1	8	33,820	0.8960	0.0550	0.8975	0.0025	0
GSM8K, M=20, R3	2	35,352	0.8930	0.0612	0.8946	0.0054	4×10^{-16}
MMLU, M=10, R1	4	41,098	0.8915	0.0586	0.8930	0.0070	7×10^{-16}
MMLU, M=10, R3	–	–	–	–	–	–	3×10^{-16}
MMLU, M=20, R1	12	42,606	0.8951	0.0540	0.8965	0.0035	7×10^{-16}
MMLU, M=20, R3	3	40,380	0.8914	0.0598	0.8929	0.0071	7×10^{-16}
ARC-C, M=10, R1	9	29,871	0.8912	0.0564	0.8927	0.0073	1×10^{-15}
ARC-C, M=10, R3	–	–	–	–	–	–	4×10^{-15}
ARC-C, M=20, R1	10	32,764	0.8942	0.0553	0.8957	0.0043	1×10^{-15}
ARC-C, M=20, R3	4	30,143	0.8957	0.0563	0.8972	0.0028	0
HellaSwag, M=10, R1	5	15,371	0.8887	0.0588	0.8902	0.0098	8×10^{-16}
HellaSwag, M=10, R3	2	13,846	0.8963	0.0584	0.8978	0.0022	0
HellaSwag, M=20, R1	8	14,851	0.8939	0.0561	0.8954	0.0046	1×10^{-16}
HellaSwag, M=20, R3	5	15,927	0.8944	0.0566	0.8958	0.0042	6×10^{-16}
WinoGrande, M=10, R1	8	32,400	0.8940	0.0555	0.8954	0.0046	2×10^{-15}
WinoGrande, M=10, R3	3	27,186	0.8954	0.0575	0.8969	0.0031	0
WinoGrande, M=20, R1	8	37,044	0.8921	0.0569	0.8936	0.0064	6×10^{-16}
WinoGrande, M=20, R3	5	37,438	0.8953	0.0566	0.8968	0.0032	1×10^{-16}

N.3 Measuring a certificate

Procedure 7: n_{95} in replay (Appendix O).

Input: a bank of N items, a certificate, and a grid of n.

1. At each n, draw $8{,}000$ samples of n items from the bank with replacement. A sample is stored as the vector C of how often each bank item was drawn, so every pair's discordant counts are C times a fixed $0/1$ vector, for all pairs at once.
2. Run the certificate on each sample and record whether it certified the complete order and whether that order agrees with the bank order.
3. Take the running maximum of the completion rate along the grid, and interpolate $\log n$ linearly between the two grid points that bracket 0.95. On all 392 grid points of the eighteen curves the completion rate and the complete-and-correct rate are equal, so the choice does not matter here.
4. For the interval, resample each point's rate from its binomial and repeat step 3.

Code: `w9_measure.batch_rates`, `w12_interp.crossing`.

The power of a discrete test need not grow with n. With every item discordant and $q = 0.6$, the exact test at level 0.05 certifies with probability $0.6^5 = 0.078$ at $n = 5$ and $0.6^6 = 0.047$ at $n = 6$. So the running maximum cannot rest on monotonicity. On our curves it is inert, since none of the 748 steps between grid points decreases, so every n_{95} equals the first crossing of the raw curve.

O Experimental protocol in full

Data. All scores come from the correctness matrix released by Polo et al. (2024), which records the binary score of 395 open models on every item of each benchmark. The file is checked against a recorded SHA-256 digest before any use. Five benchmarks enter: GSM8K ($1{,}319$ items), MMLU ($14{,}042$ items, the union of its 57 subject files), ARC-Challenge ($1{,}172$), HellaSwag ($10{,}042$) and WinoGrande ($1{,}267$). The last three were chosen by a rule written down before any outcome on them was computed: every top-level binary-scored benchmark in the file that is not a component of an aggregate already in use. TruthfulQA is excluded by that rule because its scores are not binary, and the 57 MMLU subject files because they are components of the MMLU aggregate.

Model filtering. A model is kept for a benchmark when its accuracy lies in $[0.05, 0.95]$, so that it is neither degenerate nor saturated, and when its score vector is not an exact duplicate of a model already kept. Two models with identical score vectors are indistinguishable on every item, so keeping both would add an exact tie that no amount of data resolves.

Model sets. Each configuration takes $M \in \{10, 20\}$ models from the filtered pool and orders them by accuracy. Adjacent models must differ in accuracy by at least 0.01, which keeps every comparison away from an exact tie. Set R1 is the chain of M models with the smallest span in accuracy subject to that floor; for a fixed starting model the greedy choice of the next admissible model minimises the span, so a scan over starts finds the optimum exactly. Set R3 keeps the same floor, allows a span up to twice R1's, and maximises the coefficient of variation of the per-pair difficulty $\mathrm{Var}(D_e)/\Delta_e^2$ over adjacent pairs, so that its comparisons differ as much as possible in how many items they need. R3 is found by a local search with 400 random admissible restarts and swap-based hill climbing, seeded at 7 and started from R1. Both rules read only accuracies and paired statistics of the published matrix, and both were fixed before any certificate was run on them.

The replay law. Every sampling experiment draws items independently and uniformly, with replacement, from a bank. The item law G is therefore the bank's empirical law, the true accuracies are the bank accuracies, and the true order is the bank order, so every validity statement is checked against a known truth. The lower bounds are evaluated at the same law.

Estimating n_{95}. For each configuration and rule we evaluate, on a geometric grid of n spaced 6% apart, the probability that the rule certifies the complete ranking and that the ranking is correct. Each grid point uses $8{,}000$ independent draws, generated as multinomial item counts so that the discordant counts of all K pairs follow from two matrix products. n_{95} is read off by log-linear interpolation at 0.95; a running maximum is applied first, and it changes no curve (Procedure 7). A parametric bootstrap redraws each grid point from a binomial with the estimated probability, $2{,}000$ times, and reports the 2.5 and 97.5 percentiles of the crossing. The estimator has no selection step, unlike a bisection that stops at the first noisy estimate above 0.95.

Evaluating Theorem 1. The hardest pairs are ranked by the per-item divergence J of (4), and the total variation of (5) is evaluated exactly for the three hardest, by summing the binomial law of the discordant count over its effective support. The bound is the smallest n at which it reaches $1 - 2\delta$, found by bisection on an exact, monotone quantity.

Evaluating Theorem 2. For k adjacent pairs taken in order of increasing J, and equal weights, the total variation equals $\mathbb{E}_{\mathbb{P}}[(1 - \bar{L})_+]$ with $\bar{L}$ the average of the k likelihood ratios, each a function of one pair's discordant counts. We estimate that expectation from $6{,}000$ multinomial draws of n items from the bank. A sample size is excluded only when the upper end of an approximate 95% interval on the total variation is below $1 - 2\delta$. A 3% grid brackets the crossing, integer bisection inside the bracket finds the last excluded n, and the reported bound is that value plus one. The number of pairs k is chosen on one set of draws from $k \in \{1, 2, 3, 4, 5, 6, 8, 10, 12, 15, M - 1\}$ and the bound is reported on an independent set. Each reported bound is then certified by one test on fresh draws (Procedure 6), so all eighteen hold together at computational confidence 0.95, which is separate from the level δ.

Seeds. Every random stream is seeded deterministically, from a SHA-256 digest of its configuration name and purpose or, for the model-set search and the bootstrap, from a fixed integer, so every run reproduces exactly (Appendix V).

Compute. All experiments run on one workstation CPU, with no accelerator. Nothing is trained and no model is queried, since every score is read from the released matrix.

P RESULTS FOR EVERY CONFIGURATION

Table 5 gives the per-configuration whole-bank outputs and replay values behind Figure 2; the per-configuration version of Table 1 is Table 9.

Table 6 gives, for each configuration, both lower bounds, the number k of pairs mixed in Theorem 2, n_{95} for Theorem 4 with its bootstrap interval, and n_{95} for the uniform budget and for the step-down procedure. The step-down procedure reaches 95% far earlier than the other two. Its column reports descriptive completion rates (Appendix F).

Table 5: The certificate on the whole bank, and against the lower bounds of §3. FULL marks whether Theorem 4 closes the ranking on the bank, and CERT and PRE are the relations and the proved prefix the uniform-budget certificate gives. SINGLE is Theorem 1, and BEST is the larger of it and Theorem 2; the first is computed exactly and the second by Monte Carlo from the replay law. n_{95} is for Theorem 4, and the ratio is n_{95} over BEST.

Benchmark	Set	Bank	K	FULL	CERT	PRE	SINGLE	BEST	n_{95}	Ratio
GSM8K, M=20	R1	1,319	190		136	0	30,432	33,821	60,931	1.80
GSM8K, M=20	R3	1,319	190		161	0	32,180	35,353	64,155	1.81
MMLU, M=10	R1	14,042	45		37	0	34,378	41,099	67,361	1.64
MMLU, M=10	R3	14,042	45		44	3	34,552	34,552	56,055	1.62
MMLU, M=20	R1	14,042	190		172	0	34,378	42,607	77,022	1.81
MMLU, M=20	R3	14,042	190		186	7	34,378	40,381	72,808	1.80
ARC-C, M=10	R1	1,172	45		19	0	23,238	29,872	48,339	1.62
ARC-C, M=10	R3	1,172	45		33	1	25,177	25,177	40,688	1.62
ARC-C, M=20	R1	1,172	190		128	0	23,238	32,765	56,450	1.72
ARC-C, M=20	R3	1,172	190		151	1	24,648	30,144	52,820	1.75
HellaSwag, M=10	R1	10,042	45	✓	42	6	13,867	15,372	26,072	1.70
HellaSwag, M=10	R3	10,042	45	✓	44	6	13,194	13,847	22,695	1.64
HellaSwag, M=20	R1	10,042	190	✓	184	6	11,696	14,852	26,199	1.76
HellaSwag, M=20	R3	10,042	190	✓	185	6	13,272	15,928	28,240	1.77
WinoGrande, M=10	R1	1,267	45		20	0	25,544	32,401	52,286	1.61
WinoGrande, M=10	R3	1,267	45		32	1	23,433	27,187	43,707	1.61
WinoGrande, M=20	R1	1,267	190		122	0	27,979	37,045	66,618	1.80
WinoGrande, M=20	R3	1,267	190		138	0	29,764	37,439	65,989	1.76

Table 6: Lower bounds and n_{95} for every configuration. The ratio is Theorem 4's n_{95} over the larger bound, as in Table 5.

Configuration	Thm 1	Thm 2 (k)	Theorem 4 n_{95}	95% interval	uniform budget	step-down	ratio
GSM8K, M=20, R1	30,432	33,821 (8)	60,931	[60,355, 61,644]	79,925	36,963	1.80
GSM8K, M=20, R3	32,180	35,353 (2)	64,155	[63,545, 64,848]	83,598	37,218	1.81
MMLU, M=10, R1	34,378	41,099 (4)	67,361	[66,543, 67,971]	82,651	45,250	1.64
MMLU, M=10, R3	34,552	33,290 (1)	56,055	[55,121, 56,817]	70,162	35,291	1.62
MMLU, M=20, R1	34,378	42,607 (12)	77,022	[76,374, 77,734]	98,197	47,230	1.81
MMLU, M=20, R3	34,378	40,381 (3)	72,808	[71,945, 73,870]	92,885	42,830	1.80
ARC-C, M=10, R1	23,238	29,872 (9)	48,339	[47,178, 49,170]	59,268	32,517	1.62
ARC-C, M=10, R3	25,177	24,079 (1)	40,688	[39,997, 41,274]	51,515	25,409	1.62
ARC-C, M=20, R1	23,238	32,765 (10)	56,450	[55,895, 57,053]	71,469	34,951	1.72
ARC-C, M=20, R3	24,648	30,144 (4)	52,820	[52,302, 53,720]	68,093	31,682	1.75
HellaSwag, M=10, R1	13,867	15,372 (5)	26,072	[25,827, 26,354]	31,895	17,460	1.70
HellaSwag, M=10, R3	13,194	13,847 (2)	22,695	[22,260, 23,109]	28,130	14,609	1.64
HellaSwag, M=20, R1	11,696	14,852 (8)	26,199	[26,042, 26,363]	33,443	16,549	1.76
HellaSwag, M=20, R3	13,272	15,928 (5)	28,240	[27,979, 28,602]	36,232	16,931	1.77
WinoGrande, M=10, R1	25,544	32,401 (8)	52,286	[51,599, 52,968]	64,507	35,494	1.61
WinoGrande, M=10, R3	23,433	27,187 (3)	43,707	[43,121, 44,283]	54,050	28,071	1.61
WinoGrande, M=20, R1	27,979	37,045 (8)	66,618	[65,920, 67,445]	84,478	41,046	1.80
WinoGrande, M=20, R3	29,764	37,439 (5)	65,989	[65,286, 66,624]	83,661	39,540	1.76

Figures 4–6 show the whole completion curve behind each n_{95}. The solid vertical line is Theorem 1 and the dashed one the larger of the two bounds. Every curve for Theorem 4 crosses 0.95 to the right of both lines, as it must. The uniform-budget certificate lies to the right of Theorem 4 everywhere, by the 18 to 24% reported in the main text.

Sensitivity to the power target. The main text fixes the success probability at 0.95. Reading the same curves at 0.80 and 0.90 gives Table 7. The ratio falls as the target falls, from 1.61–1.81 at 0.95 to 1.35–1.63 at 0.90 and 1.05–1.43 at 0.80, because the lower bounds are fixed at the level the theorems use while the certificate needs fewer items for a weaker target. The comparison at 0.95 is

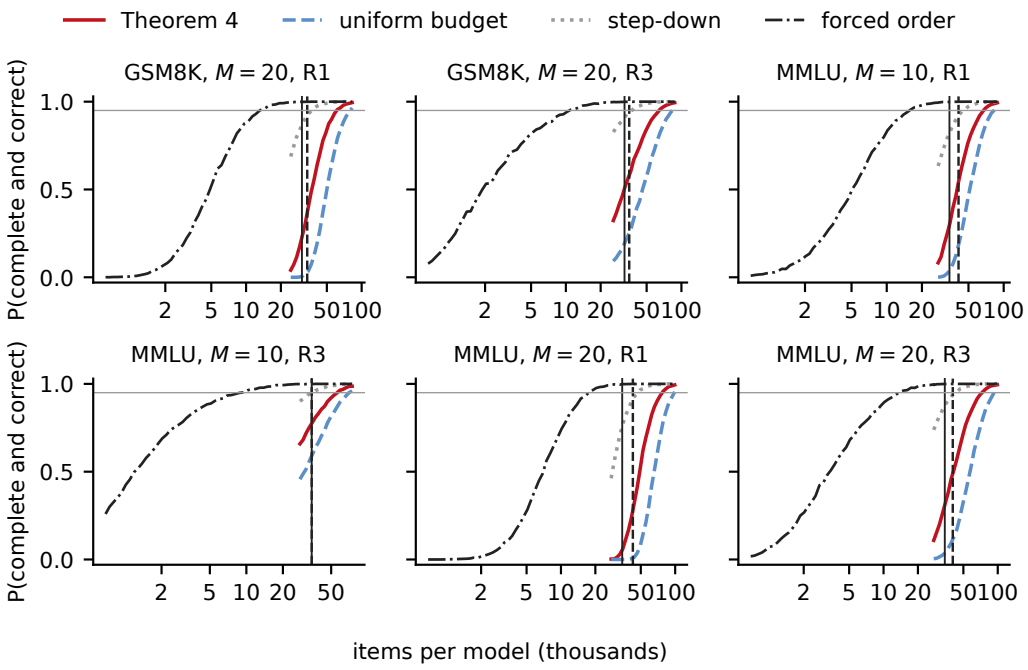

Figure 4: Probability of certifying the complete ranking correctly, GSM8K and MMLU, on a logarithmic axis. Vertical lines: Theorem 1 (solid) and the larger of Theorems 1 and 2 (dashed). Dash-dotted: probability that the order of the sample means has no inversion, which no certificate controls (Appendix L).

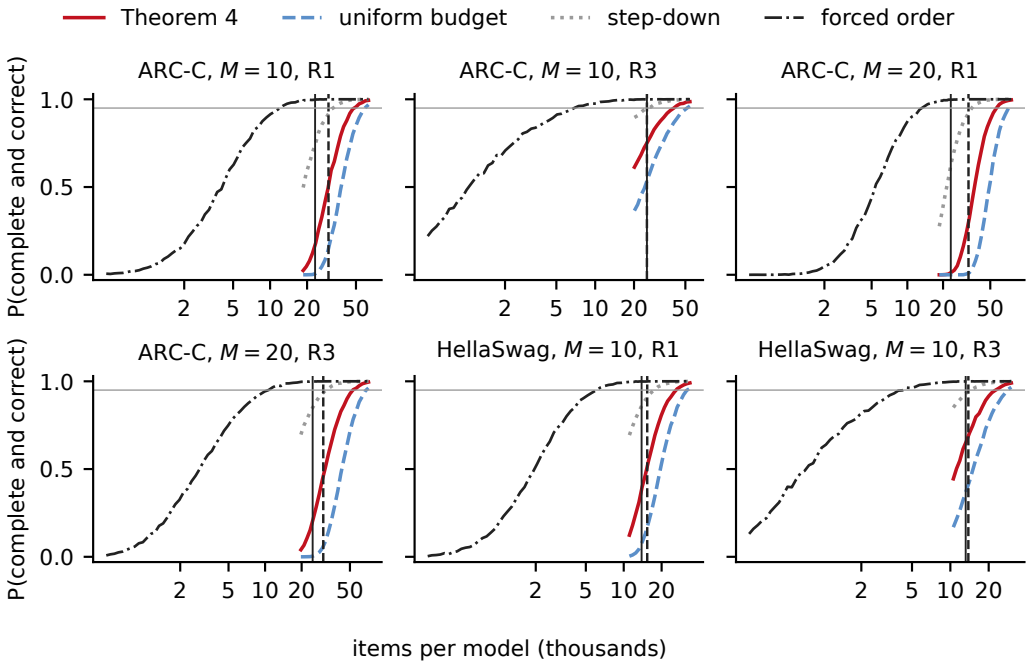

Figure 5: As Figure 4, ARC-Challenge and HellaSwag.

the one the theorems speak to, and the others are reported so that a reader can see how much of the ratio is the choice of target.

Table 7: Theorem 4 over the larger lower bound, when n is read at success probability 0.80, 0.90 and 0.95.

Configuration	power 0.80	power 0.90	power 0.95
GSM8K, M=20, R1	1.41	1.63	1.80
GSM8K, M=20, R3	1.32	1.57	1.81
MMLU, M=10, R1	1.25	1.46	1.64
MMLU, M=10, R3	1.05	1.35	1.62
MMLU, M=20, R1	1.43	1.63	1.81
MMLU, M=20, R3	1.35	1.59	1.80
ARC-C, M=10, R1	1.27	1.45	1.62
ARC-C, M=10, R3	1.09	1.36	1.62
ARC-C, M=20, R1	1.38	1.56	1.72
ARC-C, M=20, R3	1.35	1.57	1.75
HellaSwag, M=10, R1	1.28	1.50	1.70
HellaSwag, M=10, R3	1.16	1.39	1.64
HellaSwag, M=20, R1	1.42	1.60	1.76
HellaSwag, M=20, R3	1.38	1.58	1.77
WinoGrande, M=10, R1	1.25	1.44	1.61
WinoGrande, M=10, R3	1.18	1.41	1.61
WinoGrande, M=20, R1	1.42	1.59	1.80
WinoGrande, M=20, R3	1.35	1.57	1.76

Sensitivity to δ. Theorem 1 pairs a validity level δ with a success probability $1-\delta$, so changing δ moves both the bound and the target. Table 8 repeats the comparison at $\delta = 0.01$, reading Theorem 4 at success probability 0.99, and at $\delta = 0.10$, reading it at 0.90, on all eighteen configurations with $4{,}000$ draws per grid point. Each ratio here uses Theorem 1 alone, including the column at $\delta = 0.05$, which therefore exceeds the headline ratio wherever Theorem 2 is the stronger bound. The ratio is 1.36 to 1.73 at $\delta = 0.01$, 1.62 to 2.43 at 0.05 and 1.92 to 3.30 at 0.10. It falls as δ falls on every configuration, which is what a heuristic count of the multiplicity cost suggests: the certificate pays about $\log((M-1)/\delta)$ per comparison while the single-pair bound pays about $\log(1/\delta)$, and the difference between the two matters less the smaller δ is. A stricter guarantee therefore costs more items in absolute terms but brings the certificate closer to the bound.

Table 8: Theorem 4 against Theorem 1 at three levels. At each δ the bound is computed at that level and the certificate is read at success probability $1-\delta$.

	$\delta = 0.01$			$\delta = 0.05$	$\delta = 0.10$		
Configuration	Thm 1	$n_{0.99}$	ratio	ratio	Thm 1	$n_{0.90}$	ratio
GSM8K, M=20, R1	60,872	91,852	1.51	2.00	18,473	48,998	2.65
GSM8K, M=20, R3	64,369	98,649	1.53	1.99	19,535	49,276	2.52
MMLU, M=10, R1	68,767	104,106	1.51	1.96	20,869	53,588	2.57
MMLU, M=10, R3	69,114	97,567	1.41	1.62	20,975	40,259	1.92
MMLU, M=20, R1	68,767	110,832	1.61	2.24	20,869	62,893	3.01
MMLU, M=20, R3	68,767	109,838	1.60	2.12	20,869	57,732	2.77
ARC-C, M=10, R1	46,483	71,975	1.55	2.08	14,107	38,779	2.75
ARC-C, M=10, R3	50,361	68,372	1.36	1.62	15,284	29,980	1.96
ARC-C, M=20, R1	46,483	80,191	1.73	2.43	14,107	46,519	3.30
ARC-C, M=20, R3	49,303	77,908	1.58	2.14	14,963	42,450	2.84
HellaSwag, M=10, R1	27,738	39,070	1.41	1.88	8,418	20,595	2.45
HellaSwag, M=10, R3	26,391	36,267	1.37	1.72	8,010	17,227	2.15
HellaSwag, M=20, R1	23,394	38,290	1.64	2.24	7,100	21,971	3.09
HellaSwag, M=20, R3	26,548	42,168	1.59	2.13	8,057	22,518	2.79
WinoGrande, M=10, R1	51,094	79,184	1.55	2.05	15,506	41,513	2.68
WinoGrande, M=10, R3	46,872	67,409	1.44	1.87	14,225	33,872	2.38
WinoGrande, M=20, R1	55,965	96,539	1.72	2.38	16,984	54,424	3.20
WinoGrande, M=20, R3	59,537	95,409	1.60	2.22	18,068	53,188	2.94

Q CHOOSING THE MODELS ON ONE PART OF THE BANK

The R1 and R3 rules read bank accuracies and paired statistics to choose the models. The replay experiments are unaffected, since there the bank is the item law and the chosen models are fixed

before any resampling. Certifying the chosen models on the same bank is different: as a statement about the item population, a procedure that chooses its models on the items it certifies them on is not the fixed-set procedure of the theorems. Table 9 separates the two roles. A random 30% of each bank, fixed by a seed before anything is computed, is given to R1 and R3 unchanged, including the accuracy filter and the removal of duplicate score vectors, which also see only that part; the chosen models are frozen, and every certificate runs on the other 70%, which is then an i.i.d. sample from the item law independent of the choice.

No complete ranking closes on the held-out items. Every held-out part is smaller than every whole-bank n_{95} in Figure 2, which makes that unsurprising, but those values belong to different model sets and describe a probability, so they do not predict the outcome of one draw. The uniform-budget certificate still certifies 40% to 98% of the relations, and on HellaSwag the proved prefix reaches eight models. The valid directional Holm procedure of Appendix F, whose edges are always a subset of the uniform-budget certificate's, certifies as many relations in ten configurations and up to four fewer in the others. The chosen model sets overlap little with the whole-bank sets, since the rules are sensitive to small changes in accuracy, so the two columns describe different leaderboards.

Table 9: Models chosen on 30% of the bank, certified on the other 70%. *Shared*: models in common with the whole-bank configuration. Held-out columns: whether Theorem 4 closes, relations certified by the uniform-budget certificate at δ/K, its proved prefix, and relations certified by Holm over the $2K$ directional nulls. Whole-bank columns are descriptive.

				held-out items			whole bank	
Configuration	Items	Shared	Thm 4	δ/K	Prefix	Dir. Holm	δ/K	Thm 4
GSM8K, M=20, R1	923	1		121/190	0	119	136	
GSM8K, M=20, R3	923	8		148/190	0	146	161	
MMLU, M=10, R1	9,829	0		36/45	0	36	37	
MMLU, M=10, R3	9,829	1		42/45	3	42	44	
MMLU, M=20, R1	9,829	3		166/190	0	166	172	
MMLU, M=20, R3	9,829	3		182/190	3	181	186	
ARC-C, M=10, R1	820	0		20/45	0	19	19	
ARC-C, M=10, R3	820	0		32/45	0	32	33	
ARC-C, M=20, R1	820	0		111/190	0	110	128	
ARC-C, M=20, R3	820	4		139/190	0	139	151	
HellaSwag, M=10, R1	7,029	0		40/45	4	40	42	✓
HellaSwag, M=10, R3	7,029	3		44/45	8	44	44	✓
HellaSwag, M=20, R1	7,029	2		180/190	1	179	184	✓
HellaSwag, M=20, R3	7,029	5		183/190	5	183	185	✓
WinoGrande, M=10, R1	887	0		18/45	0	17	20	
WinoGrande, M=10, R3	887	1		27/45	1	27	32	
WinoGrande, M=20, R1	887	1		115/190	0	115	122	
WinoGrande, M=20, R3	887	2		129/190	0	125	138	

R SENSITIVITY TO HOW THE LEADERBOARDS ARE BUILT

The eighteen configurations fix two construction choices: adjacent models differ in accuracy by at least 0.01, and M is 10 or 20. To see whether the ratio depends on them, we rebuilt the R1 leaderboard on the two largest banks for every gap floor in $\{0.005, 0.01, 0.02\}$ and every $M \in \{5, 10, 15, 20\}$, and reran the whole pipeline: the exact Theorem 1 bound, the Theorem 2 bound found by the same search and certified on $200{,}000$ fresh draws (Bonferroni over all 48 possible tests), and n_{95} for Theorem 4 and the uniform-budget certificate with $2{,}000$ draws per grid point. HellaSwag has no twenty-model set with a floor of 0.02, so 23 leaderboards result, and every mixture bound was certified at the first attempt.

Figure 7 and Table 10 give the result. The bounds range from about $2{,}200$ to $190{,}000$ items, yet the ratio of Theorem 4 to the larger bound stays between 1.43 and 1.83. It rises with M, from 1.43–1.52 at $M = 5$ to 1.62–1.69 at $M = 10$ and 1.70–1.83 at $M = 15$ and 20, and at each M the three floors differ by at most 0.13. The saving of Theorem 4 over the uniform-budget certificate grows with M in the same way, from 11%–16% at $M = 5$ to 20%–23% at $M = 20$. The configuration with MMLU, $M = 20$ and floor 0.01 is the paper's MMLU R1 leaderboard, and its ratio here, 1.83 with fewer draws, agrees with the 1.81 of Table 5. With $2{,}000$ draws per point, 6 grid steps decrease by Monte Carlo noise, all of them where the curve is already between 0.965 and 0.998, so the running

maximum changes no crossing: on all 23 curves the raw first crossing of 0.95 equals the reported n_{95} for both rules.

Table 10: The sensitivity study in full. Span is the accuracy range of the chosen models; Theorem 2 is shown where it exceeds Theorem 1; the last two columns are n_{95} over the larger bound.

Bank	M	gap floor	span	Thm 1	Thm 2	n_{95} Thm 4	ratio	uniform budget
MMLU	5	0.005	0.020	58,370	62,827	92,584	1.47	1.75
MMLU	5	0.01	0.040	22,881	26,877	39,160	1.46	1.66
MMLU	5	0.02	0.080	6,781	7,794	11,147	1.43	1.68
MMLU	10	0.005	0.048	101,306	124,621	201,559	1.62	2.00
MMLU	10	0.01	0.093	34,378	40,948	68,056	1.66	2.02
MMLU	10	0.02	0.184	8,151	9,797	16,142	1.65	2.00
MMLU	15	0.005	0.078	123,284	162,899	277,961	1.71	2.12
MMLU	15	0.01	0.146	34,378	43,805	74,346	1.70	2.14
MMLU	15	0.02	0.293	8,157	9,120	16,131	1.77	2.25
MMLU	20	0.005	0.106	137,063	189,679	322,429	1.70	2.18
MMLU	20	0.01	0.199	34,378	42,078	77,165	1.83	2.34
MMLU	20	0.02	0.411	8,151	10,035	17,823	1.78	2.27
HellaSwag	5	0.005	0.020	32,673	37,057	53,090	1.43	1.62
HellaSwag	5	0.01	0.040	8,684	9,630	14,371	1.49	1.68
HellaSwag	5	0.02	0.080	2,249	2,444	3,711	1.52	1.75
HellaSwag	10	0.005	0.046	45,064	51,110	85,060	1.66	2.03
HellaSwag	10	0.01	0.092	13,867	15,660	26,499	1.69	2.02
HellaSwag	10	0.02	0.187	3,313	4,240	6,911	1.63	1.99
HellaSwag	15	0.005	0.073	41,766	54,739	93,848	1.71	2.14
HellaSwag	15	0.01	0.146	13,386	15,061	27,278	1.81	2.27
HellaSwag	15	0.02	0.322	5,070	5,702	10,123	1.78	2.25
HellaSwag	20	0.005	0.100	41,766	57,993	98,865	1.70	2.19
HellaSwag	20	0.01	0.219	11,696	14,838	26,368	1.78	2.31
HellaSwag	20	0.02	no admissible set					

S WHAT EACH BANK PROVES

Figures 8–10 show, for each configuration, which of the K relations the uniform budget certifies on the whole bank. Rows and columns are models in bank order, and a blue cell at (j, k) means j is proved above k. The unproved relations sit next to the diagonal, between models whose accuracies are close, and a proved prefix of length m is a top-left block whose rows are blue across the whole matrix.

S.1 ADJACENT COMPARISONS ON THE WHOLE BANK

For every configuration the table lists each adjacent pair in bank order: the accuracy gap, the share s of items on which the two models disagree, the share a of those that favour the higher model, the exact one-sided p-value on the whole bank, and whether that p-value clears the level of Theorem 4 and the level of the uniform-budget certificate. A configuration closes under Theorem 4 exactly when every pair, adjacent or not, clears $\delta/(M-1)$.

GSM8K, M=20**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0121	0.288	0.521	2.2e-1		
2–3	0.0106	0.317	0.517	2.6e-1		
3–4	0.0144	0.271	0.527	1.7e-1		
4–5	0.0121	0.309	0.520	2.3e-1		
5–6	0.0144	0.299	0.524	1.8e-1		
6–7	0.0106	0.255	0.521	2.4e-1		
7–8	0.0121	0.264	0.523	2.1e-1		
8–9	0.0114	0.248	0.523	2.2e-1		
9–10	0.0114	0.231	0.525	2.1e-1		
10–11	0.0114	0.218	0.526	2.0e-1		
11–12	0.0159	0.213	0.537	1.2e-1		
12–13	0.0106	0.132	0.540	1.6e-1		
13–14	0.0144	0.114	0.563	7.1e-2		
14–15	0.0106	0.109	0.549	1.4e-1		
15–16	0.0106	0.111	0.548	1.4e-1		
16–17	0.0106	0.105	0.551	1.3e-1		
17–18	0.0114	0.108	0.552	1.2e-1		
18–19	0.0106	0.092	0.557	1.2e-1		
19–20	0.0106	0.076	0.570	9.7e-2		

GSM8K, $M{=}20$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0538	0.327	0.582	3.6e-4	✓	
2–3	0.0387	0.281	0.569	4.7e-3		
3–4	0.0334	0.350	0.548	2.3e-2		
4–5	0.0114	0.348	0.516	2.6e-1		
5–6	0.0281	0.290	0.548	3.3e-2		
6–7	0.0288	0.211	0.568	1.3e-2		
7–8	0.0281	0.251	0.556	2.4e-2		
8–9	0.0106	0.335	0.516	2.7e-1		
9–10	0.0364	0.144	0.626	3.1e-4	✓	
10–11	0.0371	0.280	0.566	6.2e-3		
11–12	0.0243	0.118	0.603	6.4e-3		
12–13	0.0174	0.107	0.582	3.2e-2		
13–14	0.0167	0.206	0.540	1.0e-1		
14–15	0.0129	0.058	0.610	3.4e-2		
15–16	0.0144	0.093	0.577	5.2e-2		
16–17	0.0136	0.103	0.566	7.2e-2		
17–18	0.0144	0.110	0.566	6.7e-2		
18–19	0.0114	0.034	0.667	1.8e-2		
19–20	0.0182	0.076	0.620	1.0e-2		

MMLU, $M{=}10$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0102	0.229	0.522	6.1e-3		
2–3	0.0103	0.124	0.542	2.8e-4	✓	✓
3–4	0.0102	0.175	0.529	2.1e-3	✓	
4–5	0.0102	0.297	0.517	1.4e-2		
5–6	0.0103	0.318	0.516	1.6e-2		
6–7	0.0100	0.320	0.516	1.8e-2		
7–8	0.0109	0.253	0.521	5.4e-3	✓	
8–9	0.0108	0.272	0.520	7.6e-3		
9–10	0.0100	0.192	0.526	3.5e-3	✓	

MMLU, $M{=}10$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0224	0.110	0.602	6.0e-16	✓	✓
2–3	0.0242	0.131	0.593	1.0e-15	✓	✓
3–4	0.0298	0.127	0.618	1.3e-23	✓	✓
4–5	0.0104	0.345	0.515	1.9e-2		
5–6	0.0234	0.357	0.533	1.9e-6	✓	✓
6–7	0.0212	0.218	0.549	3.8e-8	✓	✓
7–8	0.0149	0.066	0.613	2.9e-12	✓	✓
8–9	0.0182	0.219	0.542	2.1e-6	✓	✓
9–10	0.0212	0.132	0.580	2.7e-12	✓	✓

MMLU, $M{=}20$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0102	0.230	0.522	6.2e-3		
2–3	0.0106	0.208	0.525	3.1e-3		
3–4	0.0102	0.084	0.560	1.8e-5	✓	✓
4–5	0.0100	0.145	0.535	9.4e-4	✓	
5–6	0.0103	0.168	0.530	1.6e-3	✓	
6–7	0.0113	0.265	0.521	5.1e-3		
7–8	0.0118	0.260	0.523	3.2e-3		
8–9	0.0112	0.266	0.521	5.3e-3		
9–10	0.0102	0.229	0.522	6.1e-3		
10–11	0.0103	0.124	0.542	2.8e-4	✓	
11–12	0.0102	0.175	0.529	2.1e-3	✓	
12–13	0.0102	0.297	0.517	1.4e-2		
13–14	0.0103	0.318	0.516	1.6e-2		
14–15	0.0100	0.320	0.516	1.8e-2		
15–16	0.0109	0.253	0.521	5.4e-3		
16–17	0.0108	0.272	0.520	7.6e-3		
17–18	0.0100	0.192	0.526	3.5e-3		
18–19	0.0108	0.198	0.527	2.2e-3	✓	
19–20	0.0101	0.284	0.518	1.3e-2		

MMLU, $M{=}20$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0302	0.089	0.670	4.5e-34	✓	✓
2–3	0.0427	0.199	0.607	2.6e-30	✓	✓
3–4	0.0207	0.099	0.604	3.3e-15	✓	✓
4–5	0.0221	0.161	0.569	3.4e-11	✓	✓
5–6	0.0183	0.229	0.540	3.1e-6	✓	✓
6–7	0.0113	0.066	0.585	1.2e-7	✓	✓
7–8	0.0130	0.093	0.570	2.3e-7	✓	✓
8–9	0.0105	0.191	0.527	2.4e-3	✓	
9–10	0.0157	0.106	0.574	6.0e-9	✓	✓
10–11	0.0100	0.285	0.518	1.3e-2		
11–12	0.0155	0.235	0.533	8.4e-5	✓	✓
12–13	0.0125	0.071	0.589	1.3e-8	✓	✓
13–14	0.0193	0.191	0.550	9.3e-8	✓	✓
14–15	0.0106	0.332	0.516	1.5e-2		
15–16	0.0100	0.320	0.516	1.8e-2		
16–17	0.0188	0.127	0.574	2.4e-10	✓	✓
17–18	0.0375	0.225	0.584	3.1e-21	✓	✓
18–19	0.0365	0.298	0.561	1.1e-15	✓	✓
19–20	0.0406	0.274	0.574	2.1e-20	✓	✓

ARC-C, $M{=}10$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0102	0.142	0.536	2.0e-1		
2–3	0.0102	0.140	0.537	2.0e-1		
3–4	0.0102	0.183	0.528	2.3e-1		
4–5	0.0102	0.193	0.527	2.3e-1		
5–6	0.0102	0.188	0.527	2.3e-1		
6–7	0.0102	0.225	0.523	2.5e-1		
7–8	0.0102	0.205	0.525	2.4e-1		
8–9	0.0102	0.215	0.524	2.4e-1		
9–10	0.0102	0.201	0.525	2.4e-1		

ARC-C, $M{=}10$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0324	0.087	0.686	1.1e-4	✓	✓
2–3	0.0119	0.031	0.694	1.4e-2		
3–4	0.0137	0.038	0.682	1.1e-2		
4–5	0.0162	0.054	0.651	1.1e-2		
5–6	0.0119	0.032	0.684	1.7e-2		
6–7	0.0179	0.112	0.580	4.0e-2		
7–8	0.0119	0.036	0.667	2.2e-2		
8–9	0.0111	0.136	0.541	1.7e-1		
9–10	0.0102	0.244	0.521	2.6e-1		

ARC-C, $M{=}20$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0111	0.151	0.537	1.8e-1		
2–3	0.0119	0.137	0.544	1.5e-1		
3–4	0.0102	0.125	0.541	1.8e-1		
4–5	0.0102	0.159	0.532	2.1e-1		
5–6	0.0111	0.137	0.540	1.7e-1		
6–7	0.0102	0.147	0.535	2.0e-1		
7–8	0.0102	0.159	0.532	2.1e-1		
8–9	0.0102	0.142	0.536	2.0e-1		
9–10	0.0102	0.140	0.537	2.0e-1		
10–11	0.0102	0.183	0.528	2.3e-1		
11–12	0.0102	0.193	0.527	2.3e-1		
12–13	0.0102	0.188	0.527	2.3e-1		
13–14	0.0102	0.225	0.523	2.5e-1		
14–15	0.0102	0.205	0.525	2.4e-1		
15–16	0.0102	0.215	0.524	2.4e-1		
16–17	0.0102	0.201	0.525	2.4e-1		
17–18	0.0111	0.219	0.525	2.3e-1		
18–19	0.0102	0.213	0.524	2.4e-1		
19–20	0.0102	0.203	0.525	2.4e-1		

ARC-C, $M{=}20$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0324	0.087	0.686	1.1e-4	✓	✓
2–3	0.0119	0.031	0.694	1.4e-2		
3–4	0.0137	0.038	0.682	1.1e-2		
4–5	0.0111	0.059	0.594	7.4e-2		
5–6	0.0119	0.055	0.609	5.2e-2		
6–7	0.0137	0.084	0.582	6.5e-2		
7–8	0.0128	0.054	0.619	3.8e-2		
8–9	0.0111	0.071	0.578	9.4e-2		
9–10	0.0111	0.137	0.540	1.7e-1		
10–11	0.0102	0.225	0.523	2.5e-1		
11–12	0.0102	0.196	0.526	2.3e-1		
12–13	0.0102	0.239	0.521	2.6e-1		
13–14	0.0102	0.201	0.525	2.4e-1		
14–15	0.0179	0.103	0.587	3.4e-2		
15–16	0.0145	0.083	0.588	5.2e-2		
16–17	0.0401	0.158	0.627	3.4e-4	✓	
17–18	0.0307	0.176	0.587	7.3e-3		
18–19	0.0427	0.203	0.605	7.2e-4	✓	
19–20	0.0606	0.320	0.595	1.4e-4	✓	✓

HellaSwag, $M{=}10$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0101	0.053	0.594	7.5e-6	✓	✓
2–3	0.0101	0.073	0.569	1.1e-4	✓	✓
3–4	0.0101	0.048	0.606	2.2e-6	✓	✓
4–5	0.0101	0.077	0.565	1.6e-4	✓	✓
5–6	0.0110	0.101	0.554	3.0e-4	✓	✓
6–7	0.0101	0.084	0.560	2.9e-4	✓	✓
7–8	0.0103	0.113	0.545	1.2e-3	✓	
8–9	0.0103	0.120	0.543	1.6e-3	✓	
9–10	0.0102	0.132	0.538	2.8e-3	✓	

HellaSwag, $M{=}10$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0263	0.061	0.717	1.1e-27	✓	✓
2–3	0.0131	0.027	0.744	2.0e-16	✓	✓
3–4	0.0123	0.032	0.690	2.7e-12	✓	✓
4–5	0.0113	0.021	0.768	1.3e-15	✓	✓
5–6	0.0119	0.025	0.735	2.0e-14	✓	✓
6–7	0.0103	0.100	0.551	6.6e-4	✓	✓
7–8	0.0101	0.124	0.541	2.3e-3	✓	
8–9	0.0419	0.112	0.686	5.5e-37	✓	✓
9–10	0.0397	0.093	0.713	3.8e-40	✓	✓

HellaSwag, $M{=}20$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0181	0.064	0.641	3.5e-13	✓	✓
2–3	0.0122	0.069	0.589	1.7e-6	✓	✓
3–4	0.0101	0.064	0.579	3.7e-5	✓	✓
4–5	0.0102	0.077	0.566	1.3e-4	✓	✓
5–6	0.0102	0.078	0.565	1.5e-4	✓	✓
6–7	0.0108	0.087	0.562	1.5e-4	✓	✓
7–8	0.0103	0.093	0.555	4.2e-4	✓	
8–9	0.0101	0.070	0.572	8.1e-5	✓	✓
9–10	0.0101	0.094	0.554	5.5e-4	✓	
10–11	0.0108	0.091	0.559	1.9e-4	✓	✓
11–12	0.0103	0.075	0.569	9.7e-5	✓	✓
12–13	0.0105	0.088	0.560	2.3e-4	✓	✓
13–14	0.0107	0.101	0.553	4.3e-4	✓	
14–15	0.0109	0.092	0.559	1.9e-4	✓	✓
15–16	0.0107	0.123	0.543	1.3e-3	✓	
16–17	0.0114	0.122	0.546	6.3e-4	✓	
17–18	0.0130	0.121	0.554	9.2e-5	✓	✓
18–19	0.0152	0.129	0.559	1.2e-5	✓	✓
19–20	0.0141	0.184	0.538	5.1e-4	✓	

HellaSwag, $M{=}20$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0252	0.063	0.701	1.1e-24	✓	✓
2–3	0.0127	0.033	0.694	7.2e-13	✓	✓
3–4	0.0111	0.024	0.727	4.3e-13	✓	✓
4–5	0.0103	0.029	0.676	8.7e-10	✓	✓
5–6	0.0105	0.042	0.625	1.8e-7	✓	✓
6–7	0.0125	0.062	0.601	2.5e-7	✓	✓
7–8	0.0106	0.101	0.552	4.8e-4	✓	
8–9	0.0105	0.134	0.539	2.3e-3	✓	
9–10	0.0101	0.109	0.546	1.2e-3	✓	
10–11	0.0124	0.071	0.588	1.5e-6	✓	✓
11–12	0.0113	0.083	0.568	5.3e-5	✓	✓
12–13	0.0178	0.071	0.625	1.2e-11	✓	✓
13–14	0.0107	0.123	0.543	1.3e-3	✓	
14–15	0.0114	0.122	0.546	6.3e-4	✓	
15–16	0.0283	0.126	0.612	7.7e-16	✓	✓
16–17	0.0409	0.131	0.656	2.3e-30	✓	✓
17–18	0.0296	0.153	0.597	1.7e-14	✓	✓
18–19	0.0367	0.154	0.619	3.0e-21	✓	✓
19–20	0.0641	0.172	0.687	1.3e-55	✓	✓

WinoGrande, $M{=}10$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0103	0.084	0.561	1.2e-1		
2–3	0.0103	0.181	0.528	2.1e-1		
3–4	0.0103	0.159	0.532	2.0e-1		
4–5	0.0103	0.174	0.529	2.1e-1		
5–6	0.0103	0.206	0.525	2.3e-1		
6–7	0.0103	0.230	0.522	2.4e-1		
7–8	0.0103	0.233	0.522	2.4e-1		
8–9	0.0103	0.249	0.521	2.5e-1		
9–10	0.0103	0.220	0.523	2.4e-1		

WinoGrande, $M{=}10$**, R3**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0316	0.071	0.722	1.5e-5	✓	✓
2–3	0.0189	0.060	0.658	4.0e-3	✓	
3–4	0.0103	0.036	0.644	3.6e-2		
4–5	0.0150	0.073	0.602	3.1e-2		
5–6	0.0110	0.200	0.528	2.1e-1		
6–7	0.0103	0.228	0.522	2.4e-1		
7–8	0.0103	0.217	0.524	2.3e-1		
8–9	0.0300	0.115	0.630	1.0e-3	✓	✓
9–10	0.0466	0.271	0.586	8.5e-4	✓	✓

WinoGrande, $M{=}20$**, R1**

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0103	0.171	0.530	2.1e-1		
2–3	0.0103	0.181	0.528	2.1e-1		
3–4	0.0110	0.164	0.534	1.8e-1		
4–5	0.0103	0.072	0.571	1.0e-1		
5–6	0.0103	0.096	0.554	1.4e-1		
6–7	0.0103	0.084	0.561	1.2e-1		
7–8	0.0103	0.181	0.528	2.1e-1		
8–9	0.0103	0.159	0.532	2.0e-1		
9–10	0.0103	0.174	0.529	2.1e-1		
10–11	0.0103	0.206	0.525	2.3e-1		
11–12	0.0103	0.230	0.522	2.4e-1		
12–13	0.0118	0.252	0.524	2.2e-1		
13–14	0.0103	0.272	0.519	2.6e-1		
14–15	0.0103	0.263	0.520	2.6e-1		
15–16	0.0118	0.264	0.522	2.2e-1		
16–17	0.0110	0.226	0.524	2.2e-1		
17–18	0.0103	0.268	0.519	2.6e-1		
18–19	0.0142	0.281	0.525	1.8e-1		
19–20	0.0118	0.361	0.516	2.6e-1		

WinoGrande, M=20, R3

Pair	Gap	s	a	p on bank	at $\delta/(M-1)$	at δ/K
1–2	0.0142	0.044	0.661	1.1e-2		
2–3	0.0142	0.153	0.546	1.1e-1		
3–4	0.0126	0.068	0.593	5.3e-2		
4–5	0.0126	0.158	0.540	1.4e-1		
5–6	0.0110	0.032	0.675	1.9e-2		
6–7	0.0103	0.036	0.644	3.6e-2		
7–8	0.0150	0.073	0.602	3.1e-2		
8–9	0.0103	0.165	0.531	2.0e-1		
9–10	0.0118	0.125	0.547	1.3e-1		
10–11	0.0103	0.230	0.522	2.4e-1		
11–12	0.0103	0.233	0.522	2.4e-1		
12–13	0.0118	0.242	0.524	2.1e-1		
13–14	0.0103	0.272	0.519	2.6e-1		
14–15	0.0103	0.263	0.520	2.6e-1		
15–16	0.0103	0.290	0.518	2.7e-1		
16–17	0.0229	0.266	0.543	6.4e-2		
17–18	0.0205	0.278	0.537	9.1e-2		
18–19	0.0253	0.286	0.544	5.2e-2		
19–20	0.0339	0.305	0.556	1.6e-2		

T BASELINES IN FULL

T.1 THE EXACT TEST AGAINST EMPIRICAL BERNSTEIN

Figures 11–13 and Table 11 compare the two per-pair tests under the uniform budget on all eighteen configurations, counting complete and correct rankings. Both tests use the level δ/K per pair, and the Bernstein radius is (15). Empirical Bernstein needs 1.36 to 1.57 times as many items as the exact test to reach 0.95. The gain is largest on HellaSwag, 1.49 to 1.57, which is also where the hardest comparison has the smallest discordant share, $s = 0.12$ to 0.13 against at least 0.22 on every other configuration. That fits the form of the two tests, since the exact test uses only the discordant items, while (15) carries a term of order $\log(2/\delta_e)/n$ that does not shrink with the variance. At five fixed budgets the exact test certifies at least as often in all 90 cells and strictly more often in 56. On the six configurations we started from, eight of the thirty cells tie because neither test certifies anything at that budget.

T.2 THE VALIDITY SWEEP

The synthetic sweep draws n items for M models whose accuracies are set exactly. With probability ρ an item is easy and every model scores one; otherwise each model scores one independently with the probability that makes its accuracy come out right. That makes scores correlated across models within an item and keeps items independent. In the mixed shape the models are 0.04 apart except the top two, which are 0.0005, 0.002 or 0.01 apart. In the uniform shape every adjacent gap takes that value. Figure 14 and Table 12 give, for each rule, the fraction of $2{,}000$ runs in which a certified relation pointed the wrong way.

Table 11: n_{95} under both tests and their ratio, and the completion rate of each test at five fixed budgets, written exact / Bernstein.

Configuration	n_{95} exact	n_{95} Bernstein	ratio	complete, exact / Bernstein, at n in thousands 10	20	40	80	140
GSM8K, M=20, R1	79,346	109,491	1.38	0.00/0.00	0.00/0.00	0.16/0.00	0.96/0.63	1.00/0.99
GSM8K, M=20, R3	83,843	116,198	1.39	0.00/0.00	0.03/0.00	0.38/0.04	0.93/0.62	1.00/0.99
MMLU, M=10, R1	82,444	117,049	1.42	0.00/0.00	0.00/0.00	0.15/0.00	0.93/0.52	1.00/0.99
MMLU, M=10, R3	70,722	103,243	1.46	0.02/0.00	0.28/0.01	0.69/0.26	0.97/0.82	1.00/1.00
MMLU, M=20, R1	98,274	134,128	1.36	0.00/0.00	0.00/0.00	0.00/0.00	0.80/0.15	1.00/0.97
MMLU, M=20, R3	93,822	130,017	1.39	0.00/0.00	0.00/0.00	0.10/0.00	0.87/0.37	1.00/0.97
ARC-C, M=10, R1	59,402	83,886	1.41	0.00/0.00	0.00/0.00	0.58/0.00	1.00/0.93	1.00/1.00
ARC-C, M=10, R3	52,592	78,086	1.48	0.03/0.00	0.35/0.01	0.86/0.43	1.00/0.96	1.00/1.00
ARC-C, M=20, R1	71,971	98,282	1.37	0.00/0.00	0.00/0.00	0.13/0.00	0.98/0.72	1.00/1.00
ARC-C, M=20, R3	68,650	94,281	1.37	0.00/0.00	0.00/0.00	0.40/0.00	0.99/0.80	1.00/1.00
HellaSwag, M=10, R1	31,818	49,172	1.55	0.00/0.00	0.54/0.00	0.99/0.80	1.00/1.00	1.00/1.00
HellaSwag, M=10, R3	28,286	44,319	1.57	0.14/0.00	0.76/0.06	1.00/0.89	1.00/1.00	1.00/1.00
HellaSwag, M=20, R1	33,258	49,544	1.49	0.00/0.00	0.25/0.00	0.99/0.66	1.00/1.00	1.00/1.00
HellaSwag, M=20, R3	35,759	53,771	1.50	0.00/0.00	0.28/0.00	0.97/0.57	1.00/1.00	1.00/1.00
WinoGrande, M=10, R1	63,405	91,114	1.44	0.00/0.00	0.00/0.00	0.49/0.00	0.99/0.85	1.00/1.00
WinoGrande, M=10, R3	54,702	79,080	1.45	0.00/0.00	0.07/0.00	0.78/0.11	1.00/0.96	1.00/1.00
WinoGrande, M=20, R1	84,237	115,046	1.37	0.00/0.00	0.00/0.00	0.04/0.00	0.92/0.44	1.00/0.99
WinoGrande, M=20, R3	83,872	114,269	1.36	0.00/0.00	0.00/0.00	0.09/0.00	0.93/0.48	1.00/0.99

T.3 Weighted levels and the chain test

Table 13 gives the thirty-two cells behind Appendix G: for two budgets and four sources of a prior order, the completion rate of the chain test on that order, of Holm, of the equal split δ/K and of the floored weighted scheme, and the paired verdict of weighted against equal. The order sources are the true order, the true order with one or with three adjacent transpositions, and an order estimated from a pilot of 5% of the items. The pilot is charged to the two procedures that use an order, the chain test and the weighted scheme, and Holm and the equal split use every item.

U The models in each configuration

For reproducibility, the models of every configuration in bank order, with their bank accuracy. Names are those of the release, without its common prefix.

GSM8K, M=20, R1

Rank	Model	Accuracy
1	martyn/llama-megamerge-dare-13b	0.2851
2	Locutusque/Orca-2-13B-no_robots	0.2729
3	Zangs3011/mistral_7b_DolphinCoder	0.2623
4	WebraftAI/synapsellm-7b-mistral-v0.4-preview3	0.2479
5	Sao10K/Zephyrus-L1-33B	0.2358
6	Locutusque/Rhino-Mistral-7B	0.2214
7	hfl/chinese-alpaca-2-13b-16k	0.2108
8	speechlessai/speechless-mistral-7b-dare-0.85	0.1986
9	uukuguy/speechless-coder-ds-6.7b	0.1873
10	migtissera/SynthIA-7B-v1.3	0.1759
11	Unbabel/TowerInstruct-7B-v0.1	0.1645
12	huggingface/llama-30b	0.1486
13	stabilityai/StableBeluga-13B	0.1380
14	Danielbrdz/Barcenas-13b	0.1236
15	elinas/chronos-13b-v2	0.1130
16	PulsarAI/2x-LoRA-Assemble-Nova-13B	0.1024
17	The-Face-Of-Goonery/Huginn-13b-v1.2	0.0917
18	Secbone/llama-2-13B-instructed	0.0804
19	jondurbin/airoboros-13b	0.0697
20	LTC-AI-Labs/L2-7b-Synthia-WVG-Test	0.0591

GSM8K, M=20, R3

Table 12: The forty-eight cells of the validity sweep. "Thm 4 (runs)" is the number of runs, out of 2,000, in which Theorem 4 certified a complete ranking.

Shape	M	n	gap	ρ	Thm 4 (runs)	Thm 4	uniform budget	step-down
mixed	4	2,000	0.0005	0.0	2	0.0000	0.0070	0.0185
mixed	4	2,000	0.0005	0.5	49	0.0120	0.0055	0.0445
mixed	4	2,000	0.0020	0.0	2	0.0000	0.0065	0.0190
mixed	4	2,000	0.0020	0.5	77	0.0065	0.0045	0.0265
mixed	4	2,000	0.0100	0.0	7	0.0005	0.0005	0.0040
mixed	4	2,000	0.0100	0.5	354	0.0005	0.0000	0.0010
mixed	4	20,000	0.0005	0.0	63	0.0150	0.0090	0.0410
mixed	4	20,000	0.0005	0.5	65	0.0070	0.0025	0.0315
mixed	4	20,000	0.0020	0.0	114	0.0060	0.0025	0.0160
mixed	4	20,000	0.0020	0.5	176	0.0035	0.0025	0.0075
mixed	4	20,000	0.0100	0.0	930	0.0000	0.0000	0.0005
mixed	4	20,000	0.0100	0.5	1885	0.0000	0.0000	0.0000
mixed	10	2,000	0.0005	0.0	0	0.0000	0.0005	0.0090
mixed	10	2,000	0.0005	0.5	6	0.0005	0.0000	0.0345
mixed	10	2,000	0.0020	0.0	0	0.0000	0.0005	0.0070
mixed	10	2,000	0.0020	0.5	3	0.0005	0.0015	0.0285
mixed	10	2,000	0.0100	0.0	0	0.0000	0.0000	0.0005
mixed	10	2,000	0.0100	0.5	25	0.0000	0.0000	0.0015
mixed	10	20,000	0.0005	0.0	21	0.0035	0.0000	0.0395
mixed	10	20,000	0.0005	0.5	28	0.0045	0.0010	0.0355
mixed	10	20,000	0.0020	0.0	45	0.0015	0.0000	0.0150
mixed	10	20,000	0.0020	0.5	54	0.0000	0.0000	0.0140
mixed	10	20,000	0.0100	0.0	1029	0.0000	0.0000	0.0000
mixed	10	20,000	0.0100	0.5	1333	0.0000	0.0000	0.0000
uniform	4	2,000	0.0005	0.0	0	0.0000	0.0335	0.0340
uniform	4	2,000	0.0005	0.5	0	0.0000	0.0005	0.0005
uniform	4	2,000	0.0020	0.0	0	0.0000	0.0220	0.0220
uniform	4	2,000	0.0020	0.5	0	0.0000	0.0005	0.0005
uniform	4	2,000	0.0100	0.0	0	0.0000	0.0055	0.0065
uniform	4	2,000	0.0100	0.5	391	0.0000	0.0000	0.0000
uniform	4	20,000	0.0005	0.0	0	0.0000	0.0310	0.0315
uniform	4	20,000	0.0005	0.5	27	0.0000	0.0000	0.0000
uniform	4	20,000	0.0020	0.0	0	0.0000	0.0070	0.0080
uniform	4	20,000	0.0020	0.5	1379	0.0000	0.0000	0.0000
uniform	4	20,000	0.0100	0.0	42	0.0000	0.0000	0.0000
uniform	4	20,000	0.0100	0.5	2000	0.0000	0.0000	0.0000
uniform	10	2,000	0.0005	0.0	0	0.0000	0.0245	0.0245
uniform	10	2,000	0.0005	0.5	0	0.0000	0.0015	0.0015
uniform	10	2,000	0.0020	0.0	0	0.0000	0.0100	0.0110
uniform	10	2,000	0.0020	0.5	0	0.0000	0.0005	0.0010
uniform	10	2,000	0.0100	0.0	0	0.0000	0.0005	0.0025
uniform	10	2,000	0.0100	0.5	0	0.0000	0.0000	0.0000
uniform	10	20,000	0.0005	0.0	0	0.0000	0.0180	0.0180
uniform	10	20,000	0.0005	0.5	0	0.0000	0.0000	0.0005
uniform	10	20,000	0.0020	0.0	0	0.0000	0.0015	0.0015
uniform	10	20,000	0.0020	0.5	0	0.0000	0.0000	0.0000
uniform	10	20,000	0.0100	0.0	0	0.0000	0.0000	0.0000
uniform	10	20,000	0.0100	0.5	1481	0.0000	0.0000	0.0000

Rank	Model	Accuracy
1	perlthoughts/Chupacabra-7B-v2.03-128k	0.5049
2	sr5434/CodegebraGPT-10b	0.4511
3	xxyyy123/Mistral7B_adaptor_v1	0.4124
4	RatanRohith/MistralBeagle-RS-7B-V0.1	0.3791
5	OpenBuddy/openbuddy-falcon-40b-v16.1-4k	0.3677
6	crumb/apricot-wildflower-20	0.3397
7	abdulrahman-nuzha/finetuned-Mistral-5000-v1.0	0.3108
8	umd-zhou-lab/claude2-alpaca-13B	0.2828
9	OpenAssistant/llama2-70b-oasst-sft-v10	0.2722
10	Sao10K/Zephyrus-L1-33B	0.2358
11	speechlessai/speechless-mistral-7b-dare-0.85	0.1986
12	PeanutJar/Mistral-v0.1-PeanutButter-v0.0.2-7B	0.1744
13	PocketDoc/Dans-AdventurousWinds-7b	0.1569
14	Undi95/ReMM-v2.2-L2-13B	0.1403
15	Undi95/ReMM-v2.1-L2-13B	0.1274
16	elinas/chronos-13b-v2	0.1130
17	Gryphe/MythoMix-L2-13b	0.0993
18	davzoku/cria-llama2-7b-v1.3	0.0849
19	NewstaR/Koss-7B-chat	0.0735
20	DevaMalla/llama_7b_lora	0.0553

Table 13: Completion rates of the chain test, Holm, the equal split and the weighted levels, 600 paired replicates per cell. Verdict: weighted against equal by a paired two-standard-error rule, with – for inconclusive.

Configuration	n	order	chain test	Holm	equal	weighted	verdict
GSM8K, M=20, R1	40,000	exact	0.977	0.977	0.178	0.582	better
GSM8K, M=20, R1	40,000	swap1	0.000	0.973	0.148	0.448	better
GSM8K, M=20, R1	40,000	swap3	0.000	0.970	0.157	0.268	better
GSM8K, M=20, R1	40,000	pilot	0.108	0.972	0.147	0.252	better
GSM8K, M=20, R1	70,000	exact	1.000	1.000	0.905	0.977	better
GSM8K, M=20, R1	70,000	swap1	0.000	0.998	0.878	0.945	better
GSM8K, M=20, R1	70,000	swap3	0.000	1.000	0.885	0.897	–
GSM8K, M=20, R1	70,000	pilot	0.332	1.000	0.873	0.908	better
GSM8K, M=20, R3	40,000	exact	0.957	0.957	0.342	0.645	better
GSM8K, M=20, R3	40,000	swap1	0.000	0.953	0.372	0.587	better
GSM8K, M=20, R3	40,000	swap3	0.000	0.960	0.350	0.482	better
GSM8K, M=20, R3	40,000	pilot	0.538	0.967	0.327	0.603	better
GSM8K, M=20, R3	70,000	exact	1.000	1.000	0.873	0.978	better
GSM8K, M=20, R3	70,000	swap1	0.000	0.998	0.887	0.927	better
GSM8K, M=20, R3	70,000	swap3	0.000	1.000	0.872	0.903	better
GSM8K, M=20, R3	70,000	pilot	0.752	1.000	0.857	0.953	better
MMLU, M=10, R1	40,000	exact	0.908	0.902	0.143	0.457	better
MMLU, M=10, R1	40,000	swap1	0.000	0.930	0.162	0.275	better
MMLU, M=10, R1	40,000	swap3	0.000	0.912	0.165	0.157	–
MMLU, M=10, R1	40,000	pilot	0.127	0.930	0.173	0.145	–
MMLU, M=10, R1	70,000	exact	0.998	0.998	0.877	0.965	better
MMLU, M=10, R1	70,000	swap1	0.000	0.998	0.850	0.872	–
MMLU, M=10, R1	70,000	swap3	0.000	0.995	0.842	0.782	worse
MMLU, M=10, R1	70,000	pilot	0.313	0.998	0.852	0.842	–
MMLU, M=10, R3	40,000	exact	0.980	0.980	0.685	0.833	better
MMLU, M=10, R3	40,000	swap1	0.000	0.967	0.668	0.755	better
MMLU, M=10, R3	40,000	swap3	0.000	0.970	0.665	0.700	better
MMLU, M=10, R3	40,000	pilot	0.662	0.957	0.675	0.787	better
MMLU, M=10, R3	70,000	exact	1.000	1.000	0.952	0.982	better
MMLU, M=10, R3	70,000	swap1	0.000	0.995	0.950	0.958	–
MMLU, M=10, R3	70,000	swap3	0.000	0.998	0.965	0.950	–
MMLU, M=10, R3	70,000	pilot	0.853	1.000	0.947	0.970	better

MMLU, M=10, R1

Rank	Model	Accuracy
1	`OpenBuddyEA/openbuddy-llama-30b-v7.1-bf16`	0.5763
2	`diffnamehard/Psyfighter2-Noromaid-ties-Capybara-13B`	0.5661
3	`kingbri/chronolima-airo-grad-l2-13B`	0.5558
4	`Undi95/Nous-Hermes-13B-Code`	0.5456
5	`jondurbin/airoboros-c34b-2.1`	0.5354
6	`CallComply/zephyr-7b-beta-128k`	0.5251
7	`HyperbeeAI/Tulpar-7b-v0`	0.5151
8	`hfl/chinese-alpaca-2-13b-16k`	0.5042
9	`Charlie911/vicuna-7b-v1.5-lora-timedial`	0.4934
10	`PygmalionAI/pygmalion-2-7b`	0.4834

MMLU, M=10, R3

Rank	Model	Accuracy
1	`kyujinpy/PlatYi-34B-Llama`	0.7857
2	`KnutJaegersberg/Deacon-34b-Adapter`	0.7633
3	`AA051610/A0106`	0.7391
4	`AA051611/limb`	0.7092
5	`cognitivecomputations/yayi2-30b-llama`	0.6988
6	`OpenAssistant/llama2-70b-oasst-sft-v10`	0.6755
7	`jeonsworld/CarbonVillain-en-10.7B-v1`	0.6543
8	`TomGrc/FusionNet_passthrough`	0.6394
9	`Dans-DiscountModels/Dans-07YahooAnswers-7b`	0.6211
10	`Mihaiii/Metis-0.1`	0.6000

MMLU, M=20, R1

Rank	Model	Accuracy
1	bongchoi/MoMo-70B-LoRA-V1.1	0.6618
2	fblgit/UNA-SOLAR-10.7B-Instruct-v1.0	0.6516
3	janhq/supermario-slerp	0.6410
4	perlthoughts/Chupacabra-8x7B-MoE	0.6308
5	perlthoughts/Chupacabra-7B-v2.03-128k	0.6208
6	jae24/openhermes_dpo_norobot_0201	0.6105
7	oh-yeontaek/llama-2-13B-LoRA-assemble	0.5993
8	euclaise/Ferret-7B	0.5875
9	OpenBuddyEA/openbuddy-llama-30b-v7.1-bf16	0.5763
10	diffnamehard/Psyfighter2-Noromaid-ties-Capybara-13B	0.5661
11	kingbri/chronolima-airo-grad-l2-13B	0.5558
12	Undi95/Nous-Hermes-13B-Code	0.5456
13	jondurbin/airoboros-c34b-2.1	0.5354
14	CallComply/zephyr-7b-beta-128k	0.5251
15	HyperbeeAI/Tulpar-7b-v0	0.5151
16	hfl/chinese-alpaca-2-13b-16k	0.5042
17	Charlie911/vicuna-7b-v1.5-lora-timedial	0.4934
18	PygmalionAI/pygmalion-2-7b	0.4834
19	davzoku/cria-llama2-7b-v1.3	0.4727
20	wang7776/Mistral-7B-Instruct-v0.2-sparsity-20	0.4625

MMLU, M=20**, R3**

Rank	Model	Accuracy
1	brucethemoose/Yi-34B-200K-DARE-merge-v5	0.7776
2	Mihaiii/Pallas-0.5-LASER-exp2-0.1	0.7474
3	jondurbin/spicyboros-70b-2.2	0.7047
4	Sao10K/Euryale-L2-70B	0.6839
5	bongchoi/MoMo-70B-LoRA-V1.1	0.6618
6	cookinai/BruinHermes	0.6435
7	PulsarAI/OpenHermes-2.5-neural-chat-v3-3-Slerp	0.6322
8	one-man-army/una-neural-chat-v3-3-P2-OMA	0.6192
9	diffnamehard/Mistral-CatMacaroni-slerp-gradient	0.6087
10	notbdq/alooowso	0.5931
11	PulsarAI/2x-LoRA-Assemble-Nova-13B	0.5830
12	Undi95/MLewd-Chat-v2-13B	0.5676
13	Undi95/Emerald-13B	0.5550
14	IGeniusDev/llama13B-quant8-testv1-openorca-customdataset	0.5357
15	CallComply/zephyr-7b-beta-128k	0.5251
16	HyperbeeAI/Tulpar-7b-v0	0.5151
17	DopeorNope/LaOT	0.4963
18	NewstaR/Koss-7B-chat	0.4588
19	itsliupeng/openllama-7b-base	0.4222
20	ashercn97/giraffe-7b	0.3816

ARC-C, M=10**, R1**

Rank	Model	Accuracy
1	diffnamehard/Mistral-CatMacaroni-slerp-gradient	0.6553
2	decapoda-research/Antares-11b-v1	0.6451
3	huggingface/llama-65b	0.6348
4	Mihaiii/Pallas-0.5-LASER-0.6	0.6246
5	Undi95/ReMM-v2.1-L2-13B	0.6143
6	migtissera/Tess-7B-v1.4	0.6041
7	Secbone/llama-2-13B-instructed	0.5939
8	xzuyn/Alpacino-SuperCOT-13B	0.5836
9	SciPhi/SciPhi-Self-RAG-Mistral-7B-32k	0.5734
10	HyperbeeAI/Tulpar-7b-v0	0.5631

ARC-C, M=10**, R3**

Rank	Model	Accuracy
1	udkai/Turdus	0.7338
2	cookinai/BruinHermes	0.7014
3	janhq/supermario-slerp	0.6894
4	superlazycoder/NeuralPipe-7B-slerp	0.6758
5	openaccess-ai-collective/DPOpenHermes-7B	0.6596
6	simonveitner/MathHermes-2.5-Mistral-7B	0.6476
7	xxyyy123/Mistral7B_adaptor_v1	0.6297
8	PeanutJar/Mistral-v0.1-PeanutButter-v0.0.2-7B	0.6177
9	YeungNLP/firefly-llama2-13b-v1.2	0.6067
10	SanjiWatsuki/neural-chat-7b-v3-3-wizardmath-dare-me	0.5964

ARC-C, M=20**, R1**

Rank	Model	Accuracy
1	nfaheem/Marcoroni-7b-DPO-Merge	0.7304
2	jondurbin/bagel-dpo-34b-v0.2	0.7193
3	jondurbin/spicyboros-70b-2.2	0.7073
4	Weyaxi/Helion-4x34B	0.6971
5	Walmart-the-bag/Solar-10.7B-Cato	0.6869
6	superlazycoder/NeuralPipe-7B-slerp	0.6758
7	maywell/PiVoT-SUS-RP	0.6655
8	diffnamehard/Mistral-CatMacaroni-slerp-gradient	0.6553
9	decapoda-research/Antares-11b-v1	0.6451
10	huggingface/llama-65b	0.6348
11	Mihaiii/Pallas-0.5-LASER-0.6	0.6246
12	Undi95/ReMM-v2.1-L2-13B	0.6143
13	migtissera/Tess-7B-v1.4	0.6041
14	Secbone/llama-2-13B-instructed	0.5939
15	xzuyn/Alpacino-SuperCOT-13B	0.5836
16	SciPhi/SciPhi-Self-RAG-Mistral-7B-32k	0.5734
17	HyperbeeAI/Tulpar-7b-v0	0.5631
18	KnutJaegersberg/Qwen-14B-Llamafied	0.5520
19	defog/sqlcoder-34b-alpha	0.5418
20	wang7776/Llama-2-7b-chat-hf-10-sparsity	0.5316

ARC-C, M=20**, R3**

Rank	Model	Accuracy
1	udkai/Turdus	0.7338
2	cookinai/BruinHermes	0.7014
3	janhq/supermario-slerp	0.6894
4	superlazycoder/NeuralPipe-7B-slerp	0.6758
5	rishiraj/oswald-2x7b	0.6647
6	mrfakename/NeuralOrca-7B-v1	0.6527
7	jondurbin/bagel-7b-v0.1	0.6391
8	bofenghuang/vigostral-7b-chat	0.6263
9	Dans-DiscountModels/Dans-07YahooAnswers-7b	0.6152
10	migtissera/Tess-7B-v1.4	0.6041
11	Secbone/llama-2-13B-instructed	0.5939
12	totally-not-an-llm/EverythingLM-13b-V3-peft	0.5836
13	SciPhi/SciPhi-Self-RAG-Mistral-7B-32k	0.5734
14	HyperbeeAI/Tulpar-7b-v0	0.5631
15	ashercn97/manatee-7b	0.5452
16	TaylorAI/Flash-Llama-7B	0.5307
17	WeOpenML/Alpaca-7B-v1	0.4906
18	openlm-research/open_llama_7b	0.4599
19	digitous/Javalion-R	0.4172
20	cognitivecomputations/yayi2-30b-llama	0.3567

HellaSwag, M=10**, R1**

Rank	Model	Accuracy
1	Mihaiii/Metis-0.3	0.8480
2	cognitivecomputations/dolphin-2.2.1-mistral-7b	0.8380
3	TIGER-Lab/TIGERScore-13B	0.8279
4	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r8-gate_up_down	0.8179
5	luffycodes/vicuna-class-shishya-ac-hal-13b-ep3	0.8078
6	HyperbeeAI/Tulpar-7b-v1	0.7969
7	llm-agents/tora-7b-v1.0	0.7868
8	jondurbin/airoboros-7b	0.7765
9	luffycodes/llama-shishya-7b-ep3-v1	0.7663
10	uukuguy/speechless-code-mistral-7b-v2.0	0.7561

HellaSwag, M=10**, R3**

Rank	Model	Accuracy
1	Sao10K/WinterGoddess-1.4x-70B-L2	0.9011
2	flemmingmiguel/NeuDist-Ro-7B	0.8748
3	superlazycoder/NeuralPipe-7B-slerp	0.8617
4	speechlessai/speechless-mistral-7b-dare-0.85	0.8493
5	xxyyy123/Mistral7B_adaptor_v1	0.8381
6	HenryJJ/Instruct_Mistral-7B-v0.1_Dolly15K	0.8262
7	Mihaiii/Pallas-0.5-LASER-0.6	0.8160
8	HuggingFaceH4/starchat-beta	0.8059
9	itsliupeng/openllama-7b-base	0.7640
10	klosax/open_llama_13b_600bt_preview	0.7243

HellaSwag, $M{=}20$**, R1**

Rank	Model	Accuracy
1	Sao10K/WinterGoddess-1.4x-70B-L2	0.9011
2	Weyaxi/SauerkrautLM-UNA-SOLAR-Instruct	0.8830
3	Sao10K/Euryale-L2-70B	0.8707
4	FelixChao/ExtremeDolphin-MoE	0.8607
5	AA051610/A0106	0.8505
6	lu-vae/llama2-13B-sharegpt4-orca-openplatypus-8w	0.8404
7	migtissera/Tess-XS-v1-3-yarn-128K	0.8296
8	The-Face-Of-Goonery/Chronos-Beluga-v2-13bfp16	0.8194
9	martyn/llama2-megamerge-dare-13b-v2	0.8093
10	Aeala/GPT4-x-Alpasta-13b	0.7992
11	TheBloke/Wizard-Vicuna-7B-Uncensored-HF	0.7885
12	huggingface/llama-7b	0.7782
13	922-CA/monika-ddlc-7b-v1	0.7678
14	WeOpenML/Alpaca-7B-v1	0.7571
15	GOAT-AI/GOAT-7B-Community	0.7463
16	hyunseoki/ko-ref-llama2-13b	0.7356
17	klosax/open_llama_13b_600bt_preview	0.7243
18	fireballoon/baichuan-vicuna-chinese-7b	0.7112
19	yeontaek/WizardCoder-Python-13B-LoRa	0.6960
20	ehartford/dolphin-2.2-yi-34b-200k	0.6818

HellaSwag, $M{=}20$**, R3**

Rank	Model	Accuracy
1	Sao10K/WinterGoddess-1.4x-70B-L2	0.9011
2	mlabonne/NeuralMarcoro14-7B	0.8759
3	Weyaxi/OpenHermes-2.5-neural-chat-v3-3-openchat-3.5-1210-Slerp	0.8632
4	flemmingmiguel/Distilled-HermesChat-7B	0.8521
5	simonveitner/MathHermes-2.5-Mistral-7B	0.8419
6	jondurbin/bagel-7b-v0.1	0.8314
7	CHIH-HUNG/llama-2-13b-FINETUNE3_3.3w-r16-gate_up_down	0.8189
8	OpenBuddy/openbuddy-mixtral-7bx8-v16.3-32k	0.8083
9	JosephusCheung/Pwen-14B-Chat-20_30	0.7978
10	TinyPixel/testmodel2	0.7878
11	mosaicml/mpt-7b	0.7753
12	Charlie911/vicuna-7b-v1.5-lora-mixed-datasets-time-unit	0.7641
13	GOAT-AI/GOAT-7B-Community	0.7463
14	hyunseoki/ko-ref-llama2-13b	0.7356
15	klosax/open_llama_13b_600bt_preview	0.7243
16	yeontaek/WizardCoder-Python-13B-LoRa	0.6960
17	qblocks/codellama_7b_DolphinCoder	0.6550
18	heegyu/RedTulu-Uncensored-3B-0719	0.6255
19	KnutJaegersberg/Qwen-1_8B-Llamafied	0.5887
20	uukuguy/speechless-coder-ds-6.7b	0.5246

WinoGrande, $M{=}10$**, R1**

Rank	Model	Accuracy
1	beberik/rawr	0.7956
2	Dans-DiscountModels/Dans-07YahooAnswers-7b	0.7853
3	OpenBuddyEA/openbuddy-llama-30b-v7.1-bf16	0.7751
4	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r8-q_k_v_o	0.7648
5	camel-ai/CAMEL-13B-Combined-Data	0.7545
6	maywell/Synatra-RP-Orca-2-7b-v0.1	0.7443
7	NousResearch/CodeLlama-34b-hf	0.7340
8	lmsys/vicuna-13b-v1.5-16k	0.7238
9	uukuguy/speechless-code-mistral-7b-v2.0	0.7135
10	jondurbin/airoboros-7b-gpt4-1.4.1-qlora	0.7032

WinoGrande, M=10, R3

Rank	Model	Accuracy
1	nfaheem/Marcoroni-7b-DPO-Merge	0.8524
2	flemmingmiguel/NeuDist-Ro-7B	0.8208
3	superlazycoder/NeuralPipe-7B-slerp	0.8019
4	PulsarAI/OpenHermes-2.5-neural-chat-v3-3-Slerp	0.7916
5	bn22/OpenHermes-2.5-Mistral-7B-MISALIGNED	0.7766
6	KnutJaegersberg/Qwen-14B-Llamafied	0.7656
7	BELLE-2/BELLE-Llama2-13B-chat-0.4M	0.7553
8	ashercn97/manatee-7b	0.7451
9	umd-zhou-lab/recycled-wizardlm-7b-v2.0	0.7151
10	fireballoon/baichuan-vicuna-chinese-7b	0.6685

WinoGrande, M=20, R1

Rank	Model	Accuracy
1	Weyaxi/Bagel-Hermes-2x34b	0.8477
2	Weyaxi/SauerkrautLM-UNA-SOLAR-Instruct	0.8374
3	Doctor-Shotgun/mythospice-limarp-70b	0.8272
4	shadowml/Marcoro14-7B-ties	0.8161
5	perlthoughts/Starling-LM-alpha-8x7B-MoE	0.8058
6	beberik/rawr	0.7956
7	Dans-DiscountModels/Dans-07YahooAnswers-7b	0.7853
8	OpenBuddyEA/openbuddy-llama-30b-v7.1-bf16	0.7751
9	CHIH-HUNG/llama-2-13b-FINETUNE4_3.8w-r8-q_k_v_o	0.7648
10	camel-ai/CAMEL-13B-Combined-Data	0.7545
11	maywell/Synatra-RP-Orca-2-7b-v0.1	0.7443
12	NousResearch/CodeLlama-34b-hf	0.7340
13	Weyaxi/Platypus-Nebula-v2-7B	0.7222
14	codellama/CodeLlama-34b-Python-hf	0.7119
15	teknium/Mistral-Trismegistus-7B	0.7017
16	ashercn97/giraffe-7b	0.6898
17	shareAI/CodeLLaMA-chat-13b-Chinese	0.6788
18	fireballoon/baichuan-vicuna-chinese-7b	0.6685
19	yeontaek/WizardCoder-Python-13B-LoRa	0.6543
20	cognitivecomputations/dolphin-2.2-yi-34b-200k	0.6425

WinoGrande, M=20, R3

Rank	Model	Accuracy
1	udkai/Turdus	0.8666
2	nfaheem/Marcoroni-7b-DPO-Merge	0.8524
3	jondurbin/spicyboros-70b-2.2	0.8382
4	jarradh/llama2_70b_chat_uncensored	0.8256
5	janhq/supermario-slerp	0.8129
6	superlazycoder/NeuralPipe-7B-slerp	0.8019
7	PulsarAI/OpenHermes-2.5-neural-chat-v3-3-Slerp	0.7916
8	bn22/OpenHermes-2.5-Mistral-7B-MISALIGNED	0.7766
9	CHIH-HUNG/llama-2-13b-FINETUNE5_4w-r4-q_k_v_o	0.7664
10	Gryphe/MythoMix-L2-13b	0.7545
11	maywell/Synatra-RP-Orca-2-7b-v0.1	0.7443
12	hfl/chinese-alpaca-2-13b-16k	0.7340
13	Weyaxi/Platypus-Nebula-v2-7B	0.7222
14	codellama/CodeLlama-34b-Python-hf	0.7119
15	teknium/Mistral-Trismegistus-7B	0.7017
16	hyunseoki/ko-ref-llama2-13b	0.6914
17	fireballoon/baichuan-vicuna-chinese-7b	0.6685
18	togethercomputer/GPT-JT-6B-v0	0.6480
19	wenge-research/yayi-7b	0.6227
20	uukuguy/speechless-coder-ds-6.7b	0.5888

V WHERE EVERY NUMBER COMES FROM

Every number in the paper is computed by a released script from the public score matrix, and nothing is typed by hand. Table 14 maps each claim to the script that produces it and the file it writes. Random streams are seeded from a SHA-256 digest of the configuration and the purpose of the stream, so a rerun reproduces every Monte Carlo number to the last digit on the same software versions. The exact computations use no randomness at all.

Table 14: Scripts and artifacts behind each claim. Paths are relative to the released repository.

Claim	Script	Artifact
Theorem 1 bound in Table 5, Procedure 5	`m11_r3.py`	recomputed inside `w17_headline.py`
Theorem 2 bound and its k	`w13_mixture.py`	`w13-mixture/mixture_v2.json`
Completion curves for three rules	`w9_measure.py`	`w9-measure/measure.json`
n_{95} and its bootstrap interval	`w12_interp.py`	`w9-measure/interp.json`
Ratios 1.61–1.81, savings over the uniform-budget certificate, Table 5 rows	`w17_headline.py`	`w17-headline/headline.json`, `table_rows.tex`
Single-comparison gap and its decomposition	`w14_decompose.py`	`w9-measure/decompose.json`
Proposal–validation certificate	`w11_split.py`	`w11-split/split_correct.json`
Validity sweep, 48 cells	`w9_validity.py`	`w9-measure/validity.json`
Proposition 5 and the table of Appendix F	`w23_holm_exact.py`	`w23-holm/holm_exact.json`
Exact test against empirical Bernstein	`w20_bernstein.py`	`w20-bernstein/bernstein.json`
Weighted levels, chain test, Holm	`m8_baselines.py`	`m8-baselines/baselines_v2.json`
Table 2, Table 4, truncation check	`w27_bound_checks.py`	`w27-bound-checks/bound_checks.json`
Closed-form checks of Appendix B	`m7_exact_tv.py`	printed
Sensitivity to δ	`w19_delta.py`	`w19-delta/delta.json`
Cycles under unequal allocation	`w21_transitivity.py`	`w21-transitivity/rates.json`
Forced empirical order	`w22_forced.py`	`w22-forced/forced.json`
Certification of the Theorem 2 bounds	`w24_mixture_verify.py`	`w24-mixture-verify/verify.json`
Gap-floor and M sensitivity	`w26_sensitivity.py`	`w26-sensitivity/ sensitivity.json`
Held-out certificates, Table 1	`w25_select_validate.py`	`w25-select-validate/select_validate.json`
i.i.d. instance of Proposition 5	`w23_holm_exact.py`	`w23-holm/holm_iid.json`
Figures 1, 2 and 3, Table 1	`fig_main.py`	`figures/fig_*.pdf`, `generated/tab_summary.tex`
MMLU order right at 17,600 items (§1)	`w22_forced.py`	`w22-forced/forced.json`
Every appendix figure and generated table	`w18_appendix.py`	`figures/app_*.pdf`, `generated/*.tex`

W What each procedure outputs and guarantees

Table 15 lists every procedure in the paper by what it outputs and which error it controls, so that comparisons between them are made at equal guarantees. All rows assume a fixed set of models and i.i.d. shared items, and "controls" means that the probability of outputting a relation that is an inversion is at most δ, unconditionally, under any dependence between models.

Table 15: Procedures by output and guarantee. The last two rows are reported only for comparison.

Procedure	Output	Guarantee	Where
Theorem 4, all pairs at $\delta/(M-1)$	complete order or nothing	controls, proved	§4.2
Uniform budget, each pair at δ/K	certified edges, partial order	controls, proved	§2.4
Both on one sample, each at $\delta/2$	complete order, else partial	controls jointly, proved	§2.4
Proposal–validation certificate	complete order or nothing	controls, proved; proposal items count	§4.5, App. J
Weighted levels, floor	certified edges	controls for any order fixed in advance	App. G
Chain test on a frozen order	complete order or nothing	controls if the order is independent of the items	App. G
Holm over $2K$ directional nulls	certified edges	controls, proved; never beats the uniform-budget certificate	App. F
Empirical Bernstein, any scores in $[0, 1]$	certified edges	controls, proved	App. K
Holm on data-selected p-values	certified edges	does *not* control (Prop. 5)	§4.3
Holm on data-selected p-values, complete mode	complete order or nothing	descriptive comparison	App. F
Sorting the sample means	complete order	none; ranking risk only	App. L

The lower bounds of Section 3 apply to any procedure that meets the stated validity and success requirements, including adaptive or sequential ones whose item count is bounded by n.

X Assumptions and what each guarantee covers

This appendix states, in one place, the setting each result is proved or measured in.

The item law. Every guarantee concerns a population of items from which the evaluation draws independently, and it certifies the ranking under that population. The certificates need no other assumption. What they control is the sampling error of the evaluation, the part of a leaderboard's uncertainty that depends on how many items are scored.

Replay. The measured n_{95} values are computed under each bank's own empirical law, where the truth is known exactly, and they give the items a certificate needs on a population distributed like the bank. The certificate and the lower bounds are evaluated under the same law, so their ratio compares the two on one statistical problem. For the unknown item population, Appendix C gives lower confidence bounds for comparison sets fixed on independent items; they are 7.9 to 979 times smaller than the plug-in values on the same items.

The lower bounds. Theorem 1 is exact for one comparison at every n, and on these banks the certificate meets it within 0.6 to 1.9% when only one comparison matters. Theorem 2 prices many comparisons through a uniform mixture of single-pair tilts over the hardest k adjacent pairs, one admissible family of alternatives. The stronger of the two bounds lies below the smallest achievable requirement, so the factor 1.61 to 1.81 bounds from above what any procedure could save on these configurations.

Which models are ranked. The theorems are stated for a fixed set of models. For replay, the bank is the item law and the chosen models are fixed before any resampling. For population statements, Table 1 and Appendix Q choose the models on 30% of each bank and certify them on the other 70%, and the whole-bank results are descriptive of the observed matrices.

Binary scores. Theorem 3 and the lower bounds use the three-point form of a paired difference. Appendix K gives the certificate for any score in $[0, 1]$, with the empirical Bernstein radius in place of the exact test.

Shared items and adaptivity. The certificates score every model on the same n items. The lower bounds hold for adaptive and sequential procedures as well, because an adaptive transcript is a function of the drawn items, so Theorem 2 also bounds what spending more items on hard pairs could save. Appendix L describes the cycles that unequal allocation creates.

Step-down procedures. Proposition 5 concerns Holm applied to p-values whose direction is chosen by the data, used as a certificate of individual directions. Holm over pre-specified directional nulls is valid (Appendix F). The complete-order mode of the data-selected version gives 0.0056 on the same construction, and the paper certifies complete orders with Theorem 4, which is proved.

Conditional and unconditional error. Every guarantee bounds the probability that a wrong certificate is output, over the draw of the items. In the synthetic cell with $M = 4$, $n = 20{,}000$, a gap of 0.0005 and independent scores, Theorem 4 certified in 63 of $2{,}000$ runs and was wrong in 30 of them, which is within δ unconditionally. At low power a certificate is read together with its power curve.

Cost. The lower bounds count independent item draws per model. Model calls and tokens are a separate cost, which Appendix I relates to items.

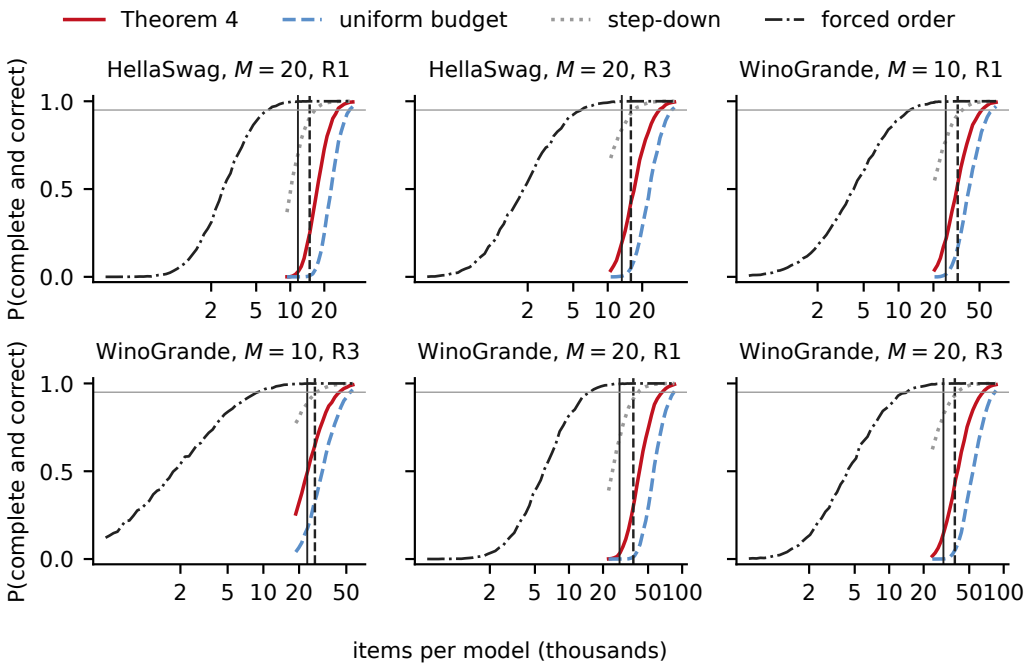

Figure 6: As Figure 4, the remaining HellaSwag and WinoGrande configurations.

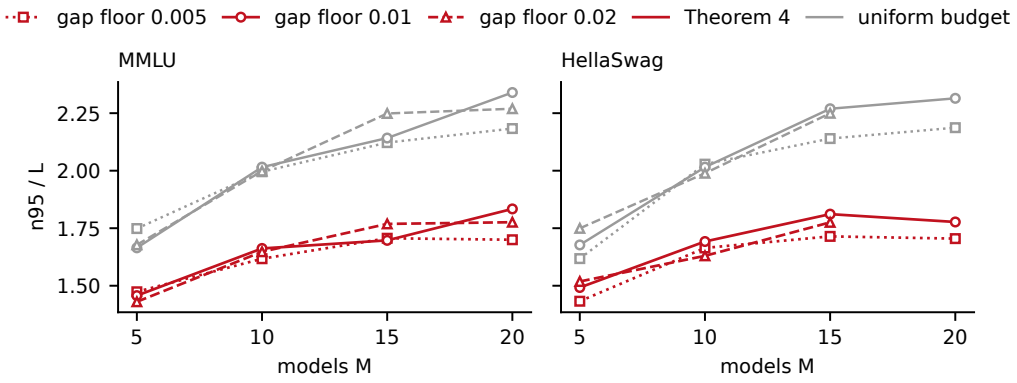

Figure 7: n_{95} over the larger lower bound L against the number of models, for three gap floors, on MMLU and HellaSwag. Red: Theorem 4. Grey: the uniform-budget certificate.

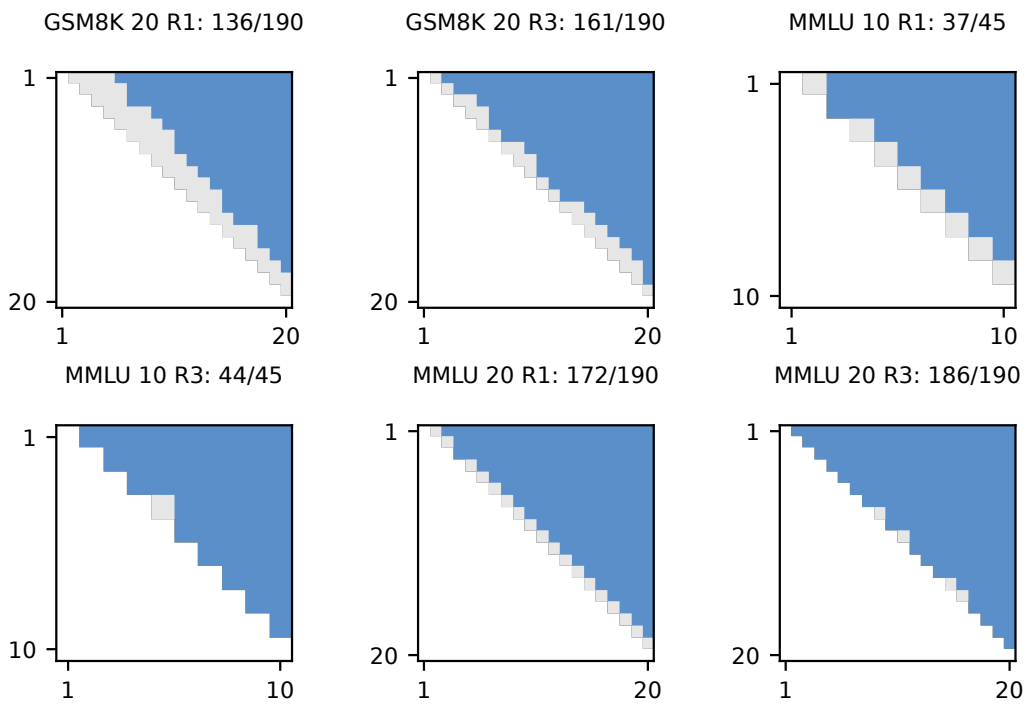

Figure 8: Relations the uniform-budget certificate certifies on the whole bank, GSM8K and MMLU. Titles give the certified count out of K.

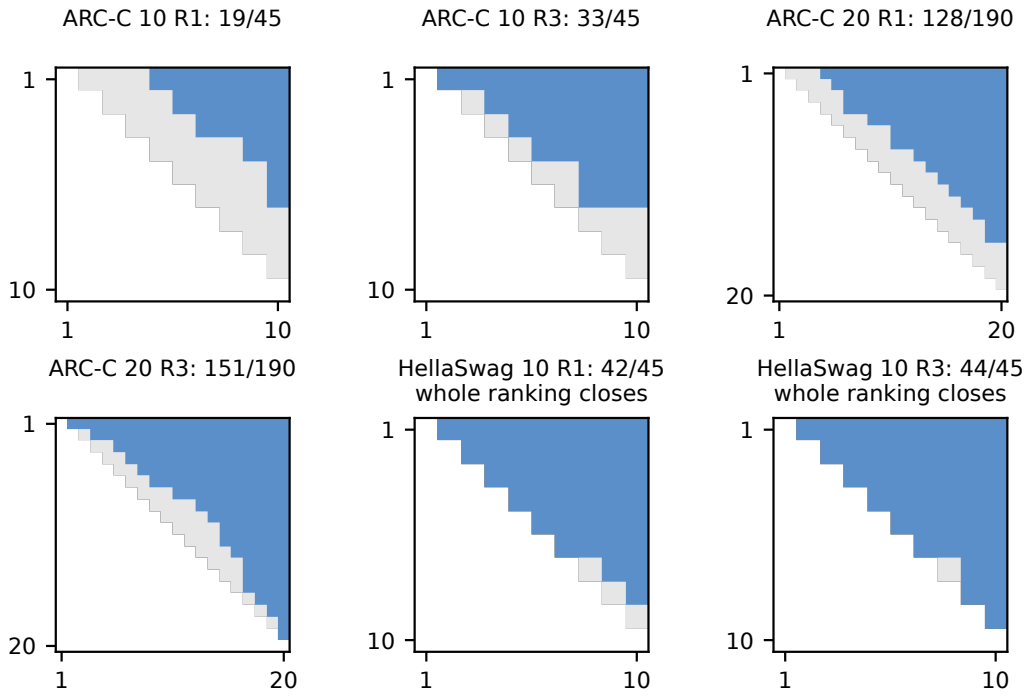

Figure 9: As Figure 8, ARC-Challenge and HellaSwag.

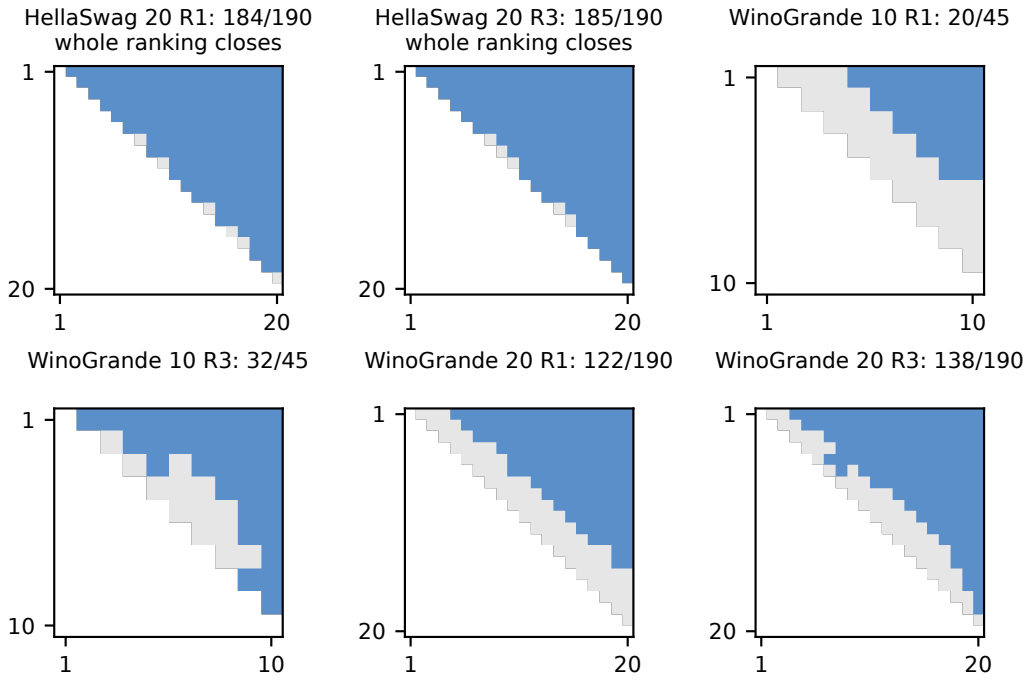

Figure 10: As Figure 8, the remaining configurations.

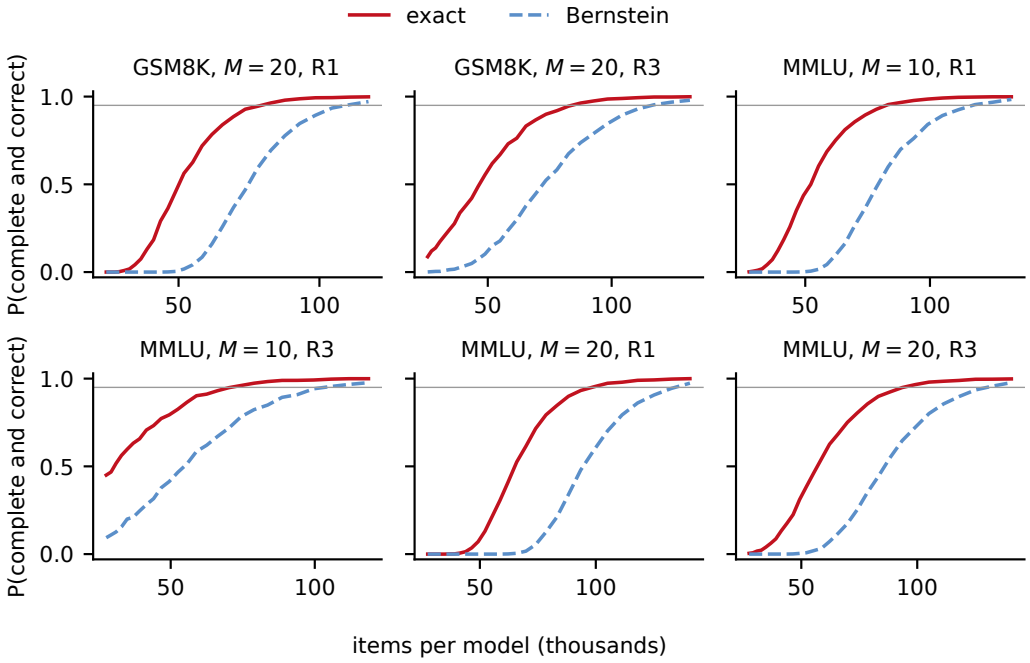

Figure 11: Probability of certifying the complete ranking correctly under the exact test and under empirical Bernstein, both at δ/K per pair, GSM8K and MMLU.

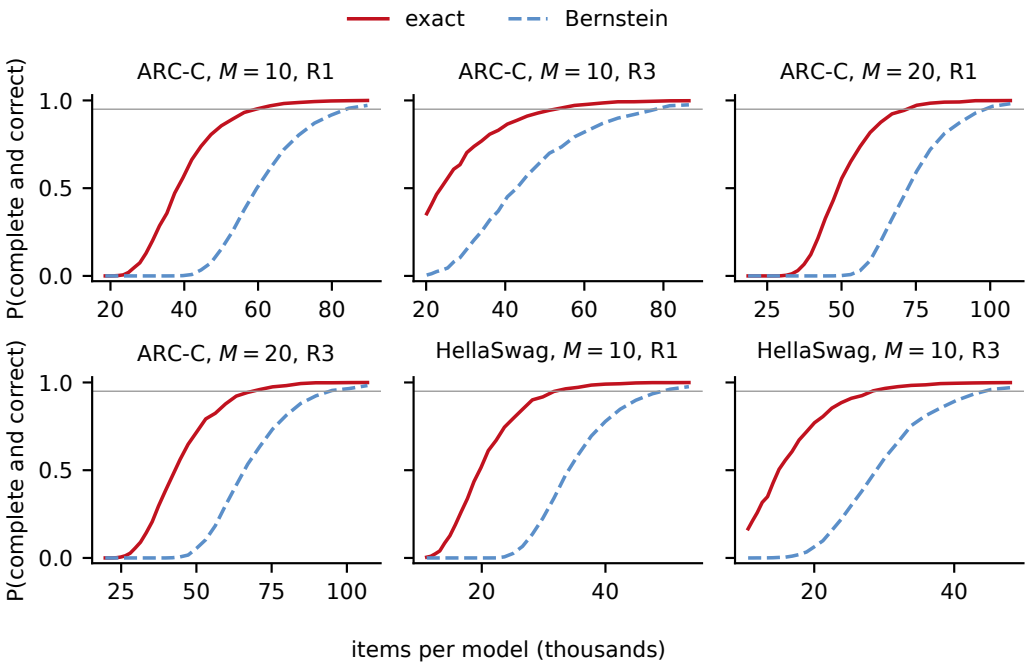

Figure 12: As Figure 11, ARC-Challenge and HellaSwag.

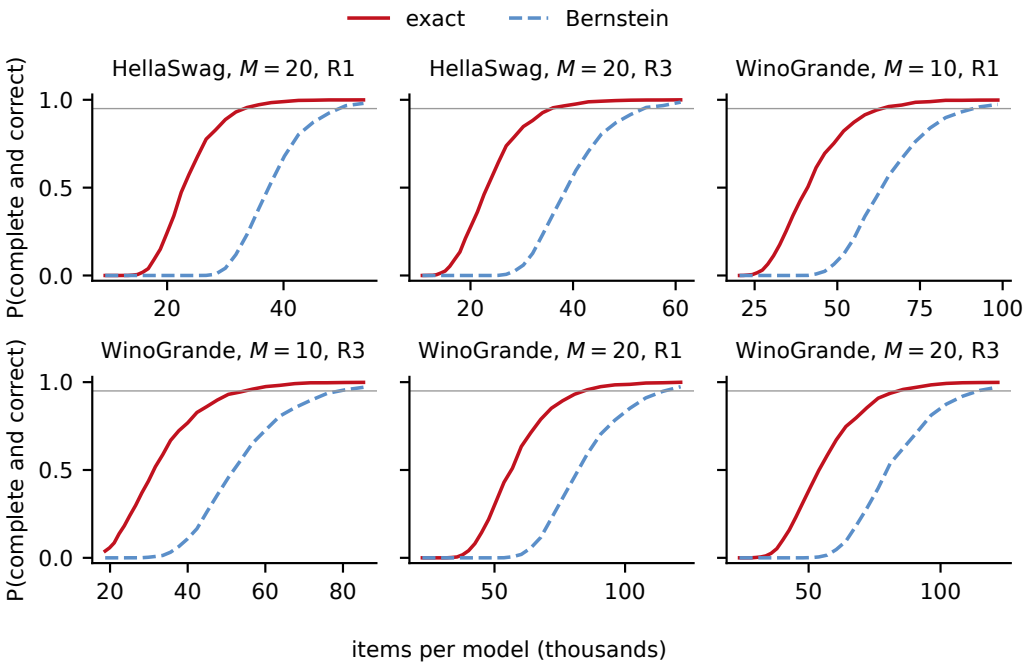

Figure 13: As Figure 11, the remaining HellaSwag and WinoGrande configurations.

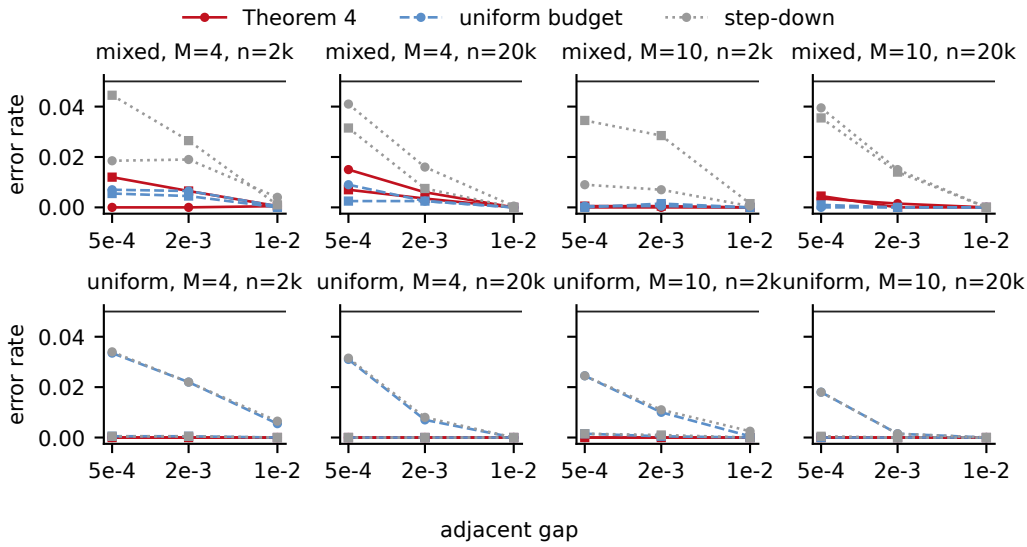

Figure 14: Fraction of runs with a certified relation in the wrong direction. Circles: $\rho = 0$; squares: $\rho = 0.5$. The horizontal line is $\delta = 0.05$.