RANK CONFIDENCE SEQUENCES: ANYTIME-VALID LEADERBOARDS

Anonymous authors
Paper under double-blind review

ABSTRACT

Leaderboards rank models by their average scores on benchmark items, and they are consulted repeatedly while the evaluation is still running. Existing confidence intervals for a model's rank control their error rate only if they are computed once, after a number of items chosen in advance. If they are recomputed as results arrive, and the evaluation stops once they look decisive, their error rate exceeds its nominal level. Anytime-valid methods keep their guarantees at all sample sizes simultaneously and hence under any stopping rule. They exist for the accuracy of one model, for one pair of models and for the set of models that may be best. For pairwise battles they also give ranks. None gives ranks when all models are scored on the same items, which makes their scores dependent. We construct rank confidence sequences: for every model, a set of ranks that contains its true rank, simultaneously for all models and at all times, at a chosen error level α, in finite samples. The construction combines betting e-processes, one for each ordered pair of models, with closed testing over the possible orderings of the models. It allows any dependence between the models' scores on an item. The method has two advantages. A leaderboard can be inspected after every item without inflating its error rate. The evaluation of each model can stop as soon as the question asked about it is answered, which saves compute. When results are examined only once, halfway through or later, little power is lost relative to fixed-sample methods. The paper quantifies these advantages in simulations and on public leaderboard data.

1 INTRODUCTION

The problem. Confidence intervals for leaderboard ranks are designed to be computed once, but leaderboards are consulted many times while the evaluation is still running. A benchmark leaderboard scores each of M models on a common set of test items, such as exam questions, and ranks the models by their average scores. The averages come from finitely many items, so the ranking is uncertain. Several methods therefore attach a confidence interval to each model's rank (Mogstad et al., 2024; Al Mohamad et al., 2022; Neuhof & Benjamini, 2024). At confidence level 95%, these intervals contain the true ranks of all models simultaneously with probability at least 95%, some only asymptotically. The guarantee assumes a specific protocol. The number of items is chosen before any results are seen. The intervals are then computed once, after all of those items have been scored. We call such intervals *fixed-sample*.

Repeated looks. Inspecting a leaderboard repeatedly as results arrive breaks the fixed-sample guarantee. In practice, models are scored item by item or batch by batch, and the standings are examined along the way. We call each examination a *look*. Because evaluation is expensive (Ghosh et al., 2026), it is cut short when the ranking seems clear and extended only when two models seem close. Each look, however, is another chance for a fixed-sample interval to be wrong. Stopping when the intervals seem decisive also tends to stop exactly when one of them is wrong. The probability that at least one look produces a false statement therefore exceeds the nominal level. This effect of repeated significance tests has long been known, and Armitage et al. (1969) computed its size. In our simulation with six equally accurate models examined 200 times, fixed-sample rank sets at nominal level 5% make at least one false statement in 30% to 51% of runs (Section 5, experiment E1). The fixed-sample guarantee is not wrong: it covers one look, not the way leaderboards are used.

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.

What is needed. A leaderboard that is monitored needs rank intervals that are *anytime-valid*. Fix a level $\alpha \in (0, 1)$ in advance. A procedure is anytime-valid if the probability that it is wrong at any look, over the whole evaluation, is at most α. Equivalently, its error guarantee holds at all sample sizes simultaneously. It then holds at whatever sample size the evaluation stops, even when the decision to stop depends on the data seen so far. For a single mean, intervals with this property are called *confidence sequences* (Howard et al., 2021; Waudby-Smith & Ramdas, 2024). Two further features of leaderboards matter. First, the statement must cover the ranks of all models at once, so the method must correct for making many comparisons. Second, all models are scored on the same items, so their scores are dependent. An item that is easy for one model tends to be easy for the others, and the strength of this dependence is unknown.

What exists. To our knowledge, no existing method gives anytime-valid confidence sets for the ranks of all models when the models are scored on the same items. Existing rank intervals are fixed-sample. They certify pairwise comparisons of the form "model j is better than model l" with a multiple-testing procedure and convert them into rank intervals (Mogstad et al., 2024; Al Mohamad et al., 2022; Neuhof & Benjamini, 2026a;b). The anytime-valid methods that exist answer other questions. Some track the accuracy of one model (Zhou et al., 2026; Hsu & Shekhar, 2026), and one compares a single pair of models (Kotawala, 2026). Arnold et al. (2026) certify which models may be the best, a statement about rank one only. Gu et al. (2026) give anytime-valid rank sets for a different kind of data, pairwise battles. In a battle, two models answer the same prompt and a judge picks the better answer, with the pair chosen adaptively. Gu et al. (2026) correct for multiplicity by Bonferroni, testing each of the $M(M-1)$ ordered pairs at level $\alpha/\{M(M-1)\}$. The logical relations among the pairs are used only afterwards, to extend certified comparisons by transitivity. Group-sequential designs with Pocock boundaries (Pocock, 1977; Arviv et al., 2026) stop pairwise comparisons early, but only at looks scheduled in advance and without a correction across pairs. A side-by-side comparison of these methods is in Appendix F.

What we do. We construct *rank confidence sequences*. For every model, a rank confidence sequence gives a set of ranks that contains the model's true rank. This holds for all models at once and at every look, with probability at least $1 - \alpha$. A model's *ability* is its mean score. Its *true rank* is one plus the number of models with strictly higher ability (Section 2). The ability can be the model's average over a fixed benchmark. It can also be the model's expected score on items drawn from a population. The guarantee holds in finite samples, with no asymptotic approximation. It requires per-item scores in $[0, 1]$. On a fixed benchmark it also requires that the items be evaluated in a random order, which the evaluator controls. It makes no assumption about how the scores of different models are correlated. The method keeps a list of candidate orderings of the models, rankings in which ties are allowed, and reports only what all surviving orderings agree on. For each pair of models, it measures the evidence that one scores higher than the other as the wealth of an imaginary gambler betting on it. It discards an ordering once the average of these wealths over the comparisons the ordering denies has grown large. Section 3 gives the details.

Contributions.

1. **A leaderboard can be read at any time without inflating its error.** With probability at least $1 - \alpha$, all rank statements are correct simultaneously, at every look and under any stopping rule (Theorem 4.2; Section 5, E1 and E2).
2. **A leaderboard can state how much of its ranking the data support.** Besides certified comparisons and rank sets, the report gives certified top-k sets and tiers. A top-k statement certifies that a model is among the k best, or that it is not. Tiers $s = 1, 2, \ldots$ group the models by certified comparisons. A model in tier s has at least $s - 1$ models certified above it, so its true rank is s or worse (Theorems 4.3 and 4.4; Section 5, E2).
3. **Evaluation compute is saved.** The evaluation of a model can stop as soon as the question asked about it is answered, for example whether it is in the top three. The guarantee survives even though the decision to stop depends on the observed scores (Proposition 4.5; Section 5, E3).
4. **The price of these guarantees is small.** From the same pairwise evidence, the method certifies every comparison that a Bonferroni correction over the pairs certifies, and its exact version can certify more (Section 4.1). At a pre-planned look at halfway, it certifies about as many comparisons as the strongest fixed-sample procedure (Section 5, E4).

2 PROBLEM SETTING

Models, abilities, ranks. Each of $M \geq 2$ models has one scalar ability, its mean benchmark score, and its rank is one plus the number of models with strictly larger ability. Write $[M] = \{1, \dots, M\}$ for the set of models and $\theta = (\theta_1, \dots, \theta_M) \in \mathbb{R}^M$ for the vector of abilities, where $\mathbb{R}$ is the set of real numbers. Write $\#A$, or $|A|$, for the number of elements of a finite set A. The *true rank* of model j is

$$R_j(\theta) = 1 + \#\{l \in [M] : \theta_l > \theta_j\}, \tag{1}$$

abbreviated R_j when θ is clear from context. Rank 1 is best, and tied models share a rank. Two models tied for first both have rank 1. Let $\mathcal{P} = \{(j, l) \in [M]^2 : j \neq l\}$ be the ordered pairs. The *null hypothesis* for $(j, l) \in \mathcal{P}$ is $H_{jl} : \theta_j \leq \theta_l$, that j is not better than l. Rejecting H_{jl} certifies $\theta_j > \theta_l$, written $j \succ l$.

Data. All models are scored on the same items, and we assume nothing about how their scores are correlated. Item i yields the score vector $x_i = (x_{i1}, \dots, x_{iM}) \in [0, 1]^M$, one score per model. The most common case is binary correctness, where $x_{ij} = 1$ if model j answers item i correctly and $x_{ij} = 0$ otherwise. For binary scores, θ_j is the model's accuracy. More generally, scores may take any value in $[0, 1]$.

Two sampling models. Our main setting is a fixed benchmark evaluated in random order, and the guarantee holds equally when the items are a sample from a population. Time $t = 0, 1, 2, \dots$ counts the items revealed so far, and $X_t = (X_{t1}, \dots, X_{tM})$ is the score vector of the tth item revealed.

(F) Finite benchmark. The benchmark consists of N items and is a fixed matrix $x \in [0, 1]^{N \times M}$, where x_{ij} is the score of model j on item i. The ability $\theta_j = N^{-1} \sum_{i=1}^{N} x_{ij}$ is the mean score of model j on the whole benchmark. One permutation Π of $[N] = \{1, \dots, N\}$ is drawn uniformly at random and shared by all models. At time $t \leq N$, item $\Pi(t)$ is revealed. Its score vector is $X_t = x_{\Pi(t)}$, the row $\Pi(t)$ of x. The only randomness in the data is the order of evaluation.

(S) Superpopulation. $X_1, X_2, \dots$ are independent and identically distributed (i.i.d.) draws from a distribution on $[0, 1]^M$, and $\theta_j = \mathbb{E}[X_{1j}]$, where $\mathbb{E}$ denotes expectation. The joint law of X_1 across models is unrestricted.

Under both sampling models, $\mathcal{F}_t$ denotes the information available after t items. Formally, write $\sigma(\cdot)$ for the σ-algebra generated by the listed random variables. Then $\mathcal{F}_t = \sigma(\Pi(1), \dots, \Pi(t), \xi)$ under (F) and $\mathcal{F}_t = \sigma(X_1, \dots, X_t, \xi)$ under (S). Here ξ is an auxiliary random variable that lets the analyst randomize decisions. It is independent of Π under (F) and of $(X_1, X_2, \dots)$ under (S). The increasing family $(\mathcal{F}_t)_{t \geq 0}$ is called the filtration. A quantity is *adapted* if its value at time t is determined by $\mathcal{F}_t$. Equivalently, it can be computed from what has been observed by time t.

Goal. We want rank sets that cover the true ranks of all models at once, at every time, not at one pre-chosen sample size. Fix a level $\alpha \in (0, 1)$ and write $\mathbb{P}_\theta$ for probability under a data-generating law whose ability vector is θ. A *rank confidence sequence* at level α is a family of adapted random sets $\mathcal{R}_{j,t} \subseteq [M]$, one per model j and time t, with $\mathbb{P}_\theta(\forall t \geq 0,\ \forall j \in [M] :\ R_j(\theta) \in \mathcal{R}_{j,t}) \geq 1 - \alpha$ for every data-generating law.

3 RANK CONFIDENCE SEQUENCES

Overview and guarantee. After every item, the method keeps the orderings of the models, ties allowed, that the data have not yet contradicted. It reports only statements on which all surviving orderings agree. It repeats three steps after each item $t = 1, 2, \dots$. Step 1 *updates* a nonnegative measure of evidence for each ordered pair of models. Step 2 *eliminates* every ordering against which enough evidence has accumulated. Step 3 *reports* the certified dominances $j \succ l$, rank sets, top-k sets and tiers on which the remaining orderings agree. An optional Step 4 *retires* models whose evaluation can stop. With probability at least $1 - \alpha$, the true ordering, the one given by the true abilities, is never eliminated. Every report is agreed on by all surviving orderings, the true one included, so with the same probability every statement in every report is true at all times simultaneously. The user may therefore look after every item and stop by any rule.

Inputs. The method needs the scores $X_{tj} \in [0,1]$ of every model $j \in [M]$ on a common sequence of items $t = 1, 2, \ldots$, and two choices made before any data are seen: the level $\alpha \in (0,1)$ and a finite grid $\Lambda \subset (0,1)$ of betting fractions (Step 1), by default $\Lambda = \{0.03, 0.06, 0.12, 0.25, 0.5\}$. The grid affects how fast evidence accumulates and never the guarantee. Under (F), a benchmark of N items, the user also draws one uniformly random order of the items from a recorded seed before the evaluation, and evaluates every model in that order. Then $X_t = (X_{t1}, \ldots, X_{tM})$ is the score vector of the tth item in that order. Under (S), the item stream is open-ended and i.i.d., so we set $N = \infty$.

Step 1: update the pairwise wealths. For each ordered pair $(j,l) \in \mathcal{P}$ the method maintains a *wealth* $E_t^{jl} \geq 0$. After t items it measures the evidence that model j is better than model l, or equivalently the evidence against $H_{jl} : \theta_j \leq \theta_l$. It is the average capital of $|\Lambda|$ gamblers, each starting with one unit and, on every item, staking a fixed fraction $\lambda \in \Lambda$ of their capital on model j outscoring model l. Formally, for each $\lambda \in \Lambda$ put $E_{\lambda,0}^{jl} = 1$ and, on item t with paired difference $Z_t^{jl} = X_{tj} - X_{tl}$ and offset b_t^{jl} as given below,

$$E_{\lambda,t}^{jl} = E_{\lambda,t-1}^{jl}\Big(1 + \lambda\,\frac{Z_t^{jl} - b_t^{jl}}{1 + b_t^{jl}}\Big), \qquad E_t^{jl} = \frac{1}{|\Lambda|}\sum_{\lambda\in\Lambda} E_{\lambda,t}^{jl}, \tag{2}$$

where the *offset* b_t^{jl}, computed from items $1, \ldots, t-1$ only, bounds the conditional mean of Z_t^{jl} under H_{jl}:

$$b_t^{jl} = \begin{cases} 0 & \text{under (S)}, \\ \max\Big\{-0.99,\ \min\Big\{1,\ \dfrac{-S_{t-1}^{jl}}{N-t+1}\Big\}\Big\} & \text{under (F)}, \end{cases} \qquad S_{t-1}^{jl} = \sum_{s<t} Z_s^{jl}. \tag{3}$$

Here S_{t-1}^{jl} is the cumulative paired difference over the first $t-1$ items. Under H_{jl}, $\mathbb{E}[Z_t^{jl} \mid \mathcal{F}_{t-1}] \leq b_t^{jl}$. For i.i.d. items this follows from $\mathbb{E}[Z_t^{jl}] = \theta_j - \theta_l \leq 0$. On a finite benchmark the next item is drawn uniformly from the $N - t + 1$ items not yet revealed, whose differences sum to $N(\theta_j - \theta_l) - S_{t-1}^{jl} \leq -S_{t-1}^{jl}$. The truncation to $[-0.99, 1]$ keeps each factor in (2) nonnegative and well defined. Thus, under H_{jl}, each factor has conditional expectation at most one, so the bets are fair or unfavorable and the wealth cannot grow in expectation. By Ville's inequality (Ville, 1939), its probability of ever reaching $1/\alpha$ is at most α. Such a wealth is an *e-process* (Ramdas et al., 2023), a nonnegative process whose expected value stays at most one under any stopping rule. Its reciprocal acts as a p-value that remains valid when it is checked after every item. When $\theta_j > \theta_l$, the wealth grows exponentially if, for some $\lambda \in \Lambda$, the factor in (2) has a positive expected logarithm. The best bet size depends on the unknown gap $\theta_j - \theta_l$, and averaging over a grid avoids relying on a single one. The pairs (j,l) and (l,j) have separate wealths.

Step 2: eliminate orderings and certify dominances. An ordering of the models is eliminated once the wealths that should stay small if that ordering were true average to $1/\alpha$ or more. The statement "j is better than l" is certified once every ordering that says otherwise has been eliminated. This is the principle of *closed testing* (Marcus et al., 1976), with one test per ordering. A *weak order* W on $[M]$ is a ranking of the models in which ties are allowed. Write $j \prec_W l$ if W ranks l strictly above j, and $j \sim_W l$ if it ties them. Write $j \preceq_W l$ when W ranks l at least as high as j, including a tie. Formally $\preceq_W$ is a reflexive, transitive and total relation on $[M]$. We represent W by a vector of levels $(v_1, \ldots, v_M)$, with $j \preceq_W l$ if and only if $v_j \leq v_l$. Let $\mathcal{W}$ be the set of all weak orders on $[M]$. It has 13 elements for $M = 3$ and 545,835 for $M = 8$. The abilities determine one of them, the *true ordering* $W(\theta)$. It satisfies $j \preceq_{W(\theta)} l$ if and only if $\theta_j \leq \theta_l$. The rank (1) depends on θ only through the true ordering. Under W the rank of model j is $R_j(W) = 1 + \#\{l \in [M] : j \prec_W l\}$. If W were the true ordering, the hypotheses H_{jl} that hold would be exactly those of the pairs in $T(W) = \{(j,l) \in \mathcal{P} :\ j \preceq_W l\}$. By Step 1, none of the wealths of these pairs would be likely to become large. The evidence against W is therefore the average of these wealths. Let E_t^W denote that average. We eliminate W permanently when E_t^W reaches $1/\alpha$, and let $\mathcal{C}_t$ be the set of surviving orderings:

$$E_t^W = \frac{1}{|T(W)|}\sum_{(j,l)\in T(W)} E_t^{jl}, \qquad \mathcal{C}_t = \Big\{W \in \mathcal{W} :\ \max_{s\leq t} E_s^W < \frac{1}{\alpha}\Big\}. \tag{4}$$

An average of e-processes on the same items is again an e-process, whatever their dependence, as for e-values (Vovk & Wang, 2021). The set $\mathcal{C}_t$ of surviving orderings is therefore a confidence set for the true ordering that is valid at all times at once, $\mathbb{P}_\theta(\forall t \geq 0 :\ W(\theta) \in \mathcal{C}_t) \geq 1 - \alpha$ (Theorem 4.2). The certified dominances and the rank sets are what all surviving orderings agree on,

$$\mathcal{D}_t = \big\{(j,l) \in \mathcal{P} :\ l \prec_W j \text{ for all } W \in \mathcal{C}_t\big\}, \qquad \mathcal{R}_{j,t} = \big\{R_j(W) :\ W \in \mathcal{C}_t\big\}. \tag{5}$$

Because elimination is permanent, $\mathcal{C}_t$ can only shrink and $\mathcal{D}_t$ can only grow as t increases.

Computing Step 2. Step 2 can be computed in three ways. Algorithm 1 enumerates the surviving weak orders. It alone gives the exact rank sets and is practical up to about $M = 8$. Algorithm 2 is a polynomial-time shortcut for any M that may certify fewer comparisons. Algorithm 3 computes the same $\mathcal{D}_t$ by integer programming at larger M. It solved every instance in our experiments. The shortcut certifies (j, l) once the pooled statistic $B_t^{jl} = E_t^{jl} + \sum_{m \in [M] \setminus \{j,l\}} \min\{E_t^{jm}, E_t^{ml}\}$ reaches $M(M-1)/\alpha$. Each term $\min\{E_t^{jm}, E_t^{ml}\}$ is indirect evidence that $\theta_j > \theta_l$ through a third model m. It is large only if $\theta_j > \theta_m$ and $\theta_m > \theta_l$ are both well supported. The shortcut certifies a subset of $\mathcal{D}_t$ and at least what a Bonferroni correction over the $M(M-1)$ pairs certifies, namely every pair whose wealth has reached $M(M-1)/\alpha$, a rule we call ***e-Bonferroni***. The *exact test*, computed by Algorithm 1 or 3, can certify more, because testing whole orderings uses the logical links among the pairwise hypotheses. For example, H_{jl} and H_{lm} together imply H_{jm}, so combinations of true and false hypotheses that no ordering allows need not be guarded against.

Step 3: report. Except for the exact rank sets obtained by enumeration, every reported quantity is a function of the certified set $\mathcal{D}_t$, so all of them are correct whenever $\mathcal{D}_t$ contains no false dominance. The subscript t is dropped below when the time is fixed.

Dominances. Each $(j, l) \in \mathcal{D}_t$ is reported as "j is better than l", written $j \succ l$.

Ranks. The rank interval of model j is $[L_{j,t}, U_{j,t}]$ with

$$L_{j,t} = 1 + \#\{l : (l,j) \in \mathcal{D}_t\}, \qquad U_{j,t} = M - \#\{l : (j,l) \in \mathcal{D}_t\}. \tag{6}$$

Enumeration can instead report the *exact rank set* $\mathcal{R}_{j,t} \subseteq [L_{j,t}, U_{j,t}]$, which may be smaller.

Top-k sets. For $k \in [M]$, model j is certified inside the top k if $U_j \leq k$ and outside it if $L_j > k$. Ties can put more than k models inside.

Tiers. A *chain above* j of length h is a sequence of models $l_1, \ldots, l_h$ such that (l_1, l_2), …, (l_{h-1}, l_h) and (l_h, j) all belong to $\mathcal{D}$. Let $h(j)$ be the greatest such length, with $h(j) = 0$ if no model is certified better than j. The *tier* of j is $\mathrm{tier}(j) = 1 + h(j)$, and tier s is the set $\{j : \mathrm{tier}(j) = s\}$. Tiers are computed by the recursion $\mathrm{tier}(j) = 1 + \max\{\mathrm{tier}(l) : (l,j) \in \mathcal{D}\}$, with $\mathrm{tier}(j) = 1$ if the set is empty. A model in tier s has true rank s or worse.

If $\mathcal{D}$ ever contains a cycle, such as $j \succ l$ and $l \succ j$, or no ordering survives, an error has occurred, which happens with probability at most α. Report it as such.

Step 4 (optional): retire models. A model may be dropped from further evaluation at any time, by any rule that uses only the scores observed so far, and the guarantee still holds. Let $A_t \subseteq [M]$ be the set of models scored on item t, with $A_{t+1} \subseteq A_t$, chosen from the scores observed on items $1, \ldots, t-1$. The wealth of (j, l) is updated by (2) while both j and l belong to A_t. Afterwards it is frozen, $E_t^{jl} = E_{t-1}^{jl}$, and continues to be used in Steps 2 and 3. Natural rules retire j once its top-k status is certified, once every pair involving j is certified in one direction, or once $U_j - L_j$ is below a target. The item order is not affected by retirement, and a retired model is not brought back.

Scope and cost. The model set must be fixed in advance, items must not be chosen adaptively, and every model that has not been retired must have a score on each item. Pairwise-battle data are out of scope. The per-item update is negligible next to running the models. It costs $O(M^2)$ per item, 25 μs for 20 models on one CPU core. Appendix A states Algorithms 1–3 with implementation details and costs. Appendix D gives the limits of the guarantee.

4 What can be established, and why it matters

The guarantee of Section 3 follows from three results, one per subsection, all proved in Appendix C.

4.1 ALL CERTIFIED DOMINANCES ARE CORRECT, AT ALL TIMES

What needs a guarantee. Every report except the exact rank sets is computed from the certified set $\mathcal{D}_t$, so the central requirement is that, with probability at least $1-\alpha$, no $\mathcal{D}_t$ at any time t contains a false dominance, a pair (j,l) with $\theta_j \le \theta_l$. A pair with $\theta_j > \theta_l$ is a true dominance. We can guarantee this over all pairs and times at once, even though the wealths share items and may be dependent in unknown ways. The guarantee starts from one property of the pairwise wealths, that of an e-process (Ramdas et al., 2023). Identify a hypothesis H with the set of data-generating laws it allows and let $\mathbb{E}_P$ denote expectation under the law P. A *stopping time* is a random time τ such that whether $\tau = t$ can be decided from the information $\mathcal{F}_t$ available at time t. Formally, an *e-process* for H is a nonnegative adapted process $(E_t)_{t\ge 0}$ with $\mathbb{E}_P[E_\tau] \le 1$ for every $P \in H$ and every bounded stopping time τ. An e-process satisfies Ville's inequality: $P(\exists t \ge 0 :\ E_t \ge 1/\alpha) \le \alpha$ for every $P \in H$, where $P(\cdot)$ is probability under the law P. This is the counterpart of Markov's inequality that holds over all times at once. Under both sampling models every wealth of Section 3 is an e-process for its hypothesis H_{jl}, whatever the dependence between models. This is the betting construction of Waudby-Smith & Ramdas (2024). Because all models share one item order, all wealths are e-processes for the *same* filtration. We record this as an assumption.

Assumption 4.1. There is a filtration $(\mathcal{F}_t)$ and, for each $(j,l) \in \mathcal{P}$, a nonnegative adapted process (E_t^{jl}) with $E_0^{jl} = 1$ that is an e-process for H_{jl} with respect to $(\mathcal{F}_t)$ under every data-generating law for which $\theta_j \le \theta_l$. No condition is placed on the joint behavior of the $M(M-1)$ processes.

Result. With probability at least $1-\alpha$ the true ordering is never eliminated, and as long as it survives no certified dominance is false. Write $\theta^\star$ for the true ability vector and $W^\star = W(\theta^\star)$ for the true ordering. Here $\mathcal{C}_t$ and $\mathcal{D}_t$ are the surviving and certified sets of (4) and (5).

Theorem 4.2 (Validity)**.** *Under Assumption 4.1, for every data-generating law with parameter $\theta^\star$ and $W^\star = W(\theta^\star)$, and with $\mathbb{P}$ denoting probability under that law,*

$$\mathbb{P}\big(\forall t \ge 0 :\ W^\star \in \mathcal{C}_t\big) \ge 1-\alpha. \tag{7}$$

Consequently the family-wise error rate, the probability of certifying any false dominance, is controlled uniformly over time,

$$\mathbb{P}\big(\exists t \ge 0,\ \exists (j,l) \in \mathcal{D}_t :\ \theta_j^\star \le \theta_l^\star\big) \le \alpha. \tag{8}$$

Both statements also hold when t is replaced by any $\mathbb{N}_0$-valued random time, stopping or otherwise, where $\mathbb{N}_0 = \{0,1,2,\dots\}$.

The proof applies Ville's inequality to $E_t^{W^\star}$, an average of wealths of true hypotheses. While $W^\star$ survives, the certified set is a strict partial order (no pair in both directions, closed under transitivity), so the report never contradicts itself.

Relation to existing results. Theorem 4.2 combines closed testing over weak orders with averaging of e-processes. We use the partitioning principle of Finner & Strassburger (2002), where the orderings partition the parameter space and one test is carried out per cell, as in the fixed-sample rank intervals of Al Mohamad et al. (2021). Shaffer (1986) shows that only sets of hypotheses that can be true together need be tested. Here these are exactly the sets $T(W)$, one per weak order. Averaging e-values (nonnegative statistics with expectation at most one under the null), the fixed-time counterpart of e-processes, remains valid under arbitrary dependence (Vovk & Wang, 2021; Wang, 2025), while products generally require independence, which shared items rule out. Our result differs from the two closest ones as follows. Arnold et al. (2026) average e-processes and close over all subsets of models, which certifies the set of models that may be best and no ranks. Hartog & Lei (2025) close over all sets of hypotheses, the full closure, which controls the family-wise error for arbitrary hypotheses but yields no rank statements. We test one averaged e-process per weak order, which yields rank sets. Nothing is lost relative to standard corrections. On the same wealths, the exact test certifies every pair that the full closure certifies, and the shortcut every e-Bonferroni rejection used by Gu et al. (2026). Appendix B.3 states both results.

4.2 RANKS, TOP-k SETS AND TIERS INHERIT THE GUARANTEE

What needs a guarantee. Users read ranks, top-k sets and tiers, which are built from the dominances. These reported quantities must also be correct without costing additional error. To cover

the exact test, the shortcut and e-Bonferroni at once, call a family $(\mathcal{D}_t)_{t\geq 0}$ of random subsets of $\mathcal{P}$ a *certified set* if its *good event* $\mathcal{G} = \{\forall t \geq 0,\ \forall (j,l) \in \mathcal{D}_t :\ \theta_j^\star > \theta_l^\star\}$ has probability at least $1-\alpha$. All three are certified sets. Taking the transitive closure adds no error. Let $\overline{\mathcal{D}}_t$ be the transitive closure of $\mathcal{D}_t$, which adds (j,m) whenever (j,l) and (l,m) are present, repeatedly. On $\mathcal{G}$, the transitive closure contains only true dominances. True dominances form a strict partial order, so the closure is certified on the same event.

Result. Rank intervals, top-k sets and tiers are correct on $\mathcal{G}$, and exact rank sets while $W^\star$ survives.

Theorem 4.3 (Rank-set duality)**.** *(a)* Projection. *Under Assumption 4.1, the rank sets $\mathcal{R}_{j,t}$ of* (5) *satisfy $\mathbb{P}(\forall t, \forall j :\ R_j(\theta^\star) \in \mathcal{R}_{j,t}) \geq 1-\alpha$. (b)* Counting. *Let $(\mathcal{D}_t)$ be any certified set and compute $L_{j,t}$ and $U_{j,t}$ from it by* (6)*. On $\mathcal{G}$, $L_{j,t} \leq R_j(\theta^\star) \leq U_{j,t}$ for all j and all t, and the intervals only improve if $\mathcal{D}_t$ is replaced by $\overline{\mathcal{D}}_t$. (c)* Comparison. *For $\mathcal{D}_t$ from* (5)*, $\mathcal{R}_{j,t} \subseteq [L_{j,t}, U_{j,t}]$, and the inclusion is strict for some wealth configurations.*

By part (b), on $\mathcal{G}$ every top-k certificate of Step 3 is correct, for all j, k and t simultaneously.

Theorem 4.4 (Tier guarantee)**.** *Let $(\mathcal{D}_t)$ be a certified set and let $\mathrm{tier}_t(j)$ be the tier of model j computed from $\overline{\mathcal{D}}_t$ as in Step 3 of Section 3. On $\mathcal{G}$, simultaneously for all j and all t,*

$$R_j(\theta^\star) \ \geq\ L_{j,t} \ \geq\ \mathrm{tier}_t(j), \tag{9}$$

where $L_{j,t}$ is computed from $\overline{\mathcal{D}}_t$, and the number of tiers is at most the number of distinct values among $\theta_1^\star, \ldots, \theta_M^\star$. On $\mathcal{G}$, if $\mathcal{D}_t \subseteq \mathcal{D}_{t+1}$ then $\mathrm{tier}_t(j) \leq \mathrm{tier}_{t+1}(j)$, so a model's tier index never decreases as evidence accrues.

The guarantee is only (9). A lower tier index does not imply a certified pairwise dominance, and tier 1 also holds every model with no certified comparison.

Relation to existing results. The counting part of Theorem 4.3 comes from Al Mohamad et al. (2022) and Mogstad et al. (2024), and Gu et al. (2026) apply it at all times. They also define the same tiers without a guarantee, which Theorem 4.4 adds. Appendix B.4 gives more on tiers.

4.3 Evaluation can stop early, model by model

What needs a guarantee. Retiring a model, that is, dropping it from further evaluation, is data-dependent. Once a model stops, its future scores are not observed and every pair involving it has its wealth frozen. The guarantee must remain valid under both effects for any rule based only on observed scores. After retirement the analyst observes less information than $\mathcal{F}_t$, which contains the scores of all models on all revealed items.

Result. Freezing each wealth at the retirement of either of its models preserves Assumption 4.1, and with it every result that rests on that assumption.

Proposition 4.5 (Retirement preserves validity)**.** *Under* (F) *or* (S)*, let the item order in* (F) *be drawn before evaluation and remain unaffected by later decisions. At each time t let $A_t \subseteq [M]$ be the set of models scored on item t, chosen as a function of the scores observed on items $1, \ldots, t-1$ and of ξ, with $A_1 = [M]$, $A_{t+1} \subseteq A_t$, and $A_{N+1} = \emptyset$ under* (F)*. Model j is retired at $\sigma_j = \sup\{t : j \in A_t\}$ ($\sigma_j = \infty$ if never), so that item σ_j is included in its record, and retirement is permanent under both sampling models. Let $\mathbf{1}\{\cdot\}$ denote the indicator of the event in braces and $s \wedge t = \min\{s,t\}$, and let $\mathcal{F}_t^{\mathrm{obs}} = \sigma\big(\xi,\ \mathbf{1}\{j \in A_s\},\ \mathbf{1}\{j \in A_s\}X_{sj} :\ s \leq t,\ j \in [M]\big)$ be the observed-data filtration, generated by which models were scored and the scores observed. Call a bet* predictable *for a filtration if the bet on item t is determined by the information at time $t-1$. Then $\mathcal{F}_t^{\mathrm{obs}} \subseteq \mathcal{F}_t$, and each σ_j is an $(\mathcal{F}_t^{\mathrm{obs}})$-stopping time and hence an $(\mathcal{F}_t)$-stopping time. Replacing each E_t^{jl} by $E_{t\wedge\sigma_j\wedge\sigma_l}^{jl}$, with any $(\mathcal{F}_t^{\mathrm{obs}})$-predictable bets, preserves Assumption 4.1 and every result resting on it.*

Relation to existing results. Retirement times and bets chosen from the observed scores alone are also stopping times and predictable bets for $(\mathcal{F}_t)$. All frozen wealths therefore remain e-processes for one common filtration, while the procedure never needs an unobserved score. Freezing at such

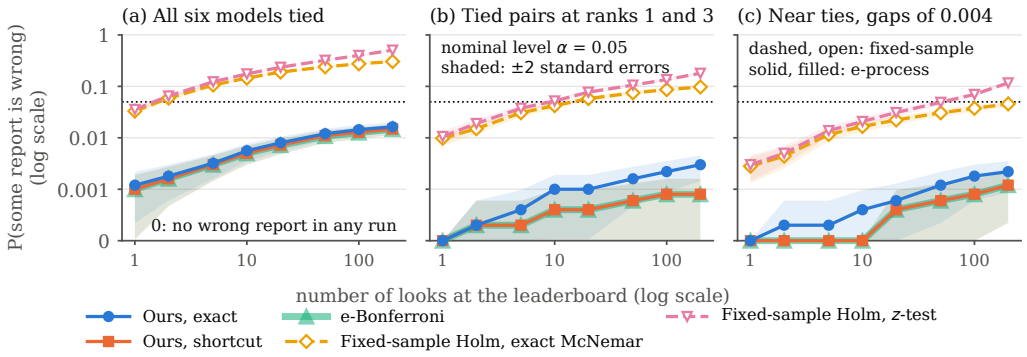

Figure 1: Probability that some report is wrong against the number of looks, $\alpha = 0.05$. Most fixed-sample curves (dashed) exceed α as looks accumulate. E-process curves (solid) stay below α. Both axes use a log scale: tick labels are actual values a factor of ten apart, except 0.

times matters. Tavyrikov et al. (2026) show that averaging running maxima, which freezes each e-process at its best past value, need not control the family-wise error. Existing sequential evaluation methods stop evaluating one model (Zhou et al., 2026; Hsu & Shekhar, 2026) or one pair at a time (Kotawala, 2026; Gu et al., 2026; Arviv et al., 2026). Ours stops models one by one inside a simultaneous guarantee over all ranks. Appendix D says when a frozen pair may be updated again.

5 Experiments

Four experiments, E1 to E4, test monitoring, certification on real data, compute savings and the price of anytime validity. The real-data experiments use per-item results of 395 large language models (LLMs) from the Open LLM Leaderboard (Maia Polo et al., 2024; Beeching et al., 2023). Setups are in Appendix E, with further experiments on cross-model dependence and on what the multiplicity correction and the bet contribute.

E1. Only fixed-sample rank sets break under monitoring. With six tied models checked at 200 looks and nominal level 5%, fixed-sample rank sets err in 30% to 51% of runs and our exact test in 1.6% (Figure 1a). The models answer 2000 simulated i.i.d. items, with equally spaced looks. The fixed-sample rank sets are the standard construction, Holm's correction (Holm, 1979) over paired tests of all $M(M-1)$ ordered pairs, converted into rank intervals by (6). They err in 30% of runs with the exact McNemar test and 51% with the paired z-test. With tied pairs at ranks 1 and 3 the rates are 9.7%, 18% and ours 0.3%. With no exact ties but gaps of 0.004, they are 4.6%, 11.7% and 0.2%. At a single pre-planned look both fixed-sample procedures are valid, at 3.3% and 3.6% with all six tied, so the inflation is caused by monitoring alone and not by the normal approximation.

E2. The data support only part of the ranking early. Figure 2 follows the twenty highest-scoring models across six benchmarks, while Table 1 summarizes the halfway and final results. On MMLU, halfway through the benchmark, 50% of the 190 true dominances, the pairs whose order on the full benchmark is strict, are certified and the twenty models form 3.0 certified tiers. HellaSwag is similar, while the other four benchmarks separate fewer pairs at halfway. The difficulty comes from choosing twenty models near the top. When twenty models are spread across the leaderboard, 91% of their true dominances on MMLU are already certified halfway through. We also used the top eight models. Of all 900 runs, one had a wrong report, far below the 5% level.

E3. Model-by-model stopping saves evaluation. The top twenty are chosen with the full benchmark in hand, which an operator cannot do, so we repeat on all 395 released models with no further selection. Most are far from the top-3 boundary, so their evaluation can stop early. Certifying top-3 membership for all 395 uses 6.2% of the full MMLU evaluation, against 31% for the twenty highest-scoring alone. In absolute terms the 395-model run costs about four times as much. In simulation, retiring each model once its own top-3 question is answered costs 22% against 84% for stopping

Table 1: True dominances certified, as a percentage of those that exist, and mean tiers at halfway. Ours checked at every 1%, 50 item orders. No report was wrong in any of these runs.

	Top twenty (179–190 dominances)			All 395 models (≈ 77,700)		
Benchmark	half	end	tiers	half	end	tiers
MMLU	50%	99%	3.0	86%	99.6%	21.2
HellaSwag	49%	96%	3.2	77%	99.2%	20.4
GSM8K	22%	96%	2.9	78%	98.9%	10.8
WinoGrande	2%	85%	2.0	26%	95.5%	4.2
ARC-Challenge	0%	82%	1.1	48%	97.0%	6.8
TruthfulQA	17%	82%	2.0	55%	97.0%	6.1

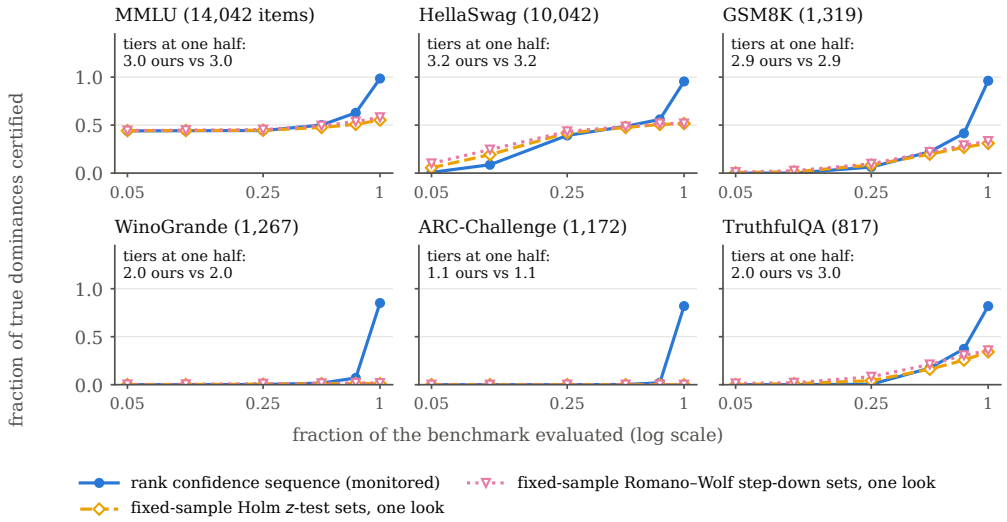

Figure 2: Twenty highest-scoring models, 50 item orders: fraction of true dominances certified against fraction evaluated. Ours (Algorithm 2) checked at every 1%, Holm and Romano–Wolf at one look. Each panel gives the mean tiers at halfway, ours against Romano–Wolf.

every model at the same moment, and the cost barely depends on k. The full leaderboard is also better resolved, with 86% of the true dominances on MMLU certified halfway and 99.6% by the end (Table 1). No report was wrong in any of these runs.

E4. Anytime validity costs little at a pre-planned look. At a pre-planned look at halfway, ours matches or exceeds the strongest fixed-sample procedure, the bootstrap step-down rank sets of Mogstad et al. (2024), based on Romano & Wolf (2005), on five of the six benchmarks and trails by four percentage points on TruthfulQA (Figure 2). Beyond halfway the offset of (F), which uses the finiteness of the benchmark, increasingly favors ours, so the later separation should not be read as a pure power comparison.

6 Limitations and Conclusion

Limitations. At an early pre-planned look, fixed-sample procedures certify more. Models arriving over time (Fischer & Ramdas, 2024) and adaptive item selection (Zhou et al., 2026) are not covered. Validity covers the benchmark as given. Contamination and transfer to other tasks lie outside it.

Conclusion. Rank confidence sequences give shared-item leaderboards rank statements valid for all models at all times. In our experiments they stayed valid under monitoring where fixed-sample rank sets did not, matched or exceeded them at a pre-planned look at halfway on five of six benchmarks, and settled every model's top-3 status with 6.2% of a full MMLU evaluation.

AI USE STATEMENT

We used generative AI tools to draft and polish draft sections of the paper, assist in finding related work, and check and format references. We also used generative AI for code assistance, including debugging, editing, and polishing codes. We did not use generative AI to formulate mathematical claims, develop the proofs, design the methodology or experiments, implement the proposed method, generate synthetic data, or interpret results. All AI-assisted content was reviewed by the authors, who take responsibility for the final work.

REPRODUCIBILITY STATEMENT

Complete proofs of every stated result are in Appendix C, with the scope of the guarantee and the open problems in Appendix D. The method is specified in full in Section 3. The reference implementation, the scripts that run each experiment with their random seeds, and the result files behind every number in Section 5 and Appendix E are in the supplementary code archive. The real-data experiments use publicly released per-item results (Maia Polo et al., 2024). The archive includes a script that checks the downloaded data against a stored checksum and builds the score matrices. The archive is available for review at `https://anonymous.4open.science/r/rank-confidence-sequences-supplement`.

REFERENCES

Diaa Al Mohamad, Erik van Zwet, Aldo Solari, and Jelle Goeman. Simultaneous confidence intervals for ranks using the partitioning principle. *Electronic Journal of Statistics*, 15(1):2608–2646, 2021. doi: 10.1214/21-EJS1847.

Diaa Al Mohamad, Jelle J. Goeman, and Erik W. van Zwet. Simultaneous confidence intervals for ranks with application to ranking institutions. *Biometrics*, 78(1):238–247, 2022.

Noga Alon. Ranking tournaments. *SIAM Journal on Discrete Mathematics*, 20(1):137–142, 2006.

P. Armitage, C. K. McPherson, and B. C. Rowe. Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society. Series A (General)*, 132(2):235–244, 1969.

Sebastian Arnold, Georgios Gavrilopoulos, Benedikt Schulz, and Johanna Ziegel. Sequential model confidence sets. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2026. Advance article, doi:10.1093/jrsssb/qkag066; first posted as arXiv:2404.18678, 2024.

Ofir Arviv, Kristjan Greenewald, Yotam Perlitz, Hadar Mulian, Michal Shmueli-Scheuer, and Leshem Choshen. Stop guessing when to stop testing: Efficient model evaluation with just enough data. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 871–881. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.43.

Angel Rodrigo Avelar Menendez, Yufeng Liu, and Xiaowu Dai. Prompt-dependent ranking of large language models with uncertainty quantification. arXiv:2603.03336, 2026.

Edward Beeching, Clémentine Fourrier, Nathan Habib, Sheon Han, Nathan Lambert, Nazneen Rajani, Omar Sanseviero, Lewis Tunstall, and Thomas Wolf. Open LLM Leaderboard (2023–2024). Hugging Face, 2023. URL `https://huggingface.co/spaces/open-llm-leaderboard-old/open_llm_leaderboard`. Per-item results as released with tinyBenchmarks, snapshot of January 2024; accessed 17 September 2026.

Sam Bowyer, Laurence Aitchison, and Desi R. Ivanova. Position: Don't use the CLT in LLM evals with fewer than a few hundred datapoints. In *Proceedings of the 42nd International Conference on Machine Learning, Position Paper Track*, volume 267 of *Proceedings of Machine Learning Research*, pp. 81143–81184. PMLR, 2025. URL `https://openreview.net/forum?id=YhZ2PY2nZa`.

Onrina Chandra and Min-ge Xie. Finite-sample valid rank confidence sets for a broad class of statistical and machine learning models. arXiv:2512.00316, 2025.

Pierre Charbit, Stéphan Thomassé, and Anders Yeo. The minimum feedback arc set problem is NP-hard for tournaments. *Combinatorics, Probability and Computing*, 16(1):1–4, 2007.

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? Try ARC, the AI2 reasoning challenge. arXiv:1803.05457, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. arXiv:2110.14168, 2021.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in Banach spaces. In *Proceedings of the 31st Conference on Learning Theory*, volume 75 of *Proceedings of Machine Learning Research*, pp. 1493–1529, 2018.

Helmut Finner and Klaus Strassburger. The partitioning principle: A powerful tool in multiple decision theory. *Annals of Statistics*, 30(4):1194–1213, 2002.

Lasse Fischer and Aaditya Ramdas. Admissible online closed testing must employ e-values. arXiv:2407.15733, 2024.

Avijit Ghosh, Yifan Mai, Georgia Channing, and Leshem Choshen. AI evals are becoming the new compute bottleneck. EvalEval Coalition Blog, April 29, 2026. URL `https://evalevalai.com/research/2026/04/29/eval-costs-bottleneck/`. Accessed 22 August 2026.

Runzhe Gu, Wenguang Sun, Bowen Gang, and Xintao Xia. Anytime-valid inference for online ranking of large language models. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*. PMLR, 2026. URL `https://openreview.net/forum?id=7Y8xRnGQ47`.

Will Hartog and Lihua Lei. Family-wise error rate control with e-values. arXiv:2501.09015, 2025.

Reinhard Heckel, Nihar B. Shah, Kannan Ramchandran, and Martin J. Wainwright. Active ranking from pairwise comparisons and when parametric assumptions do not help. *Annals of Statistics*, 47(6):3099–3126, 2019.

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

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

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

Chia-Yu Hsu and Shubhanshu Shekhar. Efficient sequential evaluation of large language models. arXiv:2607.17409, 2026.

Jason C. Hsu. *Multiple Comparisons: Theory and Methods*. Chapman & Hall, 1996.

Qi Huangfu and J. A. Julian Hall. Parallelizing the dual revised simplex method. *Mathematical Programming Computation*, 10(1):119–142, 2018.

Anany Kotawala. Resolution diagnostics for paired LLM evaluation. arXiv:2605.30315, 2026.

Jiachun Li, David Simchi-Levi, and Will Wei Sun. Low rank for rank: Uncertainty-aware task-specific LLM ranking under sparse pairwise comparisons. arXiv:2605.29395, 2026.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3214–3252, 2022.

Felipe Maia Polo, Lucas Weber, Leshem Choshen, Yuekai Sun, Gongjun Xu, and Mikhail Yurochkin. tinyBenchmarks: Evaluating LLMs with fewer examples. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 34303–34326. PMLR, 2024.

Ruth Marcus, Eric Peritz, and K. R. Gabriel. On closed testing procedures with special reference to ordered analysis of variance. *Biometrika*, 63(3):655–660, 1976.

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

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

Leon Mirsky. A dual of Dilworth's decomposition theorem. *American Mathematical Monthly*, 78 (8):876–877, 1971.

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. *Review of Economic Studies*, 91(1):476–518, 2024.

Bitya Neuhof and Yuval Benjamini. Confident feature ranking. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1468–1476. PMLR, 2024.

Bitya Neuhof and Yuval Benjamini. Quantifying ranking uncertainty in LLM benchmarks. arXiv:2607.16259, 2026a.

Bitya Neuhof and Yuval Benjamini. Rank intervals for leaderboards: A hierarchical framework for model evaluation. arXiv:2606.08679, 2026b.

Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977. doi: 10.1093/biomet/64.2.191.

Aaditya Ramdas, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023.

Joseph P. Romano and Michael Wolf. Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469):94–108, 2005.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 8732–8740, 2020.

Juliet Popper Shaffer. Modified sequentially rejective multiple test procedures. *Journal of the American Statistical Association*, 81(395):826–831, 1986.

Shivalika Singh, Yiyang Nan, Alex Wang, Daniel D'souza, Sayash Kapoor, Ahmet Üstün, Sanmi Koyejo, Yuntian Deng, Shayne Longpre, Noah A. Smith, Beyza Ermis, Marzieh Fadaee, and Sara Hooker. The leaderboard illusion. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, volume 38, 2025. URL `https://openreview.net/forum?id=4Ae8edNqm0`.

Yury Tavyrikov, Jelle J. Goeman, and Rianne de Heide. Carefree multiple testing with e-processes. *Electronic Journal of Statistics*, 20(2):3178–3189, 2026.

Jean Ville. *Étude critique de la notion de collectif*. Gauthier-Villars, 1939.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, et al. SciPy 1.0: Fundamental algorithms for scientific computing in Python. *Nature Methods*, 17(3):261–272, 2020.

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

Ruodu Wang. The only admissible way of merging arbitrary e-values. *Biometrika*, 112(2):asaf020, 2025.

Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, pp. 20204–20214, 2020.

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

Peter H. Westfall. Multiple testing of general contrasts using logical constraints and correlations. *Journal of the American Statistical Association*, 92(437):299–306, 1997.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4791–4800, 2019.

Zhijian Zhou, Zesheng Ye, Zhaorun Chen, Bo Li, and Feng Liu. CELEUS: Certifiable and efficient LLM evaluation via e-processes. arXiv:2606.20820, 2026.

APPENDIX CONTENTS

A Method details **15**

B Supporting theory **17**

B.1 Validity and growth of each pairwise wealth 17

B.2 Supplement to the validity theorem 19

B.3 Testing only orderings loses nothing, and the shortcut is valid 19

B.4 Tiers and retirement 20

C Proofs **21**

C.1 Pairwise wealths 21

C.2 Validity, structure and power 24

C.3 Closure, ranks, tiers and retirement 26

C.4 Integer programming and hardness of exact certification 29

D Scope of the guarantee and open problems **30**

E Experiments **31**

E.1 Does monitoring break a fixed-sample rank set, and does ours survive it? 32

E.2 How much of a real leaderboard can be certified? 33

E.3 How much compute does retirement save? 34

E.4 What does anytime validity cost? 35

E.5 Does the guarantee depend on how model errors are correlated? 35

E.6 What does the multiplicity correction contribute? 36

F Related work in detail **37**

A METHOD DETAILS

Step 2 of Section 3 can be carried out three ways, and this appendix gives all three. Enumeration checks every ordering directly and is practical while the number of orderings is manageable. Integer programming computes the same answer at any size. A closed-form shortcut is fast but may certify less. All three are run after every update. If they are run less often, the running maximum in (4) is taken over those runs only, so no more orderings are eliminated. The certified set is then contained in $\mathcal{D}_t$, each rank set contains $\mathcal{R}_{j,t}$, and the guarantee of Theorem 4.2 still holds. Algorithm 1 is the enumeration.

Algorithm 1 Exact certification by enumeration (small M). Run after every update.

Require: wealths E_t^{jl}; level α; the set $\mathcal{C}$ of surviving weak orders (initially $\mathcal{W}$)

for each weak order $W \in \mathcal{C}$ **do**
  $e \leftarrow \frac{1}{|T(W)|} \sum_{(j,l)\in T(W)} E_t^{jl}$ ▷ average wealth over the pairs with $j \preceq_W l$
  if $e \geq 1/\alpha$ **then** remove W from $\mathcal{C}$ ▷ permanent
  end if
end for
$\mathcal{D} \leftarrow \{(j,l) \in \mathcal{P} : \ l \prec_W j \text{ for every } W \in \mathcal{C}\}$ ▷ certified: j better than l
$\mathcal{R}_j \leftarrow \{1 + \#\{l : j \prec_W l\} : \ W \in \mathcal{C}\}$ for each j ▷ rank set of model j
return $\mathcal{C}, \mathcal{D}, (\mathcal{R}_j)_j$

Beyond about eight models enumeration is impractical. The closed-form rule of Algorithm 2 then certifies a subset of $\mathcal{D}_t$ at $O(M)$ operations per pair, and Algorithm 3 below computes $\mathcal{D}_t$ exactly. Transitivity supplies the second term of the shortcut's statistic. Take any ordering that puts l at least as high as j, and any third model m. Either the ordering puts m at least as high as j, and then the wealth E^{jm} counts against it, or it puts m below j and hence below l, and then E^{ml} counts against it. Either way the ordering is charged at least $\min\{E^{jm}, E^{ml}\}$.

Algorithm 2 Shortcut certification (any M). Run after every update.

Require: wealths E^{jl}; level α; the set $\mathcal{D}$ of certified pairs (initially empty)

for each ordered pair $(j,l) \notin \mathcal{D}$ **do**
  $B \leftarrow E^{jl} + \sum_{m \neq j,l} \min\{E^{jm},\, E^{ml}\}$
  if $B \geq M(M-1)/\alpha$ **then** add (j,l) to $\mathcal{D}$ ▷ permanent
  end if
end for
replace $\mathcal{D}$ by its transitive closure ▷ if $j \succ m$ and $m \succ l$ then $j \succ l$; same good event, no added error probability
return $\mathcal{D}$

Exact certification for any M. Algorithm 3 computes $\mathcal{D}_t$ without enumerating orderings, by asking an integer program whether any surviving ordering places l at least as high as j. A weak order W is encoded by $y \in \{0,1\}^{\mathcal{P}}$ with $y_{ab} = 1$ exactly when $(a,b) \in T(W)$. The $0/1$ vectors with $y_{ab} + y_{ba} \geq 1$ for all $a \neq b$, which is totality, and $y_{ab} + y_{bc} - y_{ac} \leq 1$ for all distinct a, b, c, which is transitivity, are exactly the weak orders. A *call time* is a time at which an algorithm is run, and a *wealth matrix* is the array $(E_s^{ab})_{(a,b)\in\mathcal{P}}$ of pairwise wealths at one call time. With a run after every update the call times up to t are $1, \ldots, t$. The order W survives call time s when $E_s^W < 1/\alpha$, equivalently when $g_s(y) := \sum_{(a,b)\in\mathcal{P}} y_{ab}(E_s^{ab} - 1/\alpha) < 0$. For a set S of call times let

$$z^\star(S) = \min\Big\{ z : \ z \geq g_s(y) \text{ for all } s \in S,\ y \in \{0,1\}^{\mathcal{P}} \text{ a weak order},\ y_{jl} = 1 \Big\}, \qquad (10)$$

an integer program with one continuous variable. Then $(j,l) \in \mathcal{D}_t$ exactly when $z^\star(\{1, \ldots, t\}) \geq 0$, and a minimizer with $z^\star < 0$ is a surviving ordering with $(j,l) \in T(W)$, a *witness* that $(j,l) \notin \mathcal{D}_t$. Survival needs the strict inequality $g_s(y) < 0$ and certification its complement $z^\star \geq 0$, so in exact arithmetic no tolerance is needed. Algorithm 3 keeps a pool of witnesses and drops one the first time its average wealth E_t^W reaches $1/\alpha$. A witness W *covers* every pair in $T(W)$, and (10) is

solved only for pairs that no witness in the pool covers. The program starts with few call-time constraints and adds them only when needed. This is safe because a maximum over fewer times is no larger, so $z^\star(S) \geq 0$ for any $S \subseteq \{1, \ldots, t\}$ already certifies the pair. Otherwise the minimizer is checked against every call time, and a time at which it fails is added to S. In exact arithmetic Algorithm 3 computes $\mathcal{D}_t$ (Proposition A.1). Under numerical optimization the implementation, which solves the programs with HiGHS (Huangfu & Hall, 2018) through SciPy (Virtanen et al., 2020), is conservative. It certifies a pair only when the solver's lower bound on the optimum exceeds a tolerance of $10^{-9}M(M-1)/\alpha$, relative to the scale of g_s. It accepts an ordering as a witness only when it survives every call time by that margin, and it leaves a numerically ambiguous pair unresolved until the next call. Hence within the solver's numerical guarantees it never certifies more than $\mathcal{D}_t$. No such unresolved case arose in any run reported here.

Algorithm 3 Exact certification by integer programming (any M). Run after every update.

Require: wealths E_s^{ab} at every call time $s \leq t$; level α; the certified set $\mathcal{D}$, the pool $\mathcal{Q}$ of witnesses (surviving weak orders found so far) and the set S of constraint times, all initially empty

remove from $\mathcal{Q}$ every W with $E_t^W \geq 1/\alpha$ ▷ rejected for good
for each $(j,l) \notin \mathcal{D}$ with $(j,l) \notin T(W)$ for all $W \in \mathcal{Q}$, smallest E_t^{jl} first **do**
  if $B_t^{jl} \geq M(M-1)/\alpha$ **then** add (j,l) to $\mathcal{D}$ and **continue** ▷ Proposition B.8
  end if
  loop
    solve (10) for $S \cup \{t\}$: optimum $z^\star$, minimizer W
    if $z^\star \geq 0$ **then** add (j,l) to $\mathcal{D}$ and **break**
    end if
    if $E_s^W < 1/\alpha$ for every call time $s \leq t$ **then** add W to $\mathcal{Q}$ and **break**
    else add to S a call time s with $E_s^W \geq 1/\alpha$
    end if
  end loop
end for
return $\mathcal{D}$

Proposition A.1 (Algorithm 3 is exact)**.** *For every $M \geq 2$, every $\alpha \in (0,1)$ and every sequence of wealth matrices, one per call time, Algorithm 3 in exact arithmetic returns the set $\mathcal{D}_t$ of Algorithm 1 at every call time t, and each inner loop ends after at most t solves.*

Exact certification is hard in the worst case. A decision problem is *coNP-hard* when every problem whose "no" answers can be checked in polynomial time reduces to it. Such a problem has no polynomial-time algorithm unless $\mathrm{P} = \mathrm{NP}$.

Proposition A.2 (Certification is coNP-hard in the worst case)**.** *Deciding from a wealth matrix and a level α whether a given pair belongs to $\mathcal{D}_1$, the certified set when the certifier is called once, on that matrix alone, is coNP-hard. The wealth matrices of the reduction take the values 0, $2^{\nu-1}$ and 2^ν for an integer $\nu \geq 1$, have polynomial encoding size, and are produced by the general update* (11) *of Appendix B.1 with $N = \infty$ from polynomially many binary items with predictable bets in $\{0,1\}$.*

The reduction is from the feedback arc set problem in tournaments (Alon, 2006; Charbit et al., 2007). A tournament is a directed graph with exactly one arc between every two vertices, and the problem asks for an ordering of its vertices with the fewest arcs pointing backward. The reduction uses bets in $\{0,1\}$, and Appendix D says what remains open for the default mixture bet. Each integer program (10) has $M(M-1)$ binary variables, one continuous variable and $M(M-1)(M-2)$ transitivity constraints. Algorithm 3 solved every instance in our experiments, and Appendix E.6 reports its running time.

An implementation stores $\log E_{\lambda,t}^{jl}$ instead of $E_{\lambda,t}^{jl}$ to avoid overflow. As for cost, the update is $O(M^2)$ per item, Algorithm 2 is $O(M^3)$ and Algorithm 1 is $O(M^2)$ per surviving ordering. In NumPy on one core the update costs $25\,\mu$s per item for 20 models and $47\,\mu$s for 50, the shortcut 0.07 ms and 0.6 ms per call, and the enumeration 58 ms per call for 8 models while every ordering survives, falling as they are discarded. At each call Algorithm 3 solves no integer program for a pair

that is already certified, that the shortcut certifies or that a witness in the pool covers. For any other pair it solves at least one and at most t (Proposition A.1).

B SUPPORTING THEORY

In Appendices B to E, the true ability vector $\theta^\star$ is also written θ, as in Sections 2 and 3.

B.1 VALIDITY AND GROWTH OF EACH PAIRWISE WEALTH

A large wealth must be improbable when the hypothesis it argues against is in fact true, however often and whenever the analyst looks. Fix $(j, l) \in \mathcal{P}$, write $Z_t = Z_t^{jl}$, and consider the general form of the update (2), $E_0^{jl} = 1$ and

$$E_t^{jl} = E_{t-1}^{jl}\Big(1 + \lambda_t \, \frac{Z_t - b_t}{1 + b_t}\Big), \tag{11}$$

in which the offset $b_t \in (-1, 1]$ and the bet $\lambda_t \in [0, 1]$ are *predictable*, that is, determined by the information $\mathcal{F}_{t-1}$. Each $E_{\lambda,t}^{jl}$ of (2) has this form with $\lambda_t \equiv \lambda$, and so does the grid average E_t^{jl}, with λ_t the mean of Λ weighted by the wealths $E_{\lambda,t-1}^{jl}$ (Remark B.3). The factor in (11) is nonnegative, and its conditional expectation given $\mathcal{F}_{t-1}$ is at most one whenever $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le b_t$. The wealth is then a nonnegative *supermartingale*: given the past, its expected next value is at most its current value. It is therefore an e-process for every predictable bet and stays one when frozen at a stopping time. The offsets (3) satisfy that condition for every law with $\theta_j \le \theta_l$, and the wealth of (2) is an average over $\lambda \in \Lambda$ of such processes. The choice of bets therefore affects power and never validity.

The pairwise construction follows the betting approach of Waudby-Smith & Ramdas (2024), including, under (F), its version for sampling without replacement. Waudby-Smith & Ramdas (2020) built earlier confidence sequences for that setting without betting.

Lemma B.1 (Pairwise betting supermartingale)**.** *Let P be a law under which $\mathbb{E}_P[Z_t \mid \mathcal{F}_{t-1}] \le b_t$ almost surely for every t. Then (E_t^{jl}) of* (11) *is a nonnegative $(\mathcal{F}_t)$-supermartingale under P with $E_0^{jl} = 1$, for every predictable choice of (λ_t). If the process is frozen at an $(\mathcal{F}_t)$-stopping time σ, that is, replaced by $E_{t\wedge\sigma}^{jl}$ where $t \wedge \sigma = \min\{t, \sigma\}$, the conclusion is unchanged.*

Proposition B.2 (Both sampling models)**.** *In each case below, for every $(j, l) \in \mathcal{P}$, every predictable bet (λ_t) with values in $[0, 1]$, and every data-generating law under which H_{jl} is true, (E_t^{jl}) is a nonnegative $(\mathcal{F}_t)$-supermartingale started at 1, hence an e-process for H_{jl}. This holds whatever the other models' parameters and whatever the dependence between models.*

(a) Under (F)*, with ξ independent of Π, for $t \le N$, with*

$$b_t = \max\Big\{-1 + \delta,\ \min\Big\{1,\ \frac{-S_{t-1}}{N - t + 1}\Big\}\Big\}, \qquad S_{t-1} = \sum_{s<t} Z_s, \tag{12}$$

for any fixed $\delta \in (0, 1]$, and $E_t^{jl} = E_N^{jl}$ for $t > N$.

(b) Under (S)*, with $b_t \equiv 0$.*

Remark B.3 (Bets). Validity holds for every predictable (λ_t), so the bet is chosen for power, and a constant bet is a poor choice when gaps differ across pairs. For binary scores under (S) (the finite case behaves the same away from the end of the benchmark), let $\Delta = \theta_j - \theta_l > 0$ and let $\rho = \mathbb{P}(Z \ne 0)$ be the *discordance rate*, the probability that the two models score differently on an item. Here Z stands for a generic paired difference Z_t^{jl}. The expected log-growth of a constant bet is $\mathbb{E}\log(1 + \lambda Z) \approx \lambda\Delta - \lambda^2\rho/2$: it is maximized at $\lambda^\star = \Delta/\rho$, which is in fact the exact maximizer when $\Delta < \rho$, and is *negative* beyond about $2\lambda^\star$. By the strong law of large numbers, a constant bet with negative expected log-growth drives the wealth of a pair with small Δ/ρ, close relative to its discordance rate, to zero almost surely. Such a pair may therefore never be certified, however long the evaluation runs (an early crossing of the threshold remains possible). Two predictable alternatives avoid this, and a third mitigates it. The first is the plug-in bet $\lambda_t = \min\{\bar\lambda, \max\{0, \widehat{\mathbb{E}}_{t-1}Y / \widehat{\mathbb{E}}_{t-1}Y^2\}\}$. Here

$Y_s = (Z_s - b_s)/(1 + b_s)$, $\widehat{\mathbb{E}}_{t-1}$ is the empirical average over items $1, \dots, t-1$, and $\bar{\lambda} < 1$ is fixed. This bet approximates GRAPA, the "growth rate adaptive to the particular alternative" bet (Waudby-Smith & Ramdas, 2024), which maximizes the average log-growth over the past items. Clipping to $[0, \bar{\lambda}] \subset [0, 1]$ meets the requirement of Lemma B.1, and the cap $\bar{\lambda} < 1$ keeps every factor at least $1 - \bar{\lambda} > 0$. The bet sets $\lambda_t = 0$ whenever $\widehat{\mathbb{E}}_{t-1} Y^2 = 0$ (in particular $\lambda_1 = 0$). The second is online Newton steps (Cutkosky & Orabona, 2018; Waudby-Smith & Ramdas, 2024), a second-order gradient update of λ_t on the running log-wealth, projected onto $[0, \bar{\lambda}]$. The third is a *mixture*. For constants $\lambda^{(1)}, \dots, \lambda^{(|\Lambda|)}$ let $E_t^{(k)} = \prod_{s \le t}(1 + \lambda^{(k)} Y_s)$ be the wealth of the kth constant; the mixture is the average $|\Lambda|^{-1} \sum_k E_t^{(k)}$. It is an average of supermartingales for one filtration, hence a supermartingale. It is also the wealth of the single predictable bet $\lambda_t = \sum_k E_{t-1}^{(k)} \lambda^{(k)} / \sum_k E_{t-1}^{(k)}$, the wealth-weighted mean of the grid (its denominator is positive because every $\lambda^{(k)} < 1$), so Lemma B.1 applies verbatim. A mixture over a fixed grid covers the range of Δ/ρ the grid spans but not arbitrarily small ratios. Once Δ/ρ falls below about half the smallest grid value, every component has negative expected log-growth and the mixture wealth also tends to zero (Proposition B.4). The mixture trails the best constant on the grid by at most a factor $|\Lambda|$ in wealth, that is, by $\log |\Lambda|$ in the threshold, and has no learning phase. The bet for pair (j, l) may depend on the whole past of all models.

Under (S) the log-wealth of a constant bet is a sum of independent bounded terms, which yields the growth rate of every bet on the grid, a consistency condition for the mixture and explicit certification-time bounds, with no restriction on the dependence between models within an item. Fix $(j, l) \in \mathcal{P}$ with $\Delta = \theta_j - \theta_l > 0$, write $Z_t = Z_t^{jl}$, $E_t = E_t^{jl}$ for the grid average of (2) and $E_t^{\lambda} = E_{\lambda,t}^{jl}$ for the wealth of the constant bet λ, and let $\lambda_{\min}$ be the smallest element of the grid $\Lambda = \{\lambda^{(1)}, \dots, \lambda^{(|\Lambda|)}\} \subset (0, \frac{1}{2}]$. Let $\mu(\lambda) = \mathbb{E} \log(1 + \lambda Z_1)$, $\rho = \mathbb{E}[Z_1^2]$, which is the discordance rate of Remark B.3 for binary scores, $c_\lambda = \log \frac{1+\lambda}{1-\lambda}$, $v_\lambda = \mathrm{Var} \log(1 + \lambda Z_1)$ and $\eta = \log(|\Lambda| M(M-1)/\alpha)$. Let $\tau^B = \inf\{t : E_t \ge M(M-1)/\alpha\}$ be the e-Bonferroni certification time of (j, l), and let τ^{sc} and τ be the times at which (j, l) enters $\mathcal{D}_t^{\mathrm{sc}}$ and $\mathcal{D}_t$, respectively, where $\mathcal{D}_t^{\mathrm{sc}}$ is the set certified by the shortcut (Proposition B.8). On every path the exact test certifies no later than the shortcut, and the shortcut no later than e-Bonferroni.

Proposition B.4 (Growth, consistency and certification time under (S))**.** *Assume* (S)*, no retirement, and* $\Delta > 0$*.*

(a) Growth. *For* $\lambda \in (0, 1)$*,* $t^{-1} \log E_t^{\lambda} \to \mu(\lambda)$ *almost surely;* μ *is concave with* $\mu(0) = 0$*, and for* $\lambda \le \frac{1}{2}$*,* $\lambda\Delta - \lambda^2 \rho \le \mu(\lambda) \le \lambda\Delta$*, so* $\mu(\lambda) > 0$ *whenever* $0 < \lambda < \Delta/\rho$*.*

(b) Consistency. *If* $\mu(\lambda_{\min}) > 0$ *then* $\tau^B < \infty$ *almost surely and* $(j, l) \in \mathcal{D}_t$ *for all* $t \ge \tau^B$*; if* $\mu(\lambda_{\min}) < 0$ *then every bet on the grid has negative growth and* $E_t \to 0$ *almost surely. The first case holds whenever* $\lambda_{\min} < \Delta/\rho$*; for binary scores,* $\mu(\lambda) = \frac{1}{2} \rho\, c_\lambda \big(\Delta/\rho - x_0(\lambda)\big)$ *with* $x_0(\lambda) = \log \frac{1}{1-\lambda^2} / c_\lambda$*, so it holds if and only if* $\Delta/\rho > x_0(\lambda_{\min})$*, which is* $\Delta/\rho > 0.015$ *for the default grid. For binary scores no finite grid is consistent for every* $\Delta > 0$*: pairs with* $0 < \Delta/\rho < x_0(\lambda_{\min})$ *have negative growth on every grid bet. Consistency for every positive gap would require bets approaching zero, which we do not pursue here.*

(c) Certification time. *Let* $\lambda \in \Lambda$ *have* $\mu = \mu(\lambda) > 0$ *and write* $c = c_\lambda$*. For every integer* $t \ge 2\eta/\mu$*,*

$$\mathbb{P}(\tau^B > t) \le \exp\Big(-\frac{t\mu^2}{2c^2}\Big) \quad \textit{and} \quad \mathbb{P}(\tau^B > t) \le \exp\Big(-\frac{t\mu^2}{8v_\lambda + \frac{4}{3}c\mu}\Big), \qquad v_\lambda \le 4\lambda^2\rho. \tag{13}$$

Define the rate and the horizon

$$a_\lambda = \max\Big\{\frac{\mu^2}{2c^2},\ \frac{\mu^2}{8v_\lambda + \frac{4}{3}c\mu}\Big\}, \qquad t^\star(\varepsilon) = \max\Big\{\frac{2\eta}{\mu},\ \frac{\log(1/\varepsilon)}{a_\lambda}\Big\}.$$

Then, for every $\varepsilon \in (0, 1)$*,* $\mathbb{P}(\tau^B \le \lceil t^\star(\varepsilon) \rceil) \ge 1 - \varepsilon$*, and* $\mathbb{E}[\tau^B] \le 2\eta/\mu + 1/a_\lambda + 2$*; the same bounds hold for* τ^{sc} *and* τ*.*

(d) Default grid. *If* $0.06 \le \Delta/\rho \le 1$*, some grid bet lies in* $[\Delta/(4.2\rho), \Delta/(2\rho)]$*, for which* $0.18\,\Delta^2/\rho \le \mu \le \Delta^2/(2\rho)$ *and* $c \le 1.1\,\Delta/\rho$*, and then*

$$t^\star(\varepsilon) \le \frac{11.2\,\rho\eta + \min\{74.7,\, 270\rho\}\log(1/\varepsilon)}{\Delta^2}, \quad \mathbb{E}[\tau^B] \le \frac{11.2\,\rho\eta + \min\{74.7,\, 270\rho\}}{\Delta^2} + 2, \tag{14}$$

so certification takes $O(\rho\,\Delta^{-2}\log(|\Lambda|M^2/(\alpha\varepsilon)))$ *items. For binary scores with* $x_0(0.03) < \Delta/\rho < 0.06$ *only the general bound above applies, with the exact growth* $\mu(0.03) = \frac{1}{2}\rho\,c_{0.03}(\Delta/\rho - x_0(0.03))$*, which tends to zero as* $\Delta/\rho \downarrow x_0(0.03) \approx 0.015$*: a fixed grid has no uniform rate over its positive-growth pairs.*

B.2 Supplement to the validity theorem

Remark B.5 (The restriction must be stated through $T(W)$). Every subset $\mathcal{S} \subseteq \mathcal{P}$ of the one-sided hypotheses H_{jl} is *jointly satisfiable*, since $\theta_1 = \cdots = \theta_M$ makes all of them true. Requiring only that an intersection be consistent with some weak order therefore excludes nothing. The restriction that does exclude intersections is that of Shaffer (1986): $\mathcal{S}$ needs a test only if it can be the *exact* set of true hypotheses, that is, $\mathcal{S} = T(W)$ for some W. The full closure tests the $2^{M(M-1)} - 1$ intersection hypotheses $\bigcap_{(j,l)\in\mathcal{S}} H_{jl}$, one per nonempty $\mathcal{S} \subseteq \mathcal{P}$. Of these, $|\mathcal{W}|$ remain: 13 of 63 for $M = 3$, 75 of 4095 for $M = 4$. For $M = 2$ nothing is saved.

The family of intersections $\bigcap_{(j,l)\in T(W)} H_{jl}$, $W \in \mathcal{W}$, that Algorithm 1 tests is not itself closed under intersection. This does no harm: the proof of Theorem 4.2 uses only that the cell $\Theta_{W^\star} = \{\theta : W(\theta) = W^\star\}$ containing the truth is rejected with probability at most α. The restriction to exactly realizable intersections is also that of Westfall (1997).

Proposition B.6 (The output is a partial order)**.** *On the event* $\{\mathcal{C}_t \neq \emptyset\}$*, which contains the event in* (7)*,* $\mathcal{D}_t$ *is a strict partial order on* $[M]$*: irreflexive, asymmetric and transitive. If* $\mathcal{C}_t = \emptyset$ *then* $\mathcal{D}_t = \mathcal{P}$*; this has probability at most* α *uniformly in* t*, and it reveals that an error has occurred.*

B.3 Testing only orderings loses nothing, and the shortcut is valid

Testing only the orderings instead of every set of pairwise hypotheses restricts the closed test. Both claims of Section 4.1 about what this restriction costs hold on every data set, not merely with high probability. Both compare procedures applied to the same wealths.

Proposition B.7 (Restriction never hurts)**.** *Let* $\mathcal{D}_t^{\mathrm{full}}$ *be the output of the unrestricted closed test that rejects* H_{jl} *once* $\max_{s\le t} |\mathcal{S}|^{-1} \sum_{(a,b)\in\mathcal{S}} E_s^{ab} \ge 1/\alpha$ *for every set* $\mathcal{S} \subseteq \mathcal{P}$ *that contains* (j,l)*. Then* $\mathcal{D}_t^{\mathrm{full}} \subseteq \mathcal{D}_t$ *for all* t *on every sample path. For some wealth configurations, exhibited in Example B.11, the restricted test certifies a pair that the full closure does not.*

Proposition B.8 (Transitivity pooling; validity of Algorithm 2)**.** *Let*

$$B_t^{jl} = E_t^{jl} + \sum_{m\in[M]\setminus\{j,l\}} \min\{E_t^{jm}, E_t^{ml}\}$$

be the pooled statistic of Section 3 and let $\mathcal{D}_t^{\mathrm{sc}} = \{(j,l) \in \mathcal{P} : \max_{s\le t} B_s^{jl} \ge M(M-1)/\alpha\}$ *be the set that Algorithm 2 certifies before closure. Then* $\mathcal{D}_t^{\mathrm{sc}} \subseteq \mathcal{D}_t$ *for all* t *on every sample path, so* $\mathcal{D}^{\mathrm{sc}}$ *inherits* (8)*. It costs* $O(M)$ *per pair per update and contains the e-Bonferroni rejections* $\{\max_{s\le t} E_s^{jl} \ge M(M-1)/\alpha\}$.

Relation to existing results. To our knowledge both propositions are new. Proposition B.7 has a known relative. It is the sequential counterpart of the fixed-sample observation of Shaffer (1986) that logical relations among hypotheses permit a milder multiplicity correction at no cost in error. It differs in that it holds for e-processes, on every sample path and at all times, and in that its comparator $\mathcal{D}^{\mathrm{full}}$ is the e-value closed test of Hartog & Lei (2025). Proposition B.8 has, to our knowledge, no direct counterpart. Its closest relative is the certification logic of Gu et al. (2026), which corrects by Bonferroni over the pairs (e-Bonferroni in our terms) and uses transitivity only after certification, to propagate certified pairs. The pooled statistic B_t^{jl} uses transitivity before certification (Appendix A). Because $B_t^{jl} \ge E_t^{jl}$, the shortcut contains their rejections, at a threshold that stays conservative relative to the exact test.

Lemma B.9 (Transitive closure adds no error)**.** *Let $(\mathcal{D}_t)$ be a certified set with good event $\mathcal{G}$, and let $\overline{\mathcal{D}}_t$ be the transitive closure of $\mathcal{D}_t$, the set of $(j,l) \in \mathcal{P}$ joined by a directed path $j = v_0, \dots, v_r = l$, $r \geq 1$, with every $(v_{i-1}, v_i) \in \mathcal{D}_t$. On $\mathcal{G}$, for every t, every $(j,l) \in \overline{\mathcal{D}}_t$ satisfies $\theta^\star_j > \theta^\star_l$; in particular $\overline{\mathcal{D}}_t$ is a strict partial order and $\mathcal{D}_t$ contains no directed cycle. Hence $(\overline{\mathcal{D}}_t)$ is a certified set with the* same *good event: closing under transitivity adds rejections but never turns an error-free sample path into an erroneous one, so it leaves the family-wise error guarantee unchanged.*

Corollary B.10 (Certified top-k sets)**.** *Let $(\mathcal{D}_t)$ be a certified set with $L_{j,t}, U_{j,t}$ as in* (6)*. For $k \in [M]$ let $I_{k,t} = \{j : U_{j,t} \leq k\}$ and $O_{k,t} = \{j : L_{j,t} > k\}$. On $\mathcal{G}$, for all k and t simultaneously, every $j \in I_{k,t}$ has $R_j(\theta^\star) \leq k$ and every $j \in O_{k,t}$ has $R_j(\theta^\star) > k$.*

Example B.11 (Strictness, $M = 3$)**.** Let $\theta_1 > \theta_2 > \theta_3$ and suppose that at some time t the wealths are $E^{13} \geq 6/\alpha$, $E^{12} = a$ and $E^{23} = c$, with the three wealths of true hypotheses all zero and with no earlier time at which E^{12} or the mean over $\mathcal{P} \setminus \{(1,3)\}$ was larger. Of the eight weak orders with $1 \preceq_W 2$, six contain $(1,3)$ and are eliminated through E^{13}; the other two place model 3 strictly below model 1 and hence, by transitivity, below model 2, so their means are $a/3$ and $a/4$. The restricted test therefore certifies $1 \succ 2$ once $a \geq 4/\alpha$, whereas the full closure must also reject $\mathcal{P} \setminus \{(1,3)\}$, needing $a + c \geq 5/\alpha$, and e-Bonferroni needs $a \geq 6/\alpha$. Whenever $c < 1/\alpha$, the restricted test is strictly stronger than both, for $4/\alpha \leq a < 5/\alpha - c$: with $\alpha = 0.1$, $E^{13} = 60$, $a = 42$ and $c = 5$, it certifies $1 \succ 2$ while the full-closure statistic stands at 9.4 against a threshold of 10 and E^{12} at 42 against 60.

The bound of Proposition B.8 is crude in two places that could be tightened. It divides by $M(M-1)$ for every ordering, although $|T(W)| = \binom{M}{2} + \#\{\text{tied pairs}\}$ reaches that value only when every pair is tied. It also has no step-down, that is, no lowering of the threshold for the remaining orders once some orders are rejected. The exact test has neither weakness, so what either change could recover is at most what the exact test gains over e-Bonferroni. In simulation the exact test certifies about 7% more true dominances than e-Bonferroni after 500 items and at most 3% more by 4000. On the eight highest-scoring models of the real leaderboard it certifies 2–5% more up to three quarters of a benchmark and nearly the same number at the end (Appendix E.6).

Remark B.12 (Weights must not depend on time). The plain average in (4) may be replaced by any weighted form $E^W_t = \sum_{(j,l) \in T(W)} w^W_{jl} \, E^{jl}_t$ with fixed weights $w^W_{jl} \geq 0$ summing to one over $T(W)$, which the proof of Theorem 4.2 covers without change; the weights may depend on W. What is *not* valid is letting them adapt: a predictably re-weighted average $\sum_k w_{k,t} E^{(k)}_t$ of supermartingales is not a supermartingale in general, so the weights in (4) must not depend on t. Propositions B.7 and B.8 are stated for the plain average.

B.4 Tiers and retirement

An *antichain* of a strict partial order $\mathcal{D}$ is a set of models no two of which are related by $\mathcal{D}$. The tier guarantee of Section 4.2 is part (c) below, with its monotonicity in (d); the rest is standard order theory.

Theorem B.13 (Tier guarantee, full statement)**.** *Let $\mathcal{D}$ be a strict partial order on $[M]$ and compute tiers from $\mathcal{D}$ as in Step 3 of Section 3. Parts (a), (b) and (e) hold for every such $\mathcal{D}$. In (c) and (d), $\mathrm{tier}_t(j)$ is the tier computed from the transitive closure $\overline{\mathcal{D}}_t$ of a certified set on the event $\mathcal{G}$, where $\overline{\mathcal{D}}_t$ is a strict partial order by Lemma B.9, and (d) assumes in addition $\mathcal{D}_t \subseteq \mathcal{D}_{t+1}$; the subscript is dropped when t is fixed. On $\mathcal{G}$ these tiers equal those computed from $\mathcal{D}_t$ itself: $\mathcal{D}_t \subseteq \overline{\mathcal{D}}_t$, and each pair of $\overline{\mathcal{D}}_t$ is joined by a chain in $\mathcal{D}_t$, which has no cycle on $\mathcal{G}$, so longest chains have the same length.*

(a) Partition into antichains. *The tiers partition $[M]$, tiers $1, \dots, \max_j \mathrm{tier}(j)$ are all nonempty, and no certified dominance holds between two models of the same tier.*

(b) Order. *If $(j,l) \in \mathcal{D}$ then $\mathrm{tier}(j) < \mathrm{tier}(l)$. Every model in tier $s \geq 2$ is dominated by some model in tier $s - 1$.*

(c) Guarantee. *On $\mathcal{G}$, simultaneously for all j and all t, the bound* (9) *holds, that is, $R_j(\theta^\star) \geq L_{j,t} \geq \mathrm{tier}_t(j)$, where $L_{j,t}$ is computed from $\overline{\mathcal{D}}_t$; the number of tiers is at most the number of distinct values among $\theta^\star_1, \dots, \theta^\star_M$.*

(d) Monotonicity. *If $\mathcal{D}_t \subseteq \mathcal{D}_{t+1}$ then $\mathrm{tier}_t(j) \le \mathrm{tier}_{t+1}(j)$: a model is never promoted as evidence accrues.*

(e) Minimality. *No partition of $[M]$ into antichains of $\mathcal{D}$ has fewer blocks (Mirsky, 1971).*

Remark B.14 (What tiers do not say). With $\mathcal{D} = \{(a, b)\}$ on $\{a, b, c\}$ the tiers are $\{a, c\}$ and $\{b\}$, and c is comparable to neither, which is why $\mathrm{tier}(j) < \mathrm{tier}(l)$ does not imply that j is certified better than l. The intuitive definition "same tier if and only if neither dominates the other" fails because incomparability is not transitive: with $\mathcal{D} = \{(a, c)\}$, a and b are incomparable, b and c are incomparable, and $a \succ c$.

The cost of retirement is bounded under (S) for the rule that retires a model once every pair involving it is resolved. The top-k rule, under which a pair can be frozen before it is certified, is measured in Appendix E.3 and not bounded here.

Corollary B.15 (Cost of all-pairs retirement under (S))**.** *Assume* (S)*, bets that are functions of each pair's own past (the mixture is), and the* all-pairs rule*, which retires model j at the first time every pair $\{j, l\}$ is certified in one direction. For an unordered pair let $q(j, l)$ be the ordered pair in the true direction, and let $\mathcal{P}^+ = \{(j, l) : \theta_j > \theta_l\}$. For $q = (j, l) \in \mathcal{P}^+$ write $\Delta_q = \theta_j - \theta_l$, $\rho_q = \mathbb{E}[(Z_1^{jl})^2]$ and $\mu_q(\lambda) = \mathbb{E}\log(1 + \lambda Z_1^{jl})$. These are the quantities Δ, ρ and $\mu(\lambda)$ of Proposition B.4 computed for the pair q. Assume that every $q \in \mathcal{P}^+$ involving j has a bet $\lambda_q \in \Lambda$ with $\mu_q = \mu_q(\lambda_q) > 0$, which holds if and only if $\mu_q(\lambda_{\min}) > 0$, and let $a_q = a_{\lambda_q}$ be its rate in Proposition B.4(c) and τ_q^B the e-Bonferroni time of q computed without retirement. If model j is tied with no other model, put $\kappa_j = \max_{l \ne j} 2\eta/\mu_{q(j,l)}$ and $\beta_j = \max_{l \ne j} 1/a_{q(j,l)}$. Then $\sigma_j \le \max_{l \ne j} \tau^B_{q(j,l)}$ on every path, $\mathbb{P}(\sigma_j > t) \le (M - 1)e^{-t/\beta_j}$ for every integer $t \ge \kappa_j$, and*

$$\mathbb{E}[\sigma_j] \le \kappa_j + \beta_j\big(\log(M - 1) + 1\big) + 2. \tag{15}$$

If every model is untied and every true pair satisfies $0.06 \le \Delta_q/\rho_q \le 1$, then with $\Delta_j = \min_{l \ne j} |\theta_j - \theta_l|$ and $\rho_j = \max_{l \ne j} \rho_{q(j,l)}$,

$$\mathbb{E}[\sigma_j] \le \frac{11.2\,\rho_j\eta + \min\{74.7,\, 270\rho_j\}\big(\log(M - 1) + 1\big)}{\Delta_j^2} + 2, \tag{16}$$

and consequently $\mathbb{E}[\sum_j \sigma_j] = O\big(\sum_j \rho_j \Delta_j^{-2} \log(|\Lambda| M^2/\alpha)\big)$. A model tied with another model is never retired under the all-pairs rule on the good event $\mathcal{G}$; over any finite evaluation horizon it therefore incurs the full planned cost.

C Proofs

Time is discrete, $t \in \mathbb{N}_0 = \{0, 1, 2, \dots\}$, and $\mathbf{1}\{A\}$ denotes the indicator of an event A. We write A^c for the complement of an event A and abbreviate "almost surely" to a.s. A random variable is *$\mathcal{F}_t$-measurable* if its value is determined by the information $\mathcal{F}_t$. Thus adapted means $\mathcal{F}_t$-measurable at each t, and predictable means $\mathcal{F}_{t-1}$-measurable. Conditional expectations of nonnegative random variables are always defined, possibly infinite, and the tower property holds for them; we use this without comment. In each proof, the bracket at the right of a line states what justifies that line.

C.1 Pairwise wealths

Proof of Lemma B.1. Write $\phi_t = 1 + \lambda_t(Z_t - b_t)/(1 + b_t)$, so that $E_t^{jl} = \prod_{s \le t} \phi_s$ by the definition of the wealth in (11).

Nonnegativity and measurability.

$$\begin{aligned}
\frac{Z_t - b_t}{1 + b_t} &\ge \frac{-1 - b_t}{1 + b_t} = -1 && [Z_t \ge -1 \text{ and } 1 + b_t > 0] \\
\phi_t &\ge 1 - \lambda_t \ge 0 && [\text{previous line and } \lambda_t \in [0, 1]] \\
E_t^{jl} = \prod_{s \le t} \phi_s &\ge 0, \text{ and is } \mathcal{F}_t\text{-measurable} && [Z_s, b_s, \lambda_s \text{ are } \mathcal{F}_s\text{-measurable for } s \le t]
\end{aligned}$$

Bound on one factor. To avoid any integrability question for the signed quantity $\lambda_t(Z_t - b_t)/(1+b_t)$, rewrite ϕ_t as a sum of two nonnegative terms:

$$\phi_t = 1 - \lambda_t + \frac{\lambda_t}{1+b_t}\,(1+Z_t)$$

[algebra; both terms are nonnegative since $\lambda_t \le 1$, $1+Z_t \ge 0$, $1+b_t > 0$]

$$\mathbb{E}_P[\phi_t \mid \mathcal{F}_{t-1}] = 1 - \lambda_t + \frac{\lambda_t}{1+b_t}\big(1 + \mathbb{E}_P[Z_t \mid \mathcal{F}_{t-1}]\big) \qquad [\lambda_t \text{ and } b_t \text{ are } \mathcal{F}_{t-1}\text{-measurable}]$$

$$\le\ 1 - \lambda_t + \frac{\lambda_t}{1+b_t}(1+b_t) = 1 \qquad [\mathbb{E}_P[Z_t \mid \mathcal{F}_{t-1}] \le b_t \text{ and } \lambda_t/(1+b_t) \ge 0]$$

Supermartingale.

$$\mathbb{E}_P[E_t^{jl} \mid \mathcal{F}_{t-1}] = E_{t-1}^{jl}\,\mathbb{E}_P[\phi_t \mid \mathcal{F}_{t-1}] \qquad [E_{t-1}^{jl} \ge 0 \text{ is } \mathcal{F}_{t-1}\text{-measurable}]$$

$$\le\ E_{t-1}^{jl} \qquad [\text{bound on one factor}]$$

$$\mathbb{E}_P[E_t^{jl}] \ \le\ \mathbb{E}_P[E_{t-1}^{jl}] \ \le\ \ldots \ \le\ E_0^{jl} = 1 \qquad [\text{tower property; induction on } t]$$

so the process is integrable and is a nonnegative supermartingale started at 1.

Frozen process. Let σ be an $(\mathcal{F}_t)$-stopping time.

$$E_{t\wedge\sigma}^{jl} - E_{(t-1)\wedge\sigma}^{jl} = \mathbf{1}\{\sigma \ge t\}\big(E_t^{jl} - E_{t-1}^{jl}\big)$$

[both sides vanish on $\{\sigma < t\}$ and agree on $\{\sigma \ge t\}$]

$$\mathbb{E}_P\big[E_{t\wedge\sigma}^{jl} - E_{(t-1)\wedge\sigma}^{jl}\,\big|\,\mathcal{F}_{t-1}\big] = \mathbf{1}\{\sigma \ge t\}\,\big(\mathbb{E}_P[E_t^{jl} \mid \mathcal{F}_{t-1}] - E_{t-1}^{jl}\big) \ \le\ 0$$

[$\{\sigma \ge t\} = \{\sigma \le t-1\}^c \in \mathcal{F}_{t-1}$; E_t^{jl}, E_{t-1}^{jl} integrable by the previous block; supermartingale property]

so the stopped process is again a nonnegative supermartingale started at 1. □

Proof of Proposition B.2. By Lemma B.1 it suffices to verify, whenever $\theta_j \le \theta_l$, that b_t is predictable with $b_t \in (-1, 1]$ and that $\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \le b_t$.

(a) Under (F), let $z_i = x_{ij} - x_{il}$, so that H_{jl} reads $\sum_{i=1}^N z_i \le 0$. Fix $t \le N$.

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \frac{\sum_{i=1}^N z_i - S_{t-1}}{N-t+1}$$

[given $\mathcal{F}_{t-1}$, $\Pi(t)$ is uniform on the $N-t+1$ unrevealed items, and ξ is independent of Π]

$$\le\ \frac{-S_{t-1}}{N-t+1} \qquad [\textstyle\sum_{i=1}^N z_i \le 0]$$

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \ \le\ 1 \qquad [Z_t \le 1, \text{ scores lying in } [0,1]]$$

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] \ \le\ \min\Big\{1, \frac{-S_{t-1}}{N-t+1}\Big\} \ \le\ b_t$$

[the two lines above; taking the maximum with $-1+\delta$ in eq. (12) can only raise the bound]

$$-1 < -1+\delta \ \le\ b_t \ \le\ 1$$

[eq. (12): the clipping, as Lemma B.1 requires; the minimum with 1 and $\delta \le 1$]

b_t is $\mathcal{F}_{t-1}$-measurable [it is a function of $S_{t-1} = \sum_{s<t} Z_s$]

For $t > N$ the process is constant. All $M(M-1)$ processes are adapted to the single filtration generated by the shared permutation; independence of permutations across pairs is neither available nor needed.

(b) Under (S), with $b_t \equiv 0$:

$$\mathbb{E}[Z_t \mid \mathcal{F}_{t-1}] = \mathbb{E}[X_{tj} - X_{tl}] \qquad [X_t \text{ is independent of } \mathcal{F}_{t-1} = \sigma(X_1, \ldots, X_{t-1}, \xi)]$$

$$= \theta_j - \theta_l \ \le\ 0 = b_t \qquad [\text{definition of } \theta;\ H_{jl}]$$

Nothing is assumed about the joint law of $(X_{t1}, \ldots, X_{tM})$, and λ_t may depend on the past of all models, since only $\mathcal{F}_{t-1}$-measurability is used. □

Proof of Proposition B.4. Under (S) with $b_t \equiv 0$ the factor of the constant bet λ on item s is $1+\lambda Z_s$; write $\ell_s = \log(1+\lambda Z_s)$, so that $\log E_t^\lambda = \sum_{s\le t} \ell_s$.

(a) Growth. Let $f(u) = \log(1+u) - u + u^2$; then $f(0) = 0$ and $f'(u) = u(1+2u)/(1+u)$, which is ≤ 0 on $[-\frac{1}{2}, 0]$ and ≥ 0 on $[0, \frac{1}{2}]$, so $f \ge 0$ on $[-\frac{1}{2}, \frac{1}{2}]$.

$$\ell_1, \ell_2, \dots \text{ are i.i.d. with } |\ell_s| \le \log \tfrac{1}{1-\lambda} \qquad [Z_s \text{ i.i.d., } |Z_s| \le 1, \lambda < 1]$$

$$t^{-1} \textstyle\sum_{s\le t} \ell_s \to \mathbb{E}\ell_1 = \mu(\lambda) \text{ a.s.} \qquad [\text{strong law of large numbers}]$$

$$\mu \text{ is concave on } [0,1), \text{ and } \mu(0) = 0$$

$$[\lambda \mapsto \log(1+\lambda z) \text{ is concave for each } z \in [-1,1]; \text{ expectation preserves concavity}]$$

$$u - u^2 \le \log(1+u) \le u \quad \text{for } u \in [-\tfrac{1}{2}, \tfrac{1}{2}] \qquad [f \ge 0; \text{ concavity of log}]$$

$$\lambda\Delta - \lambda^2\rho \le \mu(\lambda) \le \lambda\Delta \quad \text{for } \lambda \le \tfrac{1}{2}$$

$$[u = \lambda Z_1 \in [-\tfrac{1}{2}, \tfrac{1}{2}]; \text{ take expectations, } \mathbb{E}Z_1 = \Delta, \mathbb{E}Z_1^2 = \rho]$$

(b) Consistency. Suppose $\mu(\lambda_{\min}) > 0$.

$$\log E_t \ge \log E_t^{\lambda_{\min}} - \log|\Lambda| \to +\infty \text{ a.s.} \qquad [\text{the mixture averages nonnegative terms; (a)}]$$

$$\tau^B < \infty \text{ a.s.} \qquad [\tau^B \text{ is the first time } E_t \ge M(M-1)/\alpha]$$

$$(j,l) \in \mathcal{D}_t \text{ for all } t \ge \tau^B$$

[e-Bonferroni uses the running maximum and is contained in $\mathcal{D}_t^{\mathrm{sc}} \subseteq \mathcal{D}_t$ (Proposition B.8); $\mathcal{D}_t$ is monotone]

Suppose $\mu(\lambda_{\min}) < 0$ and let $\lambda \in \Lambda$ with $\lambda > \lambda_{\min}$.

$$\mu(\lambda_{\min}) \ge \tfrac{\lambda_{\min}}{\lambda}\mu(\lambda) + \big(1 - \tfrac{\lambda_{\min}}{\lambda}\big)\mu(0) \qquad [\text{concavity of } \mu]$$

$$\mu(\lambda) \le \tfrac{\lambda}{\lambda_{\min}}\mu(\lambda_{\min}) < 0 \qquad [\mu(0) = 0; \text{ rearrange}]$$

$$E_t \to 0 \text{ a.s.} \qquad [\text{(a): each of the } |\Lambda| \text{ components has } \log E_t^\lambda \to -\infty]$$

For binary scores, with $p_\pm = \mathbb{P}(Z_1 = \pm 1) = (\rho \pm \Delta)/2$ and $\rho > 0$:

$$\mu(\lambda) = p_+ \log(1+\lambda) + p_- \log(1-\lambda) = \tfrac{1}{2}\rho \log(1-\lambda^2) + \tfrac{1}{2}\Delta\, c_\lambda$$

$$[Z_1 \in \{-1,0,1\}; \log(1+\lambda) \pm \log(1-\lambda) \text{ equal } \log(1-\lambda^2) \text{ and } c_\lambda]$$

$$= \tfrac{1}{2}\rho\, c_\lambda\big(\Delta/\rho - x_0(\lambda)\big), \qquad x_0(\lambda) = \log \tfrac{1}{1-\lambda^2}/c_\lambda$$

$$[\text{factor out } \tfrac{1}{2}\rho c_\lambda > 0]$$

$$\mu(\lambda_{\min}) > 0 \iff \Delta/\rho > x_0(\lambda_{\min})$$

$$[\text{previous line; } x_0(0.03) = 0.000900/0.0600 = 0.0150]$$

(c) Certification time. Fix $\lambda \in \Lambda$ with $\mu > 0$ and put $V_s = \ell_s - \mu$.

$$\ell_s \in [\log(1-\lambda), \log(1+\lambda)], \text{ an interval of length } c \qquad [|Z_s| \le 1]$$

$$\{\tau^B > t\} \subseteq \{E_t < \tfrac{M(M-1)}{\alpha}\} \subseteq \{E_t^\lambda < \tfrac{|\Lambda|M(M-1)}{\alpha}\} = \{\textstyle\sum_{s\le t} \ell_s < \eta\}$$

$$[\text{first crossing; } E_t \ge E_t^\lambda/|\Lambda|; \text{ definition of } \eta]$$

$$\mathbb{P}\big(\textstyle\sum_{s\le t} \ell_s \le t\mu - x\big) \le \exp\big(-2x^2/(tc^2)\big), \quad x > 0$$

$$[\text{Hoeffding's inequality, i.i.d. summands with range } c]$$

$$\mathbb{P}(\tau^B > t) \le \exp\big(-t\mu^2/(2c^2)\big) \quad \text{for integer } t \ge 2\eta/\mu$$

$$[\text{the two previous lines with } x = t\mu - \eta \ge t\mu/2; \text{ the bound decreases in } x]$$

$$|\log(1+u)| \le 2|u| \quad \text{for } u \in [-\tfrac{1}{2}, \tfrac{1}{2}]$$

$$[u \ge 0\text{: } \log(1+u) \le u; \ u < 0\text{: } -\log(1+u) \le \tfrac{1}{1+u} - 1 = \tfrac{-u}{1+u} \le 2|u|]$$

$$v_\lambda \le \mathbb{E}\ell_1^2 \le 4\lambda^2\rho \qquad [\text{previous line with } u = \lambda Z_1]$$

$$\mathbb{P}\big(\textstyle\sum_{s\le t} V_s \le -x\big) \le \exp\Big(-\frac{x^2}{2(tv_\lambda + cx/3)}\Big), \quad x > 0$$

$$[\text{Bernstein's inequality: } V_s \text{ i.i.d., centered, } |V_s| \le c]$$

$$\mathbb{P}(\tau^B > t) \le \exp\Big(-\tfrac{(t\mu/2)^2}{2(tv_\lambda + ct\mu/6)}\Big) = \exp\Big(-\tfrac{t\mu^2}{8v_\lambda + \frac{4}{3}c\mu}\Big) \quad \text{for integer } t \ge 2\eta/\mu$$

$$[\{\tau^B > t\} \subseteq \{\textstyle\sum_{s\le t} V_s < -(t\mu - \eta)\} \text{ with } t\mu - \eta \ge t\mu/2; \text{ the bound decreases in } x]$$

which is (13); since both bounds hold, $\mathbb{P}(\tau^B > t) \le e^{-a_\lambda t}$ for every integer $t \ge 2\eta/\mu$, with a_λ the larger of the two rates. Write $a = a_\lambda$.

$$\mathbb{P}(\tau^B > \lceil t^\star(\varepsilon)\rceil) \le \varepsilon \qquad [\lceil t^\star\rceil \ge 2\eta/\mu \text{ and } a\lceil t^\star\rceil \ge \log(1/\varepsilon)]$$

$$\mathbb{E}[\tau^B] = \textstyle\sum_{t\ge 0}\mathbb{P}(\tau^B > t) \le t_0 + \sum_{t\ge t_0} e^{-at} \le t_0 + 1 + \frac{1}{a}, \quad t_0 = \lceil 2\eta/\mu\rceil$$

[probabilities are at most one below t_0; geometric series; $1/(1-e^{-a}) \le 1 + 1/a$]

$$\mathbb{E}[\tau^B] \le 2\eta/\mu + 1/a + 2 \qquad [t_0 \le 2\eta/\mu + 1]$$

the bounds hold for τ^{sc} and τ [$\tau \le \tau^{\mathrm{sc}} \le \tau^B$ on every path, Proposition B.8]

(d) Default grid. Let $I = [\Delta/(4.2\rho), \Delta/(2\rho)]$. If $0.06 \le \Delta/\rho \le 0.126$ then $0.03 \in I$, since $0.03 \le \Delta/(2\rho)$ iff $\Delta/\rho \ge 0.06$ and $0.03 \ge \Delta/(4.2\rho)$ iff $\Delta/\rho \le 0.126$. If $0.126 < \Delta/\rho \le 1$, sort the grid increasingly and let $\lambda^{(k)}$ be the largest grid bet at most $\Delta/(4.2\rho)$; either $\lambda^{(k)} = \Delta/(4.2\rho) \in I$, or $\lambda^{(k)} < \Delta/(4.2\rho) \le 1/4.2 < 0.5$, so that $\lambda^{(k+1)}$ exists and, consecutive grid ratios being at most $2.084 < 2.1$, $\Delta/(4.2\rho) < \lambda^{(k+1)} \le 2.084\,\lambda^{(k)} < \Delta/(2\rho)$. In both cases some grid bet λ lies in I, and for it:

$$\mu \ge \lambda(\Delta - \lambda\rho) \ge 0.18\,\Delta^2/\rho$$

[(a); $\lambda(\Delta - \lambda\rho)$ is concave in λ, with values $0.181\,\Delta^2/\rho$ and $0.25\,\Delta^2/\rho$ at the ends of I]

$$\mu \le \lambda\Delta \le \tfrac{\Delta^2}{2\rho}, \qquad c \le 2.2\lambda \le 1.1\,\tfrac{\Delta}{\rho}$$

[(a); $c_\lambda/\lambda = 2\,\mathrm{artanh}(\lambda)/\lambda$ increases to 2.197 at $\lambda = \frac{1}{2}$]

$$\tfrac{2\eta}{\mu} \le \tfrac{11.2\,\rho\eta}{\Delta^2}, \qquad \tfrac{2c^2}{\mu^2} \le \tfrac{2(1.21)}{0.0324}\,\Delta^{-2} = 74.7\,\Delta^{-2} \qquad \text{[substitute the bounds on } \mu \text{ and } c]$$

$$8v_\lambda \le 32\lambda^2\rho \le \tfrac{8\Delta^2}{\rho}, \qquad \tfrac{4}{3}c\mu \le \tfrac{4}{3}(1.1)\tfrac{1}{2}\,\tfrac{\Delta^3}{\rho^2} \le 0.734\,\tfrac{\Delta^2}{\rho}$$

[$\lambda \le \Delta/(2\rho)$; the bounds on c and μ; $\Delta \le \rho$ in this regime]

$$\tfrac{8v_\lambda + \frac{4}{3}c\mu}{\mu^2} \le \tfrac{8.734\,\Delta^2/\rho}{0.0324\,\Delta^4/\rho^2} \le 270\,\tfrac{\rho}{\Delta^2} \qquad \text{[the two previous lines]}$$

and (14) follows from (c) with $1/a_\lambda \le \min\{74.7, 270\rho\}\Delta^{-2}$. For binary scores the exact growth of $\lambda = 0.03$ is the formula in (b), with $c_{0.03} = 0.0600$. □

C.2 Validity, structure and power

Proof of Theorem 4.2. Fix the data-generating law P, with parameter $\theta^\star$ and $W^\star = W(\theta^\star)$, and write $E^W = (E_t^W)_{t\ge 0}$ for the process of (4).

$E^{W^\star}$ is an e-process under P. Let τ be a bounded $(\mathcal{F}_t)$-stopping time.

$$(j,l) \in T(W^\star) \;\Rightarrow\; \theta_j^\star \le \theta_l^\star \qquad \text{[definition of } T(W)]$$

$$\Rightarrow\; \mathbb{E}_P\big[E_\tau^{jl}\big] \le 1 \qquad \text{[Assumption 4.1; } \tau \text{ is a stopping time for the common filtration]}$$

$$\mathbb{E}_P\big[E_\tau^{W^\star}\big] = \sum_{(j,l)\in T(W^\star)} w_{jl}^{W^\star}\,\mathbb{E}_P\big[E_\tau^{jl}\big]$$

[eq. (4) in its weighted form (Remark B.12); linearity, the weights being deterministic]

$$\le \sum_{(j,l)\in T(W^\star)} w_{jl}^{W^\star} = 1 \qquad [w_{jl}^{W^\star} \ge 0, \text{ the previous line, and the weights sum to one}]$$

Only linearity of expectation is used; the joint law of the pairwise processes plays no role. That all of them are e-processes for the *same* filtration is needed, because τ must be a stopping time for each of them at once.

Ville's inequality. Let $\tau^\star = \inf\{t : E_t^{W^\star} \ge 1/\alpha\}$ and fix a deterministic n.

$$\tau^\star \text{ is an } (\mathcal{F}_t)\text{-stopping time} \qquad [E^{W^\star} \text{ is adapted}]$$

$$E_{\tau^\star\wedge n}^{W^\star} \;\ge\; \alpha^{-1}\mathbf{1}\{\tau^\star \le n\} \qquad \text{[nonnegativity on } \{\tau^\star > n\}\text{; the threshold on } \{\tau^\star \le n\}]$$

$$1 \;\ge\; \mathbb{E}_P\big[E_{\tau^\star\wedge n}^{W^\star}\big] \qquad \text{[the e-process property above, at the bounded stopping time } \tau^\star \wedge n]$$

$$\ge\; \alpha^{-1}\mathbb{P}(\tau^\star \le n) \qquad \text{[the pathwise bound of the previous display]}$$

$$\mathbb{P}(\tau^\star < \infty) \;\le\; \alpha \qquad \text{[let } n \to \infty\text{]}$$
$$\{\exists t : W^\star \notin \mathcal{C}_t\} = \{\tau^\star < \infty\} \qquad \text{[definition of } \mathcal{C}_t\text{]}$$

which is (7).

Family-wise error. On $\{\tau^\star = \infty\}$, for every t:

$$\begin{aligned} W^\star &\in \mathcal{C}_t && \text{[definition of } \tau^\star \text{ and definition of } \mathcal{C}_t\text{]} \\ (j,l) \in \mathcal{D}_t &\;\Rightarrow\; l \prec_W j \text{ for every } W \in \mathcal{C}_t && \text{[eq. (5)]} \\ &\;\Rightarrow\; l \prec_{W^\star} j && [W^\star \in \mathcal{C}_t] \\ &\;\Rightarrow\; \theta^\star_j > \theta^\star_l && [W^\star = W(\theta^\star)] \end{aligned}$$

So the event in (8) is contained in $\{\tau^\star < \infty\}$, whose probability is at most α by Ville's inequality above.

Random times. Both events are of the form $\{\forall t : \dots\}$, so they imply the corresponding statement at any $\mathbb{N}_0$-valued random time, whether or not it is a stopping time. □

Proof of Proposition B.6. Suppose $\mathcal{C}_t \neq \emptyset$ and let $W \in \mathcal{C}_t$ be arbitrary.

Irreflexive. $\mathcal{D}_t \subseteq \mathcal{P}$, which contains no pair (j,j).

Asymmetric.

$$\begin{aligned} (j,l),(l,j) \in \mathcal{D}_t &\;\Rightarrow\; l \prec_W j \text{ and } j \prec_W l && \text{[eq. (5)]} \\ &\;\Rightarrow\; \text{a contradiction} && [l \prec_W j \text{ means } l \preceq_W j \text{ and not } j \preceq_W l] \end{aligned}$$

Transitive. Let $(j,l),(l,m) \in \mathcal{D}_t$.

$$\begin{aligned} m &\neq j && \text{[asymmetry]} \\ l \prec_W j \text{ and } m &\prec_W l && \text{[eq. (5)]} \\ m &\preceq_W j && [\preceq_W \text{ is transitive}] \\ j \preceq_W m &\text{ is impossible} && [j \preceq_W m \preceq_W l \text{ would give } j \preceq_W l, \text{ contradicting } l \prec_W j] \\ m &\prec_W j && \text{[the two previous lines]} \\ (j,m) &\in \mathcal{D}_t && [W \in \mathcal{C}_t \text{ was arbitrary; eq. (5)}] \end{aligned}$$

Empty confidence set. If $\mathcal{C}_t = \emptyset$:

$$\begin{aligned} \mathcal{D}_t &= \mathcal{P} && \text{[the condition in eq. (5) is vacuous]} \\ \{\exists t : \mathcal{C}_t = \emptyset\} &\subseteq \{\exists t : W^\star \notin \mathcal{C}_t\} && [W^\star \in \mathcal{C}_t \text{ would make } \mathcal{C}_t \text{ nonempty}] \\ \mathbb{P}(\exists t : \mathcal{C}_t = \emptyset) &\;\le\; \alpha && \text{[eq. (7)]} \end{aligned}$$

□

Proof of Proposition B.7. Let $(j,l) \in \mathcal{D}_t^{\text{full}}$ and let W be any weak order with $(j,l) \in T(W)$; put $\mathcal{S} = T(W)$.

$$\begin{aligned} \mathcal{S} \subseteq \mathcal{P} \text{ and } \mathcal{S} &\ni (j,l) && \text{[definition of } T(W) \text{ and the choice of } W] \\ \max_{s \le t} |\mathcal{S}|^{-1} \sum_{(a,b)\in\mathcal{S}} E_s^{ab} &\;\ge\; 1/\alpha && [(j,l) \in \mathcal{D}_t^{\text{full}}\text{: the full closure rejects every such } \mathcal{S}] \\ E_s^W = |\mathcal{S}|^{-1} \sum_{(a,b)\in\mathcal{S}} E_s^{ab} && && \text{[eq. (4) with arithmetic-mean weights]} \\ \max_{s \le t} E_s^W &\;\ge\; 1/\alpha, \text{ so } W \notin \mathcal{C}_t && \text{[the two previous lines; definition of } \mathcal{C}_t] \\ (j,l) &\in \mathcal{D}_t && [W \text{ with } (j,l) \in T(W) \text{ was arbitrary; eq. (5)}] \end{aligned}$$

Example B.11 gives a configuration in which the inclusion is strict; it has been checked by enumerating the 13 weak orders and the 63 subsets. □

Proof of Proposition B.8. Let $(j,l) \in \mathcal{D}_t^{\mathrm{sc}}$ and choose $s \le t$ with $B_s^{jl} \ge M(M-1)/\alpha$. Let W be any weak order with $j \preceq_W l$, that is $(j,l) \in T(W)$, and let $m \in [M] \setminus \{j,l\}$.

Claim: $j \preceq_W m$ or $m \preceq_W l$.

$$\begin{aligned} \text{neither } j \preceq_W m \text{ nor } m \preceq_W l &\Rightarrow m \prec_W j \text{ and } l \prec_W m && \text{[totality of } \preceq_W\text{]} \\ &\Rightarrow l \prec_W j && \text{[transitivity]} \\ &\Rightarrow \text{a contradiction with } j \preceq_W l && [l \prec_W j \text{ means not } j \preceq_W l] \end{aligned}$$

Pooling.

$$T(W) \ni (j,l) \text{ and, for each } m \in [M] \setminus \{j,l\},\ (j,m) \text{ or } (m,l)$$

[definition of $T(W)$ and the claim]

these $M-1$ pairs are distinct

[two could coincide only if $(j,m) = (m',l)$ for some $m, m' \notin \{j,l\}$, which would force $m' = j$]

$$\sum_{(a,b)\in T(W)} E_s^{ab} \ge E_s^{jl} + \sum_{m\in[M]\setminus\{j,l\}} \min\{E_s^{jm}, E_s^{ml}\} = B_s^{jl}$$

[all e-processes are nonnegative; the pair chosen for m contributes at least the minimum]

$$E_s^W \ge \frac{B_s^{jl}}{M(M-1)} \ge \frac{1}{\alpha} \qquad \text{[eq. (4) with } |T(W)| \le M(M-1)\text{; the choice of } s]$$

$$W \notin \mathcal{C}_t \qquad \text{[definition of } \mathcal{C}_t\text{, since } s \le t]$$

$$(j,l) \in \mathcal{D}_t \qquad [W \text{ with } (j,l) \in T(W) \text{ was arbitrary; eq. (5)}]$$

The same time s serves for all W, which is why the running maximum may be taken over B^{jl}. The e-Bonferroni rejections are contained in $\mathcal{D}_t^{\mathrm{sc}}$ because $B_s^{jl} \ge E_s^{jl}$. The cost is one sum of $M-2$ minima per pair. □

C.3 Closure, ranks, tiers and retirement

Proof of Lemma B.9. Work on $\mathcal{G}$ and fix t. Let $(j,l) \in \overline{\mathcal{D}}_t$, joined by the directed path $v_0 = j, v_1, \ldots, v_r = l$ with $(v_{i-1}, v_i) \in \mathcal{D}_t$ for all i.

$$\theta^\star_{v_{i-1}} > \theta^\star_{v_i} \quad \text{for every } i \qquad [(v_{i-1}, v_i) \in \mathcal{D}_t \text{ and definition of } \mathcal{G}]$$

$$\theta^\star_j = \theta^\star_{v_0} > \theta^\star_{v_1} > \cdots > \theta^\star_{v_r} = \theta^\star_l \qquad \text{[chain the previous line]}$$

$\mathcal{D}_t$ has no directed cycle

[a cycle through j is such a path from j to itself and would give $\theta^\star_j > \theta^\star_j$]

$\overline{\mathcal{D}}_t$ is a strict partial order

[transitive by construction; irreflexive because $(j,j) \in \overline{\mathcal{D}}_t$ would be a directed cycle]

the good event of $(\overline{\mathcal{D}}_t)$ contains $\mathcal{G}$

[the second line holds on $\mathcal{G}$ for every t and every $(j,l) \in \overline{\mathcal{D}}_t$]

the good event of $(\overline{\mathcal{D}}_t)$ is contained in $\mathcal{G}$ $\qquad [\mathcal{D}_t \subseteq \overline{\mathcal{D}}_t \text{ for every } t]$

so the two events coincide. □

Proof of Theorem 4.3. (a) On the event in (7), for all j and t:

$$W^\star \in \mathcal{C}_t \qquad \text{[eq. (7)]}$$

$$R_j(\theta^\star) = R_j(W^\star) \qquad \text{[the rank depends on } \theta \text{ only through } W(\theta)]$$

$$\in \mathcal{R}_{j,t} \qquad \text{[definition of } \mathcal{R}_{j,t} \text{ and the first line]}$$

(b) Work on $\mathcal{G}$ and fix j and t.

$$(l,j) \in \mathcal{D}_t \Rightarrow \theta^\star_l > \theta^\star_j \qquad \text{[definition of } \mathcal{G}]$$

$$\#\{l : \theta^\star_l > \theta^\star_j\} \ge \#\{l : (l,j) \in \mathcal{D}_t\} \qquad \text{[previous line]}$$

$$R_j(\theta^\star) \ge 1 + \#\{l : (l,j) \in \mathcal{D}_t\} = L_{j,t} \qquad \text{[eq. (1) and eq. (6)]}$$

$$(j,l) \in \mathcal{D}_t \;\Rightarrow\; l \neq j \text{ and } \theta^\star_l < \theta^\star_j \qquad [\mathcal{D}_t \subseteq \mathcal{P}\text{; definition of } \mathcal{G}]$$
$$\#\{l : \theta^\star_l > \theta^\star_j\} \;\le\; (M-1) - \#\{l : (j,l) \in \mathcal{D}_t\}$$
[each l with $(j,l) \in \mathcal{D}_t$ is one of the $M-1$ other models and is not counted on the left]
$$R_j(\theta^\star) \;\le\; M - \#\{l : (j,l) \in \mathcal{D}_t\} = U_{j,t} \qquad [\text{eq. (1) and eq. (6)}]$$

Ties cause no difficulty: models tied with j are counted on neither side. Since $\mathcal{D}_t \subseteq \overline{\mathcal{D}}_t$ and $\overline{\mathcal{D}}_t$ is certified with the same good event (Lemma B.9), using $\overline{\mathcal{D}}_t$ can only raise L and lower U.

(c) Let $W \in \mathcal{C}_t$; recall $R_j(W) = 1 + \#\{l : j \prec_W l\}$.

$$(l,j) \in \mathcal{D}_t \;\Rightarrow\; j \prec_W l, \text{ so } l \text{ is counted in } R_j(W) \qquad [\text{eq. (5)}]$$
$$R_j(W) \;\ge\; 1 + \#\{l : (l,j) \in \mathcal{D}_t\} = L_{j,t} \qquad [\text{previous line}]$$
$$(j,l) \in \mathcal{D}_t \;\Rightarrow\; l \prec_W j, \text{ so } l \text{ is not counted} \qquad [\text{eq. (5); } l \prec_W j \text{ excludes } j \prec_W l]$$
$$R_j(W) \;\le\; M - \#\{l : (j,l) \in \mathcal{D}_t\} = U_{j,t}$$
[at most the $M-1$ other models are counted, and those with $(j,l) \in \mathcal{D}_t$ are not]

Hence $\mathcal{R}_{j,t} \subseteq [L_{j,t}, U_{j,t}]$.

Strictness. Take $M = 3$, $\alpha = 0.1$ and, as hypothetical values at some time with no earlier crossing, $E^{12} = 10$, $E^{13} = 90$, $E^{21} = 0.2$, $E^{32} = 21$, $E^{23} = E^{31} = 0$. Enumeration shows that exactly four weak orders survive, namely $3 \sim 2 \prec 1$, $3 \prec 2 \sim 1$, $2 \prec 3 \prec 1$ and $3 \prec 2 \prec 1$, writing each from its lowest-ranked model to its highest (so $3 \sim 2 \prec 1$ ranks model 1 first and ties models 2 and 3), and that $\mathcal{D}_t = \{(1,3)\}$. Hence $U_{1,t} = 2$, whereas model 1 has rank 1 in every surviving weak order, so $\mathcal{R}_{1,t} = \{1\}$. The gain comes from the tie convention: H_{12} cannot be rejected while the tie $1 \sim 2$ survives, yet a tie at the top still gives rank 1. Projection sets need not be intervals: with $(E^{12}, E^{13}, E^{21}, E^{23}, E^{31}, E^{32}) = (0, 5, 0, 31, 5, 31)$ and $\alpha = 0.1$, again with no earlier crossing, exactly $1 \sim 2 \prec 3$ and $3 \prec 1 \sim 2$ survive, so $\mathcal{R}_{3,t} = \{1, 3\}$ while $[L_{3,t}, U_{3,t}] = [1, 3]$. □

Proof of Corollary B.10. On $\mathcal{G}$, for all j, k and t at once:

$$j \in I_{k,t} \;\Rightarrow\; R_j(\theta^\star) \;\le\; U_{j,t} \;\le\; k \qquad [\text{Theorem 4.3(b); definition of } I_{k,t}]$$
$$j \in O_{k,t} \;\Rightarrow\; R_j(\theta^\star) \;\ge\; L_{j,t} \;>\; k \qquad [\text{Theorem 4.3(b); definition of } O_{k,t}]$$

□

Proof of Theorem B.13. Let $\mathcal{D}$ be a strict partial order on the finite set $[M]$. For (c) work on $\mathcal{G}$, fix t, and take $\mathcal{D} = \overline{\mathcal{D}}_t$. Part (d) is deterministic and is proved for any nested pair whose closures are strict partial orders.

a chain cannot repeat an element — [a repetition would give $(v,v) \in \mathcal{D}$ by transitivity]
$h(j) \le M-1$ is well defined — [previous line]

(b), first claim. Let $(j,l) \in \mathcal{D}$ and let $l_1, \ldots, l_h$ be a chain above j.

$l_1, \ldots, l_h, j$ is a chain above l — [$(l_h, j) \in \mathcal{D}$ by the chain, $(j,l) \in \mathcal{D}$ by assumption]
$h(l) \;\ge\; h(j) + 1$ — [take $h = h(j)$]
$\mathrm{tier}(j) < \mathrm{tier}(l)$ — [tier $= 1 + h$]

(a).

the tiers partition $[M]$ — [they are the level sets of the function tier]
no two models of one tier are related by $\mathcal{D}$ — [first claim of (b)]
tiers $1, \ldots, \max_j \mathrm{tier}(j)$ are nonempty
[downward induction from the second claim of (b), proved next from the first claim only]

(b), second claim. Let $h(j) = h \ge 1$, with a longest chain $l_1, \ldots, l_h$ above j.

$h(l_h) \;\ge\; h - 1$ — [$l_1, \ldots, l_{h-1}$ is a chain above l_h]
$h(l_h) \;\le\; h - 1$ — [$(l_h, j) \in \mathcal{D}$ and the first claim]

$$\mathrm{tier}(l_h) = \mathrm{tier}(j) - 1, \text{ and } l_h \text{ dominates } j \qquad \text{[the two previous lines; } (l_h, j) \in \mathcal{D}\text{]}$$

(c). Let $l_1, \dots, l_h$ be a longest chain above j, so $h = h(j)$.

$$(l_i, j) \in \mathcal{D} \quad \text{for every } i \qquad \text{[transitivity along the chain]}$$

$$L_{j,t} = 1 + \#\{l : (l, j) \in \mathcal{D}\} \;\geq\; 1 + h = \mathrm{tier}_t(j) \qquad \text{[the } l_i \text{ are distinct]}$$

$$R_j(\theta^\star) \;\geq\; L_{j,t} \qquad \text{[Theorem 4.3(b) applied to } \overline{\mathcal{D}}_t\text{, on } \mathcal{G}\text{]}$$

For the last assertion let j have the largest tier, s; a longest chain above j, followed by j, consists of s models with strictly decreasing $\theta^\star$ on $\mathcal{G}$.

(d).

$$\mathcal{D}_t \subseteq \mathcal{D}_{t+1} \;\Rightarrow\; \overline{\mathcal{D}}_t \subseteq \overline{\mathcal{D}}_{t+1} \qquad \text{[a directed path in } \mathcal{D}_t \text{ is one in } \mathcal{D}_{t+1}\text{]}$$

$$\Rightarrow \text{ every chain above } j \text{ at time } t \text{ is one at time } t+1 \qquad \text{[previous line and the definition of a chain]}$$

$$\Rightarrow\; \mathrm{tier}_t(j) \leq \mathrm{tier}_{t+1}(j) \qquad \text{[}h \text{ cannot decrease, and tier} = 1 + h\text{]}$$

(e). Recall that an antichain is a set of models no two of which are related by $\mathcal{D}$. Let s be the number of tiers.

there is a chain of s elements [a longest chain above a model of maximal tier, followed by that model]

an antichain meets a chain in at most one element [two elements of a chain are related by $\mathcal{D}$]

every partition into antichains has at least s blocks [each block contains at most one element of that chain]

the tiers attain s [they are s antichains by (a); Mirsky (1971)]

□

Proof of Proposition 4.5. Let $(\mathcal{F}_t)$ be the filtration of Section 2, which contains the scores of all models on the items drawn so far, observed or not. Under (F) the order Π is drawn before evaluation, so retirement does not change which item is revealed next.

Induction on t: A_t is $\mathcal{F}_{t-1}$-measurable and $\mathcal{F}_t^{\mathrm{obs}} \subseteq \mathcal{F}_t$.

$$A_1 = [M] \text{ is } \mathcal{F}_0\text{-measurable} \qquad \text{[deterministic]}$$

$$A_s \text{ is } \mathcal{F}_{s-1}\text{-measurable},\; s \leq t \;\Rightarrow\; \mathbf{1}\{j \in A_s\},\; \mathbf{1}\{j \in A_s\}X_{sj} \text{ are } \mathcal{F}_t\text{-measurable},\; s \leq t$$
[induction hypothesis; X_{sj} is $\mathcal{F}_s$-measurable and $\mathcal{F}_s \subseteq \mathcal{F}_t$]

$$\Rightarrow\; \mathcal{F}_t^{\mathrm{obs}} \subseteq \mathcal{F}_t \qquad \text{[}\mathcal{F}_t^{\mathrm{obs}} \text{ is generated by } \xi \text{ and these variables]}$$

$$\Rightarrow\; A_{t+1} \text{ is } \mathcal{F}_t\text{-measurable} \qquad \text{[}A_{t+1} \text{ is a measurable function of them and of } \xi\text{]}$$

Stopping times, bets, frozen wealths.

$$\{\sigma_j \leq t\} = \{j \notin A_{t+1}\} \qquad \text{[the sets } A_t \text{ are nested, so } j \in A_s \text{ exactly for } s \leq \sigma_j\text{]}$$

$$\in \mathcal{F}_t^{\mathrm{obs}} \qquad \text{[}A_{t+1} \text{ is a function of } \xi \text{ and of the scores observed on items } 1, \dots, t\text{]}$$

$$\sigma_j \text{ and } \sigma_j \wedge \sigma_l \text{ are } (\mathcal{F}_t)\text{-stopping times} \qquad \text{[previous line makes them } (\mathcal{F}_t^{\mathrm{obs}})\text{-stopping times, and } \mathcal{F}_t^{\mathrm{obs}} \subseteq \mathcal{F}_t\text{]}$$

every $(\mathcal{F}_t^{\mathrm{obs}})$-predictable bet is $(\mathcal{F}_t)$-predictable [$\mathcal{F}_{t-1}^{\mathrm{obs}} \subseteq \mathcal{F}_{t-1}$]

$$(E_t^{jl}) \text{ is a nonnegative } (\mathcal{F}_t)\text{-supermartingale started at } 1 \qquad \text{[Proposition B.2 with these bets, under any law with } \theta_j \leq \theta_l\text{]}$$

$$(E_{t\wedge\sigma_j\wedge\sigma_l}^{jl}) \text{ is a nonnegative } (\mathcal{F}_t)\text{-supermartingale started at } 1 \qquad \text{[last sentence of Lemma B.1 at the stopping time } \sigma_j \wedge \sigma_l\text{]}$$

Because the sets A_t are nested, both models were scored on every item up to $\sigma_j \wedge \sigma_l$, so S_{t-1} in (12) is computable and the frozen process depends on observed scores only. The analysis runs in the larger filtration; the procedure never needs the unobserved scores. The frozen processes therefore satisfy Assumption 4.1, and every result of Section 4 and Appendix B that rests on it continues to hold. □

Proof of Corollary B.15. Under the all-pairs rule a pair $\{j,l\}$ is frozen at $\sigma_j \wedge \sigma_l$, and σ_l is the first time every pair at l, $\{j,l\}$ included, is resolved (likewise σ_j). This rule depends only on observed scores, as Proposition 4.5 requires. So a pair is frozen only after it is resolved. Until then both models are active, every Z_s^{jl} is observed, and the bet depends on the pair's own past, which is the same with and without retirement. So until resolution the pair's wealths equal those computed without retirement. If $\{j,l\}$ were unresolved at time $\tau^B_{q(j,l)}$ it would be unfrozen, its wealth in the true direction would equal the unretired one, which has crossed $M(M-1)/\alpha$, and e-Bonferroni $\subseteq \mathcal{D}_t$ would resolve it. Hence:

$$\sigma_j \le \max_{l\neq j} \tau^B_{q(j,l)} \qquad \text{[every pair at } j \text{ is resolved by its own } \tau^B_{q(j,l)}\text{, as shown above]}$$

$$\mathbb{P}(\sigma_j > t) \le \textstyle\sum_{l\neq j} \mathbb{P}(\tau^B_{q(j,l)} > t) \le (M-1)e^{-t/\beta_j} \quad \text{for integer } t \ge \kappa_j$$

[union bound; Proposition B.4(c) at each pair, since $t \ge 2\eta/\mu_q$ and $a_q \ge 1/\beta_j$]

$$\mathbb{E}[\sigma_j] \le t_1 + \textstyle\sum_{t\ge t_1}(M-1)e^{-t/\beta_j} \le t_1 + 1 + \beta_j, \quad t_1 = \lceil \max\{\kappa_j, \beta_j \log(M-1)\}\rceil$$

[geometric series, $\sum_{t\ge t_1} e^{-t/\beta_j} \le e^{-t_1/\beta_j}(1+\beta_j)$, and $(M-1)e^{-t_1/\beta_j} \le 1$]

$$\mathbb{E}[\sigma_j] \le \kappa_j + \beta_j\big(\log(M-1)+1\big) + 2 \qquad [t_1 \le \max\{\kappa_j, \beta_j\log(M-1)\}+1]$$

which is (15). Under the conditions of (16), Proposition B.4(d) gives, for every pair q at j, $2\eta/\mu_q \le 11.2\rho_q\eta/\Delta_q^2 \le 11.2\rho_j\eta/\Delta_j^2$ and $1/a_q \le \min\{74.7, 270\rho_q\}/\Delta_q^2 \le \min\{74.7, 270\rho_j\}/\Delta_j^2$, since $\rho_q \le \rho_j$, $\Delta_q \ge \Delta_j$ and $\min\{74.7, 270\rho\}$ is nondecreasing in ρ; summing over j with $\min\{74.7, 270\rho_j\} \le 270\rho_j$ and $\log(M-1) \le \eta$ gives the order statement. On $\mathcal{G}$ no tied pair is ever certified (Theorem 4.2), so a model with a tied partner is never retired. □

C.4 Integer programming and hardness of exact certification

Proof of Proposition A.1. Identify $y \in \{0,1\}^{\mathcal{P}}$ with the relation $\{(a,b) : y_{ab} = 1\}$ on $[M]$, and write $[t] = \{1,\dots,t\}$ for the call times so far. Write $\mathbf{1}\{T(W)\}$ for the indicator vector of $T(W)$, whose entries are $\mathbf{1}\{(a,b) \in T(W)\}$ for $(a,b) \in \mathcal{P}$. A *total preorder* is a reflexive, transitive and total relation, that is, a weak order as defined in Section 3.

$$y \text{ satisfies totality and transitivity} \iff y = \mathbf{1}\{T(W)\} \text{ for a unique weak order } W$$

[a relation is a total preorder iff it is total and transitive; $T(W)$ determines W]

$$E_s^W < 1/\alpha \iff \textstyle\sum_{(a,b)\in T(W)} E_s^{ab} < |T(W)|/\alpha \iff g_s(y) < 0$$

[eq. (4) with arithmetic-mean weights; $|T(W)| = \sum_{ab} y_{ab}$]

$$W \in \mathcal{C}_t \iff \max_{s\in[t]} g_s(y) < 0 \qquad [\mathcal{C}_t \text{ discards } W \text{ at the first call with } E_s^W \ge 1/\alpha]$$

$$(j,l) \in \mathcal{D}_t \iff z^\star([t]) \ge 0$$

[eq. (5): no $W \in \mathcal{C}_t$ has $(j,l) \in T(W)$; the minimum in (10) is over a finite set and is attained]

$$z^\star(S) \le z^\star([t]) \quad \text{for every } S \subseteq [t] \qquad \text{[for each } y \text{ a maximum over fewer times is no larger]}$$

The algorithm maintains three invariants at every call time t. First, every order in the pool lies in $\mathcal{C}_t$. This holds because an order enters the pool only after passing every survival constraint and leaves at the first call at which it fails one, and $\mathcal{C}_t$ is decreasing in t. Second, a pair covered by the pool is not in $\mathcal{D}_t$, because the covering order has $\max_{s\in[t]} g_s < 0$. Third, a pair added to $\mathcal{D}$ is in $\mathcal{D}_t$, by Proposition B.8 when it is added through B_t^{jl}, and by the last two lines above when it is added because $z^\star(S \cup \{t\}) \ge 0$. Every pair not in $\mathcal{D}$ is covered when the call ends. Every pair in $\mathcal{D}$ stays in $\mathcal{D}_{t'}$ for $t' \ge t$ because $\mathcal{D}_{t'}$ is monotone, so never revisiting a certified pair loses nothing. Hence the returned set is $\mathcal{D}_t$. Each pass through the loop for one pair either stops or adds to S a call time at which the minimizer W has $g_s \ge 0$. That time is not in $S \cup \{t\}$, because $\max_{s\in S\cup\{t\}} g_s(y) = z^\star(S\cup\{t\}) < 0$ for the minimizer. Each failed pass therefore adds a new time from $[t-1]$, so the loop stops after at most t solves. □

Proof of Proposition A.2. The reduction is from the feedback arc set problem in tournaments. Given a tournament G on $[M]$, one arc between every two vertices, and an integer K, the problem is to decide whether some linear ordering of $[M]$ has at most K *backward* arcs, arcs $a \to b$ with b placed above a. This problem is NP-hard (Alon, 2006; Charbit et al., 2007). Write $C = \binom{M}{2}$.

An ordering and its reverse have C backward arcs between them, so some ordering has at most $\lfloor C/2 \rfloor$, and instances with $K \geq \lfloor C/2 \rfloor$ are trivial. The problem therefore stays NP-hard with $1 \leq K \leq \lfloor C/2 \rfloor - 1$, which we assume.

Construction. Take $M' = M + 2$ models, the vertices of G and two more, j and l. Let $C' = \binom{M'}{2}$, let ν be the smallest integer with $2^\nu \geq C'$, and set $c = 2^\nu(K + \frac{1}{2} + M)/C'$ and $\alpha = 1/c$, which lies in $(0, 1)$ since $2^\nu \geq C'$ and $K + \frac{1}{2} + M > 1$. The wealth matrix is: for vertices $a \neq b$, $E^{ab} = 2^\nu$ if $a \to b$ is an arc of G and $E^{ab} = 0$ otherwise; $E^{jl} = 0$ and $E^{lj} = 2^\nu$; and $E^{ja} = E^{aj} = E^{la} = E^{al} = 2^{\nu-1}$ for every vertex a. Recall that $(a, b) \in T(W)$, that is $a \preceq_W b$, reads "b at least as good as a", so an arc $a \to b$ contributes 2^ν exactly when W places b at least as high as a. The instance has $M'(M' - 1)$ wealth entries in $\{0, 2^{\nu-1}, 2^\nu\}$ with $2^\nu < 2C'$ and a rational α with numerator and denominator below $2C'(K + M + 1)$, all of $O(\log M)$ bits, so it is constructed in polynomial time. We show that $(j, l) \notin \mathcal{D}_1$ if and only if G has an ordering with at most K backward arcs. Since the feedback arc set problem is NP-hard, deciding $(j, l) \in \mathcal{D}_1$ is coNP-hard.

If. Let an ordering of $[M]$ have $B \leq K$ backward arcs and let W rank the vertices by it, all strictly, with j strictly below every vertex and l strictly above.

$$\textstyle\sum_{(a,b)\in T(W)} E^{ab} = 2^\nu B + M \cdot 2^{\nu-1} + M \cdot 2^{\nu-1} + 0 = 2^\nu(B + M)$$

[backward arcs; one of $(j, a), (a, j)$ per vertex, likewise for l; $(j, l) \in T(W)$ costs $E^{jl} = 0$]

$$E^W = 2^\nu(B + M)/C' \;\leq\; 2^\nu(K + M)/C' \;<\; c \quad [|T(W)| = C', \text{no ties}; B \leq K]$$

so W survives, $(j, l) \in T(W)$, and $(j, l) \notin \mathcal{D}_1$.

Only if. Let W survive with $(j, l) \in T(W)$. Ties are allowed, so let B be the number of backward arcs between vertices that W orders strictly, n_v the number of tied pairs of vertices, and n_{tie} the number of tied pairs of all kinds, so that $|T(W)| = C' + n_{\text{tie}}$. A tied pair of vertices costs $2^\nu + 0$. A vertex strictly ordered against j costs $2^{\nu-1}$ and one tied with j costs 2^ν, likewise for l. The pair $\{j, l\}$ costs 0 if $j \prec_W l$ and 2^ν if the two are tied. Hence

$$\textstyle\sum_{(a,b)\in T(W)} E^{ab} \;\geq\; 2^\nu\big(B + n_v\big) + 2^\nu M + 2^{\nu-1}(n_{\text{tie}} - n_v)$$

[every tie outside the vertex pairs adds at least $2^{\nu-1}$ to the strict baseline $2^\nu(B + M)$]

$$\frac{2^\nu(B + M + \frac{n_{\text{tie}}}{2} + \frac{n_v}{2})}{C' + n_{\text{tie}}} < c = \frac{2^\nu(K + \frac{1}{2} + M)}{C'} \quad [W \text{ survives}; |T(W)| = C' + n_{\text{tie}}]$$

$$B + \tfrac{n_v}{2} + n_{\text{tie}}\Big(\tfrac{1}{2} - \tfrac{K+\frac{1}{2}+M}{C'}\Big) < K + \tfrac{1}{2} \quad \text{[multiply out and cancel } 2^\nu M]$$

$$K + \tfrac{1}{2} + M \;\leq\; \tfrac{M(M-1)}{4} + M - \tfrac{1}{2} < \tfrac{(M+2)(M+1)}{4} = \tfrac{C'}{2} \;\; [K \leq \lfloor C/2 \rfloor - 1 \leq M(M-1)/4 - 1]$$

$$B + \tfrac{n_v}{2} < K + \tfrac{1}{2} \quad \text{[the coefficient of } n_{\text{tie}} \text{ is positive]}$$

Now refine W to a linear ordering of the vertices: keep the strict comparisons and order each class of tied vertices so that at most half of the arcs inside it are backward. This is possible because an ordering of the class and its reverse have every internal arc backward exactly once. The arcs inside the classes number n_v, so the ordering has at most $B + n_v/2 < K + \frac{1}{2}$ backward arcs, hence at most K. Allowing ties therefore cannot help an ordering survive.

Realizability. With $N = \infty$ the factor of the update (11) is $1 + \lambda Z$ with $Z \in \{-1, 0, 1\}$ for binary scores, so with bets in $\{0, 1\}$ every factor is 0, 1 or 2 and every wealth is 0 or a power of two. Give each ordered pair (a, b) its own items, on which every other pair bets 0. With bet 1, k items on which model a scores 1 and model b scores 0 make $E^{ab} = 2^k$. With bet 1, one item on which a scores 0 and b scores 1 makes $E^{ab} = 0$. At most $(\nu + 1)M'(M' - 1)$ items are needed, the schedule is fixed in advance, hence predictable, and the certifier is called once, after the last item. The default mixture bet is not covered: its factors are at least $\frac{1}{2}$, so its wealths never vanish. □

D SCOPE OF THE GUARANTEE AND OPEN PROBLEMS

Every result stated in this paper is proved in Appendix C. Four kinds of machine check accompany the proofs but do not replace them. The counts of weak orders, 13 of 63 intersections for $M = 3$

and 75 of 4095 for $M = 4$, are verified by exhaustive enumeration, as is the three-model example of Appendix B. That the exact test certifies everything the full closure certifies is checked at $M = 3$ by enumerating all 63 subsets for each of many random wealth matrices. The remaining containments, between e-Bonferroni, the shortcut and the exact test, are checked on random wealth matrices at $M = 3$, 4 and 5. The fourth averages the finite-benchmark wealth exactly over all item orders and confirms that its expectation is at most one at every time. These are implementation checks on finitely many inputs.

What the guarantee does not cover. The method needs a scalar ability per model, and pairwise-battle data do not supply one. In battle data a platform picks two models, both answer the same prompt, and a judge picks the winner. The natural parameter is then $p_{jl} = \mathbb{P}(j \text{ beats } l)$, and "$p_{jl} > \frac{1}{2}$" is not transitive in general: j may usually beat l, and l usually beat m, while m usually beats j. Lemma B.9 and the rank (1) rest on transitivity of $>$ on $\mathbb{R}$. The method also needs the set of models fixed before any data arrive, because a new entrant enlarges $\mathcal{W}$ and changes every $T(W)$. Models arriving over time would require online closed testing, which admits new hypotheses as they arrive (Fischer & Ramdas, 2024), or an error budget split in advance among future entrants. Finally, the method needs the item order not to depend on the data. Lemma B.1 allows a pair to be frozen at a stopping time under both sampling models, but not the choice of *which item* to evaluate next: under (S) that breaks the independence of X_t from $\mathcal{F}_{t-1}$, and under (F) it breaks the uniformity of $\Pi(t)$. The inverse-probability-weighted signal of Zhou et al. (2026) is a natural repair and is not worked out here. Pausing a pair and resuming it is covered under (F) only if the skipped items are scored retroactively before its wealth is updated again, since S_{t-1} in (12) needs every revealed difference. Scoring the skipped items forgoes the saving that pausing was meant to buy.

What is open inside it. Deciding membership in $\mathcal{D}_t$ is coNP-hard over the wealth matrices that the update (11) can produce with some predictable bets (Proposition A.2), by a reduction that uses bets in $\{0, 1\}$. The default mixture bet keeps every wealth strictly positive, and whether it can produce such hard instances is unknown. Nothing in this paper depends on the answer: Algorithm 3 solved every instance in our experiments, and Algorithm 2 is the polynomial fallback. On the power side, Proposition B.4 and Corollary B.15 give consistency, certification times and the expected cost of retirement. That cost bound covers the all-pairs rule under (S). It does not cover the top-k rule behind the compute savings we report, under which a pair can be frozen before it is certified. Validity under either rule follows from Proposition 4.5, which needs no such bound. What remains open is a rate under (F), a cost bound for the top-k rule, and a rate for the exact test sharper than the bound it inherits from e-Bonferroni.

E Experiments

The data, the designs and the definition of an error are common to every experiment and are given first. Each subsection after that asks one question and answers it. Appendices E.1 to E.4 carry the numbers reported in Section 5, in the order in which it reports them, and Appendices E.5 and E.6 add experiments that Section 5 does not report.

The real leaderboard. The data are the per-item results of 395 Open LLM Leaderboard models, as released with tinyBenchmarks (Maia Polo et al., 2024; Beeching et al., 2023), on six benchmarks: MMLU (Hendrycks et al., 2021) (14,042 items), HellaSwag (Zellers et al., 2019) (10,042), GSM8K (Cobbe et al., 2021) (1319), WinoGrande (Sakaguchi et al., 2020) (1267), ARC-Challenge (Clark et al., 2018) (1172) and TruthfulQA (Lin et al., 2022) (817, whose per-item scores can take any value in $[0, 1]$). The estimand is (F): the truth is the ranking by mean score on the whole benchmark, so every certified dominance can be checked against it. Three model subsets are used, the 20 and the 8 highest-scoring models by mean score over the six benchmarks and the *spread subset* of 20 models evenly spaced through the ranking of all 395, together with all 395 released models with no further selection. Items are revealed in a uniformly random order, redrawn 50 times per setting, and the procedure reports every 1% of the benchmark, with checkpoints at fractions 0.05, 0.10, 0.25, 0.50, 0.75 and 1.00. Throughout, $\alpha = 0.05$, the bets are the mixture grid of Section 3, and the width of a rank interval is $U_{j,t} - L_{j,t}$, zero for a model whose rank is resolved. Certification uses the shortcut for the twenty-model subsets and for all 395, and enumeration for the eight-model subset.

The simulations. Two simulated designs recur. The first streams i.i.d. items under (S), and the models share each item's difficulty, so that they are positively dependent. Item t has difficulty $d_t \sim N(0,1)$, and model j, with logistic ability ω_j, answers it correctly with probability $1/(1+e^{-(\omega_j - d_t)})$, independently of the other models given d_t. Its six models have logistic abilities $(0.5, \ldots, 0.5)$, all tied at accuracy 0.602; $(1, 1, 0.6, 0.6, 0.2, -0.2)$, two tied pairs; $(1, 0.98, 0.62, 0.60, 0.22, 0.20)$, near ties; $(1, 0.8, 0.6, 0.4, 0.2, 0)$, evenly spread; or, in the bet comparison of Appendix E.6, $(0.5, 0.4, 0.3, 0.2, 0.1, 0)$, a close race with gaps of about 0.02. True accuracies lie between 0.46 and 0.70. This design is used wherever a known tie is needed, because ties are where a multiple-comparison procedure is tested hardest. The second design is (F). It draws one benchmark of $N = 5000$ items from the same logistic model, and the truth is each model's accuracy on those items. It has either 20 models with accuracies from 0.40 to 0.78 and adjacent gaps between 0.016 and 0.022, or 6 models from 0.50 to 0.70 with gaps of about 0.04, or 6 models whose two leaders have equal logistic ability. Replication counts are given with each experiment, and shaded bands in the figures are two standard errors.

What counts as an error. A run counts as wrong if any report it issues, at any monitored time, certifies a dominance that is false. A rank interval that misses a true rank implies such a dominance, by Theorem 4.3(b), so this single count covers the dominances and the intervals together; the exact rank sets $\mathcal{R}_j$ of (5) are not scored separately. It is the error rate faced by an analyst who is free to stop at any look, which is the quantity the guarantee bounds.

The fixed-sample competitors. Three constructions stand in for current practice. Holm rank sets apply Holm's step-down correction (Holm, 1979) to the $M(M-1)$ one-sided paired tests and convert the rejections into rank intervals by the counting formula (6). A *step-down* correction tests the hypotheses in order of their evidence and relaxes the threshold as hypotheses are rejected. Holm rank sets are the simultaneous rank intervals of Neuhof & Benjamini (2024), which Neuhof & Benjamini (2026a) also compute for a leaderboard. Here the paired test is either a paired z-test or an exact one-sided McNemar test (McNemar, 1947), which is a sign test on the items where the two models' scores differ. Romano–Wolf step-down rank sets (Romano & Wolf, 2005; Mogstad et al., 2024) set the threshold by a bootstrap of the studentized maximum, the largest over pairs of the mean paired difference divided by its standard error. The bootstrap is a centered multiplier bootstrap with 200 draws, which reweights the centered per-item differences by independent random multipliers. All are recomputed at each look where they are monitored.

Implementation. The implementation, a few hundred lines of Python, follows Section 3 alone. Its automated tests run the machine checks listed in Appendix D. A separate simulation shows that the bound of Theorem 4.2 is nearly attained: with four tied models of accuracy 0.5 scored independently, 400 i.i.d. items under (S) checked after every item, a fixed bet of 0.5 and $\alpha = 0.2$, the true weak order is rejected at some time in 18.4% of 4000 runs, with a standard error of 0.6%.

E.1 Does monitoring break a fixed-sample rank set, and does ours survive it?

Setup. Six models of the first simulated design, all tied, with tied pairs at ranks 1 and 3, or with near ties, are evaluated on 2000 items, and every procedure sees the same data. The fixed-sample rank sets are recomputed at each of a number of equally spaced looks, which is the horizontal axis of Figure 1. Ours is checked at the same looks, although its guarantee would also cover a check after every item. There are 5000 replications.

Monitoring destroys the guarantee of a fixed-sample rank set and leaves ours intact. At 200 looks the fixed-sample sets make a false statement in 30% of runs with the exact McNemar test and 51% with the paired z-test, against 1.6% for our exact test. At a single pre-planned look the same two procedures are valid, at 3.3% and 3.6%, so what fails is the repeated look and not the normal approximation. The effect does not need exact ties. With tied pairs at ranks 1 and 3 the figures are 9.7% and 18% against our 0.3%, and when no two models are exactly tied but some gaps are 0.004 they are 4.6% and 11.7% against our 0.2%.

Table 2: All 395 released models, no further selection, 50 item orders per benchmark. True dominances certified, as a percentage of those that hold on the whole benchmark; tiers at the end. Holm uses the paired z-test.

		0.25		0.50		1.00		tiers
Benchmark	items	ours	Holm	ours	Holm	ours	Holm	ours / Holm
MMLU	14,042	76	77	86	83	99.6	88	225 / 26
HellaSwag	10,042	62	63	77	74	99.2	81	165 / 24
GSM8K	1319	63	66	78	75	98.9	83	124 / 14
WinoGrande	1267	5	6	26	17	95.5	34	53 / 5
ARC-Challenge	1172	18	22	48	39	97.0	56	68 / 8
TruthfulQA	817	19	35	55	53	97.0	66	69 / 13

E.2 How much of a real leaderboard can be certified?

Setup. The real leaderboard described above, in two forms: the three selected subsets of twenty and eight models, and all 395 released models with no further selection. Selecting the twenty highest-scoring models is legitimate under (F), because the score matrix is fixed before the item order is drawn, but it is not a selection an operator could make prospectively, which is why the unselected run matters.

Half a benchmark certifies up to half of the true dominances among close models, and 26% to 86% of them among all 395 models. At half of MMLU the twenty highest-scoring models fall into 3.0 certified tiers with 50% of the 190 true dominances certified and a mean rank-interval width of 9.5, while the spread subset reaches 11.2 tiers. At three quarters of MMLU ours certifies 63% of the top-twenty true dominances against 51% for Holm and 54% for Romano–Wolf, with 4.9 tiers against 3.3 and 3.6. Across the 900 runs that the three subsets, six benchmarks and 50 orders make up, exactly one report was false, in one item order of the HellaSwag eight-model set, well below $\alpha = 5\%$.

The whole leaderboard. Removing the selection makes certification easier, even though the correction grows. With all 395 models, 86% of the 77,802 true dominances of MMLU are certified at half the benchmark and 99.6% at the end, against 83% and 88% for Holm, and the certified partial order has 225 tiers against Holm's 26. Table 2 gives the six benchmarks; WinoGrande is the hard case, at 26% of true dominances at half. Ours never certified a false dominance in 300 runs monitored at about a hundred looks each, and neither did the Holm sets, which use the paired z-test and are recomputed at the six checkpoints. With 395 models the shortcut threshold is $M(M-1)/\alpha = 3{,}112{,}600$, but the correction grows only as $2\log M$ while the number of easy pairs grows as M^2. These percentages come from the shortcut, so by Proposition B.8 the exact test would certify at least as many. Romano–Wolf is omitted here because its bootstrap over 155,630 pairs is not affordable at this M.

Why the end of a benchmark flatters our sets. The large gap at fraction 1.0 in Figure 2 is not a power difference. Under (F) a pair is resolved with certainty once the unrevealed items can no longer overturn it, and the offset (12) captures most of this. The fixed-sample constructions instead treat the benchmark as a sample. They therefore certify only 55–58% of the top-twenty true dominances on MMLU at the point where the ranking is in fact known exactly. Our own sets fall short of the full ranking for pairs whose margin on the whole benchmark is one or two items. The bet uses only the null's bound on the conditional mean of the next difference. A null that has become impossible, because the revealed items already settle the pair, is therefore not by itself infinite evidence. After that point the offset sits at its lower clip $-1+\delta$, with $\delta = 0.01$, and each item multiplies the wealth by at most $1 + \lambda(2-\delta)/\delta$. Pairs with a margin of one or two items have very few such items. The fair comparison is therefore at fractions up to one half, where the three procedures are close and the fixed-sample sets are ahead early on.

Table 3: Certifying every model's top-3 status, all 395 models, 50 item orders per benchmark. Cost is model-item evaluations divided by $395N$; settled is the mean fraction of models whose top-3 status is certified at the end.

Benchmark	cost	settled
MMLU	6.2%	100.0%
HellaSwag	15.2%	100.0%
GSM8K	16.7%	99.2%
WinoGrande	47.3%	99.5%
ARC-Challenge	36.9%	99.5%
TruthfulQA	27.3%	97.7%

E.3 How much compute does retirement save?

Setup. Cost is counted in model-item evaluations against a fixed benchmark of $N = 5000$ items, divided by what a full evaluation of every model would cost, with 300 replications and reports every 25 items. Five retirement rules are compared: no retirement; stopping everything once every model's top-k status is certified; retiring a model once its own top-k status is certified, which is the rule of Section 3; the same but only once every pair between the model and a model whose status is still open is certified, so that a retirement can never block another certification; and, for the full ranking, retiring a model once every pair involving it is certified. The group-sequential comparison is a Pocock design (Pocock, 1977), as used by Arviv et al. (2026), who add items in batches of 100. Our implementation compares a z-statistic with one constant critical value at each of ten equally spaced looks fixed in advance, without a multiplicity correction, and a second version adds a Bonferroni correction over the 380 ordered pairs. Both versions close their certifications under transitivity.

Retiring each model as soon as its own question is answered costs about a fifth of a full evaluation, and stopping everything at one moment costs four times as much. Certifying the top-3 status of all 20 simulated models costs 22% of a full run under the retirement rule against 84% under the common stopping rule, with no report ever wrong. With six models and $k = 2$ the three top-k rules cost 44%, 22% and 30%: the cautious fourth rule gained nothing, since the plain rule already reached the goal in every run, as did no retirement at all. Certifying the entire ranking is far more expensive, 51% of a full run for six models and 90% for twenty, because adjacent gaps of about 0.02 are resolved only when the finite benchmark is nearly exhausted. The mixture bet cost less than a fixed bet of 0.25 in every setting but one, for example 51% against 63% for the full ranking of six models. The exception is two leaders of equal logistic ability among six models, where the fixed bet was slightly cheaper under the three top-k rules.

The cost barely depends on k. Between $k = 2$ and $k = 16$ among twenty close models the cost stays at 21–24% of a full run, because it is set by the models adjacent to the kth place and not by k itself. At $k = 1$ and $k = 19$ it falls to 18% and 16%, where only one side of the boundary has to be resolved. Ours reaches the goal in 97.5–99.5% of runs at every k, over 200 replications.

On the real leaderboard. The top-3 goal costs 31% of a full MMLU run for the twenty highest-scoring models and 4% for the spread subset, and 57% of GSM8K and 68% of HellaSwag for the top twenty, reached in 96% to 100% of orders. Where the top twenty are indistinguishable, on ARC-Challenge and WinoGrande, retirement saves almost nothing, as expected. Table 3 shows that applying the same rule to all 395 models costs a smaller fraction still, because most models on the leaderboard are far from the top-3 boundary and retire almost at once, whereas the twenty highest-scoring are near-ties by construction. The comparison is of fractions and not of compute: 6.2% of the 395-model evaluation of MMLU is about four times the absolute cost of 31% of the twenty-model one. Every model's status is settled in 100% of orders on MMLU and HellaSwag. On the other four benchmarks some models near the boundary stay unsettled, about two on ARC-Challenge and WinoGrande, three on GSM8K and nine on TruthfulQA. The all-or-nothing goal is therefore reached in only 0–2% of orders, while 97.7–99.5% of models are settled. No report was wrong in any of these runs.

Table 4: Fixed-sample constructions against ours, twenty highest-scoring models, 50 item orders. Top: true dominances certified at one half, as a percentage of those that exist. Bottom: probability that some report, at any look, was false.

	MMLU	HellaSwag	GSM8K	WinoGrande	ARC-Challenge	TruthfulQA[a]
True dominances certified at one half of the benchmark (%)						
Ours, monitored	50	49	22	2	0	17
Holm, McNemar	47	47	19	1	0	61[b]
Holm, paired z-test	48	48	20	1	0	16
Romano–Wolf	49	49	22	1	0	21
Probability of a false statement at some look						
Ours, monitored	0	0	0	0	0	0
Holm, McNemar	0	0.02	0	0	0	1.00[b]
Holm, paired z-test	0	0.02	0.04	0	0.02	0
Romano–Wolf	0.02	0.02	0.02	0.04	0.02	0

[a]Per-item scores in $[0, 1]$, not binary. [b]The sign test of McNemar's procedure targets the median of the paired differences, which is not the mean-score estimand for non-binary scores. Its certifications on TruthfulQA are not comparable and every order contained a false one. This is a test mismatch, not a monitoring effect.

Against a group-sequential design. The Pocock design is as cheap as ours and answers the question far less often. With the twenty simulated models it reaches the goal in 56% of runs against our 99%, because a z-test cannot use the fact that the benchmark is finite and its looks must be fixed in advance. With no multiplicity correction and only asymptotic validity, it also errs in 2.7% of runs at that cost, and in 3.7% without retirement. The same holds across the sweep over k, where the uncorrected design reaches the goal in 56–68% of runs with a false statement in 3–4%. A Bonferroni correction supplies the missing multiplicity correction but removes the saving. The corrected version costs 32% and reaches the goal in 3% of runs. With two leaders of equal logistic ability among six models, whose accuracies on the realized benchmark differ only by chance, the uncorrected design errs in 1% of runs, and in 2% without retirement, while ours errs in none, at an equal cost of 12%. All costs are averaged over every run, whether the goal was reached or not.

E.4 What does anytime validity cost?

Setup. The comparison is made at the one look where a fixed-sample procedure is valid, so that it measures only the power given up for the right to monitor. In simulation the first design runs to 2000 items, with either Holm construction run once and our exact test monitored at 200 looks. On the real leaderboard the three fixed-sample constructions are compared with ours on the twenty highest-scoring models, with Romano–Wolf computed at the six checkpoints only, since its bootstrap is too slow for every look.

In simulation the price of anytime validity is a fraction of one dominance. After 2000 items with two tied pairs, both Holm constructions certify 12.98 of 13 true dominances and ours 12.79; with near ties the figures are 11.97 and 11.62 of 15. On the real leaderboard the fixed-sample sets lead by up to eight percentage points at a quarter of a benchmark, and by half that lead is gone: ours matches or exceeds Romano–Wolf on five of the six benchmarks and is behind by four points on TruthfulQA. Table 4 shows Romano–Wolf to be the strongest of the three. Beyond one half ours pulls ahead for a reason unrelated to power, given in Appendix E.2. This cost buys the right to look: the same fixed-sample sets, once monitored, err in up to 4% of orders (McNemar on TruthfulQA aside), and ours in none. The monitoring penalty is smaller here than in the tied simulations because the gaps among the top twenty are mostly nonzero.

E.5 Does the guarantee depend on how model errors are correlated?

Setup. Each regime holds every model's marginal accuracy fixed and varies only the dependence among its item-level scores, so that any difference between regimes is attributable to dependence alone. For six models and 4000 i.i.d. items under (S), with the accuracies of the first simulated design as marginals, binary scores come from a Gaussian copula, $X_{tj} = \mathbf{1}\{a_j\sqrt{\gamma}\,G_t + \sqrt{1-\gamma}\,\varepsilon_{tj} \leq$

Table 5: Four dependence regimes with identical marginal accuracies. "Anytime error" uses six tied models; the other columns use evenly spread logistic abilities after half the items, with the probability of reaching the top-3 goal in parentheses.

Dependence	Anytime error	True dominances	Mean rank width	Tiers	Top-3 cost (goal)
Independent	2.0%	70%	1.49	3.3	41% (49%)
Positive, $\gamma = 0.5$	1.5%	76%	1.18	3.7	32% (79%)
Positive, $\gamma = 0.9$	1.6%	96%	0.22	5.3	16% (100%)
Mixed sign, $\gamma = 0.8$	1.6%	80%	1.01	3.9	46% (26%)

$\Phi^{-1}(\theta_j)\}$ with G_t and ε_{tj} independent standard normal, Φ the standard normal distribution function, $\gamma \in [0, 1)$ the weight of the shared factor G_t and $a_j \in \{-1, 1\}$ the sign with which model j loads on it. This gives $\mathbb{P}(X_{tj} = 1) = \theta_j$ for every γ and every choice of signs. The four regimes are independence ($\gamma = 0$), positive equicorrelation at $\gamma = 0.5$ and $\gamma = 0.9$ with every $a_j = 1$, and a mixed-sign regime at $\gamma = 0.8$ in which three models load positively and three negatively; the resulting binary correlations are about 0, 0.32, 0.70 and ± 0.55. Monitoring is every 20 items, with the mixture bet and the exact certifier, 5000 replications for the tied configurations and 2000 for the others, and the same code runs unchanged in every regime.

Validity is unchanged across dependence structures and speed is not. With six tied models the anytime false-statement probability is 2.0%, 1.5%, 1.6% and 1.6% across the four regimes, and with two tied pairs it is 0.6%, 0.4%, 0.4% and 0.5%, all at $\alpha = 0.05$. Table 5 collects the four regimes. Speed depends strongly on the regime. With evenly spread logistic abilities at half the items the procedure certifies 70% of the true dominances under independence, with a mean rank-interval width of 1.49 and 3.3 tiers, against 96%, 0.22 and 5.3 under strong positive dependence, where certifying every model's top-3 status costs 16% of a full evaluation instead of 41%. The mixed-sign regime is less favorable by design, since pairs within a group benefit from positive dependence while pairs across groups are negatively dependent and harder to resolve: it certifies more true dominances than independence at the halfway point but its top-3 cost is 46% and the goal is reached in only 26% of runs. With near ties, independence and the two positive regimes certify 74%, 79% and 80% at half. Here 80% is a practical ceiling, because the three near-tied pairs, with gaps of about 0.004, are almost never resolved within 4000 items. The mixed-sign regime is slightly slower than independence, at 72%. The regimes differ in the variance of the paired difference, $\mathrm{Var}(Z_t^{jl}) = \mathrm{Var}(X_{tj}) + \mathrm{Var}(X_{tl}) - 2\,\mathrm{Cov}(X_{tj}, X_{tl})$: positive covariance makes the paired signal less noisy and negative covariance more noisy, so a pair with margin 0.075 has paired-difference variance about 0.45, 0.30 and 0.14 under independence, $\gamma = 0.5$ and $\gamma = 0.9$, and its median first-certification time falls from 28% to 18% to 9% of the items.

E.6 What does the multiplicity correction contribute?

Setup. Every certifier receives the same wealth paths, so only the multiplicity correction varies. Four certifiers are compared: e-Bonferroni with closure, which is the certification logic of Gu et al. (2026); the shortcut of Algorithm 2; the exact test of Algorithm 1; and a variant of the exact test whose weights in (4) are uniform over the pairs of $T(W)$ in the same or in adjacent levels of W, not over all of them. The simulated comparison uses six models over 4000 items monitored every 20, with 2000 replications; every error rate was at most 1.9%. The real comparison reuses the wealths of Appendix E.2 unchanged.

The exact test certifies more pairs than e-Bonferroni, mostly early in evaluation, and never fewer. On simulated data the exact test certifies about 7% more true dominances than e-Bonferroni after 500 items and at most 3% more by 4000, because a threshold enters the sample size an e-process needs only through its logarithm: halving the multiplicity factor saves $\log 2$ divided by the growth rate, a fixed number of items. The counts are 5.10 against 4.77 of 15 true dominances after 500 items with near ties and 5.81 against 5.45 with evenly spread logistic abilities, becoming 12.01 against 12.00 and 13.23 against 12.90, respectively, by 4000. On the real wealths the exact test certifies 2–5% more than e-Bonferroni among the eight highest-scoring models up to three quarters of a benchmark and nearly the same number at the end, ahead at 20% of the looks and behind at

Table 6: True dominances certified from identical wealths, so that only the correction differs; means over 50 item orders and the six benchmarks, and for twenty models also over the top and spread subsets. The last two columns count the looks at which the row beat or trailed the reference.

	fraction of the benchmark						looks vs. e-Bonferroni	
	0.05	0.10	0.25	0.50	0.75	1.00	ahead	behind
Eight highest-scoring models, 27.5 *true dominances,* 31,600 *looks*								
e-Bonferroni with closure (reference)	2.15	3.11	4.86	7.51	12.62	26.04	reference	
Shortcut (Algorithm 2)	2.15	3.11	4.86	7.51	12.63	26.04	118	0
Exact (Algorithm 1)	2.19	3.21	4.99	7.90	13.02	26.11	6,320	0
Exact, adjacent-level weights	2.04	2.35	3.90	4.97	7.87	26.20	831	15,052
Twenty models, top and spread subsets pooled, 187.8 *true dominances,* 63,200 *looks*								
e-Bonferroni with closure (reference)	27.0	37.9	64.5	89.5	110.7	177.6	reference	
Shortcut (Algorithm 2)	27.0	37.9	64.5	89.5	110.7	177.6	473	0

none. Table 6 gives the four certifiers side by side. The shortcut and e-Bonferroni with closure are nearly identical in these experiments. Here transitivity pooling guarantees that nothing is lost but brings no measurable gain in power. Over 600 further simulated runs the shortcut was ahead at 121 of 120,000 looks and behind at none. Adjacent-level weights are valid by Remark B.12 but lose most of the power early, certifying 0.3 and 0.4 pairs at 500 items, because for a strict order they average only the $M - 1$ adjacent pairs and give zero weight to the non-adjacent pairs, whose large wealths are what eliminates orderings far from the truth.

The exact test and its adjacent-level variant enumerate every weak order, 545,835 of them for eight models but about 2.7×10^{21} for twenty, which is why only the two closed-form certifiers appear in the lower half of Table 6.

Beyond eight models, by integer programming. Computed by integer programming, the exact test certifies only slightly more than the shortcut. On the twelve real settings at $M = 20$ the shortcut trails Algorithm 3 by 1.4 true dominances at half a benchmark, with a range of 0.1 to 2.9, and by 0.3 at the end. Table 7 reports every setting. The shortcut is never ahead, as Proposition B.8 requires. The gap is largest while certification is in progress, as on the spread subsets at one half, and nearly vanishes once the shortcut has certified nearly everything, as at the end of every spread subset, or while nothing is certified yet, as on ARC-Challenge and WinoGrande among the top twenty. Neither certifier was ever wrong, and Algorithm 3 solved every instance. Cost is modest: a run with looks every 1% of a benchmark needs about 294 integer programs at 26 ms each, eight seconds in all, on one CPU core with HiGHS through SciPy and the relative optimality gap set to zero. On simulated runs of 3000 items with looks every 20, the times are about 0.2, 15 and 63 seconds at $M = 10$, 20 and 30, with 85, 388 and 935 programs; there the exact test and the shortcut certify the same 44.7 of 45, 189.3 of 190 and 431.3 of 435 true dominances.

The bet matters more than the correction. A bet with negative growth can keep a pair from ever being certified, whatever the correction. A fixed bet of 0.25 never certifies pairs whose gap is about 0.02: with six models at accuracies 0.50 to 0.60 it certifies 6.5 of 15 true dominances after 5000 items and 7.0 after 20,000, against 14.0 for the mixture bet, 13.5 for the plug-in bet and 13.0 for online Newton steps. This is the negative-growth effect of Remark B.3, which no multiplicity correction can undo. The price of adapting is paid early and only where the fixed bet happens to be well tuned: with accuracies evenly spaced from 0.50 to 0.70, where 0.25 is close to the log-optimal bet $\lambda^\star = \Delta/\rho$, the fixed bet leads after 500 items, 6.9 pairs against 5.8, but the mixture overtakes it by 2000 and leads clearly by 4000, 13.2 against 11.1. The mixture was the best adaptive bet at every checkpoint of every setting, which is why it is the default. These runs use 1000 replications, or 300 for the close race at 20,000 items, and every error rate was at most 3.2%.

F Related work in detail

Table 8 places the closest methods side by side; the rest of this appendix says in words what the columns compress, and then relates the construction to the statistical tools on which it builds.

Table 7: The exact test by integer programming against the shortcut, at $M = 20$ with 50 item orders per row, drawn independently of those behind Table 1, so the shortcut columns differ slightly from it. True dominances certified, rank width at one half, and what the exact test cost to run.

		0.25		0.50		1.00		width at 0.50		cost	
Benchmark, subset	true	exact	shortcut	exact	shortcut	exact	shortcut	exact	shortcut	programs	s
MMLU, top	190	84.1	84.1	93.6	92.4	187.3	187.1	9.64	9.76	244	6
HellaSwag, top	190	77.4	76.3	94.9	94.1	181.7	181.3	9.51	9.59	292	8
GSM8K, top	190	14.4	13.4	45.1	43.1	183.2	182.8	14.49	14.69	331	9
WinoGrande, top	180	0.1	0.1	3.3	3.1	153.9	153.2	18.67	18.69	238	6
ARC-Challenge, top	179	0.0	0.0	0.3	0.2	148.2	147.1	18.97	18.98	213	6
TruthfulQA, top	190	1.0	0.6	33.2	30.9	156.0	155.6	15.68	15.91	285	7
MMLU, spread	190	165.5	164.7	173.3	172.9	189.6	189.5	1.67	1.71	217	6
HellaSwag, spread	190	118.7	116.8	146.9	145.5	187.5	187.5	4.31	4.45	304	8
GSM8K, spread	190	139.2	136.9	162.7	161.3	189.4	189.3	2.73	2.87	308	8
WinoGrande, spread	186	42.9	40.6	92.6	89.8	184.0	183.8	9.74	10.02	360	10
ARC-Challenge, spread	189	70.2	67.2	116.3	114.3	188.8	188.8	7.37	7.57	358	9
TruthfulQA, spread	190	79.6	76.3	124.7	122.6	185.7	185.4	6.53	6.74	373	10

Table 8: Closest existing methods. Monitoring: valid under continuous looks and any stopping rule. Finite-sample: no asymptotics. Structure: the correction uses the relations among pairs.

	Rank sets, simultaneous	Monitoring	Finite-sample	Structure	Early stop of evaluation
Fixed-sample rank inference					
Mogstad et al. (2024)	yes	no	no	no	no
Al Mohamad et al. (2022)	yes	no	yes	no	no
Al Mohamad et al. (2021)	yes	no	yes	yes	no
Neuhof & Benjamini (2024)	yes	no	yes[a]	no	no
Neuhof & Benjamini (2026a)	yes	no	yes[a]	no	no
Neuhof & Benjamini (2026b)	no	no	yes[a]	no	no
Li et al. (2026); Avelar Menendez et al. (2026)	yes	no	no	no	no
Chandra & Xie (2025)	yes	no	yes (simulable noise)	n/a	no
Sequential evaluation					
Zhou et al. (2026); Hsu & Shekhar (2026)	no (one model)	yes	yes	n/a	yes
Kotawala (2026)	no (pairs)	yes	yes	no correction[b]	yes (pairwise)
Gu et al. (2026), battles	yes	yes	yes	no (Bonferroni)	yes (pairwise)
Arnold et al. (2026)	no (superior set)	yes	yes	no (generic closure)	no
Arviv et al. (2026)	no (pairs)	pre-set looks	no	no correction	yes
Heckel et al. (2019)	no (partition at stopping)	own rule only	yes	no (union bound)	yes
This paper	yes	yes	yes	yes	yes

[a]Given finite-sample pairwise tests. [b]Bonferroni and Holm only in its separate fixed-sample analysis.

Fixed-sample rank sets are valid for one look at one pre-chosen sample size, and the constructions of Mogstad et al. (2024), Al Mohamad et al. (2022), Neuhof & Benjamini (2024) and their leaderboard descendants (Neuhof & Benjamini, 2026a;b; Li et al., 2026; Avelar Menendez et al., 2026) set a critical value for a fixed sample size, mostly by asymptotic approximation. They then apply the same counting formula as our (6). The first difference is the certified set $\mathcal{D}$ to which the formula is applied. Ours is valid at all times and uses the order structure inside the multiplicity correction. The second is that enumeration adds the exact rank sets of Theorem 4.3(a), which can be strictly

tighter than counting (Theorem 4.3(c)). Al Mohamad et al. (2021) instead test one configuration of the ranks at a time with the partitioning principle, which uses the order structure and gives rank sets without counting, at one fixed sample size. Our exact test is the anytime-valid counterpart of their procedure.

Group-sequential stopping needs its looks fixed in advance and a normal approximation, and as applied so far it has no correction across pairs. Ours needs neither and corrects across all pairs. Arviv et al. (2026) report large savings from Pocock boundaries (Pocock, 1977), one constant critical value at each pre-scheduled look, on pairwise comparisons. They leave multiplicity across pairs open, and Appendix E.3 measures the cost of this. Active-ranking algorithms (Heckel et al., 2019) choose which comparisons to make adaptively, which we do not. They guarantee their answer only at their own stopping time, when they output a grouping of the models. Ours guarantees every intermediate report under any stopping rule.

Among the anytime-valid methods, the two closest differ from ours in what they are able to certify and in how they treat the order structure. Gu et al. (2026) control the family-wise error rate by Bonferroni over the ordered pairs. They use transitivity only to propagate certified preferences and need transitive preferences only for that step. They assume i.i.d. prompts and give no bound for a fixed benchmark sampled without replacement. In ours transitivity instead follows from scalar abilities. It restricts the sets of true hypotheses that the multiplicity procedure has to test at all, which is why the exact test never certifies less than the full e-value closure of Hartog & Lei (2025). On the same wealths, the shortcut never certifies less than the Bonferroni correction of Gu et al. (2026), because its statistic is the pair's own wealth plus nonnegative terms. Arnold et al. (2026) certify which models may be best, a statement about rank one only, through a closure over all subsets of models that uses no relation among the pairwise hypotheses. Ours gives simultaneous rank inference for common-item benchmark scores, with validity on a fixed benchmark that rests only on the random item order.

The construction builds on several established statistical literatures. Multiple comparisons with the best (Hsu, 1996) is the classical form of "which models are tied for first", and our certified tier 1 is its anytime-valid analog. Step-down resampling (Romano & Wolf, 2005) is the fixed-sample engine behind Mogstad et al. (2024). Betting martingales (Waudby-Smith & Ramdas, 2024), time-uniform confidence sequences (Howard et al., 2021) and Ville's inequality (Ville, 1939) supply the sequential machinery. Fischer & Ramdas (2024) show that online closed testing procedures that cannot be uniformly improved must use e-values. This is a natural route to models arriving over time. On the evaluation side, Miller (2024) argues for error bars on benchmark scores, computed from item-level results, Bowyer et al. (2025) argue against the normal approximation at small sample sizes, and Chiang et al. (2024) describe the battle format we leave out. Singh et al. (2025) document how data-dependent practices, such as testing many private variants and disclosing only the best, shape public leaderboards. Time-uniform guarantees cover a data-dependent choice of when to stop or report one evaluation, but not the selection of which model variants to disclose.