Peeking Without Penalty: Anytime-Valid Sequential Evaluation of Language Models

Andrew Sheng-Han Huang*
National Yang Ming Chiao Tung University
guitarshhuang1111@gmail.com

Abstract

Benchmark accuracy comparisons are the field's primary instrument for measuring progress in language models, yet they are routinely run with a statistical procedure that would be rejected in any introductory sequential analysis course: evaluate a fixed (often budget-dictated) number of items, look at the running difference whenever convenient, and stop when the answer looks clear. We show by simulation that this ubiquitous "peeking" workflow inflates the false-positive rate of model comparisons from a nominal 5% to 40.1%. We then adapt modern safe anytime-valid inference to the specific structure of language-model evaluation, yielding SAVE (**S**equential **A**nytime-**V**alid **E**valuation): a protocol that streams benchmark items in random order, maintains paired e-processes and confidence sequences, and stops the moment a verdict – *superiority*, *practical equivalence*, or *budget exhausted* – is certified, with the same error guarantee no matter how often the experimenter looks. A sampling-without-replacement variant certifies what the *full benchmark* would conclude while evaluating only part of it, and e-value aggregation (e-BH) extends the guarantee to leaderboards of many simultaneous comparisons. Across 31 pairwise comparisons of seven open-weight models (2B–120B class) on MMLU and GSM8K subsets, SAVE returns a certified verdict on every pair (no replayed run ends in budget exhaustion) at a median of 33% of benchmark items (33% of token cost); on the 20 pairs that the full benchmark itself separates, the median cost falls to 24%, while the remaining 11 pairs cost 83% and are mostly returned as certified practical equivalence – a verdict a fixed-n test cannot deliver at all. SAVE contradicts the full-benchmark verdict in 7 of 31,000 replayed protocol runs, and its empirical false-positive rate in matched null simulations stays below the nominal level. We will release the protocol as a small library, and argue that "evaluate until certain" should replace "evaluate everything, then squint" as the default workflow for model comparison.

1 Introduction

Every consequential claim in modern language-model research – "our method beats the baseline," "the new checkpoint is better," "model A leads the leaderboard" – is, at bottom, a statement that one model's expected score on some distribution of tasks exceeds another's. The community's instrument for such statements is the benchmark evaluation: run both models on N items, compare means. The instrument is trusted, ubiquitous, and expensive; a single full sweep of a modern suite can consume thousands of GPU-hours [23, 25], and the cost compounds with reasoning models that emit long chains of thought per item, with multi-seed runs, and with development loops that re-evaluate after every change.

Given the cost, it is striking how statistically casual the standard workflow is. Three practices are near-universal:

*Incoming graduate student, National Taiwan University (from September 2026). Code, raw model outputs, and end-to-end reproduction scripts are available from the author and will be released publicly; see the disclosure paragraph before the references.

1. **Fixed-size evaluation without power analysis.** N is set by what the benchmark ships with or what the budget allows, not by the effect size of interest. A recent audit found that 11 of 40 published pairwise comparisons on the Open LLM Leaderboard, and 4 of 9 adjacent-rank pairs in the MMLU-Pro top ten, are statistically unresolved at conventional levels under an i.i.d. model [17] – the reported orderings are, in plain terms, noise.
2. **Silent peeking.** Development is intrinsically sequential: practitioners watch partial results stream in, stop runs that look decided, extend runs that look close, and re-run "one more seed" when a result is almost significant. Each look is a hypothesis test; classical fixed-n error guarantees are void under such optional stopping [16, 40]. Unlike A/B testing platforms, which confronted this a decade ago, evaluation harnesses provide no protection.
3. **Unadjusted multiplicity.** Leaderboards make $\binom{M}{2}$ simultaneous claims; selecting and publicizing the significant-looking ones is textbook selection bias, yet no deployed leaderboard we are aware of controls any family-wise error notion across its induced pairwise comparisons – methodology for certified rankings exists [21], but it is fixed-sample and is not in production.

The statistical literature has, in the last few years, produced exactly the toolkit this situation calls for: *safe anytime-valid inference* (SAVI) – e-processes, betting-based confidence sequences, and e-value multiple testing – which delivers error guarantees that hold *simultaneously over all data-dependent stopping times* [10, 12, 32, 41, 44]. Under SAVI, peeking is not a sin to be policed but the intended mode of use: the experimenter is encouraged to look constantly and stop the moment the evidence suffices. The fit to language-model evaluation is, we argue, almost embarrassingly good, for three reasons specific to the setting:

- **Evaluation outcomes are bounded and paired.** Per-item scores lie in $[0, 1]$ and both models answer the same items, so the relevant statistic is a paired difference in $[-1, 1]$ – precisely the regime where nonparametric betting methods are tight and assumption-free.
- **Benchmarks are finite populations.** The quantity practitioners actually act on is "what would the leaderboard say if we ran all N items." Sampling items without replacement admits confidence sequences that converge to *certainty* as the benchmark is exhausted [43], so one can certify the full-benchmark verdict from a fraction of it.
- **Language models agree on most items.** The variance of the paired difference is bounded by the models' disagreement rate, which we measure to be far below one for real model pairs. Decisions are therefore systematically cheaper than unpaired worst-case planning suggests – often by an order of magnitude.

Contributions. We make the case, and the artifact:

1. We formalize model comparison as sequential inference on paired bounded differences and instantiate SAVE, a stopping protocol with three certified verdicts (superiority, practical equivalence within a margin δ, budget exhaustion with a confidence sequence), valid under arbitrary peeking (§3). A without-replacement variant targets the exact full-benchmark gap, and e-BH aggregation [41] controls false discovery across all leaderboard pairs at any common stopping time (§3.4). To our knowledge this is the first evaluation protocol to combine anytime-valid guarantees under arbitrary optional stopping with a without-replacement target that certifies the full-benchmark verdict and e-value multiplicity control across all leaderboard pairs at arbitrary stopping times; the ingredients are imported with attribution, and concurrent work (§2) brings one or two of them to evaluation, but not the combination.
2. We quantify the cost of the status quo: continuous monitoring with naive z-tests at nominal $\alpha = 0.05$ yields 40.1% false positives in matched null simulations, while SAVE stays at 3.4% under identical monitoring, with stopping times competitive with an *oracle* fixed-n design that is told the true effect size in advance (§4).
3. We evaluate the protocol end-to-end on real models rather than only in simulation: seven open-weight models (2B–120B class) on 1,000 MMLU items and 300 GSM8K items, with deterministic decoding and fully logged outputs, giving 31 pairwise comparisons. SAVE certifies a verdict

on all of them at a median of 33% of items (33% of measured token cost), falling to 24% on the 20 pairs the full benchmark itself resolves, contradicts the full-benchmark verdict in 7 of 31,000 replays, and resolves 29% of the 21 MMLU leaderboard pairs within the first 100 items (§5). The library and all raw per-item logs will be released publicly.

We do not claim novel concentration inequalities: the martingale machinery is imported from the sequential-inference literature with attribution, and the transfer of anytime-valid ideas to model comparison has independent precedent outside language modeling [27, 37]. The contribution is the specific instantiation – recognizing that the evaluation crisis described by recent audits [17, 28, 42] and the mature SAVI toolkit are two halves of one solution – together with the finite-population and multiplicity extensions, the design choices, guarantees, measurements, and software that make the transfer practical.

2 Related work

Statistics of LM evaluation. A growing line of work documents that benchmark conclusions are noisier than their presentation suggests. Miller [28] derives fixed-sample standard errors, clustering corrections, and paired tests for evals; Wang [42] decomposes evaluation noise into data and prediction components via all-pairs paired analysis; Sclar et al. [35] shows prompt-format variance alone can swamp claimed gaps. Closest to our motivation, Kotawala [17] audits published leaderboards for statistical resolution and finds much of it lacking, and – in a sensitivity analysis we regard as independent support for the present paper – shows that replacing the fixed-n paired threshold by a time-uniform one derived from a paired-Bernoulli mixture e-process inflates the required sample size by roughly 2.15× and flips one further MMLU-Pro adjacent-rank pair from resolved to unresolved. That analysis prices anytime-validity as a *diagnostic*; it does not monitor an e-process to a data-dependent stopping time, and it issues no verdict. Li et al. [21] goes further on the multiplicity axis, certifying task-specific ranks and top-K membership via multiplier-bootstrap calibration, but again at a pre-specified sample size. Our protocol is the constructive complement to this literature: it consumes the same paired structure and outputs certified verdicts at data-dependent stopping times.

Sample-efficient evaluation. A complementary literature reduces the *number* of items: IRT-based benchmark compression [25], adaptive item selection [2, 5], and finite-population estimation with active querying and historical factor models [38, 45]. A parallel strand allocates a fixed budget across *models* using bandit machinery – successive rejects with paired comparisons [24], or best-model identification with low-rank score imputation [38]. These methods choose *which* items or models to run under a *fixed* budget, generally relying on auxiliary models of item difficulty fit on historical data, and their guarantees are for a pre-specified sample size; notably, Tolochinsky et al. [38] does sample without replacement from the benchmark and targets finite-population means, but its intervals are valid at a predetermined horizon rather than uniformly over stopping times. SAVE is orthogonal on both axes: it requires no historical data or difficulty model (items are simply randomized), and it decides *when to stop*, with guarantees that survive optional stopping. The approaches compose naturally – an informative item ordering or an adaptive model scheduler changes SAVE's stopping time, not its validity – and we view their combination as promising future work. Judge-noise correction via prediction-powered inference [3, 9] addresses an orthogonal error source (imperfect scoring) and could be layered on our outcome stream.

Group-sequential and alpha-spending designs. Interim analysis with controlled error long predates SAVI: Pocock [31] and O'Brien and Fleming [30] derived boundaries that spend a fixed α over a pre-specified number of looks, Lan and DeMets [19] freed the timing of the looks via alpha-spending functions, and the machinery is textbook material in clinical trials [15]. Concurrent work applies exactly this machinery to model evaluation [4]. What group-sequential designs still require is a pre-specified maximum sample size and adherence to the (re)computed look schedule; anytime-valid methods remove both inputs. Because the two families are natural competitors, we compare them head-to-head – together with the mixture sequential probability ratio test [16] – in §4.1.

Sequential and anytime-valid testing in machine learning. Sequential analysis begins with Wald [40]; confidence sequences with Darling and Robbins [8], Lai [18], Robbins [33]. Racing algorithms

applied Hoeffding-style sequential elimination to model selection three decades ago [6, 26, 29], and Mathieu et al. [27] modernized this for deep reinforcement learning, using group sequential tests with family-wise error control to compare agents with as few training runs as possible; we revive the same idea with tighter, nonparametric, *anytime-valid* machinery [12, 44] and LM-specific structure. SAVI has begun to reach evaluation-adjacent problems in AI: certifying agent trajectories from black-box verifier scores [34], auditing systems for failure modes under adaptive data collection [46], and – closest to us in spirit – comparing two robot policies with betting martingales and early stopping [37], which reports large reductions in evaluation burden but addresses superiority alone under an i.i.d. model, without a finite-population target or multiplicity control. In the LM era, sequential stopping has mostly been used *within* a single model's generation or repeated sampling – terminating self-consistency draws once the modal answer is clear [14, 20], or certifying answer sufficiency mid-generation [1] – which concerns inference-time compute for one model on one item, not statistical comparison of models across a benchmark. Anytime-valid methods are standard in industrial A/B testing [16, 32]; deployed evaluation harnesses, which face a structurally identical problem with stronger pairing, still offer no such protection.

Concurrent work. Three preprints appeared after this manuscript was completed (2026-06-12) and are concurrent rather than prior. Arviv et al. [4] (2026-07-09) applies *classical* group sequential testing with a Pocock spending function to model evaluation, defining six stopping rules – efficacy, an equivalence margin, precision, threshold crossing, futility, and diminishing returns – and demonstrating paired comparison and ranking at impressive scale on the Open VLM Leaderboard (206 vision-language models across 31 multimodal benchmarks), with roughly an 80% reduction in evaluation cost. The difference is the error model rather than the ambition: alpha-spending budgets a *pre-specified* number of interim looks on a *pre-specified* maximum sample size, so validity is lost under the unplanned, arbitrarily frequent peeking that characterizes real development loops (measured in §4.1), whereas e-processes are valid at every stopping time by construction; that work also does not treat the benchmark as a finite population, and explicitly defers multiplicity across pairwise comparisons. Their equivalence-margin rule and our TOST-style equivalence verdict are independent arrivals at the same practical need. Hsu and Shekhar [13] (2026-07-19) constructs betting- and RIPr-based confidence sequences for LM evaluation and designs active querying rules to shrink them, which is the closest work to ours in statistical machinery; it targets a *single* model's capability rather than a paired comparison, is evaluated on simulated correctness vectors, and is described by its authors as a preliminary version. Its querying rules and our stopping protocol are complementary: better item selection shortens SAVE's stopping time without affecting its validity. Li [22] (2026-06-13) uses betting e-processes in LM evaluation pipelines for a different purpose – attributing observed score drift to the system under test versus a changing judge. Taken together, these confirm that sequential evaluation is an idea whose time has arrived; none of them combines anytime-validity with a finite-population target and multiplicity control, which is the specific gap SAVE fills.

3 Setting and protocol

3.1 Setup

A benchmark is a finite set $Q = \{q_1, \ldots, q_N\}$ of items. A scored evaluation of model M assigns $s_M(q) \in [0, 1]$ (binary correctness in our experiments; partial credit changes nothing below). For two models A, B, define the paired per-item difference

$$D_i = s_A(q_i) - s_B(q_i) \in [-1, 1], \qquad X_i = \tfrac{1}{2}(D_i + 1) \in [0, 1].$$

Two targets of inference are useful, and the protocol supports both:

- the **finite-population gap** $\mu_N = \frac{1}{N}\sum_{i=1}^{N} D_i$ – "what the full benchmark would report" – relevant when the benchmark itself is the instrument of record (leaderboards, regression gates);
- the **distributional gap** $\mu = \mathbb{E}[D]$ under i.i.d. item sampling – relevant when the benchmark is treated as a sample from a broader task distribution.

Items are evaluated in uniformly random order (without replacement for μ_N; with replacement, or as an i.i.d. stream, for μ). Randomizing order is essential and free: benchmarks are typically grouped

by subject, and any deterministic order would violate the exchangeability the guarantees rest on. Decoding is deterministic (greedy, fixed seed) so that $s_M(q)$ is a fixed attribute of (M, q); §7 discusses stochastic decoding.

3.2 Anytime-valid primitives

Definition 1 (Confidence sequence)**.** *A $(1-\alpha)$ confidence sequence for a parameter θ is a sequence of sets $(\mathrm{CS}_t)_{t\geq 1}$, each computable from the first t observations, with $\mathbb{P}(\exists t : \theta \notin \mathrm{CS}_t) \leq \alpha$.*

Definition 2 (e-process)**.** *An e-process for a (possibly composite) null H_0 is a nonnegative process $(\mathcal{E}_t)_{t\geq 0}$, $\mathcal{E}_0 = 1$, upper-bounded by a supermartingale under every distribution in H_0. By Ville's inequality [39], $\mathbb{P}_{H_0}(\exists t : \mathcal{E}_t \geq 1/\alpha) \leq \alpha$, so rejecting when $\mathcal{E}_t \geq 1/\alpha$ is valid at any data-dependent stopping time.*

Both objects make "peeking" free by construction: the error budget is spent over the whole time horizon at once. The cost is a slightly wider interval than a fixed-n confidence interval at any single t – the iterated-logarithm price [33] – which §4 shows is modest at the effect sizes that matter in practice.

We instantiate both primitives by betting [36, 44]. For the one-sided null $H_0 : \mathbb{E}[X] \leq m$ with $X_t \in [0, 1]$, define the capital process

$$\mathcal{E}_t(m) \;=\; \prod_{s\leq t}\big(1 + \lambda_s\,(X_s - m)\big), \qquad \lambda_s \in [0,\, c/m], \tag{1}$$

with λ_s predictable (a function of $X_1, \ldots, X_{s-1}$) set by the empirical-variance plug-in of Waudby-Smith and Ramdas [44] and truncation $c = 0.75$. Since $\lambda_s \geq 0$, each factor has conditional expectation $1 + \lambda_s(\mathbb{E}[X_s] - m) \leq 1$ under every $P \in H_0$, so $\mathcal{E}_t(m)$ is an e-process for the composite null.

Reported two-sided confidence sequences are obtained by inverting the *same* one-sided capital processes: $\mathrm{CS}_t = \{m : \max_{s\leq t} \mathcal{E}_s(m) < 2/\alpha\}$, with (1) run at level $\alpha/2$ on X for the lower endpoint and on $1 - X$ for the upper – and with the without-replacement instantiation (2) below when targeting μ_N – so the reported interval always rests on the same argument, in the same sampling regime, as the verdicts. Because the plug-in truncates λ_s at $c \leq c/m$ for every $m \leq 1$, λ_s does not depend on m; hence $\log \mathcal{E}_t(m)$ is strictly decreasing in m, the un-rejected set is an interval, and the inversion is computed exactly by nested refinement. The closed-form empirical-Bernstein confidence sequence of Waudby-Smith and Ramdas [44], applied to X and $1 - X$, is retained for one internal purpose: the i.i.d. equivalence exit consults an interval at every t, where repeated inversion would be wasteful, and both constructions are valid $(1-\alpha)$ confidence sequences (details in Appendix B). We verify all guarantees empirically in §4 and in an executable test suite.

Without replacement. When targeting μ_N, the null "population mean $\leq m$" pins down the conditional mean of the next draw given the history: with $S_{t-1} = \sum_{s<t} X_s$,

$$m_t \;=\; \frac{Nm - S_{t-1}}{N - t + 1}, \tag{2}$$

and the capital process (1) with m replaced by the drifting m_t is again an e-process [43]. Two boundary cases make the finite-population logic vivid: if $m_t < 0$ the null is already arithmetically impossible (rejection with certainty); if $m_t > 1$ it can no longer be falsified and betting stops. As the benchmark is consumed the interval collapses to a point, so any nonzero gap is certified by exhaustion at the latest – i.i.d. machinery has no analogue of this.

3.3 The Save protocol

Algorithm 1 combines three certified exits. Superiority fires on either one-sided e-process at level $\alpha/2$; practical equivalence fires when the confidence sequence is contained in the margin $(-\delta, \delta)$ – implemented for the finite-population target as an anytime-valid two-one-sided-tests (TOST) construction, i.e. rejection of both shifted nulls $\mu \leq -\delta$ and $\mu \geq \delta$ by their own without-replacement e-processes at level $\alpha/2$ each, the sequential analogue of classical equivalence testing; and if the

Algorithm 1 SAVE: anytime-valid paired comparison of models A and B

input benchmark Q ($|Q| = N$), level α, equivalence margin δ, budget $T \le N$
draw a uniformly random permutation π of Q
for $t = 1, 2, \ldots, T$ **do**
  evaluate both models on item $q_{\pi(t)}$; observe D_t, set $X_t = (D_t+1)/2$
  update $\mathcal{E}_t^A$ ▷ e-process for $H_0 : \mu \le 0$, i.e. "A not better," at level $\alpha/2$
  update $\mathcal{E}_t^B$ ▷ e-process for $H_0 : \mu \ge 0$ (apply (1) to $1 - X$), level $\alpha/2$
  update the $(1-\alpha)$ confidence sequence $[L_t, U_t] \ni \mu$ ▷ by inverting $\mathcal{E}^A, \mathcal{E}^B$: $\{m : \max_{s \le t} \mathcal{E}_s(m) < 2/\alpha\}$, $\alpha/2$ per side
  if $\mathcal{E}_t^A \ge 2/\alpha$ **then return** "$A \succ B$" with $[L_t, U_t]$ ▷ then $L_t \ge 0$ automatically
  else if $\mathcal{E}_t^B \ge 2/\alpha$ **then return** "$B \succ A$" with $[L_t, U_t]$ ▷ then $U_t \le 0$ automatically
  else if $[L_t, U_t] \subset (-\delta, \delta)$ **then return** "equivalent within δ" with $[L_t, U_t]$
  end if
end for
return "undecided" with the running interval $[L_T, U_T]$

budget runs out first, the protocol returns the running confidence sequence – which, by construction, is a valid interval *at the budget stopping time too*. There is no failure mode in which the experimenter walks away empty-handed: every exit is a quantified statement.

Proposition 1 (Verdict validity)**.** *Under random item order, for any data-dependent stopping rule and any horizon: (i) if $\mu \le 0$, $\mathbb{P}(\text{declare } A \succ B) \le \alpha/2$, and symmetrically; in particular the probability of declaring any false winner is at most α when $\mu = 0$; (ii) if $|\mu| \ge \delta$, $\mathbb{P}(\text{declare equivalence}) \le \alpha$; (iii) with probability $\ge 1 - \alpha$, every reported interval contains μ simultaneously for all t, where the reported interval $[L_t, U_t] = \{m : \max_{s \le t} \mathcal{E}_s(m) < 2/\alpha\}$ is obtained by inverting the same e-process family used in (i), instantiated in the same sampling regime; (iv) the reported interval is consistent with the verdict by construction: a superiority declaration implies $0 \notin [L_t, U_t]$, and a TOST equivalence declaration implies $[L_t, U_t] \subseteq [-\delta, \delta]$. The same statements hold verbatim for μ_N under the without-replacement instantiation (2).*

The proof (Appendix A) is a direct assembly of Ville's inequality applied to (1), the composite-null supermartingale property, and the confidence-sequence guarantee; we state it to fix exactly which workflow is being certified, since the protocol's value lies in the guarantee surviving behavior – continuous monitoring, early stopping, budget cuts – that breaks classical analyses.

Remark 1 (How long does it take?)**.** *For betting e-processes, rejection under $|\mu| = \Delta > 0$ occurs after $O\big(\frac{\sigma^2}{\Delta^2}\big(\log\log\frac{\sigma^2}{\Delta^2} + \log\frac{1}{\alpha}\big)\big)$ items, matching the iterated-logarithm lower bound up to constants [12, 44]. The paired variance obeys $\sigma^2 = \mathbb{E}[D^2] - \mu^2 \le \rho - \mu^2$ where $\rho = \mathbb{P}(D \ne 0)$ is the models' disagreement rate. This is the operative quantity for language models: pairs that agree on 90% of items have $\sigma^2 \le 0.1$, an order of magnitude smaller than the unpaired worst case, and stopping times shrink proportionally (§4, §5).*

Remark 2 (Choosing the margin δ)**.** *A confidence sequence's width at the full budget T is of order $w(T) = \sqrt{8\hat{\sigma}^2 \log(2/\alpha)/T}$. Certifying equivalence requires not only width below 2δ but the interval to sit centered inside the band despite the $O(\sqrt{\hat{\sigma}^2/t})$ fluctuation of its midpoint; a practical rule that matches our simulations is $\delta \gtrsim w(T)$. With $\hat{\sigma}^2 \approx 0.3$ and $T = 12{,}000$, $w(T) \approx 0.027$: $\delta = 0.02$ is never certified, $\delta = 0.03$ in 42% of tied runs, $\delta = 0.05$ in 97% (§4). Practitioners should set δ to the smallest gap they would act on, then check it clears the floor for their budget; if not, the honest output is the interval, not a verdict.*

3.4 Many models: anytime-valid leaderboards

A leaderboard over M models induces $K = \binom{M}{2}$ comparisons. Each pair (j, k) carries the pair-level e-value $\mathcal{E}^{(j,k)} = \frac{1}{2}(\mathcal{E}^A + \mathcal{E}^B)$ from its paired stream: under the pair's tie null ($\mu_{jk} = 0$) both one-sided capital processes have unit initial value and nonpositive drift, so their average has expectation

at most one (the maximum would not – averaging is the price of testing "some difference" with two one-sided bets). The direction reported for a resolved pair is the side whose process crossed its own threshold. Because e-values evaluated at stopping times remain e-values, and the e-BH procedure of Wang and Ramdas [41] controls FDR at level α under *arbitrary dependence*, the following is valid at any common budget t, chosen adaptively: sort the K running e-values in decreasing order, find the largest k with $e_{(k)} \geq K/(\alpha k)$, and declare those k pairs resolved. No assumption about the dependence between pairs sharing a model is needed – exactly the situation in which p-value-based corrections become either invalid or hopelessly conservative. The same e-values support racing-style schedulers (stop sampling pairs once resolved, reallocate budget to unresolved ones [26]); our experiments use the simplest uniform round-robin scheduler and leave adaptive allocation as future work.

3.5 Cost accounting

Items are not equally priced: measured per-item cost varies with prompt length and, dramatically, with generated tokens (reasoning traces). SAVE therefore reports savings in both *items* and *measured tokens* (prompt + generated, summed over both models, from logged counts). All real-data savings in §5 are token-weighted as well as item-weighted; the two can differ when stopping happens to favor cheap or expensive items, though in our runs they track closely.

4 Simulation study

Paired differences are simulated as ternary variables ($D \in \{-1, 0, 1\}$) parameterized by gap Δ and disagreement rate ρ – exactly the structure of binary-scored model comparison. All numbers below are from 10^4–2×10^4 replications per condition (produced by the reproduction scripts; full tables in Appendix E).

The cost of peeking, measured. Under the null ($\Delta = 0$, $\rho = 0.3$), an experimenter who computes a paired z-test every 20 items at nominal $\alpha = 0.05$ over a horizon of 2,000 items reaches a false positive with probability 40.1% – eight times nominal. A single pre-registered fixed-n test at the horizon is fine (4.9%) but answers the wrong workflow: nobody evaluates exactly once at a pre-committed n. SAVE, monitored after *every single item* over the same horizon, rejects falsely in 3.4% of runs – below its nominal 5%, with the slack reflecting finite-horizon conservatism (the guarantee covers infinite horizons). Figure 1 traces the divergence as a function of items inspected: the naive curve climbs without bound (Robbins' law of the iterated logarithm guarantees it reaches 1 eventually [33]), while the e-process curve plateaus.

Efficiency against an oracle. The classical alternative needs the unknown Δ to size the experiment. We therefore compare SAVE's data-dependent stopping time to a fixed-n design told the true (Δ, ρ) and sized for 90% power (Figure 2). Across $\Delta \in [0.03, 0.2]$ and $\rho \in \{0.2, 0.4\}$, SAVE's median stopping time tracks the oracle within a factor of 0.7–1.4: e.g. at $(\Delta, \rho) = (0.05, 0.2)$ the protocol decides at a median of 680 items versus the oracle's 831; at $(0.08, 0.4)$, 580 versus 647. The protocol pays the iterated-logarithm premium only at the smallest effects, and in exchange requires no effect size guess at all: the same procedure that stops at 58 items for $\Delta = 0.2$ keeps sampling at $\Delta = 0.02$. In practice the oracle does not exist, and mis-specifying it is costly in the familiar two ways (wasted compute, or an underpowered run that ends "unresolved").

Agreement is the budget. Fixing $\Delta = 0.05$ and varying disagreement: at $\rho = 0.06$ (models agree on 94% of items) the median decision takes 248 items; at $\rho = 0.8$, 5,393 items – a 20× spread driven entirely by paired variance (Figure 6, left). Remark 1's σ^2/Δ^2 scaling predicts this almost exactly, and §5 confirms the same law on real model pairs. The practical reading: evaluating two similar checkpoints against each other is *cheap*; it is dissimilar models on heterogeneous benchmarks that cost.

Equivalence and its floor. Under exact ties ($\Delta = 0$, $\rho = 0.3$, horizon 12,000), a margin of $\delta = 0.05$ certifies equivalence in 96.8% of runs (median 3,727 items); $\delta = 0.03$ in 41.9%; $\delta = 0.02$ – below the width floor of Remark 2 – in none. False winners are declared in 3.2–3.5% of tied runs, within

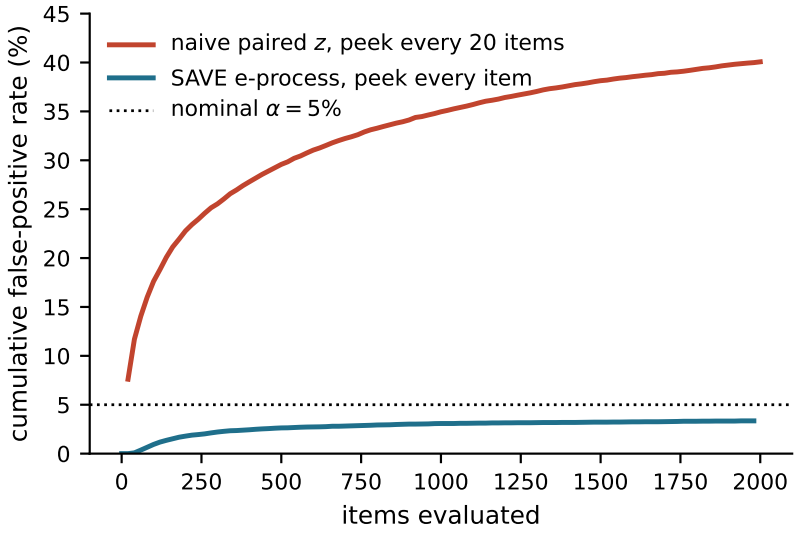

Figure 1: Continuous monitoring breaks the naive test and not the e-process. Cumulative probability of a (false) rejection under the null ($\Delta = 0$, disagreement $\rho = 0.3$, 2×10^4 simulated paired streams) as a function of items evaluated. The naive paired z-test peeked every 20 items crosses its nominal 5% within a few hundred items and keeps climbing; SAVE (monitored after every item) stays below nominal at all horizons.

the $\alpha = 5\%$ budget. Ties are expensive relative to wins (no signal accumulates), which is itself an argument for explicit margins: without them, near-ties silently consume unbounded budget.

Exhaustion: the without-replacement dividend. On a finite benchmark of $N = 1{,}000$ with true gap 0.03 (disagreement 0.2), the i.i.d. e-process usually cannot certify within the benchmark (median stopping time = N, resolution rate 10%), while the without-replacement variant certifies at a median of 778 items (resolution rate 100%): as the remaining-item mass shrinks, the null's conditional mean (2) drifts away from the data and capital compounds. For gap 0.05 the medians are 601 versus 1,000. The dividend extends to ties: under an exact tie with 10% disagreement on $N = 1{,}000$, the TOST exit certifies equivalence at margin $\delta = 0.03$ in 99% of runs at a median of 681 items – a certificate the i.i.d. construction essentially never grants (§4, equivalence floor). Small gaps and near-ties on finite benchmarks are exactly where leaderboards live, and where the finite-population view pays.

4.1 Comparison with group-sequential designs

Interim analysis with controlled error is not new: Pocock [31] and O'Brien and Fleming [30] control the Type I error of a pre-planned sequence of looks, Lan and DeMets [19] freed the timing of the looks via alpha-spending, and concurrent work applies this machinery to model evaluation [4]. The right question is therefore not whether sequential testing helps – it demonstrably does – but what anytime-validity buys over the classical constructions. We compare SAVE against Pocock and O'Brien–Fleming boundaries, Lan–DeMets alpha-spending recomputed for the actual look schedule, and the mixture sequential probability ratio test (mSPRT) of Johari et al. [16], on the same ternary paired streams as above (our boundary computations reproduce the published Pocock and O'Brien–Fleming constants of Jennison and Turnbull [15] to within 0.0011; full tables in Appendix F).

Where GST wins, and we say so. When the design is correctly specified – the horizon fixed in advance and the look schedule honored – group-sequential tests are more powerful than SAVE at small effects: at $(\Delta, \rho) = (0.02, 0.2)$ over a correctly sized horizon, Pocock achieves power 0.72 using 2,822 items on average where SAVE reaches 0.32 using 3,312; at $\Delta = 0.03$ the comparison is 0.97/1,843 versus 0.72/2,389. The mSPRT ($\tau = 0.05$) is likewise more sample-efficient on i.i.d. streams (power 0.91 at $\Delta = 0.03$ with 1,839 items) with Type I error still controlled (1.0–4.4% across our null conditions). We report this because it is true, and because it locates SAVE's value precisely:

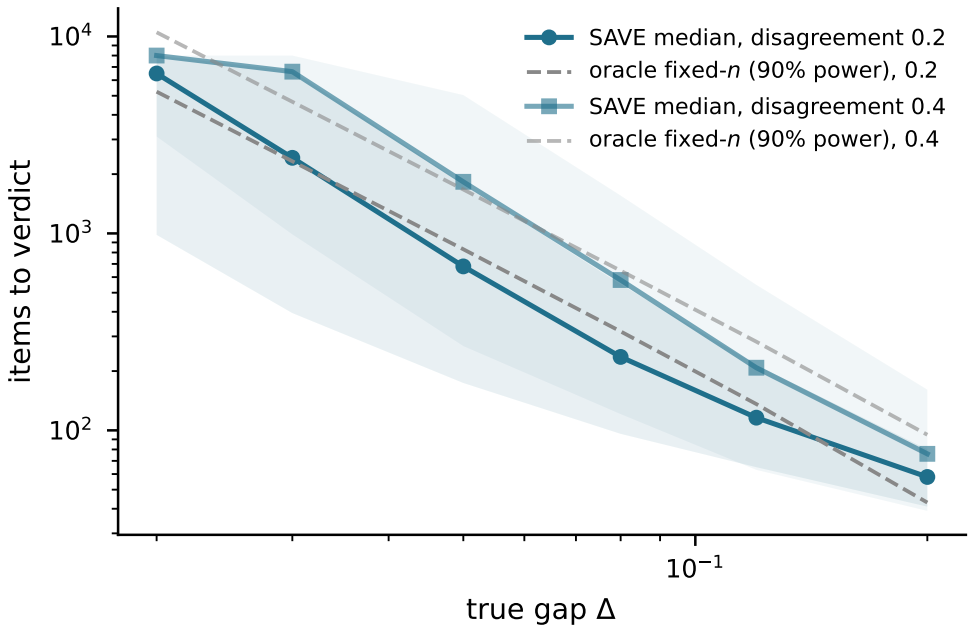

Figure 2: Adaptive stopping is competitive with an oracle fixed-n design. Median SAVE stopping time (solid, with 10–90% band) versus effect size Δ, for two disagreement rates; dashed: fixed-n sized at 90% power with the true (Δ, ρ) revealed. Log–log axes.

not unconditional efficiency, but freedom from inputs the classical designs cannot do without. We also tested – and falsified – the conjecture that GST's normal approximation degrades on sparse paired streams: under the null, GST's Type I error never exceeded 5.3% even at 99.5% agreement (Appendix F), so none of what follows rests on approximation quality.

The schedule is part of the guarantee. Group-sequential boundary constants are computed for a pre-registered number of looks, and in practice they are frequently written into an analysis plan and then consulted more often than planned. Reusing a $K = 5$ Pocock boundary while actually looking after every batch drives Type I error from 4.8% (the planned five looks) to 16.1% at 1,901 looks – more than three times nominal (Figure 3a); the O'Brien–Fleming boundary degrades from 5.0% to 8.0%. To give the classical family its strongest form: if the Lan–DeMets spending boundary is *recomputed* for the schedule actually used, Type I error stays at 4.8–5.1% throughout – alpha-spending genuinely repairs the "how often" degree of freedom, and we verified that it does. The criticism therefore cannot stop at look counts; it must move to the horizon. SAVE, monitored after every single item on the same streams, sits at 3.1%.

The horizon is part of the guarantee. Alpha-spending's information fraction is $t = n/n_{\max}$: the freedom in look timing is purchased by fixing a maximum sample size in advance. We fix the true gap at $\Delta = 0.03$ ($\rho = 0.2$; oracle 90%-power fixed-n = 2,325) and let the analyst guess $n_{\max}$, with a shared pool of 8,000 items: Pocock's power ranges from 15.2% at $n_{\max} = 250$ through 47.4% at $n_{\max} = 1{,}000$ to 97.5% at $n_{\max} = 4{,}000$ (Figure 3b,c). To match SAVE – which takes no $n_{\max}$ input and delivers 94.8% power at 2,896 items on average – the analyst must set $n_{\max}$ to 4,000, i.e. 1.72× the oracle sample size; guessing 1,000 (0.43× oracle) collapses power to 47.4%. Over-reserving carries its own cost: at $n_{\max} = 8{,}000$ the early boundaries are so conservative that GST consumes 2,350 items on average against 1,824 at $n_{\max} = 4{,}000$ – a power/speed dial that SAVE simply does not have. And the natural reaction to an $n_{\max}$ guessed too small – extend the run and test again – breaks the guarantee: a Pocock design with $n_{\max} = 500$ (Type I 5.3%) extended to 2,000 items and declared significant if either stage rejects has combined Type I error 8.9%; SAVE on the same horizon sits at 2.9%, and it has no notion of "stages" to begin with – the same e-process simply keeps running.

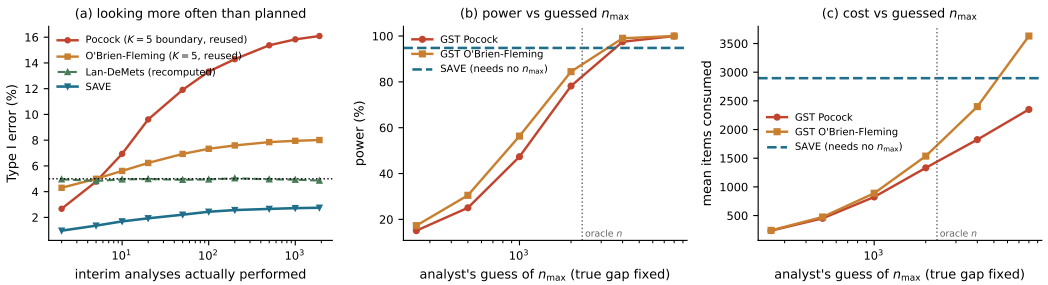

Figure 3: What anytime-validity buys. (a) A group-sequential boundary computed for $K = 5$ looks fails when the analyst actually looks more often (Pocock reaches 16.1%); Lan–DeMets spending recomputed for the actual schedule remains valid, but only by fixing $n_{\max}$ in advance. (b,c) Fixing the true gap and varying the analyst's guess of $n_{\max}$: GST must guess 1.72× the oracle sample size to match SAVE; guessing small collapses power, guessing large wastes items. SAVE (horizontal dashed line) takes no such input.

Table 1: Models, full-benchmark accuracies, measured mean cost per item (prompt + generated tokens), and answer-extraction failures (no parseable answer within the token budget; scored 0).

	MMLU subset (N=1000)			GSM8K subset (N=300)		
Model	acc. %	tok/item	unpar.	acc. %	tok/item	unpar.
Gemma4-31B	83.2	124	73	–	–	–
Gemma4-E2B	53.3	345	96	85.7	413	0
GPT-OSS-120B	90.6	260	2	96.3	327	0
Llama4-Scout	80.1	623	43	97.0	596	0
Qwen3.6-27B	85.5	122	0	–	–	–
Qwen3.6-35B-A3B-Coding	82.2	122	1	96.0	351	0
Qwen3.6-35B-A3B	81.9	122	5	96.0	358	0

Granularity. A group-sequential design cannot stop before its first scheduled look. At $\Delta = 0.2$ ($\rho = 0.2$) SAVE reaches a verdict in 60 items on average while $K = 5$ GST spends 800 (13.3×); at $\Delta = 0.12$ the comparison is 131 versus 800. The crossover sits near $\Delta \approx 0.05$. This matters for LM leaderboards specifically, because most real pairwise gaps are large (§5): the regime where SAVE dominates is the regime evaluations mostly live in.

5 Real-model study

Setup. We evaluate seven open-weight instruction-tuned models spanning roughly 2B–120B-parameter classes (three dense, four mixture-of-experts; exact checkpoints in Appendix C), served locally with deterministic greedy decoding, fixed seed, and reasoning modes disabled or minimized, on consumer hardware (Apple M5 Max, 128GB). Benchmarks: a 1,000-item uniform subsample of 14 MMLU subjects [11] spanning quantitative, logical, and humanities categories (answer-letter protocol, ≤16 generated tokens for compliant models), and 300 GSM8K test items [7] (short chain-of-thought, final-answer extraction). Every model output is logged verbatim; per-item prompt and generation token counts give measured costs. Five models run GSM8K (the two slowest dense models are MMLU-only), yielding $\binom{7}{2} = 21$ MMLU pairs and $\binom{5}{2} = 10$ GSM8K pairs. Full per-model accuracies: Table 1; the spread (MMLU 53–91%, GSM8K 86–97%) gives pair gaps from sub-point to tens of points – a realistic mix of easy calls, hard calls, and near-ties.

Protocol replay. With full correctness matrices in hand, we replay Algorithm 1 (without-replacement mode, $\alpha = 0.05$, $\delta = 0.03$) over 1,000 independent random item orders per pair, which yields the exact distribution of stopping times and verdicts – every replayed trajectory is a faithful run of the protocol on real model outputs. Ground truth for each pair is the full-benchmark verdict: the sign of the gap on all N items, with a classical paired test on all N recording whether even the complete benchmark resolves the pair.

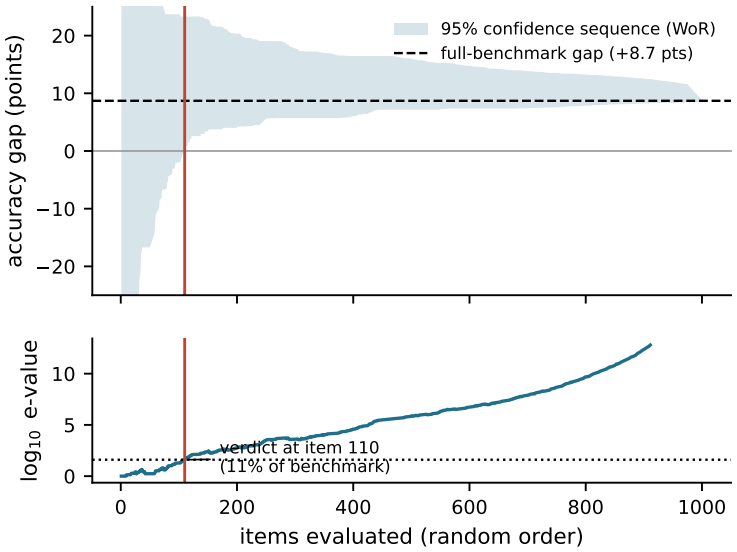

Figure 4: One comparison, end to end. A representative MMLU pair under a single random item order: the without-replacement confidence sequence (top) narrows around the full-benchmark gap while the running e-value (bottom, log scale) compounds; the protocol stops the moment the e-value crosses $2/\alpha$, far before the benchmark is exhausted. Everything to the right of the vertical line is compute the standard workflow would have spent for no change in verdict.

Headline: certified verdicts at a fraction of the cost. Across all 31 pairs, SAVE reaches a certified verdict – superiority or practical equivalence – at a median of 33% of items (33% of measured token cost), and never once runs out of benchmark without one: 0 of 31,000 replayed runs ended in budget exhaustion. This pooled figure mixes two very different regimes. On the 20 pairs that the classical full-benchmark paired z-test itself resolves at $\alpha = 0.05$ – a criterion fixed in advance, but computed from the same data, so the resulting number is conditional and is reported as such – the median cost is 24% of items (24% of tokens), with per-pair medians ranging from 3% for the largest gaps to essentially the full benchmark for the smallest resolvable ones (Figure 5). On the remaining 11 pairs the median cost is 83%, and 78% of those replays end in a certified equivalence verdict – an answer the fixed-n benchmark does not produce at all, which is why these pairs are expensive rather than failures. Pooling the two benchmarks into a single median is itself a weighting choice: taken separately, the 16 resolvable MMLU pairs certify at 16% of items while the 4 resolvable GSM8K pairs need 31%, and over all pairs the figures are 33% (MMLU, 21 pairs) and 78% (GSM8K, 10 pairs); GSM8K is dominated by near-identical models, so most of its pairs resolve as ties rather than winners. Verdict flips – a replay certifying the *opposite* winner to the full-benchmark gap – occur in 7 of 31,000 replayed runs across all pairs, and all 7 of them fall on pairs the full benchmark itself cannot separate (0 on the 20 resolvable pairs), i.e. on pairs where the sign of the realized gap is itself statistically meaningless. One subtlety deserves care: the without-replacement certificate concerns the *realized* benchmark gap μ_N, which becomes known with certainty as the benchmark is exhausted – so a pair with a tiny but nonzero μ_N can legitimately receive a certified finite-population winner that a superpopulation z-test would not grant. These near-exhaustion certificates are the non-equivalence remainder of the 11 unresolvable pairs above; the conditional 20-pair figure excludes them by construction, and the leaderboard analysis below returns to the same artifact.

The agreement law holds on real pairs. Figure 6 (right) overlays the 20 of 31 real pairs that reach a certified verdict on the simulated agreement–stopping-time law: measured disagreement rates span 0.01–0.41, and median stopping times collapse onto the σ^2/Δ^2 curve. Real model pairs cluster in the favorable regime: most agree on 86% of items, so certified comparisons are systematically cheaper than unpaired planning would predict – the empirical version of Remark 1.

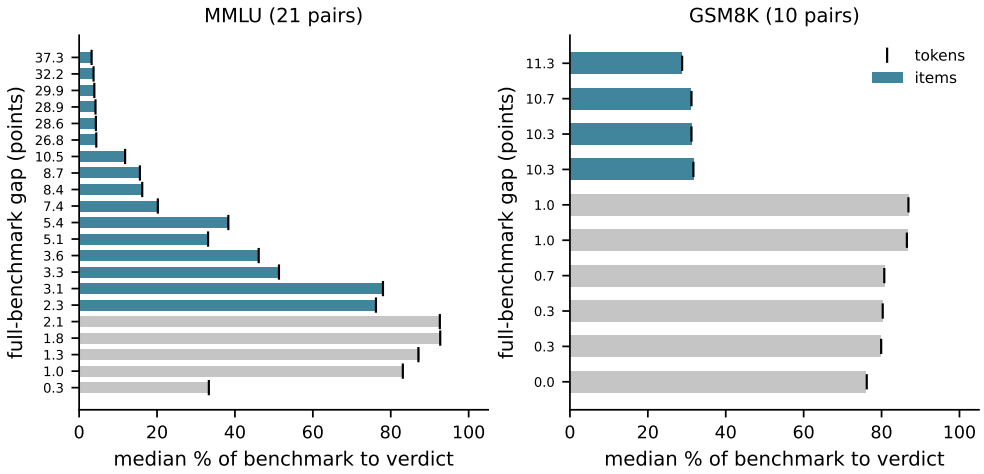

Figure 5: Real-pair savings. Median fraction of benchmark items (bars) and measured tokens (ticks) consumed before a certified verdict, per model pair over 1,000 random orders, ordered by full-benchmark gap $|\mu_N|$ (largest at top). Left: 21 MMLU pairs; right: 10 GSM8K pairs. Grey bars mark the pairs the full benchmark itself does not resolve.

Leaderboard mode. Running all pairs round-robin under a shared item budget and applying e-BH at each budget gives an anytime-valid leaderboard (Figure 7): 29% of the 21 MMLU pairs are resolved (FDR $\leq$ 5% across the whole family) within the first 100 items, rising to 100% at exhaustion. The jump to full resolution is a property of the finite-population e-process, not added power: at $n = N$ the realized gap μ_N is known exactly, and no MMLU pair has $\mu_N = 0$ – the smallest gap is 0.3 points, 3 items out of 1,000 – so nothing is left to certify. Indeed, 5 of the MMLU pairs certified at exhaustion are precisely pairs the classical full-benchmark z-test leaves unresolved. Full resolution is emphatically not the same as a difference worth acting on: that smallest-gap pair is certified because its gap is real, not because it is meaningful, and under $\delta = 0.03$ the protocol calls it a tie in every one of its 1,000 replays (Table 3). The δ-margin, not exhaustion, is the defence against hairline gaps being read as wins. A residue does appear wherever a pair is genuinely tied: on GSM8K, whose smallest realized gap is exactly zero, e-BH tops out at 71% of its 10 pairs even at exhaustion (3 of those certified being z-unresolved), matching the irresolution documented on public leaderboards [17]. We therefore read the e-BH curve at intermediate budgets, not at exhaustion. The marginal information of an evaluation item decays steeply: most of what a 1,000-item benchmark can ever say about a model roster, it says within a couple hundred items, and the protocol knows – with guarantees – *which* part it has already said.

What it costs in software. The protocol is ~470 lines of dependency-light Python over a stream of $\{-1, 0, 1\}$ outcomes; it adds microseconds per item and integrates with any harness that can yield per-item scores in random order. The library, the validity test suite (which re-derives every guarantee empirically), all raw model outputs, and every script behind every number will be released publicly, and are available from the author in the interim.

6 Practical guide

For practitioners, the protocol distilled:

1. **Randomize item order** (and log the seed). Grouped-by-subject streaming silently voids the guarantees.
2. **Pair on items.** Evaluate both models on the same items; feed differences, not separate accuracies. Pairing is where the order-of-magnitude variance reduction lives.
3. **Pick α, a margin δ you would act on, and a budget.** Check δ clears the width floor $\sqrt{8\hat{\sigma}^2 \log(2/\alpha)/T}$ (Remark 2); otherwise drop the equivalence exit.

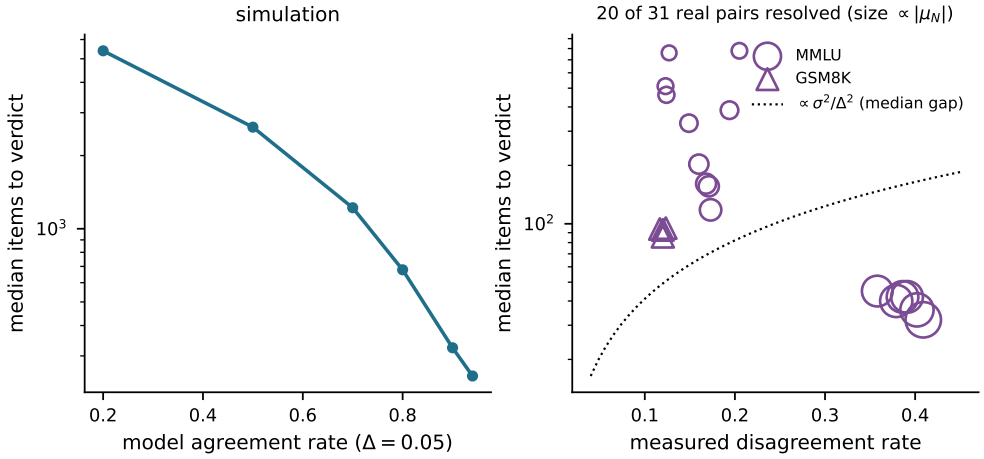

Figure 6: Disagreement, not benchmark size, sets the price of a verdict. Left: simulated median stopping time versus model agreement rate at fixed gap $\Delta = 0.05$. Right: the 20 of 31 real pairs that the full benchmark itself resolves (median stopping time vs. measured disagreement; marker size $\propto |\mu_N|$), against the $\propto \sigma^2/\Delta^2$ prediction. The remaining 11 pairs are omitted: the full benchmark does not separate them, so $|\mu_N|$ is not a meaningful Δ for the scaling law, and their stops are governed by the equivalence exit rather than by superiority.

4. **Peek freely; stop at any exit.** Report the verdict *and* the interval at stop. "Undecided at budget, gap in $[-0.4, 2.1]$ points" is a finding, not a failure. At a superiority stop, what is certified is the strict inequality $\mu > L_t \geq 0$, and L_t will typically sit just above zero at the moment the threshold is crossed – "just provably positive" is not "provably large," and a meaningful lower bound requires continuing to sample. We write the interval half-open, $(L_t, U_t]$, so that a boundary-grazing $L_t = 0$ is not misread as containing zero. And report the confidence sequence, never the running sample mean at the stop, which is biased by optional stopping (§7).
5. **On leaderboards, report e-values per pair** and the e-BH frontier, so readers see which orderings are claims and which are noise.
6. **Use without-replacement mode** whenever "the benchmark" is the population of record – it is never worse and is dramatically better near exhaustion.

7 Limitations and scope

Deterministic decoding. Our guarantees and experiments treat $s_M(q)$ as fixed under greedy decoding. Sampled decoding adds a second variance source (per-item correctness becomes Bernoulli), making outcomes $\bar{s}$ over k repeats bounded in $[0, 1]$ – the machinery applies unchanged, but the optimal split between new items and repeats, and the interaction with reasoning-model nondeterminism, is unexplored here; our preliminary analysis suggests a closed-form items-vs-repeats rule from the two variance components, which we leave to future work. **Verdicts are benchmark-relative.** SAVE certifies the gap on (the distribution behind) the chosen benchmark; it inherits, and is silent about, contamination, format sensitivity [35], and construct validity. It makes the measurement honest; it does not make the ruler good. **Iterated-logarithm premium.** At very small effects with high disagreement, the oracle fixed-n design is moderately cheaper; if an oracle is somehow available, use it. **Efficiency at small effects under a correctly specified horizon.** Relatedly, when the horizon is fixed in advance and correctly sized, group-sequential designs are more powerful than SAVE at small effects (§4.1: at $\Delta = 0.02$, Pocock power 0.72 versus SAVE's 0.32), and the mSPRT is more sample-efficient on i.i.d. streams with Type I error still controlled (1.0–4.4% in our null conditions). What SAVE buys is freedom from the $n_{\max}$ and look-schedule inputs, the finite-population target, and the equivalence exit – not unconditional statistical efficiency. **Estimation bias at the stop.** Optional stopping biases the running mean at the stopping time for any sequential procedure, and SAVE's earlier stops make the bias *larger*, not smaller: at a true gap of 0.05, the mean estimate at SAVE's stop is 0.067 (bias

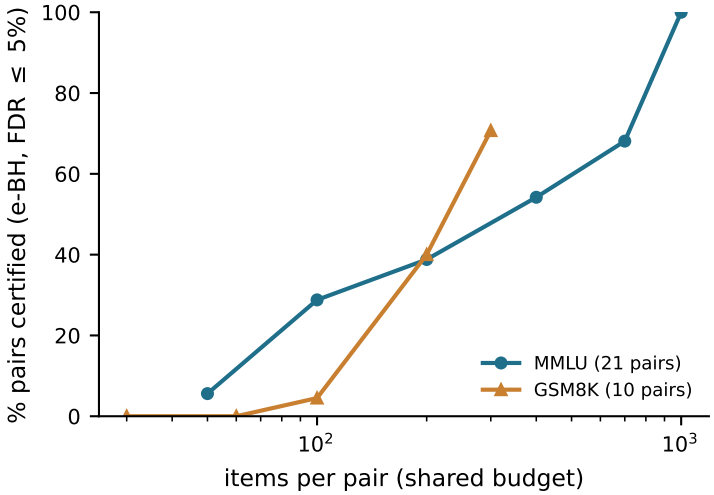

Figure 7: Anytime-valid leaderboard resolution. Fraction of all pairwise comparisons certified by e-BH (FDR ≤ 0.05, arbitrary dependence) as a function of the shared per-pair item budget.

+0.017), versus +0.003 for a $K = 5$ group-sequential design (§4.1). Report the confidence sequence, never the stopped sample mean: the reported CS covers at SAVE's own stopping times (99.9% in our checks, valid if conservative), while the naive Wald interval undercovers (93.5%). **Scheduler.** Our leaderboard experiments use uniform round-robin; adaptive budget allocation (racing) and informative item ordering [25, 45] compose with our guarantees and should compound the savings. **Scale.** Seven models, two benchmark families, English items, one hardware stack; the statistics are model-agnostic, but broader empirical sweeps – API-served frontier models, generative scoring with judges [9] – are natural extensions.

8 Conclusion

Language-model evaluation is a sequential experiment run, today, with non-sequential statistics. The mismatch is not cosmetic: it produces phantom wins at eight times the advertised error rate, leaderboard orderings that dissolve under a power analysis, and full-benchmark sweeps whose latter halves mostly confirm what their first quarters already knew. The remedy is mature, simple, and cheap. Evaluate in random order, bet on the difference, stop when certain – and peek as much as you like.

Broader impact. Cheaper certified comparisons lower the barrier to rigorous evaluation for small labs, reduce the energy footprint of routine benchmarking, and make leaderboard claims auditable. The main misuse risk is social: a "certified at $\alpha = 0.05$" stamp can be read as more than it is – a statement about one benchmark's gap, not about general capability. We mitigate by requiring the interval, the margin, and the benchmark identity to travel with every verdict.

Acknowledgments and disclosure. This manuscript was drafted with substantial assistance from an AI system (Claude, Anthropic), which also implemented the statistical library and executed the experiments under the author's direction. Every reported number is produced by the analysis scripts from logged model outputs and can be regenerated end to end from the reproduction package (to be released publicly; available from the author in the interim); the statistical guarantees are additionally checked by an executable test suite, and the bibliography was machine-validated against arXiv and Crossref records. The author reviewed the manuscript and assumes full responsibility for its content.

References

[1] Sanjeda Akter, Ibne Farabi Shihab, and Anuj Sharma. Anytime-valid answer sufficiency certificates for LLM generation via sequential information lift. *arXiv preprint arXiv:2510.06478*, 2026.

[2] Aashish Anantha Ramakrishnan, Ardavan Saeedi, Hamid Reza Hassanzadeh, Fazlolah Mohaghegh, and Dongwon Lee. Generative active testing: Efficient LLM evaluation via proxy task adaptation. *arXiv preprint arXiv:2603.19264*, 2026.

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

[4] 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. *arXiv preprint arXiv:2607.08522*, 2026.

[5] Esma Balkır, Alice Pernthaller, Marco Basaldella, José Hernández-Orallo, and Nigel Collier. Confident rankings with fewer items: Adaptive LLM evaluation with continuous scores. *arXiv preprint arXiv:2601.13885*, 2026.

[6] Mauro Birattari, Zhi Yuan, Prasanna Balaprakash, and Thomas Stützle. F-race and iterated F-race: An overview. In *Experimental Methods for the Analysis of Optimization Algorithms*, pages 311–336. Springer, 2010.

[7] 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 preprint arXiv:2110.14168*, 2021.

[8] Donald A. Darling and Herbert Robbins. Confidence sequences for mean, variance, and median. *Proceedings of the National Academy of Sciences*, 58(1):66–68, 1967.

[9] Chen Feng, Minghe Shen, Ananth Balashankar, Carsten Gerner-Beuerle, and Miguel R. D. Rodrigues. Noisy but valid: Robust statistical evaluation of LLMs with imperfect judges. *arXiv preprint arXiv:2601.20913*, 2026.

[10] Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024.

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

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

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

[14] Jingkai Huang, Will Ma, and Zhengyuan Zhou. Optimal Bayesian stopping for efficient inference of consistent LLM answers. *arXiv preprint arXiv:2602.05395*, 2026.

[15] Christopher Jennison and Bruce W. Turnbull. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC, 2000.

[16] Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Peeking at A/B tests: Why it matters, and what to do about it. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2017. doi: 10.1145/3097983.3097992.

[17] Anany Kotawala. Resolution diagnostics for paired LLM evaluation. *arXiv preprint arXiv:2605.30315*, 2026. ICML 2026 Workshop on Hypothesis Testing.

[18] Tze Leung Lai. On confidence sequences. *Annals of Statistics*, 4(2):265–280, 1976.

[19] K. K. Gordon Lan and David L. DeMets. Discrete sequential boundaries for clinical trials. *Biometrika*, 70(3):659–663, 1983.

[20] Jaeyeon Lee, Guantong Qi, Matthew Brady Neeley, Zhandong Liu, and Hyun-Hwan Jeong. ConSol: Sequential probability ratio testing to find consistent LLM reasoning paths efficiently. *arXiv preprint arXiv:2503.17587*, 2025.

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

[22] Yitao Li. Who drifted: the system or the judge? Anytime-valid attribution in LLM evaluation pipelines. *arXiv preprint arXiv:2606.15474*, 2026.

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

[24] Zifan Lyu, Chahine Nejma, Tobias Wegel, Fanny Yang, and Florian E. Dorner. Cutting LLM evaluation costs with SySRs: A bandit algorithm that provably exploits model similarity. *arXiv preprint arXiv:2606.07726*, 2026.

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

[26] Oded Maron and Andrew W. Moore. Hoeffding races: Accelerating model selection search for classification and function approximation. In *Advances in Neural Information Processing Systems*, volume 6, 1994.

[27] Timothée Mathieu, Riccardo Della Vecchia, Alena Shilova, Matheus Medeiros Centa, Hector Kohler, Odalric-Ambrym Maillard, and Philippe Preux. AdaStop: Adaptive statistical testing for sound comparisons of deep RL agents. *Transactions on Machine Learning Research*, 2024. arXiv:2306.10882.

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

[29] Volodymyr Mnih, Csaba Szepesvári, and Jean-Yves Audibert. Empirical Bernstein stopping. In *Proceedings of the 25th International Conference on Machine Learning*, 2008.

[30] Peter C. O'Brien and Thomas R. Fleming. A multiple testing procedure for clinical trials. *Biometrics*, 35(3):549–556, 1979.

[31] Stuart J. Pocock. Group sequential methods in the design and analysis of clinical trials. *Biometrika*, 64(2):191–199, 1977.

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

[33] Herbert Robbins. Statistical methods related to the law of the iterated logarithm. *Annals of Mathematical Statistics*, 41(5):1397–1409, 1970.

[34] Shuvom Sadhuka, Drew Prinster, Clara Fannjiang, Gabriele Scalia, Bonnie Berger, Aviv Regev, and Hanchen Wang. E-valuator: Reliable agent verifiers with sequential hypothesis testing. *arXiv preprint arXiv:2512.03109*, 2026.

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

[36] Glenn Shafer. Testing by betting: A strategy for statistical and scientific communication. *Journal of the Royal Statistical Society: Series A*, 184(2):407–431, 2021.

[37] David Snyder, Apurva Badithela, Nikolai Matni, George Pappas, Anirudha Majumdar, Masha Itkina, and Haruki Nishimura. Beyond binary success: Sample-efficient and statistically rigorous robot policy comparison. *arXiv preprint arXiv:2603.13616*, 2026.

[38] Elad Tolochinsky, Yaniv Tenzer, and Yaniv Romano. Valid best-model identification for LLM evaluation via low-rank factorization. *arXiv preprint arXiv:2605.10405*, 2026.

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

[40] Abraham Wald. Sequential tests of statistical hypotheses. *Annals of Mathematical Statistics*, 16(2):117–186, 1945.

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

[42] Sida Wang. Measuring all the noises of LLM evals. *arXiv preprint arXiv:2512.21326*, 2025.

[43] Ian Waudby-Smith and Aaditya Ramdas. Confidence sequences for sampling without replacement. In *Advances in Neural Information Processing Systems*, volume 33, 2020. arXiv:2006.04347.

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

[45] Skyler Wu, Yash Nair, and Emmanuel J. Candès. Efficient evaluation of LLM performance with statistical guarantees. *arXiv preprint arXiv:2601.20251*, 2026.

[46] Siyu Zhou, Patrick Vossler, Venkatesh Sivaraman, Yifan Mai, and Jean Feng. Adaptive auditing of AI systems with anytime-valid guarantees. *arXiv preprint arXiv:2605.07002*, 2026.

A Proofs

A.1 Proof of Proposition 1

(i) Superiority. Consider the declaration "$A \succ B$," which fires when $\mathcal{E}_t^A \geq 2/\alpha$ with $\mathcal{E}^A$ built from (1) at level $\alpha/2$. *I.i.d. case:* under any distribution with $\mu \leq 0$, i.e. $\mathbb{E}[X] \leq 1/2$, each factor of (1) with $m = 1/2$ satisfies $\mathbb{E}[1 + \lambda_t(X_t - 1/2) \,|\, \mathcal{F}_{t-1}] = 1 + \lambda_t(\mathbb{E}[X] - 1/2) \leq 1$ because $\lambda_t \geq 0$ is $\mathcal{F}_{t-1}$-measurable. Hence $\mathcal{E}^A$ is a nonnegative supermartingale with $\mathcal{E}_0^A = 1$, and Ville's inequality gives $\mathbb{P}(\sup_t \mathcal{E}_t^A \geq 2/\alpha) \leq \alpha/2$. The declaration is a stopping-time event contained in $\{\sup_t \mathcal{E}_t^A \geq 2/\alpha\}$. *Without-replacement case:* fix a population with mean $\mu_N^x \leq 1/2$ (in x-coordinates) and let m_t be as in (2) with $m = 1/2$. Conditionally on the first $t-1$ draws, the next draw is uniform on the remaining items, so $\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] = (N\mu_N^x - S_{t-1})/(N - t + 1) \leq m_t$, with the inequality because $\mu_N^x \leq 1/2$. Thus $\mathbb{E}[1 + \lambda_t(X_t - m_t) \,|\, \mathcal{F}_{t-1}] \leq 1$ for $\lambda_t \geq 0$ and the same Ville argument applies. The boundary handling (freeze when $m_t > 1$; certain rejection when $m_t < 0$) only ever (a) stops betting, which preserves the supermartingale property, or (b) declares rejection in states no null population can reach, which adds probability zero under the null. The symmetric claim for "$B \succ A$" applies (1) to $1 - X$. When $\mu = 0$ both nulls are true and a union bound over the two level-$\alpha/2$ tests bounds the probability of any winner declaration by α.

(ii) Equivalence. In the confidence-sequence implementation the exit fires only when $[L_t, U_t] \subset (-\delta, \delta)$; if $|\mu| \geq \delta$ then on the event $\{\forall t : \mu \in [L_t, U_t]\}$, of probability at least $1 - \alpha$, no such t exists, so a false equivalence call has probability at most α. In the TOST implementation used for the finite-population target, a tie is declared only when *both* shifted nulls $\{\mu_N \leq -\delta\}$ and $\{\mu_N \geq \delta\}$ are rejected by their level-$\alpha/2$ e-processes; if $|\mu_N| \geq \delta$ one of the two nulls is true, and rejecting a true null has probability at most $\alpha/2 \leq \alpha$ by part (i)'s argument applied at the shifted boundary (2).

(iii) Coverage. The reported interval is $[L_t, U_t] = \{m : \max_{s \leq t} \mathcal{E}_s(m) < 2/\alpha\}$, where for each fixed m the capital process (1) (respectively (2)) is run at level $\alpha/2$ on X to test "mean $\leq m$" and on $1 - X$ to test "mean $\geq m$." For any fixed m whose null is true, the corresponding process is a nonnegative supermartingale by part (i)'s argument, so Ville's inequality gives $\mathbb{P}(\exists t : \mathcal{E}_t(m) \geq 2/\alpha) \leq \alpha/2$. Taking m equal to the true mean (the boundary case of both nulls), the true mean is ever excluded from the running interval with probability at most $\alpha/2$ per side, hence at most α in total by a union bound; the running supremum makes the interval a running intersection, which only removes already-rejected values. Because this invokes only part (i)'s supermartingale property – established separately in the i.i.d. and without-replacement regimes – the statement holds verbatim for μ_N under (2). Computationally, the plug-in truncates λ_s at $c \leq c/m$ for $m \leq 1$, so λ_s does not depend on m, $\log \mathcal{E}_t(m)$ is strictly decreasing in m, the un-rejected set is an interval, and the endpoints are found exactly by nested refinement.

(iv) Verdict–interval consistency. The declaration "$A \succ B$" is the event $\mathcal{E}_t^A \geq 2/\alpha$, and $\mathcal{E}_t^A$ is precisely the member of the inverted family at the null value ($m = 1/2$ in x-coordinates, i.e. 0 in d-coordinates), at the same level, threshold, and λ sequence. The declaration therefore rejects the null value out of the interval by definition: $L_t \geq 0$, and symmetrically $U_t \leq 0$ for "$B \succ A$." Likewise a TOST equivalence declaration rejects both shifted nulls by their own members of the same family, forcing $[L_t, U_t] \subseteq [-\delta, \delta]$. Both implications are set-theoretic identities, not probabilistic events. □

A.2 Validity is also verified empirically

The test suite simulates the worst-case point null and balanced finite populations (2×10^4 streams, horizons to 4,000), measuring ever-rejection rates of 3.2% (i.i.d.) and 3.0% (without replacement) against the 5% budget, ever-miscoverage of the confidence sequence below 0.3% against its 5% budget, and e-BH false discovery proportion 0.5% against 5%, alongside power and agreement checks. Two further groups of tests target parts (iii) and (iv) directly: the without-replacement *reported* interval's ever-miscoverage of μ_N is 2.9% against the 5% budget (with separate checks on a fixed checkpoint ladder and at the protocol's own stopping times), and a verdict–interval consistency test records zero contradictions at superiority stops and zero margin violations at equivalence stops across every simulated stream. We consider executable guarantees a feature: every claim in Proposition 1 has a failing test if an implementation regression breaks it.

B Betting details

The predictable plug-in uses running moments with pseudo-counts, $\hat{\mu}_t = (1/2 + \sum_{s \le t} X_s)/(t+1)$ and $\hat{\sigma}_t^2 = (1/4 + \sum_{s \le t}(X_s - \hat{\mu}_s)^2)/(t+1)$, and

$$\lambda_t \;=\; \min\!\left(\sqrt{\frac{2\log(2/\alpha)}{\hat{\sigma}_{t-1}^2\, t\log(1+t)}},\; c\right), \qquad c = 0.75,$$

following Waudby-Smith and Ramdas [44]; since $c \le c/m_t$ whenever $m_t \le 1$, this respects the constraint $\lambda_t \in [0, c/m_t]$ of (1) (betting is frozen when (2) drifts to $m_t > 1$), and it makes λ_t independent of m_t's value, which is what renders the inversion of §3 exact. All reported intervals are obtained by that inversion: the capital processes (1) (or (2)) at level $\alpha/2$ per side with threshold $2/\alpha$, a running supremum for the time intersection, and nested refinement for the endpoints. The closed-form empirical-Bernstein confidence sequence at level $\alpha/2$ per side, applied to X and to $1 - X$ with running intersection, is used internally by the i.i.d. equivalence exit, which consults an interval at every t; it agrees with the inversion to within a factor 1.35 in width in our checks, and both are valid $(1-\alpha)$ confidence sequences, so the exit's guarantee is unaffected by which one it consults.

C Experimental details

Table 2: Model roster. Exact serving tags reproduce the runs verbatim; all decoding greedy, fixed seed, zero-shot.

Model	serving tag	type	size class	reasoning	benches
Gemma4-E2B	`gemma4:e2b-mlx-bf16`	dense (efficient)	2B-eff.	off	MMLU, GSM8K
Qwen3.6-27B	`qwen3.6:27b-mlx-bf16`	dense	27B	off	MMLU
Gemma4-31B	`gemma4:31b-mlx-bf16`	dense	31B	off	MMLU
Qwen3.6-35B-A3B	`qwen3.6:35b-a3b-mlx-bf16`	MoE	35B/3B-act.	off	MMLU, GSM8K
Qwen3.6-35B-A3B-Coding	`qwen3.6:35b-a3b-coding-bf16`	MoE	35B/3B-act.	off	MMLU, GSM8K
Llama4-Scout	`llama4:scout`	MoE	109B/17B-act.	n/a	MMLU, GSM8K
GPT-OSS-120B	`gpt-oss:120b`	MoE	117B/5B-act.	low	MMLU, GSM8K

Models.

Benchmarks. MMLU subjects (test split): abstract algebra, college mathematics, high-school mathematics, formal logic, high-school statistics, econometrics, high-school macroeconomics, conceptual physics, astronomy, philosophy, machine learning, high-school computer science, international law, management; pooled (2,450 items) and subsampled to 1,000 with a fixed seed. GSM8K: 300 test items subsampled with the same seed policy. Prompts are zero-shot; the MMLU instruction requests a bare letter; the GSM8K instruction requests brief reasoning and a final line `Answer: <number>`. Extraction rules and their unit tests ship with the code; items whose model call failed after three retries (a per-run count reported in the logs; 0 failed calls; 220 of 8,500 responses (2.6%) yielded no extractable answer despite generous token budgets and are scored 0, a per-model count reported in Table 1) are scored 0 for the failing model and flagged.

Protocol parameters. $\alpha = 0.05$ split evenly across the two one-sided exits; $\delta = 0.03$; without-replacement mode with N equal to the collected benchmark size; 1,000 replay permutations per pair; e-BH at $\alpha = 0.05$ over all pairs per benchmark.

D Full per-pair results

E Full simulation tables

F Group-sequential baseline: full tables

Full head-to-head results for §4.1 (reject rate / mean items per cell), followed by the sparse-stream negative control. We had conjectured that the normal approximation underlying group-sequential boundaries would break down on sparse paired streams (models agreeing on nearly every item); the experiment does not support this – GST's Type I error never exceeds 5.3%, and sparsity only makes every procedure more conservative. Our case against reused boundaries therefore rests entirely on

Table 3: Per-pair replay results, MMLU (1,000 random orders per pair; $\alpha = 0.05$, $\delta = 0.03$). Gap is the full-benchmark difference in points; verdict legend: A = first model, B = second, U = unresolved by the full benchmark.

Pair	gap	dis.	full	stop_{50}	items	tokens	A/B/tie/none	flip
GE2B v OSS120	-37.3	0.41	B	32	3%	3%	0.00/1.00/0.00/0.00	0.000
GE2B v Q27	-32.2	0.40	B	36	4%	4%	0.00/1.00/0.00/0.00	0.000
G31 v GE2B	+29.9	0.38	A	40	4%	4%	1.00/0.00/0.00/0.00	0.000
GE2B v Q35A3c	-28.9	0.39	B	42	4%	4%	0.00/1.00/0.00/0.00	0.000
GE2B v Q35A3	-28.6	0.39	B	42	4%	4%	0.00/1.00/0.00/0.00	0.000
GE2B v L4S	-26.8	0.36	B	45	4%	4%	0.00/1.00/0.00/0.00	0.000
OSS120 v L4S	+10.5	0.17	A	118	12%	12%	1.00/0.00/0.00/0.00	0.000
OSS120 v Q35A3	+8.7	0.17	A	156	16%	16%	1.00/0.00/0.00/0.00	0.000
OSS120 v Q35A3c	+8.4	0.17	A	161	16%	16%	1.00/0.00/0.00/0.00	0.000
G31 v OSS120	-7.4	0.16	B	203	20%	20%	0.00/1.00/0.00/0.00	0.000
L4S v Q27	-5.4	0.19	B	385	38%	38%	0.00/1.00/0.00/0.00	0.000
OSS120 v Q27	+5.1	0.15	A	330	33%	33%	1.00/0.00/0.00/0.00	0.000
Q27 v Q35A3	+3.6	0.12	A	462	46%	46%	1.00/0.00/0.00/0.00	0.000
Q27 v Q35A3c	+3.3	0.12	A	512	51%	51%	0.99/0.00/0.01/0.00	0.000
G31 v L4S	+3.1	0.20	A	779	78%	78%	0.99/0.00/0.01/0.00	0.000
G31 v Q27	-2.3	0.13	B	760	76%	76%	0.00/0.93/0.07/0.00	0.000
L4S v Q35A3c	-2.1	0.20	U	927	93%	93%	0.00/0.89/0.11/0.00	0.002
L4S v Q35A3	-1.8	0.19	U	927	93%	93%	0.00/0.71/0.29/0.00	0.000
G31 v Q35A3	+1.3	0.14	U	873	87%	87%	0.39/0.00/0.61/0.00	0.001
G31 v Q35A3c	+1.0	0.13	U	831	83%	83%	0.17/0.00/0.83/0.00	0.004
Q35A3c v Q35A3	+0.3	0.01	U	331	33%	33%	0.00/0.00/1.00/0.00	0.000

Table 4: Per-pair replay results, GSM8K (1,000 random orders per pair; $\alpha = 0.05$, $\delta = 0.03$). Gap is the full-benchmark difference in points; verdict legend: A = first model, B = second, U = unresolved by the full benchmark.

Pair	gap	dis.	full	stop_{50}	items	tokens	A/B/tie/none	flip
GE2B v L4S	-11.3	0.12	B	86	29%	29%	0.00/1.00/0.00/0.00	0.000
GE2B v OSS120	-10.7	0.12	B	93	31%	31%	0.00/1.00/0.00/0.00	0.000
GE2B v Q35A3c	-10.3	0.12	B	94	31%	31%	0.00/1.00/0.00/0.00	0.000
GE2B v Q35A3	-10.3	0.12	B	95	32%	32%	0.00/1.00/0.00/0.00	0.000
L4S v Q35A3c	+1.0	0.04	U	261	87%	87%	0.14/0.00/0.86/0.00	0.000
L4S v Q35A3	+1.0	0.04	U	260	87%	86%	0.13/0.00/0.87/0.00	0.000
OSS120 v L4S	-0.7	0.03	U	242	81%	81%	0.00/0.02/0.98/0.00	0.000
OSS120 v Q35A3c	+0.3	0.04	U	241	80%	80%	0.00/0.00/1.00/0.00	0.000
OSS120 v Q35A3	+0.3	0.04	U	239	80%	80%	0.01/0.00/0.99/0.00	0.000
Q35A3c v Q35A3	+0.0	0.03	U	228	76%	76%	0.00/0.00/1.00/0.00	0.000

the look schedule and the horizon (§4.1), not on approximation quality. The boundary machinery itself is validated against the published constants of Jennison and Turnbull [15] (maximum absolute deviation 0.0011 over twelve constants) and by independent Monte Carlo crossing simulations.

Table 5: Stopping behavior vs. effect size and disagreement (4,000 reps each; horizon 8,000). Right column: oracle fixed-n at 90% power given the true effect.

disagr.	effect	win rate	stop_{50}	stop_{10-90}	oracle n
0.2	0.02	0.585	6487	983–8000	5244
0.2	0.03	0.943	2420	394–6691	2325
0.2	0.05	1.000	680	174–1781	831
0.2	0.08	1.000	236	96–552	318
0.2	0.12	1.000	116	65–215	136
0.2	0.20	1.000	58	41–80	43
0.4	0.02	0.219	8000	3105–8000	10497
0.4	0.03	0.585	6627	989–8000	4660
0.4	0.05	0.977	1828	268–5041	1671
0.4	0.08	1.000	580	121–1557	647
0.4	0.12	1.000	208	63–548	282
0.4	0.20	1.000	76	39–161	95

Equivalence certification under exact ties ($\Delta = 0$, $\rho = 0.3$, horizon 12,000, 3,000 reps).

margin δ	tie rate	stop_{50}	stop_{90}	false winner
0.02	0.000	–	–	0.0350
0.03	0.419	10136	11654	0.0320
0.05	0.968	3727	5908	0.0317

Table 6: Head-to-head comparison under a correctly specified horizon of $4,000$ items (reject rate / mean items consumed; $5,000$ replications per cell). GST boundaries use $K = 5$ equally spaced looks; mSPRT is the mixture sequential probability ratio test with prior scale τ. Rows with effect 0.00 report Type I error.

disagr.	effect	SAVE	Pocock	OBF	$\text{mSPRT}_{.02}$	$\text{mSPRT}_{.05}$	$\text{mSPRT}_{.15}$
0.2	0.00	0.03/3,905	0.05/3,904	0.05/3,973	0.02/3,968	0.03/3,905	0.03/3,880
0.2	0.02	0.32/3,312	0.72/2,822	0.79/3,176	0.55/3,051	0.52/2,966	0.41/3,185
0.2	0.03	0.72/2,389	0.97/1,843	0.99/2,418	0.93/1,971	0.91/1,839	0.85/2,092
0.2	0.05	1.00/861	1.00/988	1.00/1,615	1.00/862	1.00/678	1.00/736
0.2	0.08	1.00/296	1.00/802	1.00/1,023	1.00/429	1.00/284	1.00/272
0.2	0.12	1.00/131	1.00/800	1.00/800	1.00/242	1.00/140	1.00/116
0.2	0.20	1.00/60	1.00/800	1.00/800	1.00/114	1.00/59	1.00/42
0.4	0.00	0.03/3,884	0.04/3,912	0.05/3,975	0.01/3,983	0.03/3,917	0.04/3,847
0.4	0.02	0.13/3,682	0.42/3,363	0.50/3,608	0.23/3,680	0.27/3,471	0.21/3,516
0.4	0.03	0.33/3,295	0.77/2,672	0.84/3,085	0.59/3,081	0.61/2,786	0.52/2,907
0.4	0.05	0.81/2,042	1.00/1,492	1.00/2,118	0.98/1,741	0.98/1,356	0.97/1,439
0.4	0.08	1.00/734	1.00/897	1.00/1,487	1.00/870	1.00/567	1.00/529
0.4	0.12	1.00/270	1.00/801	1.00/947	1.00/499	1.00/283	1.00/225
0.4	0.20	1.00/91	1.00/800	1.00/800	1.00/254	1.00/125	1.00/83

Negative control: Type I error (%) on sparse paired streams under the null, by model agreement rate (horizon $1,000$, $K = 5$ looks for GST). We conjectured the normal approximation behind GST would fail here; it does not – sparsity only makes every method conservative.

agreement	Pocock	OBF	Lan–DeMets	SAVE (continuous)
0.995	0.93	2.37	0.93	0.00
0.990	2.51	4.61	2.51	0.02
0.980	3.70	5.31	3.91	0.32
0.950	4.75	5.00	4.85	1.07
0.900	4.93	4.92	4.88	1.64
0.800	5.03	5.07	5.03	2.32
0.600	4.93	4.95	4.93	2.84