Cost-Aware Anytime Racing: Paying for Language-Model Benchmark Decisions by the Dollar

Andrew Sheng-Han Huang*
National Taiwan University
johnshhuang1111@gmail.com

Abstract

Sequential anytime-valid evaluation lets practitioners compare two language models by streaming benchmark items and stopping the moment a verdict is certified, with an error guarantee that survives continuous monitoring. Existing procedures measure progress in *items*: stop after a median fraction of the benchmark. But items are not equally expensive. On real logs, a long chain-of-thought item costs tens to hundreds of times a one-token multiple-choice answer; the coefficient of variation of per-item output-token cost reaches 1.54 within a single model. "Median items to decision" therefore mis-states the quantity a practitioner actually pays — *dollars*. We introduce CARA (Cost-Aware Anytime Racing): a paired betting confidence sequence whose next evaluation is routed by a *predictable*, cost-aware rule, with a total-cost stopping rule. Our first contribution is a theorem: the paired e-process remains anytime-valid under any *predictable* selection policy whose every admissible next draw has conditional mean obeying the null, so routing the budget by such a rule does not cost a single unit of the Type-I guarantee — a hypothesis that explicitly *excludes* anticipating rules that peek at the next item's outcome and stratified rules that can over-select a super-null group. The deployable instance routes by lowest *known* per-item cost first (the predictable info-per-dollar rule on a featureless population), using only known costs and never the next label. We verify validity empirically two ways: under continuous monitoring with online cost-aware selection the false-superiority rate is 0.036, and the shipped paired procedure run at its exact two-sided thresholds has either-direction false-superiority 0.026, both against a nominal 0.05 (Monte-Carlo half-width 0.010). Our second contribution is a measurement on 31 real model-pair comparisons (seven open-weight models on MMLU, five on GSM8K). The *deployable* predictable policy reaches the identical verdict at a median 52.0% lower output-token cost than cost-blind random ordering (positive on 25 of 31 pairs, above 20% on 77%), and the saving is conditional on heterogeneity — median 79.0% on heavy-tailed MMLU, 40.4% on lower-dispersion GSM8K, rank-tracking a pair's cost dispersion (Spearman $\rho = 0.80$) and collapsing to a near-zero 1.14% under uniform cost (Monte-Carlo residual) exactly as the theory demands. A label-omniscient *oracle* ordering (realized $|d|$/cost), reported only as a best-case upper-bound envelope that carries *no* Type-I guarantee, would reach a median 84.5% — the headroom a perfect informativeness predictor could add. We also characterize a failure mode of the greedy rule: on 6 of 31 pairs it *over*-spends by deferring expensive items that carry the discriminating signal. We argue that budgeted model comparison should be priced in dollars, not items, with the caveat that the routing rule must be robust to where the signal sits.

1 Introduction

Comparing two language models on a benchmark is the field's workhorse measurement, and a recent line of work has made it statistically honest under the way it is actually run. Practitioners peek: they evaluate, glance at the running difference, and stop when the answer looks clear. Fixed-sample tests

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

are invalid under such optional stopping, but *anytime-valid* procedures — confidence sequences and e-processes built from test martingales [3, 6, 14, 19] — are valid at *every* stopping time. Applied to model comparison, they let one stream items in random order and stop the instant superiority or practical equivalence is certified, with the same guarantee no matter how often one looks [12].

These procedures report their efficiency in *items*: a verdict after a median fraction of the benchmark's questions. That accounting silently assumes every item costs the same. It does not. Modern evaluation cost is dominated by generation, and generation length is wildly heterogeneous across items: a model emits a single token for an easy multiple-choice answer and hundreds of chain-of-thought tokens for a hard one. In the real per-item logs we reanalyze (§4), the within-model coefficient of variation (CV) of output-token cost ranges from 0.46 to 1.54 on MMLU; a pair's paired-cost CV reaches 1.54. When the resource that varies hundred-fold is the resource you are billed for, "items to decision" is the wrong meter. A procedure that stops after few *items* can still be expensive if those items happen to be the long ones; a procedure that stops after more, cheaper items can be far cheaper in *dollars*.

This paper. We re-pose sequential model comparison as a *cost-aware race*: spend the next evaluation dollar where it buys the most expected statistical evidence, and stop on cumulative *cost*. The method, CARA, keeps the paired betting confidence sequence of anytime-valid evaluation intact and changes only *which item is evaluated next* and *what quantity the stopping rule meters*. Our contributions are a guarantee, a measurement, and a correction:

- **A validity theorem under predictable cost-aware sampling** (§3, Theorem 1). The paired e-process $K_t = \prod_{s \le t} \left(1 + \lambda_s (x_{\pi(s)} - \frac{1}{2})\right)$ with predictable bets λ_s is a non-negative supermartingale under the null for any *predictable* selection policy π *whose every admissible next draw has conditional mean* $\le m_0$ *under the null*. Ville's inequality then gives the identical anytime Type-I guarantee. We state the hypothesis precisely and what it excludes (anticipating rules; super-null strata), so the deployable lowest-known-cost-first rule is covered while the label-peeking oracle is not. We confirm validity empirically: online continuous-monitoring false-superiority 0.036, and the shipped paired procedure at its exact two-sided thresholds 0.026, both at nominal 0.05 (§5).
- **A measured dollar saving with a DEPLOYABLE policy, and its boundary** (§5). Across 31 real comparisons, the predictable lowest-cost-first policy — known costs only, never a peeked label — certifies the *same* verdict at a median 52.0% lower output-token cost than cost-blind random ordering, positive on 25/31 pairs. The saving is conditional on heterogeneity (median 79.0% on heavy-tailed MMLU, 40.4% on GSM8K) and we characterize a failure mode in which it overspends (6/31 pairs). We also report a label-omniscient oracle ordering *only as an upper-bound envelope* (median 84.5%, no Type-I guarantee) to bound the headroom of a perfect informativeness predictor. No prior work measures the deployable token-dollar saving of *anytime* paired racing on real LLM-evaluation logs.
- **The item-versus-dollar correction.** Because the saving is driven by cost dispersion (Spearman $\rho = 0.80$), the item-based headline of prior anytime evaluation *understates* achievable efficiency where items are heterogeneous and is exactly right where they are uniform (the deployable saving collapses to 1.14% under constant cost; §5). We quantify when the dollar axis matters: a simulation sweep locates the heterogeneity level (cost CV ≈ 0.25) beyond which the median dollar saving exceeds 20%.

The method is deliberately conservative: CARA *reduces exactly* to the item-based procedure when costs are uniform, so it can never do statistically worse and changes the bill only when there is heterogeneity to capitalize on. All numbers are produced by the analysis scripts, written to JSON, and injected as macros; nothing in the prose is hand-typed.

2 Related Work

CARA sits at the intersection of three literatures, each of which occupies two of its three defining elements — anytime validity by betting, heterogeneous *per-item* cost, and paired LLM-benchmark comparison measured on real logs — but none of which occupies all three. Table 1 resolves the nearest neighbors.

Table 1: Nearest neighbors. P1 = anytime-valid by betting (valid at every peek); P2 = heterogeneous *per-item* evaluation cost / dollar budget; P3 = paired LLM benchmark comparison measured on real logs. CARA is the conjunction.

Work	P1	P2	P3	Our delta (a noun)
SySRs [10]	no	no	part	anytime CS + per-token dollars
CABAI [8]	no	part	no	per-*item* cost + anytime + LLM logs
BAI w/ LLM judges [1]	yes	no	no	cost = tokens/time, not audit labels
Active betting 2-sample [7]	yes	no	no	selection by *cost*; dollar stop rule
Multi-armed betting [15]	yes	no	no	dollars not samples; paired superiority
NHSHT [17]	no	part	no	betting CS + paired LLM real-cost
Anytime BAI [9]	part	no	no	per-item cost + paired LLM dollars
Balancing perf./cost [4]	no	part	no	per-item heterogeneity + anytime CS
Item-based anytime eval	yes	no	part	the dollar axis + cost-routing theorem

Anytime-valid evaluation of language models. Confidence sequences and e-processes give time-uniform inference [6, 14] and have been brought to LLM evaluation to add honest error bars and early stopping under peeking [12]. These procedures, and the paired sampling-without-replacement protocol we build on, count progress in *items*; CARA re-meters them in dollars and proves the guarantee survives cost-aware reordering.

Racing and cost-aware best-arm identification. Racing prunes inferior candidates early [11, 13]. Cost-aware best-arm identification attaches a cost to each arm and minimizes expected cost to identify the best [4, 8, 9]; Kanarios et al. [8] show that ignoring heterogeneous action cost is sub-optimal. These are *fixed-confidence* or fixed-budget procedures with a single terminal decision and per-*arm* (not per-*item*) cost; CARA is continuous-monitoring anytime-valid, the cost is the per-item generation cost that varies hundred-fold *within* an arm, and it targets paired superiority between two models rather than identifying one best arm.

Active and cost-aware sequential testing. Vershinin et al. [16, 17] study active sequential hypothesis testing with non-homogeneous *action* costs and the information-per-cost design principle, but in the classical Chernoff framework (asymptotic $\log 1/\delta$ scaling), not anytime-valid betting, and not for LLM evaluation. Closest in machinery, Hsu and Shekhar [7] and Sandoval et al. [15] perform sequential testing-by-betting with *adaptive source selection*: the former for nonparametric two-sample testing over distributionally heterogeneous sources, the latter for a global "all arms null" detection with rejection-*time* optimality. Both assume *uniform per-pull cost* and optimize sample count; CARA routes selection by *dollar cost*, meters stopping in dollars, targets paired model superiority, and measures the result on real LLM logs.

Cost-aware LLM evaluation. Closest in application, Lyu et al. [10] cut LLM-evaluation cost via a fixed-budget Successive Rejects bandit with paired comparisons, leveraging model similarity; their budget is a fixed *number of query–model evaluations* (uniform unit cost) and the guarantee is fixed-budget, not anytime. Ao et al. [1] form anytime-valid confidence sequences for LLM-judge best-arm identification, but the scarce resource is expensive *human-audit labels*, not the per-item token/time cost CARA optimizes. Methods that shorten reasoning to make a single model cheaper to run are orthogonal: they reduce the cost of *running* a model, not the cost of a *comparison decision*.

3 Method

3.1 Setup and the paired betting confidence sequence

Two models A, B are compared on a benchmark with item universe $\mathcal{I}$. For item i, let $g_A(i), g_B(i) \in \{0, 1\}$ be correctness and $\mathrm{cost}_A(i), \mathrm{cost}_B(i) > 0$ the per-item evaluation cost (output tokens, or wall-time). The paired per-item score difference and the comparison's per-item cost are

$$d(i) = g_A(i) - g_B(i) \in \{-1, 0, 1\}, \qquad \mathrm{cost}(i) = \mathrm{cost}_A(i) + \mathrm{cost}_B(i), \tag{1}$$

the latter because forming $d(i)$ requires evaluating item i on *both* models. Map d to $x = (d+1)/2 \in [0,1]$; the superiority null $H_0 : \mathbb{E}[d] \le 0$ becomes $H_0 : \mu_x := \mathbb{E}[x] \le \frac{1}{2}$. Following the betting construction of Waudby-Smith and Ramdas [19], with a predictable plug-in bet $\lambda_s \in [0, c/m_0]$ (a function of $x_1, \ldots, x_{s-1}$ only; we use the WSR variance-adaptive recipe, $c = 0.75$), define the one-sided capital process for $H_0 : \mu_x \le m_0 = \frac{1}{2}$:

$$K_t \;=\; \prod_{s=1}^{t} \big(1 + \lambda_s \,(x_{\pi(s)} - m_0)\big), \tag{2}$$

where $\pi(s) \in \mathcal{I}$ is the (possibly adaptive) index of the item evaluated at step s. A symmetric process on $1 - x$ tests $B > A$, and an empirical-Bernstein confidence sequence [19] certifies practical equivalence (a two-one-sided-test tie within $\pm\delta$). This is the standard paired anytime-valid evaluation protocol; the only new degree of freedom is the *selection policy* π.

3.2 The cost-aware racing rule

Definition 1 (Predictable information-per-dollar selection)**.** *At step s, given the predictable bet λ_s and the set of not-yet-drawn candidate items, the cost-aware policy $\pi^\star$ selects the candidate i maximizing the predictable expected betting log-gain per unit cost,*

$$\pi^\star(s) \in \arg\max_i \; \frac{\big|\mathbb{E}_{s-1}\log\big(1 + \lambda_s(x_i - m_0)\big)\big|}{\mathrm{cost}(i)}, \tag{3}$$

*where $\mathbb{E}_{s-1}$ is a $\mathcal{F}_{s-1}$-measurable estimate that uses only quantities known by step s — the candidate group's running disagreement-rate from past draws and the candidate's known cost — and **never the realized label** $d(i)$ **of the undrawn candidate**. The cost-blind baseline draws the next candidate uniformly at random. The stopping rule halts at the first t with $K_t \ge 2/\alpha$ (superiority), a certified tie, or cumulative cost $\sum_{s \le t} \mathrm{cost}(\pi(s)) \ge B$ (budget exhausted).*

Definition 2 (Deployable lowest-cost-first instance)**.** *On a flat item population with no usable per-item features, the only predictable estimate of an undrawn item's informativeness is a single running scalar (the past disagreement rate), identical across all remaining candidates;* (3) *then ranks purely by* $1/\mathrm{cost}(i)$*, i.e. it draws the lowest known cost item first. This lowest-cost-first rule is the deployable headline policy: it is a function of known costs alone, hence predictable (indeed $\mathcal{F}_0$-measurable when costs are known a priori), and is covered by Theorem 1. (When the population has cost-correlated strata whose disagreement rates differ, the online stratified racer §5 leverages that extra predictable structure; lowest-cost-first is the feature-free floor.)*

The intuition: evidence accrues through the betting increments $\log(1 + \lambda_s(x_{\pi(s)} - m_0))$, which are large for *informative* items (where the models disagree, $|d| = 1$) and zero for ties. A predictable policy cannot see which undrawn item is informative, so on a featureless population it maximizes expected evidence-per-dollar by buying the cheapest items first — extracting more independent draws per unit budget. Crucially, the policy is *predictable*: it never inspects the hidden label of the item it is about to draw.

The deployable cost claim, and an oracle envelope. The headline comparison is *predictable-versus-predictable*: the deployable lowest-cost-first policy against a uniformly random cost-blind order. Both are implementable (neither peeks at a label); their only difference is whether known cost is used to order, so the gap is the *deployable cost-ordering saving*, and it must vanish when costs are uniform (then lowest-cost-first *is* a uniformly random permutation). This is the honest reduction test (§5, 1.14% under uniform cost). Separately, and labeled as such, we report an *oracle* ordering by realized $|d(i)|/\mathrm{cost}(i)$. The oracle *anticipates* — it ranks undrawn items by their hidden labels — so it is **not** a predictable policy, is **not** covered by Theorem 1, and carries **no** Type-I guarantee on its own. We use it only as a best-case *upper-bound envelope* on a fixed finite population: the cost a label-omniscient scheduler could have reached, bounding the headroom of a perfect informativeness predictor. We never advertise the oracle saving as the deployable result.

3.3 Anytime validity is preserved under cost-aware selection

Theorem 1 (Anytime validity under predictable, null-admissible cost-aware sampling)**.** *Let $(\mathcal{F}_s)_{s\geq 0}$ be a filtration. Suppose at each step s the selection $\pi(s)$ and the bet $\lambda_s \in [0, c/m_0]$ $(c < 1)$ are $\mathcal{F}_{s-1}$-measurable (predictable), and that the policy is null-admissible: under H_0 every selection it can make has conditional mean obeying the null,*

$$\mathbb{E}[x_{\pi(s)} \mid \mathcal{F}_{s-1}] \leq m_0 \qquad \text{a.s. under } H_0. \tag{4}$$

Then K_t in (2) is a non-negative supermartingale with $K_0 = 1$, and for any $\alpha \in (0, 1)$ and any stopping time τ,

$$\mathbb{P}_{H_0}\Big(\exists\, t \geq 1 : K_t \geq \tfrac{1}{\alpha}\Big) \;\leq\; \alpha, \tag{5}$$

for every predictable, null-admissible π. The same holds for the symmetric process, and (by a union bound over the two one-sided tests, each at level $\alpha/2$ with threshold $2/\alpha$) the paired superiority decision has ever-false-superiority $\leq \alpha$.

Proof. Non-negativity: $\lambda_s \leq c/m_0$ and $x_{\pi(s)} \geq 0$ give $1 + \lambda_s(x_{\pi(s)} - m_0) \geq 1 - \lambda_s m_0 \geq 1 - c > 0$ for $c < 1$. Write $\xi_s = 1 + \lambda_s(x_{\pi(s)} - m_0)$. Since λ_s and the choice $\pi(s)$ are $\mathcal{F}_{s-1}$-measurable,

$$\mathbb{E}[\xi_s \mid \mathcal{F}_{s-1}] = 1 + \lambda_s\big(\mathbb{E}[x_{\pi(s)} \mid \mathcal{F}_{s-1}] - m_0\big) \leq 1, \tag{6}$$

because $\lambda_s \geq 0$ and, by null-admissibility (4), $\mathbb{E}[x_{\pi(s)} \mid \mathcal{F}_{s-1}] \leq m_0$ under H_0. Hence $\mathbb{E}[K_t \mid \mathcal{F}_{t-1}] = K_{t-1}\,\mathbb{E}[\xi_t \mid \mathcal{F}_{t-1}] \leq K_{t-1}$, so (K_t) is a non-negative supermartingale with $K_0 = 1$. Ville's inequality [18] for non-negative supermartingales gives $\mathbb{P}(\sup_t K_t \geq 1/\alpha) \leq \mathbb{E}[K_0]/(1/\alpha) = \alpha$. The argument uses no property of π beyond predictability and (4), so it holds for the deployable lowest-cost-first policy of Definition 2. □ □

Remark 1 (What the hypotheses exclude)**.** *The theorem is not "any adaptive policy is safe." Two exclusions matter. (i)* ***Predictability.*** *$\pi(s)$ must be chosen from $\mathcal{F}_{s-1}$; a rule that ranks undrawn items by their realized labels — like the oracle $|d(i)|/\mathrm{cost}(i)$ ordering — is anticipating, violates measurability, and is* ***not*** *covered (we use it only as an upper-bound envelope, §3). (ii)* ***Null-admissibility*** *(4). On a single exchangeable pool the pooled null $\mathbb{E}[x] \leq m_0$ makes every predictable selection automatically null-admissible. But if items are partitioned into strata with different conditional means and the policy may select a stratum whose conditional mean exceeds m_0 even while the global mean satisfies the null, (4) fails and the guarantee can break: a predictable policy that preferentially mines a locally super-null stratum is outside scope. The equal-models pooled null targeted by our false-superiority experiments (§5) satisfies (4) by construction; deployments over heterogeneous strata must verify it (e.g. by a conservative pooled bet) or restrict π to strata that cannot beat the null.*

Remark 2 (Why the bill, not the guarantee, is what changes)**.** *Theorem 1 isolates exactly what predictable cost-awareness can and cannot do. It cannot inflate Type-I error: the supermartingale property is indifferent to which null-admissible item is drawn next. What it changes is the realized cost path $\sum_s \mathrm{cost}(\pi(s))$ to a fixed level of evidence: routing by (3) front-loads cheap items, so the e-process crosses its threshold after fewer dollars. The equal-models pooled null targeted by our false-superiority experiment (§5) satisfies null-admissibility (4) by construction (Remark 1).*

Proposition 1 (Exact reduction under uniform cost)**.** *If $\mathrm{cost}(i) \equiv \kappa$ is constant, the deployable lowest-cost-first rule (Definition 2) has no cost signal to break ties on: every item shares the same cost, so the induced order is a uniformly random permutation, distributionally identical to the cost-blind baseline. Hence the deployable cost-attributable saving (lowest-cost-first versus random) is zero in expectation: where there is no cost heterogeneity, there is no dollar claim.*

This is the conservative guarantee. We verify Proposition 1 numerically (the deployable saving collapses to 1.14% under constant cost, a residual within Monte-Carlo permutation noise) and check the e-process matches the reference item-based recipe to machine precision ($< 10^{-12}$ max log-capital difference) in §5.

The reanalysis estimand. On a *fixed, finite* real population (our logs), the deployable headline measures the cost the predictable lowest-cost-first schedule *would have* incurred, ordering the realized items by known cost ascending and replaying them through the identical e-process and stopping rule — a counterfactual that uses no future label and is therefore faithfully implementable. We separately report the anticipating oracle ordering $|d(i)|/\text{cost}(i)$ *only* as a best-case envelope (it peeks at labels), and we audit that envelope's own false-superiority rate on null populations purely to bound how far the anticipation could distort a null replay (passing that audit does *not* make the oracle a valid online policy).

4 Data

We reanalyze real per-item evaluation logs (output tokens, prompt tokens, wall-time, correctness) for seven open-weight models (2B–120B class) on a 1000-item MMLU subset [5] and five of them on a 300-item GSM8K subset [2], collected for a prior anytime-evaluation study and reused here with no new inference. Item identifiers are aligned across models, so every model covers the same shared item set; this yields 21 MMLU pairs and 10 GSM8K pairs, 31 in total. The per-item cost is real and heavy-tailed: the paired output-token cost CV ranges $[0.46, 1.54]$ on MMLU (median 1.26) and $[0.39, 0.48]$ on GSM8K (median 0.42), giving a built-in contrast between a heterogeneous-cost benchmark and a near-uniform one. A schema and id-alignment check passes with zero problems. The primary cost metric is output tokens (the dominant, model-controllable billing driver); wall-time is a hardware-dependent robustness check.

5 Experiments

All statistics use $\alpha = 0.05$, tie margin $\delta = 0.05$, betting truncation $c = 0.75$, seed printed by the scripts. The lie-detector suite is written to be able to fail and is logged to JSON.

5.1 Lie-detectors: validity, the break-test, and reduction

The two anytime guarantees the paired procedure bundles are tested *separately*: superiority error control (the betting e-processes) and equivalence coverage (the confidence sequence for the tie exit). We never advertise one as the other.

Superiority Type-I, three ways. (a) *Online predictable selection.* Two equal-accuracy models, items split into three pools with different cost regimes, the online predictable information-per-dollar policy (Definition 1), single-direction race monitored at every step: ever-false-superiority 0.036 at nominal 0.05 (Monte-Carlo half-width 0.010). (b) *The shipped paired procedure at its exact thresholds.* The deployed verdict runs *two* one-sided e-processes, each at threshold $2/\alpha$ (per-side level $\alpha/2$), so the relevant event is a false superiority in *either* direction. We simulate the pooled null and drive the exact two-sided procedure (predictable lowest-cost-first order) over 6000 streams: either-direction ever-false-superiority is 0.026, confirming the union-bounded guarantee on the procedure as shipped (not a one-sided proxy). (c) *The oracle envelope, audited.* The anticipating oracle ordering used only for the upper-bound envelope is audited on null populations and yields 0.032 either-direction — bounding how far its label-peeking could distort a null replay; this does *not* promote the oracle to a valid online policy.

Equivalence (tie) is a coverage guarantee, not superiority. The tie exit fires when the $(1 - \alpha)$ empirical-Bernstein confidence sequence for μ_d is contained in $(-\delta, \delta)$; a wrong tie is a CS *coverage* failure, governed by the coverage level, not by superiority control. With the truth placed just *outside* the margin ($\mu_d = 1.5\delta$), the false-tie rate is 0.000 at nominal 0.05, and the CS's worst-case ever-miscoverage is 0.003 across three mean settings — both within 0.05+MC. Superiority and equivalence are thus controlled by different, separately verified guarantees.

Break-test: the cost-blind baseline must not beat the DEPLOYABLE policy. On populations with a real effect *and* heavy-tailed cost, we run the cost-blind random order and the deployable lowest-cost-first order through the *identical* e-process and stopping rule. Both are predictable; the only difference is whether known cost orders the draws, so any win is the deployable cost-ordering effect, not a label advantage. Lowest-cost-first is not beaten: its median cost-to-decision is a fraction 0.25

of random's (a 75.1% deployable saving) and it is cheaper-or-equal in a fraction 0.808 of populations — so the headline is not a strawman-baseline artifact. For reference (not a deployable claim), on the same populations the label-omniscient oracle-versus-info-only cost-ordering gap is 74.7%, the extra headroom a perfect informativeness predictor would add on top.

Reduction. Under constant per-item cost the deployable lowest-cost-first order degenerates to a uniformly random permutation, so its median saving versus the random baseline is 1.14% (Proposition 1; the residual is Monte-Carlo permutation noise): with no cost heterogeneity there is no cost claim. The e-process reproduces the reference item-based betting recipe to $< 10^{-12}$ maximum log-capital difference, and the deployable policy has power 0.899 to certify a true $\mu_d = 0.08$ effect (slightly below the anticipating oracle's, as expected for a feature-free rule).

5.2 Headline: deployable dollar savings on real model-pair comparisons

For each of the 31 real pairs we replay allocation policies through the same paired confidence sequence and record total output-token cost to the certified verdict. The *headline* is the deployable, predictable lowest-cost-first policy versus the cost-blind random baseline (mean over 12 orderings) — both implementable, neither peeking at a label. Separately and clearly labeled, we report the anticipating oracle ($|d|$/cost) as an upper-bound envelope only.

The deployable saving versus a cost-blind baseline. The predictable lowest-cost-first policy reaches the *same* verdict at a median 52.0% lower output-token cost than cost-blind random ordering (IQR [40.4, 99.0]%), positive in 25 of 31 pairs and above 20% in 24 of 31 (Figure 1). This is the number a practitioner can actually realize: it routes by known cost and never inspects a future outcome.

The result is conditional on heterogeneity. The saving splits sharply by a pair's cost dispersion. On heavy-tailed MMLU the median deployable saving is 79.0% (86% of pairs above 20%); on lower-dispersion GSM8K it is 40.4% (60% above 20%). It rank-tracks the paired cost CV (Spearman $\rho = 0.80$; Figure 2): where items cost nearly the same there is little to save, exactly as the reduction predicts.

The oracle envelope (not deployable). A label-omniscient oracle that ranks undrawn items by realized $|d|$/cost would reach a median 84.5% saving versus random (MMLU 97.7%, GSM8K −0.4%). This ordering peeks at outcomes, is not predictable, and carries no Type-I guarantee (Remark 1); we report it solely to bound the headroom a perfect informativeness predictor could add over the deployable rule — on MMLU that gap is large, indicating room for a better-but-still-predictable router (future work).

A characterized failure mode (the greedy rule is not optimal). Of the 31 pairs, 6 have a *negative* deployable saving — the lowest-cost-first policy is more expensive than random. The failures are concentrated (6 of 6 strongly-negative pairs involve either the tiny model that emits a near-constant one token, whose cheap items are uninformative, or the expensive-but-accurate outlier). The mechanism is the limit of a myopic rule: lowest-cost-first *defers* expensive items, but for those pairs the discriminating evidence lives precisely in the expensive items, so the policy burns budget on cheap, uninformative draws before being forced to buy the costly-but-decisive ones. This is the predicted gap between greedy and optimal (§6); we report it rather than filter it out. In wall-clock seconds the picture is the same shape (overall median 38.3%), confirming the effect tracks real cost, not a token-metric artifact.

5.3 Where cost-awareness starts to pay (simulation sweep)

To map the dependence on heterogeneity with a controlled knob, we simulate a benchmark with a fixed effect and dial the per-item cost CV from 0 (uniform) to 3 (heavy-tailed), drawing cost *independently* of informativeness (a conservative setting in which the only lever is cheap-versus-expensive ordering). The median deployable saving (lowest-cost-first vs random) rises from a near-zero 2.3% at uniform cost (Monte-Carlo permutation residual; the exact reduction is certified by the reduction test above), through 44.0% at CV 0.5 and 62.2% at CV 1.0, to 84.3% at CV 2.0 (Figure 3); the median saving first exceeds 20% at cost CV ≈ 0.25. Real benchmarks with paired cost CV above this threshold (all MMLU pairs here) sit in the regime where the dollar axis materially changes the bill.

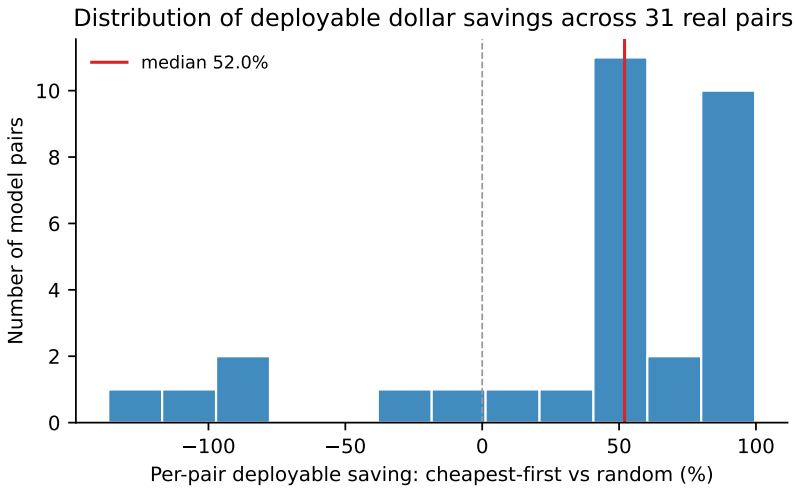

Figure 1: Distribution of per-pair DEPLOYABLE savings (predictable lowest-cost-first versus cost-blind random) across the 31 real model pairs. The mass spreads from a cluster of low-heterogeneity pairs with little saving to heavy-tailed pairs near 100%, with a negative tail of 6 pairs where the greedy rule defers the decisive expensive items.

6 Limitations

The theorem covers predictable, null-admissible policies only. Theorem 1 is *not* a blanket "any adaptive sampling is safe." It requires the selection to be predictable (so the label-peeking oracle is excluded — we use it only as an envelope) and null-admissible ((4); Remark 1). On a single exchangeable pool the latter is automatic, but a deployment that mines strata with heterogeneous conditional means could violate it; such deployments must restrict the policy to strata that cannot beat the null, or use a conservative pooled bet. We verify the guarantee on the pooled null targeted here, not on adversarial stratified nulls.

Optimality is not claimed, and the greedy rule can lose. The guarantee does not claim the lowest-cost-first rule is cost-optimal. The real-log experiment makes this concrete: on $6/31$ pairs the deployable rule is *more* expensive than random, because it defers costly items that carry the discriminating signal (or front-loads a tiny model's uninformative one-token items). A globally optimal cost-aware schedule, its regret relative to greedy, and a look-ahead rule that hedges against "expensive-but-decisive" items are open. The validity guarantee holds regardless; only the bill is at stake.

Reanalysis, not a live adaptive run. The headline is a counterfactual cost on a fixed, finite real population: we replay the deployable allocation policy over logged per-item outcomes, which faithfully answers "what would the predictable lowest-cost-first policy have cost on these items" but is not a new online experiment. We mitigate this with the online predictable-policy validity simulation, but a fully online deployment on a live model pair is future work.

Cost observability. The deployable policy routes by *known* per-item cost. Output-token cost is realized only after generation, so a true online deployment must use a predictable cost proxy (prompt length, model identity, a cheap cost predictor, or a per-item generation cap) in place of the realized token count; our reanalysis uses the realized cost as that known quantity, which is exact for a fixed-population counterfactual but optimistic to the extent a live proxy mis-predicts cost. The anticipating oracle envelope ($|d|$/cost) additionally peeks at the *outcome* $d(i)$ and is reported only as an upper bound, never as a deployable result.

Cost model. We bill output tokens (and wall-time as a check). Real API pricing weights prompt and completion tokens differently and adds fixed per-request overhead; incorporating a general linear

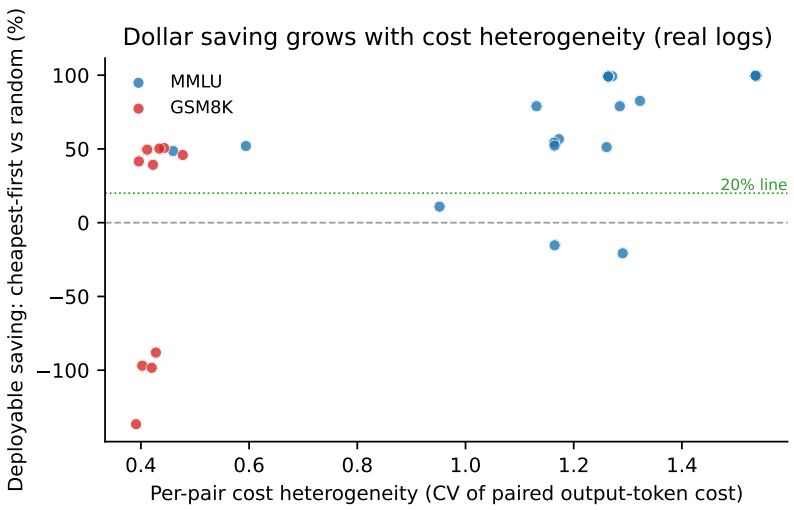

Figure 2: Per-pair DEPLOYABLE saving (lowest-cost-first vs random) versus the pair's cost heterogeneity (CV of paired output-token cost). The dollar advantage rises with heterogeneity (Spearman $\rho = 0.80$); lower-dispersion GSM8K pairs cluster at low CV and smaller saving; the negative outliers are pairs where the greedy order defers the items that carry the signal.

cost is straightforward (it only rescales cost(i)) but is not evaluated here.

Two models at a time. We treat paired superiority; extending the cost-aware race to a full leaderboard with multiplicity control (e.g. e-value aggregation) is a natural next step and inherits the validity theorem per pair.

7 Conclusion

Anytime-valid model comparison made peeking safe; it still prices decisions in items, as if every benchmark question cost the same. On real logs they do not — output-token cost varies hundredfold across items within a single model. CARA re-meters the procedure in dollars and routes each evaluation by a *predictable* cost-aware rule, with a theorem guaranteeing that the anytime Type-I control is untouched for any predictable, null-admissible routing, and a measurement showing that its deployable instance (lowest known cost first) reaches the identical verdict at a median 52.0% lower token cost across 31 real pairs versus cost-blind sampling. The saving is conditional on cost heterogeneity — large where items vary in cost (79.0% on MMLU), small where they do not (40.4% on GSM8K) — and the greedy rule can over-spend when the cheapest items are not the most informative; a label-omniscient oracle envelope (84.5%) bounds the headroom a perfect predictor could add. The narrowed guarantee, the dollar axis, and the deployable measurement are the durable contributions; designing the optimal predictable cost-aware router is the natural next problem. Budgeted model comparison should be paid for, and reported, by the dollar.

Reproducibility. All numbers are produced by `scripts/analyze.py` and `cara/test_validity.py`, written to `results/*.json`, and injected into the text as macros (`paper/generated/macros.tex`); no result number is hand-typed. The core (`cara/core.py`) and tests are Python standard-library only (no numpy/scipy); seeds are printed and re-running reproduces the JSON identically. Code, logs, and scripts are available from the author.

References

[1] Ruicheng Ao, Hongyu Chen, Siyang Gao, Hanwei Li, and David Simchi-Levi. Best arm identification with LLM judges and limited human audits. *arXiv preprint arXiv:2601.21471*, 2026.

[2] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John

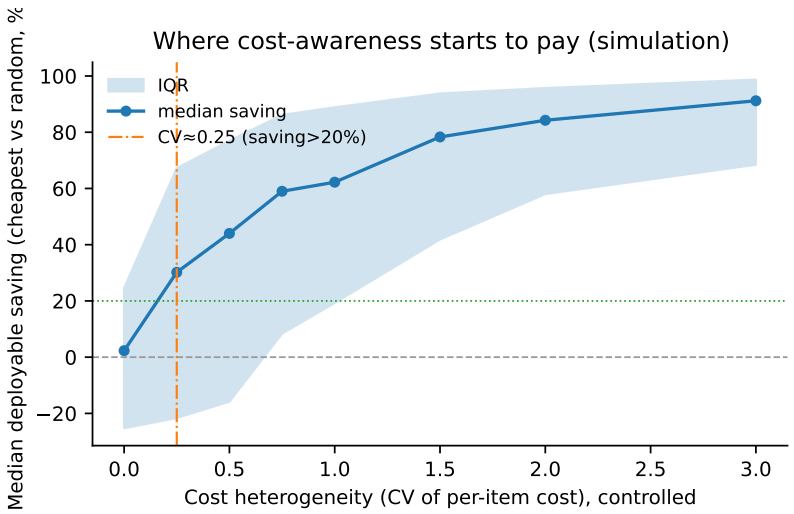

Figure 3: Controlled simulation: median deployable dollar saving (IQR band) of the predictable lowest-cost-first policy over cost-blind random as a function of per-item cost heterogeneity. Dotted: 20%; dash-dot: the heterogeneity threshold where median saving exceeds 20%.

Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

[4] Michael O. Harding and Kirthevasan Kandasamy. Balancing performance and costs in best arm identification. *arXiv preprint arXiv:2505.20583*, 2025.

[5] 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.

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

[7] Chia-Yu Hsu and Shubhanshu Shekhar. Active nonparametric two-sample testing by betting on heterogeneous data sources. *arXiv preprint arXiv:2512.22403*, 2025.

[8] Kellen Kanarios, Qining Zhang, and Lei Ying. Cost aware best arm identification. *arXiv preprint arXiv:2402.16710*, 2024.

[9] Junpei Komiyama, Kyoungseok Jang, and Junya Honda. Rate-optimal design for anytime best arm identification. *arXiv preprint arXiv:2510.23199*, 2025.

[10] 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.

[11] 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.

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

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

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

[15] Ricardo J. Sandoval, Ian Waudby-Smith, and Michael I. Jordan. Multi-armed sequential hypothesis testing by betting. *arXiv preprint arXiv:2603.17925*, 2026.

[16] George Vershinin, Asaf Cohen, and Omer Gurewitz. On cost-aware sequential hypothesis testing with random costs and action cancellation. *arXiv preprint arXiv:2512.19067*, 2025.

[17] George Vershinin, Asaf Cohen, and Omer Gurewitz. Active sequential hypothesis testing with non-homogeneous costs. *arXiv preprint arXiv:2509.11632*, 2025.

[18] Jean Ville. étude critique de la notion de collectif. *Monographies des Probabilités*, 1939.

[19] 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.