ANYTIME-VALID FALSE DISCOVERY RATE CONTROL FOR ADAPTIVE FOUNDATION MODEL EVALUATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Foundation model leaderboards are adaptive experiments: evaluators repeatedly choose model comparisons, allocate samples to promising claims, and stop when results look publishable, so standard fixed-time p-value procedures do not generally retain valid false-discovery guarantees after data-dependent stopping. We formulate leaderboard evaluation as sequential FDR control over directed pairwise model claims, and show that fixed-time BH fails under adaptive stopping while pairwise e-processes combined with e-BH control FDR at arbitrary stopping times. The main practical obstacle is power: greedy allocation concentrates nearly all budget on one comparison, leaving other true alternatives undiscovered. We make this quantitative—greedy allocation exhibits a $1/K$ power ceiling up to an explicit finite-sample correction (K true alternatives), allocation-aware sampling recovers power, and we prove a $\Theta(m/s)$ separation in discovery budget between top-k and uniform sampling that holds for any FDR-controlling rule (not only e-BH). We propose allocation-aware anytime-valid protocols (top-k round-robin, threshold reallocation) and validate them on synthetic adaptive leaderboards and adaptive replays of Chatbot Arena and MT-Bench, showing inflated fixed-time error rates, stable e-BH control, and substantial power recovery.

1 INTRODUCTION

Leaderboards turn benchmark measurements into claims about model superiority. Pairwise human-preference platforms such as Chatbot Arena make this workflow especially explicit: models are repeatedly compared, aggregate evidence is updated, and a public ranking changes over time Chiang et al. (2024). These claims are rarely produced by a fixed, prespecified experiment. Model developers inspect intermediate results, rerun close comparisons, try additional tasks, and report claims after evidence looks strong enough. This workflow creates a sequential and adaptive multiple-testing problem.

The paper's central message is:

> *Validity requires anytime inference; usefulness also requires allocation-aware sampling.*

Figure 1 summarizes the resulting two-layer framework.

We make the following contributions:

1. We formalize adaptive foundation-model evaluation as sequential FDR control over directed pairwise claims, show fixed-time BH fails under adaptive stopping (Proposition 1), and establish the *validity layer*: pairwise e-processes with e-BH control FDR at any stopping time under arbitrary adaptive allocation (Proposition 2; known machinery Wang & Ramdas (2022)).
2. We prove that greedy allocation has a $1/K$ starvation ceiling up to an explicit finite-sample correction, design allocation-aware protocols (top-k round-robin, threshold reallocation) that recover power with unconditional e-BH FDR control, and prove a $\Theta(m/s)$ sample-complexity separation between top-k and uniform sampling that holds for *any* FDR-controlling rule under a sampling constraint (Propositions 3, 5).

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

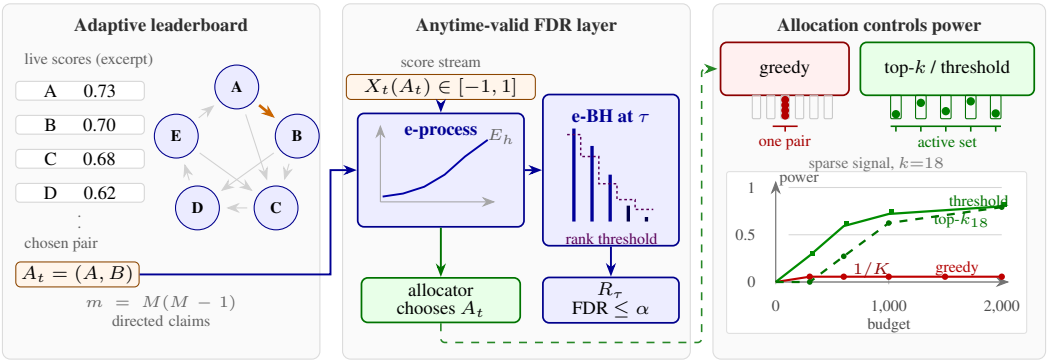

Figure 1: Anytime-valid FDR for adaptive leaderboard evaluation. A live leaderboard induces many directed pairwise claims; selected comparisons update per-pair e-processes, and e-BH at any stopping time τ reports R_τ with FDR $\leq \alpha$ under any predictable allocation. The allocator uses current evidence to choose the next pair A_t: greedy concentrates budget and starves at power $\leq 1/K$, while top-k and threshold reallocation spread effort and recover power. A formal $\Theta(m/s)$ budget separation holds between top-k and uniform sampling; uniform is not shown in the inset.

3. We delimit the information geometry (γ^{-2} weak-signal ceiling; optional $O(M \log M)$ transitivity-based certification) and validate the framework on synthetic leaderboards and adaptive replays of Chatbot Arena and MT-Bench.

2 Adaptive Leaderboard Model

Let there be M models. For each ordered pair $h = (i, j)$, define the one-sided claim

$$H_{ij}: \text{ model } i \text{ is not better than model } j.$$

Rejecting H_{ij} corresponds to announcing the leaderboard claim $i > j$. Let $\mathcal{H}$ be the set of all directed pairwise hypotheses, with $m = |\mathcal{H}| = M(M-1)$.

At each evaluation round t, the evaluator chooses a pair

$$A_t \in \mathcal{H}$$

using all information in the past filtration $\mathcal{F}_{t-1}$. The evaluator then observes a bounded score difference

$$X_t(A_t) \in [-1, 1].$$

For a null pair h, the null condition is

$$\mathbb{E}[X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] \leq 0. \tag{1}$$

This covers one-sided model-comparison claims: under the null, the selected score difference has nonpositive conditional mean. The allocation rule may be arbitrary. It can choose the most promising pair, spread budget across an active set, remove discovered pairs, or depend on any previous p-values, e-values, or scores.

At a stopping time τ, the evaluator reports a rejection set $R_\tau \subseteq \mathcal{H}$. The target error criterion is

$$\mathrm{FDR}(\tau) = \mathbb{E}\left[\frac{|R_\tau \cap \mathcal{H}_0|}{\max(|R_\tau|, 1)}\right] \leq \alpha. \tag{2}$$

This adapts the classical FDR criterion Benjamini & Hochberg (1995) to reported pairwise claims. The methods control statistical errors from sampling, adaptivity, optional stopping, and multiplicity; they do not certify benchmark quality (contamination, bias, semantic alignment), which is studied separately Xu et al. (2024); Jain et al. (2025); White et al. (2025).

3 Fixed-Time Inference Is Not Anytime-Valid

Standard p-values and fixed-time multiple-testing corrections, including BH Benjamini & Hochberg (1995), are designed for a nonadaptive analysis time. Leaderboard practice violates this premise:

comparisons are revisited, promising differences receive more evaluation budget, and results are often reported after an apparently publishable discovery appears.

The first claim is intentionally conservative.

Proposition 1 (Fixed-time BH is not anytime-valid)**.** *There exists a global-null adaptive evaluation process and a stopping time τ such that applying a fixed-time BH rule at τ has error probability above the nominal level α. A proof sketch and toy construction are given in Appendix A.*

A fixed-time guarantee controls rejection probability at a prespecified analysis time, not after repeatedly checking and stopping when a rejection first appears. A toy construction (Appendix A) repeats valid fixed-time analyses under a global null and stops at the first rejecting look; in a clean independent-look version the probability of at least one rejection over T looks is $1 - (1 - \alpha)^T$, far above α.

The simulations use a more leaderboard-like stress test: the evaluator repeatedly samples the pair with the smallest current fixed-time p-value and stops when a fixed-time rule reports a discovery. Under the global null (where FDR=FWER), fixed-time BH reaches FDR/FWER around 0.3–0.4 at the data-dependent stopping time, and unadjusted p-values can approach certain false discovery under min-p stopping. The structural point is that min-p sampling plus stop-on-discovery turns repeated fixed-time analyses into an error-amplification mechanism, motivating evidence processes that remain valid at stopping times Howard et al. (2021); Grünwald et al. (2024).

4 Anytime-Valid FDR via E-Processes and E-BH

For each hypothesis h, let $E_h(t)$ be a nonnegative e-process adapted to the global filtration Grünwald et al. (2024). For a null $h \in \mathcal{H}_0$, assume

$$\mathbb{E}[E_h(\tau)] \;\le\; 1 \tag{3}$$

for every stopping time τ. Thus $E_h(\tau)$ is a valid e-value no matter when the evaluator stops.

In the bounded-difference model, a simple one-sided betting factor is

$$1 + \lambda X_t(h), \qquad \lambda \in [0, 1].$$

When h is null and selected at time t,

$$\mathbb{E}[1 + \lambda X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] \;\le\; 1.$$

Multiplying these factors over the times at which h is sampled yields a test supermartingale/e-process for h. A finite convex mixture over several lambdas is also an e-process. The simulations use an equal-weight mixture over

$$\{0.1, 0.2, 0.4, 0.6, 0.8\}.$$

At time τ, e-BH sorts the e-values in descending order and rejects the top k hypotheses, where

$$k \;=\; \max\{r :\; E_{(r)}(\tau) \ge m/(\alpha r)\}. \tag{4}$$

This is the e-BH procedure of Wang and Ramdas Wang & Ramdas (2022); Ramdas & Wang (2025). The rank-dependent threshold is less conservative than testing each hypothesis at the e-Bonferroni threshold m/α: larger rejection sets are allowed to use the smaller threshold $m/(\alpha r)$.

The validity statement is the formal backbone of the paper.

Proposition 2 (e-BH controls FDR at adaptive stopping times)**.** *If every null hypothesis has a valid e-process with respect to the global adaptive filtration, then e-BH applied at any stopping time controls FDR at level α. The proof is given in Appendix B.*

The proof is short because optional stopping converts each null $E_h(\tau)$ into a valid e-value. The e-BH FDR guarantee then applies to the stopped collection of e-values Wang & Ramdas (2022). Dependence among pairs and adaptive allocation do not invalidate the argument, provided each null e-process is valid under the allocation rule.

This result is not the paper's main theoretical novelty. It is the validity layer that makes adaptive leaderboard inference possible. Empirically, e-BH and e-Bonferroni remain at 0.000 FDR/FWER

in the displayed global-null overnight grid, and e-mix-BH remains stable across the subsequent global-null validity checks. The main remaining challenge is power.

We measure power as the expected true-positive rate over the non-null pairwise claims,

$$\text{power} \;=\; \mathbb{E}\left[\frac{|R_\tau \cap \mathcal{H}_1|}{|\mathcal{H}_1|}\right]. \tag{5}$$

This definition separates validity from usefulness. E-BH can control FDR under arbitrary adaptive allocation, but it cannot discover alternatives that receive too little evaluation budget.

5 Allocation-Aware Power

The simplest adaptive evaluator samples the currently most promising pair. This is a useful p-hacking stress test, but it is a poor discovery protocol. It can spend nearly all future budget on a single already-promising comparison, starving other true alternatives.

The finite-budget starvation statement captures this mechanism.

Proposition 3 (Greedy allocation power upper bound)**.** *Under the adaptive leaderboard model with e-BH at level α, burn-in b samples per pair satisfying $(1+\lambda_{\max})^b < 1/\alpha$, and greedy allocation (sample the pair with the largest current e-value), the expected power over K true alternatives satisfies*

$$\mathbb{E}[\text{power}_{\text{greedy}}] \;\le\; 1/K + \delta_{\text{lock}}(T, b, \gamma, \lambda) + \alpha/(mK), \tag{6}$$

where δ_{lock} is an explicit Hoeffding–Azuma bound (on the log-e gap) that vanishes as the burn-in b and drift γ grow. The bound is structural *rather than non-asymptotically tight: greedy discovers at most one true alternative with high probability, regardless of whether the post-burn-in leader is a true alternative or a null. The full proof is in Appendix D.*

This is a structural upper bound that identifies $1/K$ as the greedy power ceiling. At the experimental burn-in $b = 2$, the Hoeffding–Azuma term δ_{lock} is numerically loose, so we do not claim it is already negligible; we verify the ceiling empirically. In the baseline sparse-signal setting, greedy e-mix-BH discovers roughly one true pair out of 18, so power remains around $0.056 = 1/18$ from budget 300 through budget 2000, and the max-pair sample fraction grows to about 0.94, meaning almost all incremental evidence is spent on one pair—consistent with Proposition 3. Section 6.3 confirms this scaling directly: greedy power tracks $1/K$ across $K \in \{7, 13, 18, 28\}$.

The lesson is that anytime-valid inference is necessary but not sufficient. A useful protocol must also allocate evidence across the set of plausible discoveries.

5.1 Top-k Round-Robin Allocation

The first allocation-aware protocol selects an active set of k promising comparisons and samples them in round-robin order. We report two variants: (i) `top_k_round_robin`, with practical dynamic re-ranking at every step; and (ii) `top_k_stable_round_robin`, with a fixed refresh interval ρ, where the active set is recomputed every ρ rounds and then round-robined until the next refresh. We report the stable variant as the main theorem-backed protocol: it gives a clean allocation statement while keeping the practical behavior of top-k.

Top-k is theoretically clean because it gives direct lower bounds on each active pair's sample count.

Proposition 4 (Top-k linear e-growth)**.** *Under round-robin allocation over an active set S, each active true alternative with drift $\gamma_h > 0$ has $\log E_h(n) \ge \gamma_h \lfloor n/|S| \rfloor - O(\sqrt{n \log(1/\delta)})$ with probability $\ge 1-\delta$; the periodic-reload variant (`top_k_stable_round_robin`) preserves this up to reload effects. Proofs in Appendix C.*

Active true alternatives thus reach the e-BH discovery scale once their linear log-e growth dominates the rank threshold $m/(\alpha r)$ (or the conservative m/α for named alternatives). Empirically, top-k_{18} reaches 0.795 power at budget 2000 and outperforms greedy at all budgets; $k \in \{10, 18, 25\}$ all substantially outperform greedy.

5.2 Sample Complexity of Allocation Policies

The linear-growth propositions show that each active pair's evidence grows, but they do not characterize how much budget is needed for guaranteed discovery. The following sample complexity result bridges this gap and formalizes the paper's central message that allocation matters as much as evidence strength.

Proposition 5 (Allocation-dependent e-BH sample complexity)**.** *For $m = M(M-1)$ directed hypotheses, s true alternatives (drift γ), define the r-discovery budget T_r (minimum post-burn-in samples for e-BH to discover $\geq r$ truths with probability $\geq 1-\delta$). Then $T_s^{\text{uniform}} = \Omega(m\,\gamma^{-2}\log(m/(\alpha\delta)))$ scales with m, while $T_s^{\textit{top-}k} = \mathcal{O}(k\,\gamma^{-2}\,\text{polylog})$ (for $k \geq s$) scales with k, so $T_s^{\text{uniform}}/T_s^{\textit{top-}k} = \Theta(m/k)$ when $s = o(m)$ and $k = \Theta(s)$. Greedy achieves at most one discovery at any finite budget. Appendix F gives the concentration upper bounds; Appendix G a matching change-of-measure lower bound for any FDR-controlling rule (under $\alpha s < 1$), so the separation is genuine up to logarithmic factors.*

The practical implication: the budget should scale with the number of plausible claims (k), not the total comparisons (m); setting k larger than the expected number of true alternatives costs only a proportional k/s slowdown.

The lower-bound side is the paper's strongest independent result, so we state it at the main-text level.

Lemma 1 (Universal per-hypothesis discovery floor, any FDR rule)**.** *For* any *sequential rule controlling FDR at level α (not only e-BH; no e-process) that discovers alternative h with probability $\geq 1-\beta$, $\mathbb{E}_\theta[N_h(\tau)] \geq \Omega(\gamma^{-2}\log(1/(\alpha s)))$ in the sparse regime $\alpha s < 1-\beta$. The proof (Appendix G) uses only FDR control plus the Kaufmann change-of-measure, no e-values. Thus no FDR-valid procedure—e-values, p-values, or otherwise—beats top-k over uniform by more than logarithmic factors under the sampling constraint; allocation-aware sampling is a necessity forced by the FDR definition, not a convenience of e-BH (scope: uniform/top-k sampling constraint; not a minimax converse over arbitrary allocations).*

5.3 Threshold Reallocation

Top-k allocation asks the user to choose k. A more adaptive protocol is threshold reallocation: (i) maintain e-mix-BH discoveries over all pairs; (ii) once a pair is discovered, remove it from the active sampling pool; (iii) reallocate future samples to unresolved comparisons. This protocol matches practical leaderboard intuition: once the evidence for a comparison is strong enough to report, do not spend more budget on it.

We distinguish two variants in practice: (i) *open-loop* threshold reallocation (high-power default), which removes pairs using the full `e_mix_bh` rejection set; and (ii) *guarded* threshold reallocation, which removes only pairs crossing the e-Bonferroni bar m/α. Both variants carry an end-to-end guarantee (Appendix E): because reallocation is an *allocation* rule rather than a rejection rule, the reported e-BH set controls FDR at level α unconditionally (Proposition E.2), and any still-active true alternative that remains in the candidate pool receives a linear sample share and is discovered once its log-e drift dominates the rank threshold. The variants differ only in an auxiliary property: the guarded rule additionally guarantees, on a $1-\alpha$ event, that every retired pair is truly non-null (a purity guarantee on the removed set), while open-loop retires on the lower e-BH bar, freeing budget sooner—which is why it is the empirically stronger default at low budget.

Proposition 6 (Guarded threshold reallocation)**.** *Under guarded threshold reallocation (e-Bonferroni removals), with probability $\geq 1-\alpha$ no null is removed; any active true alternative then reaches $E_h \geq m/(\alpha r)$ after $O(K_{\max}\gamma^{-2}\log(m/(\alpha r\delta)))$ post-removal samples, so e-BH discovers at least r truths with probability $\geq 1-\alpha-\delta$ (Appendix E). Reallocation changes only future allocation, not the rejection rule, so the reported set remains e-BH and controls FDR unconditionally (Proposition E.2); threshold is a recovery protocol, while the formal $\Theta(m/s)$ separation (Proposition 5) is top-k vs uniform.*

Experiments support this. Open-loop threshold reallocation reaches e-mix-BH power ≈ 0.80 at budget 2000 and is much stronger than top-k at low budget (0.507 vs 0.097 at budget 500); at low sparse-evidence budgets it *beats a truth-aware even-spreading oracle* (Appendix Table 8).

5.4 Dense Weak-Signal Limitation

Dense weak signals expose a different limitation. In a setting where many models have small advantages, oracle allocation itself reaches only about 0.520 power at budget 2000 (300 reps). Top-$k18$ and threshold reallocation reach 0.486 and 0.502, respectively, close to the oracle ceiling. This indicates that allocation is no longer the main bottleneck.

Proposition 7 (Small effects impose a γ^{-2} discovery scale)**.** *Any FDR-valid rule rejecting a γ-advantage alternative with probability $\geq 1-\beta$ must allocate $\Omega(\gamma^{-2}\,\mathrm{kl}(1-\beta, \alpha s))$ samples to it (Appendix G). Dense weak signals are information-limited, not allocation-limited.*

In the `dense_weak` setting with $K = 28$ true alternatives, the oracle e-mix-BH power is only 0.520 at budget 2000, while top-k_{18} and threshold reallocation reach 0.486 and 0.502, respectively (Appendix Table 8). The separate one-leader grid in Appendix Table 5 shows the same information-coordinate transition: at budget 2000, oracle e-mix-BH power rises from 0.140 at $\gamma = 0.06$ to 0.539 at $\gamma = 0.08$, while uniform allocation remains at 0.000. Thus allocation recovery applies when starvation is the bottleneck, not when weak-signal information is.

5.5 Structured Certification Under Transitivity

The allocation results above treat the $m = M(M-1)$ directed claims as an unstructured family. Real leaderboards carry strong structure: when the true preferences form a consistent total order $\pi^\star$, the claims are not independent—$i \succ j$ and $j \succ k$ imply $i \succ k$. We now state what can and cannot be gained from this transitivity. Two facts discipline the discussion. First, *naive transitive closure of an e-BH rejection set over the full m pairs does not control closure FDR*: under a single false directional rejection (claiming $i \succ j$ when in fact $j \succ i$), the closure can carry up to $M-1$ false implied claims, and conflicting edges can even create cycles, so the closed set need not be a partial order. Closure-level error control therefore requires controlling the *comparators* directly, not closing an unstructured FDR set. Second, the cost of that control is an FWER-type (no-false-comparator) requirement, which is the price of a clean closure guarantee.

Proposition 8 (Transitive-order certification)**.** *Under an unknown strict total order on M models, a predictable comparison scheme using $N_c = O(M \log M)$ nodes—each backed by two directional e-processes with split error budget $\sum_{s,d} \alpha_{s,d} \leq \alpha$—achieves anytime no-false-comparator control ($\Pr\{\exists s : d_s \text{ false}\} \leq \alpha$, Ville + union bound) and closure-level FDR $\leq \alpha$ at total budget $\widetilde{O}(M \log M \cdot \gamma^{-2})$. The $O(M \log M)$ comparison count is optimal (matching the $\Omega(M \log M)$ sorting lower bound). This is an optional structured regime; the unstructured e-BH family (Proposition 2) controls FDR without any transitivity assumption, and is the primary recommendation.*

This is a certification statement (FWER-level comparator control), not a multiplicity reduction of the unstructured e-BH family; the $O(M \log M)$ comparison count is optimal (matching the $\Omega(M \log M)$ sorting lower bound). It governs the structured certification task, while Proposition 5's $\Theta(m/s)$ separation governs the flat m-family; using transitivity as a sampling constraint within the unstructured e-BH family is FDR-safe but does not improve power (Section 6.4).

Remark (Scope: exact vs. approximate transitivity). Proposition 8 is scoped to exact transitivity; under a δ-transitive side condition the closure FDP degrades to $\leq \alpha + \delta$, and in strongly cyclic regimes the guarantee is vacuous. The unstructured e-BH family controls FDR without transitivity, and is the primary recommendation.

6 Experiments

The experiments are synthetic adaptive leaderboards designed to isolate the statistical issues in Sections 3–5. Unless otherwise noted, there are $M = 8$ models and $m = 56$ directed pairwise hypotheses. The default sparse-signal setting has ability vector

$$(0.45, 0.30, 0.18, 0, 0, 0, 0, 0),$$

which yields $s = 18$ true directed alternatives. The score differences are bounded in $[-1, 1]$, with noise level $\sigma = 0.35$, burn-in of two samples per directed pair, and target level $\alpha = 0.05$. The fixed-λ e-process uses $\lambda = 0.4$; the mixture e-process averages over $\lambda \in \{0.1, 0.2, 0.4, 0.6, 0.8\}$. Full parameter details (signal settings, sampling policies) are in Appendix Table 4.

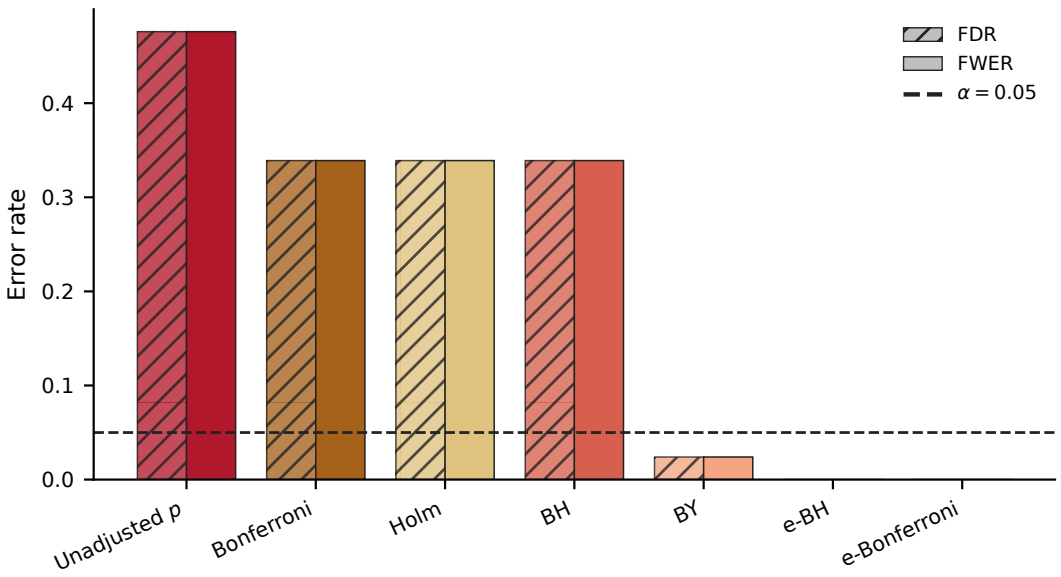

Figure 2: Global-null FDR/FWER under adaptive stopping: fixed-time p/BH inflate above α; e-BH/e-Bonferroni stay at zero.

6.1 Fixed-Time Failure under Global Null

The overnight grid tests the failure mechanism under the global null, where every rejection is false and therefore FDR equals FWER. The evaluator repeatedly samples promising pairs and stops under fixed-time rules. In these runs, unadjusted p-values can approach certain false discovery under min-p stopping, and fixed-time BH reaches FDR/FWER around 0.3–0.4 under BH-style stopping at budget 1000. By contrast, e-BH and e-Bonferroni remain at 0.000 FDR/FWER in the displayed global-null grid. Figure 2 shows the separation.

6.2 Starvation and Allocation Recovery

A finite mixture over betting fractions preserves validity (convex mixtures of e-processes are e-processes) and helps the smallest budgets (e-mix-BH power 0.044 vs fixed-λ 0.007 at budget 150), but both plateau at $0.056 \approx 1/18$ after budget 300: tuning λ is not the bottleneck, allocation is (Appendix Figure 4).

The allocation grid isolates this. Greedy e-mix-BH stays at power 0.056 through budget 2000 while concentrating $\sim 94\%$ of incremental evidence on a single pair (Appendix Figure 7), matching Proposition 3; uniform collapses. Allocation-aware samplers break the plateau—at budget 2000 oracle round-robin reaches 0.961, top-k_{18} 0.795, threshold reallocation 0.800—and threshold reallocation *beats a truth-aware even-spreading oracle at lower sparse-evidence budgets*, since concentration dominates spreading (Appendix Table 8). Figure 3 shows the recovery; global-null validity (e-BH FDR/FWER $\equiv 0$) is unaffected by the allocator.

6.3 Scaling and Sensitivity

A k/s sweep validates Proposition 5: under-allocation ($k < s$) caps power near k/s, over-allocation ($k > s$) slows growth by the predicted k/s factor, and the phase transition sits at $k \approx s$; e-BH/e-mix-BH FDR and FWER stay at or near zero across all (k, s, budget), so validity is unaffected by k. Two further sweeps pin the scaling laws. Varying the number of true alternatives $K \in \{7, 13, 18, 28\}$ at fixed $M = 8$ (Table 1), greedy e-mix-BH power tracks $1/K$ digit-for-digit $(0.143, 0.077, 0.056, 0.036)$, while threshold reallocation holds 0.81–0.97 across K and top-k_{18} drops only at $K = 28$ where $k < s$ under-covers—exactly the k/s budget logic. Varying the family size $M \in \{8, 16, 24, 32, 50\}$ (Table 2, up to m=2450 claims, budget scaled $\propto m/s$ to hold per-alternative information fixed), the greedy $1/K$ plateau persists, uniform collapses as m/s grows

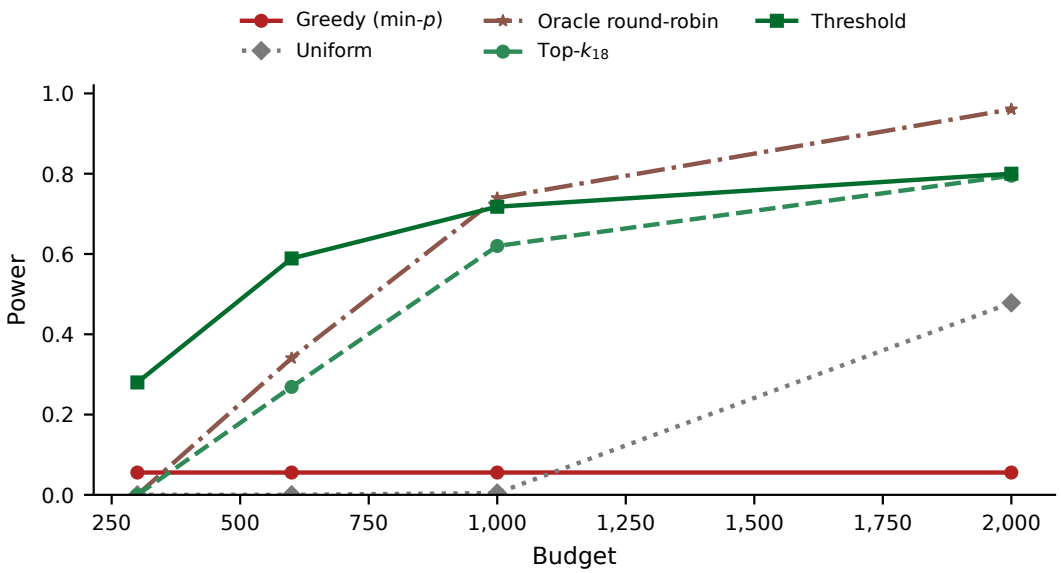

Figure 3: e-mix-BH power by sampler: allocation-aware (oracle, top-k, threshold) recover; greedy/uniform stay starved.

Table 1: Greedy e-mix-BH power tracks $1/K$ within MC error (budget 2000, 200 reps). Top-k18 and threshold stay above the ceiling; e-mix-BH FWER is 0 for greedy.

K	$1/K$	Greedy	Top-k18	Threshold	Greedy max-pair
7	0.143	0.143	0.994	0.971	0.94
13	0.077	0.077	0.937	0.892	0.94
18	0.056	0.056	0.783	0.796	0.94
28	0.036	0.036	0.530	0.815	0.94

(the empirical signature of the $\Theta(m/s)$ separation), and allocation-aware samplers track the oracle with an advantage that *widens* with M—the conclusions are ecological, not M=8 artifacts.

6.4 Real-data adaptive replays and robustness

We replay the protocol on two real pairwise-preference sources with complementary signal structure: Chatbot Arena Chiang et al. (2024) (sparse near-binary human preferences; twenty models, $m = 242$ of the 380 directed pairs with recorded battles, 109 empirical alternatives) and MT-Bench GPT-4 single-answer grading Zheng et al. (2023) (dense-weak continuous scores $X = (\text{score}_i - \text{score}_j)/9 \in [-1, 1]$; twenty models, $m = 380$, 149 alternatives). The evaluator adaptively chooses pairs and the stopping time, sampling i.i.d. with replacement from the empirical battle pool—validating the statistical mechanism under realistic signal, sparsity, and noise (not a live deployment; no model API calls).

Validity. A symmetrized-null stress test (nonzero outcomes sign-flipped, so every rejection is false, FDR = FWER) shows fixed-time BH inflates with the look horizon—FWER $0.051 \to 0.084 \to 0.104 \to 0.138$ over budgets $\{500, 1000, 2000, 5000\}$ at M=6, 2.8× the nominal α—while swapping only the stopper to e-BH holds FWER ≤ 0.048; at M=20, e-BH keeps FWER ≤ 0.017. *Validity is determined by the stopper, not the allocator*, and the failure is cumulative (Appendix Figure 8).

Power. On Arena (Table 3), greedy locks at $0.009 = 1/109$; threshold reallocation recovers $(0.14 \to 0.25 \to 0.40)$ and, at low sparse-evidence budgets, *beats a truth-aware even-spreading oracle*. On MT-Bench (Appendix Table 7), greedy locks at $0.007 = 1/149$, threshold recovers $(0.12 \to 0.24 \to 0.44)$, FDR ≤ 0.0033; excluding GPT-4 leaves conclusions intact. The $1/K$ plateau and threshold-beats-oracle recovery reproduce on both real-data regimes.

Robustness and limits. High-rep checks keep FWER ≤ 0.04 (Appendix Figs. 5, 6). In the $K = 28$ dense-weak setting, the oracle reaches only 0.520 power at budget 2000 (Appendix Table 8),

Table 2: Ecological validity across family size M (e-mix-BH power, mean $\pm$ binomial MC SE; reps $= 300$ for $M \leq 32$, reps $= 100$ for $M{=}50$, which is compute-limited). Three separated leaders + nulls at every M; the budget is scaled $\propto m/s$ (anchored so $M{=}8 \rightarrow 2000$) to hold the per-alternative information coordinate roughly constant and isolate the effect of family size $m = M(M{-}1)$. The greedy plateau tracks $1/K$ digit-for-digit across $M \in \{8, 16, 24, 32, 50\}$ (Propositions 3 and D.2), uniform allocation collapses as m/s grows (the $\Theta(m/s)$ separation, Proposition 5), and the allocation-aware samplers (threshold, top-$k{=}s$) recover power toward the oracle. e-BH FDR stays at ≈ 0 throughout (validity layer, Proposition 2).

M	m	s	m/s	budget	$1/K$	greedy	uniform	threshold	top-$k{=}s$	oracle
8	56	18	3.11	2000	0.0556	0.056 ($\pm$0.013)	0.497 ($\pm$0.029)	0.801 ($\pm$0.023)	0.780 ($\pm$0.024)	0.968 ($\pm$0.010)
16	240	42	5.71	3673	0.0238	0.024 ($\pm$0.009)	0.000 ($\pm$0.001)	0.663 ($\pm$0.027)	0.612 ($\pm$0.028)	0.904 ($\pm$0.017)
24	552	66	8.36	5377	0.0152	0.015 ($\pm$0.007)	0.000 ($\pm$0.000)	0.605 ($\pm$0.028)	0.557 ($\pm$0.029)	0.821 ($\pm$0.022)
32	992	90	11.02	7086	0.0111	0.011 ($\pm$0.006)	0.000 ($\pm$0.000)	0.538 ($\pm$0.029)	0.460 ($\pm$0.029)	0.744 ($\pm$0.025)
50	2450	144	17.01	10938	0.0069	0.007 ($\pm$0.008)	0.000 ($\pm$0.000)	0.453 ($\pm$0.050)	0.319 ($\pm$0.047)	0.527 ($\pm$0.050)

Table 3: Anytime-valid adaptive replay on Chatbot Arena preferences, extended from 6 to 20 models (e-mix-BH power, 300 reps, mean $\pm$ binomial MC SE). Real human-preference battles from `lmarena-ai/arena-human-preference-55k`; $m = 242$ directed hypotheses with 109 empirical alternatives (full-pool advantage > 0.02), median 123 battles/pair. Greedy allocation locks at the $1/109 \approx 0.009$ starvation plateau (Propositions 3 and D.2, now on real preferences) regardless of budget; uniform collapses; threshold reallocation recovers power and *beats a naive oracle that knows the truths but spreads budget evenly* (concentration $>$ even-spreading in the sparse-evidence regime). e-BH FDR ≤ 0.002 (FWER ≤ 0.043) across all configs; under the symmetrized-null stress test e-BH holds FWER ≤ 0.017 at $M{=}20$.

allocator (e-value + e-BH stopper)	budget 2500	5000	10000	$1/K$ ref
greedy (max e-mix)	0.009 ($\pm$0.005)	0.009 ($\pm$0.005)	0.009 ($\pm$0.005)	$1/109$=0.009
greedy (min-p)	0.009 ($\pm$0.005)	0.009 ($\pm$0.005)	0.009 ($\pm$0.005)	$1/109$=0.009
uniform all	0.000 ($\pm$0.000)	0.003 ($\pm$0.003)	0.063 ($\pm$0.014)	—
top-k stable	0.077 ($\pm$0.015)	0.087 ($\pm$0.016)	0.091 ($\pm$0.017)	—
oracle (knows truths, even spread)	0.002 ($\pm$0.003)	0.072 ($\pm$0.015)	0.241 ($\pm$0.025)	—
threshold reallocation	**0.139 ($\pm$0.020)**	**0.248 ($\pm$0.025)**	**0.399 ($\pm$0.028)**	—

illustrating the γ^{-2} scale of Proposition 7. Transitivity as a sampling constraint is FDR-safe but does not improve power (Appendix Table 6).

7 Related Work

Classical FDR control Benjamini & Hochberg (1995) assumes a fixed family and prespecified analysis time; online FDR methods (SAFFRON/LORD Ramdas et al. (2018); Fisher (2024), e-LOND Xu & Ramdas (2024)) let hypotheses arrive over time. Our setting differs: the pairwise-claim family is finite, but the evaluator adaptively resamples, reallocates budget, and stops data-dependently, making anytime-valid e-processes and e-BH Grünwald et al. (2024); Wang & Ramdas (2022) the natural validity layer; we use the simple factor $1 + \lambda X_t(h)$ (adaptive betting Waudby-Smith & Ramdas (2024) is a drop-in upgrade) since the bottleneck is allocation (§5.4). The closest frame is multi-arm testing by betting Sandoval et al. (2026) (single global null) and best-arm identification in bandits Garivier & Kaufmann (2016); Kaufmann et al. (2016); Chen et al. (2017) (fixed-confidence stopping); our Proposition 5 adapts the change-of-measure technique to a $\Theta(m/s)$ FDR-constrained separation, so the budget scales with plausible claims (k), not m. The motivation is pairwise-preference leaderboards Chiang et al. (2024); benchmark-contamination work Xu et al. (2024); Jain et al. (2025); White et al. (2025) is complementary.

Limitations

We target adaptive-sampling/stopping error, not contamination, semantic validity, or nonstationarity; claims scale as $M(M-1)$ and dense weak signals are information-limited. Replays use empirical snapshots with replacement; MT-Bench has shared question effects and GPT-4 is both judge and contestant, so exclude-GPT-4 supports allocation conclusions, not absolute rankings. Threshold has unconditional e-BH FDR and conditional power but no matching top-k/uniform separation; the lower bound assumes the stated sampling constraint and $\alpha s < 1 - \beta$.

REPRODUCIBILITY STATEMENT

We provide an anonymous artifact for review containing a self-contained Python implementation of the allocation policies (greedy, top-k, threshold), the anytime-valid e-processes, the e-BH stopper, and the adaptive replay harness, together with the signal configurations, sampling schedules, and random seeds used for every figure and table. The artifact is available at `https://anonymous.4open.science/r/anytime-fdr-review-3D05/`. The core code uses only the Python standard library; real-data replays use optional downloads of public static snapshot files and make no live model API calls. All experiments are simulations over publicly available data—Chatbot Arena battle snapshots Chiang et al. (2024) and the `lmsys/mt-bench` GPT-4 grades Zheng et al. (2023)—and synthetic signal models; conditional on the specified snapshot files and recorded seeds, the reported analyses are reproducible and do not depend on proprietary or deprecated endpoints. Each Monte-Carlo estimate reports its repetition count (200–300 reps, fewer only at $M \in \{32, 50\}$ where compute-bound) and a binomial standard error where relevant, so reported power and FWER/FDR values can be verified within sampling error. All formal statements are accompanied by full proofs in Appendices A–G. Compute requirements are modest and no accelerator is needed: the synthetic grids finish in minutes per configuration on a standard workstation, and the adaptive replays at $M = 20$ run in roughly 5–13 minutes per repetition batch using inner parallelism; the $M \in \{32, 50\}$ scaling tables are the only compute-bound configurations, which is why they use fewer repetitions. All experiments (synthetic grids, Arena and MT-Bench replays, scaling tables) can be reproduced on a commodity laptop in under 48 hours without a GPU.

AI USE STATEMENT

During the preparation of this work, the author used OpenAI Codex to assist with prototype code, language revision, manuscript formatting, and cross-file consistency checks related to the methodology, experiments, and presentation stages. The author specified the research questions and assumptions, independently verified the mathematical derivations and executable tests, checked all reported results against the released artifacts, edited the manuscript, and takes full responsibility for its content.

REFERENCES

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Lijie Chen, Jian Li, and Mingda Qiao. Towards instance optimal bounds for best arm identification. In *Conference on Learning Theory (COLT)*, volume 65 of *Proceedings of Machine Learning Research*, pp. 535–592, 2017. URL `https://proceedings.mlr.press/v65/chen17b.html`.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Banghua Zhu, Hao Zhang, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. Chatbot arena: An open platform for evaluating LLMs by human preference. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 8359–8388, 2024. URL `https://proceedings.mlr.press/v235/chiang24b.html`.

Aaron Fisher. Online false discovery rate control for lord++ and saffron under positive, local dependence. *Biometrical Journal*, 66(1), 2024. doi: 10.1002/bimj.202300177.

Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Conference on Learning Theory (COLT)*, volume 49 of *Proceedings of Machine Learning Research*, pp. 998–1027, 2016. URL `https://proceedings.mlr.press/v49/garivier16a.html`.

Peter Grünwald, Rianne de Heide, and Wouter M. Koolen. Safe testing. *Journal of the Royal Statistical Society: Series B*, 86(5):1091–1128, 2024. doi: 10.1093/jrsssb/qkae011.

Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform, nonparametric, nonasymptotic confidence sequences. *The Annals of Statistics*, 49(2):1055–1080, 2021. doi: 10.1214/20-AOS1991.

Naman Jain et al. Livecodebench: Holistic and contamination free evaluation of large language models for code. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=chfJJYC3iL.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best-arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17(1):1–42, 2016. URL https://jmlr.org/papers/v17/kaufman16a.html.

Aaditya Ramdas and Ruodu Wang. Hypothesis testing with e-values. *Foundations and Trends in Statistics*, 1(1-2):1–390, 2025. doi: 10.1561/3600000002.

Aaditya Ramdas, Tijana Zrnic, Martin J. Wainwright, and Michael I. Jordan. Saffron: An adaptive algorithm for online control of the false discovery rate. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4286–4294, 2018. URL https://proceedings.mlr.press/v80/ramdas18a.html.

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

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. doi: 10.1111/rssb.12489.

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. doi: 10.1093/jrsssb/qkad009.

Colin White, Samuel Dooley, et al. Livebench: A challenging, contamination-limited LLM benchmark. In *International Conference on Learning Representations*, 2025. ICLR 2025 Spotlight.

Cheng Xu, Shuhao Guan, Derek Greene, and M.-Tahar Kechadi. Benchmark data contamination of large language models: A survey. *arXiv preprint arXiv:2406.04244*, 2024.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, volume 238 of *Proceedings of Machine Learning Research*, pp. 3997–4005, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. Judging LLM-as-a-judge with MT-bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Table 4: Experimental setup. Top: core simulation defaults. Bottom: signal settings (ability vectors and resulting number of true directed alternatives s) and sampling policies.

Core defaults			
Models M / hypotheses m	8 / 56 $= M(M-1)$	Noise σ / benchmark sd	0.35 / 0.02
Target level α	0.05	Score range	$[-1, 1]$
Burn-in b	2 per directed pair	Fixed bet λ	0.4
Mixture λ set	$\{0.1, 0.2, 0.4, 0.6, 0.8\}$	Stop reference p	0.05
Signal settings (ability vector $\rightarrow$ true alternatives s)			
global_null	$(0, 0, 0, 0, 0, 0, 0, 0) \rightarrow 0$	baseline_3	$(0.45, 0.30, 0.18, 0, 0, 0, 0, 0) \rightarrow 18$
weak_3	$(0.35, 0.25, 0.15, 0, 0, 0, 0, 0) \rightarrow 18$	very_weak_3	$(0.30, 0.20, 0.10, 0, 0, 0, 0, 0) \rightarrow 18$
strong_3	$(0.55, 0.40, 0.25, 0, 0, 0, 0, 0) \rightarrow 18$	dense_weak	$(0.45, 0.30, 0.18, 0.05, 0.03, 0.02, 0.01, 0) \rightarrow 28$
Sampling policies			
greedy_min_p	smallest current fixed-time p-value	uniform_all	uniform over all directed pairs
oracle_alt_rr	round-robin over true alternatives (upper bound)	top_k_rr / top_k_stable_rr	top-k, refresh every $\rho = 50$ (stable)
epsilon_greedy	$\epsilon \in \{0.1, 0.2\}$	threshold_reallocate	remove discovered pairs, reallocate

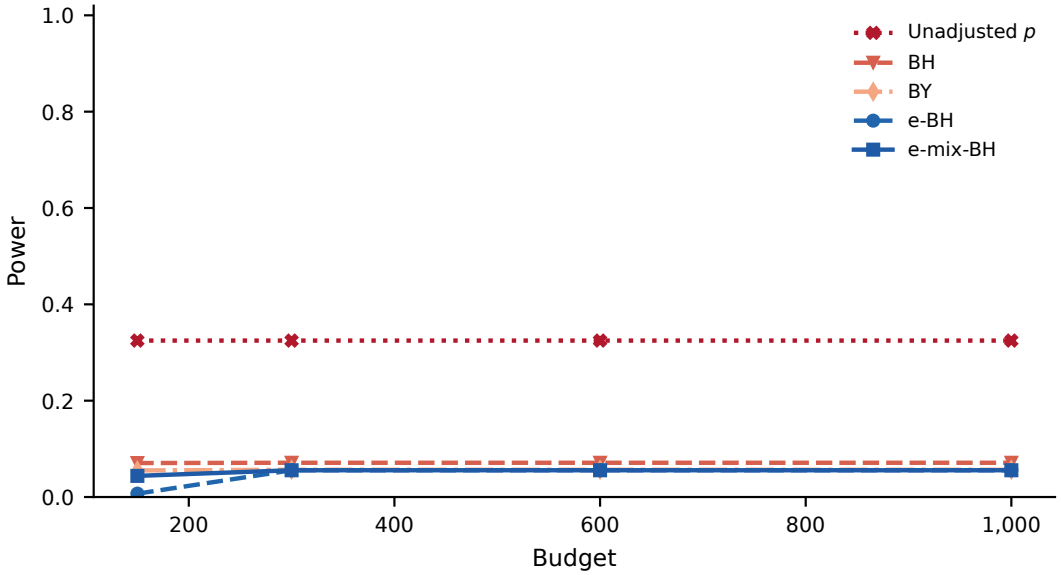

Figure 4: Sparse-signal power by budget: fixed-λ and mixture e-BH both plateau near $1/18 \approx 0.056$ after budget 300.

Deferred Experimental Figures and Tables

This appendix collects robustness and setup exhibits that support, but are not central to, the main empirical narrative.

A Fixed-Time BH under Adaptive Stopping

Proposition A.1 (Fixed-time BH is not anytime-valid)**.** *There exists a global-null adaptive evaluation process and a stopping time τ such that applying a fixed-time BH rule at τ has error probability strictly above its nominal level α.*

Toy construction

At each look $t = 1, \ldots, T$, construct a fresh batch of valid global-null p-values. Let R_t be the rejection set from a fixed-time BH procedure at look t. Choose a stopping time

$$\tau = \inf\{t \leq T : R_t \text{ is nonempty}\},$$

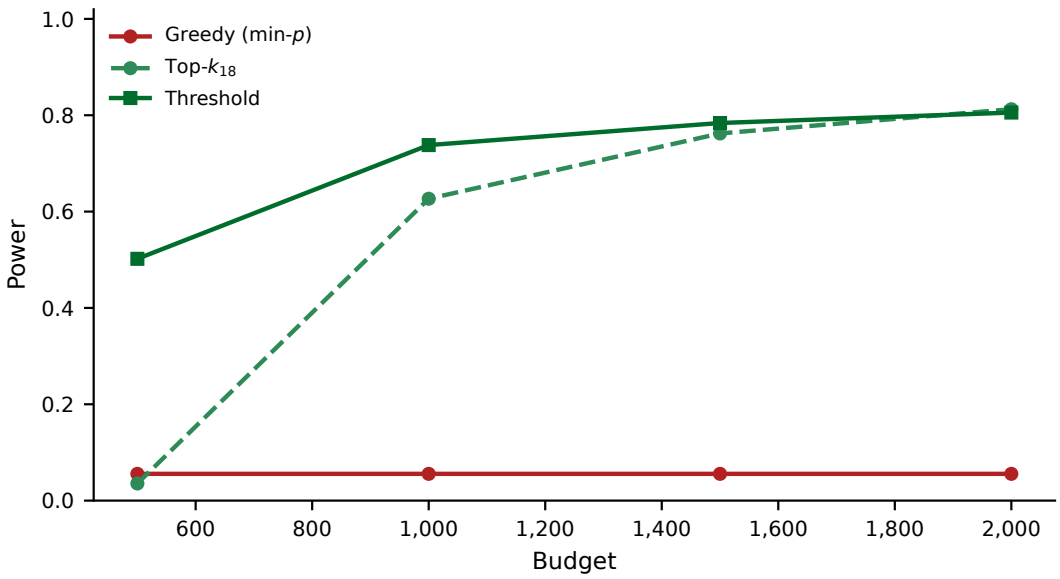

Figure 5: Baseline sparse-signal power by budget across samplers; greedy stays at the $1/18$ plateau.

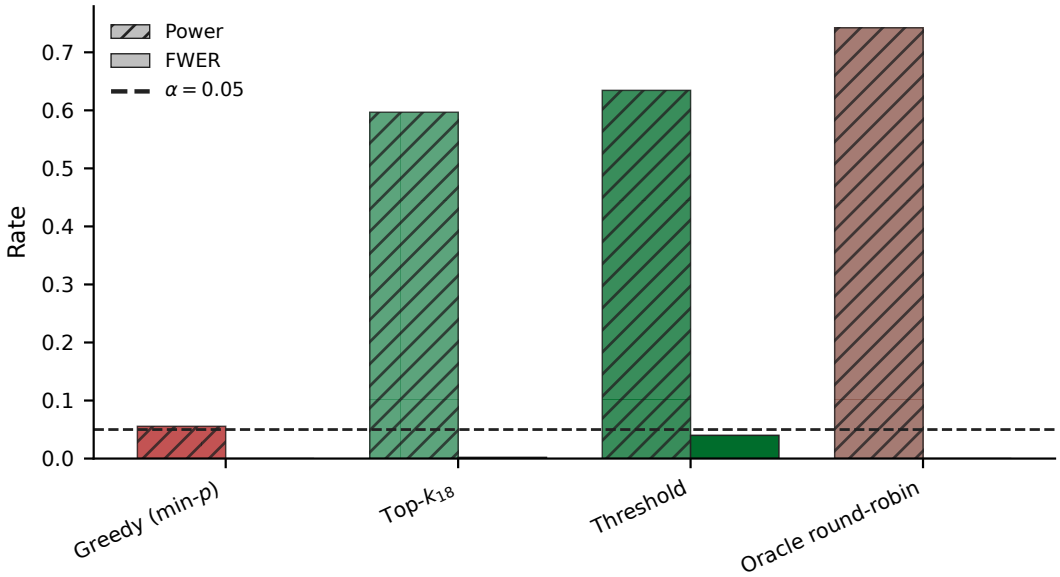

Figure 6: Very-weak targeted check (500 reps): allocation-aware samplers retain power with FWER near zero.

with $\tau = T$ if no look rejects. If the rejection events are independent and each has probability close to α, then

$$\begin{aligned}\Pr(R_\tau \text{ is nonempty}) &= \Pr(\exists\, t \le T :\ R_t \text{ nonempty}) \\ &\approx\ 1 - (1-\alpha)^T.\end{aligned}$$

This toy construction is an appendix example; the main text relies on the adaptive simulation for the leaderboard-specific failure mechanism.

B ANYTIME-VALID E-BH

For each hypothesis $h \in \mathcal{H}$, let $E_h(t)$ be adapted to the global filtration. For every null $h \in \mathcal{H}_0$, assume $\mathbb{E}[E_h(\tau)] \le 1$ for every stopping time τ. At time τ, sort e-values in descending order:

Table 5: Small-effect information ceiling: e-mix-BH power of the oracle (round-robin over true alternatives) and threshold reallocation versus the per-alternative information coordinate $B\gamma^2/K$ (one leader, $K = 7$ true alternatives, reps=300, mean $\pm$ binomial MC SE). Power transitions with $B\gamma^2/K$, not with B or γ alone, supporting the γ^{-2} discovery scale of Proposition 7; the adaptive (WSR) oracle shows the same information-coordinate transition, with finite-sample power differences from the mixture oracle.

γ	budget	$B\gamma^2/K$	oracle e-mix	oracle e-adapt	threshold	uniform
0.02	500	0.029	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.001 ($\pm$0.002)	0.000 ($\pm$0.000)
0.02	1000	0.057	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.002 ($\pm$0.003)	0.000 ($\pm$0.000)
0.02	2000	0.114	0.000 ($\pm$0.001)	0.000 ($\pm$0.000)	0.009 ($\pm$0.005)	0.000 ($\pm$0.000)
0.04	500	0.114	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.007 ($\pm$0.005)	0.000 ($\pm$0.000)
0.04	1000	0.229	0.001 ($\pm$0.002)	0.001 ($\pm$0.002)	0.025 ($\pm$0.009)	0.000 ($\pm$0.000)
0.04	2000	0.457	0.017 ($\pm$0.007)	0.006 ($\pm$0.005)	0.084 ($\pm$0.016)	0.000 ($\pm$0.000)
0.06	500	0.257	0.000 ($\pm$0.000)	0.001 ($\pm$0.002)	0.025 ($\pm$0.009)	0.000 ($\pm$0.000)
0.06	1000	0.514	0.013 ($\pm$0.007)	0.011 ($\pm$0.006)	0.093 ($\pm$0.017)	0.000 ($\pm$0.000)
0.06	2000	1.029	0.140 ($\pm$0.020)	0.065 ($\pm$0.014)	0.195 ($\pm$0.023)	0.000 ($\pm$0.000)
0.08	500	0.457	0.000 ($\pm$0.001)	0.007 ($\pm$0.005)	0.082 ($\pm$0.016)	0.000 ($\pm$0.000)
0.08	1000	0.914	0.069 ($\pm$0.015)	0.039 ($\pm$0.011)	0.189 ($\pm$0.023)	0.000 ($\pm$0.000)
0.08	2000	1.829	0.539 ($\pm$0.029)	0.324 ($\pm$0.027)	0.274 ($\pm$0.026)	0.000 ($\pm$0.000)
0.10	500	0.714	0.002 ($\pm$0.003)	0.010 ($\pm$0.006)	0.138 ($\pm$0.020)	0.000 ($\pm$0.000)
0.10	1000	1.429	0.253 ($\pm$0.025)	0.164 ($\pm$0.021)	0.255 ($\pm$0.025)	0.000 ($\pm$0.000)
0.10	2000	2.857	0.883 ($\pm$0.019)	0.723 ($\pm$0.026)	0.360 ($\pm$0.028)	0.000 ($\pm$0.000)
0.15	500	1.607	0.043 ($\pm$0.012)	0.172 ($\pm$0.022)	0.298 ($\pm$0.026)	0.000 ($\pm$0.000)
0.15	1000	3.214	0.901 ($\pm$0.017)	0.795 ($\pm$0.023)	0.438 ($\pm$0.029)	0.000 ($\pm$0.000)
0.15	2000	6.429	1.000 ($\pm$0.000)	0.986 ($\pm$0.007)	0.519 ($\pm$0.029)	0.000 ($\pm$0.001)
0.20	500	2.857	0.415 ($\pm$0.028)	0.585 ($\pm$0.028)	0.434 ($\pm$0.029)	0.000 ($\pm$0.000)
0.20	1000	5.714	0.999 ($\pm$0.002)	0.984 ($\pm$0.007)	0.580 ($\pm$0.029)	0.000 ($\pm$0.000)
0.20	2000	11.429	1.000 ($\pm$0.000)	0.983 ($\pm$0.007)	0.635 ($\pm$0.028)	0.008 ($\pm$0.005)

Table 6: Transitivity pruning as a sampling constraint (e-mix-BH power, reps=300, mean $\pm$ binomial MC SE). Pairs whose direction is implied by the transitive closure of the current e-BH rejection set are retired from the active sampling pool only; the rejection set is unchanged, so FDR control (Proposition 2) is preserved exactly (global-null FWER ≤ 0.01). Pruning does not improve power: it is neutral in the sparse one-leader setting (within ≈ 1 SE) and slightly harmful in dense settings (up to ≈ 3 SE at high budget), because greedy/threshold allocation already concentrates budget and retiring implied pairs removes useful sampling targets. This negative result motivates using transitivity for certification (Proposition 8) rather than for sampling efficiency. "Pruned frac" is the mean fraction of pairs retired by transitivity during sampling.

phase	budget	greedy	threshold	pruning	pruned frac	oracle	prun.$-$thr.
global_null	500	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000	0.000 ($\pm$0.000)	+0.000
global_null	1000	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000	0.000 ($\pm$0.000)	+0.000
global_null	2000	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000	0.000 ($\pm$0.000)	+0.000
K7_one_leader	500	0.143 ($\pm$0.020)	0.930 ($\pm$0.015)	0.924 ($\pm$0.015)	0.172	1.000 ($\pm$0.000)	-0.006
K7_one_leader	1000	0.143 ($\pm$0.020)	0.951 ($\pm$0.012)	0.952 ($\pm$0.012)	0.207	1.000 ($\pm$0.000)	+0.001
K7_one_leader	2000	0.143 ($\pm$0.020)	0.968 ($\pm$0.010)	0.959 ($\pm$0.011)	0.224	1.000 ($\pm$0.000)	-0.009
K18_three_leaders	500	0.056 ($\pm$0.013)	0.507 ($\pm$0.029)	0.502 ($\pm$0.029)	0.178	0.059 ($\pm$0.014)	-0.006
K18_three_leaders	1000	0.056 ($\pm$0.013)	0.731 ($\pm$0.026)	0.695 ($\pm$0.027)	0.324	0.732 ($\pm$0.026)	-0.036
K18_three_leaders	2000	0.056 ($\pm$0.013)	0.810 ($\pm$0.023)	0.755 ($\pm$0.025)	0.429	0.968 ($\pm$0.010)	-0.055
K28_dense_ladder	500	0.036 ($\pm$0.011)	0.305 ($\pm$0.027)	0.295 ($\pm$0.026)	0.161	0.000 ($\pm$0.000)	-0.011
K28_dense_ladder	1000	0.036 ($\pm$0.011)	0.539 ($\pm$0.029)	0.516 ($\pm$0.029)	0.334	0.284 ($\pm$0.026)	-0.023
K28_dense_ladder	2000	0.036 ($\pm$0.011)	0.823 ($\pm$0.022)	0.736 ($\pm$0.025)	0.554	0.638 ($\pm$0.028)	-0.087
K28_dense_weak	500	0.036 ($\pm$0.011)	0.296 ($\pm$0.026)	0.293 ($\pm$0.026)	0.162	0.000 ($\pm$0.000)	-0.002
K28_dense_weak	1000	0.036 ($\pm$0.011)	0.449 ($\pm$0.029)	0.412 ($\pm$0.028)	0.295	0.264 ($\pm$0.025)	-0.037
K28_dense_weak	2000	0.036 ($\pm$0.011)	0.504 ($\pm$0.029)	0.472 ($\pm$0.029)	0.403	0.522 ($\pm$0.029)	-0.032

$E_{(1)}(\tau) \geq \cdots \geq E_{(m)}(\tau)$. E-BH rejects the top k hypotheses where $k = \max\{r : E_{(r)}(\tau) \geq m/(\alpha r)\}$.

Proposition B.1 (e-BH controls FDR at stopping times)**.** *If each null hypothesis has a valid e-process with respect to the global adaptive filtration, then e-BH applied at any stopping time τ controls FDR at level α.*

Proof sketch. Optional stopping gives $\mathbb{E}[E_h(\tau)] \leq 1$ for every null h, so the stopped values are valid e-values. The e-BH guarantee applies to arbitrary collections of e-values whose null expectations are at most one Wang & Ramdas (2022). Therefore e-BH at τ controls FDR. $\square$

Table 7: Anytime-valid adaptive replay on MT-Bench GPT-4 single-answer grading, a second real data source qualitatively distinct from Chatbot Arena (e-mix-BH power, 300 reps, mean $\pm$ binomial MC SE). Public `lmsys/mt-bench` GPT-4 grades on 80 multi-turn questions; the 20 highest-mean-score models form $m = 380$ directed hypotheses with 149 empirical alternatives (full-pool score advantage > 0.02). Each directed observation is the bounded score difference $X = (\text{score}_i - \text{score}_j)/9 \in [-1, 1]$ over a shared (question, turn), 160 per pair—the paper's bounded-score model directly, with no win/loss conversion. Greedy allocation locks at the $1/149 \approx 0.007$ starvation plateau (Propositions 3 and D.2, now on continuous LLM-judge scores); uniform collapses under the dense-weak signal; threshold reallocation recovers power with budget and again *beats a naive oracle that knows the truths but spreads budget evenly* at the evaluated sparse-evidence budgets. e-mix-BH empirical FDR stays ≤ 0.0033, far below α=0.05 (the e-BH FDR guarantee holds; empirical FWER rises with power for the threshold allocator since e-BH controls FDR rather than FWER, and under the symmetrized-null stress test e-BH holds FWER ≤ 0.014). MT-Bench is an automated-evaluation replay distribution, not an independent ground truth: GPT-4 is both judge and a top contestant, so we report an exclude-GPT-4 sensitivity in the text.

allocator (e-value + e-BH stopper)	budget 2500	5000	10000	$1/K$ ref
greedy (max e-mix)	0.007 ($\pm$0.005)	0.007 ($\pm$0.005)	0.007 ($\pm$0.005)	$1/149$=0.007
greedy (min-p)	0.007 ($\pm$0.005)	0.007 ($\pm$0.005)	0.007 ($\pm$0.005)	$1/149$=0.007
uniform all	0.000 ($\pm$0.000)	0.000 ($\pm$0.000)	0.000 ($\pm$0.001)	—
top-k stable	0.060 ($\pm$0.014)	0.066 ($\pm$0.014)	0.067 ($\pm$0.014)	—
oracle (knows truths, even spread)	0.000 ($\pm$0.000)	0.000 ($\pm$0.001)	0.258 ($\pm$0.025)	—
threshold reallocation	**0.116 ($\pm$0.018)**	**0.236 ($\pm$0.025)**	**0.441 ($\pm$0.029)**	—

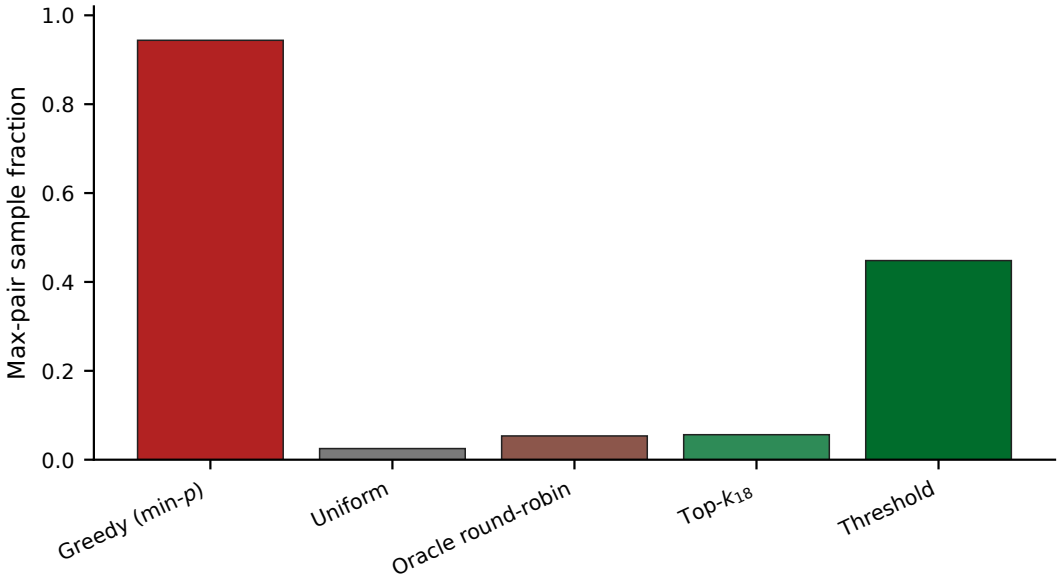

Figure 7: Max-pair sample fraction: greedy concentrates $\sim$94% on one pair; allocation-aware spreads budget.

Lemma B.1 (Pairwise bounded-difference betting factor)**.** *For bounded pairwise differences $X_t(h) \in [-1, 1]$, under the one-sided null $\mathbb{E}[X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] \leq 0$, the factor $1 + \lambda X_t(h)$, $\lambda \in [0, 1]$, has conditional expectation at most one when h is sampled.*

Proof. The factor is nonnegative because $X_t(h) \geq -1$ and $\lambda \leq 1$. Under the null,

$$\begin{aligned}\mathbb{E}[1 + \lambda X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] &= 1 + \lambda\, \mathbb{E}[X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] \\ &\leq\ 1.\end{aligned}$$

Multiplying selected-pair factors gives a nonnegative test supermartingale/e-process for each null pair, matching the safe-testing/e-process framing in Grünwald et al. (2024). □

Lemma B.2 (Finite mixtures remain valid)**.** *Let $E_h^{(1)}(t), \ldots, E_h^{(L)}(t)$ be valid e-processes for null h, and let $w_\ell \geq 0$ with $\sum_\ell w_\ell = 1$. Then $E_h^{\text{mix}}(t) = \sum_\ell w_\ell E_h^{(\ell)}(t)$ is also a valid e-process.*

Table 8: Threshold reallocation versus a naive oracle that knows the true alternatives but spreads budget uniformly over all of them (e-mix-BH power, reps=300, values as mean ± binomial MC SE). Adaptive retirement of resolved pairs concentrates budget on unresolved comparisons and can exceed the naive oracle; in dense-weak settings threshold approaches the oracle ceiling rather than exceeding it. e-BH controls FDR at α by Proposition 2 (FWER may exceed α when many true discoveries make $V/\max(R,1)$ small, which is consistent with FDR control).

phase	budget	oracle	threshold	top-k18	greedy	thr.-oracle Δ
K7_one_leader	500	1.000 (±0.000)	0.916 (±0.016)	0.821 (±0.022)	0.143 (±0.020)	-0.084
K7_one_leader	1000	1.000 (±0.000)	0.951 (±0.012)	0.980 (±0.008)	0.143 (±0.020)	-0.049
K7_one_leader	2000	1.000 (±0.000)	0.967 (±0.010)	0.993 (±0.005)	0.143 (±0.020)	-0.033
K13_two_leaders	500	0.825 (±0.022)	0.780 (±0.024)	0.453 (±0.029)	0.077 (±0.015)	-0.046
K13_two_leaders	1000	0.962 (±0.011)	0.861 (±0.020)	0.881 (±0.019)	0.077 (±0.015)	-0.101
K13_two_leaders	2000	0.999 (±0.002)	0.903 (±0.017)	0.935 (±0.014)	0.077 (±0.015)	-0.096
K18_three_leaders	500	0.056 (±0.013)	0.507 (±0.029)	0.097 (±0.017)	0.056 (±0.013)	+0.452
K18_three_leaders	1000	0.734 (±0.025)	0.734 (±0.026)	0.622 (±0.028)	0.056 (±0.013)	-0.000
K18_three_leaders	2000	0.963 (±0.011)	0.804 (±0.023)	0.786 (±0.024)	0.056 (±0.013)	-0.159
K28_dense_ladder	500	0.000 (±0.000)	0.306 (±0.027)	0.037 (±0.011)	0.036 (±0.011)	+0.306
K28_dense_ladder	1000	0.285 (±0.026)	0.537 (±0.029)	0.414 (±0.028)	0.036 (±0.011)	+0.252
K28_dense_ladder	2000	0.640 (±0.028)	0.819 (±0.022)	0.532 (±0.029)	0.036 (±0.011)	+0.179
K28_dense_weak	500	0.000 (±0.000)	0.298 (±0.026)	0.038 (±0.011)	0.036 (±0.011)	+0.298
K28_dense_weak	1000	0.260 (±0.025)	0.441 (±0.029)	0.372 (±0.028)	0.036 (±0.011)	+0.181
K28_dense_weak	2000	0.520 (±0.029)	0.502 (±0.029)	0.486 (±0.029)	0.036 (±0.011)	-0.018

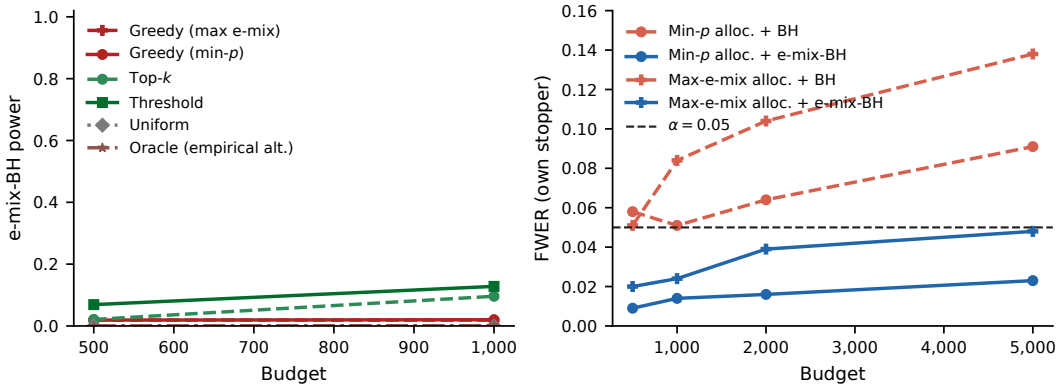

Figure 8: Arena adaptive replay at $M{=}6$ (small-M regime where fixed-time failure is sharpest). Left: e-mix-BH power by sampler/budget (greedy starves at $1/14$; allocation-aware recovers, threshold reaches 0.86). Right: symmetrized-null FWER in the 2×2 ablation (BH inflates to 0.138; e-BH stays $\leq \alpha$).

Proof. For any stopping time τ, $\mathbb{E}[E_h^{\mathrm{mix}}(\tau)] = \sum_\ell w_\ell \mathbb{E}[E_h^{(\ell)}(\tau)] \leq \sum_\ell w_\ell = 1$. This is the finite-mixture closure used for the mixture over lambdas in the simulations. □

C Top-k Linear Growth

Let S be an active set of size K, sampled in round-robin order after burn-in. For a true alternative $h \in S$, let $Y_{h,s} = \log e_{h,s}$ be the log e-factor on the s-th sample of h. Assume $a_h \leq Y_{h,s} \leq b_h$ and $\mathbb{E}[Y_{h,s} \mid \mathcal{F}_{h,s-1}] \geq \gamma_h > 0$. Let $N_h(n)$ be the number of post-burn-in samples allocated to h after n total post-burn-in samples. Round-robin gives $N_h(n) \geq \lfloor n/K \rfloor$.

Proposition C.1 (Active alternatives have linear log-e growth)**.** *With probability at least $1-\delta$, for every true alternative $h \in S$,*

$$\log E_h(n) \;\geq\; \gamma_h N_h(n) - (b_h - a_h)\sqrt{N_h(n)\,\log(K/\delta)/2}.$$

In particular, $\log E_h(n) \geq \gamma_h \lfloor n/K \rfloor - (b_h - a_h)\sqrt{\lfloor n/K \rfloor\,\log(K/\delta)/2}$.

Proof sketch. Define martingale differences $Z_{h,s} = Y_{h,s} - \mathbb{E}[Y_{h,s} \mid \mathcal{F}_{h,s-1}]$. The increments are bounded in an interval of width $b_h - a_h$. Hoeffding–Azuma gives, for each h,

$$\Pr\Big(\textstyle\sum_{s=1}^{N} Z_{h,s} \leq -(b_h - a_h)\sqrt{N\log(1/\delta_h)/2}\Big) \;\leq\; \delta_h.$$

Using $\mathbb{E}[Y_{h,s} \mid \mathcal{F}_{h,s-1}] \geq \gamma_h$ and setting $\delta_h = \delta/K$, apply a union bound over at most K active hypotheses. □

Proposition C.2 (Periodic-reload top-k preserves allocation guarantees)**.** *Let A_t be the active set at post-burn step t. For pair h, let $R_h(n) = \sum_{s=1}^{n} \mathbf{1}\{h \in A_s\}$, the number of post-burn steps where h is active. Under a periodic-reload top-k policy with reload period ρ and cap size $K \leq K_{\max}$, $N_h(n) \geq \lfloor R_h(n)/K_{\max} \rfloor$. Hence if $R_h(n) \geq \beta_h n$, then $N_h(n) \geq \lfloor \beta_h n/K_{\max} \rfloor$.*

Since $R_h(n)$ is under sampler control, this form separates allocation design from inference: if the refresh rule keeps a true alternative active for most steps (small churn), its sample count still grows linearly and Proposition C.1 applies.

Corollary C.1 (e-BH discovery scale)**.** *Let $r \geq 1$. If at least r active true alternatives satisfy*

$$\gamma_h \lfloor n/K \rfloor - (b_h - a_h)\sqrt{\lfloor n/K \rfloor \log(K/\delta)/2} \;\geq\; \log(m/(\alpha r)),$$

then, on the event in Proposition C.1, those alternatives each have $E_h(n) \geq m/(\alpha r)$. Hence the r-th largest e-value is at least $m/(\alpha r)$, and e-BH rejects at least r hypotheses. To guarantee that a named true alternative is itself rejected without an additional rank condition, use the conservative sufficient condition $E_h(n) \geq m/\alpha$.

D Greedy Allocation Power Upper Bound

Throughout this appendix we assume that for each true alternative h, when sampled, $\mathbb{E}[\log(1 + \lambda X_t(h)) \mid \mathcal{F}_{t-1}, A_t = h] \;\geq\; \gamma_{\log} > 0$ (a per-sample log-e drift lower bound). This is a direct condition on the log-e process; it does *not* follow from the score-mean condition $\mathbb{E}[X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] \geq \gamma$ alone, since $\log(1 + \lambda \cdot)$ is concave and Jensen gives $\mathbb{E}[\log(1 + \lambda X)] \leq \log \mathbb{E}[1 + \lambda X]$. For null pairs, $\mathbb{E}[\log(1 + \lambda X_t(h)) \mid \mathcal{F}_{t-1}, A_t = h] \leq 0$ holds by Jensen and the null condition (1). We write $w := \log((1 + \lambda)/(1 - \lambda))$ for the log-e increment width.

Proposition D.1 (Greedy power upper bound)**.** *Under the adaptive leaderboard model with e-BH at level α, burn-in b samples per pair satisfying $(1 + \lambda_{\max})^b < 1/\alpha$, and greedy allocation (each round sample the pair with the largest current e-value), the expected power over K true alternatives satisfies*

$$\mathbb{E}[\mathrm{power}_{\mathrm{greedy}}] \;\leq\; 1/K + \delta_{\mathrm{lock}}(T, b, \gamma, \lambda) + \alpha/(mK),$$

where $\gamma_{\log} > 0$ is the per-sample log-e drift lower bound assumed above, $w := \log\big((1+\lambda)/(1-\lambda)\big)$ is the log-e increment width, and

$$\delta_{\mathrm{lock}} \;\leq\; (m-1)\, T\, \exp\!\big(-2(D_0 + \gamma_{\log} T)^2/(Tw^2)\big),$$

with $D_0 > 0$ the log-e gap between the post-burn-in leader and any competitor. $\delta_{\mathrm{lock}} \to 0$ as the burn-in b (hence D_0) and drift γ grow.

Proof. Step 1 (deterministic burn-in cap). Each sampled betting factor satisfies $1+\lambda X_t \leq 1+\lambda_{\max}$ (since $X_t \leq 1$), so after burn-in b,

$$E_h(b) \;\leq\; (1 + \lambda_{\max})^b \;<\; 1/\alpha \;\leq\; m/(\alpha r) \quad \text{for every } r \leq m,$$

because the loosest e-BH threshold is $m/(\alpha m) = 1/\alpha$. Hence no pair is rejectable by e-BH at burn-in, and any pair that receives no post-burn-in samples keeps $E_h < 1/\alpha$ forever and is never discovered. Therefore true discoveries can only come from pairs that receive sustained post-burn-in sampling.

Step 2 (log-e lock-in). Suppose a true alternative $h^\star$ is the post-burn-in leader. For any other pair $h \neq h^\star$, define the log-e gap $D_h(t) := \log E_{h^\star}(t) - \log E_h(t)$. While $h^\star$ is sampled, $\log E_{h^\star}$ increases by $\log(1+\lambda X_s)$, a bounded increment in $[\log(1-\lambda), \log(1+\lambda)]$ of width $w = \log((1+\lambda)/(1-\lambda))$ with conditional mean at least $\gamma_{\log} = \log(1+\lambda\gamma) > 0$—a constant drift. $\log E_h$ is frozen because h is not sampled, so D_h is a submartingale with drift $\gamma_{\log}$ and bounded increments. By the Hoeffding–Azuma maximal inequality,

$$\Pr\big(\exists\, t \leq T : D_h(t) \leq 0 \;\big|\; D_h(0) = D_0 > 0\big) \\ \leq\; T\, \exp\!\big(-2(D_0 + \gamma_{\log} T)^2/(Tw^2)\big).$$

A union bound over the $m-1$ competitors gives δ_{lock}. On the lock-in event (probability $\geq 1-\delta_{\text{lock}}$), $h^\star$ is sampled at every post-burn-in step and every other true alternative is never sampled, hence (Step 1) never discovered: at most one true alternative is discovered.

Step 3 (a null leader does not discover—Ville). The argument does not require the leader to be a true alternative. If the leader $h^\star$ is a null pair, then under the null $\mathbb{E}[1+\lambda X \mid \mathcal{F}_{t-1}, A_t = h^\star] \leq 1$, so $E_{h^\star}$ is a nonnegative test supermartingale (Lemma B.1; the mixture version is preserved by Lemma B.2). By Ville's inequality, $\Pr(\sup_t E_{h^\star}(t) \geq m/\alpha) \leq \alpha/m$. So a null leader is rejected with probability at most α/m, and while it leads, no true alternative is sampled and none is discovered.

Step 4 (merge). On the good event (lock-in and no null crossing, probability $\geq 1 - \delta_{\text{lock}} - \alpha/m$), at most one true alternative is discovered, so $\text{power} \leq 1/K$. On the complement, $\text{power} \leq 1$. Therefore

$$\begin{aligned}\mathbb{E}[\text{power}] &\leq (1 - \delta_{\text{lock}} - \alpha/m) \cdot (1/K) + (\delta_{\text{lock}} + \alpha/m) \cdot 1 \\ &\leq 1/K + \delta_{\text{lock}} + \alpha/(mK).\end{aligned}$$

□

Remark. For the benchmark burn-in $b = 2$, $\lambda_{\max} = 0.8$, $\gamma \approx 0.1$: $\gamma_{\log} = \log(1.08) \approx 0.077$, $w = \log 9 \approx 2.20$, and $2\gamma_{\log}^2 T/w^2 \approx 5$ at $T = 2000$, so δ_{lock} is numerically loose at this burn-in. Proposition D.1 is therefore a structural upper bound ($1/K + \delta$ with $\delta \to 0$ as $b, \gamma \to \infty$); the empirical greedy plateau $\text{power} \approx 0.056 = 1/18$ across budgets 300–2000 supports that the true excess over $1/K$ is small in practice.

Proposition D.2 (Greedy power tight lower bound)**.** *In the setting of Proposition D.1, assume additionally that the true alternatives have log-e drifts $\gamma_{\log,h} \geq \gamma_{\log}^\star > 0$ (let $h^\star$ denote one of largest drift), and that the burn-in and budget satisfy $b(\gamma_{\log}^\star)^2 \gg w^2 \log m$ and $(T-bm)\gamma_{\log}^\star \gg \log(m/\alpha)$. Then*

$$\mathbb{E}[\text{power}_{\text{greedy}}] \;\geq\; 1/K - \varepsilon_{\text{lead}} - \delta_{\text{lock}} - \varepsilon_{\text{cross}},$$

with $\varepsilon_{\text{lead}}, \varepsilon_{\text{cross}} \to 0$ under these asymptotics. Combined with Proposition D.1, $\mathbb{E}[\text{power}_{\text{greedy}}] = 1/K \pm o(1)$: greedy discovers essentially exactly one true alternative, so its power is tight at $1/K$.

Proof. Let $h^\star$ be a true alternative with log-e drift $\gamma_{\log}^\star > 0$. Each log-e increment lies in $[\log(1-\lambda), \log(1+\lambda)]$ (width w); for a null pair, $\mathbb{E}[\log(1+\lambda X)] \leq \log \mathbb{E}[1+\lambda X] \leq 0$ by Jensen and the null condition (1).

Step 1 (the post-burn-in leader is a true alternative with high probability). After burn-in, $\log E_{h^\star}(b)$ has mean $b\gamma_{\log}^\star > 0$ while every null's $\log E_h(b)$ has nonpositive mean (Jensen and the null condition (1)). By Hoeffding–Azuma on the bounded increments, for any *null* competitor h,

$$\Pr(\log E_h(b) \geq \log E_{h^\star}(b)) \;\leq\; \exp\!\big(-b(\gamma_{\log}^\star)^2/(2w^2)\big).$$

A union bound over the $m - K$ null competitors gives

$$\begin{aligned}\Pr(\text{some null beats every true alternative at burn-in}) &\leq (m-K)\exp\left\{-\frac{b(\gamma_{\log}^\star)^2}{2w^2}\right\} \\ &=:\; \varepsilon_{\text{lead}} \to 0\end{aligned}$$

when $b(\gamma_{\log}^\star)^2/w^2 \gg \log m$. (We do not require $h^\star$ specifically to lead: if another true alternative leads, the same argument applies to it; greedy discovers whichever true alternative is the leader, which still yields power $1/K$.)

Step 2 (lock-in). Conditionally on $h^\star$ being the leader, the lock-in argument of Proposition D.1 Step 2 applies unchanged: with probability at least $1 - \delta_{\text{lock}}$, $h^\star$ is sampled at every post-burn-in step.

Step 3 (the leader crosses the e-BH threshold). Write $\log E_{h^\star}(T) = \log E_{h^\star}(b) + \sum_{s=bm+1}^{T} \mathbf{1}\{A_s = h^\star\} \log(1+\lambda X_s)$. The burn-in term $\log E_{h^\star}(b)$ has mean $b\gamma_{\log}^\star > 0$; on the Step-2 lock-in event the

post-burn-in sum has mean $(T-bm)\gamma^\star_{\log}$, so $\log E_{h^\star}(T)$ has mean at least $(T-bm)\gamma^\star_{\log}$ (the burn-in contribution only helps). By Hoeffding–Azuma on the bounded increments of the full trajectory,

$$\Pr\big(\log E_{h^\star}(T) < \log(m/\alpha)\big) \le \exp\left\{-\frac{2}{Tw^2}\big((T-bm)\gamma^\star_{\log} - \log(m/\alpha)\big)^2\right\} =: \varepsilon_{\text{cross}} \to 0$$

when $(T-bm)\gamma^\star_{\log} \gg \log(m/\alpha)$. On the complementary event $E_{h^\star}(T) \ge m/\alpha$, so $h^\star$ is rejected by e-BH (the e-Bonferroni threshold m/α implies e-BH rejection).

Step 4 (merge). On the event that $h^\star$ leads, locks in, and crosses—probability at least $1-\varepsilon_{\text{lead}}-\delta_{\text{lock}}-\varepsilon_{\text{cross}}$ by a union bound—greedy discovers at least $h^\star$, i.e. at least one true alternative, so power $\ge 1/K$. Therefore $\mathbb{E}[\text{power}] \ge (1-\varepsilon_{\text{lead}}-\delta_{\text{lock}}-\varepsilon_{\text{cross}})/K \ge 1/K - (\varepsilon_{\text{lead}}+\delta_{\text{lock}}+\varepsilon_{\text{cross}})$. □

Remark. The lower bound is tight in a *separated* asymptotic regime (large burn-in and drift): $\varepsilon_{\text{lead}}$ and $\varepsilon_{\text{cross}}$ vanish when $b(\gamma^\star_{\log})^2/w^2 \to \infty$ and $(T-bm)\gamma^\star_{\log} \to \infty$. Like the upper bound's δ_{lock}, these finite-sample constants are loose at the benchmark $b=2$; the matching property ($1/K+\delta$ above vs $1/K-\varepsilon$ here, both with vanishing error) is what makes Propositions D.1–D.2 a tight characterization, corroborated experimentally by the digit-for-digit $1/K$ greedy plateau across $K \in \{7, 13, 18, 28\}$ (Section 6.3) and on Arena data (Section 6.4).

E Threshold Reallocation

Threshold reallocation combines an evidence-driven sampler and an evidence-based rejection rule. The practical protocol has two variants: (i) *open-loop* reallocation, removing pairs using current procedure discoveries (high power, conditional analysis); and (ii) *guarded* reallocation, removing only after an e-value passes an e-Bonferroni bar, giving an unconditional false-removal bound. This section gives the guarded version a formal guarantee.

Let m be the number of hypotheses and α the target level. At each step, for each pair h compute a valid e-process value $E_h(t)$ and its e-Bonferroni rejection set $A_t = \{h : E_h(t) \ge m/\alpha\}$. Only pairs in A_t^c remain in the adaptive active sampling pool for the next step.

Lemma E.1 (Guarded false-removal probability)**.** *For a true-null pair h, define the crossing time $\tau_h = \inf\{t \ge 0 : E_h(t) \ge m/\alpha\}$. By Ville's inequality for nonnegative e-processes with $E_h(0) = 1$, $\Pr_h(\tau_h < \infty) \le \alpha/m$. Hence, by a union bound over at most m null hypotheses, $\Pr(\exists\, h \in \mathcal{H}_0 : \tau_h < \infty) \le \alpha$. So with probability at least $1-\alpha$, every pair removed by the guarded rule is truly non-null.*

Proposition E.1 (Guarded threshold reallocation, end-to-end guarantee)**.** *Consider guarded threshold reallocation after burn-in. At every post-burn step t, the sampler forms an active candidate pool $B_t \subseteq \mathcal{H} \setminus A_t$ of size at most $K_{\max}$, samples from B_t in round-robin order, and removes pairs only through $A_t = \{h : E_h(t) \ge m/\alpha\}$. Let U be a set of r true alternatives. Suppose that during the next n post-burn draws,*

$$R_h(n) := \sum_{t=1}^{n} \mathbf{1}\{h \in B_t\} \;\ge\; \beta n \qquad \textit{for every } h \in U,$$

and that, when sampled, each $h \in U$ has log-e increments $Y_{h,s}$ bounded in an interval of width w_h and satisfying $\mathbb{E}[Y_{h,s} \mid \mathcal{F}_{h,s-1}] \ge \gamma_h > 0$. Then with probability at least $1-\alpha-\delta$,

$$\log E_h(n) \;\ge\; \gamma_h \left\lfloor \frac{\beta n}{K_{\max}} \right\rfloor - w_h \sqrt{\left\lfloor \frac{\beta n}{K_{\max}} \right\rfloor \log(r/\delta)/2} \qquad \textit{for all } h \in U,$$

and no null pair has been removed. If the right-hand side is at least $\log(m/(\alpha r))$ for all $h \in U$, then e-BH rejects at least r true alternatives by time n.

Proof. Lemma E.1 gives an event $\mathcal{F}$ of probability at least $1-\alpha$ on which no null pair ever crosses the e-Bonferroni removal bar, hence guarded reallocation never removes a true null. On $\mathcal{F}$, removal

can only retire true alternatives. For each $h \in U$, the round-robin rule over pools of size at most $K_{\max}$ gives

$$N_h(n) \;\geq\; \Big\lfloor R_h(n)/K_{\max} \Big\rfloor \;\geq\; \Big\lfloor \beta n/K_{\max} \Big\rfloor.$$

Apply the Hoeffding–Azuma argument of Proposition C.1 to the r alternatives in U, using δ/r for each alternative and then union bounding. This yields the displayed simultaneous log-e lower bound with conditional probability at least $1-\delta$. Intersecting this event with $\mathcal{F}$ gives probability at least $1-\alpha-\delta$. If all r alternatives have $E_h \geq m/(\alpha r)$, then the r-th largest e-value is at least $m/(\alpha r)$, so the e-BH rank rule rejects at least r hypotheses; because every member of U is non-null, these are at least r true discoveries. □

The high-power default in the main text is *open-loop* reallocation, which retires pairs using the full `e_mix_bh` rejection set rather than the stricter e-Bonferroni bar. This is the empirically strongest protocol, and it also carries a clean end-to-end guarantee. The key observation is that reallocation is an *allocation* rule, not a *rejection* rule: it decides where future samples go, never which discoveries are reported, so its validity is inherited from the e-process/e-BH layer for free.

Proposition E.2 (Open-loop threshold reallocation: unconditional FDR and conditional power)**.** *Run open-loop threshold reallocation: after each update, irrevocably retire from the active sampling pool every pair in the current e-BH rejection set R_t (its e-value is then frozen), and allocate future draws round-robin over the surviving active pairs. Report R_T, the e-BH set at the budget T (computed over the frozen e-values of retired pairs and the current e-values of active pairs).*

1. *(Unconditional FDR) For every budget T, the reported set R_T controls $\mathrm{FDR}(T) \leq \alpha$.*
2. *(Conditional power, no false-removal penalty) Let $U \subseteq \mathcal{H}_1$ be any r true alternatives that are still active (not yet retired) at some post-burn-in step, and suppose they remain in the active pool for at least a β fraction of the next n draws, with active pool size at most $K_{\max}$. If each $h \in U$ has bounded log-e increments with conditional mean at least $\gamma_h > 0$, then with probability at least $1-\delta$ the simultaneous linear-growth bound of Proposition E.1 holds for all $h \in U$, so each E_h reaches at least $m/(\alpha r)$ and e-BH discovers at least these r alternatives by time n.*

Proof. (1) Index the filtration so that $\mathcal{F}_t$ contains all observations through step t. After the step-t update the rejection set R_t is $\mathcal{F}_t$-measurable; the rule retires the pairs in R_t and selects the next draw A_{t+1} among the surviving pairs as an $\mathcal{F}_t$-measurable function, so the allocation is predictable in the sense of Section 2. Each null e-process stays valid under any predictable allocation, since $\mathbb{E}[1+\lambda X_t(h) \mid \mathcal{F}_{t-1}, A_t = h] \leq 1$ regardless of how A_t was chosen. A pair retired at the stopping time $\tau_h = \inf\{t : h \in R_t\}$ has its e-process stopped at τ_h; by optional stopping for nonnegative supermartingales, $E_h(\tau_h)$ remains a valid e-value, as does $E_h(T)$ for pairs still active at T. The reported set is e-BH applied to these stopped e-values, so Proposition 2 gives $\mathrm{FDR}(T) \leq \alpha$. The e-values are stopped at different, data-dependent times—one stopping time per pair—but the e-BH FDR bound requires only that each null input be a valid e-value marginally (it holds under arbitrary dependence among the e-values), not that all inputs share a common stopping time, so heterogeneous retirement does not affect validity. Reallocation is an allocation rule: it never overrides or augments the e-BH rejection decision, so validity is unconditional.

(2) The conclusion concerns the still-active alternatives in U, whose e-values are not frozen and continue to accumulate. Under round-robin allocation over pools of size at most $K_{\max}$, conditioning on the event that each $h \in U$ stays in the active pool for a β fraction of n draws gives the count $N_h(n) \geq \lfloor \beta n/K_{\max} \rfloor$ as in Proposition E.1, and the Hoeffding–Azuma argument yields the simultaneous log-e lower bound with probability at least $1-\delta$. Once each $E_h \geq m/(\alpha r)$, the r-th largest e-value is at least $m/(\alpha r)$, so the e-BH rank rule (Lemma F.2) rejects at least r hypotheses, including the r members of U, which are all non-null. A true alternative is retired only after entering R_t, i.e. only once already discovered, so it is never starved by its own removal; false removals of null pairs cannot retire any member of $U \subseteq \mathcal{H}_1$ and only free budget. Hence no $1-\alpha$ term (such as the guarded variant incurs via Lemma E.1) is needed for the power of the active true alternatives. □

Remark (The power claim is conditional, and concerns active alternatives). Proposition E.2(1) is an *unconditional* FDR statement; part (2) is a *conditional* power statement, predicated on the active

true alternatives receiving a round-robin share of the draws (the $\beta, K_{\max}$ condition), exactly as in Proposition E.1. The conclusion is that the r *still-active* alternatives in U are discovered by the budget; it is not a durability claim that every pair ever retired remains in R_T; R_T is the e-BH snapshot at the budget, the standard analysis-time rejection set. This matches the simulator, which reports e-BH discoveries at the budget T.

Remark (What open-loop does and does not give, relative to guarded). Propositions E.1 and E.2 separate three concerns. (i) *FDR validity* is unconditional and identical for both variants—it is inherited from Proposition 2 because reallocation is an allocation rule. (ii) *Power for discovered true alternatives* is $1-\delta$ for both; the guarded variant's $1-\alpha-\delta$ bound in Proposition E.1 is the price of an additional property, not of validity or power. (iii) The property guarded buys and open-loop does not is a *purity guarantee on the removed set*: on a $1-\alpha$ event, every pair the guarded rule retires is truly non-null (Lemma E.1), which matters if downstream code assumes retirement certifies a true alternative. Open-loop instead removes on the lower e-BH bar, so it retires pairs sooner—freeing budget earlier, which is why it is empirically stronger at low budget—at the cost of allowing null false removals within the FDR guarantee. The experiments use open-loop as the high-power default and report guarded as the purity-certified variant.

F Allocation-Dependent Sample Complexity for E-BH Discovery

This section gives the proof of Proposition 5.

Let $m = M(M-1)$ be the number of directed pairwise hypotheses. Let $\mathcal{H}_1$ be the set of $s = |\mathcal{H}_1|$ true alternatives, each with conditional mean log-e-increment drift $\gamma_h \geq \gamma > 0$ when sampled. Assume the log-e-increments are bounded: $a_h \leq Y_{h,t} = \log(1+\lambda X_t(h)) \leq b_h$, $w_h = b_h - a_h$, $w = \max_h w_h$. Define the r-discovery budget $T_r(P,\delta)$ for a protocol P as the minimum post-burn-in sample count n such that with probability at least $1-\delta$, e-BH discovers at least r true alternatives.

Lemma F.1 (Per-hypothesis log-e growth with high probability)**.** *For a single true alternative h, if h is sampled N_h times with conditional mean drift γ_h, then for any $\delta_h \in (0,1)$,*

$$\Pr\big(\log E_h \;\leq\; \gamma_h N_h - w_h\sqrt{N_h \log(1/\delta_h)/2}\big) \;\leq\; \delta_h.$$

Proof. $\log E_h(t)$ is a sum of bounded martingale differences with conditional mean at least γ_h. By Hoeffding–Azuma, the deviation probability is bounded by $\exp(-2(\gamma_h N_h - \log E_h)^2/(N_h w_h^2))$. Setting this to δ_h and solving for $\log E_h$ yields the bound (see Proposition C.1 for the construction). □

Lemma F.2 (e-BH discovery threshold)**.** *Let $E_{(1)} \geq \cdots \geq E_{(m)}$ be the sorted e-values. If at least r e-values satisfy $E_h \geq m/(\alpha r)$, then e-BH rejects at least r hypotheses.*

Proof. If $E_{(r)} \geq m/(\alpha r)$, then $k = \max\{j :\; E_{(j)} \geq m/(\alpha j)\}$ satisfies $k \geq r$ because $E_{(r)} \geq m/(\alpha r)$ and the threshold $m/(\alpha j)$ is decreasing in j. □

Proposition F.1 (Uniform allocation: sufficient budget)**.** *Under uniform allocation, the r-discovery budget satisfies*

$$T_r^{\text{uniform}}(\delta) \;\geq\; \frac{mw^2}{2\gamma^2}\log\frac{m}{\alpha r\delta} \;+\; \mathcal{O}\big(\sqrt{T_r^{\text{uniform}}(\delta)\log m}\big).$$

Proof. Under uniform allocation, each pair receives $N_h = \lfloor n/m \rfloor$ samples. By Lemma F.1 with $\delta_h = \delta/m$ and a union bound, with probability at least $1-\delta$ every true alternative satisfies

$$\log E_h(n) \;\geq\; \gamma\lfloor n/m\rfloor - w\sqrt{\lfloor n/m\rfloor\,\log(m/\delta)/2}.$$

For e-BH to discover at least r alternatives (Lemma F.2), at least r true alternatives must exceed $m/(\alpha r)$. A necessary condition is

$$\gamma\lfloor n/m\rfloor - w\sqrt{\lfloor n/m\rfloor\,\log(m/\delta)/2} \;\geq\; \log(m/(\alpha r)).$$

Let $x = \sqrt{\lfloor n/m\rfloor}$. This is a quadratic $\gamma x^2 - w\sqrt{\log(m/\delta)/2}\,x - \log(m/(\alpha r)) \geq 0$. Solving and squaring and ignoring floor/lower-order terms yields the stated bound; the dominant term is $(mw^2)/(2\gamma^2)\log(m/(\alpha r\delta))$. □

Remark (Scope of Proposition F.1). Proposition F.1 is *not* a lower bound on the budget any FDR-controlling strategy must spend. It is the budget at which the Hoeffding–Azuma sufficient condition for r discoveries becomes feasible under uniform allocation. A true matching lower bound via change-of-measure (in the style of Kaufmann et al. Kaufmann et al. (2016)) is given in Appendix G.

Proposition F.2 (Top-k allocation upper bound)**.** *Under top-k round-robin allocation with active set size $k \geq s$, the s-discovery budget satisfies*

$$T_s^{\textit{top-k}}(\delta) \;\leq\; \frac{kw^2}{2\gamma^2}\Big[\sqrt{\log(k/\delta)} + \sqrt{\log(k/\delta) + (4\gamma/w^2)\log(m/(\alpha s))}\Big]^2 \;+\; o(k).$$

Proof. Under top-k round-robin over an active set of size k, each active pair receives $N_h \geq \lfloor n/k \rfloor$ samples. By Lemma F.1 with $\delta_h = \delta/k$ and a union bound, with probability at least $1-\delta$ every true alternative in the active set satisfies

$$\log E_h(n) \;\geq\; \gamma\lfloor n/k \rfloor - w\sqrt{\lfloor n/k \rfloor \, \log(k/\delta)/2}.$$

To guarantee s discoveries via Lemma F.2 it suffices that all s true alternatives satisfy $E_h \geq m/(\alpha s)$, requiring $\gamma\lfloor n/k \rfloor - w\sqrt{\lfloor n/k \rfloor \, \log(k/\delta)/2} \geq \log(m/(\alpha s))$. Solving the same quadratic in $x = \sqrt{\lfloor n/k \rfloor}$ yields the bound. When $k = s$, the dominant term is $(sw^2)/(2\gamma^2)\log(m/(\alpha s\delta))$, smaller than the uniform lower bound by a factor of m/s. □

Proposition F.3 (Separation: upper-bound ratio)**.** *When $s = o(m)$ and $k = \Theta(s)$, the ratio of the sufficient budgets satisfies*

$$T_s^{\text{uniform}}(\delta) \,/\, T_s^{\textit{top-k}}(\delta) \;=\; \mathcal{O}(m/k) \;\to\; \infty \quad \textit{as } m/k \to \infty.$$

Proof. Take the ratio of the dominant logarithmic terms from Propositions F.1 and F.2. For large m relative to k, the pre-factor m/k dominates asymptotically. □

Proposition F.4 (Greedy allocation lower bound)**.** *There exists a configuration with $s \geq 2$ true alternatives, each with drift $\gamma > 0$, such that for any finite budget T, greedy allocation achieves*

$$\Pr\big(|R_T^{\text{greedy}} \cap \mathcal{H}_1| \leq 1\big) \;\geq\; 1 - \epsilon.$$

for arbitrarily small $\epsilon > 0$ by choosing the burn-in lead appropriately.

Proof. Let h_1, h_2 be two true alternatives with identical drift. After burn-in B, let $D_0 = \log E_{h_1}(B) - \log E_{h_2}(B) > 0$. Let $D_t = \log E_{h_1}(t) - \log E_{h_2}(t)$; when h_1 is sampled, $D_{t+1} - D_t$ is bounded with expected value γ. By the Hoeffding–Azuma maximal inequality, $\Pr(\min_{0\leq t\leq T} D_t \leq 0) \leq T\exp(-2(D_0+\gamma T)^2/(Tw^2))$. For fixed T, ϵ, choose D_0 large enough (via burn-in) that this is at most ϵ. On the complementary event, h_1 is sampled at every post-burn-in step and h_2 receives no additional evidence, so at most h_1 is discovered. □

G Matching Lower Bound for Allocation-Dependent E-BH Discovery

Section F gives concentration-derived upper bounds on the budgets sufficient for e-BH discovery. This section gives a matching lower bound via a change-of-measure argument that holds for *any* FDR-controlling discovery rule at a stopping time (not only e-BH) under a stated sampling constraint, upgrading the separation to $\Theta(m/k)$.

Work in the Gaussian model: a sample of hypothesis h is $X \sim \mathcal{N}(\mu_h, \sigma^2)$, with $\mu_h = \gamma > 0$ for a true alternative and $\mu_h = 0$ for a null. Let θ be a configuration in which a fixed set S of s hypotheses are alternatives and the remaining $m - s$ are null. For $h \in S$, let θ'_h be the configuration identical to θ except that h is null. The per-sample KL between θ and θ'_h is $\mathrm{KL}_h = \gamma^2/(2\sigma^2)$ at h and 0 elsewhere. The evaluator runs *any* discovery rule that controls FDR at level α at an a.s. finite stopping time τ; the rule need not be e-BH, and the argument below uses no e-process.

Condition (C1): $\alpha s < 1 - \beta$, where $1 - \beta$ is the per-hypothesis power target. The weaker $\alpha s < 1$ already suffices for Lemma G.1 to give a nontrivial per-null bound, but the sample lower bound in Lemma G.2 needs $\alpha s < 1 - \beta$ so that the binary KL below is well-defined and monotone.

Lemma G.1 (FDR-implied per-null rejection bound)**.** *Under θ'_h, for* any *procedure controlling FDR at level α (not only e-BH),*

$$\Pr_{\theta'_h}(h \in R_\tau) \;\le\; \alpha s.$$

Proof. Under θ'_h the hypothesis h is null and at most $s - 1$ of the other hypotheses are alternatives, so on the event $h \in R_\tau$ we have $V_\tau \ge 1$ and $|R_\tau| = V_\tau + T_\tau$ with $T_\tau \le s - 1$; hence the false-discovery proportion satisfies $V_\tau / \max(|R_\tau|, 1) = V_\tau/(V_\tau + T_\tau) \ge V_\tau/(V_\tau + s - 1) \ge 1/s$. Therefore, pointwise,

$$\mathbf{1}\{h \in R_\tau\} \;\le\; s \cdot \frac{V_\tau}{\max(|R_\tau|, 1)}.$$

Taking expectations under θ'_h and using FDR control $\mathbb{E}_{\theta'_h}[V_\tau/\max(|R_\tau|, 1)] \le \alpha$ gives $\Pr_{\theta'_h}(h \in R_\tau) \le s \cdot \mathrm{FDR}_{\theta'_h}(\tau) \le \alpha s$. This uses only the definition of FDR control—no e-BH rank-threshold structure and no e-process. In particular, Condition (C1), $\alpha s < 1$, is not an artifact of e-BH but the intrinsic sparsity regime in which a nontrivial per-null bound is available from FDR control alone. □

Lemma G.2 (Per-hypothesis sample lower bound)**.** *If the procedure discovers h with probability at least $1 - \beta$ under θ, then*

$$\mathbb{E}_\theta[N_h(\tau)] \;\ge\; (2\sigma^2/\gamma^2)\,\mathrm{kl}(1 - \beta,\, \alpha s),$$

so the right-hand side is $\Omega\big((\sigma^2/\gamma^2)\log(1/(\alpha s))\big)$ under (C1).

Proof. Apply the change-of-measure inequality of Kaufmann, Cappé & Garivier Kaufmann et al. (2016), valid for any a.s. finite stopping time τ and any event $\mathcal{E} \in \mathcal{F}_\tau$:

$$\textstyle\sum_{h'} \mathbb{E}_\theta[N_{h'}(\tau)]\,\mathrm{KL}_{h'} \;\ge\; d_{\mathrm{kl}}\big(\Pr_\theta(\mathcal{E}),\, \Pr_{\theta'_h}(\mathcal{E})\big).$$

Since θ and θ'_h differ only at h, all other $\mathrm{KL}_{h'} = 0$, giving $\mathbb{E}_\theta[N_h(\tau)]\,\gamma^2/(2\sigma^2) \ge \mathrm{kl}(\Pr_\theta(\mathcal{E}), \Pr_{\theta'_h}(\mathcal{E}))$. Take $\mathcal{E} = \{h \in R_\tau\}$: $\Pr_\theta(\mathcal{E}) \ge 1 - \beta$ and $\Pr_{\theta'_h}(\mathcal{E}) \le \alpha s < 1 - \beta$ by Lemma G.1. The binary KL $d_{\mathrm{kl}}(p, q)$ is strictly decreasing in q for $q < p$ and strictly increasing in p (since $\partial d_{\mathrm{kl}}/\partial q = (q - p)/(q(1 - q)) < 0$ for $q < p$), hence $d_{\mathrm{kl}}(\Pr_\theta(\mathcal{E}), \Pr_{\theta'_h}(\mathcal{E})) \ge d_{\mathrm{kl}}(1 - \beta, \alpha s)$. □

Corollary G.1 (Bounded-preference information ceiling)**.** *The same per-hypothesis lower bound holds for any bounded one-parameter pairwise preference family whose null member is θ_0 and whose alternative member θ_γ has mean advantage γ, whenever the per-sample KL satisfies*

$$\mathrm{KL}(\theta_\gamma, \theta_0) \;\le\; C\gamma^2$$

for small γ. In that case every FDR-controlling sequential discovery rule with per-hypothesis power at least $1 - \beta$ must allocate

$$\mathbb{E}_{\theta_\gamma}[N_h(\tau)] \;\ge\; C^{-1}\gamma^{-2}\,\mathrm{kl}(1 - \beta, \alpha s)$$

samples to h in the sparse regime $\alpha s < 1 - \beta$.

Proof. Repeat the proof of Lemma G.2. The change-of-measure inequality gives

$$\mathbb{E}_{\theta_\gamma}[N_h(\tau)]\;\mathrm{KL}(\theta_\gamma, \theta_0) \ge \mathrm{kl}(\Pr_{\theta_\gamma}(h \in R_\tau), \Pr_{\theta_0}(h \in R_\tau)).$$

The first probability is at least $1 - \beta$ by the power requirement and the second is at most αs by Lemma G.1. The assumed local KL upper bound then yields the claim. Standard bounded Bernoulli, trinomial win/loss/tie, and smooth Bradley–Terry local alternatives satisfy $\mathrm{KL}(\theta_\gamma, \theta_0) = \mathcal{O}(\gamma^2)$ around interior null parameters. □

Proposition G.1 (Matching allocation-dependent separation)**.** *For any discovery rule that controls FDR at level α over the parameter class with* at most *s alternatives (so θ'_h lies in the guaranteed class), run with a fixed budget T (stopping time $\tau \le T$ a.s.), under (C1) and a per-hypothesis power target $\Pr_\theta(h \in R_\tau) \ge 1-\beta$, with a sampling constraint (uniform: each pair receives $\le T/m$ samples; top-k round-robin over an active set of size $k \ge s$: each active pair receives $\le T/k$), the minimum budget $T^\star$ to discover a given alternative satisfies*

$$\begin{aligned} T^\star_{\text{uniform}} &= \Omega\big(m\,(\sigma^2/\gamma^2)\log(1/(\alpha s))\big), \\ T^\star_{\textit{top-}k} &= \Omega\big(k\,(\sigma^2/\gamma^2)\log(1/(\alpha s))\big), \end{aligned}$$

matching the sufficient-budget upper bounds of Propositions F.1–F.2 up to logarithmic factors (the lower-bound log is $\log(1/\alpha s)$ vs. the upper-bound $\log(m/\alpha s\delta)$, differing by $\mathcal{O}(\log m)$). Combined with them, $T^\star_{\text{uniform}}/T^\star_{\textit{top-}k} = \widetilde{\Theta}(m/k)$ in the sparse regime $s = o(m)$, $k = \Theta(s)$.

Proof. Under uniform allocation each hypothesis receives at most T/m samples, so $N_h(\tau) \le T/m$; Lemma G.2 gives $T/m \ge (2\sigma^2/\gamma^2)\,\mathrm{kl}(\dots)$, i.e. $T^\star_{\text{uniform}} \ge (2m\sigma^2/\gamma^2)\,\mathrm{kl}(\dots)$. Under top-$k$ round-robin over an active set of size k, each active hypothesis receives at most T/k, giving $T^\star_{\text{top-}k} \ge (2k\sigma^2/\gamma^2)\,\mathrm{kl}(\dots)$. The lower-bound logarithmic factor $\log(1/(\alpha s))$ versus the upper-bound $\log(m/(\alpha s\delta))$ differs by $\mathcal{O}(\log m)$, which does not affect the $\Theta(m/k)$ ratio. □

Remark (Scope). Proposition G.1 holds for *any* FDR-controlling discovery rule at a stopping time: Lemma G.1 uses only the definition of FDR control (no e-BH rank-threshold structure), and Lemma G.2 uses the Kaufmann change-of-measure inequality with the observation KL $\gamma^2/(2\sigma^2)$ (no e-process). Two scope limits remain. First, the lower bound is stated under a *sampling constraint* (uniform, or top-k round-robin of size k): an unconstrained adaptive allocator that concentrates samples on the true alternatives need not incur the m factor, so this is not a minimax converse over arbitrary allocations. Second, $\alpha s < 1$ (Condition C1) is the intrinsic sparsity regime in which FDR control yields a nontrivial per-null bound (Lemma G.1), not an artifact of e-BH.