EVIDENCE-GATED RESEARCH: STATISTICALLY CONTROLLED MODEL ADOPTION IN ADAPTIVE SEARCH

Anonymous authors
Paper under double-blind review

ABSTRACT

An accepted model becomes the reference for later proposals in adaptive search, so an erroneous replacement can alter hypotheses that have not yet been generated. Evidence-Gated Research (EGR) studies model adoption when improvement must hold across a predeclared set of environments. EGR freezes each challenger before decision evidence is examined, constructs anytime-valid environment-specific evidence under predictable sampling, and passes their intersection–union evidence to an online adoption controller. Under a conditional mean bound for each sampled evidence unit, the procedure controls false discovery rate for the declared all-environment adoption target even when earlier adoptions change later challengers and null hypotheses. In a 5,000-trajectory closed-loop benchmark, development-only e-LOND has persistent-target FDR 0.621, whereas no persistent false-adoption path is observed for the audited EGR variants in that finite run. In matched replay over 600 challenger–incumbent pairs, Stagewise EGR and fixed-anytime IUT have identical adoption events by construction; at $h = 20$, Stagewise uses 56.1% less decision evidence by stopping resolved components and avoiding later stages after failure. EGR-ALLOC further reallocates saved evidence to unresolved environments. Three-environment public-data and controlled neural studies show similar reductions in decision-evidence cost. These results characterize when sequential evidence can support statistically controlled model replacement in an adaptive search.

1 INTRODUCTION

Model development repeatedly asks whether a new candidate should replace the current model. AutoML and hyperparameter optimization formalize this search (Baratchi et al., 2024), while research agents extend it to code changes, training runs, and experiment selection (Chan et al., 2025; Toledo et al., 2025; Lu et al., 2026; Aygün et al., 2026; Ghareeb et al., 2026). Tree-search and multi-agent systems use accepted results to guide later proposals (Aygün et al., 2026; Ghareeb et al., 2026). An unsupported replacement can therefore affect comparisons that have not yet been proposed.

The replacement decision has two statistical requirements. Evidence must address the environments where an improvement is required, and error control must account for repeated candidates. A test on development data alone cannot establish a gain in an audit environment. Testing each candidate with an anytime-valid rule also leaves multiplicity across candidates unaddressed. Once these requirements are met, the remaining question is how to avoid spending observations on environments that already support adoption.

EGR freezes each challenger before its decision evidence is revealed. It then collects evidence across the declared environments and applies an online threshold to their combined result. The next challenger may depend on everything learned so far; the current comparison must remain valid conditional on the history that produced it.

Contributions. We formulate model replacement as an online sequence of history-dependent, moving-incumbent hypotheses. We give sufficient conditions under which frozen-candidate, multi-environment decision evidence yields conditionally valid candidate e-values under predictable routing and component stopping, and show that these values can be passed to e-LOND for FDR control despite adoption-dependent future hypotheses; a stricter threshold gives FWER control. We then

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.

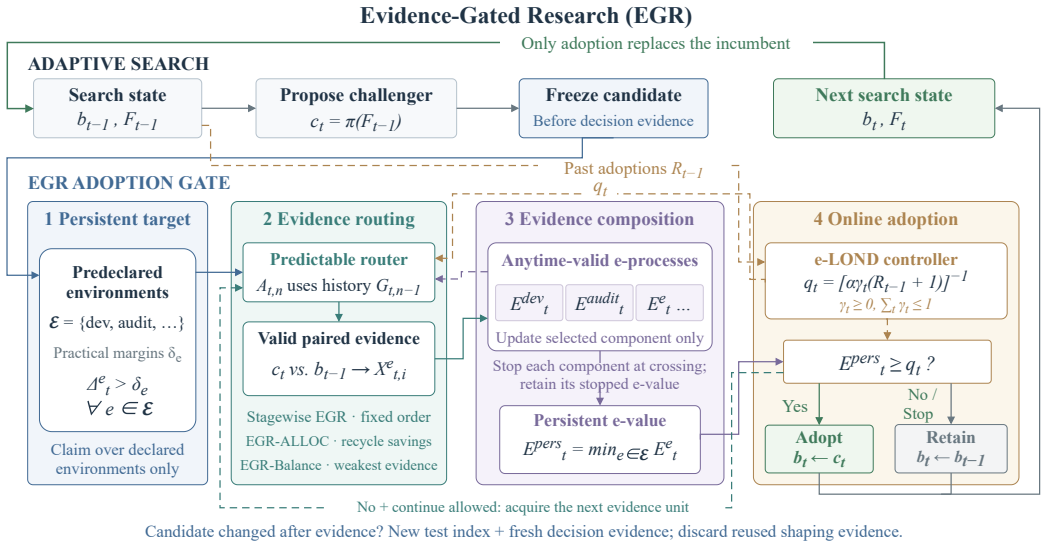


Figure 1: EGR places the statistical decision at model replacement. The challenger is fixed before decision evidence is opened; environment-specific e-processes are updated under predictable routing, combined for the declared persistent target, and passed to an online controller. Only an adoption changes the incumbent.

separate evidence acquisition from validity: Stagewise EGR short-circuits failed intersection–union tests, EGR-ALLOC recycles saved budget, and EGR-Balance targets weak evidence components. Closed-loop experiments evaluate path-altering errors, matched replay isolates acquisition cost on identical pairs, and leakage/dependence diagnostics expose failures of the evidence contract.

2 Related Work

The reusable holdout addresses repeated adaptive access to validation data (Dwork et al., 2015). Learn-then-test (LTT) casts risk calibration as multiple testing (Angelopoulos et al., 2025), and adaptive LTT uses e-processes for sequential hyperparameter selection (Zecchin et al., 2025). Model selection under temporal shift compares performance in changing environments (Han et al., 2024). Automated-science systems generate increasingly complex sequences of proposals (Lu et al., 2026; Aygün et al., 2026; Ghareeb et al., 2026). EGR studies the replacement decisions within such a sequence: each accepted model becomes the incumbent, and each gain must hold across a declared set of environments.

Always-valid A/B inference permits continuous monitoring and optional stopping (Johari et al., 2022). Betting-based e-processes provide the sequential evidence used here (Ramdas et al., 2023; Waudby-Smith and Ramdas, 2024; Pandeva et al., 2024; Ramdas and Wang, 2025). Given valid candidate e-values, e-LOND controls online FDR (Xu and Ramdas, 2024); e-GAI methods improve spending efficiency (Zhang et al., 2025), and online e-closure/compound-e methods offer further improvements (Xu et al., 2026). EGR does not modify the online controller itself: it constructs candidate evidence and specifies where that evidence may be collected so that an existing controller can be used after the challenger and incumbent have been generated by the evolving search state.

The persistent-benefit target has the classical intersection–union structure: every environment-specific alternative must hold (Berger and Hsu, 1996). EGR combines this intersection–union test (IUT) construction with anytime-valid component evidence in a search whose incumbent changes after adoption.

Table 1 separates evidence construction from online control. In EGR, the incumbent in $H_t(c_t, b_{t-1})$ records earlier adoption decisions, so later hypotheses depend on the realized search history.

Table 1: Statistical objects addressed by related frameworks: aLTT (Zecchin et al., 2025), e-LOND (Xu and Ramdas, 2024), and e-GAI (Zhang et al., 2025). Entries describe the cited formulations; "Not central" means that the feature is not an explicit object of analysis there, and does not rule out later extensions.

Object	aLTT	e-LOND / e-GAI	EGR
Hypothesis object	Candidate / hyperparameter	Supplied e-value stream	Challenger vs. current incumbent
Prior-adoption dependence	Not central	Not explicitly modeled	Explicit moving incumbent
Multi-env. intersection target	Not native	Not constructed	Explicitly constructed
Candidate freeze / restart	Not central	Outside controller	Explicit freeze/restart condition
Within-candidate evidence	Sequential	Candidate e-value supplied	Predictable multi-env. routing
Across-candidate control	Family selection	Online FDR	Online control of adoption stream

3 Persistent Adoption in an Adaptive Search Loop

Let $\mathcal{F}_{t-1}$ be the history when challenger t is frozen. It includes earlier tests, the scout data used to construct this challenger, and any proposal randomization. Thus

$$c_t = \pi(\mathcal{F}_{t-1}), \tag{1}$$

and both c_t and incumbent b_{t-1} are $\mathcal{F}_{t-1}$-measurable. Fix a finite, nonempty set of environments $\mathcal{E}$. For each e, let P_t^e be the reference distribution specified before decision sampling. Define the observed paired gain and its target mean by

$$X_{t,i}^e = U^e(c_t; Z_{t,i}^e) - U^e(b_{t-1}; Z_{t,i}^e), \qquad \Delta_t^e = \int [U^e(c_t; z) - U^e(b_{t-1}; z)]\, P_t^e(dz). \tag{2}$$

The reference law, utilities, and margin $\delta_e \geq 0$ are fixed given $\mathcal{F}_{t-1}$. The margin is the smallest gain that would justify replacement; it is zero when any positive gain matters. The benchmarks fix margins before each run. Assumption 1 below links the declared target to the observations actually sampled.

We call a challenger *persistently beneficial with respect to the declared environment set* if $\Delta_t^e > \delta_e$ in every declared environment. The persistent null is

$$H_t^{\mathrm{pers}} = \bigcup_{e \in \mathcal{E}} H_t^e, \qquad H_t^e : \Delta_t^e \leq \delta_e. \tag{3}$$

Write $\nu_t^e = \mathbf{1}\{\Delta_t^e \leq \delta_e\}$ and $\nu_t = \max_e \nu_t^e$. These truth indicators are $\mathcal{F}_{t-1}$-measurable; this does not mean they are known to the planner. If D_t indicates adoption, let $R_T = \sum_{t \leq T} D_t$ and $V_T = \sum_{t \leq T} \nu_t D_t$. We target

$$\mathrm{FDR}(T) = \mathbb{E}\left[\frac{V_T}{R_T \vee 1}\right], \qquad \mathrm{FWER}(T) = \mathbb{P}(V_T \geq 1). \tag{4}$$

"Persistent" refers to this all-environment target, not to undeclared future shifts. Earlier adoptions may change the next challenger and its null, provided that each comparison is fixed before its decision evidence is revealed.

4 Evidence-Gated Research

4.1 Anytime-valid components and persistent composition

Let $\mathcal{G}_{t,n}$ contain the full history after n allocations for candidate t, with $\mathcal{G}_{t,0} = \mathcal{F}_{t-1}$. The router chooses $A_{t,n} \in \mathcal{E}$ using $\mathcal{G}_{t,n-1}$. Set $N_{t,n}^e = \sum_{s \leq n} \mathbf{1}\{A_{t,s} = e\}$. On $\{A_{t,n} = e\}$, the revealed unit is $X_{t,N_{t,n}^e}^e$. All within-candidate stopping times below use this joint history.

Assumption 1 (Sequentially valid evidence unit)**.** *Selected updates are bounded. On* $\{\nu_t^e = 1, A_{t,n} = e\}$*, the acquisition mechanism satisfies*

$$\mathbb{E}[X^e_{t,N^e_{t,n}} - \delta_e \mid \mathcal{G}_{t,n-1}] \le 0. \tag{5}$$

The selected variable is needed only on $\{A_{t,n} = e\}$ *and can be set to zero elsewhere.*

This condition is stronger than a marginal mean bound. For example, conditional on $\mathcal{F}_{t-1}$, a fresh independent draw $Z^e \sim P_t^e$ gives $\mathbb{E}[X^e - \delta_e \mid \mathcal{G}_{t,n-1}, A_{t,n} = e] = \Delta_t^e - \delta_e \le 0$ under H_t^e; independent sampling with replacement from a frozen empirical pool is the corresponding empirical-distribution case. With dependent observations, the bound must instead hold for the actual update unit, such as a block or cluster, given the full joint history. Merely aggregating correlated rows does not establish it. Section 5.3 tests failures caused by invalid evidence units and post-evidence candidate selection.

Let $C_e > 0$ be a predeclared bound satisfying $|X^e_{t,i} - \delta_e| \le C_e$ for every admissible environment-e evidence unit, and set $Y^e_{t,i} = (X^e_{t,i} - \delta_e)/C_e \in [-1, 1]$. For a fixed finite grid $0 \le \lambda_k < 1$, define (empty products equal one)

$$E^e_{t,m} = \frac{1}{K}\sum_{k=1}^{K}\prod_{i=1}^{m}(1 + \lambda_k Y^e_{t,i}). \tag{6}$$

Proposition 1 (Component e-process)**.** *For an environment sampled at every step, Assumption 1 makes each betting product in Eq.* (6) *a nonnegative supermartingale on its true-null event. Their mixture starts at one. At any almost-surely finite stopping time* τ*, it satisfies* $\mathbb{E}[\nu^e_t E^e_{t,\tau} \mid \mathcal{F}_{t-1}] \le \nu^e_t$.

Theorem 1 extends this result to interleaved sampling. Other conditionally valid component e-processes can replace the bounded mixture.

Lemma 1 (Persistent candidate e-value)**.** *Suppose each nonnegative stopped value* E^e_t *satisfies* $\mathbb{E}[\nu^e_t E^e_t \mid \mathcal{F}_{t-1}] \le \nu^e_t$*. Then*

$$E^{\mathrm{pers}}_t = \min_{e \in \mathcal{E}} E^e_t \tag{7}$$

satisfies $\mathbb{E}[\nu_t E^{\mathrm{pers}}_t \mid \mathcal{F}_{t-1}] \le \nu_t$*. No independence across environments is required.*

Under the persistent null, at least one component null is true, and the minimum cannot exceed that component. Appendix A makes this argument explicit when the true component depends on history.

Candidate freezing and evidence reuse. Let E_1, E_2 be independent, each equal to 2 with probability $1/2$ and 0 otherwise. Each has mean one, but selecting the larger value gives $\mathbb{E}[\max(E_1, E_2)] = 3/2 > 1$. A challenger selected using decision evidence therefore needs a fresh test conditional on that selection history. Restarting means discarding the selection evidence from the new test and advancing the candidate index; it does not reset the online error budget or erase earlier tests.

4.2 Predictable routing and evidence allocation

Write $\widetilde{E}^e_{t,n} = E^e_{t,N^e_{t,n}}$ on the global routing clock.

Theorem 1 (Validity under predictable routing)**.** *If* $A_{t,n}$ *is* $\mathcal{G}_{t,n-1}$*-measurable and Assumption 1 holds, each routed betting product, and its mixture* $\widetilde{E}^e_{t,n}$*, is a nonnegative supermartingale on* $\{\nu^e_t = 1\}$*. At any almost-surely finite joint-history stopping time* κ^e_t*, the value* $E^e_t = \widetilde{E}^e_{t,\kappa^e_t}$ *satisfies the bound in Lemma 1.*

Let τ_t be the finite time at which the candidate decision ends. A component is frozen when it first crosses the preassigned threshold q_t, or at τ_t if it has not crossed. Unopened components retain value one. We combine these stopped values, which need not cross simultaneously. Taking the largest value observed retrospectively would not in general give an e-value. Appendix A.5 specifies the stopping rule and its validity.

Stagewise EGR tests environments in a fixed order, proceeding only after the current component crosses q_t within its cap B; a failed stage ends the comparison without adoption. **EGR-ALLOC**

Table 2: Operational conditions used by the guarantee. The statement is local to the declared target and evidence unit.

Condition	Required?	Role
Candidate frozen before evidence	Yes	Prevents candidate-time leakage
Joint-filtration validity	Yes	Links null to selected updates
Predictable allocation; a.s.-finite stopping	Yes	Preserves e-process validity
Cross-environment independence	No	Minimum composition is pointwise
Raw-unit dependence	Not automatic	Must be handled in the evidence unit

recycles earlier savings: with total cap $|\mathcal{E}|B$, spending through stage j (zero indexed) cannot exceed $(j+1)B$, reserving B for each unopened environment. **EGR-Balance** samples the unresolved environment with the smallest current log evidence. All three freeze completed components.

As a deterministic reference, start all counts at zero and suppose log evidence grows as $\ell_e(n_e) = n_e r_e$, $r_e > 0$, and define $S = \sum_e r_e^{-1}$ and $r_{\max} = \max_e r_e$.

Proposition 2 (First-order balancing guarantee)**.** *If each observation is sent to a current minimizer of $n_e r_e$, then after N allocations,*

$$\frac{N}{S} - r_{\max} \le \min_e n_e r_e \le \frac{N}{S}. \tag{8}$$

For $L \ge 0$, any rule reaching level L in every component needs at least SL observations; greedy balancing needs at most $\lceil S(L + r_{\max})\rceil$. This is a deterministic reference, not a stochastic optimality guarantee.

4.3 Across-candidate control

Fix $\alpha \in (0,1)$ and a deterministic nonnegative schedule with $\sum_{t\ge1} \gamma_t \le 1$. The e-LOND rule (Xu and Ramdas, 2024) sets

$$\alpha_t = \alpha\gamma_t(R_{t-1}+1), \qquad q_t = \alpha_t^{-1}, \tag{9}$$

and permits adoption only when $E_t^{\mathrm{pers}} \ge q_t$. If $\gamma_t = 0$, set $q_t = \infty$ and do not adopt. The experiments normalize $\gamma_t \propto [t\log^2(t+1)]^{-1}$ over a predeclared horizon; the schedule is not reset after failed tests or candidate changes.

Theorem 2 (Persistent adoption control)**.** *Suppose the declared comparison is fixed given $\mathcal{F}_{t-1}$, Assumption 1 holds, routing is predictable, and component stopping times are almost surely finite in the joint history. Combine the stopped values by Eq.* (7)*. Then $\mathbb{E}[\nu_t E_t^{\mathrm{pers}} \mid \mathcal{F}_{t-1}] \le \nu_t$. With Eq.* (9)*, $\mathrm{FDR}(T) \le \alpha$ for every deterministic horizon T, even when earlier adoptions change later hypotheses. Using $q_t = 1/(\alpha\gamma_t)$ instead gives $\mathrm{FWER}(T) \le \alpha$.*

The proof bounds each false adoption's contribution to the final discovery proportion before taking expectations. The next frozen history includes the completed test and any new scout data. Only adoption replaces the incumbent; declining a challenger leaves it unchanged. Changes made after examining decision evidence require a new candidate index and fresh conditionally valid evidence. Alternative online controllers require their own guarantees.

5 Experiments

The experiments separate errors in the adoption target, evidence construction, and repeated testing from the cost of acquiring valid evidence. We call a method *audited* when its criterion covers every declared environment. Reported rates are Monte Carlo estimates, not substitutes for Theorem 2. "0 obs." means empirical FDR is zero because no trajectory contains a false adoption. We also report a one-sided 95% Clopper–Pearson upper bound $u_N = 1 - 0.05^{1/N}$ for the path event $\mathbb{P}(V_T > 0)$; since $V_T/(R_T \vee 1) \le \mathbf{1}\{V_T > 0\}$, this is a conservative upper bound on FDR. Here $u_{5000} = 5.99 \times 10^{-4}$, $u_{800} = .00374$, and $u_{500} = .00597$. Appendices B–C detail the protocols and uncertainty.

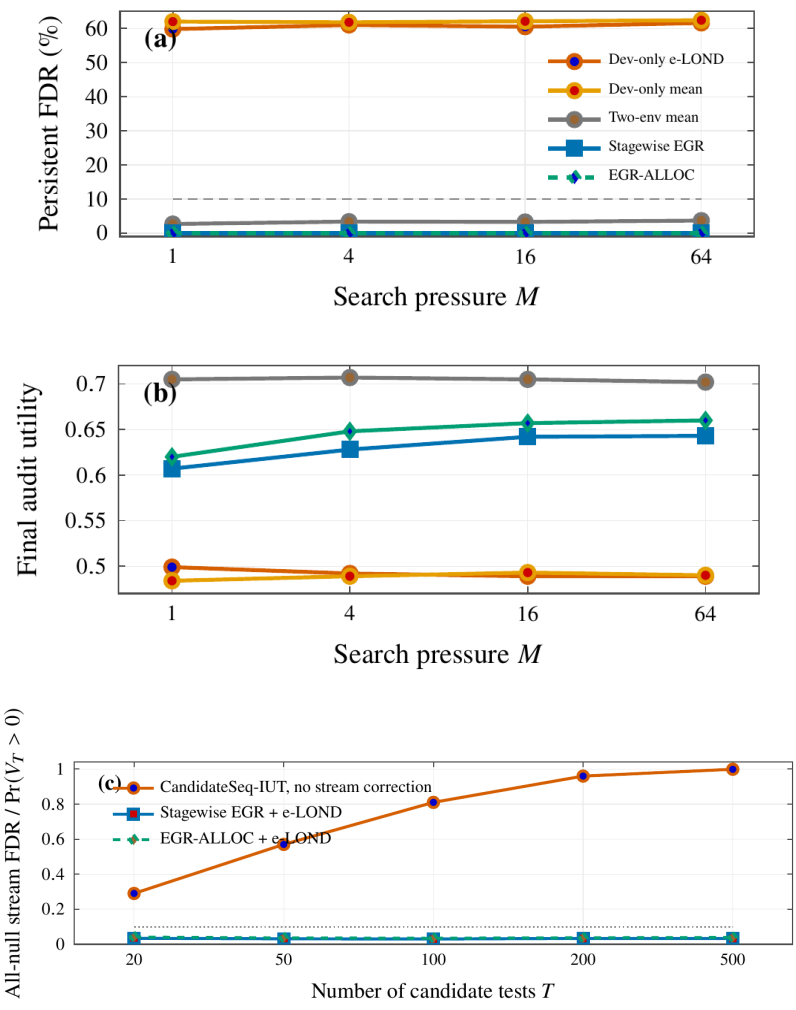


Figure 2: Target mismatch and stream multiplicity. (a) Persistent FDR; (b) final audit utility. (c) All-null stream error versus candidate tests T; here every rejection is false, so FDR $= \mathbb{P}(V_T > 0)$. Candidate-level reuse without stream correction accumulates error, whereas EGR with online correction remains controlled in the tested regimes.

5.1 Target mismatch and repeated adoption

In the synthetic search, one coordinate improves utility in both environments; the other improves development utility at the audit environment's expense. Each of $T = 20$ steps scouts M incumbent mutations, and an accepted challenger becomes the next parent. At $M = 16$, with 5,000 trajectories, $\delta = 0.01$, and cap $B = 120$ per environment, development-only e-LOND has persistent FDR 0.621 despite testing its own development target validly. In a separate all-null stream, repeatedly applying a two-environment anytime test without stream correction gives both FDR and $\mathbb{P}(V_T > 0)$ of 0.999 by $T = 500$. Correcting the target alone therefore does not control repeated adoption.

EGR-ALLOC reaches persistent-alternative power .251 at 153.7 evidence units per candidate, versus .188 at 240.0 for fixed audited IUT. These policies also produce different future candidates through their adoption histories, so this comparison does not isolate acquisition efficiency. Matched replay in Section 5.5 removes that difference. Neither EGR variant has an observed persistent false-adoption path among the 5,000 trajectories; the one-sided path-event upper bound is 5.99×10^{-4}.

Table 3: Headline closed-loop results at $M = 16$. Evidence is the mean paired decision updates per candidate; "0 obs." uses the finite-run convention defined at the start of Section 5. Baselines: IUT (Berger and Hsu, 1996); e-LOND (Xu and Ramdas, 2024).

Method	Persistent FDR	Pers.-alt. power	Audit utility	Evidence / candidate
Dev mean	.6240	.907	.493	240.0
Dev e-LOND	.6210	.461	.489	221.6
Two-env mean	.0340	.867	.705	240.0
Fixed IUT + e-LOND	0 obs.	.188	.641	240.0
Stagewise EGR	0 obs.	.196	.642	145.5
EGR-ALLOC	0 obs.	.251	.657	153.7

5.2 Stream-level controller sensitivity

We hold candidate evidence fixed and feed identical persistent e-value streams to four controllers at $T = 500$ with 5% alternatives. Persistent-alternative power is .619 for e-LOND, .890 for e-LORD, .888 for e-SAFFRON, and .415 for the e-FWER rule; the corresponding empirical FDRs are .00015, .00055, .00055, and .00020. The power differences in this comparison come from the downstream spending rules. Online e-closure/compound-e procedures (Xu et al., 2026) were not implemented in this diagnostic, so their power remains to be compared. These experiments do not extend Theorem 2 beyond e-LOND; each alternative controller requires its own validity conditions.

5.3 Evidence validity, candidate freezing, and adaptive allocation

At within-cluster correlation $\rho = .95$, treating dependent raw rows as fresh sequential units gives a candidate-null crossing probability of .728. Aggregating each cluster into one bounded evidence unit reduces it to .0127. This difference illustrates why the marginal mean condition alone is insufficient: the unit used for each update must satisfy the joint-history condition in Assumption 1.

Candidate-time leakage occurs when a challenger is chosen using its own decision evidence. To separate this effect from optional stopping, we construct development-favorable challengers exactly on the audit boundary and evaluate their component e-values at a fixed endpoint. Every variant is a true persistent null, so a valid e-value must satisfy $\mathbb{P}(E \geq h) \leq 1/h$ by Markov's inequality. With $L = 64$ post-evidence variants and $h = 10$, the crossing rates over 100,000 trials are .352% for Frozen, 20.089% for Leak+reuse, and .335% for Leak+restart. Only Leak+reuse exceeds the $1/h = 10\%$ reference bound. At $h = 20$, the rates in the same order are .052%, 3.802%, and .063%. Reusing the evidence that selected the challenger distorts calibration; a fresh restart brings the observed rate back close to the frozen baseline. Appendices B–C give the protocol and threshold sweeps.

Across $|\mathcal{E}| \in \{2, 3, 5\}$ and $h \in \{5, 10, 20, 50, 100\}$, EGR-Balance uses observed evidence to choose each sample. Its mean evidence cost divided by that of an oracle with known simulator evidence rates ranges from .970 to 1.034. This is an empirical comparison; Proposition 2 concerns deterministic progress only.

5.4 Multi-environment and larger neural searches

Public-data searches use a fixed, stratified scikit-learn digits snapshot, local incumbent-based proposals, and independent draws with replacement from empirical decision pools. The three-environment extension declares clean development, Gaussian-noise audit, and dropout+noise audit; a stronger corruption is reserved for hidden evaluation. Error control is relative to the fixed empirical distributions, not an unknown deployment population.

A second study uses a generated classification problem with 40,000 samples, 50 features, five classes, and a 16-model MLP bank. It tests the procedure at a larger controlled scale; neither study establishes performance on prospective image data, language models, or fully agentic systems.

Relative to fixed IUT, Stagewise EGR reduces mean decision evidence by 67.6% in the three-environment digits study and 62.9% in the 40k benchmark. EGR-ALLOC recycles budget saved at earlier stages while keeping the same total cap; its observed powers are .3349 and .2666, respectively.

Table 4: Three-environment audited comparisons. "0 obs." uses the common finite-run convention in Section 5; Monte Carlo intervals are reported in Appendix C. Baselines: IUT (Berger and Hsu, 1996); e-LOND (Xu and Ramdas, 2024).

Benchmark	Method	FDR	Power	Evidence
Digits 3-env.	Fixed IUT + e-LOND	0 obs.	.3137	1200.0
($N = 800$)	Stagewise EGR	0 obs.	.3221	388.8
	EGR-ALLOC	0 obs.	.3349	392.6
Controlled 40k	Fixed IUT + e-LOND	0 obs.	.2149	1200.0
($N = 500$)	Stagewise EGR	0 obs.	.2380	444.6
	EGR-ALLOC	0 obs.	.2666	440.9

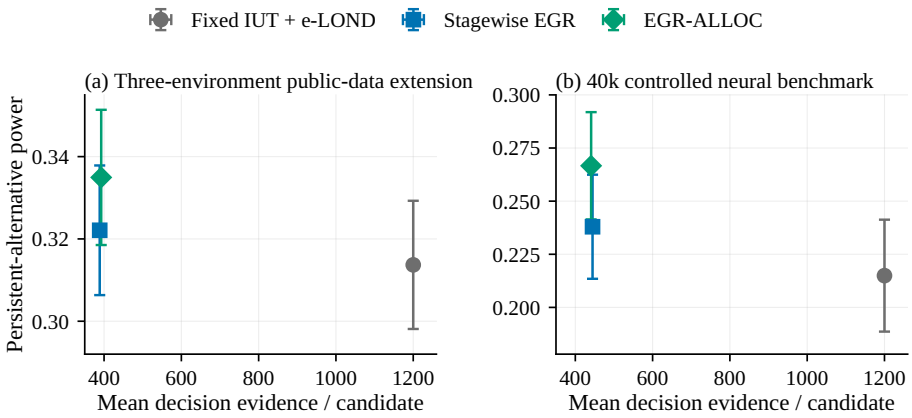


Figure 3: Persistent-alternative power versus decision-evidence cost in the three-environment studies. Error bars show 95% Monte Carlo intervals; exact numerical summaries are in Table 4.

None of the three audited methods has an observed persistent false-adoption path in these finite runs. Appendix C reports sensitivity to the proposal policy.

5.5 Matched replay isolates acquisition efficiency

We freeze all 600 ordered non-self pairs from the 25-model neural grid and replay identical paired evidence streams. Figure 4 shows the full sweep $h \in \{5, 10, 20, 50, 100\}$; $h = 20$ is a representative interior point. Fixed-anytime IUT and Stagewise EGR adopt on the same event at every h: every component crosses the threshold within its own cap (Appendix A.8). At $h = 20$ their persistent-alternative power is therefore identical (.6477), while Stagewise uses 702.4 rather than 1600.0 observations per candidate, a 56.1% saving. The paired saving is 897.6 observations (95% pair-bootstrap CI 878.1–917.3). EGR-ALLOC reaches power .6699 at 715.7 observations, increasing power over Stagewise by .0222 (95% CI .0159–.0290).

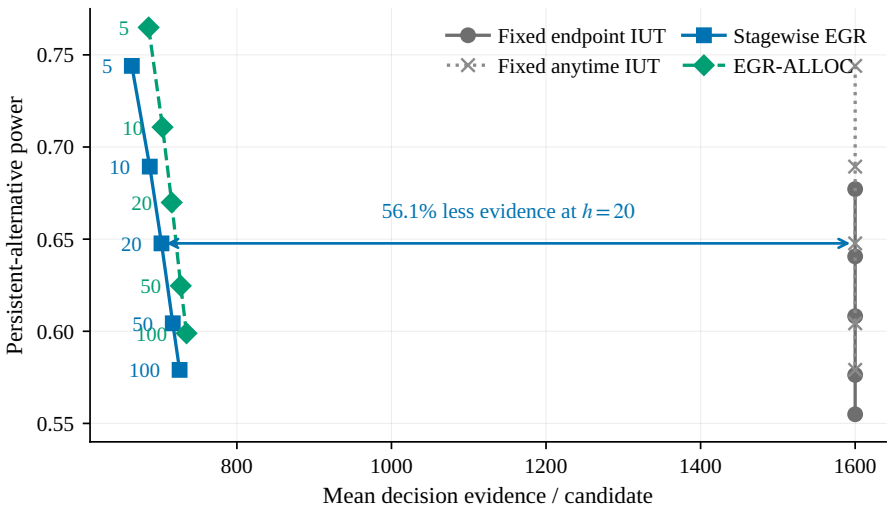


Figure 4: Matched frozen-candidate replay over 600 challenger–incumbent pairs. Stagewise EGR has the same alternative crossing event as fixed-anytime IUT while placing evidence sequentially; EGR-ALLOC can recycle saved early-stage budget. At $h = 20$, Stagewise uses 56.1% less evidence at identical power.

6 Discussion and Limitations

EGR's guarantee depends on the comparison being fixed before its decision evidence is examined and on the conditional mean bound holding for each sampled unit. Neither a favorable development result nor a valid isolated test is sufficient for repeated all-environment adoption. The leakage and clustered-data diagnostics show two ways in which the evidence itself can fail these conditions.

The experiments also distinguish search quality from test efficiency. Different adoption rules produce different incumbents and future proposals, so closed-loop power is policy-dependent rather than a candidate-level efficiency comparison. Matched replay removes that confounding: fixed-anytime IUT and Stagewise have identical adoption events on identical streams, and the measured saving comes from component stopping and short-circuiting after an earlier stage fails.

EGR is therefore a validity framework for state-changing model replacement, not a new online multiple-testing controller. Its guarantees use existing e-value machinery once candidate-level evidence has been constructed correctly; the distinctive statistical object is a challenger–incumbent comparison whose challenger, incumbent, and null may all depend on previous adoption decisions. Stagewise savings partly reflect short-circuiting an intersection–union decision, while EGR-ALLOC and EGR-Balance study how remaining evidence can be redistributed beyond that basic saving. The empirical guarantees concern fixed decision pools or controlled distributions. Prospective deployment data, fully agentic proposal loops, and training-time overhead remain to be evaluated; any later test still requires evidence satisfying the conditional bound given the enlarged history.

7 Conclusion

EGR tests model replacement against a declared all-environment target while accounting for repeated candidates. Predictable sampling and valid component stopping supply conditional e-values; online control then bounds false adoption across the changing-incumbent search. The experiments show evidence savings under these conditions and identify failures when the target, evidence unit, or selection history is handled incorrectly.

Ethics Statement. This work uses synthetic, controlled, and standard public benchmark data and does not involve human subjects or sensitive personal data. The statistical guarantee applies only to the predeclared environments, utilities, margins, evidence units, and sampling conditions. If those choices omit deployment-relevant harms or if audit feedback leaks into candidate construction, the procedure can give assurance about an incomplete target. Deployment use therefore requires domain-specific evaluation and human oversight in addition to the statistical gate.

Reproducibility Statement. Appendices A–C give complete proofs, experimental protocols, and uncertainty diagnostics. The supplement includes runnable code, source arrays/data, manifests, paper-facing outputs, and experiment-level checksums for E27/E29 and E30–E33. The submission package additionally records the headline closed-loop protocol parameters available in the manuscript and SHA-256 hashes for its paper-facing artifacts, together with a checksum of the complete supplemental archive.

AI Use Statement. Generative AI tools were used to refine hypotheses and exposition, assist with proof and code drafting, suggest and critique experimental designs, support literature search, and prepare or edit figures and manuscript text. The authors checked the mathematical claims, citations, reported numerical outputs, the experiment code and artifacts included with the submission, and final figures, and take responsibility for the submitted content.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998.

Eser Aygün, Anastasiya Belyaeva, Gheorghe Comanici, et al. An AI system to help scientists write expert-level empirical software. *Nature*, 654:909–916, 2026. doi: 10.1038/s41586-026-10658-6.

Mitra Baratchi, Can Wang, Steffen Limmer, Jan N. van Rijn, Holger H. Hoos, Thomas Bäck, and Markus Olhofer. Automated machine learning: Past, present and future. *Artificial Intelligence Review*, 57:122, 2024. doi: 10.1007/s10462-024-10726-1.

Roger L. Berger and Jason C. Hsu. Bioequivalence trials, intersection–union tests and equivalence confidence sets. *Statistical Science*, 11(4):283–302, 1996. doi: 10.1214/ss/1032280304.

Jun Shern Chan, Neil Chowdhury, Oliver Jaffe, James Aung, Dane Sherburn, Evan Mays, Giulio Starace, Kevin Liu, Leon Maksin, Tejal Patwardhan, Aleksander Madry, and Lilian Weng. MLE-Bench: Evaluating machine learning agents on machine learning engineering. In *International Conference on Learning Representations*, 2025.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. The reusable holdout: Preserving validity in adaptive data analysis. *Science*, 349(6248):636–638, 2015. doi: 10.1126/science.aaa9375.

Ali E. Ghareeb, Benjamin Chang, Ludovico Mitchener, et al. A multi-agent system for automating scientific discovery. *Nature*, 655:497–505, 2026. doi: 10.1038/s41586-026-10652-y.

Elise Han, Chengpiao Huang, and Kaizheng Wang. Model assessment and selection under temporal distribution shift. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 17374–17392, 2024.

Ramesh Johari, Pete Koomen, Leonid Pekelis, and David Walsh. Always valid inference: Continuous monitoring of A/B tests. *Operations Research*, 70(3):1806–1821, 2022. doi: 10.1287/opre.2021.2135.

Chris Lu, Cong Lu, Robert Tjarko Lange, Yutaro Yamada, Shengran Hu, Jakob Foerster, David Ha, and Jeff Clune. Towards end-to-end automation of AI research. *Nature*, 651:914–919, 2026. doi: 10.1038/s41586-026-10265-5.

Teodora Pandeva, Patrick Forré, Aaditya Ramdas, and Shubhanshu Shekhar. Deep anytime-valid hypothesis testing. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 622–630, 2024.

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, Peter Grünwald, Vladimir Vovk, and Glenn Shafer. Game-theoretic statistics and safe anytime-valid inference. *Statistical Science*, 38(4):576–601, 2023. doi: 10.1214/23-STS894.

Edan Toledo, Karen Hambardzumyan, Martin Josifoski, Rishi Hazra, Nicolas Baldwin, Alexis Audran-Reiss, Michael Kuchnik, Despoina Magka, Minqi Jiang, Alisia Lupidi, et al. AI research agents for machine learning: Search, exploration, and generalization in MLE-bench. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-1186.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 86(1):1–27, 2024. doi: 10.1093/jrsssb/qkad009.

Ziyu Xu and Aaditya Ramdas. Online multiple testing with e-values. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pages 3997–4005, 2024.

Ziyu Xu, Lasse Fischer, and Aaditya Ramdas. Improving online FDR procedures via online analogs of e-closure and compound e-values. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pages 7552–7561, 2026.

Matteo Zecchin, Sangwoo Park, and Osvaldo Simeone. Adaptive learn-then-test: Statistically valid and efficient hyperparameter selection. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 74018–74036, 2025.

Yifan Zhang, Zijian Wei, Haojie Ren, and Changliang Zou. e-GAI: e-value-based generalized α-investing for online false discovery rate control. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pages 76308–76331, 2025.

A Proofs

A.1 Histories, null events, and stopping conventions

We work under a fixed data-generating law. The declared reference kernel P_t^e, challenger, incumbent, utilities, and margins are measurable with respect to $\mathcal{F}_{t-1}$. Consequently, Δ_t^e, ν_t^e, and ν_t are also $\mathcal{F}_{t-1}$-measurable. This is a property of the statistical target under the fixed law, not a claim that the planner can evaluate its truth. Proposal randomization and scout observations belong to this history before decision sampling begins. At the next candidate, $\mathcal{F}_t$ contains the completed test and the data used to construct the next challenger.

The filtration $\mathcal{G}_{t,n}$ records all environments jointly, with $\mathcal{G}_{t,0} = \mathcal{F}_{t-1}$. A stopping time in this appendix is relative to that filtration, unless a single-environment clock is explicitly specified. After termination, extend each process by holding it constant. The environment set is finite and nonempty, the betting grid is finite and fixed before sampling, and every stopped value used in a decision is nonnegative. Bounds involving a null event are written with its indicator, so that no restriction is imposed on alternative histories. In particular, the precise meaning of conditional component validity is

$$\mathbb{E}[\nu_t^e E_t^e \mid \mathcal{F}_{t-1}] \leq \nu_t^e. \tag{10}$$

This formulation also avoids treating the set of true nulls as a deterministic subset of candidate indices.

A.2 Proof of Proposition 1

Fix t and e and sample this environment at every step. Abbreviate $\nu = \nu_t^e$ and define

$$L_m^k = \prod_{i=1}^{m}(1 + \lambda_k Y_{t,i}^e), \qquad L_0^k = 1.$$

Since $|Y_{t,i}^e| \le 1$ and $0 \le \lambda_k < 1$, each factor lies in $[1 - \lambda_k, 1 + \lambda_k]$. Thus L_m^k is nonnegative, adapted, and bounded above by 2^m at a fixed m, which ensures integrability. The variable ν is measurable at time zero. Assumption 1 gives

$$\begin{aligned}\mathbb{E}[\nu L_m^k \mid \mathcal{G}_{t,m-1}] &= \nu L_{m-1}^k + \lambda_k L_{m-1}^k \mathbb{E}[\nu Y_{t,m}^e \mid \mathcal{G}_{t,m-1}] \\ &\le \nu L_{m-1}^k.\end{aligned}$$

Hence $(\nu L_m^k)_m$ is a nonnegative supermartingale, starting at ν. Averaging over the fixed betting grid shows that $M_m = \nu E_{t,m}^e$ is also a nonnegative supermartingale with $M_0 = \nu$.

Let τ be an almost-surely finite stopping time. For each deterministic integer b, bounded optional sampling yields

$$\mathbb{E}[M_{\tau \wedge b} \mid \mathcal{F}_{t-1}] \le M_0 = \nu.$$

As $b \to \infty$, $M_{\tau \wedge b} \to M_\tau$ almost surely. Conditional Fatou's lemma, using nonnegativity, gives

$$\mathbb{E}[M_\tau \mid \mathcal{F}_{t-1}] \le \liminf_{b \to \infty} \mathbb{E}[M_{\tau \wedge b} \mid \mathcal{F}_{t-1}] \le \nu.$$

No uniform-integrability assertion or finite expected stopping time is needed for this inequality. A finite observation cap is a sufficient implementation-level condition for almost-sure finiteness. Interleaved sampling requires the joint-clock argument below, rather than an assumption that every local update index is eventually reached. □

A.3 Proof of Lemma 1

Order the environments as $e_1, \ldots, e_d$ before sampling. Define

$$B_j = \{\nu_t^{e_j} = 1\} \cap \bigcap_{i<j} \{\nu_t^{e_i} = 0\}, \qquad j = 1, \ldots, d.$$

The events B_j are $\mathcal{F}_{t-1}$-measurable, disjoint, and partition $\{\nu_t = 1\}$. They select the first true component without looking at decision evidence. Pointwise,

$$\nu_t E_t^{\text{pers}} \le \sum_{j=1}^{d} \mathbf{1}_{B_j} E_t^{e_j}.$$

Because $\mathbf{1}_{B_j} \le \nu_t^{e_j}$ and $\mathbf{1}_{B_j}$ is measurable given the frozen history,

$$\begin{aligned}\mathbb{E}[\nu_t E_t^{\text{pers}} \mid \mathcal{F}_{t-1}] &\le \sum_{j=1}^{d} \mathbf{1}_{B_j} \mathbb{E}[\nu_t^{e_j} E_t^{e_j} \mid \mathcal{F}_{t-1}] \\ &\le \sum_{j=1}^{d} \mathbf{1}_{B_j} = \nu_t.\end{aligned}$$

On alternative histories the left side is zero. The argument needs neither independent environments nor simultaneous component stopping. It uses only the component bounds and the pointwise minimum.□

A.4 Proof of Theorem 1

Fix t, e, k. Define an increment that is meaningful even when environment e has never been sampled:

$$W_{t,n}^e = \begin{cases} (X_{t,N_{t,n}^e}^e - \delta_e)/C_e, & A_{t,n} = e, \\ 0, & A_{t,n} \ne e. \end{cases}$$

After candidate termination it is also zero. Since the choice of environment is known before its next outcome is observed, Assumption 1 implies

$$\mathbb{E}[\nu_t^e W_{t,n}^e \mid \mathcal{G}_{t,n-1}] \le 0.$$

Let $\widetilde{L}_n^k = \prod_{s\le n}(1+\lambda_k W_{t,s}^e)$. Zero increments contribute a factor of one, so

$$\widetilde{L}_n^k = \prod_{i=1}^{N_{t,n}^e}(1+\lambda_k Y_{t,i}^e), \qquad \widetilde{L}_n^k = \widetilde{L}_{n-1}^k(1+\lambda_k W_{t,n}^e).$$

The product is adapted, nonnegative, and bounded by 2^n at fixed n. Therefore

$$\begin{aligned}\mathbb{E}[\nu_t^e \widetilde{L}_n^k \mid \mathcal{G}_{t,n-1}] &= \nu_t^e \widetilde{L}_{n-1}^k + \lambda_k \widetilde{L}_{n-1}^k \mathbb{E}[\nu_t^e W_{t,n}^e \mid \mathcal{G}_{t,n-1}] \\ &\le \nu_t^e \widetilde{L}_{n-1}^k.\end{aligned}$$

Averaging over k proves the same inequality for $\nu_t^e \widetilde{E}_{t,n}^e$. Bounded optional sampling at $\kappa_t^e \wedge b$, followed by conditional Fatou as in Proposition 1, proves Eq. (10). Different environments may use different stopping times, provided each is a stopping time for the same joint filtration. Lemma 1 then applies. □

A.5 Threshold crossings and frozen component values

The candidate threshold q_t is known at $\mathcal{F}_{t-1}$. Let τ_t be the almost-surely finite global decision time and set

$$T_t^e = \inf\{n \ge 0 : \widetilde{E}_{t,n}^e \ge q_t\}, \qquad \kappa_t^e = T_t^e \wedge \tau_t,$$

with $\inf \varnothing = \infty$. An adapted process crossing a threshold known at time zero defines a stopping time; hence κ_t^e is a finite joint-history stopping time. The terminal value used in the candidate minimum is

$$E_t^e = \widetilde{E}_{t,\kappa_t^e}^e.$$

The implementation freezes a component at its first crossing and routes only to unresolved components. This is equivalent to holding its stopped process constant. A component that is never opened stays at its initial value one. A component that does not cross contributes its value when the candidate is declined or its budget expires. Thus, for a finite threshold, all components having crossed by τ_t is equivalent to the minimum of their frozen terminal values exceeding or equaling q_t. This includes crossings at different times.

The retained value must not be replaced by the largest value observed after the fact. For example, a nonnegative martingale with $M_0 = 1$ and $M_1 \in \{0, 2\}$ equiprobably has $\mathbb{E}[M_1] = 1$, but $\mathbb{E}[\max(M_0, M_1)] = 3/2$. A retrospective maximum can therefore fail the e-value mean bound. A crossing event can still define a valid sequential test; the proof above obtains the corresponding e-value by stopping at the crossing, or at the cap if there is none. No monotonicity of the unstopped evidence path is assumed.

A.6 Proof of Proposition 2

Let $d = |\mathcal{E}|$, start with $n_e(0) = 0$, and define the final progress after N allocations by $p_e = n_e(N)r_e$ and $m_N = \min_e p_e$. Since $r_e > 0$,

$$N = \sum_e \frac{p_e}{r_e} \ge m_N \sum_e \frac{1}{r_e} = m_N S,$$

which gives the upper bound $m_N \le N/S$ for any allocation.

For the greedy rule, consider the last update assigned to a sampled environment e. Immediately before that update, its progress equals the then-current minimum, which is no greater than m_N because all coordinates are nondecreasing. Its last update adds exactly r_e, so $p_e \le m_N + r_e$. An unsampled environment has $p_e = 0$ and satisfies the same bound. Ties between current minimizers do not affect this argument. Consequently,

$$N = \sum_e \frac{p_e}{r_e} \le m_N S + d, \qquad m_N \ge \frac{N-d}{S} \ge \frac{N}{S} - r_{\max},$$

where the last step uses $d \le Sr_{\max}$. For a target $L \ge 0$, reaching every component requires $N \ge SL$. Conversely, choosing $N = \lceil S(L + r_{\max}) \rceil$ makes the derived lower bound at least L. This proves both claims. Before every component reaches L, a current minimizer is unresolved, so excluding already completed components does not change the guarantee up to that time. Random log-evidence paths need not be monotone or have constant positive rates; the proposition makes no claim for those paths. □

A.7 Proof of Theorem 2

Theorem 1 gives conditional validity of each stopped component. Lemma 1 therefore yields

$$\mathbb{E}[\nu_t E_t^{\mathrm{pers}} \mid \mathcal{F}_{t-1}] \le \nu_t, \qquad \mathbb{E}[\nu_t E_t^{\mathrm{pers}}] \le \mathbb{E}[\nu_t] \le 1. \tag{11}$$

The null indicator is retained inside the expectation throughout the argument. In particular, the identity of true nulls need not be fixed before the search begins.

FDR. Write $E_t = E_t^{\mathrm{pers}}$. Fix a deterministic horizon T. If $D_t = 1$, then $R_T \ge R_{t-1} + 1$ and the adoption rule implies $1 \le \alpha\gamma_t(R_{t-1} + 1)E_t$. If $D_t = 0$, its contribution is zero. Thus, pathwise,

$$\frac{\nu_t D_t}{R_T \vee 1} \le \frac{\nu_t D_t}{R_{t-1} + 1} \le \alpha\gamma_t \nu_t E_t.$$

This also holds for $\gamma_t = 0$, since such an index cannot produce an adoption. Summing and applying Eq. (11),

$$\begin{aligned} \mathrm{FDR}(T) &= \mathbb{E}\left[\sum_{t=1}^{T} \frac{\nu_t D_t}{R_T \vee 1}\right] \\ &\le \alpha \sum_{t=1}^{T} \gamma_t \mathbb{E}[\mathbb{E}[\nu_t E_t \mid \mathcal{F}_{t-1}]] \\ &\le \alpha \sum_{t=1}^{T} \gamma_t \mathbb{E}[\nu_t] \le \alpha \sum_{t=1}^{T} \gamma_t \le \alpha. \end{aligned}$$

This is the e-LOND counting argument (Xu and Ramdas, 2024), with the conditional construction supplying valid null-weighted evidence for the adaptive hypotheses. It uses no independence between candidates, and it does not assume that $R_T \vee 1 \ge R_{t-1} + 1$ on paths where $D_t = 0$.

FWER. Now use the separate rule with threshold $1/(\alpha\gamma_t)$, and let D_t, R_t, V_t refer to the history generated by that rule. For $\gamma_t > 0$, conditional Markov's inequality, restricted to the true-null event, gives

$$\begin{aligned} \mathbb{E}[\nu_t D_t \mid \mathcal{F}_{t-1}] &\le \alpha\gamma_t \mathbb{E}[\nu_t E_t \mid \mathcal{F}_{t-1}] \\ &\le \alpha\gamma_t \nu_t. \end{aligned}$$

For $\gamma_t = 0$ the left side is zero by convention. Taking expectations and applying the union bound yields

$$\mathbb{P}(V_T \ge 1) \le \sum_{t=1}^{T} \mathbb{P}(\nu_t D_t = 1) \le \alpha \sum_{t=1}^{T} \gamma_t \mathbb{E}[\nu_t] \le \alpha.$$

The FWER rule can generate a different search path from e-LOND; its guarantee is proved along its own path under the same conditional assumptions. □

The deterministic spending schedule is fixed for the whole declared search. A candidate modified after seeing decision evidence must be assigned a new index, with the earlier index retained in the history even if no adoption occurred. Resetting the schedule after unsuccessful attempts is not covered. For a predeclared finite horizon, the normalized weights may be extended by zero beyond that horizon; separately renormalized runs do not automatically share one overall error budget.

A.8 Why matched Stagewise and fixed-anytime IUT have equal power

The replay claim is a pathwise property of its coupling, separate from the FDR theorem. Fix one candidate pair, a threshold $h > 1$, a cap B for each of d environments, and a complete potential evidence sequence for each environment. Require these same sequences to be used under both acquisition rules; opening an environment later does not change its sequence. Define

$$a_e = \inf\{m \in \{1, \dots, B\} : E_m^e \geq h\}, \qquad \ell_e = \min(a_e, B),$$

with $a_e = \infty$ when there is no crossing. The fixed-anytime comparator collects all dB observations and adopts precisely on

$$\mathcal{A} = \bigcap_{e=1}^{d} \{a_e \leq B\}.$$

Stagewise opens environment j only when all earlier environments have crossed. If $\mathcal{A}$ occurs, every stage opens and crosses, so Stagewise adopts. If $\mathcal{A}$ fails, the first noncrossing stage uses its cap and terminates without adoption. Both rules therefore have the same adoption indicator on every coupled path, under nulls and alternatives alike.

Their evidence costs need not match. For Stagewise the exact cost is

$$C_{\text{stage}} = \sum_{j=1}^{d} \ell_j \prod_{i<j} \mathbf{1}\{a_i \leq B\} \leq dB = C_{\text{fixed}}.$$

Savings arise from stopping completed stages at their first crossing and from not opening later stages after a failure. This identity proves equal power for the matched replay, not for separate closed-loop runs whose candidates and thresholds can differ. It also does not apply to fixed-*endpoint* IUT, which tests $E_B^e \geq h$ and can miss an earlier crossing. EGR-ALLOC changes later-stage caps by recycling savings, so its adoption event can differ from both matched-cap rules.

B Detailed Experimental Protocol

B.1 Common conventions and synthetic closed-loop setup

Empirical FDR is estimated by computing $\text{FDP} = V_T/(R_T \vee 1)$ for each independent trajectory and averaging. "Any false" is $\mathbb{P}(V_T > 0)$. Let $S_T = \sum_{t \leq T}(1 - \nu_t)$ be the number of proposed persistent alternatives and $A_T^{\text{true}} = \sum_{t \leq T}(1 - \nu_t)D_t$ the number adopted. Reported closed-loop persistent-alternative power is the Monte Carlo mean of $A_T^{\text{true}}/(S_T \vee 1)$. Because S_T depends on the realized proposal path, this is a policy-level search metric rather than a candidate-level power comparison across methods; Section 5.5 supplies the latter on fixed pairs. Evidence counts paired decision-evidence updates across declared environments; scout data used only for candidate proposal are excluded. Component e-processes use the precommitted betting grid $\Lambda = \{.03, .06, .10, .18, .30, .45, .65, .82\}$. For horizon T, e-LOND uses $\gamma_t \propto [t \log^2(t+1)]^{-1}$ normalized over $t = 1, \dots, T$; this schedule is fixed before simulation and is not selected by outcome. Monte Carlo standard errors are the across-replicate sample standard deviation divided by the square root of the number of nonmissing replicates. Zero observed path-level events use the one-sided 95% Clopper–Pearson convention stated in Section 5.

The latent stable score in the synthetic closed loop is $s(x) = 0.20 + 0.90x_0 - 0.20x_0^2 - 0.06x_1^2$, with development and audit logits $\ell_{\text{dev}}(x) = s(x) + \rho x_1$ and $\ell_{\text{audit}}(x) = s(x) - \kappa x_1$, using $\rho = 1.20$ and $\kappa = .45$. Bernoulli utilities use the logistic transform. Mutations have Gaussian standard deviation .55 and are clipped to $[-2, 2]^2$; the incumbent starts at $(0, 0)$. At each step, M mutations are evaluated with four scout observations each; scout observations are never reused as decision evidence. Decision evidence uses common-uniform paired Bernoulli gains with $\delta = .01$. The main grid is $M \in \{1, 4, 16, 64\}$ with 2,000 trajectories per cell; the $M = 16$ headline uses 5,000 trajectories. Long-stream diagnostics use $T \in \{20, 50, 100, 200, 500\}$, $B = 60$, and 2,000 streams per displayed cell.

B.2 Validity and allocation diagnostics

Controller comparisons feed identical persistent candidate e-value streams to e-LOND, e-LORD, e-SAFFRON, and an e-FWER rule. At $T = 500$ and 5% alternatives, powers are .619, .890, .888, and .415, with empirical FDR .00015, .00055, .00055, and .00020. Clustered-dependence experiments vary within-cluster correlation $\rho \in \{0, .25, .5, .75, .95\}$ over 20,000 null trials per setting; the naive construction treats dependent rows as fresh sequential units, whereas the block-valid construction aggregates each cluster into one bounded evidence unit.

To isolate candidate-time leakage from optional stopping, E33 evaluates component e-values at a fixed endpoint with $B = 800$. It uses the same synthetic score $s(x)$, development/audit logits, practical margin $\delta = .01$, paired Bernoulli gain mechanism, and precommitted betting grid as the main synthetic study. The incumbent is $x = (0, 0)$. For each trial and each candidate-pool size $L \in \{1, 4, 16, 64\}$, we sample $x_1 \sim \mathrm{Unif}[.8, 1.2]$ and solve for the smaller admissible x_0 such that the true audit gain equals exactly δ. The resulting development gains lie in [.2766,.3522], while every challenger remains a true persistent null because $\Delta^{\mathrm{audit}} = \delta$.

E33 compares three protocols over 100,000 trials per L and diagnostic thresholds $h \in \{5, 10, 20, 50\}$. **Frozen** chooses one variant using four development-only scout observations and only then opens decision evidence. **Leak+reuse** opens independent decision-evidence streams for all L variants, selects the variant with the largest fixed-endpoint persistent e-value, and reuses those same observations for the decision. **Leak+restart** makes the same post-evidence selection but treats the leaked observations as candidate-construction data, discards them, and starts a fresh candidate/test index with new decision evidence. Independent streams across variants are deliberate so that the diagnostic isolates candidate-time selection bias. Deployment error rates are outside this calculation.

Allocation studies vary $|\mathcal{E}| \in \{2, 3, 5\}$ (with an additional $|\mathcal{E}| = 8$ calibration), diagnostic threshold $h \in \{5, 10, 20, 50, 100\}$, and compare EGR-Balance with an oracle that knows simulator evidence rates.

B.3 Public-data adaptive search studies

The original fixed stratified digits split contains 808 training, 269 search, 270 development, 270 audit, and 180 hidden examples. Standardization is fit on training only. The RBF-SVC family has 64 configurations, with $\log C \in \{-4, -3, \ldots, 3\}$ and $\log \gamma \in \{-6, -5.5, \ldots, -2.5\}$. The neural family contains 25 one-hidden-layer ReLU MLPs formed by crossing widths $\{4, 8, 16, 32, 64\}$ with train-noise augmentation $\{0, .2, .4, .6, .8\}$. Search uses $T = 12$, $M = 8$ local grid proposals, and 32 scout examples. The RBF-SVC run uses 500 trajectories; the neural rerun uses 1,000 trajectories, $B = 800$, $\delta = .005$, and $\alpha = .10$. Decision evidence is sampled with replacement from fixed empirical development/audit pools; hidden data are evaluation only.

E30 reuses the neural digits architecture but declares three environments: clean development, Gaussian-noise audit, and dropout+noise audit. A stronger corruption is hidden evaluation only. Runs use $N = 800$, $T = 12$, $M = 8$, nominal $B = 400$ per environment, $\delta = .005$, and $\alpha = .10$. EGR-ALLOC uses the cumulative pooled-budget rule described in Section 4.2. EGR-Balance is also evaluated. Because decision evidence is sampled with replacement from fixed empirical pools, validity is relative to those empirical distributions.

E31 fixes the E30 model/evidence bank and adoption rules while changing only the predictable proposal policy: local, wide-local, or global. Each cell uses $N = 400$ trajectories. This isolates empirical sensitivity to proposal dynamics without changing the candidate evidence construction.

B.4 Controlled neural benchmark and matched replay

E32 uses a reproducibly generated 40,000-sample, 50-feature, five-class controlled classification problem and a 16-model MLP bank indexed by hidden width and train-noise augmentation. Three declared environments are clean development, strong additive-noise audit, and dropout+noise audit; a stronger hidden environment is evaluation only. Runs use $N = 500$, $T = 15$, $M = 8$, $B = 400$, $\delta = .0025$, and $\alpha = .10$. E32 is designed as a scale/mechanism stress test; public-image and LLM-agent settings are not represented.

For the matched frozen-candidate replay, all 600 ordered non-self pairs from the 25-model neural grid are frozen. Each pair is evaluated with 200 replicated evidence streams at $h \in \{5, 10, 20, 50, 100\}$ and per-environment cap $B = 800$. Within each pair/replicate, identical development and audit streams are reused across fixed-endpoint IUT, fixed-anytime IUT, Stagewise EGR, and EGR-ALLOC. Therefore fixed-anytime IUT and Stagewise EGR have identical alternative crossing events by construction and differ only in evidence usage.

C Additional Results and Diagnostics

C.1 Original two-environment public-data results

Table 5: Original public-data adaptive searches. RBF-SVC uses $N = 500$ trajectories and Neural MLP uses $N = 1000$; "0 obs." follows the common Clopper–Pearson convention in Section 5. The IUT baseline follows Berger and Hsu (1996); online e-LOND follows Xu and Ramdas (2024).

Benchmark	Metric	Fixed IUT	Stagewise EGR	EGR-ALLOC
RBF-SVC	FDR	0 obs.	0 obs.	0 obs.
	Power	.522	.554	.583
	Evidence	1200.0	567.5	584.3
Neural MLP	FDR	0 obs.	.0005	0 obs.
	Power	.319	.355	.374
	Evidence	1600.0	782.3	781.0

C.2 Three-environment uncertainty and balancing

For E30, fixed IUT, Stagewise EGR, and EGR-ALLOC each have 0 observed any-false paths in 800 trajectories; the one-sided 95% Clopper–Pearson upper bound is .00374. Power (95% MC interval) is .3137 [.2981,.3293], .3221 [.3064,.3379], and .3349 [.3185,.3514]. Evidence intervals are 1200 exactly, 388.8 [387.1,390.4], and 392.6 [390.3,394.9]. EGR-Balance attains power .3687 [.3517,.3858] with evidence 1106.4 [1105.4,1107.5] and one observed false-adoption path ($\mathbb{P}(V_T > 0) = .00125$); we report it separately as an allocation diagnostic.

For E32, fixed IUT, Stagewise, and EGR-ALLOC have 0 observed any-false paths in 500 trajectories (one-sided 95% Clopper–Pearson upper bound .00597). Power intervals are .2149 [.1886,.2413], .2380 [.2135,.2625], and .2666 [.2414,.2919]; evidence intervals are 1200 exactly, 444.6 [440.2,448.9], and 440.9 [436.6,445.3].

C.3 Proposal-policy robustness

Table 6: Proposal-policy robustness (E31). Every displayed audited method has zero observed FDR in 400 trajectories per policy. Baselines: IUT (Berger and Hsu, 1996); e-LOND (Xu and Ramdas, 2024).

Policy	Metric	Fixed IUT	Stagewise EGR	EGR-ALLOC
Local	Power	.3192	.3224	.3309
	Evidence	1200.0	389.3	390.7
Wide-local	Power	.5466	.5519	.5705
	Evidence	1200.0	379.0	379.4
Global	Power	.6993	.6877	.6855
	Evidence	1200.0	379.1	378.4

The global-policy row shows that adaptive evidence placement can save observations without uniformly increasing finite-run power across proposal dynamics. The validity statement remains the conditional guarantee in Theorem 2.

C.4 Frozen-pair diagnostics

All null rates in this subsection are candidate-level crossing frequencies, not online-stream FDR.

Table 7: Frozen-pair replay at diagnostic threshold $h = 20$. Classical IUT follows Berger and Hsu (1996); the anytime e-value construction follows betting-based safe inference (Ramdas et al., 2023; Waudby-Smith and Ramdas, 2024).

Metric	Fixed endpoint	Fixed anytime	Stagewise EGR	EGR-ALLOC
Pers.-alt. power	.6082	.6477	.6477	.6699
Persistent-null crossing	.000013	.000355	.000355	.000395
Mismatch-null crossing	.000000	.001667	.001667	.001905
Evidence	1600.0	1600.0	702.4	715.7

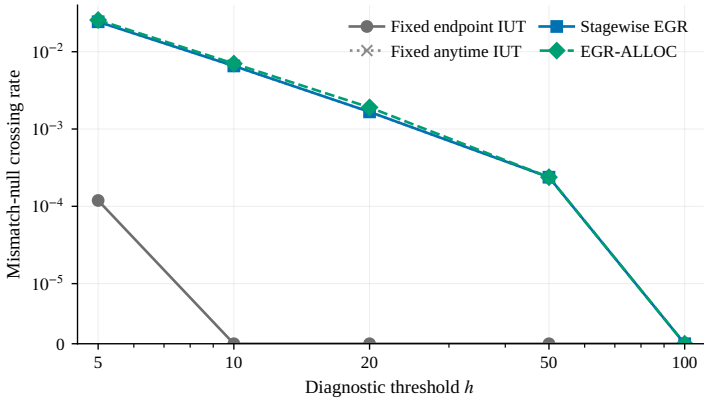


Figure 5: Frozen-pair target-mismatch crossing versus diagnostic threshold. Fixed-anytime IUT and Stagewise EGR have identical crossing events in this matched replay; EGR-ALLOC may differ because saved first-stage budget is reassigned. The vertical axis uses a symmetric-log transform around zero so exact zero crossing rates remain visible; only the nonnegative range is displayed.

C.5 Controller and evidence-unit diagnostics

At $T = 500$ with 5% alternatives and identical persistent candidate e-values, e-LOND, e-LORD, e-SAFFRON, and the e-FWER rule attain powers .619, .890, .888, and .415, respectively; empirical FDR values remain small in these runs. Downstream spending therefore changes long-horizon power while the candidate-level EGR construction is fixed. In the dependence stress test, the null crossing probability at $\rho = .95$ is .728 when highly correlated rows are treated as fresh evidence units and .0127 after cluster aggregation.

C.6 Candidate-freezing / leakage diagnostic

The supplementary candidate-freezing threshold sweep (E33) compares Frozen, Leak+reuse, and Leak+restart across candidate-pool sizes. With one available variant, Frozen and Leak+reuse coincide. At $h = 10$ and $L = 64$, the rates are .352% (Frozen), 20.089% (Leak+reuse), and .335% (Leak+restart); the reuse rate exceeds the $1/h = 10\%$ fixed-e-value reference. At $h = 20$, the corresponding rates are .052%, 3.802%, and .063%.

Full Wilson intervals for the E33 threshold sweep are provided in the supplementary E33 report.

Additional allocation calibration. At $h = 20$ and $|\mathcal{E}| = 2, 3, 5, 8$, balancing/oracle powers are .898/.904, .961/.965, .996/.996, and 1.000/1.000; mean evidence is 114.1/114.4, 144.4/142.8, 190.8/191.5, and 260.7/258.0, compared with equal-split caps 240, 360, 600, and 960. These simulator diagnostics do not strengthen Proposition 2; the proposition remains the deterministic first-order statement in Section 4.2.