VALIDATING DISCOVERED DIFFERENCES IN MODEL REPLACEMENT AUDITS

Anonymous authors
Paper under double-blind review

ABSTRACT

When an AI model is updated or replaced, we need to know how its behavior changes. Testing every possible input is impractical, so we search for behavioral differences between the original model and its replacement. But search creates two problems: a discovered difference may not hold on other inputs, while a search that finds no difference may still miss differences elsewhere. We separate what search discovers from what the evidence supports. Before search, we specify how discovered differences will be followed up on new inputs, and fresh validation determines whether there is enough evidence to support them. A false acceptance occurs when validation supports a difference that does not actually hold. Our validation controls this rate across search strategies. On Berkeley Function Calling Leaderboard (BFCL) tasks and controlled experiments, it validates differences on new inputs without false acceptances, while a simpler baseline produces a 49.44% false-acceptance rate. When search finds no difference, we instead analyze coverage of the untested inputs. Using a bound on how much the difference can change between nearby inputs, we determine when tested inputs can rule out differences elsewhere. Experiments show that limited coverage can leave existing differences undetected. Together, validation and coverage clarify what can and cannot be concluded about behavioral differences without exhaustive testing.

Code: https://anonymous.4open.science/r/vdd/

1 INTRODUCTION

AI systems may need to replace their models because of cost, or deployment needs. Replacements can include switching models, using distilled versions, or introducing routers or cascades (Hinton et al., 2015; Dekoninck et al., 2025; Casey et al., 2026). Even when overall performance is maintained or improved, replacements can change behavior on individual inputs, as observed in both standard machine learning and LLM updates (Srivastava et al., 2020; Echterhoff et al., 2024)

One way to find these differences is adaptive testing, which uses earlier results to search for inputs where two systems behave differently (Lee et al., 2020; Perez et al., 2022). However, search creates two problems. First, a difference found during search may not hold on a broader population of inputs, especially because search favors large observed differences. Second, a search that finds no difference may still miss differences on inputs that were not tested.

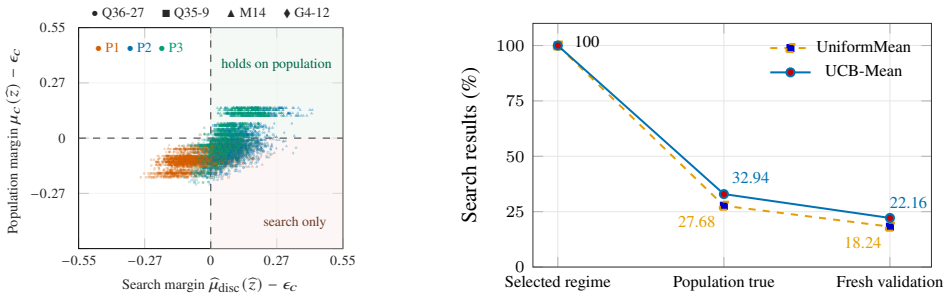
To address the first problem, we introduce a counterfactual witness interface for validating differences found during search. The key idea is to separate search from validation. Search selects a regime with an observed difference. Fresh validation then tests whether this difference holds over a pre-specified population of inputs. This population is the group of related inputs from which fresh validation samples are drawn.

We use BFCL function-calling tasks (Patil et al., 2025) to compare search and fresh validation against population ground truth, which is not used by the validation procedure. Figure 1 shows a substantial gap between search and population truth, while fresh validation is more conservative and never certifies a population-false candidate in the primary experiments.

The second problem is negative search. We analyze whether tested inputs cover the remaining input space, using a bound on how much discrepancy can change between nearby inputs. Limited coverage can therefore leave existing differences undetected. Across our experiments, E1 evaluates search and

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.

fresh validation on BFCL (Patil et al., 2025), E2 tests the validation rule under controlled ground truth, and E3 studies negative-search coverage.



(a) Search and population differences.

(b) Search, population truth, and validation.

Figure 1: Discovery is not population evidence on BFCL. (a) Search and population margins for a fixed subset of 7,201 results; lower- right points appear positive during search but are population false. (b) Aggregate selected-regime, population-true, and fresh-validation yields for UniformMean and UCB-Mean.

Relation to prior work. Prior work on model migration and backward compatibility studies how model updates can change behavior that users or downstream systems rely on. Adaptive stress testing and red teaming provide ways to search for such behavioral differences, while adaptive data analysis studies how conclusions can remain valid after data-dependent selection (Dwork et al., 2015; Rogers et al., 2020). Recent work also studies which model properties can be reliably audited under updates (Ajarra & Basu, 2026). Other approaches separate failure discovery from later validation, using crowdsourced failure reports (Cabrera et al., 2021) or adversarial falsification followed by independent validation (Basu & Chakraborty, 2026). Our setting connects these ideas by specifying, before search, the population and validation procedure used to evaluate any selected finding.

Contributions.

- **Population-bound validation.** We formalize model-replacement auditing as a post-selection problem: adaptive search selects a regime with an observed difference, and fresh validation controls false acceptance uniformly over adaptive discovery policies.
- **Search–validation separation.** On BFCL and controlled experiments, we show that search success can substantially overstate population-supported differences, and isolate uncertainty and feasibility as distinct requirements for certification.
- **Limits of negative search.** We characterize when a failed search can support broader clearance: coverage and bounded discrepancy change, rather than query count alone, determine what can be ruled out beyond the queried regimes.

2 Population-bound model-replacement audits

2.1 Define the claim before search

Matched responses and feasibility. Let f_t be the original model and s_C a candidate replacement fixed before discovery. An action $a \in \mathcal{A}$ sends the same controlled case to both systems. A world W specifies their response laws $Y_t^W(a)$ and $Y_C^W(a)$. A predeclared feasibility rule uses only the case and its metadata, $F^W(a) = \varphi_F(a) \in \{0, 1\}$, before either endpoint is queried. On feasible cases, a fixed discrepancy rule gives

$$Z_C^W(a) = \ell_C\big(a, Y_t^W(a), Y_C^W(a)\big) \in [0, M]. \tag{1}$$

The matched record is $O^W(a) = (Y_t^W(a), Y_C^W(a), F^W(a))$. If $F^W(a) = 0$, both endpoints are skipped, the responses and $Z_C^W(a)$ are null, and $F^W(a)Z_C^W(a)$ is defined as zero. The comparison is counterfactual only in the literal operational sense: it compares what the incumbent and proposed replacement return on the same case.

Each regime names one population claim. A regime $z \in \mathcal{Z}$ is a claim that discovery may select. Its kernel K_z defines the follow-up population over cases $u \in \mathcal{U}_z$, and $a(z, u)$ turns a sampled case into a matched evaluation. Expectations below are over $u \sim K_z$ and fresh response randomness:

$$\mu_C^W(z) = \mathbb{E}_{W,K_z}\left[F^W(a)Z_C^W(a)\right], \qquad p^W(z) = \mathbb{E}_{W,K_z}\left[F^W(a)\right], \tag{2}$$

$$T_{C,z}(W) = \mathbf{1}\left\{\mu_C^W(z) > \epsilon \text{ and } p^W(z) \geq 1-\rho\right\}, \tag{3}$$

$$T_C^{\exists}(W) = \mathbf{1}\{\exists z \in \mathcal{Z} : T_{C,z}(W) = 1\}. \tag{4}$$

Here $\mu_C^W(z)$ is the average feasible discrepancy with infeasible cases contributing zero, and $p^W(z)$ is the feasible-case probability. A point kernel makes a claim about one case. A broader kernel makes a claim about a population. The kernel must be fixed before search.

The audit plan. The contract C fixes the systems, discrepancy rule, regimes, kernels, thresholds ϵ, ρ, and discrepancy bound M (Appendix A.1). The plan $P = (C, B, n, \alpha)$ also fixes the discovery budget, validation sample size, and error level. A discovery policy uses its history H_B and random seed R_π to select $\widehat{z} = \psi(H_B, R_\pi)$. The verifier tests $T_{C,\widehat{z}}$: the population claim attached to the selected regime. It does not test $T_C^{\exists}$, which is the stronger claim that some witness exists anywhere in the declared family. Rejecting an entire replacement class requires the stronger minimax record in Appendix A.1.1.

2.2 Validate the selected claim on fresh cases

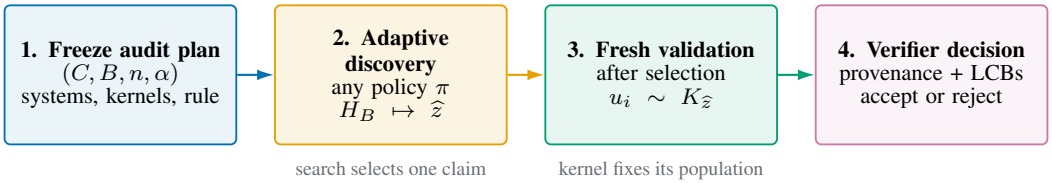


Figure 2: The audit contract is fixed before discovery. After search selects a regime, validation draws fresh cases from that regime's declared population and tests the same claim.

Validation protocol. After discovery, seal $\widehat{z}$, obtain fresh randomness, and draw $u_i \overset{\text{iid}}{\sim} K_{\widehat{z}}$. The verifier receives the audit plan, selected regime, sampler provenance, matched validation data, and confidence bounds (Appendix A.4).

For validation case i, apply the fixed feasibility and discrepancy rules. Infeasible cases remain in the denominator n with product zero; they are not removed. Define

$$G_i = \frac{F_i Z_{C,i}}{M} \in [0,1], \qquad \bar{G} = \frac{1}{n}\sum_{i=1}^n G_i, \qquad \bar{F} = \frac{1}{n}\sum_{i=1}^n F_i, \tag{5}$$

$$c_n(\alpha) = \sqrt{\frac{\log(2/\alpha)}{2n}}, \tag{6}$$

$$\mathrm{LCB}_\mu = M\big(\bar{G} - c_n(\alpha)\big), \qquad \mathrm{LCB}_p = \bar{F} - c_n(\alpha), \tag{7}$$

$$V_P(X) = 1 \iff \mathrm{LCB}_\mu > \epsilon \text{ and } \mathrm{LCB}_p \geq 1-\rho. \tag{8}$$

The two lower bounds estimate the population discrepancy and feasible-case probability required by the claim.

Theorem 1 (Selection-stable witness). *Let $\mathcal{F}_{\text{disc}}$ contain any replacement-construction data, the frozen incumbent and candidate systems, discrepancy rule, policy randomness, discovery history, and selected regime $\widehat{z}$. Conditional on this record, assume that the validation cases are i.i.d. draws from $K_{\widehat{z}}$. Assume fresh response randomness that is independent across cases and independent of $\mathcal{F}_{\text{disc}}$. If $0 \leq Z_C \leq M$ on feasible cases, with zero product otherwise, then every randomized or adaptive discovery policy satisfies*

$$\Pr\left[\mathrm{LCB}_\mu \leq \mu_C^W(\widehat{z}),\ \mathrm{LCB}_p \leq p^W(\widehat{z})\right] \geq 1-\alpha, \tag{9}$$

$$\sup_{\pi \in \Pi_B} \sup_{W \in \mathcal{W}_P} \Pr_{W,\pi,P}\left[V_P(X_{\pi,B,n}) = 1,\ T_{C,\widehat{z}}(W) = 0\right] \leq \alpha. \tag{10}$$

Here Π_B contains policies that use only the discovery record, and $\mathcal{W}_P$ contains worlds satisfying the plan's response bound and fixed validation law. Under the margin condition in Eq. (27), the same proof gives conditional acceptance probability at least $1 - \alpha$.

The theorem is a post-selection validity statement. Discovery may be arbitrarily adaptive within its declared record; validity comes from testing the selected claim on fresh data from the population fixed for that claim before discovery.

2.3 Why a failed search does not prove clearance

The positive validation result applies to one selected regime. A global statement that no witness exists must also account for regimes that were never tested.

Global clearance requires coverage. Consider a point-regime design with $\mathcal{Z} = \mathcal{A}$, so each regime is one action. Assume noiseless matched responses, $F \equiv 1$, a deterministic audit, and $L > 0$. Assume the admissible world class contains the zero-discrepancy world and every bounded nonnegative L-Lipschitz discrepancy. Let Q be the actions queried on the all-zero transcript. Define the largest distance from any action to the queried set by

$$h(Q) = \sup_{a \in \mathcal{A}} \min_{q \in Q} d(a, q), \tag{11}$$

$$h(Q) \le \epsilon/L, \qquad B \ge \mathcal{N}(\mathcal{A}, d, \epsilon/L). \tag{12}$$

If $Lh(Q) > \epsilon$ and $M > \epsilon$, there is another admissible discrepancy that is zero on every queried action but exceeds ϵ at an unqueried action (Proposition 1). The complete transcript is unchanged while the global verdict changes. Therefore worst-case clearance in this class requires an ϵ/L cover. Additional structure can reduce this requirement, but the structure must be stated.

The largest searched discrepancy is not a certificate. Search returns a selected extreme. Under the clipped-Gaussian null in Proposition 2, every regime has population discrepancy $\nu < \epsilon$, yet the raw maximum M_B exceeds any fixed $u \in (\nu, M)$ with probability approaching one as B increases. Search can identify a candidate for validation. Reusing the selected extreme as evidence for its own population claim does not remove selection bias.

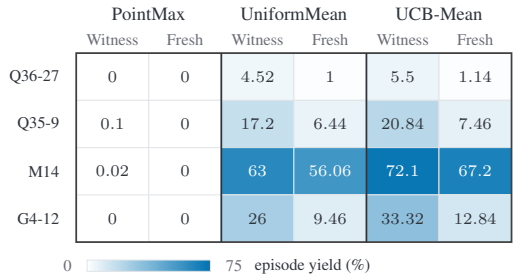
3 Experiments

We apply the audit framework of Section 2 using fixed contracts. In E1, regimes are BFCL tool-call format configurations, and K_z defines each regime's follow-up population over unseen BFCL tasks from a fixed hidden population. The full hidden population is used only to determine population ground truth and does not enter fresh validation. E1 follows search-selected differences through fresh validation on real LLMs and compares them with this ground truth. E2 tests the validation rule under controlled population truth. E3 studies negative search outcomes and asks what a search that finds no difference can rule out.

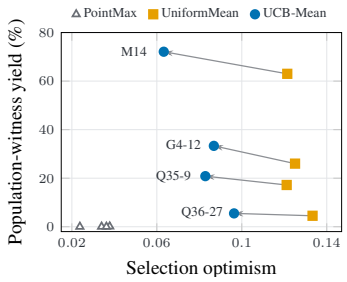
3.1 E1: how often does a discovered difference survive population validation?

Experimental setup. E1 uses BFCL function-calling tasks (Patil et al., 2025) with four target models: Qwen3.6-27B (Q36-27), Qwen3.5-9B (Q35-9), Ministral-3-14B (M14), and Gemma-4-12B (G4-12). For a regime z and a BFCL task u from its follow-up population, $a(z, u)$ formats each BFCL task for matched evaluation by the incumbent and replacement. Feasibility requires the formatted task to preserve the original tool-call semantics and remain within the input limit; $p(z)$ is the resulting feasible-case probability. Each model serves in turn as the incumbent, while the other three form a fixed convex replacement following ISQED (You, 2026). BFCL scores each function call as correct or incorrect, and we measure discrepancy as the absolute difference between the incumbent's correctness and the replacement's expected correctness. For each incumbent, we split the BFCL tasks into five folds and use each fold in turn as the hidden population, giving 20 contracts in total. Search selects a regime, FreshLCB evaluates it on fresh tasks drawn from its follow-up population, and the full hidden population provides population ground truth. We compare three search methods: PointMax, UniformMean, and UCB-Mean. Contract and implementation details are given in Appendix B.1.

Many positive search findings do not hold across their populations. UniformMean and UCB-Mean produce 34,636 selections with positive discovery mean. Of these, 22,627 (65.3%) are not population witnesses on their full hidden populations. Thus, a difference observed during search often does not persist at the population level.



(a) Target-level population-witness and FreshLCB yield.

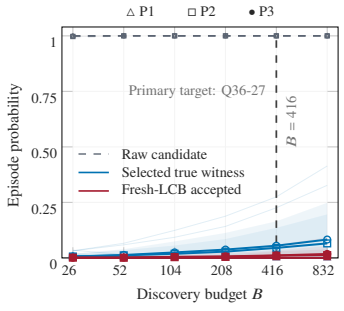


(b) Selection optimism and certificate yield.

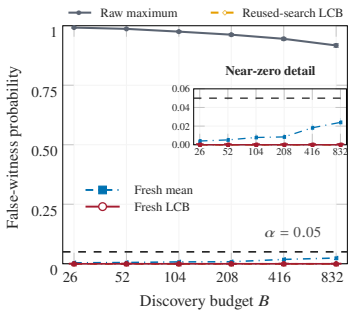
Figure 3: E1 population-witness and FreshLCB certificate yields by target and search policy. UCB-Mean increases both yields for every target, while absolute yield differs substantially across targets. Rates condition on the fixed contract mixture.

Better search finds stronger selected regimes. Compared with UniformMean, UCB-Mean increases average population-witness yield from 27.68% to 32.94% and FreshLCB certificate yield from 18.24% to 22.16% (Figure 3a). Selection optimism is the gap between the discovery estimate and the population discrepancy of the selected regime; its mean decreases from 0.125 under UniformMean to 0.082 under UCB-Mean (Figure 3b). Better search therefore finds stronger selected regimes, while certification still depends on independent fresh evidence.

Targets differ partly because witnesses are absent in some contracts. Only 12 of the 20 contracts contain any population witness, so the opportunity to find one already differs across the fixed contract mixture. Witnesses are also unevenly distributed across targets: M14 contributes 6,756 of the 12,130 population witnesses found across all search policies, and under UCB-Mean population-witness yield ranges from 5.50% for Q36-27 to 72.10% for M14 (Figure 3a). Thus, the large differences in yield across targets partly reflect differences in the availability of population witnesses in their underlying contracts, rather than search performance alone.



(a) Q36-27 stage yields.



(b) Q36-27 UCB-Mean false-witness rates.

Figure 4: Increasing discovery budget does not remove certification cost (Q36-27 only). (a) Population-witness and certificate yields across its five folds. (b) UCB-Mean false-witness rates for the decision rules; the inset resolves the near-zero rates. Only fold 2 contains population witnesses. Error bars and the inset retain the frozen-source uncertainty conventions.

More discovery budget does not remove the need for validation. For Q36-27, under UCB-Mean discovery, increasing B from 26 to 832 raises population-witness yield from 0.64% to 8.28%

(Figure 4a). Over the same sweep, FreshMean false acceptance rises from 0.40% to 2.40%, while FreshLCB records none (Figure 4b). Thus, more discovery can find more witnesses without removing the need for independent validation. Appendix B.3 reports the full fold-level results.

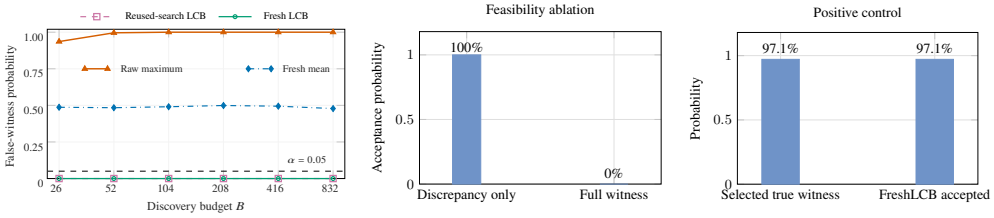
3.2 E2: FRESH SAMPLES NEED UNCERTAINTY AND FEASIBILITY CHECKS

E2 tests FreshLCB in four controlled settings where population truth is known in advance. W0 lies exactly at the population-discrepancy threshold, $\mu = \epsilon$, and is therefore not a population witness. W1 contains rare large discrepancies, but its population discrepancy remains below the threshold, $\mu < \epsilon$. W2 satisfies the discrepancy requirement, $\mu > \epsilon$, but fails the feasibility requirement, $p < 1 - \rho$. W3 satisfies both requirements and serves as a positive control. Each primary cell contains 5,000 episodes under UCB-Mean with $B = 416$ and $n = 1{,}024$. Table 1 summarizes the primary results, and Appendix B.4 gives the full simulation parameters and grids.

Table 1: E2 primary-cell calibration: 5,000 episodes per row; rates in percent. "Upper" is the archived one-sided 95% Wilson upper bound (Wilson, 1927) on false acceptance. Joint coverage concerns both FreshLCB bounds and is not reported for other rules. Acceptance is power only on true alternatives.

World	Rule	False	Upper	Joint coverage	Accepted
W0 boundary	Raw maximum	100.00	100.00	—	100.00
	Reused LCB	0.00	0.05	—	0.00
	Fresh mean	49.44	50.60	—	49.44
	FreshLCB	0.00	0.05	100.00	0.00
W1 spiky null	FreshLCB	0.00	0.05	100.00	0.00
W2 feasibility	FreshLCB	0.00	0.05	99.86	0.00
W3 positive	FreshLCB	0.00	0.05	99.92	97.10

Reliable fresh validation requires both uncertainty correction and feasibility checks. In W0, FreshMean falsely accepts 49.44% of episodes, while FreshLCB records no false acceptance (Figure 5a); FreshLCB also records none in W1. W2 isolates feasibility: although $\mu > \epsilon$, only 89% of cases are feasible, below the 90% feasibility level required in our setting. A discrepancy-only rule therefore accepts every episode, whereas the full rule, which also checks feasibility, accepts none (Figure 5b). In the positive-control W3, UCB-Mean selects a population witness in 97.1% of episodes and FreshLCB accepts all of them (Figure 5c). Finally, the two FreshLCB lower bounds jointly cover their population values in 99.86% of W2 episodes and 99.92% of W3 episodes, consistent with Theorem 1.



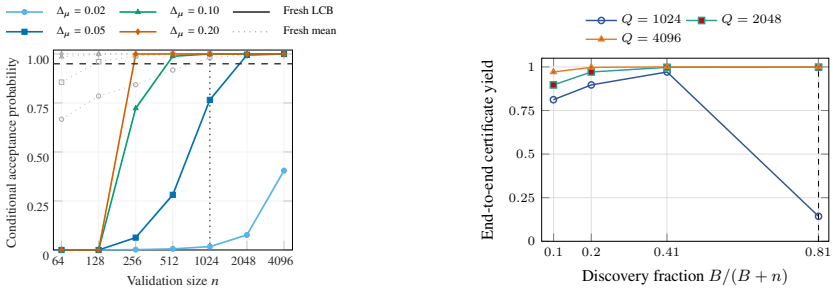
(a) W0 decision rules. **(b)** W2 feasibility ablation. **(c)** W3 positive control.

Figure 5: E2 validation-rule diagnostics: (a) W0 decision rules, (b) W2 feasibility ablation, and (c) W3 positive control with 97.1% FreshLCB acceptance.

3.3 E2: VALIDATION REQUIRES ENOUGH SAMPLES

Validation requires enough fresh samples. From Eq. 8, $\mathrm{LCB}_p = \bar{F} - c_n(\alpha)$. Even when every fresh validation case is feasible, so that $\bar{F} = 1$, the uncertainty term $c_n(\alpha)$ keeps the lower bound below one. At least $n = 185$ fresh cases are therefore needed for LCB_p to reach the required feasibility level $1 - \rho = 0.90$. Above this minimum, certification can still be difficult when μ is close to ϵ.

Population witnesses closer to the discrepancy threshold require more validation data. Figure 6a varies the fresh validation sample size n and discrepancy margin $\Delta_\mu = \mu - \epsilon$, the amount by which the population discrepancy exceeds its threshold. When $\Delta_\mu = 0.02$, FreshLCB acceptance increases from 1.70% at $n = 1{,}024$ to 40.48% at $n = 4{,}096$, so witnesses closer to the discrepancy boundary require more fresh evidence to certify. Under a fixed total query budget $Q = B + n$, this creates a trade-off between discovery and validation (Figure 6b): allocating more queries to discovery increases B but leaves fewer fresh samples n for validation. At $Q = 1{,}024$, allocating 81.25% to discovery leaves $n = 192$ and gives 14.32% certificate yield; at $Q = 2{,}048$, the same split leaves $n = 384$ and gives 99.88%. Thus, too little validation data can sharply limit end-to-end certificate yield. Appendix B.4 reports the full grids.



(a) Power versus validation size. **(b)** Certificate yield versus discovery/validation split.

Figure 6: E2 validation sample cost and discovery–validation budget trade-off. (a) FreshLCB acceptance versus validation size and discrepancy margin. (b) End-to-end certificate yield versus the discovery fraction of a fixed query budget.

Certification depends on both discrepancy and feasibility margins. Table 2 additionally varies the feasibility margin $\Delta_p = p - (1 - \rho)$, where $1 - \rho = 0.90$ in our setting. Larger discrepancy and feasibility margins generally make population witnesses easier to certify. The boundary case is especially difficult: when $\Delta_p = 0$, population feasibility exactly meets the required 90%, yet FreshLCB certifies none of the displayed cases even when the discrepancy margin is large. Thus, a population witness can satisfy both population requirements while remaining difficult to certify from finite validation data when either margin is small.

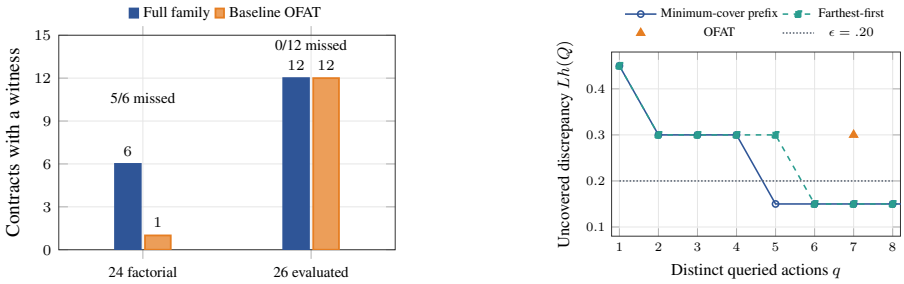
Table 2: E2 FreshLCB acceptance (%) across discrepancy and feasibility margins. Each cell contains 5,000 UCB-Mean episodes at $B = 416$ and a population witness. Bold values satisfy the sufficient-margin condition in Eq. (27).

	Feasibility margin	Discrepancy margin Δ_μ			
n	Δ_p	.02	.05	.10	.20
1,024	0.00	0.00	0.00	0.00	0.00
1,024	0.05	1.58	65.76	85.24	85.16
1,024	0.10	1.70	76.58	**100.00**	**100.00**
4,096	0.00	0.00	0.00	0.00	0.00
4,096	0.05	40.76	**100.00**	**100.00**	**100.00**
4,096	0.10	40.48	**100.00**	**100.00**	**100.00**

3.4 E3: WHAT DOES A NEGATIVE SEARCH RULE OUT?

E3 studies the negative-search question from Section 2.3: how much can we conclude from a search that finds no witness? We first use E1 to show that restricted search can miss existing population witnesses. We then study coverage directly in a controlled finite regime space. Following Section 2.3, we relate the distance from each unqueried regime to the queried set to a bound on how much discrepancy can change across that distance. This determines whether the queried set is sufficient to rule out witnesses elsewhere. Neither analysis requires new model calls.

The answer depends on the declared regime family. We first consider the 24 regimes formed by the $4 \times 3 \times 2$ combinations of return format, documentation format, and tool-call tags. We compare this full family with a one-factor-at-a-time (OFAT) (Daniel, 1973) search that starts from one baseline regime and changes one factor at a time, querying seven regimes. Across the 20 E1 contracts, six contain a population witness in the full 24-regime family, but OFAT finds a witness in only one of them (Figure 7a). E1 also evaluates two additional single-factor regimes, expanding the evaluated family to 26 regimes and OFAT to nine queries. In this expanded family, 12 contracts contain a population witness and OFAT finds all 12. Thus, the observed miss rate changes from $5/6$ in the 24-regime family to $0/12$ in the evaluated 26-regime family. A negative result in one declared regime family therefore does not automatically extend to a larger or different family.



(a) Population witnesses under two regime families. **(b)** Exact coverage in the controlled 24-point space.

Figure 7: E3 negative-search scope. (a) Full-family and baseline-OFAT population verdicts on the same 20 contracts. (b) Exact coverage of the controlled 24-point Hamming space with noiseless point kernels, $F \equiv 1$, $L = 0.15$, $\epsilon = 0.20$, and $M = 0.45$. A minimum cover uses five queries; every enumerated farthest-first path first covers at six; the seven-query OFAT set leaves $Lh = 0.30$. The value $L = 0.15$ is used only for the controlled calculation.

We use the discrepancy margin to identify witnesses missed by OFAT. For each regime, we compute the population discrepancy margin $\Delta_\mu = \mu - \epsilon$, as in Section 3.3. Because all regimes in this comparison satisfy the feasibility requirement, a positive margin identifies a population witness. We compare the largest margin among the seven regimes checked by OFAT with the largest margin among all 24 regimes. If the OFAT maximum is negative but the full-24 maximum is positive, OFAT finds no witness in the regimes it checks even though one exists elsewhere in the 24-regime family. For example, in M14 fold 1, the OFAT maximum is -0.0368, while the full-24 maximum is $+0.0569$ (Table 3). OFAT therefore misses a population witness in this contract.

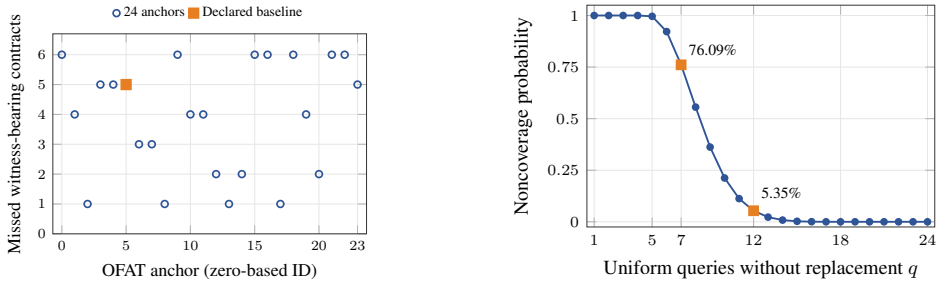
Table 3: E3 discrepancy margins for the six witness-bearing contracts in the full 24-regime family. Entries are the largest margins in each searched set; bold values indicate witnesses missed by OFAT. Full 26 and OFAT 9 have identical maxima in these contracts. Table 9 reports all 20 contracts.

Target	Fold	Full 24	OFAT 7	Full 26 / OFAT 9
Q35-9	2	+0.0163	+0.0163	+0.0631
M14	0	+0.0115	**−0.0493**	+0.1496
M14	1	+0.0569	**−0.0368**	+0.1114
M14	2	+0.0516	**−0.0348**	+0.1399
M14	4	+0.0329	**−0.0622**	+0.1221
G4-12	2	+0.0269	**−0.0170**	+0.0722

Coverage matters more than query count alone. We study this in a controlled space of 24 regimes, where two regimes are one step apart if they differ in one factor. We assume a Lipschitz bound $L = 0.15$ on discrepancy across neighboring regimes, so discrepancy can change by at most L in one step. Under the coverage bound in Eq. 20, with a discrepancy threshold of $\epsilon = 0.20$, an unqueried regime within one step of a queried zero-discrepancy regime can be ruled out, whereas one two steps away cannot be ruled out by this bound. We therefore ask how many queries are needed

to place every regime within one step of the queried set. Some sets achieve this coverage with five queries, and farthest-first achieves it with six. OFAT uses seven queries but still leaves some regimes two steps away (Figure 7b). Thus, more queries do not necessarily provide stronger coverage; where they are placed also matters.

Coverage depends on where search queries. Figure 8a repeats the natural 24-regime analysis using each regime in turn as the OFAT anchor. All 24 choices miss at least one of the six contracts that contain a population witness in the full 24-regime family, while the declared baseline misses five. Thus, changing the anchor changes which regimes OFAT checks, but no tested anchor eliminates the coverage problem. Random queries do not guarantee coverage either: in the controlled 24-regime space, 76.09% of uniformly selected seven-query sets fail to place every regime within one step of the queried set, and the noncoverage probability remains 5.35% with 12 queries (Figure 8b). These probabilities describe geometric coverage, not false-acceptance rates on the natural BFCL data. The controlled calculation assumes that discrepancy changes by at most $L = 0.15$ between neighboring regimes, an assumption that does not hold for all neighboring regimes in the natural data. Appendix B.6 reports the full coverage analysis.



(a) Natural OFAT anchor sensitivity.

(b) Exact random-query noncoverage.

Figure 8: E3 search-coverage diagnostics: (a) OFAT anchor sensitivity in the natural 24-regime family; (b) exact noncoverage probability for uniform queries in the controlled 24-regime space. Neither changing the OFAT anchor nor increasing query count alone guarantees coverage.

3.5 Scope and Limitations

E1 applies to the fixed BFCL contracts, discrepancy metric, replacements, regime populations, and thresholds. E2 tests the rule under controlled truth. E3's clearance result additionally assumes the stated metric, a known Lipschitz bound, noiseless point regimes, and full feasibility. The archived reruns check execution order and numerical replay, not adversarial sampler independence (Appendix B.5). E1 rates condition on the fixed 20-contract mixture, and zero observed false acceptance is not a zero-risk guarantee; Theorem 1 gives the stated α-level guarantee under its assumptions.

4 Conclusion

Adaptive search can find a regime where a replacement appears different from the incumbent, but the search result alone is not population evidence. The associated claim becomes evidence only after validation on fresh cases from its pre-specified population. Across BFCL and controlled experiments, we find that search-selected differences can disappear at the population level, while fresh validation can control false acceptance when uncertainty and feasibility are accounted for. A failed search raises a different problem. Global clearance requires coverage of the declared regime family, or another explicit assumption about untested regimes; query count alone does not provide that guarantee. Taken together, these results separate two roles in model-replacement audits: search determines where to look, while validation and coverage determine what the resulting evidence supports.

AI USE STATEMENT

Generative AI tools were used to help restructure and edit the manuscript, interpret and present archived experimental results, prepare LaTeX figures and tables, and write code for build and data-consistency checks. All numerical results in the paper come from archived experimental outputs; AI tools were not used to invent missing measurements. The authors are responsible for the final text, mathematical claims, experimental results, and released artifacts.

REPRODUCIBILITY STATEMENT

Section 2 specifies the audit contract and validation rule, and Appendix A gives the proofs. Appendix B lists the fixed contracts, controlled data generators, statistical units, scope comparisons, and execution limitations. The accompanying source contains the plot data and standalone LaTeX figure sources, together with checks that link the reported counts to the archived experimental outputs.

REFERENCES

Ayoub Ajarra and Debabrota Basu. On the hardness of auditing model properties under updates: Complexity of property-preserving updates. In *The 29th International Conference on Artificial Intelligence and Statistics*, 2026. URL https://openreview.net/forum?id=XuptGYEoXO.

Abhinaba Basu and Pavan Chakraborty. Proof-carrying materials: Falsifiable safety certificates for machine-learned interatomic potentials. *arXiv preprint arXiv:2603.12183*, 2026.

Ángel Alexander Cabrera, Abraham J Druck, Jason I Hong, and Adam Perer. Discovering and validating ai errors with crowdsourced failure reports. *Proceedings of the ACM on human-computer interaction*, 5(CSCW2):1–22, 2021.

Emma Casey, David Roberts, David Sim, and Ian Beaver. When your LLM reaches end-of-life: A framework for confident model migration in production systems. *arXiv preprint arXiv:2604.27082*, 2026.

Cuthbert Daniel. One-at-a-time plans. *Journal of the American statistical association*, 68(342): 353–360, 1973.

Jasper Dekoninck, Maximilian Baader, and Martin Vechev. A unified approach to routing and cascading for LLMs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 12987–13010. PMLR, 2025.

Cynthia Dwork, Vitaly Feldman, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Aaron Roth. Preserving statistical validity in adaptive data analysis. In *Proceedings of the Forty-Seventh Annual ACM Symposium on Theory of Computing*, pp. 117–126, 2015.

Jessica Maria Echterhoff, Fartash Faghri, Raviteja Vemulapalli, Ting-Yao Hu, Chun-Liang Li, Oncel Tuzel, and Hadi Pouransari. MUSCLE: A model update strategy for compatible LLM evolution. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 7320–7332. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-emnlp.430.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Ritchie Lee, Ole J. Mengshoel, Anshu Saksena, Ryan Gardner, Daniel Genin, Joshua Silbermann, Michael Owen, and Mykel J. Kochenderfer. Adaptive stress testing: Finding likely failure events with reinforcement learning. *Journal of Artificial Intelligence Research*, 69:1165–1201, 2020.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): From tool use to agentic evaluation of large language models. In *Proceedings of the 42nd International Conference on*

Machine Learning, volume 267 of *Proceedings of Machine Learning Research*, pp. 48371–48392, 2025.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 3419–3448. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.emnlp-main.225.

Ryan Rogers, Aaron Roth, Adam Smith, Nathan Srebro, Om Thakkar, and Blake Woodworth. Guaranteed validity for empirical approaches to adaptive data analysis. In *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, pp. 2830–2840. PMLR, 2020.

Megha Srivastava, Besmira Nushi, Ece Kamar, Shital Shah, and Eric Horvitz. An empirical analysis of backward compatibility in machine learning systems. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 3272–3280, 2020. doi: 10.1145/3394486.3403379.

Edwin B Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927.

Lei You. Quantifying model uniqueness in heterogeneous AI ecosystems. *arXiv preprint arXiv:2601.22977*, 2026. URL https://arxiv.org/abs/2601.22977.

A Formal details and proofs

A.1 Frozen contract and decision scope

The generic contract freezes the incumbent and candidate systems before discovery:

$$C = (f_t, s_C, \ell_C, \mathcal{A}, \mathcal{Z}, \{(\mathcal{U}_z, K_z)\}_{z\in\mathcal{Z}}, a, \varphi_F, \epsilon, \rho, M)\,, \tag{13}$$

where $M > 0$, $0 \le \epsilon < M$, and $0 \le \rho \le 1$. Any fitting, distillation, routing, or other construction step used to obtain s_C must be completed on data independent of discovery and validation.

E1 instantiates s_C as a convex combination of peer-model scalar responses. Given independent fitting actions $a_1^{P_{\mathrm{fit}}}, \dots, a_m^{P_{\mathrm{fit}}}$, it uses

$$\widehat{w} \in \arg\min_{w\in\Delta^{|\mathcal{J}_{-t}|-1}} \frac{1}{m}\sum_{i=1}^{m}\left(Y_t^W(a_i^{P_{\mathrm{fit}}}) - w^\top \Phi_{-t}^W(a_i^{P_{\mathrm{fit}}})\right)^2, \tag{14}$$

then sets $Y_C^W(a) = \widehat{w}^\top \Phi_{-t}^W(a)$ and $\ell_C(a, y, y') = |y - y'|$. The interface and validation theorem do not otherwise use convexity or the peer construction.

A.1.1 Decision scope and population records

Let $\mathcal{S}$ be a declared class of frozen candidate replacements. Let $\mu_s^W(z)$ denote Eq. (2) when candidate s and its declared discrepancy rule replace s_C and ℓ_C, and let $\mathcal{Z}_\rho^W = \{z : p^W(z) \ge 1-\rho\}$. The fixed-candidate and class-wide values are

$$V_{s_C} = \sup_{z\in\mathcal{Z}_\rho^W} \mu_C^W(z), \qquad V_{\mathcal{S}}^\star = \inf_{s\in\mathcal{S}} \sup_{z\in\mathcal{Z}_\rho^W} \mu_s^W(z). \tag{15}$$

Because $V_{\mathcal{S}}^\star \le V_{s_C}$ when $s_C \in \mathcal{S}$, a witness with $V_{s_C} > \epsilon$ rejects the frozen candidate, while $V_{\mathcal{S}}^\star > \epsilon$ requires a minimax audit over the entire replacement class. For E1, $\mathcal{S}$ may be chosen as the convex peer-router class.

The selected-regime and global tasks also require different population records:

$$X_\infty^{\mathrm{sel}}(W, z) = \left(C, z, \mu_C^W(z), p^W(z)\right), \tag{16}$$

$$X_\infty^{\mathrm{all}}(W) = \left(C, \{(z, \mu_C^W(z), p^W(z)) : z \in \mathcal{Z}\}\right). \tag{17}$$

Equality of selected-regime records preserves $T_{C,z}$; equality of global records preserves $T_C^{\exists}$. A verifier can therefore decide one selected witness without receiving the coverage needed for global clearance.

A.2 COVERAGE LOWER BOUND

Proposition 1 (Coverage lower bound). *Consider the point-regime design with $\mathcal{Z} = \mathcal{A}$, $\mathcal{U}_z = \{\star\}$, $a(z, \star) = z$, and $K_z = \delta_\star$. Fix a deterministic audit procedure that makes at most B matched queries before returning a global verdict. Let Q contain every action that the procedure queries when every matched response is zero and feasibility is one. This set includes queries from every audit stage. Assume that the admissible class contains the all-zero world. Also assume that it contains a scalar-output world with $Y_t(a) = r(a)$, $Y_C(a) = 0$, and $F(a) = 1$ for every nonnegative L-Lipschitz function $r \le M$. If $Lh(Q) > \epsilon$ and $M > \epsilon$, then two worlds produce the same complete query transcript and different task labels:*

$$T_C^{\exists}(W_0) = 0, \qquad T_C^{\exists}(W_1) = 1. \tag{18}$$

Proof. Run the deterministic audit procedure in the stipulated all-zero world W_0, where $Y_t = 0$, $Y_C = 0$, and $F = 1$. Let $Q = \{q_1, \dots, q_{B'}\}$ be the resulting set of distinct queried actions, where $B' \le B$. Define

$$r_1(a) = \min\left\{ M, L \min_{q \in Q} d(a, q) \right\}. \tag{19}$$

Distance to a set is 1-Lipschitz. Truncation at M preserves the L-Lipschitz property. Therefore, r_1 is nonnegative, bounded by M, and zero on Q. The richness assumption provides an admissible world W_1 with $Y_t = r_1$, $Y_C = 0$, and $F = 1$. The complete matched observations in W_1 equal those in W_0 on Q.

We use induction to show that the procedure queries the same actions in both worlds. The first action is fixed. Suppose the first $k - 1$ actions agree. These actions lie in Q, and both worlds return the same observations there. The histories are identical. The deterministic procedure therefore chooses the same kth action or stops with the same record. The full transcripts agree.

Every discrepancy in W_0 is zero. Hence $T_C^{\exists}(W_0) = 0$. If $\mathcal{A}$ is compact, some $a^\star$ attains the supremum in the fill distance. Otherwise, choose a sequence that approaches the supremum. The inequalities $Lh(Q) > \epsilon$ and $M > \epsilon$ ensure that some point in this sequence has discrepancy above ϵ. Choose that point as $a^\star$. The point-regime design and feasibility one give $T_C^{\exists}(W_1) = 1$. The complete query transcripts are identical, but the task labels differ. Finally, $h(Q) \le r$ means that Q is an r-net. The smallest cardinality of such a net is $\mathcal{N}(\mathcal{A}, d, r)$. □

The same construction gives the largest discrepancy compatible with an all-zero transcript:

$$\sup_{\substack{0 \le r \le M,\ \mathrm{Lip}(r) \le L \\ r|_Q = 0}} \sup_{a \in \mathcal{A}} r(a) = \min\{M, Lh(Q)\}. \tag{20}$$

Indeed, Lipschitz continuity bounds each admissible $r(a)$ by both M and $Ld(a, Q)$, and Eq. (19) attains the bound. E3 uses this equality only in its declared noiseless, fully feasible controlled class.

A.3 SELECTION INFLATION

Consider a null world in which the latent incumbent and candidate responses agree. Each observed signed difference is bounded noise:

$$\widetilde{\eta}_q = \mathrm{clip}(\eta_q, -M, M), \qquad \eta_q \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma^2). \tag{21}$$

Every regime is feasible and has discrepancy mass $\nu = \mathbb{E}|\widetilde{\eta}_q|$. Set $\nu < \epsilon < M$. Every regime then has task label zero, while the search log reports

$$M_B = \max_{q \le B} |\widetilde{\eta}_q|. \tag{22}$$

Proposition 2 (Selection inflation). *Under this null model, every u with $0 \le u < M$ satisfies*

$$\Pr(M_B \le u) = \left[2\Phi(u/\sigma) - 1\right]^B. \tag{23}$$

For any fixed $u \in (\nu, M)$, the probability $\Pr(M_B > u)$ approaches one as B grows. This statement includes $u = \epsilon$. Before clipping dominates, the maximum grows on the $\sigma\sqrt{\log B}$ scale. It eventually converges to M almost surely.

Proof. For one clipped Gaussian draw and any $0 \le u < M$,

$$\Pr(|\widetilde{\eta}_q| \le u) = \Pr(|\eta_q| \le u) = 2\Phi(u/\sigma) - 1. \tag{24}$$

Independence gives

$$\Pr(M_B \le u) = \prod_{q=1}^{B} \Pr(|\widetilde{\eta}_q| \le u) = \left[2\Phi(u/\sigma) - 1\right]^B. \tag{25}$$

For fixed $u \in (\nu, M)$, the base lies strictly between zero and one. Its Bth power therefore tends to zero. Thus $\Pr(M_B > u) \to 1$. The ordinary Gaussian maximum grows on the $\sigma\sqrt{\log B}$ scale until it reaches the clipping point. A clipped draw has positive probability of equaling M. Independence then gives $M_B \to M$ almost surely. □

A.4 Witness record and validation proof

The exact verifier-facing witness record is

$$X_{\pi,B,n} = \left(P, \widehat{z}, K_{\widehat{z}}, \tau_{\mathrm{samp}}, \{(u_i, a_i, O_i^{\mathrm{val}})\}_{i=1}^{n}, \mathrm{LCB}_\mu, \mathrm{LCB}_p\right). \tag{26}$$

The sampler record τ_{samp} may include a trusted-runner attestation, a seed commitment, and equivalent provenance. It binds the validation cases to $K_{\widehat{z}}$ after selection. A seed commitment supports replay; conditional independence additionally requires fresh post-seal entropy or an appropriate trusted-randomness assumption. The pre-search plan binds the claim family, frozen systems, and discrepancy rule before discovery.

The power part of Theorem 1 uses the following sufficient margin condition for the realized discovery record:

$$\mu_C^W(\widehat{z}) > \epsilon + 2Mc_n(\alpha), \qquad p^W(\widehat{z}) \ge 1 - \rho + 2c_n(\alpha). \tag{27}$$

Proof of Theorem 1. Condition on $\mathcal{F}_{\mathrm{disc}}$. The validation assumptions leave the cases and responses as independent draws from the fixed follow-up population specified by $K_{\widehat{z}}$ and W. Their conditional means are

$$\mathbb{E}[G_i \mid \mathcal{F}_{\mathrm{disc}}] = \frac{\mu_C^W(\widehat{z})}{M}, \qquad \mathbb{E}[F_i \mid \mathcal{F}_{\mathrm{disc}}] = p^W(\widehat{z}). \tag{28}$$

Both variables lie in $[0, 1]$. The one-sided Hoeffding inequality (Hoeffding, 1963) and the definition of $c_n(\alpha)$ imply

$$\Pr\left(\bar{G} - \frac{\mu_C^W(\widehat{z})}{M} > c_n(\alpha) \,\middle|\, \mathcal{F}_{\mathrm{disc}}\right) \le \alpha/2, \tag{29}$$

$$\Pr\left(\bar{F} - p^W(\widehat{z}) > c_n(\alpha) \mid \mathcal{F}_{\mathrm{disc}}\right) \le \alpha/2. \tag{30}$$

A union bound gives Eq. (9) with conditional probability at least $1 - \alpha$. Averaging over the discovery history preserves the guarantee.

On the simultaneous lower-bound event, acceptance implies both population requirements. A false candidate has either $\mu_C^W(\widehat{z}) \le \epsilon$ or $p^W(\widehat{z}) < 1 - \rho$. A false acceptance is therefore contained in the confidence-bound failure event. This event has probability at most α for every W and π, which proves Eq. (10).

Fix a realized $\mathcal{F}_{\mathrm{disc}}$ that satisfies Eq. (27). The lower-tail versions of the same inequalities give, with conditional probability at least $1 - \alpha$,

$$\bar{G} \ge \frac{\mu_C^W(\widehat{z})}{M} - c_n(\alpha), \qquad \bar{F} \ge p^W(\widehat{z}) - c_n(\alpha). \tag{31}$$

Subtracting one further confidence radius in Eq. (7) leaves $\mathrm{LCB}_\mu > \epsilon$ and $\mathrm{LCB}_p \ge 1 - \rho$. The verifier accepts. □

B ADDITIONAL EXPERIMENTAL DETAILS

B.1 E1 CONTRACTS AND STATISTICAL UNIT

E1 uses each of the four models as the incumbent in turn. For each incumbent, the other three models define the fixed convex replacement. Five folds per incumbent give 20 contracts. Replacement fitting is completed before any search-policy operation. Each contract fixes the incumbent, ordered peer models, convex weights, evaluator precision, baseline discrepancy, threshold, feasibility tolerance, validation sample size, and error level.For a regime z and BFCL task u, $a(z, u)$ renders u under the tool-call format specified by z, and the same rendered task is evaluated by the incumbent and replacement. A case is feasible only if this rendering preserves the task's tool-call semantics and remains within the input limit. Each fold serves once as the hidden population; search and fresh validation use the corresponding predeclared regime populations, while the full hidden fold is reserved for population ground truth.

Table 4: Fold-level frozen E1 contracts. Router weights list *peer=weight*; μ_0 is the baseline residual mean. SHA-256 values are 12-character prefixes; the vendored CSV contains the complete digests.

Target	Fold	Frozen router weights	μ_0	ϵ	Prec.	Router hash	Contract hash
Q36-27	0	Q35=0.389, G4=0.331, M14=0.279	0.226	0.276	fp8	`1be33efa55f8`	`be374e05ad42`
	1	Q35=0.244, G4=0.545, M14=0.211	0.227	0.277	fp8	`e1158735ba2e`	`8f6f08ef8f73`
	2	Q35=0.329, G4=0.421, M14=0.250	0.194	0.244	fp8	`1bb72d3ed11c`	`9c1d05159bf6`
	3	Q35=0.303, G4=0.492, M14=0.205	0.226	0.276	fp8	`59731235211b`	`fce70a2b6fe4`
	4	Q35=0.309, G4=0.470, M14=0.221	0.258	0.308	fp8	`224aa742f5fd`	`9a2cdb0367c2`
Q35-9	0	G4=0.338, M14=0.234, Q36=0.429	0.273	0.323	bf16	`6bfb518c1189`	`f32dca295752`
	1	G4=0.334, M14=0.362, Q36=0.304	0.220	0.270	bf16	`83a3013b3350`	`a5a6d5bb702a`
	2	G4=0.306, M14=0.308, Q36=0.385	0.199	0.249	bf16	`1e85c35aec45`	`72eae7e8cfa0`
	3	G4=0.334, M14=0.316, Q36=0.350	0.263	0.313	bf16	`c9e1cdacfe50`	`8b41b5738b68`
	4	G4=0.298, M14=0.362, Q36=0.341	0.280	0.330	bf16	`a13e1ebf700f`	`dc98c9776896`
M14	0	Q35=0.207, G4=0.521, Q36=0.272	0.216	0.266	fp8	`f1e67200bce2`	`72563bcc56a5`
	1	Q35=0.319, G4=0.450, Q36=0.231	0.206	0.256	fp8	`49cb096830cf`	`cabf1718a64e`
	2	Q35=0.218, G4=0.576, Q36=0.206	0.217	0.267	fp8	`3db631658f32`	`70bf1917ad65`
	3	Q35=0.312, G4=0.455, Q36=0.234	0.227	0.277	fp8	`4158bc83946e`	`02c73844e4bb`
	4	Q35=0.273, G4=0.543, Q36=0.184	0.237	0.287	fp8	`67cc70aaa957`	`6f93d6fc9d85`
G4-12	0	Q35=0.262, M14=0.456, Q36=0.283	0.204	0.254	bf16	`e513578290a0`	`46e848fd135d`
	1	Q35=0.219, M14=0.336, Q36=0.445	0.210	0.260	bf16	`085d80a36cb2`	`4e40534d0f27`
	2	Q35=0.190, M14=0.505, Q36=0.305	0.185	0.235	bf16	`07c1c06453c8`	`e5523fa55860`
	3	Q35=0.245, M14=0.338, Q36=0.417	0.203	0.253	bf16	`61d3a130ca56`	`2a5ca7f038af`
	4	Q35=0.194, M14=0.468, Q36=0.338	0.231	0.281	bf16	`445081f41e80`	`c1260deea8ea`

One replay episode is the statistical unit for the reported discrete rates, conditional on this fixed set of contracts. Each target–policy cell contains 5,000 episodes: 1,000 from each fold. Wilson intervals quantify Monte Carlo uncertainty for this fixed mixture. They do not treat the five contracts as a random sample from a larger population of deployments.

B.2 E1 COMPLETE PRIMARY-CELL RESULTS

M14 contributes 6,756 of the 12,130 selected population witnesses and 6,163 of the 8,080 FreshLCB acceptances. These counts explain the large target differences in Figure 3.

B.3 E1 BUDGET AND POLICY CHECKS

The full Q36-27 budget sweep is reported here. Table 6 gives the fold-specific counts. The preregistered Successive-Halving policy P4 adds 120,000 candidate episodes across 120 cells, using the same fixed contracts and verifier and no new LLM generation. Across the four-policy, six-budget grid, FreshLCB makes one false acceptance among 480,000 rule decisions. The decisions share contracts and random prefixes, so they are not 480,000 independent trials.

Figure 4 shows the aggregate Q36-27 curves; Table 6 gives their fold-level denominators.

Table 5: Complete primary-cell E1 results across the 12 target–policy cells. Continuous intervals are exact five-contract cluster-bootstrap 95% intervals; rate intervals are Wilson 95% intervals.

Residual levels and selection optimism

Target	Policy	Discovery mean [CI]	Oracle μ [CI]	Optimism [CI]
Q36-27	PointMax	0.204 [0.200, 0.209]	0.180 [0.174, 0.186]	0.024 [0.016, 0.033]
	UniformMean	0.334 [0.331, 0.337]	0.201 [0.190, 0.210]	0.133 [0.123, 0.145]
	UCB-Mean	0.303 [0.301, 0.304]	0.206 [0.196, 0.217]	0.096 [0.086, 0.107]
Q35-9	PointMax	0.196 [0.192, 0.199]	0.158 [0.156, 0.160]	0.038 [0.034, 0.042]
	UniformMean	0.352 [0.343, 0.361]	0.230 [0.219, 0.241]	0.121 [0.111, 0.132]
	UCB-Mean	0.323 [0.313, 0.332]	0.240 [0.228, 0.251]	0.083 [0.073, 0.094]
M14	PointMax	0.189 [0.184, 0.193]	0.155 [0.148, 0.161]	0.034 [0.030, 0.039]
	UniformMean	0.445 [0.438, 0.453]	0.324 [0.310, 0.337]	0.121 [0.108, 0.138]
	UCB-Mean	0.407 [0.400, 0.415]	0.344 [0.329, 0.358]	0.063 [0.048, 0.081]
G4-12	PointMax	0.195 [0.193, 0.197]	0.159 [0.154, 0.163]	0.036 [0.032, 0.040]
	UniformMean	0.346 [0.338, 0.354]	0.221 [0.205, 0.232]	0.125 [0.117, 0.135]
	UCB-Mean	0.317 [0.311, 0.324]	0.231 [0.213, 0.244]	0.087 [0.078, 0.098]

Decision rates (percent [95% CI])

Target	Policy	True witness	FreshLCB	False FreshLCB	FreshLCB | true
Q36-27	PointMax	0.00 [0, 0.08]	0.00 [0, 0.08]	0.00 [0, 0.08]	—
	UniformMean	4.52 [3.98, 5.13]	1.00 [0.76, 1.32]	0.00 [0, 0.08]	22.12
	UCB-Mean	5.50 [4.90, 6.17]	1.14 [0.88, 1.47]	0.00 [0, 0.08]	20.73
Q35-9	PointMax	0.10 [0.04, 0.23]	0.00 [0, 0.08]	0.00 [0, 0.08]	0.00
	UniformMean	17.20 [16.18, 18.27]	6.44 [5.79, 7.15]	0.00 [0, 0.08]	37.44
	UCB-Mean	20.84 [19.74, 21.99]	7.46 [6.76, 8.22]	0.00 [0, 0.08]	35.80
M14	PointMax	0.02 [0, 0.11]	0.00 [0, 0.08]	0.00 [0, 0.08]	0.00
	UniformMean	63.00 [61.65, 64.33]	56.06 [54.68, 57.43]	0.00 [0, 0.08]	88.98
	UCB-Mean	72.10 [70.84, 73.33]	67.20 [65.89, 68.49]	0.00 [0, 0.08]	93.20
G4-12	PointMax	0.00 [0, 0.08]	0.00 [0, 0.08]	0.00 [0, 0.08]	—
	UniformMean	26.00 [24.80, 27.23]	9.46 [8.68, 10.30]	0.00 [0, 0.08]	36.38
	UCB-Mean	33.32 [32.03, 34.64]	12.84 [11.94, 13.80]	0.00 [0, 0.08]	38.54

Each row contains 5,000 episodes at $B = 416$, $n = 1{,}024$, and $\delta = 0.05$. Search maxima and RawCandidate rates equal 1.000 and 100%, respectively, in every row; ReuseLCB accepts no episode.

Table 6: Fold-localized Q36-27 budget results. Each Fold 2 entry is *true-witness count / FreshLCB count* among 1,000 episodes. Across folds 0, 1, 3, and 4, both counts are zero for every policy and budget (12,000 policy-episodes per budget).

B	Fold 2: true / FreshLCB			Other folds, all policies
	PointMax	UniformMean	UCB-Mean	True / FreshLCB
26	35/6	32/6	32/4	0/0
52	15/2	59/4	66/8	0/0
104	3/0	93/17	124/25	0/0
208	0/0	142/24	188/35	0/0
416	0/0	226/50	275/57	0/0
832	0/0	327/64	414/89	0/0

B.4 E2 CONTROLLED CALIBRATION

E2 has 26 regimes. In regime z, independent Bernoulli variables $F \sim \mathrm{Bern}(p_z)$ and latent $Z \sim \mathrm{Bern}(q_z)$ produce $G = FZ$, so $\mu_z = p_z q_z$ and $M = 1$. The thresholds are $\epsilon = 0.10$, $\rho = 0.10$, and $\alpha = 0.05$. Table 7 defines the four worlds. W3 uses stable regime IDs $0, 1, 2, 3$. The primary cell uses UCB-Mean, $B = 416$, $n = 1{,}024$, and 5,000 episodes.

For the margin grid, every regime uses $p = 0.90 + \Delta_p$ and $q = (0.10 + \Delta_\mu)/p$. This changes the feasibility and discrepancy margins independently in the population, rather than estimating those margins from samples.

Table 1 in the main text reports the primary-cell false-acceptance and simultaneous-coverage values. The full grids above separate null calibration, margin-dependent acceptance, and budget allocation.

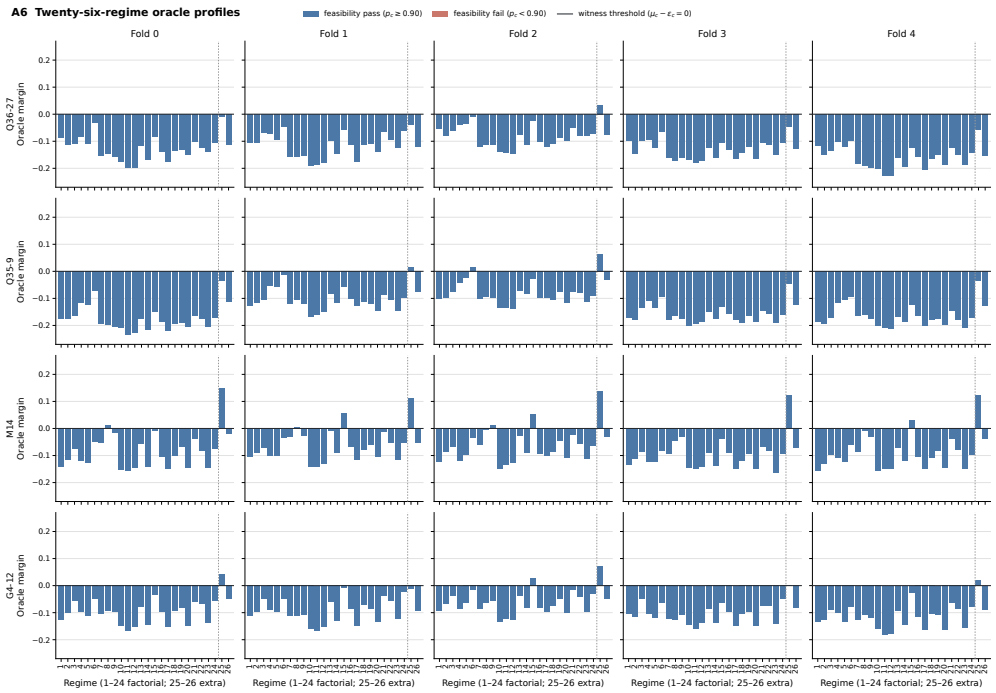


Figure 9: Population margins for all 26 regimes in each of the 20 E1 target–fold contracts. A positive margin means that the regime is a population witness; all E1 regime cells satisfy feasibility.

Table 7: Frozen E2 generators. W3 has four stable regimes and 22 W1 regimes; the other worlds assign identical parameters to all 26 regimes.

World / regime	p_z	q_z	$\mu_z = p_z q_z$	Population witness
W0 boundary	1.00	0.10	0.1000	No: strict residual boundary
W1 spiky	1.00	0.08	0.0800	No: below residual threshold
W2 feasibility	0.89	0.30	0.2670	No: feasibility fails
W3 stable	0.99	0.25	0.2475	Yes: both conditions pass
W3 other	1.00	0.08	0.0800	No: same as W1

Eq. (27) is sufficient, not necessary, so acceptance outside the outlined cells does not contradict the theorem.

B.5 E1 POST-SELECTION RERUN PROVENANCE

Layer B performs one live rerun for every Q36-27 fold–policy pair. All 15 candidate selections are sealed before validation seeds are created. Table 8 records the candidate hash, sampler hash, and difference between replayed and rerun residuals. Every rerun verdict is negative. These runs therefore check execution order and numerical replay, not the stability of positive witnesses.

The replay code derives validation seeds from a public fixed root seed and the candidate hash. This makes the run reproducible, but it does not guarantee conditional independence against a policy that can predict those seeds. Theorem 1 therefore keeps the explicit fresh-randomness assumption. In the archived live execution, 33 of 15,360 cases with $F = 0$ were still sent to the endpoints, and 11 returned nonzero raw Z. Their product FZ remains zero. This is an execution mismatch with the pre-query skip rule defined in Section 2.2; it is not evidence that the skip rule was enforced in those reruns.

Policy and population-label scope. PointMax breaks saturated ties by choosing the smallest regime ID. Small IDs are poor choices in E1 but include the positive regimes in E2 W3. Its ranking across

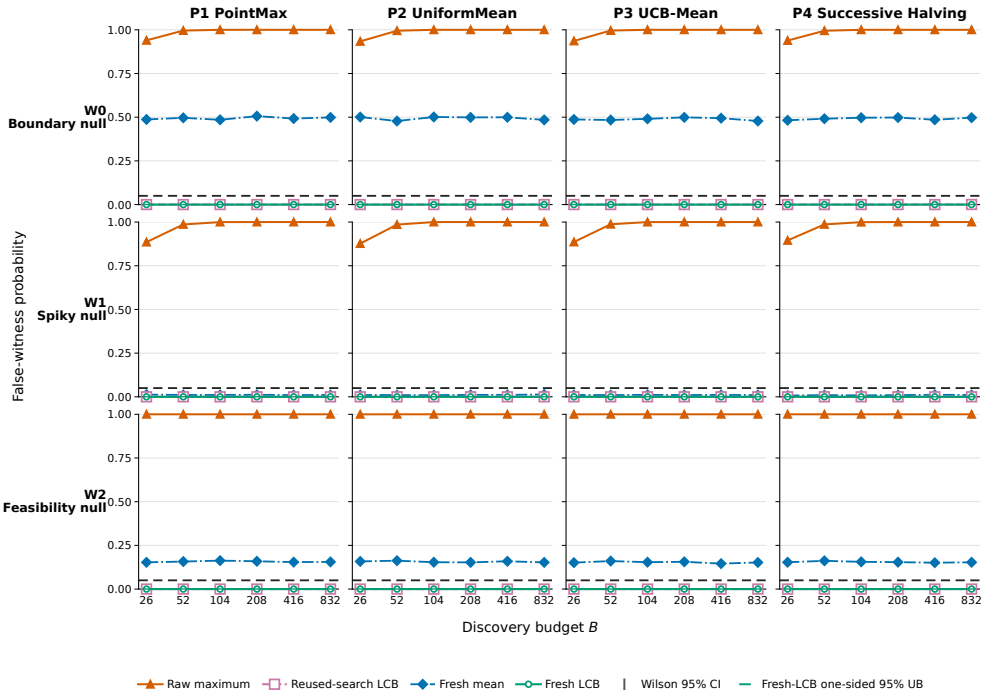


Figure 10: E2 null calibration. Rows are the boundary, rare-large-discrepancy, and feasibility nulls; columns are the four search policies. Discovery budget changes along the horizontal axis at fixed validation size. Raw maxima have high false-acceptance rates, and a fresh mean remains invalid at the boundary null. FreshLCB stays below the nominal level in the tested cells.

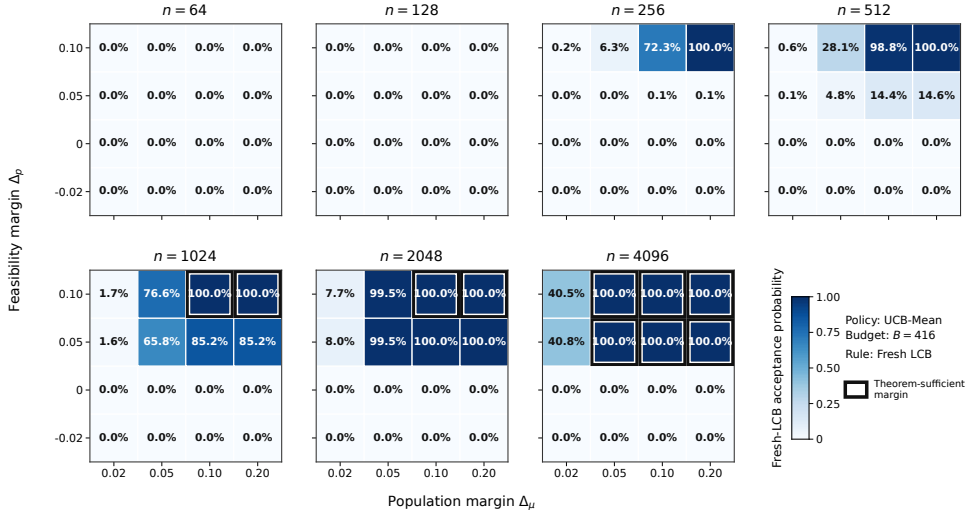


Figure 11: E2 FreshLCB acceptance under UCB-Mean at $B = 416$. Each panel fixes validation size; columns vary the discrepancy margin and rows vary the feasibility margin. Cell values are acceptance percentages. The $n = 64$ and $n = 128$ panels are zero throughout, consistent with the feasibility requirement $n \geq 185$. White outlines mark cells that satisfy the sufficient margin condition in Eq. (27).

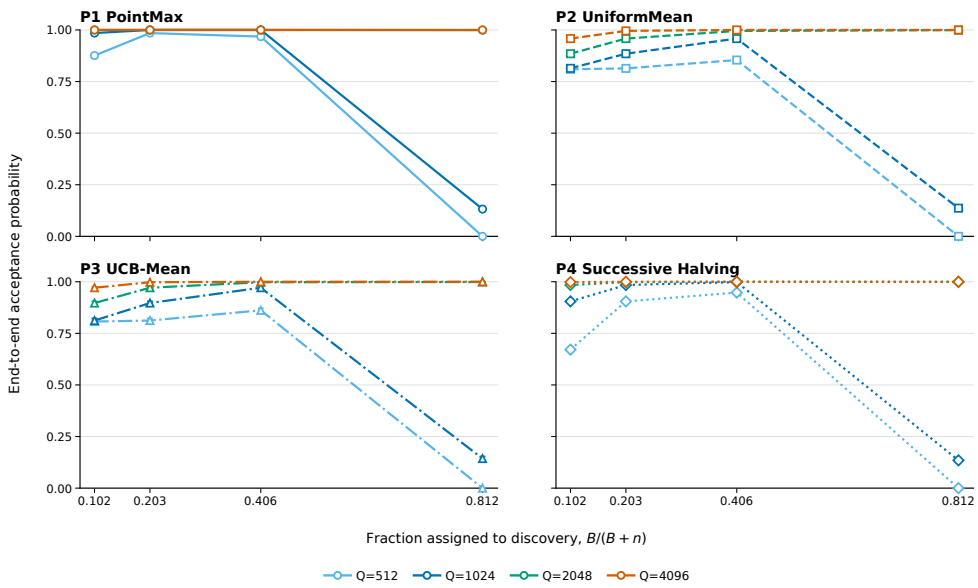


Figure 12: E2 certificate yield as the total budget is split between discovery and validation for the four controlled search policies.

Table 8: Post-selection rerun provenance: fold, eight-character hash prefixes, and absolute replay–rerun residual difference.

PointMax (P1)

F	Candidate	Sampler	$\lvert\Delta r\rvert$
0	`259406ec`	`9a6ba1cf`	0.001484
1	`3e87e7c4`	`18727628`	0.004169
2	`d54172bd`	`b71d269c`	0.000872
3	`2ee5e1e9`	`1686b123`	0.001281
4	`a31f572d`	`c9c93ae2`	0.000621

Seal<seed: 5/5; FM: 0/0 (5/5); FL: 0/0 (5/5).
Mean/max $\lvert\Delta r\rvert$: 0.001685/0.004169.

UniformMean (P2)

F	Candidate	Sampler	$\lvert\Delta r\rvert$
0	`de8f6cce`	`a5c02d81`	0.004262
1	`626406e8`	`26994db6`	0.002797
2	`fe8097ea`	`213d0b9b`	0.002195
3	`603d64b8`	`2244e0a4`	0.002521
4	`4b8f59c8`	`81327bbd`	0.000054

Seal<seed: 5/5; FM: 0/0 (5/5); FL: 0/0 (5/5).
Mean/max $\lvert\Delta r\rvert$: 0.002366/0.004262.

UCB-Mean (P3)

F	Candidate	Sampler	$\lvert\Delta r\rvert$
0	`e490e958`	`b1f6c90e`	0.015598
1	`99cf7e0a`	`934cb965`	0.001236
2	`3f723b10`	`d0a222de`	0.004391
3	`a58abedf`	`950f8d3c`	0.003323
4	`7d685b98`	`450120e0`	0.000917

Seal<seed: 5/5; FM: 0/0 (5/5); FL: 0/0 (5/5).
Mean/max $\lvert\Delta r\rvert$: 0.005093/0.015598.

FM and FL denote FreshMean and FreshLCB replay/rerun verdicts; the parenthesized count is agreement. In every row, $|\Delta\,\mathrm{LCB}_\mu| = |\Delta r|$ and $|\Delta\,\mathrm{LCB}_p| = 0$. Thus all 15 chronology checks and verdict comparisons pass, but every verdict pair is negative. Complete hashes and timestamps remain in the vendored CSV.

worlds is therefore sensitive to regime numbering, so the main policy comparison uses UniformMean and UCB-Mean. The W3 allocation plots use the fixed regime numbering of that experiment and should not be interpreted as rankings invariant to ID permutations. E1 population labels are computed on the fixed sealed population. Shared contracts and replay prefixes limit how far episode-level rates can be generalized.

B.6 E3 SCOPE AND EXACT COVERAGE

Natural regime families. E3 reuses the E1 sealed population means and feasibility probabilities at the original thresholds. The 24-regime core is $4 \times 3 \times 2$ over return format, documentation, and tool-name tags. The declared baseline is zero-based regime ID 5, with OFAT set $\{2, 3, 4, 5, 11, 17, 23\}$. The evaluated 26-regime family adds prompt and style regimes R25 and R26; its nine-query OFAT set adds the same two regimes. The 26-regime set is not a full factorial over five axes. No new model calls are used.

Released data. The accompanying source contains the plot tables, standalone figure sources, and extracted result tables used to check the reported numbers. The E3 natural analysis uses sufficient-statistic tables; the 520 original oracle JSON files are not included.

Table 9: E3 natural scope results. Entries are maximum feasible-discrepancy margins $\max_z \mu_C(z) - \epsilon_C$ within each query family. All cells meet feasibility. Positive entries therefore indicate a population witness. Bold OFAT entries mark missed factorial witnesses.

Target	Fold	Full 24	OFAT 7	Full 26	OFAT 9
Q36-27	0	-0.0332	-0.0332	-0.0085	-0.0085
Q36-27	1	-0.0458	-0.0458	-0.0375	-0.0375
Q36-27	2	-0.0095	-0.0095	+0.0338	+0.0338
Q36-27	3	-0.0662	-0.0662	-0.0463	-0.0463
Q36-27	4	-0.0983	-0.0983	-0.0560	-0.0560
Q35-9	0	-0.0715	-0.0715	-0.0370	-0.0370
Q35-9	1	-0.0140	-0.0140	+0.0167	+0.0167
Q35-9	2	+0.0163	+0.0163	+0.0631	+0.0631
Q35-9	3	-0.0945	-0.0945	-0.0450	-0.0450
Q35-9	4	-0.0934	-0.0934	-0.0339	-0.0339
M14	0	+0.0115	**-0.0493**	+0.1496	+0.1496
M14	1	+0.0569	**-0.0368**	+0.1114	+0.1114
M14	2	+0.0516	**-0.0348**	+0.1399	+0.1399
M14	3	-0.0319	-0.0816	+0.1241	+0.1241
M14	4	+0.0329	**-0.0622**	+0.1221	+0.1221
G4-12	0	-0.0329	-0.0507	+0.0441	+0.0441
G4-12	1	-0.0093	-0.0233	-0.0093	-0.0138
G4-12	2	+0.0269	**-0.0170**	+0.0722	+0.0722
G4-12	3	-0.0498	-0.0498	+0.0010	+0.0010
G4-12	4	-0.0251	-0.0788	+0.0195	+0.0195