WHEN TO SEARCH, CONTINUE, OR FALL BACK: RISK-CONTROLLED ROUTING FOR RARE-REGIME DISCOVERY

Anonymous authors
Paper under double-blind review

ABSTRACT

Scientific rare-regime discovery requires choosing a search method under limited information, uncertain performance, and costly evaluations. We formulate this choice as a sequential replacement decision: when should a learned selector replace an executable fallback, pay for another observation, or fall back? QC-PHAST Router combines a permutation-invariant history encoder, route-conditioned survival predictor, and source-bank-group calibration to assess fallback-relative p90 loss. Information access determines admissible solvers, observations are charged, and proposed hits require independent verification. Finite-stage protection assumes within-track exchangeability; latent-risk protection additionally requires joint inner-interval coverage. Neither statement guarantees lower uncapped cost or protection under family shift. A frozen external evaluation records 1,152 configurations and 147,456 paired trials across 6 information tracks. The recorded procedure passes 0 of 3 primary replacement tests and makes 0 specialist commitments. A post hoc diagnosis finds that calibration corrections exceed the largest expressible prediction gap in 3 tracks. A subsequent audit identifies inherited-history errors in continuation evidence and related deployment paths, limiting interpretation of the recorded cost comparisons. We retain these records without claiming corrected routing efficacy. The study identifies two prerequisites for useful protection: enough decision resolution to authorize a route, and faithful accounting of the observations on which that decision depends.

1 INTRODUCTION

Consider a FitzHugh–Nagumo simulator. A scientist may seek one parameter at which a stable rest state gives way to sustained oscillation. If the equations and derivatives are available, a continuation method can follow the stability boundary. If only queried criticality scores are available, an active-search policy must infer where that boundary lies. If the interface returns only a yes/no threshold predicate over a fixed library, those equation-based and scalar-score procedures are unavailable. The scientific event is unchanged, but the admissible search problem is not. Appendix Figure 4 shows the parameter-to-dynamics interpretation.

Simulator calls and scientific verification are expensive. Machine learning must decide both what to evaluate next and which search mechanism deserves the remaining budget.

Scientific campaigns need dependable discovery, not only a favorable average over easy tasks. A selector that occasionally consumes the entire budget may be worse than a slower, restartable fallback. We learn the 90th percentile of bounded, budget-normalized discovery loss, assigning no-hit outcomes maximum loss. Whether protection of that target yields lower uncapped cost is a separate empirical question.

We formulate selection as a *replacement decision*: is there enough evidence to replace a procedure that can run now? An uncertain ranking need not trigger an expensive commitment. The router can buy another observation or use the fallback, charging every observation used to choose. Noise, an absent target, and overhead can erase specialization's benefit.

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.

Two certificates remain distinct. A *protected route* clears a calibrated fallback-relative comparison; interpreting it as latent p90 improvement requires the joint-coverage assumptions in Section 5. A *verified event* has passed the domain-specific dynamical or stochastic checks in V. Route protection never substitutes for scientific verification.

The study also asks what prevents a carefully protected selector from ever acting. We distinguish three questions: are there enough independent groups to calibrate, can a predicted improvement clear the resulting correction, and do the continuation labels preserve the history being conditioned on? More trials do not answer all three.

Prior Quantum–Classical Phase-space and Stability-Threshold Search (QC-PHAST) work provides a static finite-pool regime map (Tirhekar & Bajaj, 2026). Here the learning problem is sequential: use a charged pilot to select, acquire, or fall back under censored first-hit costs. This is not a new search primitive. An optional quantum marked-set comparison remains an appendix diagnostic, not the router's central contribution.

Our contributions are:

1. **A fallback-relative target.** Learn p90 budget-normalized discovery loss, with explicit observation, computation, failure, and verification charges, rather than solver identity.
2. **A sequential replacement rule.** Separate calibrated route comparisons from policy-matched acquisition utility, with finite-stage guarantees and explicit assumptions for latent-risk protection.
3. **A held-out test and failure diagnosis.** The completed six-track evaluation records no specialist commitments. On three tracks, the correction exceeds every improvement the frozen head can express. A history audit limits the cost comparisons: these are preserved implementation records, not a corrected efficacy evaluation.

2 Related Work and Positioning

Active search seeks qualifying candidates rather than a global optimum (Garnett et al., 2012). We select among search procedures subject to information access, charged observations, and independent scientific verification.

The closest precursors address important parts of this problem. Run2Survive selects algorithms from censored runtime predictions (Tornede et al., 2020); reliable design selection can return no algorithm when evidence is insufficient (Fannjiang & Park, 2025). Budgeted conformal evidence acquisition already combines act, acquire, and abstain decisions and matches calibration to the frozen acquisition policy (Xu et al., 2026). Neither that action set nor policy-matched calibration is independently new here. Our specialization asks when evidence justifies replacing a runnable scientific-search fallback on bounded tail loss, and when a further charged observation has positive lower decision utility. It does not certify realized uncapped savings or coverage under arbitrary family shift.

Decision-aware conformal prediction already links uncertainty to risk-averse actions (Kiyani et al., 2025); calibrated policy post-processing can preserve a baseline subject to a violation-risk constraint (Joshi et al., 2026). We instead study conditional first-hit tail loss and paid observations before solver commitment. The empirical question is whether this specialization earns its cost over simpler selection and portfolio execution.

A learned selector must earn its overhead against a runnable portfolio, not merely predict the best isolated solver (Kashgarani & Kotthoff, 2021). Our finite-stage guarantees are narrower than anytime-valid risk control (Xu et al., 2024; Blot et al., 2025). Appendix A retains the detailed comparisons with optimization, landscape learning, shifted-task selection, and data-driven transition detection.

3 Problem Setting and Cost Accounting

Let a task be

$$\mathcal{T} = (\Theta, \mathcal{X}, s, \tau, V, \mathcal{A}, c, B), \tag{1}$$

where Θ is the parameter domain and $\mathcal{X} \subseteq \Theta$ is an optional finite candidate library. Finite-pool routes require $\mathcal{X}$; a continuous-only route may propose directly in Θ without one. The remaining

Table 1: Principal access families. Reduced-fidelity and noisy scalar/predicate variants remain separately labeled experimental contracts. A route is unavailable, rather than softly downweighted, when its required information is absent.

Track	Observable information	Available route families
Structured	equations, derivatives, off-pool proposals	direct certificates, root finding, continuation, continuous optimization
Scalar finite pool	queried $s(x)$ and exact verification	active/level-set search, portfolios, finite-pool coverage
Predicate finite pool	queried $O_\tau(x)$ only	random/space-filling orders and predicate-only finite-pool policies

objects are a simulator-derived score s, a scientific threshold τ fixed before test-task search, an independent verifier V, $\mathcal{A}$ the information interface, c the cost ledger, and B the total-cost budget. The verifier returns binary success, a reason code, and, for a positive certificate, an evidence level; the task freezes the minimum accepted level. A reduced-fidelity, surrogate, or noisy-predicate positive is a proposal until verified. If no qualifying candidate exists, exhausting a registered stopping rule is a valid censored outcome, not a mislabeled failure.

For route k, the ledger records

$$C_k = C_{\text{pilot}} + C_{\text{train}} + C_{\text{policy}} + C_{\text{sim}} + C_{\text{prep}} + C_{\text{oracle}} + C_{\text{EC}} + C_{\text{verify}}. \tag{2}$$

We use bounded loss

$$L_k = \begin{cases} \min\{C_k/B, 1\}, & \text{independently certified hit,} \\ 1, & \text{no hit, censoring, or scientifically invalid return.} \end{cases} \tag{3}$$

Infrastructure failure invalidates the run and is never converted into a favorable scientific label. Equation (2) defines the accounting categories, not a claim that the current experiment measures every offline expense. We report deployment cost and a separate amortized source-bank construction lower bound; the latter is not complete end-to-end cost.

Offline controlled replay versus online deployment. Offline replay hides precomputed scores behind the declared interface; it is *not* online simulator savings. Online deployment evaluates lazily against a prespecified threshold or one estimated with a fully charged pilot. Both retain empty-set and unsuccessful outcomes.

The search variable is a point in *parameter space*; each point induces dynamics in *state phase space*. We use "parameter-to-phase-space bridge" for this map and never infer intrinsic marked-set geometry from a low-dimensional projection alone.

What a scientist receives. The user supplies an event, verifier, budget, and simulator interface. The router returns a protected route comparison, a charged next observation, or the executable fallback. It is a decision layer over solvers, not a universal optimizer.

Route specification and statistical unit. Each registered route binds its access requirements, candidate generator, warm-start rule, stopping rule, cost units, verification rule, and acquisition-policy digest. A route without the required access is unavailable rather than assigned a low learned score. The source bank, not an individual configuration sampled from it, is the exchangeability unit for calibration and hierarchical inference; repeated pilot stages and stochastic continuations are nested observations, not independent scientific tasks.

Executable fallback. The fallback is fixed by access, not selected retrospectively from task outcomes. Structured tasks use a restartable structured racing portfolio; remaining finite-pool tracks use the registered hedged portfolio. The fallback map is hashed with the frozen experiment configuration, and every mapped fallback must be executable under its information mask. Calibration is stratified by track; missing, inconsistent, or inadmissible mappings fail closed.

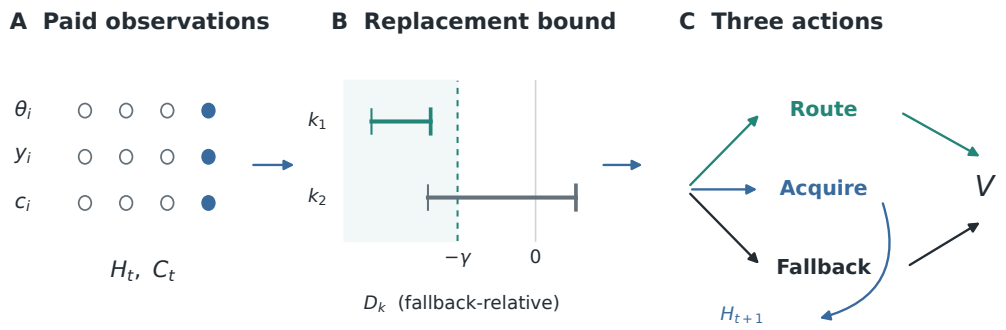


Figure 1: Protected decision at stage t. Access filtering defines the admissible route set $\mathcal{K}_a$, and the charged history yields support, fallback-relative risk, and policy-matched acquisition statistics. The router commits only if $\min_k \sup I^D_{a,k,t} \le -\gamma$, acquires one observation only if its calibrated lower value is positive, support is accepted, and the next observation fits the remaining allowance; otherwise it executes f_a. Every returned candidate faces the same independent verifier V. The intervals illustrate the rule, not fitted predictions or performance; k_1 clears the replacement threshold, whereas k_2 does not.

Algorithm 1 Fallback-relative route, continue, or fallback decision

Require: task $\mathcal{T}$, charged history H_t, frozen ensemble and calibrations, margin γ, pilot cap ρB

Remove every route that violates the task's information interface or lacks a tested executor.
if H_t already contains an independently verified hit **then**
    return that hit.
if the task is outside calibrated support **then**
    return the executable fallback f_a.
Predict conditional p90 loss and fallback-relative intervals $I^D_{a,k,t}$.
if some admissible alternative has $\sup I^D_{a,k,t} \le -\gamma$ **then**
    return the lowest predicted-risk protected route k.
if a registered stage remains and an eligible candidate fits the remaining pilot allowance **then**
    if the frozen ranking selects a candidate with $\mathrm{CVI}^-(x \mid H_t) > 0$ **then**
        Query it, charge the observation, verify any proposed hit, update H_t, and repeat.
return the executable fallback f_a.

4 QC-PHAST Router

Given a charged history H_t, the router returns one of three decisions: execute an admissible route, acquire one more candidate, or run the executable fallback. Figure 1 shows the decision order. A solver is considered only when its required information and a tested implementation are available.

Algorithmic rule. Algorithm 1 makes the order explicit. Acquisition cannot bypass support, admissibility, or pilot-cost gates.

Conditional tail-risk model. A permutation-invariant pilot encoder and route-conditioned survival heads estimate

$$R_{k,t} = Q_{0.9}(L_k \mid H_t, \mathcal{A}). \tag{4}$$

Independently seeded ensemble members predict route survival, candidate outcome distributions, and candidate p90 evaluation cost. The intended continuation target requires the same charged history for every route, with structured warm starts restricted to those observations. The historical executor did not fully preserve this state (Appendix D.10). Appendix D.1 specifies the representation, likelihood, and training supervision.

Fallback-relative calibration. Let a_i be task i's information track and f_{a_i} its executable fallback index. Define

$$D_{i,k,t} = R_{i,k,t} - R_{i,f_{a_i},t}, \qquad \widehat{D}_{i,k,t} = \widehat{R}_{i,k,t} - \widehat{R}_{i,f_{a_i},t}. \tag{5}$$

Repeated continuations provide route p90 intervals $[\ell_{i,k,t}, u_{i,k,t}]$, inducing $[d^-_{i,k,t}, d^+_{i,k,t}] = [\ell_{i,k,t} - u_{i,f_{a_i},t}, u_{i,k,t} - \ell_{i,f_{a_i},t}]$ for $k \neq f_{a_i}$; the fallback's self-comparison is exactly $[0, 0]$. Within access track a, calibrate the source-bank maximum

$$A_{g,a} = \max_{i \in g: a_i = a,\, k \in \mathcal{K}_a, t} \max\{\widehat{D}_{i,k,t} - d^-_{i,k,t}, d^+_{i,k,t} - \widehat{D}_{i,k,t}, 0\}. \tag{6}$$

The split-conformal order statistic $q_{a,1-\alpha}$ yields $I^D_{a,k,t} = [\widehat{D}_{k,t} - q_{a,1-\alpha}, \widehat{D}_{k,t} + q_{a,1-\alpha}]$ for alternatives, with $I^D_{a,f_a,t} = \{0\}$. If the requested rank exceeds the number of calibration groups, certification fails closed. Heterogeneous access tracks are reported separately and are not pooled as exchangeable samples.

Continue only for calibrated acquisition utility. For unobserved x, each ensemble member estimates the reduction in protected p90 risk after observing x, minus its p90 charge. Let $\widehat{G}^-_t(x)$ be the ensemble's raw lower gain estimate and let x^π_t be the candidate selected by the frozen ranking policy π. On an offline calibration source bank, paired continuations define the empirical utility $G^\pi_t(x^\pi_t)$ for that selected action. Concretely, this target is the paired empirical quantity $\widetilde{R}^{\mathrm{prot}}(H_t) - c(x^\pi_t)/B - \widetilde{R}^{\mathrm{prot}}(H_{t+1})$, with future risk set to zero only when the queried candidate independently verifies a terminal hit. The source-bank score is

$$B^G_g = \max_{t \in \mathcal{S}_g} \left[\widehat{G}^-_t(x^\pi_t) - G^\pi_t(x^\pi_t)\right]_+, \tag{7}$$

and its split-conformal quantile $q^G_{a,1-\alpha_G}$ gives

$$\mathrm{CVI}^-(x \mid H_t) = \widehat{G}^-_t(x) - q^G_{a,1-\alpha_G}. \tag{8}$$

Candidate ranking uses a frozen provisional route radius of zero, not the subsequently fitted route-calibration radius. The gain correction is common across candidates, so it does not change their ranking. Calibration trajectories follow the frozen ranking to the pilot cap; deployment may stop earlier, hence every deployed history is a prefix of a calibrated policy path. This construction protects actions selected by π across registered stages. It does not claim simultaneous validity for unselected candidates unless their counterfactual gains are also evaluated. The acquisition digest is bound through trajectories, certificates, checkpoints, and the test-opening manifest; off-policy histories cannot authorize deployment. This is an auxiliary utility, not an exact net-cost difference: future continuation totals already include the queried observation, while terminal credit sets future risk to zero. Its certificate concerns that registered empirical target; all-in savings require the separate held-out cost evaluation.

Support abstention and pilot cap. A shrinkage-Mahalanobis score in the frozen task embedding defines a conformal support p-value. Within each access track, location and precision use training-bank embeddings. Calibration takes the maximum score over all repeated rows from each calibration source bank, so the conformal rank contains one score per exchangeability group rather than one score per pilot row. Unsupported tasks route to fallback. The observation allowance is ρB, with ρ and γ selected on development families and then frozen. Known per-query charges enforce this allowance in the replay. Predicted p90 charges additionally screen candidates; a prediction alone cannot enforce a hard realized-cost limit for a variable-cost simulator. The total-pilot bound below also requires a bound on all pilot-side overhead, including inference and verification.

5 Protection Guarantees

The results below are finite-stage, in-distribution group statements. They are not anytime confidence sequences and do not imply formal coverage under an unseen-family shift. Proofs are given in Appendix B.

Proposition 1 (Registered-stage fallback-relative coverage)**.** *Fix the complete pilot policy π, including its candidate ranking and tie breaking. Within one information track, if calibration source-bank groups and a new source-bank group are exchangeable, the interval induced by the score in equation 6 simultaneously contains all registered continuation intervals for $D_{k,t}$ with probability at least $1-\alpha$. If those inner intervals jointly cover their latent p90 differences with probability at least $1-\beta$, then the router intervals cover the latent differences with probability at least $1-\alpha-\beta$.*

Pointwise continuation intervals do not alone establish joint $1-\beta$ coverage; the latent-risk statement requires a simultaneous inner construction or a valid multiplicity allowance. The implementation uses pointwise continuation intervals, so its outer calibration level alone must not be read as a verified joint latent-risk confidence level.

Proposition 2 (Protected routing and bounded pilot harm)**.** *On simultaneous within-track coverage, if route $\hat{k}$ is selected only when $\sup I^{D}_{a,\hat{k},t} \le -\gamma$, then $R_{\hat{k},t} + \gamma \le R_{f_a,t}$. If no route is selected, a fresh fallback restart on the fixed task, independent of pilot randomness, after a pilot of cost $C_t \le \rho B$ has normalized p90 loss at most $\min\{1, R^{\mathrm{start}}_{f_a} + \rho\}$.*

Proposition 3 (One-sided registered-policy acquisition protection)**.** *Within one information track, assume calibration groups and a new group are exchangeable and all histories are generated by the same frozen policy π. The quantile of the group score in equation 7 lower-bounds the empirical utility of the actions selected by π simultaneously over registered stages of a new group with probability at least $1-\alpha_G$. Therefore, $\mathrm{CVI}^{-}(x^{\pi}_{t} \mid H_t) > 0$ excludes nonpositive utility on that event, not nonpositive realized all-in savings.*

The route and acquisition statements concern different events. Even within the same exchangeable track, their separate error levels are not one whole-policy guarantee. With the additional inner-coverage assumption, a union bound gives joint route-risk and empirical-utility coverage of at least $1-\alpha-\beta-\alpha_G$; no independence is required. At the selected margin $\gamma = 0$, protected routing means non-inferiority on that event, not a guaranteed strictly positive saving.

6 Evaluation: Can the Router Earn Replacement?

Development, selection, and calibration split by source bank and finish before external evaluation. The development design pairs six tracks on 192 banks from twelve families: structured exact, scalar exact, predicate exact, reduced-fidelity, scalar noisy, and predicate noisy. One outcome-blind threshold/view slot per bank gives 1,152 configurations, without score-based task selection. The external design and analysis are registered before outcomes are read.

The registered confirmatory task design uses four sealed simulator families: Hodgkin–Huxley firing onset, CSTR equilibrium stability, swing-network synchrony, and Gray–Scott pattern onset. Eight banks per family, three fixed scientific thresholds, two coordinate views, and the same six tracks give 1,152 external configurations with 128 paired trials each, or 147,456 pairs per method. Three exact tracks are primary; reduced-fidelity and noisy tracks test robustness. No quantum arm is included. Appendix D.8 specifies the parameter domains, fixed thresholds, and verification.

Systems and certificates. Development systems include analytic controls and trajectory-level models: spring, pendulum, Van der Pol, Lorenz, FitzHugh–Nagumo (FHN), coupled FHN, Duffing, Brusselator, Hindmarsh–Rose, Morris–Lecar, Kuramoto, and a stochastic toggle. Evidence levels are reported separately: score-threshold hits, robust local crossings, and classified transitions. A classified Hopf-type event requires the appropriate eigenpair, nonzero frequency, transversality, and before/after dynamics; simple-zero and stochastic events use their own nondegeneracy and repeated-simulation contracts. Direct analytic construction is an expected winner where equations permit it.

Baselines. Comparators are the access-matched restartable portfolio, best fixed route, source-bank-balanced nearest-neighbor selector, random-forest p90 selector, and calibrated abstention. Search arms include coverage, cross-entropy/subset-style policies, bounded-GP active search, direct certification, differential evolution, and a deterministic cascade when admissible. Exact-refit GPs remain small-pool references; the forest selector is not Run2Survive. Tuning uses development validation banks only. Best-fixed buys no pilot; learned selectors share the charged pilot interface. Construction is reported separately and under frozen reuse horizons.

Table 2: What the evidence supports. Artifact integrity is not a guarantee of semantic correctness.

Evidence	Interpretation
Recorded actions	No specialist commitments in the frozen implementation; not proof that no useful specialist exists.
Prediction-range bound	Explains infeasibility for the frozen head and corrections; does not require valid calibration.
Costs and ablations	History-sensitive implementation outputs; external numerical impact of the defect remains unquantified.
Development repair	Tested history-capture components; not corrected training, calibration, or external efficacy.

The registered component ablations are development-validation decision replays: full, no route correction, no support abstention, no gain correction, and no continuation. They test which safeguards alter decisions on development tasks. The external study compares the frozen full router with deployable baselines; it does not establish held-out-family effects for each component separately.

Metrics and inference. For deployment cost C, external evaluation uses $\widetilde{C} = C$ for a certified success and $\widetilde{C} = \max\{C, B\}$ otherwise, retaining cost overruns. We average per-task $Q_{0.9}(\widetilde{C})$ within families, then weight families equally; this is not a pooled p90. This external cost is not the bounded loss L used for calibration. Secondary diagnostics include mean cost at or above the empirical p90, success by a fixed budget, regret to the restartable fallback, abstention, and simultaneous calibration coverage. Comparisons are paired by task and stochastic trial. Hierarchical intervals resample simulator family, source bank, task, and trial. We report deployment costs and costs adjusted by an amortized source-bank construction lower bound, not a complete offline training bill (Appendix D.6). Comparisons against the fallback add that charge *after* the failure penalty; the saved direct router–selector diagnostics instead apply the penalty to the already adjusted ledger. These are different endpoints on early unsuccessful searches and must not be mixed into one ranking (Appendix D.12). Holm correction applies to the three primary fallback-relative tests; deployable-selector comparisons are prespecified secondary analyses. We report failures and empty sets, and discard no unsuccessful configuration. The frozen implementation names its exceedance-mean field `cvar90`; we retain its values and selection rule, but do not identify it with standard CVaR when losses have atoms or ties. If failure probability exceeds 10%, $Q_{0.9}(L) = 1$, but the external p90 may exceed B. The bounded-risk certificate therefore does not by itself guarantee lower uncapped external cost. Success-by-budget and censoring accompany the primary endpoint; they do not replace it after outcomes are observed.

Recorded behavior and its limits. The frozen study records 0 specialist commitments in 147,456 router trials. There are 28,649 terminal pilot hits and 118,807 terminal fallback decisions. Only 84 trials acquire an additional observation; these overlap the terminal categories and are not extra trials. The recorded replacement tests pass on 0/3 primary tracks. Table 2 separates these observations from affected cost comparisons. The original numerical rows remain in Table 5; they are historical implementation outputs, not validated efficacy estimates.

What the result establishes. The records establish non-specialization of the frozen implementation, not that useful routes do not exist. History-dependent costs cannot establish the intended method's relative efficiency until the affected execution paths are assessed. No commitments means undefined conditional coverage, not perfect protection. Four held-out families remain four families despite the many paired trials. No external retuning is performed.

Why the same indices are not the same history. An unsuccessful pilot includes both purchased observations and rejected verification attempts. Replaying its indices with new random seeds can turn an earlier rejection into a success and stop before all already-paid work is charged. A development audit detects this mechanism, and source inspection identifies related handoffs in training, calibration, and deployment (Appendix D.10). Its external numerical impact is unquantified. This prevents an efficacy interpretation of the historical costs; it does not change the saved decision counts or the algebraic bound below.

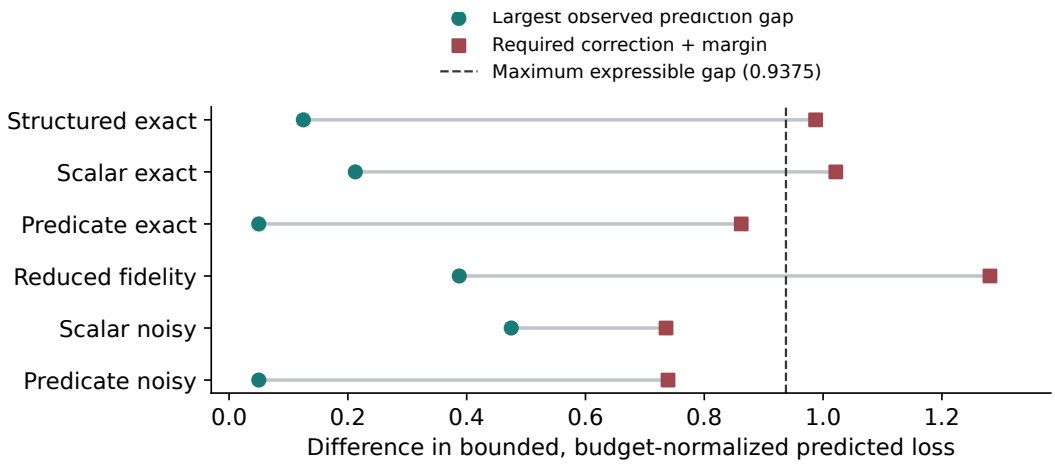


Figure 2: Decision resolution of the frozen router. Circles show the largest admissible predicted fallback-minus-specialist gap over recorded stages; squares show the required correction plus margin. Squares beyond the dashed model-range bound rule out commitment for any input. Elsewhere, observed gaps still fail to reach the required correction. This diagnoses the implemented comparison, not the validity of the calibration data or a repaired policy.

Why no specialist was authorized. The five frozen survival heads use 16 bins: their mean p90 prediction lies in $[1/16, 1]$. Thus no predicted fallback-minus-specialist improvement can exceed $15/16$. The replacement rule requires that improvement to exceed or equal $q_a + \gamma$. With $\gamma = 0$, the final corrections for structured exact, scalar exact, and reduced fidelity already exceed $15/16$, precluding commitment for any input. In the other tracks, every observed admissible prediction gap is below its correction as well (Figure 2). This post hoc check explains non-commitment without assuming that support rejection or external shift alone caused it.

Enough groups did not ensure useful decisions. The initial validation failed 24 of 162 candidate–fold–track group-count checks. Source-disjoint replacement banks raised the smallest eligible count to 31, above the required 9, and all 162 checks passed. The later prediction-range obstruction is a different failure: a finite correction need not permit any action. Appendix E recovers this distinction and its practical consequences. Removing safeguards is not a demonstrated remedy; the history-sensitive development ablations remain qualified in Table 8.

7 Limitations and Conclusion

Formal protection assumes source-bank exchangeability, not arbitrary simulator shift. Four held-out families remain four independent families regardless of trial count. Finite continuations leave p90 uncertainty; finite libraries can omit a regime; a first hit does not reconstruct a transition manifold. Equations may expose targets more cheaply than learned search. The inherited-history audit also reaches deployment source paths; its external cost impact remains unquantified (Appendix D.10).

The frozen implementation does not specialize, and corrected efficacy remains unresolved. Useful protection requires more than a confidence level: enough independent groups, a correction that permits meaningful decisions, and continuations conditioned on the actual paid history. These checks identify different reasons to stop or redesign a study. A future method must also demonstrate lower cost on a new, untouched evaluation. The present test cannot become its tuning set and still serve as confirmation.

Reproducibility Statement

The artifact includes task/split manifests, source-bank provenance, environments, seeds, executors, calibration, raw trials, failures, and analysis/rendering scripts. Independent audits check task coverage, hashes, verification, and numerical lineage. The appendix specifies statistical units, information interfaces, accounting, and the post hoc diagnostic boundary.

AI Use Statement

AI tools assisted with research and analysis, coding, understanding related work, and manuscript preparation. The authors remain responsible for the final content. This assistance included conceptual and mathematical formulation, proof drafting and critique, experimental design, implementation and testing, and interpretation of recorded results.

References

Anastasios N. Angelopoulos, Rina Foygel Barber, and Stephen Bates. Theoretical foundations of conformal prediction, 2024. URL https://arxiv.org/abs/2411.11824. Book manuscript, version 5 revised March 2026.

Vincent Blot, Anastasios Nikolas Angelopoulos, Michael Jordan, and Nicolas J.-B. Brunel. Automatically adaptive conformal risk control. In *Proceedings of the 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pp. 19–27, 2025. URL https://proceedings.mlr.press/v258/blot25a.html.

Michel Boyer, Gilles Brassard, Peter Høyer, and Alain Tapp. Tight bounds on quantum searching. *Fortschritte der Physik*, 46(4–5):493–505, 1998. doi: 10.1002/(SICI)1521-3978(199806)46:4/5⟨493::AID-PROP493⟩3.0.CO;2-P.

Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. Quantum amplitude amplification and estimation. In *Quantum Computation and Information*, volume 305 of *Contemporary Mathematics*, pp. 53–74. American Mathematical Society, 2002. doi: 10.1090/conm/305/05215.

Shachi Deshpande, Charles Marx, and Volodymyr Kuleshov. Online calibrated and conformal prediction improves Bayesian optimization. In *Proceedings of the 27th International Conference on Artificial Intelligence and Statistics*, volume 238 of *Proceedings of Machine Learning Research*, pp. 1450–1458. PMLR, 2024. URL https://proceedings.mlr.press/v238/deshpande24a.html.

David Eriksson, Michael Pearce, Jacob R. Gardner, Ryan Turner, and Matthias Poloczek. Scalable global optimization via local bayesian optimization. In *Advances in Neural Information Processing Systems*, volume 32, 2019. URL https://proceedings.neurips.cc/paper/2019/hash/6c990b7aca7bc7058f5e98ea909e924b-Abstract.html.

Clara Fannjiang and Ji Won Park. Reliable algorithm selection for machine learning-guided design. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 16078–16101, 2025. URL https://proceedings.mlr.press/v267/fannjiang25a.html.

Roman Garnett, Yamuna Krishnamurthy, Xuehan Xiong, Jeff Schneider, and Richard Mann. Bayesian optimal active search and surveying. In *Proceedings of the 29th International Conference on Machine Learning*, 2012. URL https://icml.cc/2012/papers/618.pdf.

Alkis Gotovos, Nathalie Casati, Gregory Hitz, and Andreas Krause. Active learning for level set estimation. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, 2013. URL https://www.ijcai.org/Proceedings/13/Papers/202.pdf.

Alan L. Hodgkin and Andrew F. Huxley. A quantitative description of membrane current and its application to conduction and excitation in nerve. *The Journal of Physiology*, 117(4):500–544, 1952. doi: 10.1113/jphysiol.1952.sp004764.

Liangze Jiang and Damien Teney. OOD-chameleon: Is algorithm selection for OOD generalization learnable? In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 27624–27648, 2025. URL https://proceedings.mlr.press/v267/jiang25g.html.

Sunay Joshi, Tao Wang, Hamed Hassani, and Edgar Dobriban. Risk-controlled post-processing of decision policies, 2026. URL https://arxiv.org/abs/2605.06479. arXiv preprint arXiv:2605.06479.

Haniye Kashgarani and Lars Kotthoff. Is algorithm selection worth it? comparing selecting single algorithms and parallel execution. In *AAAI Workshop on Meta-Learning and MetaDL Challenge*, volume 140 of *Proceedings of Machine Learning Research*, pp. 58–64. PMLR, 2021. URL https://proceedings.mlr.press/v140/kashgarani21a.html.

Shayan Kiyani, George J. Pappas, Aaron Roth, and Hamed Hassani. Decision theoretic foundations for conformal prediction: Optimal uncertainty quantification for risk-averse agents. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30943–30965. PMLR, 2025. URL https://proceedings.mlr.press/v267/kiyani25a.html.

Ana Kostovska, Gjorgjina Cenikj, Diederick Vermetten, Anja Jankovic, Ana Nikolikj, Urban Skvorc, Peter Korosec, Carola Doerr, and Tome Eftimov. PS-AAS: Portfolio selection for automated algorithm selection in black-box optimization. In *Proceedings of the Second International Conference on Automated Machine Learning*, volume 224 of *Proceedings of Machine Learning Research*, pp. 11/1–17, 2023. URL https://proceedings.mlr.press/v224/kostovska23a.html.

Prabha Kundur. *Power System Stability and Control*. McGraw–Hill, 1994.

Benjamin Letham, Phillip Guan, Chase Tymms, Eytan Bakshy, and Michael Shvartsman. Look-ahead acquisition functions for bernoulli level set estimation. In *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 8493–8513, 2022. URL https://proceedings.mlr.press/v151/letham22a.html.

Xuefeng Liu, Fangfang Xia, Rick L. Stevens, and Yuxin Chen. Contextual active model selection. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1581.

Zeyuan Ma, Jiacheng Chen, Hongshu Guo, and Yue-Jiao Gong. Neural exploratory landscape analysis for meta-black-box-optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/27b89ae947228a7195861e3eb307d1a8-Abstract-Conference.html.

Giang Ngo, Dang Nguyen, Dat Phan-Trong, and Sunil Gupta. Active level set estimation for continuous search space with theoretical guarantee. In *Proceedings of the 15th Asian Conference on Machine Learning*, volume 222 of *Proceedings of Machine Learning Research*, pp. 943–958, 2024. URL https://proceedings.mlr.press/v222/ngo24a.html.

Quan Nguyen and Roman Garnett. Nonmyopic multiclass active search with diminishing returns for diverse discovery. In *Proceedings of the 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, pp. 5231–5249. PMLR, 2023. URL https://proceedings.mlr.press/v206/nguyen23d.html.

Oliver O'Brien and Christoph Sünderhauf. Quantum state preparation via piecewise QSVT. *Quantum*, 9:1786, 2025. doi: 10.22331/q-2025-07-03-1786.

John E. Pearson. Complex patterns in a simple system. *Science*, 261(5118):189–192, 1993. doi: 10.1126/science.261.5118.189.

Drew Prinster, Samuel Don Stanton, Anqi Liu, and Suchi Saria. Conformal validity guarantees exist for any data distribution (and how to find them). In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 41086–41118, 2024. URL https://proceedings.mlr.press/v235/prinster24a.html.

R. Tyrrell Rockafellar and Stanislav Uryasev. Conditional value-at-risk for general loss distributions. *Journal of Banking & Finance*, 26(7):1443–1471, 2002. doi: 10.1016/S0378-4266(02)00271-6.

Anindya Sarkar, Binglin Ji, and Yevgeniy Vorobeychik. Active target discovery under uninformative priors: The power of permanent and transient memory. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5d5f4a2f5821c957ff9e4ff14ff37bb9-Abstract-Conference.html.

Felix Schur, Jesus Lago, and Tanner Fiez. The price of knowledge: Optimal algorithms for costly bandits. In *Proceedings of the 42nd Conference on Uncertainty in Artificial Intelligence*, volume 337 of *Proceedings of Machine Learning Research*, pp. 6064–6090. PMLR, 2026. URL https://proceedings.mlr.press/v337/schur26a.html.

Dale E. Seborg, Thomas F. Edgar, Duncan A. Mellichamp, and Francis J. Doyle. *Process Dynamics and Control*. John Wiley & Sons, third edition, 2010.

Shubhanshu Shekhar and Tara Javidi. Multiscale gaussian process level set estimation. In *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pp. 3283–3291, 2019. URL https://proceedings.mlr.press/v89/shekhar19a.html.

Wentao Tang. Data-driven bifurcation analysis via learning of homeomorphism. In *Proceedings of the 6th Annual Learning for Dynamics and Control Conference*, volume 242 of *Proceedings of Machine Learning Research*, pp. 1149–1160, 2024. URL https://proceedings.mlr.press/v242/tang24b.html.

Harsh Milind Tirhekar and Chandrajit Bajaj. QC-PHAST search: Classical–quantum query benchmarks for finite-pool rare-regime discovery, 2026. URL https://arxiv.org/abs/2607.21995.

Alexander Tornede, Marcel Wever, Stefan Werner, Felix Mohr, and Eyke Hüllermeier. Run2Survive: A decision-theoretic approach to algorithm selection based on survival analysis. In *Proceedings of the 12th Asian Conference on Machine Learning*, volume 129 of *Proceedings of Machine Learning Research*, pp. 737–752, 2020. URL https://proceedings.mlr.press/v129/tornede20a.html.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. Conformal thinking: Risk control for reasoning on a compute budget. In *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, 2026. URL https://arxiv.org/abs/2602.03814.

Herbert Woisetschläger, Ryan Zhang, Shiqiang Wang, and Hans Arno Jacobsen. MESS+: Dynamically learned inference-time LLM routing in model zoos with service level guarantees. In *Advances in Neural Information Processing Systems*, volume 38, 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/4dd1d9b841712bd37b833559f041530c-Abstract-Conference.html.

Qian Xie, Raul Astudillo, Peter I. Frazier, Ziv Scully, and Alexander Terenin. Cost-aware bayesian optimization via the pandora's box gittins index. In *Advances in Neural Information Processing Systems*, 2024. doi: 10.52202/079017-3669. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/d14c355d5e88cff437a6303d2d716252-Abstract-Conference.html.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Look again before you abstain: Budgeted conformal evidence acquisition for reliable vision-language model, 2026. URL https://arxiv.org/abs/2606.16667. arXiv:2606.16667v4.

Ziyu Xu, Nikos Karampatziakis, and Paul Mineiro. Active, anytime-valid risk controlling prediction sets. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-1920.

Jifan Zhang, Shuai Shao, Saurabh Verma, and Robert Nowak. Algorithm selection for deep active learning with imbalanced datasets. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi: 10.52202/075280-0422.

A Expanded Related Work

Active search, Bayesian optimization, and scientific discovery. Active search targets discovery of positives rather than global scalar optimization (Garnett et al., 2012); nonmyopic extensions can instead value diversity across multiple rare classes (Nguyen & Garnett, 2023). We study the narrower first-verified-hit objective and do not claim transition-manifold reconstruction. Level-set methods exploit numerical score geometry (Gotovos et al., 2013; Shekhar & Javidi, 2019). Continuous-domain and Bernoulli-response extensions further show that the search domain and observation channel change the admissible acquisition rule (Ngo et al., 2024; Letham et al., 2022). Bayesian optimization supplies the broader sequential decision and value-of-information language, while trust-region and cost-aware variants address high dimension and heterogeneous evaluations (Eriksson et al., 2019; Xie et al., 2024); online calibration can improve that sequential optimization loop (Deshpande et al., 2024), and costly-bandit theory makes observation charge part of the learning objective (Schur et al., 2026). Our objective differs in three coupled respects: success requires an independent scientific certificate, route access differs by task, and the calibrated decision target is fallback-relative p90 bounded discovery loss rather than simple regret. Uncapped cost is a separate test endpoint.

Algorithm selection and learned optimization. Meta-optimization, landscape representation, active model selection, and survival-based algorithm selection demonstrate several ingredients of task-conditioned policy choice (Tornede et al., 2020; Zhang et al., 2023; Liu et al., 2024; Ma et al., 2025). Portfolio composition is itself an algorithm-selection problem (Kostovska et al., 2023), and learned cost-aware routing with service guarantees is established in other domains (Woisetschläger et al., 2025). Reliable selection can also return no algorithm when evidence is insufficient (Fannjiang & Park, 2025), while active target discovery studies sequential discovery under weak priors (Sarkar et al., 2025). We therefore do not claim novelty for a Set Transformer, a survival head, adaptive acquisition selection, portfolio construction, or abstention in isolation. Budgeted conformal evidence acquisition already instantiates an act/acquire/abstain decision and shows that calibration must pass through the same frozen post-acquisition map used at deployment (Xu et al., 2026). Thus neither the three-way action set nor policy-matched acquisition calibration is independently new. Run2Survive models censored runtimes to select an algorithm; reliable design selection certifies that a returned algorithm satisfies a success criterion. Our narrower decision target is to qualify a replacement of a runnable fallback using bounded tail loss under the task's declared information interface, or buy a charged observation when its registered empirical utility has a positive lower bound. This fallback-relative scientific-search specialization is the paper's technical claim. OOD-Chameleon learns algorithm rankings on unseen shifts and datasets from dataset-level characteristics (Jiang & Teney, 2025); our family-held-out evaluation likewise tests transfer, but adds charged within-task evidence, access-conditioned admissibility, a runnable fallback, and fallback-relative tail protection. Learned routing is compared with best-fixed, similarity, random-forest p90, calibrated abstention, and no-learning portfolio/cascade diagnostics. Our fallback is deliberately a runnable portfolio: a learned selector must justify its own overhead against portfolio execution, not only against isolated solvers (Kashgarani & Kotthoff, 2021). Our finite-stage, policy-matched statement is narrower than anytime-valid active risk control and does not claim arbitrary-shift validity (Xu et al., 2024; Blot et al., 2025; Prinster et al., 2024). Risk-controlled adaptive computation also calibrates when an LLM should stop reasoning under a token budget (Wang et al., 2026). Our action and loss are different: the router replaces a runnable scientific solver, buys a physical simulator observation, or executes the fallback, and every route is judged by censored cost to an independently verified event.

Decision-aware calibration and fallback switching. Risk-averse calibration connects prediction sets to downstream decision utility (Kiyani et al., 2025); neither conformal uncertainty nor risk-aware decisions are new here. Risk-controlled policy post-processing maximizes agreement with an existing policy subject to a violation-risk constraint, using calibrated switching to a fallback (Joshi et al., 2026). Our objective instead compares conditional first-hit tail losses and permits paid simulator observations before commitment. This difference motivates the experiment; it does not by itself establish empirical superiority. A faithful survival-based selector is also distinct from the registered random-forest

p90 regressor. The current comparator set should not be read as an exhaustive evaluation of the algorithm-selection literature.

Learning and validating dynamical transitions. Machine learning can detect bifurcations from data even when governing equations are unavailable (Tang, 2024). Accordingly, a small criticality score is not treated as a classified transition. Our highest evidence level requires event-specific nondegeneracy, transversality, and before/after dynamical checks, and structured continuation remains a strong route rather than a weak baseline.

B Proofs of the Protection Statements

Proof of Proposition 1. Split-conformal ranking controls the new group's maximum residual. On that event, every registered route-stage continuation interval lies inside its router interval. Intersecting with joint inner coverage and applying a union bound gives the second statement. □

Proof of Proposition 2. Coverage gives $R_{\hat{k},t} - R_{f_a,t} \le \sup I^{D}_{a,\hat{k},t} \le -\gamma$. Conditional on the fixed task and pilot history, a fresh independent fallback restart has its from-start loss law. The fixed pilot charge adds C_t/B before truncation at one. □

Proof of Proposition 3. The source-bank maximum in equation 7 is one exchangeable scalar per calibration group. Split-conformal ranking bounds the new group's maximum optimism residual. Subtracting that bound from every registered selected-action estimate gives the simultaneous lower statement. The claim does not extend to a different acquisition policy, an unregistered stage, or an unobserved counterfactual action. □

C Claim and Evidence Boundary

Every released numerical statement is generated from immutable result shards and a machine-readable lineage record. Development tuning, calibration, external-family evaluation, and optional hardware diagnostics occupy disjoint artifact roots. External-test tasks cannot be loaded without a freeze binding the selected architecture, every ensemble checkpoint, route and acquisition contracts, calibration certificates, cost units, and selection evidence. The released result block is bound to the validated evidence digest printed in its generated status file.

D Additional Method Detail

D.1 What enters the network

The encoder receives the observation set, not a fully labeled candidate bank. With maximum parameter dimension $d_{\max}$, each padded row has $2d_{\max} + 4$ entries: candidate coordinates in the registered view; a coordinate-presence mask; a clipped log score-to-threshold ratio and its availability flag; and a signed predicate value and its availability flag. The scalar channel is zero and marked unavailable for a predicate-only interface. Padding is separately masked from attention and pooling. Per-observation costs remain in the ledger rather than being additional row features in this implementation.

A linear–GELU–layer-normalization map embeds each row. Masked self-attention blocks use residual connections, layer normalization, and a two-layer feed-forward map with intermediate width twice the hidden width. Masked mean pooling produces a permutation-invariant summary. The context network concatenates this summary with the six access flags, the observed candidate fraction, and the threshold. The fraction is the number of observations divided by pool size, not total pilot expenditure divided by budget. A learned route embedding conditions each survival head. Hard route eligibility comes from the task interface and executor registry; an auxiliary learned admissibility head cannot override those checks.

Table 3: Registered architectural choices. Their Cartesian product contains 512 settings; these are training settings, not 512 independent scientific populations.

Quantity	Values
Hidden width	$64, 96, 128, 192$
Attention heads	$4, 8$
Attention blocks	$1, 2$
Dropout	$0, 0.1$
Learning rate	$10^{-4}, 3 \cdot 10^{-4}, 10^{-3}, 3 \cdot 10^{-3}$
Weight decay	$0, 10^{-5}$
Survival endpoints	$16, 32$

D.2 Survival likelihood and ensemble prediction

Let K be the number of equally spaced output endpoints $b_r = r/K$, $r = 1, \ldots, K$. The route-conditioned hazard logits give $\lambda_{k,r} = \mathrm{sigmoid}(h_{k,r})$ and

$$\widehat{S}_k(b_r \mid H) = \prod_{q=1}^{r} (1 - \lambda_{k,q}(H)). \tag{9}$$

The implementation clamps hazards to $[10^{-6}, 1 - 10^{-6}]$ for numerical stability. For normalized event or censoring time u, the stored zero-based bin is $j = \lfloor u(K-1) \rfloor$. If $\delta = 1$ denotes a verified event, the negative log likelihood is

$$-\sum_{q<j} \log(1 - \lambda_{k,q}) - \delta \log \lambda_{k,j} - (1 - \delta) \log(1 - \lambda_{k,j}), \tag{10}$$

with zero-based hazards in this expression. Only observed, admissible route labels contribute. No-hit and budget-censored routes contribute survival likelihood; infrastructure errors are not training labels. Residual probability beyond the last endpoint is assigned to bounded loss one when forming risk summaries. The p90 is the first endpoint where the predicted CDF reaches 0.9, or one if it never does.

For J independently seeded fitted models, the deployed point prediction is

$$\widehat{R}_k(H) = \frac{1}{J} \sum_{m=1}^{J} \widehat{R}_{m,k}(H). \tag{11}$$

This is an average of member quantiles, not the quantile of a mixture distribution or a calibrated posterior interval. The separate source-bank calibration assesses its fallback-relative errors.

D.3 Training design and candidate supervision

The registered architectural grid has 512 combinations, listed in Table 3. Development validation loss selects the configuration, followed by the independently seeded five-member refit. The architectural grid is distinct from the later adaptive-policy selection and its independent calibration banks.

The training loss is survival negative log likelihood plus 0.1 times admissibility binary cross-entropy, 0.2 times candidate-class cross-entropy, and 0.05 times each of two candidate-cost smooth-L_1 losses. Candidate terms use their own validity mask. The candidate model has 17 latent classes: sixteen bins of the clipped high-fidelity log score-to-threshold ratio, plus an invalid-score class. These labels are curator-provided development supervision, including for models later deployed through predicate-only interfaces. Predicate-only deployment therefore does not mean score-free training; construction and training information belong in the reported offline cost account.

At inference, a hypothetical class is projected through the access mask before it is appended to the observation set. Predicate-only hypotheses expose only the corresponding predicate, not a numerical score. Neither candidate ranking nor this projection reads unqueried test-bank scores. The clean latent-bin model is not an independently validated noise-channel model. Selected-action calibration uses the actual registered observation process, including its corruption law; the auxiliary candidate

prediction itself should not be interpreted as the exact observable-outcome distribution in noisy or reduced-fidelity tracks.

Point cost uses a softplus output, and the upper-cost head adds a nonnegative softplus increment. The fixed replay labels are one score-check unit, or one quarter of a unit for reduced fidelity, so their point and p90 targets coincide. This training does not establish calibrated conditional p90 prediction for an arbitrary variable-cost simulator. Known replay charges and the explicit cost ledger enforce the registered allowance; a learned cost prediction alone does not.

D.4 The acquisition score and its scope

Candidate selection first removes observed or inaccessible candidates and applies the frozen candidate-subset rule. For each remaining candidate, each ensemble member evaluates the access-masked hypothetical observation for every latent class. Let $\overline{p}(z \mid x, H)$ be the ensemble-mean class probability, and let $\widehat{R}_m^{\mathrm{prot}}$ denote the member's provisional protected point risk using the frozen zero route radius. Its raw score is

$$G_m(x \mid H) = \widehat{R}_m^{\mathrm{prot}}(H) - \frac{\max_{m'} \widehat{c}_{m',0.9}(x \mid H)}{B} - \sum_z \overline{p}(z \mid x, H) \widehat{R}_m^{\mathrm{prot}}(H \cup \{(x, z)\}). \quad (12)$$

The lower member quantile $Q_{0.1}(\{G_m\}_{m=1}^J)$ is an ensemble diagnostic. It becomes an authorized acquisition statistic only after subtracting the selected-action group correction in Equation (8). Equal corrected scores are ordered by predicted p90 charge and then candidate index. The common correction cannot change that ranking.

The empirical calibration target uses paired continuations of the current and next observed history, with the explicit query penalty and verified terminal credit defined in the main text. It is not the difference of complete current and future cost ledgers. Calibration controls overstatement of this registered auxiliary target, not the causal effect of buying an observation on actual total discovery cost. The external comparison must establish the latter.

D.5 The support score and deterministic rank

Support uses the ensemble-mean task embedding. Within each access track, let μ and S be the training-embedding mean and sample covariance (denominator $n - 1$). The implementation sets

$$P = [0.9S + 0.1 \max\{\mathrm{tr}(S)/d_e, \epsilon_{\mathrm{mach}}\} I]^+, \qquad a(e) = (e - \mu)^\top P (e - \mu), \quad (13)$$

where d_e is embedding dimension and $+$ denotes the symmetric pseudoinverse. For calibration bank g, A_g is the maximum score over its registered histories. The deployed p-value is $(1 + \sum_{g=1}^{n_a} \mathbf{1}\{A_g \geq a(e)\})/(n_a + 1)$; ties are included, not randomized. The router falls back when this p-value is at or below the frozen support threshold. This calibrates an in-distribution support score; it is not a guarantee of detecting family shift.

D.6 What the construction charge covers

The complete ledger in the main text is an accounting target. The registered empirical construction adjustment covers only a reproducible lower bound: generating the development source banks, charged at one high-fidelity score-check unit plus one quarter-unit reduced-fidelity check per candidate. These are declared cost units, not a measured conversion of training time. Let $C_{\mathrm{bank}}^{\mathrm{LB}}$ denote this one-time charge and A the declared reuse horizon. The reported construction-adjusted ledger adds $C_{\mathrm{bank}}^{\mathrm{LB}}/A$ to deployment cost. The frozen artifact names this field `amortized_all_in`; that name does not make it a complete end-to-end cost measurement.

The lower bound excludes saved-score policy-replay compute, off-pool continuation and certificate generation, model-selection sweeps, and adaptive calibration evidence generation. Measured ensemble training seconds are retained separately and are not converted into simulator units. The registered horizons are $A \in \{1, 32, 1152\}$; multipliers $\{1, 2, 5, 10, 20\}$ on the lower bound probe sensitivity to missing construction charges. Neither a multiplier nor a larger reuse horizon is evidence that the omitted work has been measured. A favorable construction-adjusted comparison must therefore

Table 4: External-family search variables, registered thresholds, and scientific interpretation. These are mechanistic simulation benchmarks, not clinical, industrial, or grid-deployment validations.

System	Parameters and thresholds	What verification adds
Hodgkin–Huxley	Current $[0, 30]$; $g_{\mathrm{Na}} \in [90, 150]$; $g_{\mathrm{K}} \in [25, 50]$. $\tau = 0.01, 0.003, 0.001$.	An equilibrium-spectrum crossing must accompany sustained firing after a perturbation, with before/after spike-count contrast and a refractory-interval check.
CSTR	Damköhler number $[0.02, 0.30]$; heat release $[5, 30]$; cooling $[0.5, 5]$. $\tau = 0.05, 0.02, 0.01$.	Residual-checked equilibria and Jacobian stability are checked with perturbations and trajectories; a small stability margin alone is not a verified thermal event.
Swing network	Coupling $[0.2, 4]$; damping $[0.05, 2]$; disturbance $[0, 2]$. $\tau = 0.10, 0.05, 0.02$.	A multi-contingency order-parameter boundary is checked using a larger ensemble, coupling perturbations, terminal frequencies, and phase-cohesion diagnostics.
Gray–Scott	Feed $[0.01, 0.08]$; kill $[0.03, 0.075]$; diffusion ratio $[0.25, 0.75]$. $\tau = 0.015, 0.010, 0.005$.	A candidate near spatial-pattern onset is checked for grid refinement and persistence under perturbation; a transient coarse-grid pattern is insufficient.

survive broader evidence-generation accounting before it supports a complete end-to-end savings claim. The original values and comparison rules remain unchanged.

Inference for saved selectors is isolated when library versions differ, preserving the calibrated simulator and route runtime. Only charged pilot features and admissibility masks cross the process boundary. Inference timing includes communication; shared startup is a separate setup expense.

D.7 Statistical units and interpretation

The protected quantity is the difference of marginal p90 losses, $Q_{0.9}(L_k) - Q_{0.9}(L_f)$, not $Q_{0.9}(L_k - L_f)$. The second expression would depend on a coupling of solver randomness and answers a different question. Paired task and continuation seeds preserve the declared comparison without making the two estimands interchangeable.

For n_a independent calibration source-bank groups within track a, the conformal rank is

$$r_a = \lceil (n_a + 1)(1 - \alpha) \rceil . \tag{14}$$

If $r_a > n_a$, the implementation refuses a finite certificate. More trajectories from the same bank do not improve this group-count resolution. Independence alone also does not prove exchangeability across different simulator distributions. The formal statements require the within-track exchangeability assumption; held-out-family behavior remains empirical.

The secondary stored `cvar90` field averages all losses at or above the registered empirical p90. When many observations tie at that threshold, this may include more than the upper ten percent of probability mass. It is not generally standard expected shortfall (Rockafellar & Uryasev, 2002). The frozen selection values are preserved and labeled as a p90 exceedance mean. Reports must pair it with success-by-budget and the mass at bounded loss one, since a capped p90 alone cannot distinguish every difficult or empty-set task.

D.8 What the held-out scientific tasks ask

The four held-out families test different meanings of a verified event, not four versions of one scalar label. Each parameter candidate fixes a simulator; the scientific verifier then checks the induced dynamics. These are fixed card thresholds, not quantiles chosen to obtain a favorable number of hits.

The mechanistic context comes from conductance-based excitability, process dynamics, power-system stability, and reaction–diffusion pattern formation (Hodgkin & Huxley, 1952; Seborg et al.,

Table 5: Historical implementation outputs, not validated estimates of intended routing efficacy. Inherited-history errors affect development evidence and expose deployment paths; their external numerical impact is unquantified (Section D.10). Original family-held-out, construction-adjusted results by primary information track. Costs add an amortized source-bank construction lower bound, not a complete offline training bill; unsuccessful trials first contribute the larger of deployment cost and budget, then add construction cost. Cost overruns are retained. Tail mean is the mean cost at or above the empirical p90, not fractional-tail expected shortfall. Violation is the family-balanced rate of missing the frozen fallback-improvement margin. External interval coverage is descriptive, not a conformal guarantee under family shift. A dash denotes a quantity that is not defined for that comparator.

Method	p90 ↓	Tail mean ↓	Success@B (%)	Violation (%)	Ext. coverage (%)	Fallback (%)
Structured exact (registered gate: no pass)						
Protected portfolio	2,357	2,501	85.0	0.0	–	100.0
Best fixed route	2,784	2,927	84.9	100.0	–	100.0
Nearest neighbor	914.8	996.9	85.5	83.3	–	24.6
Random forest p90	1,129	1,422	83.5	83.3	–	39.5
Calibrated abstention	2,773	2,927	85.7	96.4	–	79.9
QC-PHAST Router	2,773	2,927	85.7	96.4	–	79.9
Scalar exact (registered gate: no pass)						
Protected portfolio	475.4	541.4	85.1	0.0	–	100.0
Best fixed route	882.7	1,022	85.1	100.0	–	0.0
Nearest neighbor	849.9	928.0	85.9	98.4	–	0.0
Random forest p90	875.6	945.4	85.8	99.0	–	2.2
Calibrated abstention	902.6	984.3	85.1	99.5	–	79.6
QC-PHAST Router	902.6	984.3	85.1	99.5	–	79.6
Predicate exact (registered gate: no pass)						
Protected portfolio	548.2	613.0	85.0	0.0	–	100.0
Best fixed route	963.5	1,012	85.3	100.0	–	0.0
Nearest neighbor	969.5	1,028	85.2	100.0	–	2.8
Random forest p90	980.6	1,042	85.2	100.0	–	6.3
Calibrated abstention	985.6	1,045	85.1	100.0	–	79.4
QC-PHAST Router	985.4	1,045	85.1	100.0	–	79.7

2010; Kundur, 1994; Pearson, 1993). Our operational thresholds and numerical checks define this benchmark's events; they are not universal transition constants supplied by those references. The extended report contains the complete cards and implementation notes. Information access determines how a search may propose a candidate, while the high-fidelity verification requirement remains separate from that proposal.

D.9 Historical implementation outputs

The following numerical rows are preserved unchanged for reproducibility. The history defect and deployment exposure described in Section D.10 limit their interpretation. They are not a corrected evaluation of the intended router or proof of comparator superiority.

D.10 When calibrated replacement becomes impossible

This diagnostic was added after the registered external analysis. It reads the frozen checkpoints, calibration certificates, saved development-validation reports, and every external router decision record. It changes no predictions, route choices, event certificates, costs, or primary tests. External action totals reconcile against the independent release audit. All 1,152 configurations and 147,456 router trials are included, including unsuccessful searches and searches that finish in the pilot.

An elementary feasibility bound. Suppose every route prediction belongs to $[\ell, u]$, and the upper fallback-relative interval is $\widehat{R}_k - \widehat{R}_f + q_a$. Replacement requires

$$\widehat{R}_f - \widehat{R}_k \geq q_a + \gamma. \tag{15}$$

Table 6: Decision feasibility under the frozen model and calibration. The largest observed gap is the maximum fallback-minus-specialist prediction over admissible alternatives and every recorded decision stage, not an outcome improvement. A correction exceeding the model's full prediction span prevents any specialist commitment.

Track	Banks	Correction q_a	Max. observed gap	Infeasible
Structured exact	39	0.987	0.125	yes
Scalar exact	38	1.021	0.212	yes
Predicate exact	37	0.862	0.050	no
Reduced fidelity	37	1.281	0.388	yes
Scalar noisy	37	0.736	0.475	no
Predicate noisy	36	0.739	0.050	no

Table 7: External decision accounting. Pilot hits and terminal fallback are disjoint terminal outcomes. Continued trials count searches buying at least one additional observation; they overlap terminal outcomes and must not be added to them. Support rejection counts decision stages, not trials.

Track	Trials	Pilot hits	Fallback	Continued	Support reject
Structured exact	24,576	4,942	19,634	9	1,508
Scalar exact	24,576	5,010	19,566	0	358
Predicate exact	24,576	4,990	19,586	0	709
Reduced fidelity	24,576	3,583	20,993	0	1,092
Scalar noisy	24,576	5,211	19,365	75	2,297
Predicate noisy	24,576	4,913	19,663	0	786

Since the left side cannot exceed $u - \ell$, $q_a + \gamma > u - \ell$ makes replacement impossible. Equality is not excluded because the implemented comparison is non-strict. This is an algebraic consequence of the rule, not a new conformal theorem. It requires neither exchangeability nor a claim about the correctness of the predictions.

The five verified checkpoints all use 16 survival bins with edges $1/16, \ldots, 1$. Each member returns a grid quantile; inference averages those five quantiles rather than taking the quantile of a mixture survival distribution. Averaging preserves the interval $[1/16, 1]$, although intermediate values need not be grid edges. The attainable prediction gap is therefore at most $15/16 = 0.9375$. Final corrections are computed from a different object: group maxima of errors against continuation intervals. Such errors need not fit inside the prediction gap. Their exceeding that gap is a loss of decision resolution, not by itself an invalid conformal calculation.

What the complete decision records show. Three tracks are algebraically infeasible under the frozen model and correction. For the other three, Table 6 compares the largest admissible predicted improvement over every saved stage with the required correction; none is sufficient. This comparison includes support-rejected histories and excludes the fallback itself and inadmissible routes. It shows that removing the support gate alone, while retaining these predictions and intervals, could not produce a commitment on those recorded histories. It does not predict a newly trained policy or histories that a different acquisition policy would generate.

Prediction saturation is not a universal explanation. All admissible predictions equal one at only 410 reduced-fidelity stages and 488 scalar-noisy stages, and at no stage in the other tracks. The more general obstruction is insufficient separation relative to the interval correction. Likewise, support rejection counts in Table 7 are not synonymous with all fallback decisions. Only 84 trials buy another observation, producing 203 acquisition actions; repeated acquisitions are not additional independent trials. Zero commitments leave commitment-conditional coverage undefined.

A development tradeoff, not a repaired external result. The five saved variants in Table 8 were evaluated during development selection. Its harm statistic is the fraction of replay trajectories whose estimated router p90 exceeds the paired fallback p90, averaged within family and then equally across families. It is not a per-query failure probability. The saved implementation reports more specialist choices, lower p90 loss, and greater harm without route correction; the gain-correction variant

Table 8: Saved development-validation decision replays for the selected candidate, each using 1,728 trajectories. These historical diagnostics are subject to the inherited-history limitation in Section D.10, not validated fixed-history ablations. Harm is the saved family-balanced fallback-relative p90 harm statistic on bounded loss; it is not the external construction-adjusted violation column.

Variant	Route actions	Acquire actions	p90 loss	Harm (%)
Full	0	317	0.0802	0.58
No route correction	1,080	4	0.0703	22.51
No support check	0	403	0.0803	0.64
No gain correction	0	15,645	0.0810	16.26
No continuation	0	0	0.0806	0.00

Stage	Information that determines the continuation
Purchased pilot	Observe x_a, x_b, x_c; record their returned values and charges. Verification of x_a rejects the proposed event.
Correct continuation	Restore those values, the rejection, and spent query/verifier budgets. Only a new permitted action buys new information.
Index-only replay	Query x_a again with new randomness. A new verification may pass, ending the replay before x_b, x_c and their already-paid costs are represented.

$$H_t = \big((x_i, o_i, c_i)_{i \le t},\ \text{verification receipts},\ \text{remaining budgets}\big) \quad \neq \quad (x_i)_{i \le t}.$$

Figure 3: Why a continuation needs the actual history. This symbolic example illustrates the audited failure mechanism; it is not a new measurement or an estimate of its frequency. An authorized retry would need an explicit policy and a new charge. It cannot silently replace an already recorded rejection.

reports more acquisitions without lower p90 loss. These records are retained, but the history defect below prevents treating their differences as validated fixed-history policy effects. They are neither independent confirmation nor held-out-family ablations and cannot justify weakening a safeguard after observing the external test.

A subsequent inherited-history audit. Figure 3 makes the distinction between the actual history and its indices explicit. A development-only audit rechecked all 5,184 frozen trajectory files and their manifest hashes. Across their saved stages, continuation losses sometimes imply a total cost below the pilot observation cost already incurred. This lower-bound check flags 87 of 864 task configurations across 20 of 144 source banks, in the Kuramoto and stochastic-toggle systems. These counts are not independent error-rate estimates: contracts, stages, routes, and repetitions are nested. The finite-pool executor replays pilot indices with a new verification seed; a previously rejected proposal can then pass, terminating before the full inherited pilot is charged. Replaying indices is therefore not equivalent to continuing from the same observations and verification history. Samples passing this cost check are not thereby proven correctly conditioned.

No outcomes were replaced or excluded. The proposed development recalibration was stopped; old normalized losses cannot recover missing per-continuation verification histories by an arithmetic adjustment. A corrected study must preserve the observations, rejected certificates, spent costs, and remaining budgets, then regenerate the affected evidence under a separately identified implementation. This audit neither re-evaluates the external executors nor quantifies the defect's impact on their reported costs. It restricts interpretation of the development diagnostics and their use for redesign; it does not change the recorded zero-commitment counts or the algebraic prediction-range obstruction.

Source inspection also identifies index-only handoffs in fixed-pilot training, adaptive calibration generation, and router and selector deployment. The latter two subtract prior verification attempts from the remaining allowance and retain their charges, but still pass indices rather than the actual observations and rejected certificates. Structured warm starts likewise read bank scores and re-verify pilot proposals. These source paths require assessment; their presence alone does not quantify a

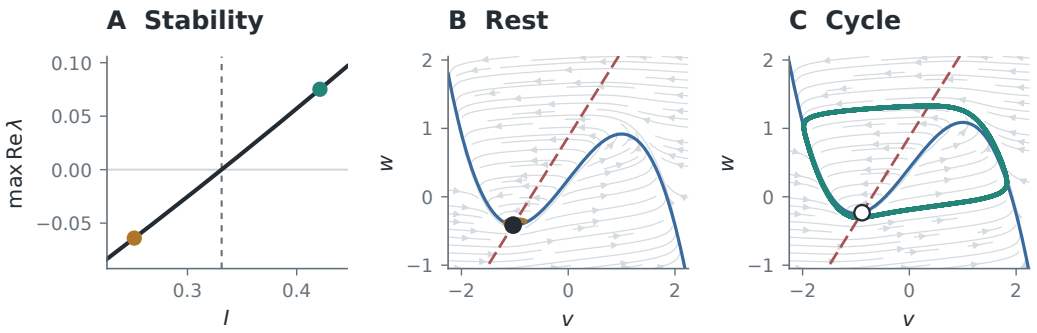


Figure 4: A computed FitzHugh–Nagumo parameter-to-phase-space bridge at $\varepsilon = 0.08$, $a = 0.70$, $b = 0.80$. (A) The leading eigenvalue's real part crosses zero near $I = 0.331$. Orange and teal points select $I^- = 0.251$ and $I^+ = 0.421$. (B,C) A perturbed trajectory returns to rest at I^- and approaches a cycle at I^+. The solid blue and dashed red curves are the $\dot{v} = 0$ and $\dot{w} = 0$ nullclines; gray curves are streamlines. Filled and open black markers denote stable and unstable equilibria. The stable transient is shown from $t = 0$; the cycle is shown after $t = 120$. This illustration is not a router-performance result.

change to any released external result. The standalone fallback is invoked without an inherited pilot, so this particular prefix-replay mechanism is absent from that call, not necessarily from every aspect of its implementation.

The separate development repair now carries actual observations and complete verification receipts into both classical finite-pool and structured continuations. Structured solvers reuse observed warm-start scores without re-verifying rejected pilot proposals; both executor types retain spent query and verification allowances. Fixed-pilot and frozen-model adaptive producers save each paid action before branching, preserve terminal pilot hits, and can resume completed actions and repetitions without buying them again. A versioned dataset reader recomputes the ledgers from these receipts and encodes the saved observations, rather than replaying pilot indices. Regression and integration tests establish these implementation properties, not routing efficacy. Quantum continuation, calibration and external execution remain outside this development repair. Corrected training evidence, refitting, policy-matched validation and fresh calibration are still required before a revised router comparison. The old numerical study remains a historical result, not corrected evidence of fixed-history routing performance.

Actionable scope of the finding. Before committing to a future external evaluation, inspect attainable prediction differences, calibration corrections, decision frequencies, and conditional residuals on development banks. More calibration groups can improve resolution of the conformal rank without necessarily making the correction small. A feasible interval is only a necessary condition: the learned model must also distinguish useful solvers, the selected events must verify, and any gain must pay for the pilot and construction. A redesigned predictor or calibration scheme requires development-only selection and a new untouched external test. This diagnostic supplies no evidence that such a redesign will succeed.

D.11 Parameter-to-dynamics illustration

D.12 Construction charges and unsuccessful searches

Let $\phi(C, B)$ equal C on a certified success and $\max\{C, B\}$ otherwise, and let K be the amortized construction lower bound. The frozen fallback-relative tables and cost sensitivities use $\phi(C, B) + K$ for a learned method. Direct router–selector diagnostics instead use $\phi(C + K, B)$ for each method. Both retain overruns, but they differ whenever an unsuccessful search stops below budget and $K > 0$. Only the former convention supplies Table 5; no direct-comparison p90 is substituted into that table. The independent audit reproduces each saved convention separately. This distinction is a limitation

Table 9: Post hoc, harmonized cost sensitivity for the router. Effect is fallback minus router; positive values favor the router. Construction is added before the failure penalty throughout this table.

Track	p90	Effect [95% interval]	Success (%)
Structured exact	2,773	-415.8 [-463.6, -335.9]	85.7
Scalar exact	831.5	-356.1 [-444.0, -205.6]	85.1
Predicate exact	916.5	-368.3 [-448.0, -214.4]	85.1
Reduced fidelity	1,041	-295.6 [-425.2, -161.1]	71.9
Scalar noisy	765.4	-336.6 [-438.3, -165.3]	90.9
Predicate noisy	923.8	-368.0 [-437.6, -233.3]	86.9

of the registered analysis, not a reason to change its outputs after observing results. A separately approved harmonized sensitivity follows in Table 9. It does not replace the registered tests.

Post hoc harmonized accounting sensitivity. This separately approved analysis applies the construction charge before the no-hit budget penalty for every method. It reuses all 1,152 configurations and 147,456 paired trials, with 5,000 paired family–bank–task–trial bootstrap replicates per track. Intervals are pointwise descriptive intervals, not new confirmatory tests. The registered primary results and their decision criteria remain unchanged.

E What survives the failed replacement test?

The study did not establish a cost-reducing router. It nevertheless distinguishes several failure mechanisms that should be checked separately when building a protected selector. This section recovers those lessons from the development and external audits. It does not reclassify implementation tests as efficacy experiments or turn the affected cost records into a clean negative estimate of the intended algorithm.

E.1 Calibration size and decision resolution are different

For n eligible source-bank groups and error level α, the deterministic split-conformal rank is $r = \lceil (n+1)(1-\alpha) \rceil$. A finite empirical correction requires $r \le n$, hence

$$n \ge \left\lceil \frac{1}{\alpha} - 1 \right\rceil. \tag{16}$$

This is the standard finite-rank requirement, not a new conformal guarantee (Angelopoulos et al., 2024). Groups must also satisfy the relevant exchangeability and policy-matching assumptions. Neither their number nor the existence of a finite order statistic establishes those assumptions.

In the initial validation audit, 1,296 trajectories did not suffice: 24 of 162 candidate–fold–track checks lacked enough eligible groups. Pilot successes had already ended those searches, so they could not supply observations of subsequent routing decisions. The smallest eligible count was 7, below the 9 required at $\alpha = 0.1$. The source-disjoint replacement produced 5,184 trajectories and passed all 162 checks; its smallest route-support and selected-action group counts were both 31. Figure 5 shows the saved route-support counts. These are metadata findings about the executed historical study. The later history defect still limits the interpretation of its continuation labels.

The subsequent obstruction is numerical, not a shortage of groups. Write the predicted fallback-minus-specialist gap as $g_{k,t} = \widehat{R}_{f,t} - \widehat{R}_{k,t}$ and the required correction plus margin as $b_a = q_a + \gamma$. Replacement requires $g_{k,t} \ge b_a$. If route predictions lie in $[\ell, u]$, then

$$b_a > u - \ell \quad \Longrightarrow \quad \text{no specialist can satisfy the replacement rule.} \tag{17}$$

This elementary bound applies to any such additive comparison, regardless of how the predictor was learned. Its value is diagnostic, not a new distribution-free theorem. More independent groups can make an empirical quantile available without making its value small enough to use. More repetitions within the same banks may refine continuation estimates but do not create new exchangeability units. Finer prediction bins likewise do not guarantee that calibration error or inner-interval uncertainty will shrink.

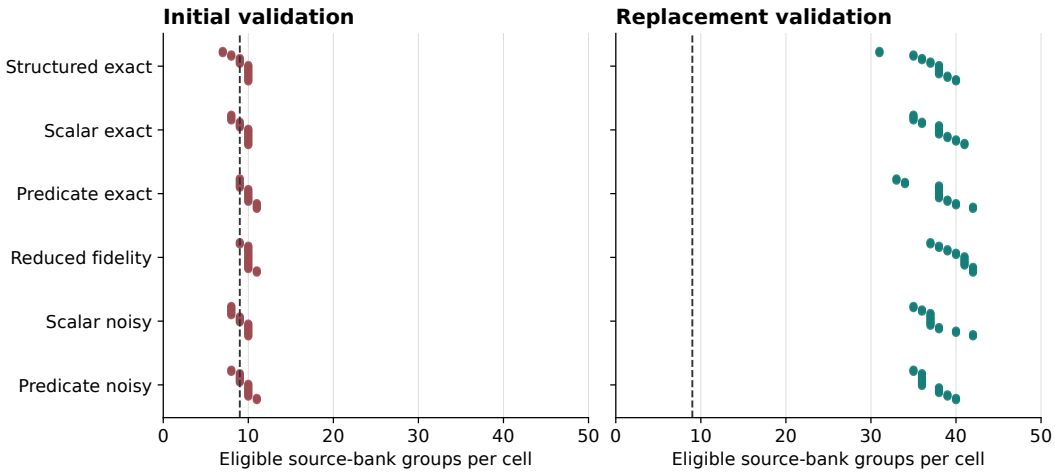


Figure 5: The sample-count bottleneck was resolved, but that did not establish routing efficacy. Each dot is one saved candidate–fold–track check of eligible route-support source banks. The dashed line is the finite-rank minimum at $\alpha = 0.1$; vertical offsets only reveal repeated counts. Candidate checks share banks and are not independent experimental replications. The two panels use different validation banks, not paired measurements of a treatment effect. Counts are recomputed from the preserved audit receipts. Passing this check says nothing about the validity of history conditioning or whether the fitted correction permits commitment.

E.2 Diagnose the reason for abstention before changing the rule

Three explanations should not be conflated. A specialist may genuinely offer no useful gain; the predictor may fail to distinguish a useful specialist; or the uncertainty correction may be too wide to authorize the predicted gain. The present evidence establishes an obstruction for the frozen comparison, not the absence of solver opportunity. In particular, inspecting every saved stage, including support-rejected stages, shows that removing support rejection alone would not authorize a route under the same predictions and correction. It does not describe the histories a different acquisition policy would generate.

A useful development report therefore pairs decision frequency with the distribution of $g_{k,t} - b_a$, prediction saturation, and fallback-relative outcomes from valid continuations. The first two describe whether the rule can act; the last asks whether acting would help. Changing α, the margin, or the selected route after seeing the external outcomes answers a new question. It cannot be presented as confirmation of the frozen rule.

E.3 A shared candidate prefix is not a shared information state

The scientific comparison conditions on returned observations and verification outcomes, not only on queried locations. A noisy score, a reduced-fidelity score, and the hidden exact score are different information. A rejected event proposal is part of what is known. If a continuation reconstructs the prefix with fresh randomness, it can change both what the solver believes and what has been charged before the solver starts.

The independent development audit inspected 83,510,208 nested route repetitions over 141,214 saved stages. It detected 1,032,698 repetitions whose implied total cost was below the already-paid pilot observation cost, across 4,901 stages, 87 task configurations, and 20 source banks. These nested counts describe the detector's findings, not an independent-sample error rate. Passing this necessary lower-bound check does not certify history correctness. The mechanism and deployment exposure are detailed in Section D.10.

For reuse, a continuation input should preserve the actual observations, channels, verification receipts, and consumed budgets. Forking two solvers must copy that state before drawing new continuation

randomness. A retry must be an explicit new paid action, not a replacement for an earlier rejection. This is a semantic requirement on the conditional learning target: reproducible hashes of an incorrectly reconstructed history do not make its labels valid.

E.4 What another researcher can use now

The directly reusable outcome is a set of discriminating checks, not a performance recommendation. First, verify that the proposed scientific event has an operational certificate, distinct from a low proxy score. Next, identify the independent calibration groups and their decision eligibility. Before an expensive external study, compare the fitted correction with the predictor's attainable and observed gaps on development histories. Audit the continuation state and costs independently of file provenance. Finally, evaluate the whole charged policy against an executable fallback, retaining unsuccessful searches and overruns.

Report what each endpoint means. The difference of two conditional marginal p90 losses is not the p90 of paired trial differences; a family-average of task p90 costs is not a pooled p90. Bounded calibration loss and uncapped deployment cost are different targets. A stored tail-summary name is not its mathematical definition. These distinctions determine whether a result answers the user's operational question, even when all computations are reproducible.

None of these checks makes a positive result inevitable. They help distinguish lack of solver opportunity, insufficient decision resolution, and invalid evidence before further compute is committed. Corrected efficacy, including its full construction cost, remains an open empirical question for a separately trained, calibrated, and independently evaluated version.

F Matched-Proposal Quantum Diagnostic

The prior QC-PHAST study motivates an optional coherent-access diagnostic (Tirhekar & Bajaj, 2026). Let a learned proposal q place marked mass $a_q = \sum_{i \in A} q_i$. Classical sampling from q requires order $1/a_q$ checks, while amplitude amplification from a prepared q state supplies an ideal order-$1/\sqrt{a_q}$ coherent-call reference even when a_q is unknown (Brassard et al., 2002). This generalizes the uniform-state BBHT reference (Boyer et al., 1998). Both arms must receive the same q and pay the same pilot and training costs. This scaling uses verifier-compatible marked mass only when the coherent oracle marks that exact set; a proxy-positive oracle additionally incurs its verification precision and cost. The amplified arm pays preparation $S(q)$, coherent predicate cost, error control, uncomputation, and verification. Structured preparation methods show why q must be paired with a concrete preparation mechanism rather than priced only by its marked mass (O'Brien & Sünderhauf, 2025). This diagnostic is not part of the primary contribution and is inadmissible without measured coherent-access evidence. A stored score array, a simulated BBHT schedule, or a successful small circuit does not establish end-to-end advantage.