CERTIFYING BRANCH AND EVENT IDENTITY IN FINITE-BUDGET NONCONVEX TRACKING

Anonymous authors
Paper under double-blind review

ABSTRACT

Finite-budget nonconvex tracking pipelines often accept candidates because they have low residuals or emerge repeatedly from local solves. That evidence is not the same as certifying branch identity or same-branch event identity: an unobserved branch, or a same-branch opposite-event solution, can remain compatible with the public transcript. Adaptive branch-aware certification (ABAC) makes branch and event identity explicit requirements of finite-budget delivery. Its evidence interface lets a selector allocate budget to branch/support probes and event-complement checks, or abstain when the evidence does not support delivery. Under a public-transcript ambiguity model with an explicit dominance-completability condition, we prove that worst-case zero-false delivery over transcript-compatible instances requires both witness types, and a constructive public-query hard family gives a tight $m + 1$ lower bound for per-location evidence acquisition with m alternative branches and one opposite-event location. We instantiate ABAC with a frozen event-margin fallback and abstention. Exploratory nominal-budget results favor coverage, but omit failed-probe costs and permit seed-derived access to the observation-error direction. Our public-input evaluation charges every probe without retuning the threshold. On 640 synthetic transitions from eight held-out seeds, both the fallback and public GCT make 17 reference-matched deliveries; the fallback also makes four false deliveries, versus zero for GCT. Mean accumulated solver cost is 85.74 versus 110.64 iterations, with every transition capped at 120. The paired seed-bootstrap 95% interval for the safe-coverage difference is $[-1.09, 1.25]$ percentage points. These results support a cost–risk tradeoff, not coverage or safety dominance. Correctness uses a best-of-192 SLSQP reference, not a global certificate; all 160 prospective states across eight seed trajectories retain their labels after 320 additional starts each. The theoretical contract remains distinct from the empirical gates and their coarse labels.

Keywords: selective certification, nonconvex tracking, branch identity, public-query auditing, abstention

1 INTRODUCTION

Nonconvex tracking and constrained decision pipelines increasingly rely on local optimization: a system generates candidates from multiple starts, checks residuals, and accepts the most plausible point when the numerical evidence looks strong. That practice is effective for finding feasible or low-objective points, but selective deployment asks for a different object. A delivered decision is not merely a good numerical point; it asserts an identity: which branch of the nonconvex landscape the point belongs to, and which active/inactive event it realizes. Low residual and repeated local convergence can support that assertion, but they do not certify it.

The broken equivalence is *candidate quality* $\neq$ *identity certification*. Multistart search improves coverage of the observed basin but does not exclude unseen branches. Residual thresholds test local feasibility but not branch dominance. A robust margin against observed competitors helps only if the relevant competitors have been observed, and an event-side check inside one branch does not repair a wrong branch. Conversely, strong branch support can still be insufficient when the same branch admits a better solution under the opposite active/inactive event.

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.

We formalize this problem as *adaptive branch-aware certification* (ABAC), a finite-budget public-transcript interface. Before delivering or abstaining, a policy may spend computation on branch probes, continuation, margin refinement, event-complement verification, or abstention, but it cannot inspect hidden reference labels. The central object is therefore an evidence-acquisition policy with selective branch/event semantics, not an optimizer update rule.

Our theory and experiments support a bounded claim. We prove that, under the public-transcript interface and dominance-completability of every unresolved location, worst-case zero-false delivery requires both branch/support and same-branch event-complement witnesses; a constructive hard family shows that per-location public evidence has a tight $m + 1$ worst-case query threshold. We instantiate the interface with a budgeted shared-probe router and freeze the calibration-selected policy before blind evaluation. The calibration-selected residual router improves safe coverage on Test A and Test B, but has lower pooled safe coverage and lower delivered-case cost than nominal-budget obs-μ GCT on the unopened eight-group audit. The separate event-margin fallback has exploratory coverage gains under nominal accounting; its public-input evaluation gives no safe-coverage gain and four additional false deliveries. Controls examine hidden-parameter access, post-hoc stability gates that fail blind evaluation, overfitting in a richer tree, and the event-complement oracle is reported as cost-unmatched rather than as an online claim. The paper therefore reports a protocol-specific risk–coverage–cost frontier, not a broad optimizer-superiority claim. Input access and cost accounting are specified for each comparison. The public-input evaluation keeps the threshold fixed but does not confirm the nominal-budget coverage advantage.

Contributions. We make four contributions.

1. We define ABAC, a public-transcript evidence-acquisition interface for selective branch/event delivery under a finite online budget.
2. We prove a two-witness necessity result and a constructive public-query indistinguishability lower bound for the defined evidence interface.
3. We instantiate the interface with a calibration-selected shared-probe router, with a prospective public-input version whose cumulative solver iterations obey the same cap as the public GCT comparator.
4. We validate the interface with separate held-out and frozen blind splits, leakage controls, overfit controls, failed post-hoc guardrails, oracle-boundary analysis, and reproducible cost accounting.

2 Related Work

Nonconvex and certified optimization. Multistart and continuation methods improve candidate generation, but candidate generation does not imply selective identity certification (18). Deterministic global optimization and branch-and-bound methods provide certificates when relaxations close (14), and convex optimization gives tractable certificate machinery in structured regimes (4). ABAC studies which public evidence to acquire before delivery under a fixed online budget, with abstention when a global certificate is unavailable. The distinction between local candidate quality and global status is also reflected in dynamic programming and equilibrium-computation theory (3; 7; 19).

Selective prediction and uncertainty quantification. Selective prediction formalizes the risk–coverage tradeoff and reject option (6; 11). Conformal prediction gives a complementary distribution-free language for prediction sets and uncertainty quantification (1; 20). ABAC uses the deliver/abstain discipline from this literature, but its correctness event is not label accuracy alone: the delivered object must carry a branch/event identity, so the certificate decomposes into branch/support and same-branch event-complement witnesses. Selective classification, time-uniform concentration, and empirical Bernstein bounds provide related risk-control tools, although none supplies ABAC's branch/event identity contract (2; 10; 13; 15; 16).

Adaptive computation and routing. Adaptive computation methods choose how much computation to spend on an input (12). ABAC also allocates computation, but the allocation is constrained by a certificate contract: a channel is useful only if it removes a public ambiguity that would otherwise make delivery unsafe. This is why richer routing is reported as a control rather than automatically promoted when calibration coverage improves. Its budgeted evidence allocation is also related

to fixed-confidence sequential design, multiple-answer exploration, and submodular selection, but ABAC adds a public-transcript safety condition (5; 8; 9; 17).

3 Public Branch/Event Certificate Interface

At a transition s, a solver produces a candidate x_0 with branch identity b_0 and event identity e_0. The selector observes a public transcript T: residuals, objective values, solver status, branch signatures, continuation traces, and any explicitly purchased branch or event-complement probes. Hidden terminal references are used only after the decision to score correctness.

Let $d(T) \in \{0, 1\}$ be the selector's decision: $d(T) = 1$ means that the candidate x_0 is delivered, and $d(T) = 0$ means that the selector abstains. A delivery is safe when x_0 has the intended branch/event identity; otherwise it is a false delivery. Let $\mathcal{A}_b(T)$ denote alternative branch locations still compatible with T, and let $\mathcal{A}_e(T)$ denote same-branch opposite-event locations still compatible with T. These are evidence ambiguity sets, not solver failure sets: a low residual can coexist with nonempty ambiguity.

The ABAC delivery contract is

$$d(T) = 1 \Rightarrow \mathcal{A}_b(T) = \varnothing \quad \text{and} \quad \mathcal{A}_e(T) = \varnothing. \tag{1}$$

The contract specifies the information needed for worst-case safe delivery. The implemented router does not compute these ambiguity sets exactly: its public residual, margin, and event gates are empirical proxies for the two witness types. Their reliability is measured by false deliveries and coverage; the necessity result does not establish soundness of these empirical gates.

4 Theory: Two Witnesses and a Public-Query Lower Bound

We use lower objective values as preferred outcomes. For an instance I, write $z \prec_I x_0$ when z is feasible and $J_I(z) < J_I(x_0)$. Let $\iota(x) = (b(x), e(x))$ denote branch/event identity. A delivery is identity-false when the delivered identity is not the identity required by the task.

Assume each compatible instance has an attained global minimum. The task accepts an identity of a global minimizer, with a fixed rule if ties require a single identity. For a feasible identity a, define its value $v_I(a) = \inf\{J_I(x) : x \text{ feasible}, \ \iota(x) = a\}$.

Definition 1 (Dominance-completability)**.** *Fix a public transcript T, an anchor x_0, and a compatible location $z \in \mathcal{A}_b(T) \cup \mathcal{A}_e(T)$. The location z is dominance-completable if there exists a transcript-compatible instance $I_{T,z}$ with exactly the same public transcript, feasible z, and $J_{I_{T,z}}(z) < v_{I_{T,z}}(\iota(x_0))$. Thus, dominance-completability is existential over compatible hidden completions; it does not assert preference in the observed instance.*

Proposition 1 (Identity-class preference implies identity falsity)**.** *Assume the delivery task requires the preferred feasible identity and that a compatible location z has $\iota(z) \neq \iota(x_0)$. If z is dominance-completable, then there is a transcript-compatible instance on which delivering x_0 is identity-false. Without strict identity-class preference (or an explicit tie rule), a different compatible identity alone does not imply false delivery.*

Proof. Dominance-completability gives a transcript-compatible instance on which the feasible location z is strictly preferred to every feasible point with identity $\iota(x_0)$. An attained global minimum therefore cannot have that identity, regardless of the tie rule among the remaining identities. Delivering x_0 is identity-false on that instance. □

Interpretation. An alternative compatible identity alone does not falsify delivery. Dominance-completability supplies a compatible instance on which the entire anchor identity class loses. Merely beating x_0 is insufficient: values $1, -1$ for two points of identity A and 0 for identity B leave A globally preferred although B beats the first point. Ambiguity sets are evidence objects, objective preference is instance-dependent, and identity falsity is a delivery outcome. The empirical gates estimate these levels; they do not prove the existential completion property.

Proposition 2 (Two-witness necessity)**.** *Under public-transcript ambiguity semantics, suppose each ambiguity location in $\mathcal{A}_b(T)$ or $\mathcal{A}_e(T)$ is dominance-completable: there exists an instance with the same public transcript in which that location is preferred to every feasible point of the delivered identity, while all quantities visible to the selector are unchanged. If there is a compatible alternative branch in $\mathcal{A}_b(T)$, or a compatible same-branch opposite-event location in $\mathcal{A}_e(T)$, then no selector that depends only on T can both deliver x_0 on the current instance and guarantee zero false deliveries over all instances compatible with T.*

Proof. By dominance-completability, a compatible alternative yields a second instance with the same public transcript in which the delivered identity cannot be globally optimal. The same construction applies to a compatible opposite-event location. Since the selector sees the same T, its action is the same on both instances; delivery is false on one of them. Thus delivering x_0 cannot have a zero-false guarantee over the transcript-compatible instance class. □

Assumption 1 (Per-location public query interface)**.** *There are m alternative branch locations and one same-branch opposite-event location. In this finite hard family the anchor and these locations are the only feasible representatives of their respective identities. One public query inspects one location and costs one unit. A query that misses the hidden competitor returns the common score $+\Delta$, with $\Delta > 0$. A policy may deliver or abstain after at most B queries.*

Theorem 1 (Public-query indistinguishability)**.** *For every deterministic policy π under the per-location public query interface, if $B < m + 1$, there exists a pair (I^+, I_j^-) such that π receives the same public transcript on both instances, the anchor candidate is correct on I^+ and false on I_j^-, and therefore π cannot deliver on I^+ while guaranteeing zero false deliveries on the family. Thus a zero-false policy required to deliver on I^+ needs $B \geq m + 1$; always abstaining does not meet this requirement. The bound is tight for this interface: querying all $m+1$ locations separates the family.*

Proof. With fewer than $m + 1$ queries, at least one location j is unqueried. In I^+, every location has score $+\Delta$. In I_j^-, only j has score $-\Delta$. Every queried location therefore returns the same value on both instances. The transcript and action coincide, but the truth of delivery differs. A policy that delivers on I^+ therefore delivers falsely on I_j^-. Conversely, querying all $m + 1$ locations separates I^+ from every I_j^-, proving tightness for this interface. □

Figure 1 displays the indistinguishable pair: the unqueried location can be an alternative branch or the same-branch opposite event. The numerical budget examples are retained in the supplement.

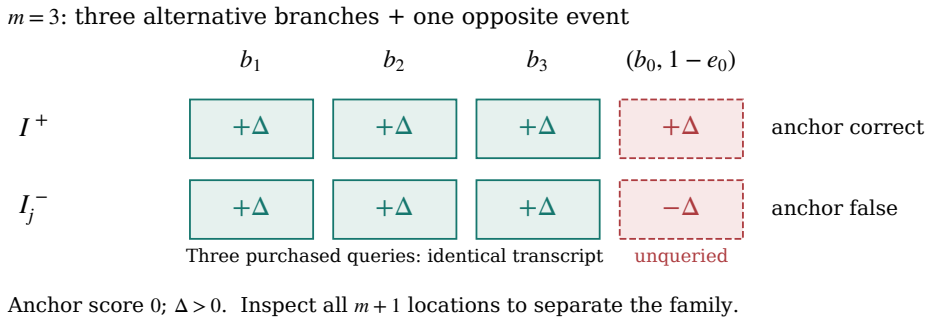


Figure 1: Constructed hard family, not experimental data. With $m = 3$ and $B = 3$, the queried scores agree on I^+ and I_j^-, while the unqueried opposite-event score changes the preferred identity. The same construction applies to an unqueried alternative branch. Worst-case zero-false delivery on I^+ requires all $m + 1$ queries in this interface.

Interpretation. The theorem rules out a tempting shortcut: a finite budget cannot be split naively between branch search and event-complement search while leaving an unresolved location and still claim worst-case zero-false delivery in this interface. It does not rule out joint relaxations, randomized risk contracts, learned priors, or solvers whose one call certifies multiple locations.

5 ALGORITHM: FROZEN SHARED-PROBE ROUTING

The ABAC instantiation uses one cheap public probe to route the remaining budget. The event-margin threshold is selected on calibration labels, but its design is informed by exploratory blind results. It remains fixed in the prospective evaluation:

1. Run a cheap six-start public probe with $p = 11$ local iterations.
2. Compute public probe features.
3. Apply the frozen nominal-risk-screened rule: route to the event-complement channel if the probe event margin is at most $q_* = 1.2333254966819576$, and otherwise route to the shared-probe GCT continuation channel. The threshold is selected on calibration data only; blind labels are never observed by the router.
4. Deliver only if the selected channel's public evidence gate passes.
5. Abstain otherwise.

The prospective implementation initializes from observed constraints, uses the normalized observed parameter direction, charges failed probes, and caps each extension by the remaining budget. Its maximum total cost is 120. Results without these input and accounting restrictions are nominal-budget diagnostics. A post-hoc stability gate and a depth-2 tree over the same trace are reported as controls: the former improves the inspected held-out split but fails blind-seed soundness, while the latter improves calibration coverage but drops held-out coverage.

6 EXPERIMENTS

The experiments compare nominal-budget and cumulative-budget protocols, with controls for parameter access, accumulated cost, and router selection.

6.1 EVALUATION PROTOCOL AND MEASURES

We separate policy selection from evaluation. The calibration split selects the router and its threshold; two held-out splits evaluate separate sets of transitions, and the blind audit evaluates the frozen policy on eight unopened seed blocks. Each transition is scored against a shared offline reference: the best accepted SLSQP solution from 192 deterministic feasible starts under hidden true parameters, using the benchmark objective, constraints, and residual filter. The obs-μ GCT is a comparator, not the source of these labels. The offline multistart solution is a heuristic reference rather than a certified global optimum. These shared labels permit paired comparison, but are not independent ground-truth certificates of branch identity. A separate 512-start convergence audit retained both branch and event labels on all 640 distinct blind states; Appendix I reports the protocol and limitations.

The public-input methods cap accumulated solver iterations at 120; feature evaluation and offline reference construction are excluded. Nominal-budget costs omit some failed probes and are not total costs. We report safe deliveries and false deliveries separately, so a method cannot improve its apparent reliability merely by abstaining more often; coverage is the number of safe deliveries divided by all transitions. We report both cost conditional on delivery and total cost per evaluated transition, since these quantities answer different questions when coverage differs. The blind audit contains 2,560 transitions organized into 32 split-by-seed units of 80 transitions, nested in eight split groups. We therefore pair methods on the same transitions and also resample the 32 units for the coverage-difference interval. Pointwise transition-level binomial intervals and tests are reported as descriptive summaries; their independence assumption is not used to claim a deployment guarantee.

Nominal-budget frontier. The strict ABAC gate is high-precision but too conservative under transfer. On the first held-out split, the nominal-budget obs-μ GCT comparator delivers $38/320$ transitions and the selected shared-probe stump delivers $40/320$; neither has an observed false identity. On the second held-out split, they deliver $9/320$ and $11/320$, respectively, again with no observed false identities. One eight-block evaluation pool is exploratory. On a separate, unopened evaluation pool, the selected router delivers 637 safe and 26 false identities, versus 732 safe and 24 false for

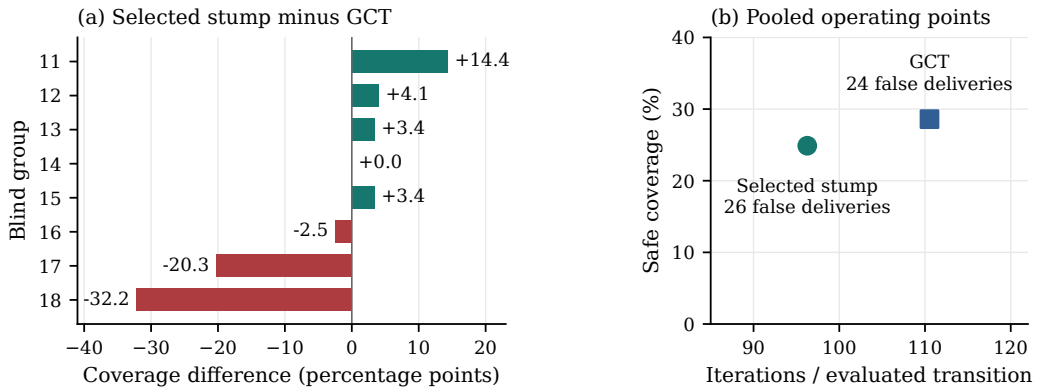


Figure 2: Selected stump versus obs-μ GCT under nominal budgets on eight blind groups, each containing 320 transitions. (a) Safe-coverage differences by group: four are positive, three negative, and one zero. (b) Aggregate safe coverage and unconditional computational cost on all 2560 transitions; labels give false-delivery counts. The aggregate coverage difference is -3.71 percentage points, with hierarchical 95% bootstrap interval $[-14.53, 5.39]$. Correctness is measured against the shared offline reference. Costs are the original recorded counters, which omit some failed probes; this figure does not establish a strict matched-budget frontier.

obs-μ GCT. The router is cheaper by 14.26 iterations per transition, but its safe coverage is lower by 3.71 percentage points. The whole-split bootstrap interval for this pool is $[-14.30, 5.16]$ percentage points and the hierarchical split/seed interval is $[-14.53, 5.39]$. We therefore do not treat the exploratory positive pooled difference as stable cross-split superiority.

Figure 2 shows why aggregation alone is insufficient: the coverage difference changes sign across blind groups. The selected router costs 96.26 iterations per evaluated transition versus 110.52 for GCT, whereas costs conditional on delivery are 99.62 and 110.62. Table 1 reports the corresponding delivery counts and conditional costs. These comparisons describe a risk–coverage–cost tradeoff rather than uniform superiority.

An exploratory public-feature fallback. The calibration-max stump above is retained as a separate baseline. We also froze a public-feature-only fallback by selecting its event-margin threshold on calibration labels using a pointwise one-sided 95% Clopper–Pearson upper bound of 5% on false delivery. The search considered 259 observed thresholds without familywise correction. The selected threshold is $q_* = 1.2333254966819576$; event-complement is used below the threshold and the shared-probe GCT continuation above it. Labels from the blind blocks are never used for routing or threshold selection. On the eight evaluation blocks this variant gives $818/21$ safe/false deliveries out of 2560 transitions, versus $732/24$ for obs-μ GCT, at recorded costs 87.36 and 110.52 iterations per transition. On the pooled set of all 16 blind blocks (5120 transitions), it gives $1811/31$ versus $1587/29$, at cost 86.08 versus 110.58. The paired safe-coverage difference is 4.38 percentage points with a split/seed bootstrap interval of approximately $[1.99, 6.80]$; the fallback has two additional false deliveries. Thus the result is an exploratory coverage–cost improvement with a small observed false-delivery cost, not a risk-certified improvement. Bonferroni correction across the searched thresholds leaves none satisfying the 5% target. The four-seed leave-one-out audit has wide cluster-level uncertainty. Because the fallback design uses information from the exploratory blind tail results, these blind comparisons are exploratory rather than confirmatory.

Table 1: Nominal-budget algorithmic frontier. These rows use the same declared online budget class, but recorded costs can omit failed probes and initialization can access hidden constraints. The event-complement oracle is shown separately because its delivered-case cost is substantially larger. Each delivery cell reports safe/false identities; the blind column pools eight held-out blind groups. Mean and maximum costs are iterations conditional on delivery, using the blind pool for GCT and the selected stump and the held-out test otherwise. The fallback's conditional cost is not reported here; its unconditional cost is reported separately.

Method	Cal.	Test A	Test B	Blind	Mean	Max
ABAC-strict	41/0	8/0	–	–	116.3	117
GCT, obs-μ	121/0	38/0	9/0	732/24	110.6	–
Start-score, $p = 14$	85/0	29/0	3/0	–	101.5	112
Progressive, balanced	99/0	28/0	10/0	–	97.0	112
Progressive, test-max	93/0	30/0	5/0	–	101.5	112
Selected stump	150/0	40/0	11/0	637/26	99.6	–
Exploratory fallback	–	–	–	818/21	–	–
Post-hoc stability	118/0	44/0	11/0	–	99.1	120
Depth-2 tree	155/0	35/0	11/0	–	97.8	120
Complement oracle	115/0	39/0	21/0	–	189.1	195

Public-input controls. GCT-HD uses hidden $\mu^\star$ when forming the uncertainty direction and is excluded as a public-input comparator. The obs-μ variant uses observed parameters for that direction and is the calibration-selected comparator. This control isolates the identified hidden-parameter input; it is not a proof against every possible leakage route. The nominal-budget implementations also use true-constraint initialization and a seed-derived direction in the shared-probe channels that equals the generator's error direction. Across 320 initial seed/cell pairs, 81 initial points differ when only observed constraints are used. In a 320-transition execution audit, seven failed probes have omitted costs and five transitions exceed 120 iterations, reaching 160; none of these five is delivered. Nominal-budget counts remain descriptive, not public-only hard-budget evidence. The prospective comparison removes these access and accounting paths.

Router complexity and transfer. A depth-2 shared-probe tree raises calibration deliveries from 150 to 155 but drops held-out test deliveries from 40 to 35. A post-hoc stability gate improves the already inspected held-out split to $44/320$, but after freezing it produces nine false deliveries on blind seed blocks. The calibration-selected one-stump router is therefore the selected policy for the nominal-budget comparison.

Figure 3 separates calibration performance from the two evaluation splits. Additional routing capacity improves the selection objective without improving either evaluation split.

Event-complement witness cost. A cost-unmatched full ABAC plus 4×20 event-complement oracle improves Test B coverage to $21/320$ with zero observed false identities, but its mean delivered-case test cost is about 189. This supports the two-witness interpretation without being promoted to the main algorithm.

Exploratory fallback and prospective transfer. The exploratory fallback improves safe coverage on the blind evaluation pool and records 21.0% fewer cost units per transition than obs-μ GCT. These nominal-budget counters do not include every failed probe. The pooled blind audit adds two false deliveries. Threshold-search multiplicity, four calibration seeds, and prior inspection of the blind tail results prevent a risk guarantee or confirmatory interpretation. We report it as an exploratory coverage–cost diagnostic.

Appendix G, Figure G.1, separates safe coverage, false deliveries, and unconditional online cost. The pooled false count increases despite the coverage and cost improvements.

With the same q_*, public-only inputs and cumulative cost accounting, eight held-out seeds yield 640 transitions. The fallback and public GCT each give 17 safe deliveries, with four and zero false deliveries, respectively; mean total costs are 85.74 and 110.64. The paired seed-bootstrap 95% interval for their safe-coverage difference is $[-1.09, 1.25]$ percentage points. All four fallback errors

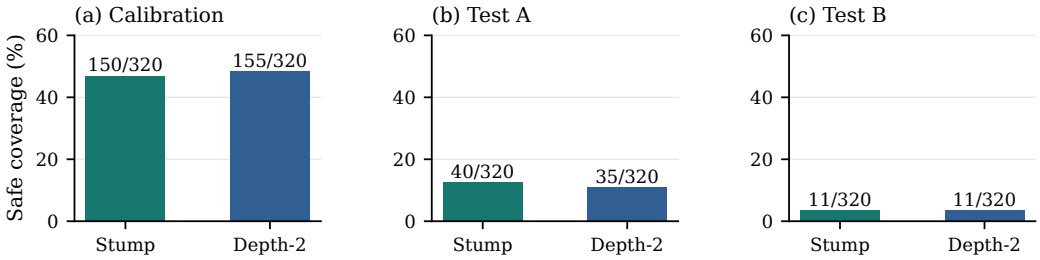


Figure 3: Router-complexity control under the shared-probe protocol. A depth-2 tree increases safe calibration deliveries from 150 to 155, but reduces Test A deliveries from 40 to 35 and leaves Test B deliveries at 11. All six cells contain 320 transitions and zero observed false deliveries. The identical vertical scales show that the calibration gain does not transfer; these counts do not establish a population-level ranking.

occur in the second four-seed block; every method abstains on the first. The fallback routes every arrival to the event-complement channel, so this regime supplies no evidence for a benefit from adaptive routing. Appendix H reports the frozen ablations and reference checks. No thresholds were retuned after correction.

7 Limitations

ABAC does not claim universal optimizer superiority. Zero observed false deliveries are not a distribution-free zero-risk guarantee. The current benchmark is controlled and should be interpreted as a certificate-interface testbed rather than broad deployment evidence. The public-query lower bound is scoped to a per-location evidence interface; a joint global certificate can escape it by changing the primitive. The cost-unmatched event-complement oracle shows remaining upside, but compressing that witness into the 120-step online budget remains the main algorithmic frontier.

8 Conclusion

ABAC makes branch and event identity part of the delivery decision rather than treating repeated local solutions as sufficient evidence. The two-witness necessity result identifies the information required for worst-case safe delivery, while the matched-budget experiments expose a risk–coverage–cost tradeoff and its remaining witness bottleneck. Under nominal budgets, the calibration-max router is cheaper but not more accurate than obs-μ GCT; the fallback gives an exploratory favorable frontier on the same block, with incomplete cost and public-input accounting. On eight held-out seeds, the public-input fallback is cheaper but adds four false deliveries without a safe-coverage gain. The exploratory positive coverage difference does not by itself establish stable transfer to the public-input evaluation. Within the declared public-transcript and finite-budget interface, certification and abstention are therefore distinct outcomes. The empirical gates define a selective operating point, not distribution-free safety or stable coverage superiority.

AI Use Statement

Generative AI tools assisted with mathematical revisions, code revision, and analysis of experimental results. The authors take responsibility for the final content.

References

[1] Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv preprint arXiv:2107.07511*, 2021.

[2] Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, 2024.

[3] Dimitri P Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 4th edition, 2012.

[4] Stephen Boyd and Lieven Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[5] Herman Chernoff. Sequential design of experiments. *The Annals of Mathematical Statistics*, 30(3):755–770, 1959. doi: 10.1214/aoms/1177706205.

[6] C. K. Chow. On optimum recognition error and reject tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46, 1970.

[7] Constantinos Daskalakis, Paul W Goldberg, and Christos H Papadimitriou. The complexity of computing a nash equilibrium. *SIAM Journal on Computing*, 39(1):195–259, 2009.

[8] Rémy Degenne and Wouter M. Koolen. Pure exploration with multiple correct answers. *arXiv preprint arXiv:1902.03475*, 2019. URL `https://arxiv.org/abs/1902.03475`.

[9] Aurélien Garivier and Emilie Kaufmann. Optimal best arm identification with fixed confidence. In *Proceedings of the 29th Conference on Learning Theory*, pp. 998–1027. PMLR, 2016.

[10] Yonatan Geifman and Ran El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

[11] Yonatan Geifman and Ran El-Yaniv. Selectivenet: A deep neural network with an integrated reject option. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2151–2159. PMLR, 2019.

[12] Alex Graves. Adaptive computation time for recurrent neural networks. *arXiv preprint arXiv:1603.08983*, 2016.

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

[14] Reiner Horst and Hoang Tuy. *Global Optimization: Deterministic Approaches*. Springer, 1996.

[15] Steven R. Howard, Aaditya Ramdas, Jon McAuliffe, and Jasjeet Sekhon. Time-uniform Chernoff bounds via nonnegative supermartingales. *Probability Surveys*, 17:257–317, 2020. doi: 10.1214/18-PS321.

[16] Andreas Maurer and Massimiliano Pontil. Empirical bernstein bounds and sample variance penalization. In *Proceedings of the 22nd Annual Conference on Learning Theory*, 2009. URL `https://arxiv.org/abs/0907.3740`.

[17] George L. Nemhauser, Laurence A. Wolsey, and Marshall L. Fisher. An analysis of approximations for maximizing submodular set functions—i. *Mathematical Programming*, 14:265–294, 1978.

[18] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer, 2006.

[19] Martin L Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, 2014.

[20] Vladimir Vovk, Alex Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

A Additional proof details

A.1 Two-witness result

Throughout this section, lower objective values are preferred. For an instance I, define $z \prec_I x_0$ iff z is feasible and $J_I(z) < J_I(x_0)$. Write $\iota(x) = (b(x), e(x))$, and call a delivery identity-false when $\iota(x_0)$ is not the identity required by the task. Assume the global minimum is attained on every compatible instance. Define $v_I(a) = \inf\{J_I(x) : x \text{ feasible}, \ \iota(x) = a\}$; the task requires a global-minimizer identity, using a fixed tie rule if necessary.

Definition A.1 (Dominance-completability)**.** *For a public transcript T, anchor x_0, and compatible location $z \in \mathcal{A}_b(T) \cup \mathcal{A}_e(T)$, z is dominance-completable when there exists an instance $I_{T,z}$ with exactly the same public transcript T and feasible z satisfying $J_{I_{T,z}}(z) < v_{I_{T,z}}(\iota(x_0))$. The construction may change only hidden quantities not fixed by T. This is an existential statement over compatible instances, not a preference claim about the observed instance.*

Proposition A.1 (Conditional witness-to-falsity implication)**.** *If $\iota(z) \neq \iota(x_0)$, the delivery task requires the preferred feasible identity, and z is dominance-completable, then there is a transcript-compatible instance on which delivery of x_0 is identity-false. Without strict identity-class preference (or an explicit tie rule), a different compatible identity alone does not imply false delivery.*

Proof. Dominance-completability supplies an instance $I_{T,z}$ preserving T with $J_{I_{T,z}}(z) < v_{I_{T,z}}(\iota(x_0))$. No feasible point of the anchor identity can attain the global minimum, which exists by assumption. Thus the anchor's identity is not required under any tie rule on the global minimizers. Delivery is therefore identity-false. The converse does not follow: a compatible alternative may be tied, infeasible after hidden constraints, or otherwise not preferred by the task. □

This separates the three levels used below: $\mathcal{A}_b(T), \mathcal{A}_e(T)$ express evidence ambiguity; dominance-completability turns ambiguity into a compatible objective-preference witness; identity falsity is the resulting delivery outcome. The empirical gates do not prove the existential completion property.

Proposition A.2 (Two-witness necessity)**.** *Let T be a public transcript and let a binary selector $d(T) \in \{0, 1\}$ depend only on T, where $d(T) = 1$ means that x_0 is delivered. Suppose every location in $\mathcal{A}_b(T) \cup \mathcal{A}_e(T)$ is dominance-completable. If either $\mathcal{A}_b(T) \neq \varnothing$ or $\mathcal{A}_e(T) \neq \varnothing$, then there exists a transcript-compatible instance on which delivering the anchor candidate is false. Consequently, no deterministic selector can deliver on all such transcripts while maintaining a worst-case zero-false guarantee over the transcript-compatible instance class.*

Proof. We prove the branch case; the same-branch event-complement case is identical after restricting the construction to branch b_0. Suppose $\mathcal{A}_b(T) \neq \varnothing$, and choose $b' \in \mathcal{A}_b(T)$. By dominance-completability there is a compatible location on b' and an instance I' with the same public transcript T on which that location is strictly preferred to the entire identity class of x_0. By the conditional witness-to-falsity proposition, delivery of x_0 is identity-false on I'. Because d observes only T, it takes the same action on the original instance and on I'. If $d(T) = 1$, the delivery is false on I'. Thus worst-case zero-false delivery is impossible unless the branch ambiguity set is empty. Repeating the argument for a compatible same-branch opposite-event point proves the second witness requirement. □

Limitations. The proposition is an ambiguity-semantics result. It does not claim that every implementation can enumerate $\mathcal{A}_b(T)$ or $\mathcal{A}_e(T)$ exactly; ABAC uses public empirical evidence gates as proxies for the two witness types. The necessity result does not prove these gates sound: their false-delivery rates are measured on the controlled benchmark.

A.2 Public-query lower bound

Assumption A.1 (Per-location public query interface)**.** *There are m alternative branch locations and one same-branch opposite-event location, which together with the anchor exhaust the feasible representatives of the distinct identities in this finite family. One query inspects exactly one location. A query that misses the hidden competitor returns $+\Delta$, where $\Delta > 0$. The anchor score is 0. A policy may query at most B locations and then deliver the anchor or abstain.*

Lemma A.1 (Unqueried-location pair)**.** *For any deterministic policy that queries at most $B < m+1$ locations, there exists a location j that is not queried on the all-positive instance.*

Proof. There are $m + 1$ locations. A budget of $B < m + 1$ inspects at most B of them, so at least one location is unqueried. □

Theorem A.1 (Tight per-location lower bound)**.** *Under the per-location public query interface, any deterministic rule that delivers on the all-positive instance I^+ and makes zero false deliveries on every instance in the constructed family requires $B \geq m+1$. The bound is tight. An always-abstain rule has zero false deliveries at zero budget, but does not satisfy the positive-delivery requirement.*

Proof. Let I^+ be the instance in which all $m + 1$ locations have score $+\Delta$; the anchor score 0 is therefore correct. Run any deterministic policy with $B < m + 1$ on I^+. By the lemma, some location j is not queried. Construct I_j^- by changing only location j to score $-\Delta$, keeping every other location at $+\Delta$. The anchor is false on I_j^-. Since j is never queried, every observed query response is $+\Delta$ on both I^+ and I_j^-. The full public transcript is identical, and a deterministic policy must take the same final action. If it delivers, it is false on I_j^-; if it abstains, it fails to deliver on the positive instance. Therefore a deterministic zero-false policy that delivers on the positive family must query all $m + 1$ locations.

For tightness, a policy with $B = m + 1$ queries every location. It delivers the anchor iff every revealed score is $+\Delta$, and otherwise abstains. This separates I^+ from every I_j^-, so the lower bound is tight for the defined interface. □

Limitations. This theorem does not rule out richer primitives that certify multiple locations jointly, randomized selective-risk contracts, learned priors, or global relaxations. It rules out only a per-location public-query shortcut that leaves one possible hidden competitor unresolved while claiming a worst-case zero-false guarantee.

B HARD-FAMILY ILLUSTRATION

Table B.1 summarizes the public-query hard-family construction used in the theorem.

Table B.1: Expected output of the hard-family diagnostic. Budget $B = m$ leaves one hidden location and therefore a paired false instance if the policy delivers on the positive instance; budget $B = m+1$ separates the family.

m	B	locations	unqueried	same transcript?	zero-false?
2	2	3	1	true	false
2	3	3	0	false	true
4	4	5	1	true	false
4	5	5	0	false	true
8	8	9	1	true	false
8	9	9	0	false	true
12	12	13	1	true	false
12	13	13	0	false	true

C ONLINE ROUTING RULE

The selected online policy is a calibration-selected one-stump shared-probe router. The rule below describes the historical residual stump, not the event-margin fallback. Historical initialization and seed-derived uncertainty directions require information excluded from the corrected public interface, and their counters omit some failed probes.

1. Execute a six-start public probe for $p = 11$ local iterations.
2. Compute public probe features, including the probe residual.

3. If the probe residual is at most $3.37053 \cdot 10^{-5}$, route to the event-complement channel; otherwise route to the GCT-style continuation channel.
4. Apply the selected channel's public certificate gate.
5. Deliver only when the selected gate passes; otherwise abstain.

The threshold is selected on calibration data only. Terminal reference labels appear only in post-hoc scoring; this does not remove the separate hidden parameter access in the historical initialization and uncertainty direction.

D THEORY AND EMPIRICAL EVIDENCE

The theoretical results and empirical controls address distinct parts of the delivery problem.

Claim	Evidence	Interpretation
Residual or multistart success does not imply identity certification.	Interface definition and two-witness proposition.	Public-transcript ambiguity, not solver failure.
Per-location evidence has a quantitative cost.	Hard-family theorem and diagnostic table.	Tight for positive delivery in the per-location interface.
Historical shared-probe routing has a nominal-budget operating point.	Held-out frontier: $40/320$ versus corrected GCT $38/320$; new frozen frontier: $637/2560$ safe and 26 false versus $732/2560$ safe and 24 false for GCT.	Recorded cost advantage; not a verified hard-budget comparison.
The comparator's direction uses observed parameters.	Hidden-$\mu^\star$ GCT artifact withdrawn; obs-μ comparator used.	The old direction is excluded; initialization is corrected only prospectively.
Richer routing can overfit.	Depth-2 tree: calibration $155/320$, held-out $35/320$; post-hoc stability gate: $775/2560$ safe and 9 false on blind blocks.	The one-stump router is the selected policy.
Event-complement evidence is valuable but expensive.	Unmatched oracle improves fresh coverage to $21/320$.	Diagnostic, not a matched-budget online method.

E SELECTIVE-RISK INTERVALS

Under independent binomial sampling, for zero-false rows with n delivered identities, the exact one-sided 95% upper bound on the false-delivery rate is $1 - 0.05^{1/n}$. The main held-out zero-false comparison has finite upper bounds: 7.58% for the corrected obs-μ GCT comparator with $38/0$ delivered/false identities and 7.22% for the selected shared-probe stump with $40/0$. On the new frozen block, the selected router has 637 safe and 26 false deliveries, while obs-μ GCT has 732 safe and 24 false deliveries. The comparison is therefore reported as a historical nominal-budget risk–coverage–cost operating point with seed-regime dependence, not as a broad optimizer-dominance claim. These are clustered bootstrap and pointwise binomial diagnostics, not distribution-free bounds for dependent stress transitions or deployment guarantees.

F PAIRED BLIND FRONTIER DIAGNOSTIC

The original eight-block blind pool is retained as an exploratory historical diagnostic. The new frozen evaluation uses eight previously unopened split groups, 32 seeds, and 2560 transitions. The selected router delivers 637 safe and 26 false identities; GCT delivers 732 safe and 24 false identities. The safe-coverage difference is -3.71 percentage points, with a whole-split 95% interval of $[-14.30, 5.16]$ and a hierarchical split/seed interval of $[-14.53, 5.39]$. The selected router costs 96.26 iterations per transition versus 110.52 for GCT. These results support a conditional cost–coverage operating point with seed-regime dependence, not stable coverage superiority.

G EXPLORATORY PUBLIC-FEATURE FALLBACK

The calibration-max stump is not silently replaced by a later rule. We also evaluated a public-feature fallback. Let r_e denote the probe event margin. We selected the threshold on calibration labels by maximizing safe deliveries subject to a pointwise one-sided 95% Clopper–Pearson upper bound of 5% on false delivery, with mean cost as the tie-break. This search examined 259 observed thresholds without a familywise correction:

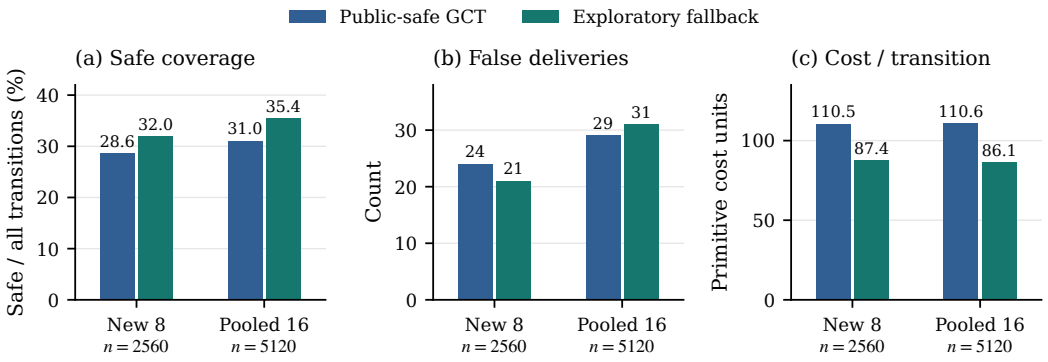


Figure G.1: Exploratory fallback results on eight new blind groups and the pooled sixteen groups, which include those eight. Safe counts are $732, 818$ (new) and $1587, 1811$ (pooled) for obs-μ GCT and the fallback, respectively. Recorded cost is per evaluated transition, not conditional on delivery, and omits some failed probes. The fallback was developed after inspection of earlier blind results; the two aggregates are overlapping, not independent replications.

$$q_f = 1.2333254966819576.$$

The route uses event-complement when $r_e \leq q_f$, and the shared-probe GCT continuation otherwise. New-block labels are used only for scoring. However, the fallback was devised after the earlier $q_{0.99}$ blind result had been examined, so all fallback comparisons below are exploratory, not confirmatory. A Bonferroni correction across the 259 searched thresholds leaves no threshold satisfying the 5% target. The earlier 0.99-quantile route is retained as an artifact ablation and is not used for the main claim.

Table G.2: Exploratory fallback on the eight new blocks (2560 transitions). Cells report safe/false deliveries.

Method	Safe/false	Coverage	Cost/transition
Obs-μ GCT	732/24	28.59%	110.52
Exploratory fallback	818/21	31.95%	87.36

The paired safe-coverage difference is $+3.36$ percentage points on the new blocks and the false-delivery difference is -3. This is an empirically favorable operating point with lower cost, not a uniform-dominance or distribution-free-safety result. It is reported separately from the calibration-max stump to preserve the selection boundary.

As a pooled exploratory summary, the same frozen rule is applied to fresh3–fresh10, which were not used for threshold selection, and pooled with new11–new18. Over 5120 blind transitions the fallback gives $1811/31$ safe/false deliveries at recorded cost 86.08 per transition, while obs-μ GCT gives $1587/29$ at 110.58. The split/seed bootstrap interval for the paired safe-coverage difference is approximately $[1.99, 6.80]$ percentage points. This broader pool has two additional false deliveries. Since the rule was developed after inspecting the earlier blind tail results, this pooled comparison is exploratory and does not provide independent confirmation.

Calibration stability. Leave-one-calibration-seed-out selection chooses thresholds in $[0.745, 1.233]$. The held-out folds contain at most one false delivery each, but their pointwise CP upper bounds remain above 5% because each fold has only about 30–37 deliveries. This is a diagnostic of finite calibration power, not evidence for a deployment-wide risk guarantee. The folds share overlapping training seeds, so they are descriptive and do not support independent cluster-level inference. Requiring a pointwise 5% CP screen separately on every calibration seed also leaves no common threshold. For prospective validation of the stricter event that a seed

contains any false delivery, 59 independent zero-failure seeds would be needed for a one-sided 95% exact-binomial upper bound below 5%. This count is a planning reference, not a result from the present data.

H PROSPECTIVE PUBLIC-INPUT COMPARISON

Before executing any new trajectory, the evaluation was expanded from four to eight unused seeds, 200–207, in two fixed four-seed blocks. The inspected historical records cover seeds 0–87. Each seed has 20 ticks and four uncertainty cells, giving 640 paired transitions but only eight trajectory clusters. The existing event-margin threshold q_* and residual threshold $3.370530775130359 \times 10^{-5}$ are unchanged. Initialization uses observed constraints; the uncertainty direction is normalized observed μ, not the seed-derived generator direction. Failed complement attempts remain charged, and every subsequent local solve is capped by the remaining budget. All methods satisfy the 120-iteration cap under independent solver-call counting. Feature overhead and offline scoring are excluded from this iteration metric.

Table H.3: Prospective corrected implementations on 640 new synthetic transitions. Safe means agreement with the shared offline reference. Costs are total solver iterations per arrival, including abstentions.

Method	Safe	False	Mean cost
Public GCT	17	0	110.64
Frozen fallback	17	4	85.74
Frozen residual stump	14	7	96.78
Always shared GCT	16	3	113.98
Always event complement	17	4	85.74

The first block yields no deliveries for any method. The second yields all counts in the table. The fallback always selects event complement, so its equality with that ablation is exact, not evidence for successful routing. Paired resampling of the eight seed clusters gives a 95% safe-coverage difference interval of $[-1.09, 1.25]$ percentage points against public GCT, and an iteration-cost difference of -24.90 with interval $[-26.04, -23.60]$. False deliveries per arrival increase by 0.625 percentage points, with interval $[0, 1.875]$. These are descriptive intervals on a small, structured seed sample, not a risk guarantee. Corrections and seed changes occur together, so the result does not identify which caused the difference from historical performance.

A secondary same-seed version control retains the historical initialization, uncertainty direction, and stopping logic without changing either threshold. It yields 15 safe and four false deliveries at actual mean cost 85.94, versus 17/0 and 110.64 for its GCT comparator. Six historical fallback transitions exceed 120 iterations, all abstentions. Thus neither implementation recovers the historical coverage advantage on these states; the public corrections alone do not explain the negative transfer result. This version control was specified after the primary comparison and is diagnostic, not a second independent confirmation.

The 160 distinct new states each use 192 deterministic feasible starts for their reference. Extending every state with 320 additional feasible starts changes no branch/event label; the expanded run accepts 50,857 of 51,200 starts. No primary labels are replaced. The label is the sign pattern of the first four coordinates and cap activity, not a unique stationary branch or a certified optimum.

I REFERENCE STABILITY ANALYSIS

Blind branch/event labels are defined by the lowest-objective accepted SLSQP solution among 192 deterministic feasible starts under each hidden true state, using the benchmark objective, constraints, and residual filter. This offline multistart reference is heuristic, not a certified global optimum; obs-μ GCT is only a comparator.

On the 640 blind seed/tick states (32 seeds, 20 ticks each), we ran 512 more feasible starts per state with the same SLSQP setup and a separate random-number stream. The added run accepted 327,068 starts and rejected 612. Its best solution retained the branch, event, and joint identity labels

on 640/640 states. Mean and maximum objective improvements were 9.48×10^{-14} and 3.85×10^{-12}; none exceeded 10^{-8}.

This supports label stability to an expanded start set, not global optimality, hidden-state model validity, or physical ground truth. The audit did not change frozen labels, policies, or routing decisions.

For the prospective set, all 160 states retain their joint labels under the expanded references. The largest objective improvement is 3.94×10^{-12}, and the five method-level safe/false counts are unchanged after rescoring. The archived 3,200 method-transition candidates were first screened under the hidden true parameters: 936 were feasible and 732 passed the reference acceptance filter. Each feasible candidate was then refined once with the same SLSQP configuration; all 936 refinements were accepted. The candidate union changed no joint label and produced no larger improvement than the expanded-start audit. These are offline checks of this fixed prospective collection, not independent samples or a global-optimality certificate; the 160 states remain nested in eight seed trajectories.

J ADDITIONAL REFERENCE STABILITY ANALYSIS

The archived table contains 192 frozen transitions, collected under two methods and three random seeds. For each method–seed pair, the evaluation covers four combinations of uncertainty level and cap, with eight transitions in each combination. The audit leaves targets, labels, and policies unchanged. It summarizes the records separately by method, seed, uncertainty level, and cap, and reports costs both per transition and per delivered case.

For ABAC, the pooled result is 40/96 safe deliveries with zero observed identity-false deliveries, total cost 6410, mean transition cost 66.77, and mean delivered-case cost 65.58. Across the twelve seed/cell clusters, safe coverage ranges from 12.5% to 75.0% and cluster total cost from 457 to 632. For GCT-HD, the corresponding values are 0/96, total cost 10652, and mean transition cost 110.96; no delivered-case mean is defined. These are descriptive stability and cost diagnostics over the archived records, not independent generalization tests. GCT-HD is the archived privileged baseline, not the obs-μ comparator used by the main comparison.