Supplementary Material: Prediction Is Not Permission Statistical Execution Authority for AI Agents

Anonymous submission

Overview

This supplement follows the current evidence chain. Section S1 gives full proofs, including the dependence-free nested decomposition and split-local certificate. Section S2 specifies the algorithm and distinguishes its guarantee from sourcewise CP, SCoRE, end-to-end CP, and the earlier full-family EPV-Opt. Section S3 records the sequential v2/v3/v4 study lineage and the frozen v4 one-time sealed protocol. Sections S4–S5 report every registered K, the v3 zero-coverage failure, task certificates, correlated replacement terms, and pipeline versioning. Section S6 gives hashes, commands, and evidence boundaries. Section S7 reports a disjoint task-specific-threshold extension and its preserved zero-gain Holm result. Section S8 gives the post-outcome candidate-order, threshold-conditioned, and matched-semantics audits. Section S9 reports the registered AgentDojo static and dynamic evidence sequence, including its negative studies, pooled calibration certificate, and independent holdout.

The former finance and broad proxy inventory is excluded from this reviewer PDF. Its source remains in `supplement_legacy_epv.tex` and is not evidence for the current claim.

S1 Formal Results and Proofs

Independent Heterogeneous Selected Risk

Theorem 1 (restated). *Let independent candidate k be eligible with probability π_k. Conditional on eligibility it is harmful with probability α_k, and its continuous score has CDF $F_{k,h}$ given harm label $h \in \{0,1\}$. Define*

$$Q_k(s) = (1-\pi_k) + \pi_k\{(1-\alpha_k)F_{k,0}(s) + \alpha_k F_{k,1}(s)\}.$$

Conditioned on an eligible winner clearing threshold λ,

$$R_{sel}(\lambda) = \frac{\sum_k \pi_k\alpha_k \int_\lambda^\infty \prod_{j\neq k} Q_j(s)\, dF_{k,1}(s)}{1-\prod_k Q_k(\lambda)}.$$

Proof. For fixed k, condition on it being eligible, harmful, and attaining score $s \geq \lambda$. It wins precisely when each competitor is ineligible or has eligible score below s. Independence gives $\prod_{j\neq k} Q_j(s)$. Multiply by $\pi_k\alpha_k$, integrate over $F_{k,1}$, and sum over mutually exclusive winners. Authorization fails only when every candidate is ineligible or below λ, which has probability $\prod_k Q_k(\lambda)$. Continuity removes tie terms. □

The identity is a mechanism model, not the experimental dependence assumption. The nested planners are handled by the next result.

Nested Correlated Candidate Expansion

Proposition 1 (restated). *Let W indicate that expanding a candidate set replaces the old winner. Then*

$$R_{new} - R_{old} = \Pr(W)\{\Pr(H_{new}=1 \mid W) - \Pr(H_{old}=1 \mid W)\}.$$

For nested counts $K_1 < \cdots < K_m$, summing this identity over transitions telescopes exactly to $R_{K_m} - R_{K_1}$.

Proof. On W^c, both systems execute the same action. On W, the expanded system executes the new winner and the old system executes the displaced winner. Expanding both risks over W and W^c cancels the unchanged term. Applying the same equality at every nested transition and summing cancels intermediate risks. No independence, score continuity, or parametric model is used. □

Threshold-Conditioned Candidate Expansion

Proposition 2 (restated). *For paired old and new selectors, let A_j be authorization, H_j the registered binary loss, $C_j = \mathbb{E}[A_j]$, $J_j = \mathbb{E}[A_jH_j]$, and $R_j = J_j/C_j$. Let W denote winner replacement. Then*

$$\begin{aligned} J_1 - J_0 = {} & \mathbb{E}[A_0A_1W(H_1-H_0)] \\ & + \mathbb{E}[A_0A_1(1-W)(H_1-H_0)] \\ & + \mathbb{E}[(1-A_0)A_1H_1] - \mathbb{E}[A_0(1-A_1)H_0]. \end{aligned}$$

For $C_1 > 0$,

$$R_1 - R_0 = \frac{J_1-J_0}{C_1} + R_0\frac{C_0-C_1}{C_1}.$$

Proof. Partition paired states into retained authorization with replacement, retained authorization without replacement, authorization entry, and authorization exit. Expanding

$A_1H_1 - A_0H_0$ on these four disjoint events gives the joint-risk identity. Since $J_j = C_jR_j$,

$$C_1(R_1 - R_0) = (J_1 - J_0) + R_0(C_0 - C_1),$$

which gives the conditional-FAR identity. No independence or score model is used. The implementation verifies both equalities to numerical tolerance for every paired audit row. □

Pairwise Ranking Is Non-Identifying

Proposition 3 (restated). *For $K > 2$, $q = \Pr(S_{harm} > S_{safe})$ does not generally identify the harm risk of the top-scoring candidate.*

Proof. Let safe scores be uniform on $[0, 1]$ and source harm probability be α. In construction A, every harmful score equals $1/2$. In construction B, a harmful score is 0 or 1 with equal probability. Both have $q = 1/2$, but their selected risks are respectively

$$\{\tfrac{1}{2}(1+\alpha)\}^K - \{\tfrac{1}{2}(1-\alpha)\}^K$$

and

$$1 - (1 - \tfrac{1}{2}\alpha)^K + (\tfrac{1}{2}\alpha)^K.$$

They differ for $K > 2$. Small continuous perturbations remove point masses without closing the gap. □

Split-Local Finite-Sample Control

Theorem 2 (restated). *Assume development, calibration, and deployment episodes are independent IID draws from one registered distribution. Freeze the selector and use development to select one threshold. Before calibration, register tasks $\mathcal{G}$, high-impact tasks $\mathcal{G}_{hi} \subseteq \mathcal{G}$, ordinary budgets α_g, high-impact budget α_{hi}, and minimum support. Apply one-sided exact binomial bounds with total error δ divided across $M = |\mathcal{G}| + |\mathcal{G}_{hi}|$ events. With probability at least $1-\delta$, every task declared certified satisfies its ordinary risk budget and every active high-impact bound. Any deployment mixture over certified tasks therefore satisfies the largest active budget.*

Proof. The threshold and complete selector are fixed before calibration outcomes. For each task ordinary loss and each registered high-impact task loss, the one-sided Clopper–Pearson bound at error δ/M covers its conditional risk for every random support count. A union bound makes all M bounds valid simultaneously. The route may then select every supported task whose bounds pass. On the simultaneous event, every selected task has risk at most its budget. The conditional risk under any mixture of selected tasks is a convex combination of their risks and cannot exceed the largest active budget. □

Development may optimize empirical coverage because calibration remains independent. Calibration is used only to certify and route the fixed threshold.

Full-Family Baseline

Full-Family Baseline Theorem. *If a finite threshold-route family is frozen before calibration and every inspected pooled, group, and high-impact event receives a simultaneous exact bound, then any calibration-feasible policy selected from that family inherits all registered bounds with probability at least $1 - \delta$.*

The proof is the same per-event exact coverage plus union bound. The v2/v4 full-family baseline inspects 418 events. It is valid but spends substantially more multiplicity than split-local certification.

Support Cost

Corollary 1 (restated). *For zero observed harms, risk budget α, error δ, and M simultaneous events,*

$$n \geq \left\lceil \frac{\log(\delta/M)}{\log(1-\alpha)} \right\rceil.$$

Proof. The zero-harm one-sided exact upper bound at error $\eta = \delta/M$ is $1 - \eta^{1/n}$. Solving $1 - \eta^{1/n} \leq \alpha$ yields the expression. □

At $\alpha = 0.10$, split-local $M = 25$ needs 59 zero-harm examples, v3 $M = 40$ needs 64, and full-family $M = 418$ needs 86. At $\alpha = 0.05$, the respective requirements are 122, 131, and 177.

S2 Algorithm and Guarantee Semantics

Split-Local EPV-Opt

Training. Fit the selected-action scorer and freeze candidate generation, eligibility, provenance, tie-breaking, and nested candidate order.

Development. From a training-fixed threshold grid, select the largest-coverage threshold satisfying the registered empirical design ceiling. This step does not claim a risk guarantee.

Fresh calibration.

1. Run the complete selector and fixed threshold.
2. For every task, compute an ordinary conditional-risk upper bound.
3. For every structurally registered high-impact task, also compute its high-impact upper bound.
4. Apply the registered simultaneous correction over all active events.
5. Authorize every supported task whose active bounds pass; send other score-clearing tasks to review.

Deployment. A selected action below threshold abstains. A score-clearing action from a certified task authorizes. A score-clearing action from an uncertified or unsupported task reviews. Any pipeline fingerprint mismatch invalidates the certificate.

Objects Controlled by Baselines

Sourcewise CP. It certifies sources before argmax. Its selected set can differ completely from the post-selection route and does not inherit a new post-argmax certificate.

SCoRE-SDR. It controls expected selected-discovery risk. Its observed conditional FAR and test UCB are common diagnostics, not its formal target.

End-to-end CP. It calibrates the final selected score and controls pooled conditional FAR. It does not require every authorized local task to have its own bound.

Full-family EPV-Opt. It jointly searches thresholds and coarse routes on calibration under a 418-event correction. It supplies the same authority semantics but is more conservative.

Split-local EPV-Opt. It spends a development split to fix the threshold, then uses calibration only for 20 ordinary task bounds and five structurally applicable high-impact bounds. Its contribution is local certificate resolution, not a new ranker.

S3 Sequential Protocol and Integrity

Study Lineage

Table S1: Sequential studies. Each fresh round was frozen before its own calibration/test outcomes. V2 results were prior development for v3/v4 rather than fresh confirmation.

Round	Episodes	Executions	Role
v2	4,500	36,000	pilot/development
v3	4,000	32,000	fresh negative certificate
v4	4,000	32,000	fresh primary hidden test

V2 established the pilot and exposed coarse-group masking. V3 used disjoint seeds but registered high-impact bounds for all tasks, which exhausted support. V4 was frozen only after preserving that failure and used another disjoint seed family. The v4 base seed is 40265000; outcome-free audit found no overlap with v2 or v3.

Frozen V4 Protocol

V4 covers 20 tasks. Each task has 100 fresh calibration and 100 one-time test episodes. Every episode executes eight unique candidate plans, for 4,000 episodes and 32,000 candidate executions. The v2 training scorer and development threshold rule are fixed. V4 calibration certifies 20 ordinary task events plus high-impact events for the five external-side-effect tasks. The family size is 25, ordinary budget 0.10, high-impact budget 0.05, total error 0.05, and minimum task support 80.

The protocol and candidate-generator SHA-256 prefixes are `a1f63730fa84a813` and `39b432cca406188b`. Full digests are stored in the frozen YAML, metadata, and artifact-integrity JSON.

Outcome and Loss Boundary

The binary proxy marks a nonempty state-changing plan as adverse when it earns no task reward. Utility subtracts action and latency cost. High impact is a pre-registered task-semantic label for external-side-effect tasks, not observed privacy, financial, or irreversible loss. Invalid DOM actions remain failed candidates and are not removed.

S4 Sealed MiniWoB Results

All Registered Candidate Counts

Table S2: Fresh v4 one-time test results across nested candidate counts. Split is split-local EPV-Opt; Full is the 418-event baseline.

K	Sel. FAR	CP Dir.	CP FAR	Full Dir.	Full FAR	Split Dir.	Split FAR
1	0.896	0.000	0.000	0.000	0.000	0.000	0.000
2	0.814	0.051	0.000	0.000	0.000	0.000	0.000
4	0.684	0.280	0.034	0.099	0.025	0.194	0.000
8	0.684	0.280	0.037	0.099	0.025	0.194	0.000

At K1/2, split-local certification returns no direct authority. At K4/8, it authorizes 19.4%, reviews 8.6%, and observes zero adverse direct actions. End-to-end CP retains higher direct coverage but only a pooled certificate.

Preserved V3 Failure

Table S3: The v3 failure and v4 correction at $K = 8$. The v3 zero-coverage row is preserved rather than omitted.

Design	Family	Direct	Review	FAR [UCB]	Utility
v3 all-task severity	40	0.000	0.279	0.000 [1.000]	0.00
v4 structural severity	25	0.194	0.086	0.000 [0.008]	367.85

V3 tested high-impact bounds on every task. With 100 examples per task and 40 simultaneous events, even zero high-impact harms had UCB 0.0647, above 0.05. V4 does not loosen that budget. It registers high-impact bounds only for tasks where the label can occur, reducing the active family to 25.

Task-Level Certificates

Table S4: Every score-clearing v4 task at $K = 8$. Test A/H is direct authorizations/harm; Review/H reports reviewed selected actions and their known reset-state outcomes.

Task	Cal. N	Cal. harm	UCB	Test A/H	Review/H	Route
click-checkboxes-large	100	5	0.147	0/0	100/2	Review
click-checkboxes-soft	65	20	0.492	0/0	63/14	Review
click-checkboxes	100	0	0.060	100/0	0/0	Authorize
click-option	99	0	0.061	100/0	0/0	Authorize
enter-password	100	0	0.060	100/0	0/0	Authorize
login-user-popup	7	4	0.960	0/0	9/5	Review
login-user	93	0	0.065	88/0	0/0	Authorize

Four tasks certify and authorize. The failing `login-user-popup` task has 4 adverse outcomes among 7 calibration actions and routes all nine score-clearing test actions to review. This directly closes the coarse-account masking failure observed in v2. Tasks with no score-clearing support abstain.

S5 Correlated Candidates and Audit

Exact Replacement Decomposition

Table S5: Dependence-free unthresholded replacement terms on the fresh v4 test. New winners are less adverse than displaced winners for K1–2 and K2–4.

Transition	$P(W)$	Old FAR	New FAR	New harm| W	Old harm| W
1→2	0.573	0.896	0.814	0.834	0.977
2→4	0.404	0.814	0.684	0.637	0.959
4→8	0.091	0.684	0.684	0.956	0.956

The observed risk changes equal the product of replacement probability and the conditional adverse-proxy gap at every transition, and they telescope from K1 to K8. This result is aligned with correlated deterministic planners; it does not invoke Theorem 1's independent-candidate model.

Versioning

The certificate fingerprint includes the protocol, candidate count, scorer digest, threshold rule, task and high-impact registries, support requirement, and tie rule. Any mutation raises a stale-certificate error. Changing which tasks receive high-impact bounds is therefore a new protocol and required the fresh v4 calibration/test collection.

S6 Artifacts, Commands, and Boundaries

Artifact Integrity

The read-only v4 audit passes all checks: exact task, episode, split, candidate, plan, method, and decision counts; protocol/generator hashes; task routes; replacement telescoping; no duplicate candidate IDs; no live API; and one-time sealed-holdout use.

The v5 audit also retains the exact registered certifier source in its `frozen_sources/` subdirectory. Its SHA-256 matches the preregistration. The current shared module adds only the later useful-completion lower-bound function used by v9/v10; the audit verifies AST identity after removing that named additive extension.

Table S6: SHA-256 prefixes for primary frozen v4 artifacts. Full hashes are in `artifact_integrity.json`.

Artifact	SHA-256 prefix
candidate_outcomes.csv	`6b025c459e74c348`
metrics.csv	`daf4c56fb2fd9b03`
decisions.csv	`58cff6a53bad4e2f`
task_certificates.csv	`459cd07223497000`
calibration.json	`cfe94af8a2fa2645`
metadata.json	`38e3503d9756d7dd`

Exact Commands

```
make sea-v4-analyze
make sea-v4-audit
make sea-v5-analyze
make sea-v5-audit
make sea-v6-audit
make agentdojo-dynamic-tables
make sea-v6-tables
make tables
make paper
make verify
```

Ordinary reproduction consumes frozen CSV files. The explicit collection target is excluded from verification and refuses to overwrite an existing candidate artifact.

Evidence Boundary

The fresh v4 evidence improves local resolution and authorization coverage. Candidates remain deterministic planners in resettable MiniWoB. The registered adverse-action taxonomy remains a task proxy. The method assumes a frozen selector and IID calibration/deployment distribution. No production, drift-robust, pointwise, privacy, financial, or irreversible-harm guarantee is claimed. End-to-end CP and SCoRE retain higher coverage/AV under different certificate targets; sourcewise CP observes zero adverse outcomes at lower coverage without a post-argmax certificate. Section S9 adds a bounded second environment but does not turn this into a multi-model deployment benchmark.

S7 V5 Taskwise Multiplicity Extension

After v4, a separate frozen study tested K-specific task thresholds and exact Holm correction. V2 train/development fixes every threshold and active event before v5. Infinite-threshold tasks always abstain; low-support registered events stay in the family with $p = 1$. Exact lower-tail tests evaluate $H_0 : R_g \geq 0.10$, and Holm controls family-wise false certification at 0.05 under arbitrary dependence. Taskwise Bonferroni uses identical thresholds and outcomes. V5 then collects 300 calibration and 100 one-time test episodes for each of 20 tasks: 8,000 disjoint-seed episodes and 64,000 candidate executions. Both pre- and post-collection integrity audits pass.

Table S7: V5 $K = 8$ routes. The global-threshold baseline and both taskwise methods use the same fresh v5 outcomes. UCB is a one-sided 95% aggregate test diagnostic, not the calibration certificate.

Method	Direct	Review	A/H	FAR [UCB]	Utility	AV
Global threshold	0.245	0.037	491/4	0.008 [0.019]	451.11	875.02
Taskwise Bonf.	0.249	0.000	499/4	0.008 [0.018]	458.61	882.52
Taskwise Holm	0.249	0.000	499/4	0.008 [0.018]	458.61	882.52

Task-specific thresholds authorize 499 of 2,000 test actions (0.2495), versus 491 (0.2455) for the v5 global-threshold split-local baseline. Both taskwise corrections observe 4/499 adverse outcomes (FAR 0.0080; aggregate test UCB 0.0183). Holm and Bonferroni certify exactly the same five tasks and produce zero Holm-only decisions because all active p-values already pass Bonferroni. This registered null result rules out an empirical multiplicity-efficiency claim here: the correction is not binding for this active family; task eligibility and support are.

The nonzero-loss case is `click-checkboxes-large`: 6/300 calibration adverse outcomes and 4/100 test adverse outcomes, with task-level test UCB 0.0892. This demonstrates that exact task certification need not rely on a zero-error calibration sample, while remaining close enough to the 0.10 budget to preclude a robustness or drift claim. The formal certificate is calibration-derived; 0.0892 is a test diagnostic. The other certified tasks cover checkbox selection, option selection, password entry, and login; their test adverse outcomes are zero. All external-side-effect tasks retain infinite thresholds and abstain; v5 establishes no high-impact execution authority. Exact task outcomes and the vector diagnostic figure remain in the frozen artifact bundle rather than the reviewer PDF.

S8 V6 Robustness and Matched-Semantics Audit

V6 is explicitly post-outcome. It uses the exhaustive v5 candidate outcomes to answer reviewer-facing robustness questions and is not a fresh confirmatory study. The audit registers 64 deterministic candidate-ID permutations, applies the same permutation across every episode, and evaluates nested prefixes $K \in \{1, 2, 4, 8\}$. The first permutation is the frozen study order; the other 63 are generated from a fixed outcome-independent seed.

Table S8: Candidate-order robustness on 64 registered permutations. Candidate order materially changes intermediate-K risk and utility, but every permutation has lower uncensored adverse-proxy risk at K4 than K1. K8 is invariant because it contains all candidates. This is a post-outcome audit.

K	Perm.	FAR mean	FAR range	Utility mean	Utility range
1	64	0.944	[0.897, 0.976]	-335.6	[-468.0, -192.3]
2	64	0.896	[0.817, 0.961]	-450.7	[-601.4, -258.3]
4	64	0.803	[0.680, 0.936]	-516.6	[-694.0, -304.5]
8	64	0.678	[0.678, 0.678]	-431.9	[-431.9, -431.9]

Table S9: Threshold-conditioned decomposition at the frozen threshold. The table reports joint-risk contributions from retained replacement and entry, their total, and the conditional-FAR denominator correction. The range across permutations includes positive and negative conditional-FAR changes even when the mean change is negative.

Transition	Old FAR	New FAR	Δ FAR mean [range]	Repl. ΔJ	Entry ΔJ	Total ΔJ	Denom.
K1–2	0.453	0.341	-0.112 [-0.511, 0.126]	-0.002	0.013	0.011	-0.259
K2–4	0.341	0.187	-0.154 [-0.524, 0.115]	-0.010	0.014	0.004	-0.183
K4–8	0.187	0.042	-0.144 [-0.626, 0.005]	-0.024	0.005	-0.020	-0.074

The row-wise implementation verifies both equalities in Proposition 2 to numerical tolerance. Retained-stable and authorization-exit terms are exactly zero in this fixed-threshold audit because adding a candidate cannot lower the selected maximum score; retained replacement, entry, and denominator shift remain observable.

Table S10: Matched-semantics v5 comparison. All rows target simultaneous task-local conditional adverse-action proxy risk with the same budgets and minimum support. Task thresholds add eight test authorizations over one pooled development threshold; Holm adds none over Bonferroni.

Method	Tasks	A/H	Coverage	FAR [UCB]	Utility	Review
Pooled threshold	5	491/4	0.245	0.008 [0.019]	451.1	0.037
Task thresholds + Bonf.	5	499/4	0.249	0.008 [0.018]	458.6	0.000
Task thresholds + Holm	5	499/4	0.249	0.008 [0.018]	458.6	0.000

This comparison isolates the benefit of task-specific thresholds without comparing unlike guarantees. The gain is real but small: the evidence supports certificate resolution and audit semantics, not a large algorithmic frontier improvement. CSA is not inserted as a numerical baseline because its formal object is an adaptive online pathwise stream, whereas this audit is sealed IID; the distinction is stated in Related Work rather than hidden by a common label.

Table S11: Post-outcome factor audit for the $K = 8$ coverage gap. Holding the development threshold fixed while retaining the original $M = 418$ correction changes nothing; reducing the fixed event family to $M = 38$ increases coarse group coverage, while the $M = 25$ task-local route trades coverage for finer local semantics. This is a diagnostic, not a fresh confirmatory result.

Design	M	Direct	Review	A/H	FAR [UCB]	Utility
Calibration search	418	0.100	0.183	199/3	0.015 [0.038]	184.7
Dev threshold, retained correction	418	0.100	0.183	199/3	0.015 [0.038]	184.7
Dev threshold, exact family	38	0.282	0.000	565/24	0.042 [0.059]	476.3
Dev threshold, task-local	25	0.245	0.037	491/4	0.008 [0.019]	451.1

The factor audit prevents attributing the full-family versus split-local gap to threshold timing alone: the retained-correction rows authorize 199/2,000 in both cases. The larger fixed-family reduction is responsible for the coarse group coverage increase, while task-local routing has the stricter object.

S9 AgentDojo Static and Dynamic Transfer Sequence

This section reports a sequential second-environment study in AgentDojo 0.1.35 (benchmark v1.2.2). Every action executes inside the resettable benchmark; no real account, payment, email, production service, or host tool is exposed. The sequence is intentionally not collapsed into one success claim: v7 is static proposal replay, v8 is a negative dynamic/provider-stress study, v9 calibrates one pooled certificate, and v10 is its separate one-time transfer test.

Table S12: Registered AgentDojo evidence lineage. Each row has a distinct role. The static and v8 studies are negative, v9 establishes only a pooled calibration certificate, and v10 does not confirm that certificate on its independent holdout.

Study	Registered scale	Outcome	Interpretation
v7 static	1,166 clusters; 75,165 context executions	No feasible certifier; zero test authorizations	Static transfer negative
v8 dynamic	120 clusters; 1,440 trajectories	Calibration infeasible; 709/720 test errors	Provider stress (667 zero-turn rows)
v9 calib.	90 clusters; 1,080 trajectories	Pooled dual certificate feasible	Registered test unopened: cost cap
v10 holdout	90 clusters; 1,080 trajectories	Pooled transfer not confirmed	Independent negative holdout

V7 static multi-model replay. Opus, Sonnet, and Fable generated proposals for 1,166 registered clusters. The artifact contains 2,904 model calls, 8,712 candidates, and 75,165 candidate–context executions. Every registered certifier is infeasible and authorizes zero test clusters. This controlled negative result tests proposal diversity without claiming interactive-agent transfer.

V8 dynamic provider stress. The first dynamic protocol runs the three model families under direct and cautious policies. Calibration support is 33 with 8 adverse outcomes, producing UCB 0.395 against the 0.10 budget. The opened test then records 709/720 provider errors and 667 zero-turn rows. These failures are retained as interface-drift evidence and are not attributed to model safety or task competence.

V9 pooled dual calibration. V9 freezes six candidates, a 50/50 ordinary/high-impact task mixture, direct validation attacks, distinct ignore-previous deployment attacks, and lexicographic selection by validation eligibility, turns, cost, and candidate ID. On 90 calibration clusters, 84 winners are selected. Authority violations are 0/84 (one-sided bound at event delta 0.025: 0.043); useful completions are 66/84 (lower bound 0.683); four provider errors count against usefulness. Both events pass the registered simultaneous gate. The registered total cap leaves insufficient budget for a complete equal-scale v9 test, so that test is not opened and no test outcome exists.

V10 independent holdout. V10 imports the exact v9 certificate and scientific pipeline without recalibration. Its seed is mechanically derived from the certificate hash. The fresh test completes all 90 clusters and 1,080 trajectories for USD 22.82. The selector authorizes 19 clusters, with 0 authority violations and 16 useful completions. The one-sided 95% authority-risk UCB is 0.146 and the useful- completion LCB is 0.641. The preregistered confirmation requires support at least 40, risk UCB at most 0.10, and utility LCB at least 0.50. It therefore does not confirm transfer despite zero observed violations and a passing utility diagnostic.

Table S13: Pooled v9 calibration certificate and v10 heldout diagnostics. "Bound" is an authority-risk upper bound and a useful-completion lower bound, respectively. The v10 intervals diagnose transfer of the frozen certificate; they are not a replacement calibration certificate.

Split	Support	Auth. viol.	Auth. bound	Useful / bound	Status
v9 calibration	84	0	0.043	66 / 0.683	feasible
v10 one-time holdout	19	0	0.146	16 / 0.641	not confirmed

The sequence supports a narrower claim than a positive external certificate: multi-model dynamic interaction and exact benchmark checkers can instantiate the authorization object, but a feasible pooled calibration certificate can lose direct support on a fresh draw. The result provides no task-local, high-impact, adaptive-attacker, production, or legal-permission guarantee.