Prediction Is Not Permission: Statistical Execution Authority for AI Agents

Anonymous submission

Abstract

In AI-agent pipelines, source-level safety need not survive adaptive selection among plans, tools, and action proposals. We study *statistical execution authority*: when the action selected by a complete agent pipeline has enough evidence to execute. We characterize *selection-induced authority risk* (SIAR) with an exact independent-candidate identity, an exact threshold-conditioned decomposition for conditional false authorization, and a counterexample showing that pairwise score–loss ranking does not identify top-one risk. Split-local EPV-Opt fixes the selector and threshold before calibration, applies simultaneous task-local and high-impact bounds, and routes unsupported actions to review or abstention. A preregistered one-time sealed MiniWoB study contains 4,000 reset-state episodes and 32,000 candidate executions. At $K = 8$, split-local EPV-Opt authorizes 0.194 of actions and observes 0/388 registered adverse authorizations (one-sided 95% test UCB 0.008, below the 0.10 target), while stronger-coverage baselines retain different guarantee semantics. A separate three-model AgentDojo sequence produces a feasible pooled calibration certificate, but its independent holdout does not confirm transfer: only 19/90 clusters authorize, giving 0 observed authority violations but a 0.146 risk UCB above the 0.10 gate; useful completion remains 16/19 with LCB 0.641. No high-impact action earns a statistical certificate. Prediction is not permission: authority must be established after selection.

Introduction

An AI agent can be right and still wrong to act. A tool agent may identify a relevant API call while lacking authority to commit it; a best-of-N planner may generate useful alternatives but select a persuasive harmful action. Prediction accuracy, confidence, and sourcewise calibration describe proposals. Execution requires a separate decision about the action selected by the complete pipeline.

Selection makes this distinction statistical. Let heterogeneous candidates pass role and provenance checks, receive scores, and compete under argmax. Even if every source has bounded marginal risk, the selected winner can have a different risk because selection favors particular score tails. We call the gap *selection-induced authority risk* (SIAR). Candidate count is not inherently harmful: adding a candidate raises risk only when the added action is more harmful than the winner it displaces on replacement events. In our real benchmark, additional candidates help rather than hurt. The relevant object is therefore the score–harm relationship of the complete selector, not K alone.

We propose *Statistical Execution Authority* as the study of when a selected agent action has sufficient statistical and provenance evidence to execute. Calibration (Guo et al. 2017), selective prediction (Chow 1970; El-Yaniv and Wiener 2010), conformal risk control (Bates et al. 2021; Angelopoulos et al. 2024), OPE (Dudík, Langford, and Li 2011; Jiang and Li 2016), and runtime guards address adjacent objects. The composition gap remains: permission rules determine eligibility, while statistical certifiers must control the realized consequence risk of the final selected output.

EPV-Opt instantiates this object. It freezes candidate generation, trusted eligibility and provenance, a score, tie-breaking, and argmax. Development data select one threshold before fresh calibration; simultaneous task and structurally applicable high-impact bounds then determine which score-clearing tasks authorize and which route to review. This split-local design avoids searching thresholds and routes on the same calibration outcomes. Changing the candidate set, registry, score, threshold rule, or task family invalidates the certificate. "Opt" denotes maximum empirical coverage inside the frozen certificate-feasible family; in the split-local case the solution is the monotone rule that authorizes every locally certified task, not a new numerical optimizer or ranker.

Contributions. First, we define post-selection agent authorization and derive an exact heterogeneous selected-risk identity with eligibility and authorization conditioning. Second, we separate unthresholded winner replacement from an exact threshold-conditioned decomposition of joint risk, conditional FAR, authorization entry, exit, and denominator shift. Third, we introduce split-local EPV-Opt, which separates threshold selection from simultaneous task/severity certification and fail-closed versioning. Fourth, we report a sealed MiniWoB study with 32,000 reset-state executions, order and matched-semantics audits, and a preregistered three-model AgentDojo transfer sequence with an independent

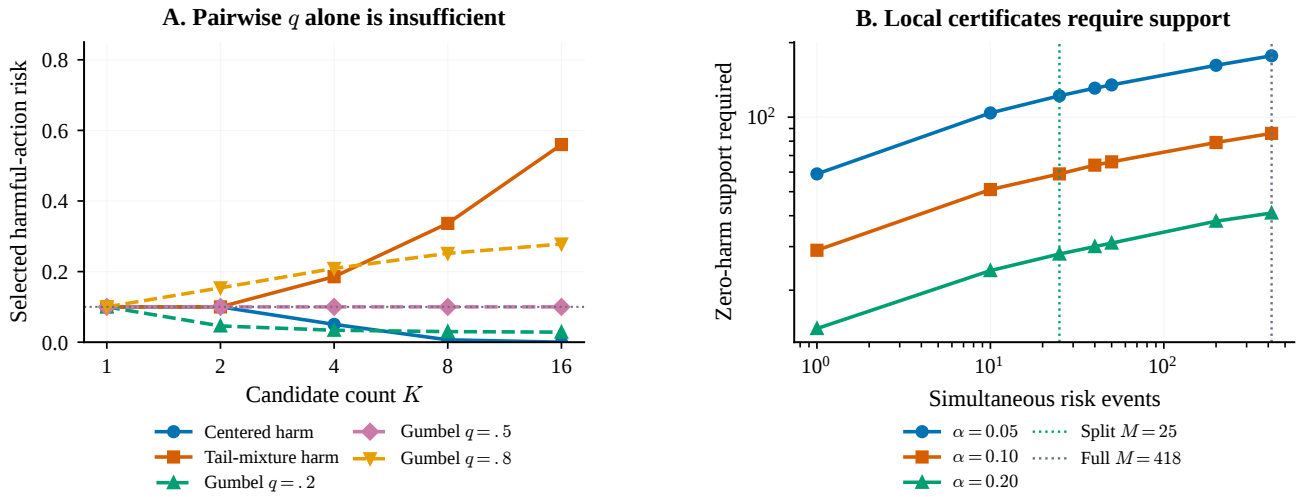

Figure 1: Candidate count alone does not determine selected-action risk. Panel A gives two score constructions with identical pairwise $q = 0.5$ but sharply different top-one risk; Gumbel curves are a registered parametric special case. Panel B shows the zero-harm support cost of simultaneous certificates and marks the split-local $M = 25$ and full-family $M = 418$ designs.

negative holdout.

Post-Selection Agent Authorization

Candidates, eligibility, and selection. At decision time t, a pipeline observes state x_t and structured proposals $\mathcal{P}_t = \{p_t^1, \ldots, p_t^K\}$. Proposal p_t^k records source q_t^k, role ρ_t^k, provenance Π_t^k, action a_t^k, score inputs, uncertainty, and cost. A trusted registry gives eligibility $E_t^k \in \{0, 1\}$; self-claimed roles are not authoritative. A frozen scorer and tie rule select

$$k_t^\star = \arg \max_{k: E_t^k = 1} S_t^k, \qquad S_t^k = s_\theta(x_t, p_t^k).$$

The winner is an action hypothesis; selection does not confer permission.

Authority risk. Let $H_t^k \in \{0, 1\}$ denote a registered adverse-action loss. At threshold λ, the authorization event is $A_{t,\lambda} = \{S_t^{k_t^\star} \geq \lambda\}$. The target risk is conditional false authorization

$$R_{sel}(\lambda) = \Pr(H_t^{k_t^\star} = 1 \mid A_{t,\lambda} = 1),$$

reported separately from joint risk $\Pr(A \cap H)$ and coverage $C(\lambda) = \Pr(A = 1)$. Sourcewise bounds $\Pr(H_t^k = 1 \mid A_t^k = 1) \leq \alpha$ do not generally imply $R_{sel} \leq \alpha$ after a new argmax.

Global, local, and high-impact authority. For a pre-registered risk group $G = g$, local authority requires $R_g = \Pr(H = 1 \mid A = 1, G = g) \leq \alpha_g$. A second label $H^{hi} \leq H$ marks a task-semantic high-impact proxy with budget α_{hi}. A pooled certificate does not imply either local statement. The decision is therefore $d_t \in \{\textsc{Authorize}, \textsc{Review}, \textsc{Abstain}\}$: review denotes insufficient local evidence, while abstention denotes failed score, eligibility, or support.

Authorization value. All-abstention is statistically safe but operationally empty. We report direct authorization, review load, missed opportunity, verification cost, and

$$\text{AV} = H_{avoid} + C_{avoid} + R_{reduce} - M_{opp} - V_{cost}.$$

The terms denote harm, cost, and risk avoided minus missed opportunity and verification overhead. EPV-Opt optimizes certified calibration coverage; AV remains an outcome metric rather than an objective tuned on the same calibration split.

EPV-Opt: Split-Local Execution Routing

EPV-Opt treats the complete selected-output map as the certification unit (Figure 2). Generation, eligibility, provenance, scoring, tie-breaking, and argmax are frozen. Development outcomes may choose one threshold; independent calibration outcomes are reserved for local certification.

Split selection and certification. Training fixes a selected score $S_i^\star$ and threshold grid Λ. On a development split, EPV-Opt selects the highest-coverage threshold $\hat{\lambda}$ whose empirical FAR is at most a registered design ceiling. This stage is not a certificate. On independent calibration data, the threshold remains fixed. For each registered task $g \in \mathcal{G}$, EPV-Opt computes a one-sided ordinary-risk bound; a high-impact bound is added only for tasks in the pre-registered family $\mathcal{G}_{hi}$ where that label can occur.

Local route optimization. Let B_g^+ and $B_g^{hi,+}$ denote bounds protected simultaneously over $M = |\mathcal{G}| + |\mathcal{G}_{hi}|$ events. The direct route is

$$\widehat{\mathcal{G}}_{auth} = \{g : N_g \geq n_{min},\ B_g^+ \leq \alpha_g,\ g \notin \mathcal{G}_{hi} \text{ or } B_g^{hi,+} \leq \alpha_{hi}\}.$$

Authorizing every certified task maximizes calibration coverage over task subsets because simultaneous task bounds compose under any mixture of the accepted tasks. AV remains an outcome metric, not a calibration objective.

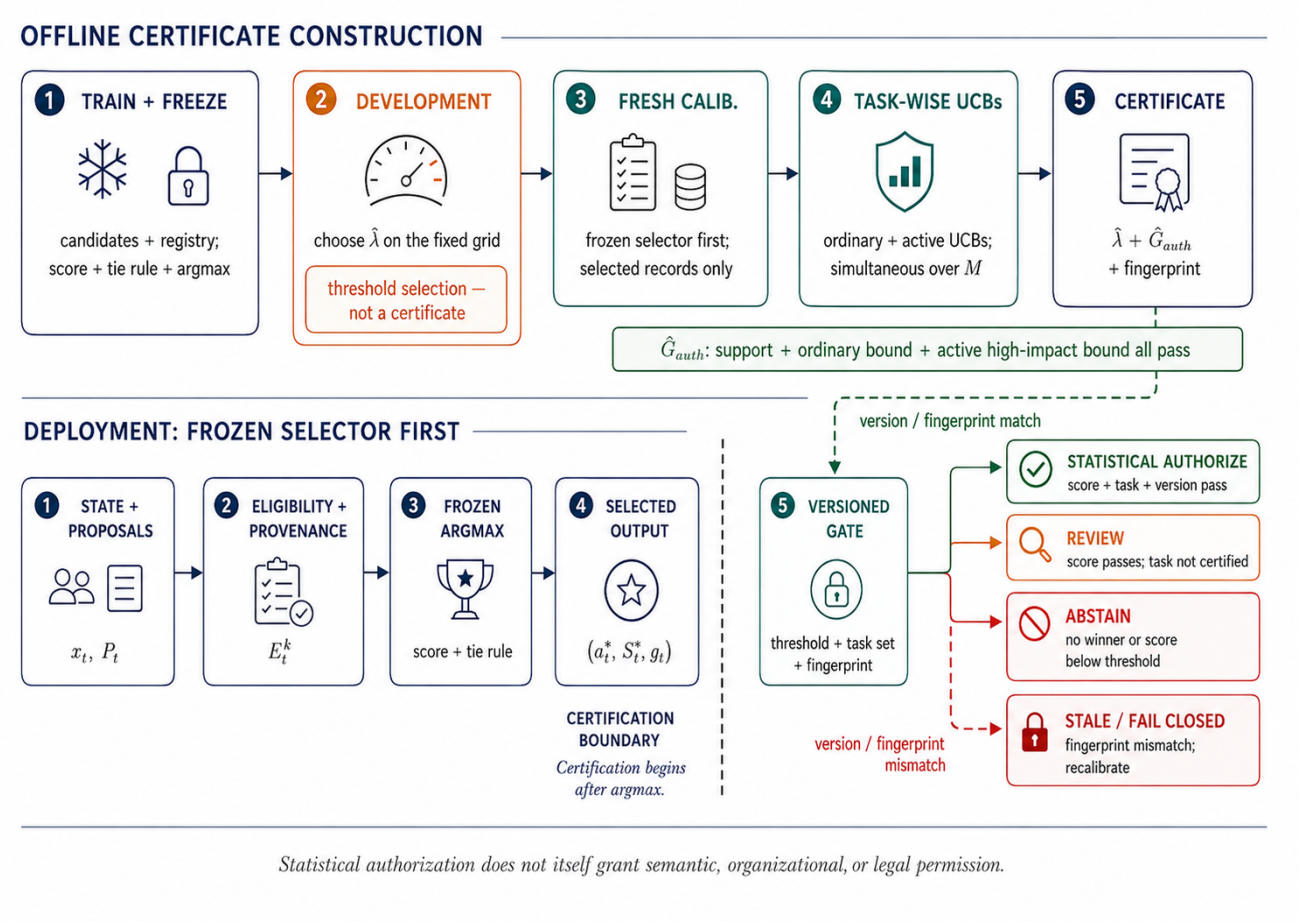

Figure 2: EPV-Opt certifies only after the frozen selector has produced its winner. A development-fixed threshold and fresh simultaneous task/severity bounds determine authorization, review, or abstention. A pipeline fingerprint makes later candidate, registry, score, threshold, or task-family changes fail closed until recalibration. The statistical certificate does not itself grant semantic, organizational, or legal permission.

Decision and versioning. At deployment, a score-clearing example in $\widehat{\mathcal{G}}_{auth}$ is authorized; a score-clearing uncertified task is reviewed; all others abstain. The certificate stores a hash of the protocol, candidate count, role registry, fixed threshold, task/severity family, tie rule, and canonical learned-model state. A mismatch raises a stale-certificate error rather than silently reusing the bound. The earlier full-family variant, which searches thresholds and coarse routes on calibration with a 418-event correction, remains a baseline and ablation.

Consequence evidence. Resettable environments can execute every candidate from the same state. Logged environments require registered overlap-aware OPE; historical logs may offer only temporally valid analogue evidence. Temporal precedence prevents future leakage but does not establish causal identification. Irreversible actions without a sandbox, support, or trusted outcome model remain review/abstain cases.

Theoretical Analysis

Theorem 1 (Independent heterogeneous selected-action risk)**.** *Let candidates be independent with eligibility probability π_k, conditional harm probability α_k, and continuous score CDF $F_{k,h}$ given eligibility and harm h. 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,

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

The numerator sums the probability that candidate k is harmful, eligible, clears λ, and beats every competitor; the denominator conditions on authorization. The IID identity $K\alpha \int F^{K-1} dF_1$ is a special case.

Proposition 1 (Unthresholded candidate-addition replacement)**.** *Let W indicate that an added candidate replaces the old winner, and let R^{win} denote the harm probability when*

a winner is always executed. Then

$$R^{win}_{new} - R^{win}_{old} = \Pr(W)\{\Pr(H_{new} = 1 \mid W) - \Pr(H_{old} = 1 \mid W)\}.$$

Thus increasing K raises uncensored winner risk only when the added winner is more adverse than the action it displaces on replacement states.

For nested candidate counts $K_1 < \cdots < K_m$, applying Proposition 1 at each transition gives the exact telescoping identity

$$R_{K_m} - R_{K_1} = \sum_{j=2}^{m} \Pr(W_j)\,\Delta_j,$$

where

$$\Delta_j = \Pr(H_{K_j} = 1 \mid W_j) - \Pr(H_{K_{j-1}} = 1 \mid W_j).$$

Unlike Theorem 1, this identity requires neither candidate independence nor continuous scores and directly matches our correlated nested planners.

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

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

and, for $C_1 > 0$,

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

The four joint-risk terms are retained-authorized replacement, retained stable action, authorization entry, and authorization exit. The final term shows why joint false authorization alone cannot explain conditional FAR when candidate expansion changes coverage. The identity is distribution-free and is verified row-wise in the released audit code.

Proposition 3 (Pairwise ranking is non-identifying)**.** *For $K > 2$, the scalar $q = \Pr(S_{harm} > S_{safe})$ does not generally identify top-one selected risk. Two score distributions can share $q = 1/2$ and still produce different R_{sel}; Figure 1A gives an explicit construction. Under the additional Gumbel random-utility family, $q = \mathrm{logistic}(\beta)$ and the selected risk has the familiar finite-β softmax form, but that map is model-specific.*

Theorem 2 (Finite-sample split-local EPV-Opt control)**.** *Assume development, calibration, and deployment episodes are independent IID draws from one registered distribution. Freeze the complete selector, choose one threshold on development, and register tasks $\mathcal{G}$ and structurally high-impact tasks $\mathcal{G}_{hi}$ before calibration. Protect ordinary task bounds and active high-impact bounds with total error δ over $M = |\mathcal{G}| + |\mathcal{G}_{hi}|$ events. With probability at least $1 - \delta$, every certified task satisfies its conditional-risk budgets in deployment. Any mixture over certified tasks therefore satisfies the largest active ordinary and high-impact budgets; unsupported tasks route to review.*

The proof is simultaneous Clopper–Pearson coverage plus mixture closure; full proofs and the more conservative finite threshold-route variant are in the supplement.

Corollary 1 (Zero-harm support)**.** *For risk budget α and M simultaneous events, a zero-harm bound requires*

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

For split-local $M = 25$, this is 59 examples at $\alpha = 0.10$ and 122 at $\alpha = 0.05$; the earlier $M = 418$ family needs 86 and 177. Local resolution and calibration search therefore have an explicit support cost.

The guarantees are conditional on authorization under the frozen IID deployment distribution. They are not pointwise coverage, do not cover drift, and expire after an unregistered pipeline change.

Sealed Evaluation

Environment and candidates. MiniWoB (Shi et al. 2017; Liu et al. 2018) permits every candidate plan to execute from the same reset state. Four deterministic planner families produce exactly eight unique plans per episode: a task-aware heuristic, single DOM actions, ordered two-step DOM plans, and plan mutations. Nested prefixes give $K \in \{1, 2, 4, 8\}$. These are controlled multi-planner candidates, not independent LLM agents.

Splits and protocol integrity. The v2 pilot provided 1,000 train and 500 development episodes for the frozen TF–IDF logistic scorer and threshold rule. After preserving v2 and a preregistered v3 zero-coverage certificate failure, v4 froze a corrected 25-event task/severity family before collecting another 4,000 disjoint-seed episodes: 2,000 fresh calibration and 2,000 one-time test. Every candidate executes from reset, producing 32,000 new outcomes. Protocol, generator, model, outcomes, and decisions have SHA-256 hashes; v2/v3/v4 seeds do not overlap. No live API or outcome-driven task removal is used.

Registered adverse-action proxy. Executing a nonempty state-changing plan without task reward is the binary adverse-action proxy. Utility is reward net of action and latency cost. Pre-registered task groups are workflow, account, and external side effect; the last is a high-impact task-semantic proxy. These labels are not observed privacy, financial, or irreversible loss.

Methods and metrics. We compare the uncensored learned selector; sourcewise CP followed by argmax; SCoRE-SDR (Bai and Jin 2026); end-to-end CP; the 418-event full-family EPV-Opt baseline; and split-local EPV-Opt. Their guarantees differ: sourcewise CP does not certify a new post-argmax distribution, SCoRE controls expected selected-discovery risk, end-to-end CP controls global conditional FAR, and split-local EPV-Opt simultaneously constrains each authorized task and active severity label. We report direct share, review, conditional and joint FAR, exact test diagnostics, utility, AV, and local support.

Second-environment transfer sequence. AgentDojo (Debenedetti et al. 2024) supplies stateful banking, workspace, travel, and Slack tools with registered prompt-injection goals. The registered sequence separates four objects: v7 static multi-model proposal replay, v8 dynamic interaction and provider-failure stress, v9 calibration of a pooled authority-and-usefulness certificate, and v10 one-time holdout transfer. V9 and v10 freeze the same six candidates (Sonnet, Opus, and Fable under direct and cautious policies), task mixture, attacks, selector, and losses. Each cluster runs a direct validation injection and a distinct deployment injection from reset state; only validation evidence can select the deployment candidate. The study tests pooled sandbox transfer, not AgentDojo leadership, task-local certification, production prevalence, or legal permission.

Results

RQ1: Does adding candidates necessarily amplify risk? No. On 2,000 fresh test episodes, the uncensored selector's adverse-proxy rate falls from 0.896 at $K = 1$ to 0.684 at $K = 4/8$. The exact correlated replacement terms are -0.082, -0.130, and 0.000 for K1–2, K2–4, and K4–8. Added winners are less adverse than displaced winners at the first two transitions and equally adverse at the last. This empirically instantiates Proposition 1; the identity itself is algebraic. In a post-outcome robustness audit, all 64 outcome-independent candidate permutations reduce uncensored risk from K1 to K4, but threshold-conditioned FAR changes can have either sign (for K1–2, range $[-0.511, 0.126]$), as Proposition 2 predicts. Candidate presentation order is therefore part of the versioned selector; changing it invalidates the certificate.

RQ2: What do post-selection methods certify? At $K = 8$, full-family EPV-Opt authorizes 197 actions (coverage 0.099), reviews 363, and observes FAR 0.025. Split-local EPV-Opt authorizes 388 (0.194), reviews 172, observes 0/388 adverse authorizations (test UCB 0.008), and raises utility from 180.66 to 367.85 and AV from 621.77 to 808.95. A post-outcome factor audit shows that a development-fixed threshold with the original $M = 418$ correction still authorizes 199 actions; the gain is therefore driven by reducing the inspected event family, not threshold timing alone. Under matched task-local semantics, task-specific thresholds add only eight authorizations over one pooled development threshold, and Holm adds none over Bonferroni. End-to-end CP still has higher coverage (0.280) and AV (915.80) at FAR 0.037; SCoRE has coverage 0.318 and AV 958.15 under expected-SDR semantics. Split-local EPV-Opt improves certificate resolution and local semantics, not every safety–utility frontier.

RQ3: Which local actions earn direct authority? At $K = 8$, four tasks pass simultaneous local bounds: `click-checkboxes`, `click-option`, `enter-password`, and `login-user`. Their calibration supports are 100, 99, 100, and 93 with zero adverse outcomes; ordinary-risk UCBs are at most 0.0646. Three score-clearing tasks fail or lack support and route to review. In particular, `login-user-popup` has 4 adverse outcomes among 7 calibration actions and receives no direct authority; its nine score-clearing test actions are reviewed. External-side-effect tasks have no score-clearing support and abstain, so high-impact execution is not established.

Table 1: Sealed $K = 8$ results. FAR denotes the registered adverse-action proxy (a state-changing plan without task reward) among authorized actions, not observed production harm; UCB is a one-sided 95% test diagnostic. AV is relative to uncensored selector utility after common overhead. Guarantee semantics differ by method.

Method	Direct	Review	FAR [UCB]	Utility	AV
Selector	1.000	0.000	0.684 [0.701]	-449.10	-8.00
Sourcewise CP	0.089	0.000	0.000 [0.017]	169.43	610.53
SCoRE-SDR	0.318	0.000	0.077 [0.097]	517.05	958.15
End-to-end CP	0.280	0.000	0.037 [0.054]	474.69	915.80
EPV-Opt-Full	0.099	0.181	0.025 [0.053]	180.66	621.77
EPV-Opt-Split	0.194	0.086	0.000 [0.008]	367.85	808.95

RQ4: What did the failed certificate teach? The preregistered v3 design applied a 0.05 high-impact bound to all 20 tasks and authorized none: with 40 simultaneous events, zero high-impact harms among 100 examples still gives UCB 0.0647. V4 preserves that failure, then tests high-impact risk only for the five tasks where the registered label can occur. This structural registration reduces the family to 25 events and restores positive task-level authority without weakening ordinary task risk.

RQ5: Does authority survive a pipeline change? No. Changing candidate count, the role registry, or learned score weights changes the canonical pipeline fingerprint, and all three audit mutations reject the stale certificate. Versioning is a by-construction invariant, not a performance comparison.

External dynamic transfer. The registered AgentDojo sequence separates proposal replay, dynamic provider stress, calibration, and an independent holdout. V7 executes 75,165 static candidate–context pairs from Opus, Sonnet, and Fable but every certifier is infeasible. V8 dynamic calibration is also infeasible and its test is dominated by provider-interface failures, which are retained as operational evidence rather than model performance. V9 then obtains a pooled dual certificate on 84 selected clusters: 0 authority violations (simultaneous UCB 0.043) and 66 useful completions (LCB 0.683). Its registered test remains unopened because the fixed cost cap could not cover the full split. V10 imports the exact pipeline on a fresh one-time holdout. It authorizes 19/90 clusters with 0 authority violations and 16 useful completions, but does not pass the preregistered confirmation gate. The joint gate requires all three conditions: support is below 40 and risk UCB 0.146 exceeds 0.10; only utility LCB 0.641 clears its 0.50 floor. The result is a preserved negative transfer finding, not a new certificate or a high-impact claim. Supplement S7–S9 gives the complete audits.

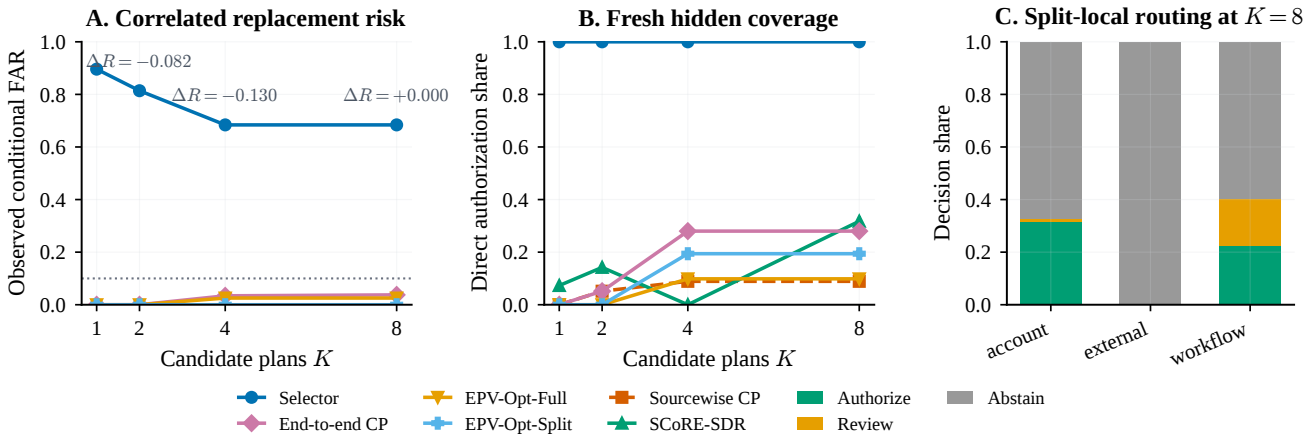

Figure 3: Sealed results separate candidate replacement, coverage, and local routing. Panel A annotates the exact dependence-free unthresholded replacement decomposition; added winners have lower adverse-proxy rates than displaced winners. Panel B compares direct authorization under distinct guarantee semantics. Panel C shows split-local EPV-Opt at $K = 8$: four tasks authorize, three score-clearing tasks review, and unsupported actions abstain. The legend is outside all axes; the dotted line is the 0.10 FAR budget.

Related Work

Runtime authority. Tool and web agents expose side effects beyond prediction error (Yao et al. 2023; Qin et al. 2024; Zhou et al. 2024; Jimenez et al. 2024; Yao et al. 2025; Debenedetti et al. 2024). Runtime guards, AgentTrust, and proof-carrying actions govern source/target permissions and audit evidence (Qin et al. 2026; Yang 2026; Wang 2026). EPV-Opt can consume such checks as eligibility; it addresses the complementary statistical risk of the action selected after heterogeneous competition.

Selective risk and action evaluation. Selective prediction and conformal risk control govern registered losses under sampling assumptions (Chow 1970; El-Yaniv and Wiener 2010; Gibbs and Candès 2021; Bates et al. 2021; Angelopoulos et al. 2024). SCRC and SCoRE treat post-selection risk (Xu, Guo, and Wei 2026; Bai and Jin 2026); SCoRE is a strong baseline here. Conformal Selective Acting (CSA) gives anytime-pathwise selective-risk control for adaptive RLVR streams (Khosravi and Huo 2026). Our setting is complementary and narrower: sealed IID calibration, a frozen heterogeneous selector, task-local routes, and versioned provenance. EPV-Opt is not a new generic confidence bound; it packages those Agent-specific objects around the final selected action. OPE and safe RL supply candidate-value evidence but do not by themselves authorize a heterogeneous test-time winner (Dudík, Langford, and Li 2011; Jiang and Li 2016; García and Fernández 2015).

Limitations and Impact

The independent heterogeneous identity assumes continuous scores, while both replacement decompositions allow arbitrary dependence. The finite-sample certificate assumes a frozen pipeline and IID calibration/deployment distribution and does not cover drift. Split-local routing reduces multiplicity from 418 to 25 registered events by spending an independent development split and structural severity registration; it does not eliminate conservative exact bounds. High-impact bounds apply only to pre-registered task-semantic categories; tasks outside that family are structurally excluded rather than certified safe. MiniWoB remains the only environment with a positive heldout statistical certificate. AgentDojo adds dynamic interaction from three model families in a recognized tool sandbox, but only for one frozen 20-task pooled mixture and two registered attack templates. Its calibration certificate does not confirm on the independent holdout because selected support contracts from 84 to 19; zero observed violations therefore still give a 0.146 UCB. Both environments use registered consequence proxies rather than observed production harm or legal permission. Split-local EPV-Opt has lower direct coverage, utility, and AV than end-to-end CP and SCoRE; matched-semantics task thresholds add only eight v5 test authorizations over a pooled threshold, and Holm adds none over Bonferroni. Task-local AgentDojo certification, independent high-impact labels, broader natural model populations, and online validity remain open.

Conservative routing can also distribute automation unevenly. Deployments should expose group support, review load, missed opportunity, and certificate expiry rather than presenting a pooled bound as permanent permission.

Conclusion

Candidate count alone does not determine selected-action risk; the score–harm relationship and complete selector do. EPV-Opt makes statistical execution authority explicit by certifying the versioned selected output and routing unsupported local actions to review or abstention. Prediction is not permission: execution authority must be established after selection.

References

Angelopoulos, A. N.; Bates, S.; Fisch, A.; Lei, L.; and Schuster, T. 2024. Conformal Risk Control. In *International Conference on Learning Representations*.

Bai, T.; and Jin, Y. 2026. Conformal Selective Prediction with General Risk Control. *arXiv preprint arXiv:2603.24704*.

Bates, S.; Angelopoulos, A. N.; Lei, L.; Malik, J.; and Jordan, M. I. 2021. Distribution-Free, Risk-Controlling Prediction Sets. *Journal of the ACM*, 68(6): 1–34.

Chow, C. K. 1970. On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1): 41–46.

Debenedetti, E.; Zhang, J.; Balunović, M.; Beurer-Kellner, L.; Fischer, M.; and Tramèr, F. 2024. AgentDojo: A Dynamic Environment to Evaluate Prompt Injection Attacks and Defenses for LLM Agents. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*.

Dudík, M.; Langford, J.; and Li, L. 2011. Doubly Robust Policy Evaluation and Learning. In *International Conference on Machine Learning*.

El-Yaniv, R.; and Wiener, Y. 2010. On the Foundations of Noise-Free Selective Classification. *Journal of Machine Learning Research*, 11: 1605–1641.

García, J.; and Fernández, F. 2015. A Comprehensive Survey on Safe Reinforcement Learning. *Journal of Machine Learning Research*, 16: 1437–1480.

Gibbs, I.; and Candès, E. J. 2021. Adaptive Conformal Inference Under Distribution Shift. In *Advances in Neural Information Processing Systems*.

Guo, C.; Pleiss, G.; Sun, Y.; and Weinberger, K. Q. 2017. On Calibration of Modern Neural Networks. In *International Conference on Machine Learning*.

Jiang, N.; and Li, L. 2016. Doubly Robust Off-Policy Value Evaluation for Reinforcement Learning. In *International Conference on Machine Learning*.

Jimenez, C. E.; Yang, J.; Wettig, A.; Yao, S.; Pei, K.; Press, O.; and Narasimhan, K. 2024. SWE-bench: Can Language Models Resolve Real-World GitHub Issues? In *International Conference on Learning Representations*.

Khosravi, H.; and Huo, X. 2026. Conformal Selective Acting: Anytime-Valid Risk Control for RLVR-Trained LLMs. *arXiv preprint arXiv:2605.20270*.

Liu, E. Z.; Guu, K.; Pasupat, P.; Shi, T.; and Liang, P. 2018. Reinforcement Learning on Web Interfaces using Workflow-Guided Exploration. In *International Conference on Learning Representations*.

Qin, S.; Zhuang, H.; Zhou, Y.; Han, Y.; and Zhang, X. 2026. AIRGuard: Guarding Agent Actions with Runtime Authority Control. *arXiv preprint arXiv:2605.28914*.

Qin, Y.; Liang, S.; Ye, Y.; Zhu, K.; Yan, L.; Lu, Y.; Lin, Y.; Cong, X.; Tang, X.; Qian, B.; et al. 2024. ToolLLM: Facilitating Large Language Models to Master 16000+ Real-world APIs. In *International Conference on Learning Representations*.

Shi, T.; Karpathy, A.; Fan, L.; Hernandez, J.; and Liang, P. 2017. World of Bits: An Open-Domain Platform for Web-Based Agents. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, 3135–3144.

Wang, Z. 2026. Proof-Carrying Agent Actions: Model-Agnostic Runtime Governance for Heterogeneous Agent Systems. *arXiv preprint arXiv:2606.04104*.

Xu, Y.; Guo, W.; and Wei, Z. 2026. Selective Conformal Risk Control. *arXiv preprint arXiv:2512.12844*.

Yang, C. 2026. AgentTrust: A Self-Improving Trust Layer for AI-Agent Actions. *arXiv preprint arXiv:2606.08539*.

Yao, S.; Shinn, N.; Razavi, P.; and Narasimhan, K. 2025. τ-bench: A Benchmark for Tool-Agent-User Interaction in Real-World Domains. In *International Conference on Learning Representations*.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *International Conference on Learning Representations*.

Zhou, S.; Xu, F. F.; Zhu, H.; Zhou, X.; Lo, R.; Sridhar, A.; Cheng, X.; Ou, T.; Bisk, Y.; Fried, D.; Alon, U.; and Neubig, G. 2024. WebArena: A Realistic Web Environment for Building Autonomous Agents. In *International Conference on Learning Representations*.

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.