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

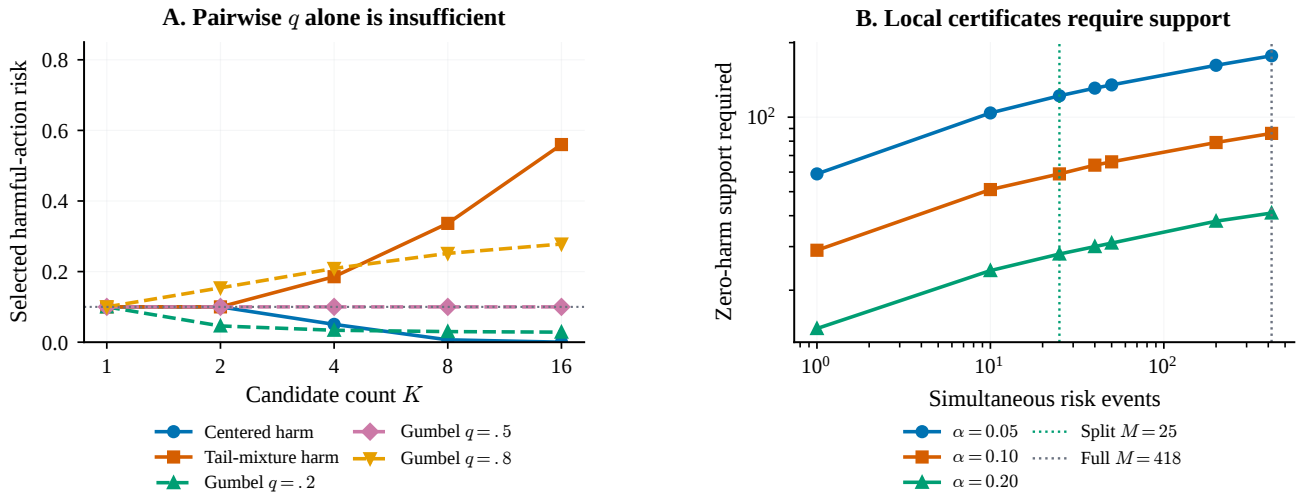

Figure 1: Candidate count alone does not determine selected-action risk. Panel A gives two score constructions with identical pairwise $q = 0.5$ but sharply different top-one risk; Gumbel curves are a registered parametric special case. Panel B shows the zero-harm support cost of simultaneous certificates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

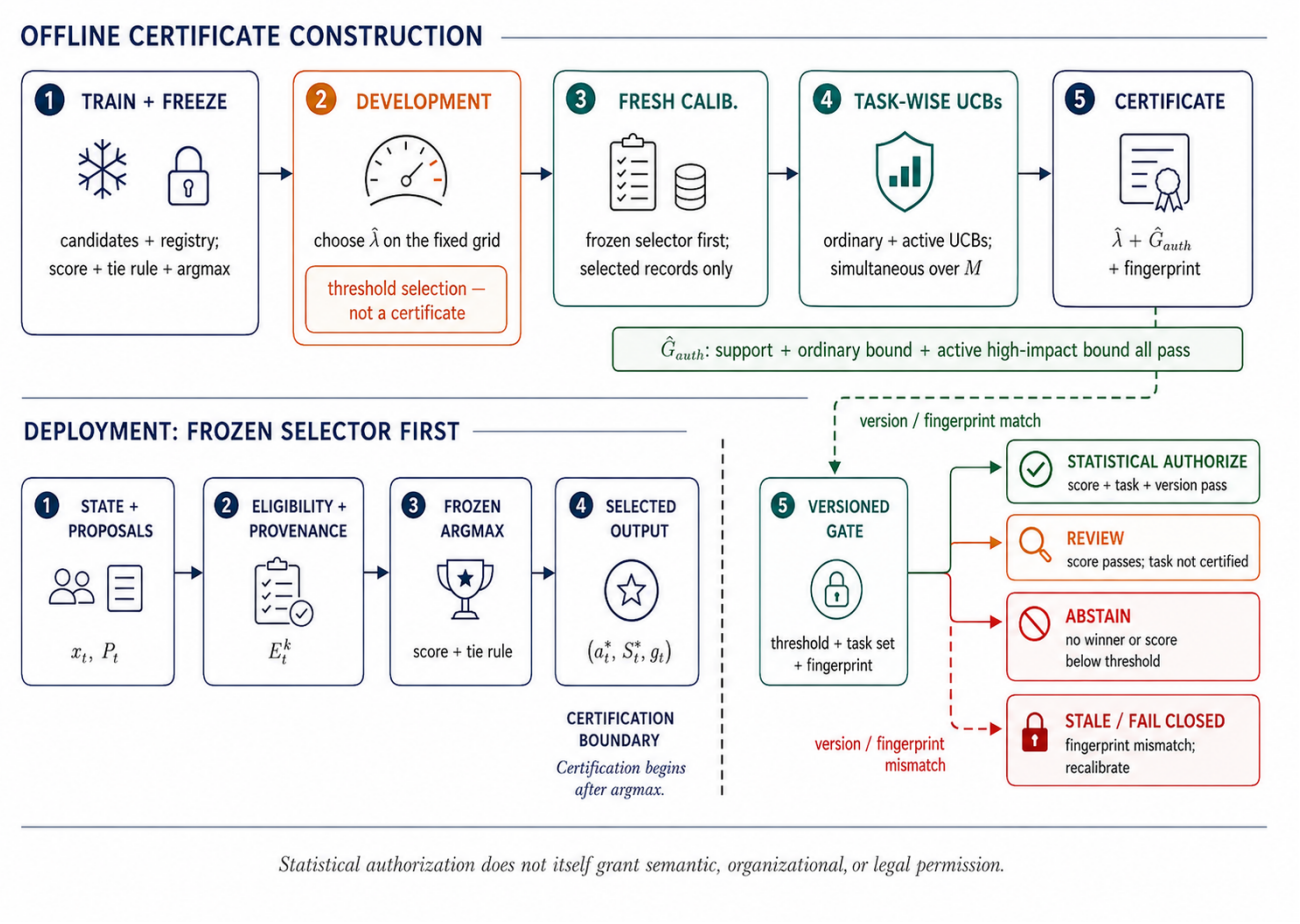

Figure 2: EPV-Opt certifies only after the frozen selector has produced its winner. A development-fixed threshold and fresh simultaneous task/severity bounds determine authorization, review, or abstention. A pipeline fingerprint makes later candidate, registry, score, threshold, or task-family changes 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.

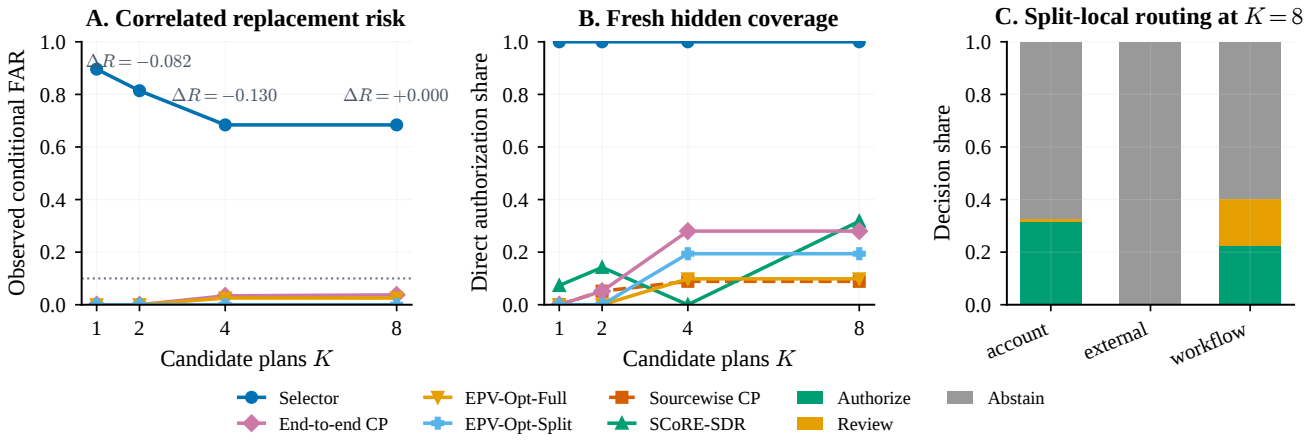

Figure 3: Sealed results separate candidate replacement, coverage, and local routing. Panel A annotates the exact dependence-free unthresholded replacement decomposition; added winners have lower adverse-proxy rates than displaced winners. Panel B compares direct authorization under distinct guarantee 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*.