AGENTVAX: FROM DOWNSTREAM HARM TO LOCAL SAFETY IN MULTI-AGENT WORKFLOWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Safety in language-model multi-agent workflows is fragmented across both evidence and control, while harmful incidents may only be confirmed after earlier decisions. This creates two coupled problems: learning future risk from what each agent actually observed, and choosing an authorized action that can still change the outcome. We find that local views provide strongly complementary safety signals: at matched inference budgets and 5% validation false-positive rate, combining views raises attack-detection TPR from 55.9% for the best single view to 73.6%, a 17.7-point improvement. We also find that the position best able to recognize risk need not be the one best able to prevent harm. Motivated by these observations, we propose **AgentVax**, which turns verified downstream incidents into supervision for earlier local decisions. AgentVax pairs events with versioned historical inputs, uses discrete-time survival analysis to predict whether and when a violating proposal will occur, and applies counterfactual replay from restored workflow states to estimate how authorized actions change harm, legitimate-task utility, and cost. These estimates train local policies under a workflow-level false-blocking constraint, while deployment uses only each agent's current authorized view. Across controlled workflow studies spanning 24 attack families, we observe directed transfer across attacks sharing safety-relevant patterns. In a separate controlled action-value study, replay-based action selection reduces decision regret from 0.150 to 0.043, a 71.3% reduction relative to risk ranking. Together, these results show how delayed downstream feedback can become earlier, locally actionable protection by learning risk from what an agent can observe and intervention from what it can still change.

1 INTRODUCTION

Language-model multi-agent workflows divide tasks across agents that read information, transform intermediate objects, plan actions, and execute tools. This expands system capability but also distributes safety-relevant evidence: a reader may know an instruction's provenance, a planner its relation to the authorized goal, and an executor the final parameters and permissions. Harmful requests or instructions embedded in untrusted content may therefore become verifiable only at a later step (Andriushchenko et al., 2025; Debenedetti et al., 2024). Once an incident is confirmed, the key question is how to use it to improve earlier decisions on future tasks without giving agents information they could not have observed.

Prior work has advanced partial-trajectory risk forecasting, adaptive guardrails, and communication-graph interventions (Xiong et al., 2026; Luo et al., 2025; Wang et al., 2025a). However, existing studies usually separately treated two problems that need to be linked for continual protection. 1) Downstream event must be grounded in each agent's historical local evidence: one global outcome does not imply that all agents had the same information. 2) Detecting risk must be separated from choosing an intervention: a high risk score does not determine whether continue, review, or block will improve the downstream outcome. We therefore ask how a verified incident can teach earlier agents both *what to predict from what they observed* and *what to do with the control they still have*.

Our experiments show that both distinctions matter. With model weights and total inference budget fixed, complementary views raise attack-detection TPR substantially more than repeated sampling of one view (Section 4). View exchanges affect detection substantially more than role-name exchanges, linking the gain to information rather than nominal roles. The experience also transfers structurally:

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

updating on attacks sharing a safety-relevant pattern but differing in goal improves held-out detection, while lexical similarity alone gives no comparable gain. At the same time, the position best able to recognize risk is not always the one with the greatest intervention benefit because permissions, remaining control, and task costs differ across stages. These observations motivate a simple principle: *learn prediction from what an agent can observe, and intervention from what it can still change*.

We propose **AgentVax** around this principle. When an incident is verified, AgentVax links the event and its occurrence time to versioned snapshots of the local inputs that participating agents actually observed. A discrete-time survival model then estimates finite-horizon probabilities of a future violating proposal. Later events define training targets but never enter earlier runtime inputs, so agents can learn from shared incidents while predictions remain grounded in their own authorized information.

AgentVax then connects prediction to action through counterfactual replay. It restores a historical workflow state and executes the alternatives authorized at that point—continue, review, or block—under a frozen joint policy. Regenerated continuations estimate how each action changes downstream harm, legitimate-task utility, and operational cost, and these action values supervise local policy updates under a workflow-level false-blocking constraint. In the controlled action-value study, replay reduces decision regret from 0.150 to 0.043, showing that explicitly estimating intervention effects produces better decisions than ranking actions by predicted risk alone.

We evaluate the design in controlled workflow studies spanning 24 attack families and eight platform strata, including chronological feedback, held-out-family evaluation, evidence and permission interventions, local adaptation, workflow replay, representation analysis, and lifecycle accounting. Across these studies, complementary evidence supports earlier warning, shared safety patterns enable directed transfer, and local adaptation preserves role-relevant information. For shared-pattern unseen attacks, role-matched adaptation raises TPR from 65.4% with pooled updates to 72.8%, while benign-task success increases from 88.7% to 91.9%. Together, the results connect distributed local evidence to a learning mechanism that turns shared downstream experience into locally actionable protection.

Contributions. Through controlled behavioral, transfer, functional, and representation analyses, we show that safety evidence is systematically distributed across local views, quantify the value of complementary evidence, identify directed transfer through shared safety patterns, and separate risk recognition from intervention ability.

We propose **AgentVax**, which combines event-to-prefix supervision, discrete-time survival prediction, executed counterfactual replay, and workflow-level policy learning to turn shared downstream incidents into local protection under each agent's authorized information and actions.

Through extensive experiments across 24 attack families, eight platform strata, chronological updates, held-out families, evidence and permission interventions, and lifecycle evaluation, we validate that complementary local evidence enables earlier warning, shared patterns support cross-attack reuse, and replay-based action values translate predictive signals into more effective local decisions.

2 Related Work

Agent and multi-agent safety. JANUS anticipates delayed risks from partial trajectories, while G-Safeguard combines anomaly detection on communication graphs with topological interventions (Xiong et al., 2026; Wang et al., 2025a). AgentVax studies workflows where local information varies across agents and time, connecting complementary warnings to each agent's ability to intervene.

Experience-driven safety learning. AGrail adaptively generates and optimizes safety checks, while AgentGuard learns conditional execution constraints from anomalous coding-agent trajectories (Luo et al., 2025; Dai & Wang, 2026). AgentVax pairs each confirmed downstream harm with multiple agents' historical local inputs and uses survival analysis to learn future-harm probabilities and timing, translating shared experience into early judgments under each agent's information constraints.

Multi-agent credit assignment. COMA constructs counterfactual baselines using a centralized critic, while C3 estimates decision contributions by evaluating alternative actions and their continuations from fixed histories (Foerster et al., 2018; Chen et al., 2026). AgentVax applies counterfactual

evaluation to locally available safety actions, learning intervention policies against a joint objective over harm, task utility, and cost under a workflow-level false-blocking constraint.

Representations of safety behavior. Representation Engineering studies high-level behavior through representation reading and intervention (Zou et al., 2023), and Arditi et al. (2024) identify a single direction mediating refusal in each model studied. We apply these tools to safety judgments made by the same model under different local views, examining how source, authorization, and goal information supports complementary warnings and participates in safety decisions.

3 Problem Formulation

We consider an organization executing a multi-agent workflow as a sequence of states and decisions. At event t, agent $i(t)$ receives a local observation $x_t = O_{i(t)}(s_t)$ determined by its messages, memory, object versions, tools, and permissions. Two agents may read the same object at different times and therefore observe different states; conversely, agents with identical available information may share a predictor. The complete state s_t is retained only for offline restoration and evaluation. During deployment, a local safety policy chooses an authorized action u_t from $\mathcal{U}_{i(t)}(x_t) \subseteq \{\text{continue}, \text{review}, \text{block}\}$.

An attack may have a harmful objective in the initial query, and earlier operations may prepare later harm. Let T_H denote the first verified prohibited proposal under the organization's authorization policy, and let $T_D \geq T_H$ denote the time at which an audit confirms the event and releases its label for learning. Proposed violations and realized harmful state changes are recorded separately. The learning problem is to use feedback available after T_D to improve future decisions from observations analogous to x_t for $t < T_H$, while respecting the benign workflow false-blocking constraint and accounting for task loss and operational cost. Attack families are evaluation groupings over multiple independent trajectories; they are not runtime modules or a required ontology for AgentVax.

4 Empirical Observations: From Local Evidence to Transferable Protection

To investigate how safety differences arise across agents processing shared information, and whether they support collaborative protection, we compare stage-specific evidence, exchange local views, and predict risk from trajectory prefixes. These comparisons reveal complementary judgments and early-warning opportunities. Cross-attack updates and chronological expansion then identify directional, bounded transfer through shared attack patterns. Local adaptation and workflow replay measure protection matched to each node's observations and controls. Finally, functional interventions and representation analyses localize safety differences and their relationship to decisions. These observations motivate sharing verified experience while preserving local information and control opportunities.

We divide trajectories into Read, Transform, Plan, and Act. A local observation contains the currently visible prefix and context; the agent runtime, or harness, specifies tools, permissions, and execution constraints. Authorization and execution logs distinguish policy-violating proposals from realized violations. Attacks are paired with legitimate high-risk tasks and split by family and task template. Detection thresholds are fixed at 5% validation false-positive rate (FPR; Appendix A.1).

4.1 Complementary Local Evidence Enables Early Warning

Local evidence varies across stages. We annotate shared attack patterns (*stage-level hazard motifs*) by action, provenance, and authorization, and examine their visibility. Source confusion is clearest during reading, goal deviation during planning, and unauthorized access during execution (Figure 1a). Deleting and restoring provenance in summaries changes downstream detection, motivating attention to instruction–data separation (Zverev et al., 2025); information-revelation and transformation controls appear in Appendix A.2.

To test whether these differences yield complementary judgments, we fix model weights and total inference budget. Multi-view detection reaches 73.6% true-positive rate (TPR), versus 55.9% for the

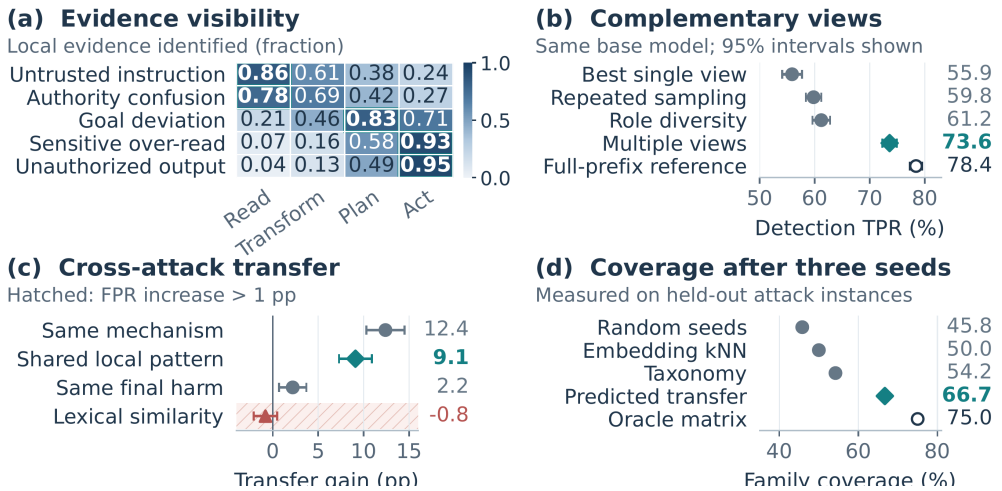

Figure 1: **Local evidence and cross-attack transfer.** (a) Identifiable fraction; outlines mark row maxima. (b) Same-model TPR (%) at 5% validation FPR, with matched inference budgets; full-prefix uses current information only. (c) Transfer gain (percentage points); hatching marks FPR increases above one point. (d) Held-out family coverage (%) after three seeds. Bars in (b,c) are 95% bootstrap intervals; other panels have no uncertainty estimate. Hollow markers denote references.

best single view and 59.8% for repeated sampling (Figure 1b). View swaps move detection profiles toward their reference by 68%, versus 19% for role swaps; identical full-prefix inputs reduce unique detections by 46%. Appendix A.2 gives single-view, budget, exchange, and error-overlap analyses and defines the profile-movement statistic.

We next forecast harm within two subsequent events using only current prefixes. Evidence aggregation achieves 0.88 AUROC versus 0.82 for a budget-matched single view, with expected calibration error (ECE) of 0.04. Unauthorized parameters revealed only late remain difficult to anticipate. Horizon, calibration, and computation controls are in Appendix A.2. This stage-dependent warning value motivates testing reuse across attacks.

4.2 Shared Attack Patterns Support Directed Transfer

To measure reuse, we update a local detector only on source family a and evaluate held-out family b under matched update budgets. The stage-s gain is $T^{(s)}_{a\to b} = \mathrm{TPR}(f_s^{+a}, D_b) - \mathrm{TPR}(f_s, D_b)$. Positive transfer requires a positive interval lower bound and at most one percentage point of additional benign FPR. Shared-pattern, different-goal pairs gain 9.1 percentage points, versus -0.8 for lexical similarity alone (Figure 1c). Pattern overlap predicts positive transfer with 0.78 AUROC, versus 0.60 for text similarity. Appendix A.3 details all four pairing types, directionality, and negative transfer.

We organize these gains into a directed analysis graph: 61% of pairs with positive transfer in either direction are asymmetric. Using historical evidence, three transfer-guided seeds cover 66.7% of families, versus 50.0% for semantic neighbors (Figure 1d). Coverage is evaluated on target tasks, not inferred from graph reachability. Chronological selection and disconnected components are detailed in Appendix A.3. These bounded relationships motivate adapting shared experience locally.

4.3 Local Adaptation Improves Workflow-Level Protection

We compare seven update strategies using identical verified cases, frozen base models, and local heads or context/runtime rules. Role-matched adaptation reaches 72.8% TPR on shared-pattern unseen attacks, versus 65.4% for pooled updates; benign task success is 91.9% versus 88.7% (Figure 2a). On disconnected families, its gain over no update is 1.7 percentage points, with an interval spanning zero. These graph-conditioned comparisons are reported separately from the end-to-end AgentVax evaluation. Appendix A.4 reports all strategies and generalization sets.

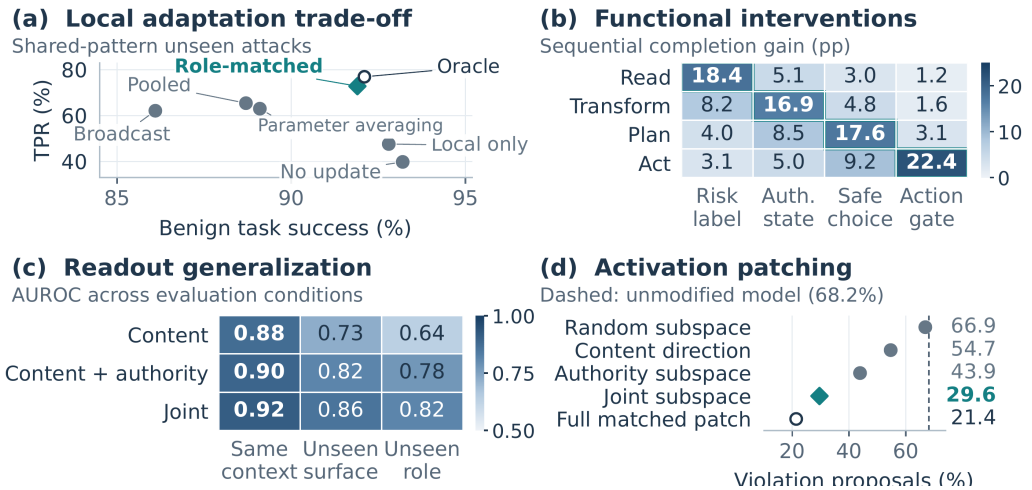

Figure 2: **Adaptation and mechanisms.** (a) Unseen shared-pattern TPR versus benign success (%); role-matched is a graph-conditioned comparator; oracle uses future labels. (b) Conditional safe-completion increments (percentage points). (c) Readout AUROC; joint: content, authority, goals, capabilities. (d) Violating proposals (%), not realized harm; lower is better; dashed: unmodified model. No uncertainty estimate is available.

Chronological workflow replay randomizes the first discovery location to test whether local gains reach execution. Role-matched updates reduce prohibited-action reachability relative to a central guard while retaining 91.7% benign task success (Appendix A.4). The resulting workflow benefit raises the question of which local functions require adaptation.

4.4 Functional and Representation-Level Analyses Explain Safety Differences

We sequentially supply correct risk labels, provenance/authority, safe actions, and deterministic enforcement. Transform, Plan, and Act benefit most from source attribution, safety decisions, and action gates, respectively (Figure 2b). These are conditional increments, not independent modules. Appendix A.5 reports the complete stage-wise intervention and representation diagnostics, with proposals kept separate from realized harm.

In the representation study, linear readouts compare matched attack–benign representations at fixed decision positions. A joint subspace encoding content hazard, authority, goals, and available actions reaches 0.82 AUROC on unseen roles, versus 0.64 for a content-hazard direction (Figure 2c). Event-aligned differences follow provenance/authorization integration; Appendix A.5 reports the readout summaries. Patching matched benign components reduces violating proposals from 68.2% to 29.6%, versus 66.9% for an equal-rank random subspace (Figure 2d). Appendix A.5 reports intervention summaries and their interpretation limits. This indicates a local intervention effect, not complete natural mediation.

These observations motivate trustworthy experience sharing, learning from original local inputs, and protection through available controls. The following method develops these principles with verified downstream harm as delayed supervision.

5 AgentVax

AgentVax connects complementary local evidence to effective control through three stages: delayed downstream events supervise historical local views, restored-state replay estimates authorized action values, and workflow-level optimization coordinates local policies. Figure 3 summarizes the resulting continual-protection loop.

Training may use complete states and a trusted verifier, but deployment uses only authorized local inputs, risk vectors, and available actions. Task actors remain frozen. This centralized-training,

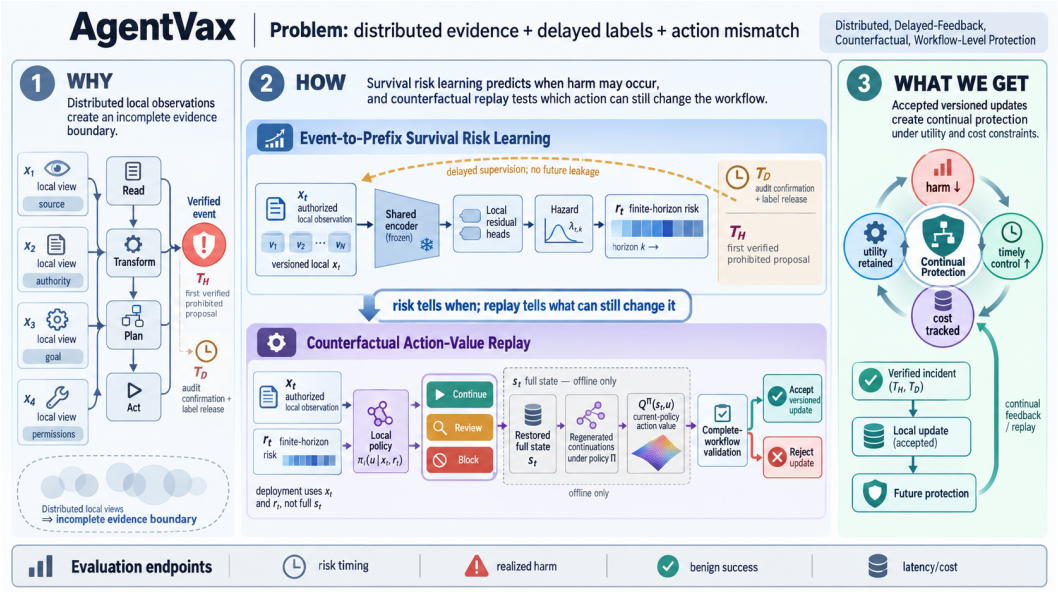

Figure 3: **AgentVax overview.** Distributed local views form an incomplete evidence boundary and produce a verified event at T_H, whose audit confirmation and label release arrive at T_D. Event-to-prefix survival learning uses versioned authorized observations x_t to produce a finite-horizon risk vector r_t; counterfactual replay combines (x_t, r_t) with authorized actions and offline-only restored state s_t to estimate $Q^{\Pi}(s_t, u)$ and validate complete-workflow updates. Accepted versioned updates close the loop toward continual protection while preserving benign utility and tracking cost.

decentralized-execution design (Foerster et al., 2018) preserves the information boundary while evaluating each update on complete workflows.

5.1 Local Risk Prediction from Delayed Feedback

At each decision we retain the actual local input $x_t = O_{i(t)}(s_t)$, including visible context, messages, memory, tool outputs, authorized objects, and their versions. The target is the first *verified violating proposal* at T_H; audit confirmation and label release occur at $T_D \geq T_H$, and realized harm has a separate timestamp. Only labels released by the update cutoff enter training, and strict early prediction uses $t < T_H$. A confirmed event supervises all participating agents' historical snapshots. Appendices B.1 and F.1 specify and test this boundary.

The target also requires a reference continuation. We fix policy μ with the AgentVax gates in logging-only mode, preserve base safeguards, and complete safety-stopped runs by isolated reference replay. For an eligible snapshot, the label is the first subsequent violating proposal under this continuation; administrative truncation is right-censored only when the stated support and conditional-independence conditions hold. Continuations that fail restoration are excluded from supervision and retained in the reported coverage accounting (Appendix B.2).

We represent likelihood and timing with discrete-time survival analysis (Gensheimer & Narasimhan, 2019). For $D_t = T_H - t$ and horizons $k = 1, \ldots, K$, define

$$\lambda_{t,k} = \Pr_{\mu}(D_t = k \mid D_t \geq k, x_t), \qquad r_t(k) = 1 - \prod_{j=1}^{k}(1 - \lambda_{t,j}). \tag{1}$$

The risk vector $r_t = (r_t(1), \ldots, r_t(K))$ is monotone in the horizon. With observed follow-up m_t capped at K and endpoint indicator δ_t, we use

$$\ell_{\text{risk}}(x_t) = -\delta_t \log \lambda_{t,m_t} - \sum_{k=1}^{m_t - \delta_t} \log(1 - \lambda_{t,k}). \tag{2}$$

An event contributes its endpoint hazard and preceding survival terms, whereas a censored record contributes observed survival terms only; records without follow-up contribute no loss. Sampling

chooses a root and event-grid location before future labels, retains matched benign authorization controls, and calibrates on an independent workload distribution (Appendix C.1, B.3).

The predictor combines a frozen shared encoder, trainable shared projection and hazard logits, agent-local residual heads, and independently fitted temperature scaling. Its probabilities are conditional on local evidence and the reference continuation, not omniscient intent. Indistinguishable inputs limit discrimination, and risk does not determine action value; Appendices C.2 and F.3 give the controls and matched-capacity alternatives.

5.2 Counterfactual Replay for Action-Value Estimation

The authorized set is $\mathcal{U}_i(x_t) \subseteq \{\text{continue}, \text{review}, \text{block}\}$. Continue permits the operation; review pauses it within current permissions; block cancels it and its declared dependents. Accounting includes review time, blocked legitimate work, and recovery (Appendix B.4). Replay restores the complete state s_t—contexts, versioned objects, queues, dependencies, tools, permissions, and execution configuration—and regenerates the downstream workflow under joint policy Π. Restoring only observations or reusing the original suffix would omit intervention consequences.

Let $L_{H,t:}$, $L_{U,t:}$, and $C_{t:}$ denote subsequent harm loss, legitimate-task loss, and operational cost. The action value is

$$Q^{\Pi}(s_t, u) = \mathbb{E}[L_{H,t:} + \lambda_u L_{U,t:} + \lambda_c C_{t:} \mid s_t, \mathrm{do}(u_t = u), \Pi]\,. \tag{3}$$

The reference policy μ defines prediction targets, whereas Π defines marginal intervention value in the presence of other guards. We estimate Q^{Π} from executed sandbox continuations, paired randomness where possible, and inverse-probability-weighted control-point samples. The local policy need not observe s_t: repeated restored states with the same x_t provide privileged training targets while deployment remains local. Appendix B.5, B.3, and F.1 specify fidelity, uncertainty, and action-ranking checks.

5.3 Workflow-Level Optimization and Continual Updates

Because an intervention can change later observations and legitimate outcomes, we optimize the complete execution distribution P^{Π}. Let $L_H(\tau)$, $L_U(\tau)$, and $C(\tau)$ be trajectory-level harm, utility loss, and cost. We solve

$$\begin{aligned} \min_{\Pi} \quad & J(\Pi) = \mathbb{E}_{\tau \sim P^{\Pi}}\big[L_H(\tau) + \lambda_u L_U(\tau) + \lambda_c C(\tau)\big], \\ \text{subject to} \quad & \mathrm{FBR}_{\mathrm{benign}}(\Pi) \le \epsilon. \end{aligned} \tag{4}$$

Here $\mathrm{FBR}_{\mathrm{benign}}$ is complete-workflow false blocking, distinct from prefix FPR; Appendix B.6 defines the accounting and timely-effective-protection criterion. This objective is the point where the observation-level warning is connected back to the workflow-level outcomes measured in the experiments.

At update round m, freeze the deployed version $\Pi^{(m)}$ and use its local distribution as the credit-assignment baseline. For replay estimates $\widehat{Q}^{(m)}$, define

$$\widehat{A}^{(m)}(s_t, u) = \sum_{v \in \mathcal{U}_i(x_t)} \pi_i^{(m)}(v \mid x_t) \widehat{Q}^{(m)}(s_t, v) - \widehat{Q}^{(m)}(s_t, u). \tag{5}$$

We fit and calibrate a candidate predictor, freeze it during policy training, and minimize

$$\mathcal{L}_{\mathrm{policy}}(\phi) = -\mathbb{E}_{\tau \sim P^{\Pi^{(m)}}} \sum_t \sum_{u \in \mathcal{U}_i(x_t)} \pi_{\phi_i}(u \mid x_t, r_t)\, \widehat{A}^{(m)}(s_t, u). \tag{6}$$

The central state constructs targets, but the learned policy consumes only (x_t, r_t); post-proposal control points can receive action supervision for realized-harm prevention without entering the strict early-forecast population. This combines counterfactual credit assignment with the local views established in the observation section.

Deployment snapshots local inputs, predicts risk, applies authorized actions, logs transitions, and buffers later verified events. Candidates are selected on full workflows and accepted independently

only when cost decreases, benign false-blocking and resource constraints hold, and task/object lineages remain disjoint. Appendices C.3, F.3, D.3, G, and F.2 address distribution change, replay error, lifecycle cost, holdouts, and common versus endogenous feedback. Better prediction alone is not sufficient evidence of workflow improvement.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Design summary. We cross eight platform strata with thirteen actor configurations (104 cells), evaluate nine methods with paired roots and common safety–utility–latency–lifecycle accounting, and use disjoint family/root splits for learning, calibration, and held-out tests. Sampling, baseline interfaces, uncertainty, and native/adapted/dual-control protocols are specified in Appendices E, E.1, and E.3; the body reports only the estimands and headline comparisons.

Comparators. The suite is Base, Rules, Current, Discovery, Memory, Binary, Central, Graph, and AgentVax; the Memory/Central/Graph variants are architecture-motivated comparators corresponding to AGrail/JANUS/G-Safeguard (Luo et al., 2025; Xiong et al., 2026; Wang et al., 2025a). Actor draws, reviewers, permissions, and calibration roots are held fixed; equal-feedback and equal-total-work views are reported separately (Appendix F.2).

Measures. We separate the first prohibited proposal at T_H, its later label release T_D, and realized harm; online prediction uses x_t, while restored s_t is offline only. Headline metrics are realized harm, proposed violations, benign success, workflow FBR, Brier score, control lead time, end-to-end p95 latency, and lifecycle work; denominators, censoring, and intervals are in Appendix E.

6.2 EFFECTIVENESS AND COMPUTATIONAL COST

Protection versus utility. Table 1 and Figure 4 show the workflow trade-off: on the source stratum, AgentVax changes harm/benign success from 53.7/87.2% to 9.5/83.8%, while Binary reaches 7.5% harm at 83.7% benign success. AgentVax's test FBR is 5.9%, above calibration; the gain over Base therefore does not establish superiority to the strong simpler guard or a guaranteed test-time FBR budget.

Platform stratum	Base	Current	Memory	Binary	Central	AgentVax
Source	53.7 / 87.2	16.5 / 83.8	8.3 / 83.7	7.5 / 83.7	9.1 / 83.9	9.5 / 83.8
Dojo	51.8 / 86.9	10.4 / 83.5	3.1 / 83.3	2.7 / 83.6	2.3 / 83.5	2.9 / 81.8
Harm	62.4 / 84.1	9.5 / 76.2	4.7 / 78.9	4.1 / 79.4	3.9 / 79.6	4.2 / 77.3
SafetyBench	49.5 / 86.4	10.9 / 84.0	8.5 / 83.6	7.4 / 83.5	7.2 / 83.4	5.4 / 81.7
Dyn	56.8 / 87.2	14.9 / 82.9	7.3 / 84.6	6.5 / 84.5	8.8 / 84.8	8.1 / 84.7
LAB	59.7 / 86.8	13.8 / 83.4	4.3 / 84.3	3.6 / 84.7	6.6 / 84.5	6.3 / 85.1
TAMAS	55.4 / 86.0	16.2 / 81.6	11.1 / 83.4	9.7 / 83.7	11.9 / 84.0	11.7 / 84.4
DUMA	52.2 / 88.1	15.0 / 85.4	7.6 / 85.6	6.9 / 85.8	8.8 / 86.1	11.6 / 85.1

Table 1: **Evaluation.** Matched workflow outcomes: realized harm / benign success (%), equally averaged over actors, roots, and worlds. Lower harm and higher success are better. Full nine-method intervals and costs are in Appendix H.

Cost. Against Central, AgentVax has lower harm in 47/104 cells, mean cellwise p95 ratio 0.91, and lifecycle-work ratio 1.83. Online speed therefore coexists with higher background cost; at equal source/M08 total work, Binary can buy more confirmed examples and still lower harm (Appendix F.2), so the ledger does not support a complete-method efficiency claim.

Transfer and adversarial scope. On four held-out source families, AgentVax and Central yield 8.8% and 9.9% harm for representative actors. This is a frozen-rule transfer check; adaptive search raises AgentVax's any-success rate from 10.5% to 63.7% as the candidate budget grows from one to sixteen, not a single-shot ASR (Appendices G, F.3, E.3).

6.3 LEARNING FROM DOWNSTREAM INCIDENTS

Does downstream feedback improve later decisions? Figure 5 separates a common confirmed-root stream from an endogenous loop. On the common stream, AgentVax changes Brier score from 0.141

to 0.137, timely effective control from 82.2% to 85.0%, and retained-task benign success from 81.9% to 84.4%; the reported checkpoints show consistent improvement on all three measures.

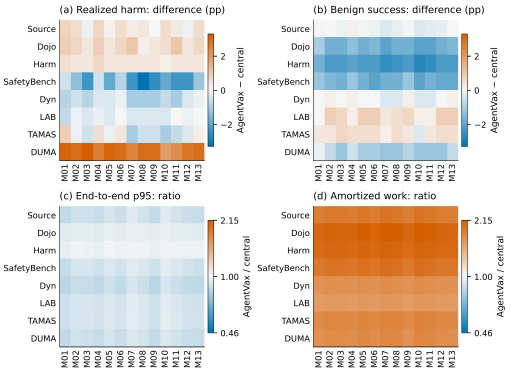

Figure 4: **Evaluation.** Same-coverage harm, utility, and cost relative to Central across platform–actor cells; top panels show differences and bottom panels ratios.

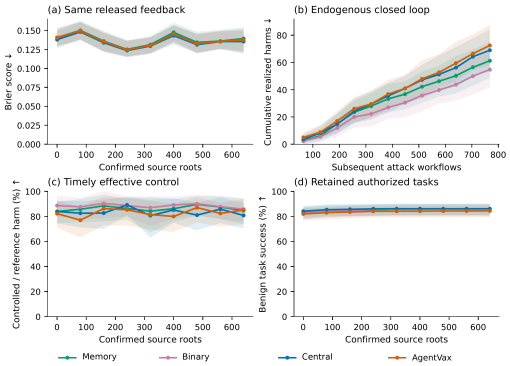

Figure 5: **Evaluation.** Common-stream and endogenous feedback responses on four source actors; bands are bootstrap intervals conditional on the stated feedback path.

Closed-loop feedback yields mean cumulative harmful-workflow counts of 72.5/69.0/61.2 for AgentVax/Central/Memory over 768 attack workflows; cleanup, delay, selective-feedback, and ordering controls are in Appendices G and F.1.

6.4 The role of local information

Does behavior use evidence rather than role names? With actors, gates, permissions, exposure, and roots fixed, removing authorization/provenance changes source early-risk and harm by -7.4/6.9 points; role renaming/sham removal do not, and restoration recovers the intact result (Figure 6).

6.5 From risk prediction to effective intervention

When does prediction justify action? Figure 7 crosses evidence with authority: at zero permission, better forecasts cannot cancel harm; replay regret 0.043 versus 0.150 for risk ranking shows that proposal probability is not action value under frozen Π.

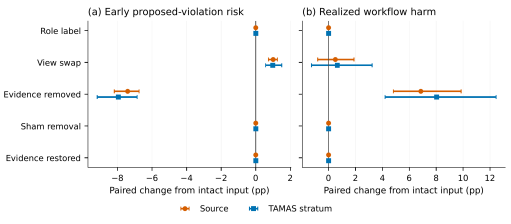

Figure 6: **Evaluation.** Paired deployment-information controls for early proposal risk and realized harm; zero is the exact negative control.

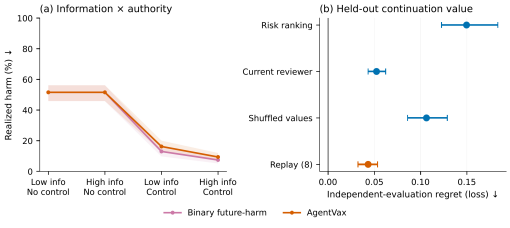

Figure 7: **Evaluation.** Prediction versus control under information-by-authority interventions; right panel reports decision regret.

7 Conclusion

AgentVax converts delayed, verified downstream incidents into local supervision and uses restored continuations to separate predictive evidence from effective control. Across the reported workflow studies, complementary views improve early warning, local adaptation supports directed transfer, and replay-derived action values improve intervention decisions under explicit utility and cost constraints.

ETHICS STATEMENT

All experiments use isolated, authorized workflows; no human subjects, personal data, production systems, or live attack services are involved. We report aggregate safety outcomes and omit deployable exploit payloads, credentials, and operational instructions. Any released implementation should preserve the sandbox, authorization, and audit boundaries described in the appendices.

AI USE STATEMENT

Generative AI tools assisted with codinng, manuscript preparation and editing, evaluation design checks, code and figure preparation, translation, and LaTeX formatting. The authors independently checked the mathematical statements, citations, numerical macros, figures, and final PDF, and take responsibility for all claims, data, code, and artifacts.

REFERENCES

Ivan Aleksandrov, German Kochnev, Sabrina Sadiekh, and Yaroslav Rogoza. DUMA-Bench: A dual-control multi-agent benchmark for evaluating LLM agent security. *arXiv:2609.24662v1*, 2026. URL https://arxiv.org/abs/2609.24662v1.

Maksym Andriushchenko, Alexandra Souly, Mateusz Dziemian, Derek Duenas, Maxwell Lin, Justin Wang, Dan Hendrycks, Andy Zou, Zico Kolter, Matt Fredrikson, Yarin Gal, and Xander Davies. AgentHarm: A benchmark for measuring harmfulness of LLM agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/c493d23af93118975cdbc32cbe7323f5-Abstract-Conference.html.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *Advances in Neural Information Processing Systems*, volume 37, 2024. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/f545448535dfde4f9786555403ab7c49-Abstract-Conference.html.

Yanjun Chen, Yirong Sun, Hanlin Wang, Jinghan Wang, Xinming Zhang, Xiaoyu Shen, Wenjie Li, and Wei Zhang. Exact is easier: Credit assignment for cooperative LLM agents. *arXiv:2603.06859v2*, 2026. URL https://arxiv.org/abs/2603.06859v2.

Wuyang Dai and Song Wang. AgentGuard: Learning execution guardrails from anomalous coding-agent trajectories. *arXiv preprint arXiv:2609.16287*, 2026. URL https://arxiv.org/abs/2609.16287.

Edoardo Debenedetti, Jie Zhang, Mislav Balunovic, Luca Beurer-Kellner, Marc Fischer, and Florian Tramer. AgentDojo: A dynamic environment to evaluate prompt injection attacks and defenses for LLM agents. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-2636. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/97091a5177d8dc64b1da8bf3e1f6fb54-Abstract-Datasets_and_Benchmarks_Track.html. Datasets and Benchmarks Track.

Jakob Foerster, Gregory Farquhar, Triantafyllos Afouras, Nantas Nardelli, and Shimon Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018. doi: 10.1609/aaai.v32i1.11794. URL https://ojs.aaai.org/index.php/AAAI/article/view/11794.

Michael F. Gensheimer and Balasubramanian Narasimhan. A scalable discrete-time survival model for neural networks. *PeerJ*, 7:e6257, 2019. doi: 10.7717/peerj.6257. URL https://doi.org/10.7717/peerj.6257.

Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102(477):359–378, 2007. doi: 10.1198/016214506000001437. URL https://doi.org/10.1198/016214506000001437.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

Monte Hoover, Vatsal Baherwani, Neel Jain, Khalid Saifullah, Joseph J. Vincent, Chirag Jain, Melissa Rad, C. Bruss, Ashwinee Panda, and Tom Goldstein. DynaGuard: A dynamic guardian model with user-defined policies. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8d2aad53fa148791ffd004a7c9cce5ad-Abstract-Conference.html.

Tanqiu Jiang, Yuhui Wang, Jiacheng Liang, and Ting Wang. AgentLAB: Benchmarking LLM agents against long-horizon attacks. *arXiv preprint arXiv:2602.16901*, 2026. URL https://arxiv.org/abs/2602.16901v1.

Ishan Kavathekar, Hemang Jain, Ameya Rathod, Ponnurangam Kumaraguru, and Tanuja Ganu. TAMAS: Benchmarking adversarial risks in multi-agent LLM systems. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 31238–31268. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1442. URL https://aclanthology.org/2026.acl-long.1442/.

Hao Li, Ruoyao Wen, Shanghao Shi, Ning Zhang, Yevgeniy Vorobeychik, and Chaowei Xiao. AgentDyn: Are your agent security defenses deployable in real-world dynamic environments? *arXiv preprint arXiv:2602.03117*, 2026. URL https://arxiv.org/abs/2602.03117v3.

Weidi Luo, Shenghong Dai, Xiaogeng Liu, Suman Banerjee, Huan Sun, Muhao Chen, and Chaowei Xiao. AGrail: A lifelong agent guardrail with effective and adaptive safety detection. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8104–8139. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.acl-long.399. URL https://aclanthology.org/2025.acl-long.399/.

Stéphane Ross and J. Andrew Bagnell. Reinforcement and imitation learning via interactive no-regret learning, 2014. URL https://arxiv.org/abs/1406.5979. arXiv preprint arXiv:1406.5979.

Mikhail Terekhov, Alexander Panfilov, Daniil Dzenhaliou, Caglar Gulcehre, Maksym Andriushchenko, Ameya Prabhu, and Jonas Geiping. Adaptive attacks on trusted monitors subvert AI control protocols. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/54b153ad8a138f4c186f21a8b7341d5e-Abstract-Conference.html.

Shilong Wang, Guibin Zhang, Miao Yu, Guancheng Wan, Fanci Meng, Chongye Guo, Kun Wang, and Yang Wang. G-Safeguard: A topology-guided security lens and treatment on LLM-based multi-agent systems. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7261–7276. Association for Computational Linguistics, 2025a. doi: 10.18653/v1/2025.acl-long.359. URL https://aclanthology.org/2025.acl-long.359/.

Xinpeng Wang, Chengzhi (Martin) Hu, Paul Röttger, and Barbara Plank. Surgical, cheap, and flexible: Mitigating false refusal in language models via single vector ablation. In *International Conference on Learning Representations*, 2025b. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/53dbd7e34fab703a639964e2d3ee9e84-Abstract-Conference.html.

Yuan Xiong, Linji Hao, Shizhu He, Yequan Wang, and Lijun Li. JANUS: Foreseeing latent risk for long-horizon agent safety. *arXiv:2607.19913v1*, 2026. URL https://arxiv.org/abs/2607.19913v1.

Fred Zhang and Neel Nanda. Towards best practices of activation patching in language models: Metrics and methods. In *International Conference on Learning Representations*, 2024. URL https://arxiv.org/abs/2309.16042v2.

Hanrong Zhang, Jingyuan Huang, Kai Mei, Yifei Yao, Zhenting Wang, Chenlu Zhan, Hongwei Wang, and Yongfeng Zhang. Agent security bench (ASB): Formalizing and benchmarking attacks and defenses in LLM-based agents. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/5750f91d8fb9d5c02bd8ad2c3b44456b-Abstract-Conference.html.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-SafetyBench: Evaluating the safety of LLM agents. *arXiv preprint arXiv:2412.14470*, 2024. URL https://arxiv.org/abs/2412.14470v1.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency. *arXiv preprint arXiv:2310.01405*, 2023. URL https://arxiv.org/abs/2310.01405.

Egor Zverev, Sahar Abdelnabi, Soroush Tabesh, Mario Fritz, and Christoph Lampert. Can LLMs separate instructions from data? and what do we even mean by that? In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/a77eadda332b6d4a9ae1e0e4024555f2-Abstract-Conference.html.

Egor Zverev, Evgenii Kortukov, Alexander Panfilov, Alexandra Volkova, Soroush Tabesh, Sebastian Lapuschkin, Wojciech Samek, and Christoph H. Lampert. ASIDE: Architectural separation of instructions and data in language models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/b77b5f1a2878fedd4705bfeafd24ca35-Abstract-Conference.html.

A Observation Evidence and Supporting Experiments

A.1 Scope, targets, and common protocol

The observation study spans 24 attack families, eight workflow templates, two domains, four model tiers, and five seeds. Within one organization, agents encounter attack-bearing email/documents, retrieval, memory, and tool outputs. Agent i observes only $x_{i,t} = O_i(\tau_{\le t}; c_i, h_i)$, with role/context c_i and runtime h_i. Read, Transform, Plan, and Act expose, respectively, incoming sources, derived content/provenance, authorized goals/plans, and proposed actions/permissions; they are stages rather than fixed agent types. Executable agent evaluation motivates this setting (Zhang et al., 2025).

For task g and policy $\mathcal{P}$, the two endpoints are

$$Y^{\text{prop}} = \mathbb{1}[\exists t : \ a_t^{\text{proposed}} \notin \mathcal{A}_{\mathcal{P}}(s_t, g)], \tag{7}$$

$$Y^{\text{real}} = \mathbb{1}[\exists t : \ s_{t+1} \in \mathcal{F}_{\mathcal{P}}]. \tag{8}$$

Gating can alter execution without changing a proposal. Authorized-task success is separate. Attack–benign pairs match domain, tools, vocabulary, and difficulty; hard negatives include authorized sensitive reads, administrative actions, permission changes, and outgoing communication. Inputs freeze before each decision, including the current-full-prefix reference. Future outcomes enter labels only. The protocol separates root incidents, tasks, templates, and families before augmentation, and keeps update, calibration, transfer-selection, and final-test sets disjoint. Final verification requires execution/provenance records, policy checks, and human review; model labels alone are insufficient.

Thresholds are fixed at 5% validation benign FPR; each aggregate is calibrated as one system. Probability calibration uses both classes, whereas FPR thresholding uses negatives. ECE concerns calibration rather than ranking (Guo et al., 2017). Tables 2(a) and 3(a) report 95% bootstrap intervals. This observation protocol uses the same disjoint update/calibration/test discipline as the main evaluation, with its resampling procedure stated separately in Appendix E.1.

A.2 Local evidence, complementary detection, and forecasting

The stage-level hazard motifs in Figure 1a are analytical shared patterns, not required runtime objects. In row order, Read/Transform/Plan/Act identifiability is $(.86, .61, .38, .24)$, $(.78, .69, .42, .27)$, $(.21, .46, .83, .71)$, $(.07, .16, .58, .93)$, and $(.04, .13, .49, .95)$. These fractions are neither TPR nor first-visibility times. Downstream TPR is 64.7% with provenance, 42.1% after deletion, and 62.9% after restoration; content-only hazards decrease by 3.4 percentage points. The manipulation distinguishes content from instruction authority (Zverev et al., 2025).

Table 2(a) fixes weights and total inference budget across local views, repeated sampling, role variants, and multi-view aggregation; validation alone fits the aggregate. View and role swaps move detection profiles by 68% and 19%, respectively, and identical full-prefix inputs reduce unique detections by 46%. Profile movement is the fractional reduction in normalized L_1 distance to the destination condition's detection profile after the swap, relative to the pre-swap distance. Unique detections are normalized over attack examples detected by at least one compared condition. Reported Jaccard overlaps use $|D_i \cap D_j|/|D_i \cup D_j|$.

Forecasting uses current prefixes and $Y_{t:t+k,m} = \mathbb{1}[\exists j \in \{1, \dots, k\} : Z_{t+j,m} = 1]$, with $k = 2$ events. Table 2(b) reports lead time in stage units; missed warnings remain in the coverage denominator and do not receive a lead-time value. Late-revealed unauthorized parameters have Read AUPRC .23. At equal total token budget, single-view self-consistency, same-view aggregation, and multi-view aggregation reach .82, .83, and .88 AUROC. Full-prefix access remains current-only.

A.3 Directed transfer and chronological expansion

Source-only updates on family a are evaluated on held-out family b, with source size, capacity, steps, and label entropy matched:

$$T_{a \to b}^{(s)} = \text{TPR}(f_s^{+a}, D_b) - \text{TPR}(f_s, D_b). \tag{9}$$

Positive transfer requires a positive 95% interval lower bound, at most one percentage point additional benign FPR, and stability across paraphrases, seeds, and another model condition. Pair relations are

Table 2: Local detection and forecasting. (a) TPR (%) and 95% bootstrap intervals; gains are percentage points relative to Plan. (b) Two-event forecasts; lead time is reported in stage units. Dashes denote inapplicable entries.

(a) Same-model detection

Condition	TPR	AUPRC	Jaccard	Gain
Read	41.8 ± 1.7	.49	–	-14.1
Transform	46.3 ± 1.5	.53	–	-9.6
Plan	55.9 ± 1.8	.62	–	0.0
Act	52.7 ± 1.6	.60	–	-3.2
Repeated $\times 4$	59.8 ± 1.4	.65	.82	3.9
Role-diverse $\times 4$	61.2 ± 1.6	.66	.75	5.3
Multi-view	73.6 ± 1.3	.78	.47	17.7
Full prefix	78.4 ± 1.1	.83	–	22.5

(b) Finite-horizon forecasts

Input	AUROC	AUPRC	ECE	Lead time (stages)
Read	.69	.47	.10	1.86
Transform	.73	.52	.09	1.51
Plan	.80	.61	.07	.92
Naive OR	.82	.63	.14	1.73
Confidence-wtd.	.86	.69	.06	1.69
Learned	.88	.72	.04	1.74
Full prefix	.92	.80	.03	2.02

fixed before outcome inspection. Table 3(a) gives the four relations. Pattern overlap, text similarity, taxonomy, and pattern-plus-history predictors achieve .78, .60, .64, and .82 AUROC; fitting and edge selection use historical pairs only. Among pairs positive in either direction, 61% are asymmetric; complex-to-simple gains are 10.7 points versus 4.1 in reverse, with 8% negative transfer among candidates.

The directed analysis graph has three major components and four isolated families among 24. Chronological selection uses verified history, evaluates the next unseen batch, and only then admits its outcomes. Table 3(b) retains all five reported three-seed conditions. Coverage is measured directly on target tasks rather than inferred from multi-hop reachability. The full-matrix oracle uses future labels and is not deployable. Neither this graph nor a motif extractor is required by AgentVax.

Table 3: Transfer and chronological expansion. (a) Gains and ΔFPR are percentage points; gain intervals are 95% bootstrap intervals, and positive pairs are percentages. A group-average FPR breach need not apply to every pair. (b) Three seeds; coverage/negative updates/FPR are percentages and precision/recall are fractions.

(a) Source–target relations

Relation	Gain	Positive (%)	ΔFPR
Same mechanism, new surface	12.4 ± 2.1	71	.4
Shared pattern, new goal	9.1 ± 1.8	63	.5
Same harm, new mechanism	2.2 ± 1.5	24	.8
Lexical similarity, new pattern	-0.8 ± 1.3	11	1.6

(b) Chronological selection

Selection	Cover.	Neg.	Prec.	Rec.	FPR
Random	45.8	12.6	.44	.56	5.0
Semantic nbrs.	50.0	10.9	.55	.71	5.8
Taxonomy	54.2	8.7	.64	.62	5.4
Transfer-based	66.7	4.3	.78	.65	5.1
Oracle	75.0	2.1	1.00	1.00	5.0

A.4 Local adaptation and workflow outcomes

Table 4(a) compares seven updates on identical verified source events with frozen base models and matched capacity/data/update budgets. The role-matched, graph-conditioned comparator updates local heads or context/runtime rules; it is not the end-to-end AgentVax method. Tests separate seen variants, shared-pattern unseen families, and disconnected families. On the latter, 40.6% TPR is 1.7 points above no update (38.9%), not pooled updates (43.1%); its interval spans zero. Oracle assignment uses future target labels.

Chronological workflow replay mixes these attacks with matched benign tasks and randomizes the first discovery node (Table 4(b)). Critical-action reachability is 18.5% for the central guard and 10.8% for role-matched adaptation. The latter's 91.7% workflow benign success differs from 91.9% in the local-update experiment. Recurrence, critical-action reachability, and unseen ASR have distinct targets and are reported with their own denominators.

Table 4: Local updates and workflow replay. (a) Seen/shared-pattern/disconnected TPR and benign success (%); cost is table-specific relative work. Parameter avg. means parameter averaging. (b) Recurrence, reachability, unseen ASR, and benign success (%); latency is relative to no defense. Outcomes, utility populations, and cost units differ between panels.

(a) Local updates

Update	Seen	Shared	Disc.	Benign	Cost
None	43.5	39.8	38.9	93.2	0.0×
Local only	71.4	47.6	39.7	92.8	.6×
Broadcast	79.1	62.1	42.5	86.1	4.0×
Pooled	80.3	65.4	43.1	88.7	3.7×
Parameter avg.	78.8	63.0	42.8	89.1	3.2×
Role-matched	82.6	72.8	40.6	91.9	1.6×
Oracle	85.4	76.9	44.0	92.1	1.4×

(b) Workflow replay

Defense	Recur.	Reach	ASR	Benign	Lat.
None	48.7	31.6	45.2	94.0	0%
Static local	36.4	24.8	38.9	91.8	+11%
Central guard	29.1	18.5	30.7	88.6	+38%
Broadcast	23.8	14.7	24.9	86.9	+17%
Object repair	31.8	25.6	42.0	93.5	+2%
Stage-matched	27.0	16.9	29.4	92.1	+9%
Role-matched	19.5	10.8	17.6	91.7	+7%

A.5 Functional interventions, representations, and verification limits

Table 5(a) sequentially supplies correct labels, provenance/authority, safe decisions, and enforcement. Its gains are order-dependent increments, not disjoint modules or an order-invariant attribution. Proposals and execution remain separate when assessing enforcement.

The readouts compare content hazard H, authority A, goal alignment G, and capability C (Table 5(b)); projected bases satisfy $U^\top U = I$ and use the dimensionality of the included feature set. The TPR column reports violating-proposal detection at the common 5% validation benign-FPR threshold. Source-sensitive representations motivate including authority (Zverev et al., 2026); refusal directions are different objects, not baselines evaluated here (Arditi et al., 2024; Wang et al., 2025b).

Table 5(c) reports five interventions against 68.2% unmodified violating proposals. Recovery is the percentage of baseline violating proposals converted to non-violating outcomes under the matched intervention; benign Δ is the percentage-point change in benign success on matched benign examples. Full matched replacement is a reference condition, and neither readout nor patching alone establishes natural causal mediation. The complementary representation test appears in Appendix F.4.

Table 5: Functional and representation diagnostics. (a) Sequential additional safe-completion gain (percentage points). (b) First three columns: AUROC; final TPR is violating-proposal detection at 5% validation benign FPR. (c) Proposals and recovery are percentages; benign Δ is a percentage-point change in benign success. Unmodified proposals: 68.2%.

(a) Conditional functional increments

Stage	Label	Provenance	Decision	Enforcement
Read	18.4	5.1	3.0	1.2
Transform	8.2	16.9	4.8	1.6
Plan	4.0	8.5	17.6	3.1
Act	3.1	5.0	9.2	22.4

(b) Readouts

Readout	Same	Surface	Role	TPR (%)
v_H	.88	.73	.64	49.2
U_{HA}	.90	.82	.78	67.5
U_{HAGC}	.92	.86	.82	72.4

(c) Representation interventions

Component	Prop.	Recov.	Benign Δ
Random	66.9	3	$-.4$
Content	54.7	29	-4.8
Authority	43.9	45	-1.9
Joint	29.6	61	-2.2
Full matched	21.4	74	-2.7

Verification boundary. Shared experience presupposes provenance/policy validation, replay, benign regression tests, and quarantine of unverified cases. Monitor attacks motivate independent adjudication (Terekhov et al., 2026). Cross-organization trust, arbitrary parameter composition, and richer actions remain outside scope.

B DATA, FEEDBACK, AND REPLAY PROTOCOLS

This appendix specifies the data, feedback, replay, and evaluation procedures for AgentVax. The organization, authorization policy, and attack access follow Section 3; an initially malicious objective is compatible with a target violation that appears later.

B.1 EVENT CLOCKS AND THE LOCAL INFORMATION BOUNDARY

A workflow is an execution trace with a predefined maximum length $T_{\max}$. An event is a logged workflow transition, such as a read, message delivery, output proposal, tool commitment, or review transition. The event vocabulary and treatment of compound operations are fixed before constructing examples. Concurrent events are ordered by a logged schedule consistent with causal dependencies; order-sensitive conclusions are tested across schedules. The active decision-maker is $i(t)$, and t identifies the checkpoint immediately before its safety action. Automatic transitions can be represented with a single available action. Event counts, stage counts, and wall-clock time are reported separately.

For each checkpoint we retain the exact serialized observation x_t, its hash, the input schema and truncation rule, and the versions of each object actually read. Effective local information includes agent identity, visible permissions, and any metadata supplied to the guard. Truncation occurs according to the same rule online and offline. An object read at version v contributes that version's bytes even if the live object later changes to $v + 1$. Future tool outputs, downstream explanations, verifier deliberations, and retrospectively reconstructed summaries are stored in a separate supervision record. None is appended to an earlier input.

The administrator's policy specifies prohibited instructions, operations, recipients, resource states, and authorization conditions. The event registry distinguishes Y_e^{prop}, indicating a violating proposal, from Y_e^{real}, indicating realized harm. A blocked proposed operation may satisfy $Y_e^{\text{prop}} = 1$ and $Y_e^{\text{real}} = 0$. A prohibited output released to a recipient may itself constitute realized harm; the policy fixes when this occurs. Deduplication links repeated descriptions to one event without collapsing separate harmful state changes.

For the primary target, T_H is the occurrence of the first verified violating proposal and T_D is its label-release time. Both have audit timestamps, and occurrence precedes confirmation. The notation $T_H - t$ uses their workflow-event coordinates; chronological training cutoffs use the recorded wall-clock release timestamp associated with T_D. These coordinates are never subtracted across units. A proposal is timestamped when it becomes available to the agent or guard, not when it is later executed. Thus, snapshots already containing that proposal are outside its strictly early-prediction risk set. A violating proposal at the initial decision can have no eligible earlier snapshot. Control evaluation still includes subsequent opportunities to reduce realized harm.

Policy checks use tool arguments, recipients, permissions, and state deltas when executable evidence exists. Semantic cases require policy-based adjudication and independent review. Confirmation can locate an earlier occurrence from immutable logs. An event with an uncertain occurrence time remains usable for event-level auditing but has no precise survival label in the minimal implementation. Interval-censored targets and independently specified repeated-event episodes are extensions. A realized-harm forecasting target requires a separately trained and reported predictor, not a relabeling of the proposal predictor's output.

B.2 REFERENCE OUTCOMES, FOLLOW-UP, AND LABEL RELEASE

The reference policy μ fixes task models, prompts, tools, base safeguards, and the continuation schedule distribution. AgentVax's additional interventions run in logging-only mode. All external effects are redirected to isolated resources with auditable semantics. Changing a task model, base safeguard, tool implementation, or relevant schedule law changes the reference condition and requires a new version and compatible labels. A restored checkpoint may already include effects of an earlier deployed intervention; μ specifies the continuation from that checkpoint, not a fictitious reconstruction of its past.

We pair each eligible local snapshot with reference continuations from its associated complete state. A confirmed deployed event identifies historical examples worth learning from and supplies an event

annotation. Its realized deployed suffix supplies a reference label only when that suffix was generated under the reference condition. Otherwise, the annotation triggers isolated μ continuations, whose outcomes determine (m_t, δ_t). This preserves the distinction between shared delayed feedback and the execution law being predicted. Multiple independent continuations from one checkpoint estimate a conditional distribution; they are kept in the same statistical cluster.

Let C_t^{adm} be the number of fully observed follow-up intervals before administrative truncation. For an eligible event-free snapshot, generate the record as follows:

1. If the reference target occurs at $1 \le D_t \le K$ before administrative truncation, set $m_t = D_t$ and $\delta_t = 1$.
2. If the full K-event window is observed without the target, set $(m_t, \delta_t) = (K, 0)$. An event beyond that window has this same finite-horizon label.
3. If the workflow legitimately completes without the target, complete the window with an absorbing no-event state. Cross-task persistent-memory effects require a separately declared multi-task window.
4. If administrative observation ends first, set $m_t = \min(K, C_t^{\mathrm{adm}})$ and $\delta_t = 0$. Apply the ordinary censored likelihood only under conditional independence of event and censoring given the effective local input, with positive observation probability at evaluated horizons.
5. If a safety intervention caused the observed stop, restore a compatible checkpoint and run the reference continuation. An unsuccessful restore produces missing reference supervision, not a negative outcome or automatically valid administrative censoring.

The observation convention includes the endpoint event when it is logged at the same boundary; ambiguous ordering is resolved from the audit trace. Records with $m_t = 0$ contribute zero loss. Infrastructure failure, unresolved labels, and evaluator timeout are not normal safe completion. Missing occurrence times, unresolved adjudications, and missing reference outcomes are reported separately by family, action, agent, and reason.

Training eligibility follows label release rather than event occurrence. A reference label becomes available only after its continuation and verification finish, and an event-triggered replay cannot precede the release of its parent feedback. Each round records feedback collection time, replay completion, dataset freeze, training completion, and deployment time. Labels not ready at dataset freeze wait for a later round. A negative window requires actual follow-up and completed adjudication; a pending event report is not evidence of safety. Evaluation predicts before updating on that task's newly released feedback.

B.3 Sampling populations and finite-budget weighting

Risk estimation is defined jointly by a continuation law and a starting-state population. Let ρ_m be the declared mixture of independent root workflows, including new events, historical events, representative benign tasks, and matched benign controls. First draw a root workflow, then draw t uniformly from the fixed grid $\{0, \dots, T_{\max} - 1\}$, before inspecting its future labels. A location contributes only if a decision exists and the specified target has not yet occurred. Reference continuations from this checkpoint define the joint law of (X, D). Equivalently, the population training objective is the expected sum of eligible losses over this fixed grid, divided by the constant $T_{\max}$.

This population differs from drawing uniformly among each realized trace's pre-violation prefixes, whose normalizing factor depends on a future outcome. We do not use that latter normalization. Unequal root or grid sampling uses known sampling weights relative to the declared target. Class balancing and matched benign augmentation change the training mixture; representative calibration and reported population weights restore the intended probability interpretation. The effective distribution of hidden states conditional on x can shift across rounds even when μ stays fixed. Calibration validity is therefore stated for both the reference version and the prefix population, with distribution shifts explicitly evaluated.

Policy training instead targets the complete decision sum on trajectories from $P^{\Pi^{(m)}}$. For root trace n with T_n available control points, draw a uniform sample $\mathcal{B}_n$ without replacement of size $\min(B, T_n)$, independently of replay outcomes. Its inclusion probability is $p_{n,t} = \min(1, B/T_n)$. If $T_n = 0$, its

policy contribution is zero. For N root traces, the estimator of Equation (6) is

$$\widehat{\mathcal{L}}_{\text{policy}} = -\frac{1}{N}\sum_{n=1}^{N}\sum_{t\in\mathcal{B}_n}\frac{1}{p_{n,t}}\sum_{u\in\mathcal{U}_{i(t)}(x_{n,t})}\pi_{\phi_{i(t)}}(u \mid x_{n,t}, r_{n,t})\,\widehat{A}^{(m)}(s_{n,t}, u). \tag{10}$$

Conditional on a trace, $\mathbb{E}[\mathbf{1}\{t \in \mathcal{B}_n\}/p_{n,t}] = 1$, so this recovers its decision sum. Averaging independent replay samples makes the target unbiased for a fixed candidate under valid restoration; choosing a candidate on the same data still requires separate validation. Risk-prioritized sampling is an optional budget allocation with a positive exploration component and known inclusion probabilities. Arbitrary priority scores are not inclusion probabilities. Failed restorations introduce outcome-dependent missingness that inverse-probability control-point weighting alone cannot correct.

B.4 Action semantics and counterfactual execution

The runtime constructs $\mathcal{U}_i(x_t)$ from the checkpoint's permissions and operation type. A mask removes unavailable actions before normalization. The guard cannot grant itself access to a new source or tool. A continue action applies the pending operation through the ordinary actor and tool interface. A block action cancels that output or operation and propagates cancellation only to tasks explicitly dependent on it. Recovery, retries, and independent branches follow the declared runtime rules.

Review permits at most one fixed review procedure per control point. Its model, prompt, accessible evidence sources, maximum calls, timeout, and timeout outcome are fixed within a comparison. It may retrieve currently authorized information, but not a later event description or a privileged central checkpoint. A review returns release or rejection; both its intermediate calls and elapsed time are charged. Independent branches may advance while an operation is paused. The scheduler and latency model account for harm that occurs during review, so review is not a free global pause.

For a sampled checkpoint and each available action, restore the same full pre-action state and run M independent continuation replicates under the frozen baseline $\Pi^{(m)}$. Let $\ell_{t,u}^{(j)}$ denote the total subsequent weighted loss in replicate j. Then

$$\widehat{Q}^{(m)}(s_t, u) = \frac{1}{M}\sum_{j=1}^{M}\ell_{t,u}^{(j)}, \qquad \ell_{t,u}^{(j)} = L_{H,t:}^{(j)} + \lambda_u L_{U,t:}^{(j)} + \lambda_c C_{t:}^{(j)}. \tag{11}$$

The stored record retains each component, the intervention's scope, seed, terminal status, and paired action differences. Common random numbers may couple compatible environmental draws across actions while preserving each branch's correct marginal distribution. Branch-dependent tool requests receive results from their own parameters and resource state; pairing never substitutes an incompatible cached response. When deterministic model seeds are not exposed by the backend, pairing uses the available shared environmental randomness while each continuation is still executed independently.

B.5 Restoration fidelity and supervision audits

A checkpoint includes object bytes and versions; all relevant agent contexts and memories; pending messages, tasks, and dependencies; tool and resource state; capabilities and authorization; model, guard, and prompt versions; the event clock; scheduling state; and recorded random streams where available. A replay may regenerate model outputs, but it must reproduce the specified pre-action condition. Suffix text from the original trajectory is never used as a substitute for new execution.

The audit suite restores an unmodified checkpoint and compares a continue branch with its original execution. Deterministic components must match hashes, tool arguments, state deltas, and dependencies. Stochastic components are checked across repeated runs for discrepancies in events, utility, and loss beyond expected sampling variation. Separate tests modify a versioned object, authorization, or queued dependency and verify that later outputs respond to the modified state. Cross-branch contamination is checked with fresh isolated resources. Failed fidelity tests exclude the affected restoration class from causal action-value claims and trigger explicit coverage reporting.

Label audits stratify by family, proposal versus realized harm, ambiguous authorization, and negative outcomes. An independent adjudicator reviews a blinded sample, including disagreements and apparent benign blocks. The audit reports disagreement, corrected labels, unresolved cases, and

timing accuracy. Tool-based outcome verification is primary when available; an LLM judge is an auditable measurement component, not ground truth by declaration. Restoration and adjudication budgets are counted separately from ordinary inference.

B.6 Loss accounting and timely effective protection

For deduplicated policy events e, let $I_e \geq 0$ be a predeclared severity weight. An executable harm loss is

$$L_H(\tau) = \sum_e I_e\big(Y_e^{\mathrm{real}} + \alpha_{\mathrm{p}} Y_e^{\mathrm{prop}}\big), \qquad \alpha_{\mathrm{p}} \geq 0. \tag{12}$$

The primary reported safety outcome remains realized harm; raw proposal and realized-harm rates are reported separately. The optional proposal penalty is fixed before evaluation and varied in sensitivity analysis. An immutable ledger assigns each increment once when incurred. If a proposal precedes a block, its proposal penalty remains; if harm has already occurred, its realized-harm increment remains. Suffix losses include only increments not yet incurred, with terminal utility loss charged at completion. This makes Equation (3) consistent with the complete loss in Equation (4).

Let b indicate a root task drawn from the benign distribution and $U(\tau) \in [0,1]$ its authorized completion score. We use $L_U(\tau) = b[1 - U(\tau)]$. For attacks embedded in an otherwise legitimate task, attacked-task completion is an additional reported outcome, not a silent change to this primary utility definition. Task mixtures and the weights λ_u, λ_c remain fixed across comparisons. Operational C combines declared scales for additional inference, tool/review calls, and latency. Training, reference replay, action replay, calibration, selection, acceptance, and storage are reported as lifecycle costs and subject to separate budgets.

For a benign workflow, let $F(\tau)$ indicate that at least one action legal in its contemporaneous context was wrongly blocked, including rejection after review. Then $\mathrm{FBR}_{\mathrm{benign}} = \mathbb{E}[F \mid b = 1]$. Review followed by release primarily incurs cost; erroneous rejection can affect both FBR and task utility. The denominator includes all sampled benign roots, not only roots that reached a selected gate. We also report benign success, because FBR does not capture every form of task degradation.

An alarm timestamp, an intervention-decision timestamp, a control-effective timestamp, a violating-proposal timestamp, a realized-harm timestamp, and a label-release timestamp answer different questions. A strict forecast alarm precedes the reference proposal event. Timely protection requires that the intervention actually becomes effective before the corresponding reference harm and reduces that harm within a fixed terminal window. Suppressing one route while harm reappears through another route within that window is not successful prevention. Paired reference scenarios establish the comparison; stochastic prevention effects are estimated across independent replicates, not inferred from one unusually favorable branch.

We report event-step and wall-clock lead times together with the fraction of reference harmful scenarios successfully protected. Failed or late interventions remain in the denominator and are reported explicitly; conditional lead-time summaries among successes cannot replace this coverage measure. Evidence for useful early forecasting is evaluated at matched benign false-block or utility levels. Decision time and control-effective time remain distinct when review, queueing, or tool latency intervenes.

C Statistical Targets and Policy-Improvement Conditions

The results below instantiate standard logarithmic-loss and finite-horizon decision identities for the information and execution conditions of AgentVax. They characterize population targets and the need for complete-policy validation. They do not establish finite-sample calibration, automatic specialization, or protection against every future attack.

C.1 Population target of the survival loss

Fix the reference version μ, the root and event-grid sampling law of Appendix B.3, and the event-free eligibility rule. Together these induce a distribution over effective local input X and reference remaining time D. For horizon k, define $p_k(x) = \Pr_\mu(D = k \mid D \geq k, X = x)$. Let R_k indicate that a record is at risk and its kth interval is observed and adjudicated. We require either

complete observation or non-informative observation conditional on X and the risk set, with positive observation probability at the horizons of interest. In particular, outcome-dependent verification or label release must be resolved by complete reference adjudication, an identified observation model, or a restricted reported population; it does not disappear by naming a record censored.

Proposition C.1 (Conditional survival target)**.** *Suppose the conditional hazards are representable and the preceding observation conditions hold. The population expectation of Equation* (2) *is minimized by $\lambda_k(x) = p_k(x)$ at every input and horizon with positive observed risk-set mass. The induced cumulative probability is the conditional probability of the target by that horizon.*

Proof. Equation (2) is a sum of Bernoulli logarithmic losses over observed at-risk intervals. Given $X = x$, $D \geq k$, and observation of interval k, the event indicator has probability $p_k(x)$ under the assumptions. Its expected loss is

$$-p_k \log \lambda_k - (1 - p_k) \log(1 - \lambda_k) = H(\mathrm{Ber}(p_k)) + \mathrm{KL}\left(\mathrm{Ber}(p_k) \,\|\, \mathrm{Ber}(\lambda_k)\right). \tag{13}$$

The divergence is nonnegative and vanishes exactly at the true probability, with the usual boundary conventions for probabilities zero and one. Observation multiplies this conditional contribution by a nonnegative risk-set weight, which is positive on identifiable intervals. Summing over k and X preserves the minimizer. Finally, conditional survival factorizes as $\Pr(D > k \mid X) = \prod_{j=1}^{k}(1 - p_j(X))$, proving the cumulative-risk statement. □

This is the standard strict propriety of logarithmic scoring applied to discrete-time survival likelihoods (Gneiting & Raftery, 2007; Gensheimer & Narasimhan, 2019). With restricted capacity, optimization instead finds a risk-set-weighted approximation. Finite data, inaccurate reference restoration, and label noise can change its quality. Temperature scaling is consequently evaluated, not assumed to establish calibration. Conditional independence given the complete state alone is insufficient when a censoring mechanism remains informative after conditioning only on the model's local input.

Both μ and the starting-state distribution matter. Let ν denote that distribution and $X = x(S)$. The conditional event law integrates reference outcomes over $\nu(ds \mid X = x)$. A change in the latent-state mixture can therefore change the correct local prediction without changing μ. Risk calibration is evaluated on a declared, representative prefix population with mature adjudication. A risk vector can still be supplied as a policy feature at post-target control points, but this proposition grants it no first-event probability interpretation outside the event-free population.

C.2 Limits imposed by local information

Proposition C.2 (Local indistinguishability)**.** *For the binary target $Y_k = \mathbf{1}\{D \leq k\}$, let $p = \Pr(Y_k = 1 \mid X = x)$. Among decisions based only on this effective local input, the minimum conditional zero–one prediction error is $\min(p, 1 - p)$. Equal effective observations under the same conditional population have the same Bayes-optimal risk, irrespective of how their outcome labels were obtained.*

Proof. A deterministic positive decision has conditional error $1 - p$ and a negative decision has error p. A randomized decision is a convex combination of those errors and cannot improve their minimum. The conditional event probability depends on the information and population being conditioned on, not on the name of the predictor that estimates it. □

Here X includes the agent identifier, accessible metadata, and history actually supplied to the guard. Identical text alone is not identical effective information if agent identities encode different priors or permissions. Same-view experiments therefore match these conditions or report them separately. Delayed labels can improve estimation of a conditional probability; they cannot make a single ambiguous observation fully informative. Additional authorized review may change the available information, and replay evaluates its cost and consequence. Neither this bound nor parameter sharing implies a one-to-one mapping between agents and abilities.

C.3 COST-TO-GO AND THE PERFORMANCE-DIFFERENCE IDENTITY

Augment s_t with time and all history needed to specify the environment's transition law. The law includes frozen task actors, the fixed review procedure, scheduling, and tool dynamics; a protection policy changes only its allowed safety decisions. Compare two policies with the same initial root distribution, environment law, action semantics, and loss weights. Charge terminal utility and any declared timeout loss before padding completed trajectories to $T_{\max}$ with zero-cost absorbing transitions. An unresolved infrastructure failure is not silently converted to safe completion.

Write c_t for the weighted loss increment, so $J(\Pi) = \mathbb{E}_\Pi \sum_{t=0}^{T_{\max}-1} c_t$. With the time index explicit, define

$$V_t^\Pi(s) = \sum_{u \in \mathcal{U}_i(x(s))} \pi_i(u \mid x(s)) Q_t^\Pi(s,u), \qquad A_t^\Pi(s,u) = V_t^\Pi(s) - Q_t^\Pi(s,u). \tag{14}$$

The action distribution includes the deployed predictor, calibration, mask, and decision rule. A deterministic rule induces a point mass. At a transition with no safety choice, the advantage of its unique action is zero. Although s is centralized, every policy being compared still acts through its permitted local observation.

Proposition C.3 (Finite-horizon performance difference)**.** *If the complete losses are integrable, then*

$$J(\Pi') - J(\Pi) = -\sum_{t=0}^{T_{\max}-1} \mathbb{E}_{s \sim d_t^{\Pi'},\, u \sim \pi'_{i(t)}(\cdot \mid x(s))} \big[A_t^\Pi(s,u) \big], \tag{15}$$

where $d_t^{\Pi'}$ is the state distribution generated by the new policy.

Proof. The Bellman equation gives

$$Q_t^\Pi(s,u) - V_t^\Pi(s) = \mathbb{E}\big[c_t + V_{t+1}^\Pi(s_{t+1}) - V_t^\Pi(s_t) \mid s_t = s, u_t = u \big].$$

Take expectations along trajectories of Π' and sum over time. The value terms telescope. The terminal value is zero after terminal losses are charged, and the common initial distribution gives $\mathbb{E}[V_0^\Pi(s_0)] = J(\Pi)$. Substituting $Q - V = -A$ yields Equation (15). □

This identity is the cost-minimization form underlying cost-to-go policy learning (Ross & Bagnell, 2014). Equation (6) uses old-policy states as a trainable surrogate for its right-hand side. Equal local observations may correspond to several complete states with different action values; the local policy learns their average action preference in its training population. Full-state replay supplies more accurate supervision without granting the execution policy extra information.

Corollary C.4 (Replay error and distribution change)**.** *Assume $|A_t^\Pi(s,u)| \le A_{\max}$ and a uniform bound $|\widehat{A}_t^\Pi - A_t^\Pi| \le \varepsilon_A$ on relevant states and actions of both policies. Let*

$$\widehat{G} = \sum_t \mathbb{E}_{s \sim d_t^\Pi,\, u \sim \pi'_{i(t)}} [\widehat{A}_t^\Pi(s,u)].$$

Then, with $\mathrm{TV}(P,Q) = \sup_E |P(E) - Q(E)|$,

$$J(\Pi') - J(\Pi) \le -\widehat{G} + T_{\max}\varepsilon_A + 2A_{\max} \sum_t \mathrm{TV}(d_t^{\Pi'}, d_t^\Pi). \tag{16}$$

Proof. For $f_t(s) = \sum_u \pi'_{i(t)}(u \mid x(s)) A_t^\Pi(s,u)$, we have $|f_t| \le A_{\max}$. Replacing its expectation under $d_t^{\Pi'}$ by one under d_t^Π changes each term by at most $2A_{\max}\,\mathrm{TV}(d_t^{\Pi'}, d_t^\Pi)$. Replacing A by $\widehat{A}$ under the old distribution changes each term by at most ε_A. Apply these inequalities to Equation (15). □

When all Q estimates at a state have error at most ε_Q and the baseline action probabilities are fixed, Equation (5) has error at most $2\varepsilon_Q$. Independent Monte Carlo continuation reduces estimation variance on evaluated checkpoints. It does not establish a uniform bound at unevaluated states. Similarly, a learned state-discrimination score is a diagnostic, not a certified total-variation bound. Equation (16) separates the two issues without assuming either is small in the experiment. Complete candidate execution is the operational check; after acceptance, the next round samples the newly deployed distribution.

C.4 Acceptance conditions and their scope

After all training and bias selection, fix a candidate Π^{cand}. On fresh acceptance roots, run it and the old policy with paired environment draws when valid. Let $\Delta J = J(\Pi^{(m)}) - J(\Pi^{\mathrm{cand}})$. The candidate is deployable only if a valid one-sided lower bound $\mathrm{LCB}(\Delta J)$ is positive, a valid upper bound $\mathrm{UCB}(\mathrm{FBR}_{\mathrm{benign}})$ is at most ϵ, and predeclared resource tests pass. Component outcomes remain visible even when their weighted sum improves.

The inference unit is an independent root task or generation-template cluster, not a prefix, counterfactual branch, or random seed. Several seeds of one root are summarized within that root. Family-generalization intervals use the declared family-level sampling estimand. Confidence procedures, levels, bounded-loss assumptions if used, and minimum acceptance sample sizes are fixed before opening the acceptance set. In particular, an observed zero false-block count does not justify a zero upper confidence bound. For independent binary root outcomes, an exact one-sided binomial bound is an available choice; clustered or averaged outcomes require a valid cluster-aware or bounded-root procedure instead.

If the lower and upper bounds simultaneously cover their targets, an accepted candidate has lower expected cost and satisfies the FBR constraint on the stated acceptance distribution. If their respective noncoverage probabilities are at most $\alpha_{J,m}$ and $\alpha_{F,m}$, the union bound limits a mistaken acceptance due to either bound by their sum, under the sampling assumptions. This is a conditional statistical statement, not a claim that a particular experiment already meets those conditions.

Training, risk calibration, action selection, acceptance, and final testing use disjoint root lineages. A failed candidate retains the old version. Adaptive retries do not repeatedly query the same acceptance outcomes: use fresh preallocated batches or a valid sequential procedure with a prespecified error budget across candidates and rounds. The final held-out test is never used for acceptance. Empirical confidence about the current evaluation population does not transfer automatically to new adversaries, shifted task mixtures, inaccurate verifiers, or unreplayable environments; Appendix G evaluates these limitations.

C.5 Fixed-path interpretation as a special case

For an auxiliary scenario in which a single fixed harmful path can only be terminated by an absorbing effective stop, let T_{stop} be its first stop time. Conditional on a reference scenario z, define $q_t(z) = \Pr(T_{\mathrm{stop}} = t \mid T_{\mathrm{stop}} \geq t, z)$. The probability chain rule yields

$$\Pr(T_{\mathrm{stop}} > m \mid z) = \prod_{t=1}^{m} [1 - q_t(z)]. \tag{17}$$

Each factor is already conditional on all earlier stops not occurring, so this identity needs no independence across agents. Its assumptions do require a fixed path and absorbing termination. Review, replanning, and parallel branches can change that path; the general method therefore uses the actual trajectory loss of Equation (4). Neither a marginal replay advantage nor a hazard of the violating proposal is the conditional effective-stop probability q_t.

D Implementation and Versioned Update Algorithms

D.1 Minimal predictor and local action policy

The task actors and shared text encoder E are frozen. The encoder consumes the serialized local observation, including authorized structured metadata. A trainable shared two-layer projection produces $z = g_{\theta_s}(E(x)) \in \mathbb{R}^d$. With residual rank q, the uncalibrated hazard logits for agent i are

$$a_i(x) = Wz + b + B_i A_i z + b_i, \qquad \lambda_i(x) = \mathrm{sigmoid}(a_i(x)), \tag{18}$$

where $W \in \mathbb{R}^{K \times d}$, $b, b_i \in \mathbb{R}^K$, $A_i \in \mathbb{R}^{q \times d}$, and $B_i \in \mathbb{R}^{K \times q}$. Ordinary weight decay regularizes the local residuals. No discrete ability labels, expert routing, transfer graph, or task-actor updates are required. A shared head receives gradients from different agents; a local residual receives only records associated with that agent. Multiple ability directions can coexist in either component.

The action network $h_{\phi_i}(z, r_t)$ is a small classifier over the three action types. Invalid actions receive a mask value of $-\infty$. For action bias $\beta_{i,u}$, the categorical policy is

$$\pi_{\phi_i,\beta_i}(u \mid x, r_t) = \frac{\exp(h_{\phi_i}(z, r_t)_u + \beta_{i,u})}{\sum_{v \in \mathcal{U}_i(x)} \exp(h_{\phi_i}(z, r_t)_v + \beta_{i,v})}, \qquad u \in \mathcal{U}_i(x). \tag{19}$$

The continue bias is fixed to zero for identifiability; review and block biases are selected jointly through complete-workflow evaluation. A categorical implementation samples this distribution online and in validation. A deterministic argmax variant must be declared explicitly: the deployment baseline is then the induced point mass, while softmax training is a differentiable surrogate. Its complete validation and performance-difference evaluation use the actual deterministic rule. A probabilistic training distribution is never silently substituted for a deployed deterministic baseline.

Once the candidate risk model is fitted and calibrated, E, g_{θ_s}, all hazard heads, and calibration temperatures are frozen for policy learning. Only action-network parameters are optimized in Equation (10); biases are subsequently selected on the action-selection split. The old complete version stays frozen throughout both stages. Local observations can be encoded during centralized training, but neither the action network nor its risk features receive full replay states. The minimal implementation stores sampled returns rather than learning an online critic.

D.2 Risk calibration and joint action-bias selection

On a separate calibration split with reference outcomes, fit a positive temperature $\mathcal{T}_i$ by minimizing the same weighted survival likelihood after replacing $a_i(x)$ by $a_i(x)/\mathcal{T}_i$ (Guo et al., 2017). Positivity can be enforced through a log-temperature parameter. Agents below a prespecified effective-sample threshold use a shared temperature. The threshold is chosen on development data, and calibration weights match the stated prefix population. Recompute cumulative risks from the calibrated hazards; this preserves horizon monotonicity. Report per-horizon reliability, Brier score, logarithmic loss, and the calibration sample coverage, including sparse agent–horizon cells.

For action selection, fix the candidate predictor and action weights. Predeclare a finite grid for each review and block bias, a coordinate order, a number of sweeps, and a total full-workflow evaluation budget. At each coordinate, hold all other biases fixed, execute each proposed setting on the selection roots, and compute complete J, benign FBR, and resource usage. Retain the feasible setting with the lowest selection loss, using a fixed tie rule. Keep the best feasible complete vector found across the search; if none is feasible, no candidate advances. The current deployed version is retained as the fallback comparator, not modified by this search.

Only one finalized predictor–policy–bias bundle proceeds to each acceptance batch. Temperatures cannot be tuned on the acceptance results, and independently chosen per-node FPR thresholds cannot replace workflow-level FBR selection. Candidate search, failed candidates, and acceptance executions all consume recorded lifecycle resources. The acceptance protocol is given in Appendix C.4.

D.3 Complete update and online execution

Algorithm D.1: Versioned AgentVax update

Inputs: fixed reference version μ; deployed bundle $\Pi^{(m)}$; versioned root buffer and released event queue; $K, T_{\max}, B, M$; declared sampling weights, losses, constraints, split manifest, and budgets.
Output: either a validated new complete bundle or the unchanged $\Pi^{(m)}$, with an audit record.

1. **Freeze the baseline and source feedback.** Snapshot the old encoder/guard versions, temperatures, masks, biases, action rule, actor configurations, and environment specification. Admit only parent feedback released by the round's collection cutoff. Resolve event identities, occurrence times, proposal/realized labels, and root lineages.
2. **Construct reference records.** Sample roots and grid locations by Appendix B.3; retrieve actual historical x_t and s_t. For each eligible event-free location, use a compatible reference suffix or restore and execute μ. Verify the outcome, derive (m_t, δ_t), and store its release time and sampling weight. Missing or zero-follow-up records are retained for coverage accounting but supply no survival loss. Freeze the risk dataset only after its admitted labels are available.

3. **Fit and calibrate the candidate predictor.** Mix new, historical, and benign root records according to the declared weights. Optimize Equation (2) on the training split. Fit temperatures on the separate calibration split. Freeze the complete candidate feature map, hazard heads, and temperatures.
4. **Collect old-policy control traces.** Run training roots under the unchanged $\Pi^{(m)}$ and save complete checkpoints at every available control point, including points after a proposal. No candidate action is used in these baseline rollouts.
5. **Estimate action values.** For each trace n, sample $\mathcal{B}_n$ and record $p_{n,t}$. For each $t \in \mathcal{B}_n$ and each $u \in \mathcal{U}_i(x_{n,t})$, restore $s_{n,t}$ for each of M replicates, execute u, and regenerate the remaining workflow under $\Pi^{(m)}$. Record the component returns and use Equation (11). Form advantages using the old actual action distribution, as in Equation (5). Record restoration failures rather than filling them with zero loss.
6. **Fit local action policies.** Compute candidate r_t from each unchanged local input. Optimize the weighted cost-sensitive objective in Equation (10), with the predictor and all old-version quantities held fixed. Stored action targets are not recomputed using the candidate downstream policy during this round.
7. **Select a complete candidate.** Jointly select action biases on the selection split using Appendix D.2. Bind predictor, temperatures, action weights, action rule, masks, and selected biases into one immutable candidate version.
8. **Validate on fresh complete workflows.** Pair old and candidate executions on the reserved acceptance roots. Apply the cost-reduction, FBR, resource, and coverage criteria specified before viewing these results. Final test roots remain inaccessible.
9. **Accept or retain.** Deploy the candidate atomically only when the acceptance conditions pass; otherwise continue with the old complete bundle. Log the decision, uncertainty, failed conditions, and all costs. Do not install a new predictor with an unvalidated old policy.
10. **Update experience storage.** Add fully logged new roots to the appropriate buffer strata. Use fixed-capacity reservoir sampling at root level to retain history, preserving each root's prefixes, versions, labels, and derived continuations together. Newly released labels wait for the next eligible update. Following acceptance, subsequent control traces naturally refresh to the new deployed state distribution.

Root strata, mixture weights, and reservoir capacities are configuration fields. Reservoir sampling within a declared stratum gives each eligible historical root the same retention probability; nonuniform mixtures require corresponding population weights. Replay-derived records never become independent roots. A held-out family, template, or object lineage cannot enter a buffer through a derived continuation. Rollback retains both model parameters and all execution/calibration configuration associated with the previous bundle.

Algorithm D.2: Local online protection

Inputs: a validated deployed bundle and a pending workflow operation.

1. Before the safety decision, serialize and hash the actual local observation, accessible object versions, operation identity, and allowed-action mask. Log the full restorable checkpoint centrally under the organization's access controls.
2. Compute the deployed feature vector, calibrated hazards, and cumulative risk vector. Select an allowed action using the deployed action rule and biases. Global replay state and later verifier output are not policy inputs.
3. Execute continue, the fixed one-round review, or block through the runtime. Record decision time, effective transition time, authorized evidence consulted during review, costs, dependent cancellations, and any realized state changes.
4. Resume the workflow according to its dependency and scheduling rules. Send potential safety events for verification; only released labels enter a later update. Version switches occur at a declared atomic boundary and are logged.

A snapshot after a repair is a new state with its own input and reference continuation. Repairing a shared object does not overwrite pre-repair training examples. A trace that already contains the target

proposal is excluded from strict first-proposal forecasting at that point, while its remaining harm and utility still matter to Algorithm D.1's action learning.

D.4 Compute, storage, and replay coverage

For local input length L, let $C_E(L)$ be frozen-encoder inference cost and C_g projection cost. With projected dimension d and residual rank q, one decision uses

$$O(C_E(L) + C_g + Kd + qd + Kq + C_{\text{policy}}(d, K)) ,$$

where C_{policy} is the declared small action network's cost. The shared hazard head costs $O(Kd)$ and the residual costs $O(qd + Kq)$. Encoder reuse is reported only when the runtime actually has a compatible cached representation for the same input hash and encoder version. Review overhead is additional and must be measured end to end.

For maximum action-set size $|\mathcal{U}| \leq 3$, a root's action-value computation requires at most $B|\mathcal{U}|M$ suffix executions, with expected cost $O(B|\mathcal{U}|MC_{\text{cont}})$ for average suffix cost C_{cont}. Reference labeling, training, calibration, coordinate search, acceptance, and failed restores are separate costs. Report raw tokens, calls, latency, hardware time, checkpoint bytes, and stored guard parameters before any monetary or amortized conversion. Amortized lifecycle cost uses the explicitly declared deployment volume, reported together with the non-amortized background total.

Successful-restore-only training targets the covered state population unless an additional missingness model identifies the full objective. Coverage is therefore reported over requested roots, checkpoints, and action branches, with reasons for missingness. Full-policy tests should include the broader supported runtime population and identify any excluded environment classes. Counterfactual-effect claims are limited to states whose restoration and outcome measurement pass the audits. This boundary is particularly important for external systems without versioned, isolatable state.

E Evaluation Design, Coverage, and Score Provenance

E.1 Sampling, outcomes, and uncertainty

All main results are produced by executed, isolated workflow runs; published results are reference-only and are never used for fitting or calibration. The 480 source roots comprise twenty roots per family across 24 families; each of seven external strata has 192 roots in twelve difficulty groups. The resulting 1,824 evaluation roots each have attack/authorized counterparts. Thirteen actors, nine methods, and three worlds yield 936 configuration cells and 1,280,448 workflow realizations, not that many independent tasks. Separate source roots support calibration (192 authorized roots, families 16–19), feedback (640 roots, families 0–15), and held-out-family evaluation (eighty roots, families 20–23). This is one split, not outer cross-validation.

Per-platform summaries average actors, roots, and worlds. The 256 hierarchical bootstrap resamples jointly sample three whole worlds, family/stratum groups, and roots within groups, sharing weights across methods, actors, and paired variants. Intervals are 95% percentile bootstrap intervals; changes and timely-control ratios are resampled jointly. Cost and p95 are point summaries; closed-loop bands condition on the recorded feedback path. The observation-study intervals in Appendix A.1 use the resampling procedure defined for that study.

The prediction endpoint is the first prohibited proposal T_H^μ under record-only reference μ; a later transition determines execution. Local x_t contains current content, goals, revealed authorization/provenance, permissions, and task-propensity metadata available at the checkpoint, but not hidden future outcomes. Complete s_t also stores unobserved evidence and event/confirmation schedules. The recorder distinguishes proposal, execution, last preventable time, warning, effective control, and T_D^μ; missed or late controls remain in coverage denominators, while undefined lead times are excluded from lead-time averages. Risk labels follow μ; action values follow the frozen deployed Π. Restored continuations recompute downstream outcomes after intervention from the corresponding workflow state.

All methods share actor draws, initial permissions, calibration roots, and a fixed authorized reviewer. Central pools current-prefix evidence and pays its communication delay; local block cancels only

a dependent branch. Workflow false blocking includes subsequently recovered authorized runs. Thus matching calibration targets does not match test utility or compute. Bounds, timestamps, proposal/execution ordering, paired restoration, and cost sums are checked for internal consistency before aggregation.

E.2 Randomization and execution controls

Keyed SHA-256 streams use seed 20260925 to fix root assignment, paired environment randomness, continuation pairing, and resampling. Content, authority, goals, sensitive-operation negatives, four control positions, permissions, and deadlines determine the observations and available actions presented at each checkpoint. Family IDs are used only as analysis strata and are never predictor inputs. Shared random streams are reused across paired interventions when the resulting branches remain compatible; branch-specific tool calls and state changes are executed from their own restored state. No seed search is used to select a preferred ranking.

Training, calibration, replay, action selection, and acceptance are executed as separate stages with disjoint root lineages where specified. Confirmed exposure changes the fitted predictor and local action policy through the update procedure in Algorithm D.1; action-value estimates use the replay budget M stated for each experiment. Workflow length, retries, review calls, and service time are recorded from the executed workflow traces and included in the lifecycle ledger.

E.3 Configuration coverage and distinct scoring protocols

P01 is the source; P02–P08 are AgentDojo, AgentHarm, Agent-SafetyBench, AgentDyn, AgentLAB, TAMAS, and DUMA-Bench strata (Debenedetti et al., 2024; Andriushchenko et al., 2025; Zhang et al., 2024; Li et al., 2026; Jiang et al., 2026; Kavathekar et al., 2026; Aleksandrov et al., 2026). These strata instantiate benchmark-derived workflow conditions under the common AgentVax endpoint and accounting protocol. AgentDyn inherits AgentDojo infrastructure; DUMA-Bench is dual-control rather than simply cooperative. Table 6 lists all actor configurations; M03/M05/M08/M10 form the prespecified mechanism subset. The risk encoder, reviewer, and judge are distinct from these task actors.

Table 6: Actor configurations used in the evaluation. M03/M05/M08/M10 form the prespecified mechanism subset; the risk encoder, reviewer, and judge are evaluated separately from the task actors.

ID	Actor label	ID	Actor label
M01	Qwen3-8B	M08	GPT-6 Sol
M02	Qwen3.6-35B-A3B	M09	GPT-6 Luna
M03	Qwen3.8-27B	M10	Claude Opus 5.5
M04	gpt-oss-20b	M11	Claude Sonnet 5
M05	gpt-oss-120b	M12	Gemini 3.8 Flash
M06	DeepSeek-V4.1-Flash	M13	Gemini 3.7 Flash
M07	GPT-6 Astra		

Native-style endpoints remain separate from both the common endpoint and published measurements (Table 7). AgentHarm's partial score is not binary ASR; the corresponding endpoint aggregates four predefined subcriteria. Source-held-out families use disjoint roots and are reported separately from cross-platform results. Likewise, 47 lower-harm cells against Central are descriptive, not a statistical win rate or a matched-utility causal effect. The complete 936-cell matrix, including intervals and costs, is retained in Appendix H.

Published AgentHarm entries use GPT-4o at different sampling budgets. In AGrail, GPT-4o names the defense configuration, not necessarily the task actor; Safe-OS accuracy and PI ASR are not general workflow safety. DynaGuard timings are mean H100/vLLM module-example times averaged over two datasets, not workflow p95 or AgentVax latency (Hoover et al., 2026). JANUS names its resulting guard Vanguard; no JANUS/Vanguard scores enter the matrix (Xiong et al., 2026).

Table 7: Scores by provenance. (a) Native-style evaluation endpoints (%); endpoints are not pooled. (b) Reference-only original-paper values: AgentHarm, ICLR 2025 p. 9 Table 2; AGrail, ACL 2025 p. 8110 Table 1; DynaGuard, ICLR 2026 p. 38 Table 22. No intervals are imputed; neither panel is a shared-protocol ranking.

(a) Native-style evaluation endpoints

Stratum	Native-style proxy	Base	Central	AgentVax
Source	Realized violation	53.7	9.1	9.5
Dojo	Attack success proxy	51.8	2.3	2.9
Harm	Partial harm score	61.4	17.1	17.8
SafetyBench	Noisy violation judgment	48.5	8.8	6.7
Dyn	Attack success proxy	56.8	8.8	8.1
LAB	Attack success proxy	59.7	6.6	6.3
TAMAS	Attack success proxy	55.4	11.9	11.7
DUMA	Attack success proxy	52.2	8.8	11.6

(b) Published reference results

Source configuration	Original metric / budget	Reported	Unit
GPT-4o (Andriushchenko et al., 2025)	harm score; greedy	72.7	%
GPT-4o (Andriushchenko et al., 2025)	harm score; best-of-5	83.7	%
GPT-4o (0-Shot) (Luo et al., 2025)	Safe-OS Norm accuracy	52.4	%
AGrail (GPT-4o) (Luo et al., 2025)	Safe-OS Norm accuracy	95.6	%
GPT-4o (0-Shot) (Luo et al., 2025)	Safe-OS PI ASR	61.9	%
AGrail (GPT-4o) (Luo et al., 2025)	Safe-OS PI ASR	0.0	%
DynaGuard-8B non-CoT (Hoover et al., 2026)	mean time per example	17.76	ms
DynaGuard-8B CoT (Hoover et al., 2026)	mean time per example	29.60	ms

F CONTROL, COST, AND MECHANISM DIAGNOSTICS

F.1 EVIDENCE, AUTHORITY, AND INFORMATION BOUNDARIES

Paired source/TAMAS controls retain roots, actor transitions, permissions, and fixed exposure. Role renaming and sham removal leave consumed features unchanged; relevant removal hides provenance/authorization, and restoration reinstates it exactly. View swaps permute four control positions while preserving the evidence union and physical permissions. The stale-version arm hides the same feature: it is a visibility proxy, not a file-hash or real snapshot test. Exact placebo/restoration zeros arise from deterministic negative-control construction. These are deployment-dependence tests, not demonstrations of how specialization was learned.

Crossing evidence access with block effectiveness separates warning from control (Figure 7). Zero permission disables intervention without changing actor safety; nonzero permission and the fixed reviewer can recover some missing information. A separate oracle boundary stipulates free, perfectly safe global cancellation, under which both policies attain zero harm/cost; this is not an AgentVax action. In 512 byte-identical local inputs with a hidden Bernoulli future event, AUROC changes from 0.500 before a noisy authorized observation to 0.902 after it. Shared labels cannot remove uncertainty absent from x_t.

Figure 9a,b varies local provenance access and central delay (0, 0.5, 1, 2, 4, 8 events) under the same physical deadlines. Review may compensate for missing local evidence; central information may arrive too late. Neither result implies optimal specialization or universal local dominance, and event delay is not measured network latency. Disabling one safety node leaves its task actor active (Table 10); contributions depend on timing, authority, and redundancy rather than local accuracy alone. Protection-position records retain missing mass: better warning need not move effective control upstream.

F.2 INDEPENDENT ACTION VALUES AND LIFECYCLE BUDGETS

The continuation test uses 240 independent states and three actions. For risk p, authority a, and utility sensitivity b, harm probabilities are p, $p(1-0.78a)$, and $p(1-a)$; utility-loss probabilities are 0, $0.10b$, and $0.80b$; deterministic costs are 0.02, 0.18, and 0.04. Harm, utility loss, and cost define the toy Q^{Π}, not the main workflow loss. Eight candidate samples select the low-budget action, an independent 256-sample stream selects the reference, and a third 2,048-sample stream evaluates both. Signed regret is selected loss minus reference loss, without clipping. Risk ranking yields 0.150 versus 0.043 for replay; this does not itself establish the workflow advantage or real restoration fidelity. The credit-assignment motivation differs from a reproduction of C3 (Chen et al., 2026).

Figure 8a,b shows amortization and budgets $M \in \{1, 2, 4, 8, 16, 32, 128\}$ with untouched evaluation randomness. Recorded samples support finite-budget diagnostics; Hoeffding radii can be loose and do not guarantee uniform LLM error. The lifecycle ledger is

$$C_{\text{life}}(N) = \overline{C}_{\text{online}} + \frac{C_{\text{setup}} + C_{\text{feedback}} + C_{\mu} + C_{\text{replay}} + C_{\text{opt}} + C_{\text{select}} + C_{\text{accept}} + C_{\text{retry}}}{N}.$$

The matrix fixes $N = 10{,}000$ and attack/authorized weights 0.25/0.75. Reference/action continuations cost 0.55/0.30 of source base-work; AgentVax models $640 \times 2 \times 3 \times 8$ action continuations ($B = 2$, three actions, $M = 8$). Selection and acceptance are separate. Table 8 preserves every online and one-time component; one source-training charge applies to each standalone target deployment, not eight charges for a jointly deployed model. Service-time p95 is computed after mixing individual workflows, never by summing or averaging component p95s. These are modeled work/time, not bills, FLOPs, or hardware measurements.

Equal feedback gives Memory/Binary/Central/AgentVax 640 confirmations. At equal modeled background work on source/M08, Binary can afford 17498.0 versus 640.0 confirmations, with harm 7.4% versus 9.4%. This assumes unlimited extra labels for the simpler method, not their availability; supply limits and real compute matching require separate evaluation.

Table 8: Complete source/M08 ledger. Online entries are work per mixed workflow; background entries are one-time modeled totals, divided by deployment volume before addition.

(a) Online work / workflow

Component	Binary	Central	AgentVax
Task	14.208	14.142	14.733
Tools	2.273	2.263	2.357
Encoding	0.540	0.538	0.560
Policy	0.021	0.021	0.022
Review	0.608	0.554	0.382
Communication	0.000	0.453	0.000
Retry	0.057	0.057	0.058

(b) One-time background work

Component	Binary	Central	AgentVax
Setup	12.0	12.0	12.0
Confirmation	9.6	9.6	9.6
Reference	6644.5	6644.5	6644.5
Replay	0.0	0.0	173966.4
Optimization	3.6	3.6	3.6
Selection	10872.9	10872.9	10872.9
Acceptance	3624.3	3624.3	3624.3
Retry	132.9	132.9	3032.3

F.3 Ablations, timing, and policy distributions

Table 9 crosses early-prefix broadcast supervision (B), replay scoring (R), and the joint loss objective (J), with identical roots/exposure and a fixed full-system threshold per actor. Utility/FBR are observed rather than recalibrated, so lower harm need not imply a matched-utility gain. Old-risk substitution uses zero exposure; shuffled values break state–value correspondence; restoration reinstates the original signal. Exact recovery is constructed. Removing a local residual changes noise, not a capacity-matched trained encoder comparison. Same-view nodes need not specialize.

Table 9: B/R/J factorial, averaged over M03/M05/M08/M10. B: prefix supervision; R: replay scoring; J: joint objective. H/U/F are harm, utility, and false blocking (%). The fixed threshold permits different operating points. Joint is a cost-sensitive score/review rule, not a trained joint optimizer.

	Source (P01)					TAMAS (P07)				
B/R/J	H	U	F	Brier	Loss	H	U	F	Brier	Loss
000	5.0	78.0	19.9	0.137	0.768	6.6	78.5	18.9	0.135	0.848
001	5.1	78.0	19.8	0.137	0.769	6.8	78.5	18.9	0.135	0.850
010	7.6	85.2	8.1	0.137	0.707	9.7	85.5	6.8	0.135	0.803
011	9.6	85.9	7.3	0.137	0.741	12.2	86.8	5.5	0.135	0.831
100	4.8	79.4	17.2	0.136	0.737	6.4	80.3	15.8	0.134	0.806
101	4.9	79.4	17.1	0.136	0.738	6.4	80.6	15.7	0.134	0.800
110	7.2	85.9	7.0	0.136	0.681	9.0	87.1	4.6	0.134	0.757
111	9.4	86.7	5.9	0.136	0.717	11.4	88.0	3.9	0.134	0.796

The main gate uses endpoint risk. With the same endpoint and other components, binary and survival heads therefore tie in workflow outcomes (Table 10); temporal modeling is not necessary for this gate. A separate 768-root test uses eight event-time bins, matched endpoint probability, and administrative censoring between bins four and eight. Binary spreads time mass uniformly, multi-horizon merges adjacent bins, and survival preserves its specified shape. The observed-follow-up loss is

$$-\delta \log\left(h_j \prod_{k<j}(1-h_k)\right) - (1-\delta)\log\left(\prod_{k\le c}(1-h_k)\right).$$

The NLL change 1.295 to 1.140 is a timing result (Figure 8c); endpoint Brier ties. Integrated Brier uses fully adjudicated follow-up over the evaluation window. No-hard-negative, uncalibrated, review-off, reviewer-quality, and slow-review arms isolate fixed-feature, operating-point, and deadline assumptions, not a guarantee that prediction improvements are actionable.

For a separate three-state, two-action discounted MDP, exact values satisfy

$$J(\Pi') - J(\Pi) = -\frac{1}{1-\gamma}\mathbb{E}_{s\sim d^{\Pi'},\, u\sim\Pi'}[A^{\Pi}(s,u)]$$

at six interpolation steps with $\gamma = 0.90$, using $A = V - Q$. The reported residual is numerical round-off. Old occupancy d^{Π} need not reproduce the difference (Figure 8d). This does not replace the main episodic objective or verify an LLM acceptance loop/uniform-error assumption.

F.4 CONTROLLED REPRESENTATIONS AND DECISION RELEVANCE

The representation study contains 1,200 roots across 24 families: 800 training, 200 selection, and 200 held-out. A sixteen-dimensional orthogonalized feature basis is extracted at fixed decision positions and organized into authority, an unused decodable variable, intent, and nuisance coordinates; the extraction specification is frozen before selection. Regularized logistic probes fitted on training features reach 0.802 event AUROC and 0.999 for the unused variable, without held-out-label tuning.

The controlled decision uses authority/intent, not the unused decodable variable. Donor matches are fixed by authority rank before outcome inspection. Authority replacement, unused-feature replacement, one magnitude-matched random direction, and exact restoration distinguish decodability from decision use (Figure 8e). Careful patching controls motivate this diagnostic (Zhang & Nanda, 2024); it does not identify a language-model circuit. The guard exposure curve (panel f) varies guard exposure while holding the actor-side representation features fixed; the intervention is interpreted as a decision-relevance diagnostic rather than a complete mediation claim.

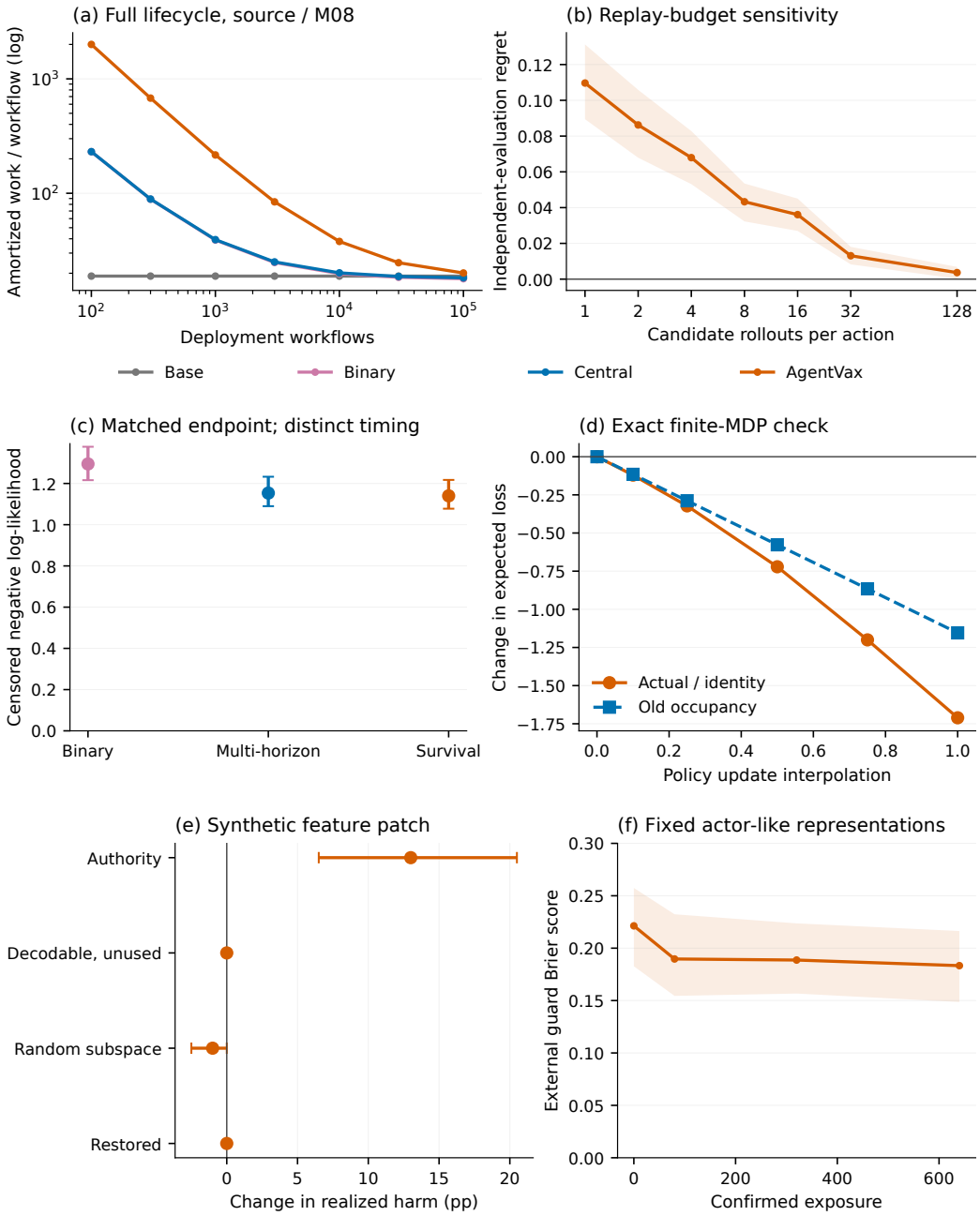

Figure 8: **Evaluation.** Decision and mechanism diagnostics (two panels per row). (a) Lifecycle amortization; work is not USD, FLOPs, or hardware time. (b) Independent action-law regret versus rollout budget. (c) Censored NLL at a matched endpoint on 768 roots. (d) Exact discounted-MDP loss change versus old-occupancy approximation. (e) Paired authority, unused-feature, random-subspace, and restoration interventions. (f) External-guard exposure with fixed actor-like features. Intervals resample each experiment's states/roots; the timing and feature interventions are diagnostic experiments rather than complete mechanism-identification claims.

G CONTINUAL FEEDBACK AND ROBUSTNESS

G.1 STREAMS, CLEANUP, AND SELECTIVE LABELS

The common stream releases 640 training roots in eight batches of eighty, then evaluates 64 disjoint test roots per checkpoint using only labels with $T_D \leq j$. Thirteen source actors plus four representatives on each of AgentDyn/TAMAS give 21 stream configurations per method; 192 reused authorized roots per platform are not new independent evidence at every checkpoint. The endogenous loop instead uses twelve batches of 64 attack roots, actual harm plus independent 25% reference audits, and confirmation delays of one to three batches. Unreleased events remain queued. Prevented branches are not labeled safe counterfactuals; audits retain μ labels, and exposure is recorded before evaluation.

Timely control uses all reference-realized-harm runs as its denominator: missed/late actions remain failures. An action after the proposal may still prevent execution and have negative lead time. Cleanup removes an old contaminated branch but not fresh roots; exposure updates affect both. Table 10 reports all eight cleanup-by-learning cells. Delay sensitivity uses 0, 1, 4, and 8 batches with label noise 0 or 0.15. The natural-only condition retains 46% of the confirmed feedback signal relative to reference auditing, and the noisy-label condition applies the stated symmetric corruption level.

G.2 ORDER, RETAINED UTILITY, AND CANDIDATE SEARCH

The same 640 labels arrive interleaved, blocked, or reversed. A recency-weighted history window over the last 160 labels controls rehearsal under each ordering. Symmetry can produce ties, which are retained (Table 10). With the defense frozen, candidate budgets 1, 4, and 16 share roots/randomness; later candidates follow a prespecified concealment schedule, and any successful candidate is retained. Increasing any-success rate (Figure 9c) follows this definition, not stronger single-shot attacks or co-evolving adversaries. Figure 9d shows the distinct delay/selection sensitivity.

Table 10: Controls: Brier for delay, benign success (%) for order, and realized harm (%) elsewhere. Panels use distinct protocols, not a common ranking.

(a) Delay and selective labels

Delay (batches)	0	1	4	8
Natural only	.138	.139	.141	.148
Reference audit	.137	.137	.138	.148

(b) Order and history rehearsal

Order	No rehearsal	Rehearsal
Interleaved	87.6	87.6
Blocked	86.7	87.6
Reversed	86.7	87.6

(c) Cleanup crossed with learning

Root	Cleanup	No learning	Learning
Fresh	No	11.9	9.4
Fresh	Yes	11.9	9.4
Old	No	11.9	9.4
Old	Yes	0.0	0.0

(d) Disable one safety node

Node	1	2	3	4
Harm	10.6	17.6	12.0	17.8

(e) Signals, heads, and equal work

Condition	Harm
No review	17.8
Uncalibrated	10.7
No hard negatives	12.8
No local residual	9.4
Stale-policy values	8.1
Shuffled action values	10.9
Old risk predictor	12.0
Matched binary head	9.4
Multi-horizon head	9.4
Equal work: Binary	7.4
Equal work: AgentVax	9.4

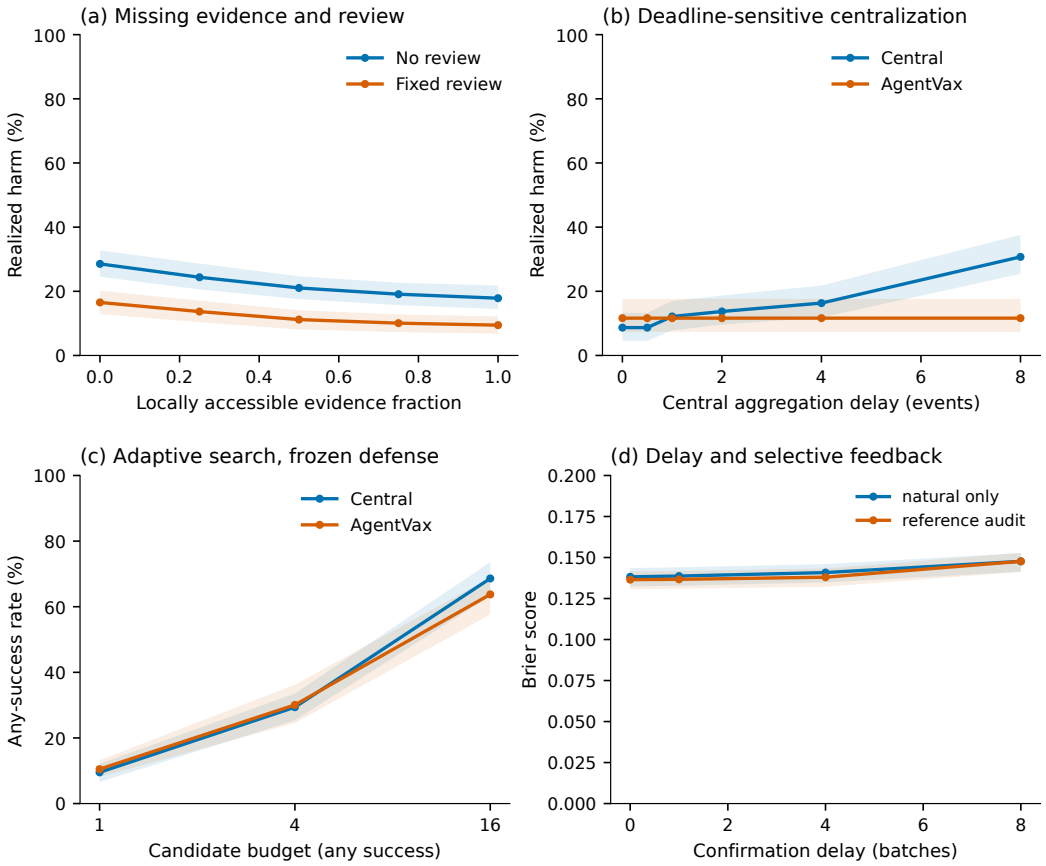

Figure 9: **Evaluation.** Information and feedback boundaries. (a) Local evidence and fixed review (source/M08). (b) Central aggregation delay (TAMAS/M08). (c) Frozen-defense any-success rate at each candidate budget, not single-shot ASR. (d) Confirmation delay and the stipulated natural-feedback penalty. Bands resample paired roots. These controls isolate information loss, confirmation delay, and candidate-search effects under the stated protocol.

H Complete Configuration Matrix

Table 11: P01 — Source. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	59.7 [54.9,64.4]	73.4 [69.1,77.7]	0.0	26.5	10.09
Rules	28.3 [24.0,32.5]	72.2 [68.3,76.9]	2.7	26.5	10.73
Cur.	17.4 [13.8,20.6]	70.8 [66.0,75.6]	6.7	27.3	10.96
Disc.	9.6 [6.9,12.6]	70.8 [66.2,75.7]	6.1	26.4	11.27
Mem.	9.2 [6.5,12.2]	70.6 [66.2,75.6]	6.4	26.5	11.24
Bin.	7.9 [5.3,10.9]	70.3 [65.0,75.1]	6.7	26.4	11.10
Cent.	9.2 [6.9,11.8]	70.6 [65.3,75.6]	6.5	30.8	11.45
Graph	16.9 [13.4,19.9]	70.7 [65.7,75.6]	7.0	27.2	11.46
AV	10.0 [7.4,13.3]	70.7 [66.4,75.1]	6.0	26.3	20.70
M02 — Qwen3.6-35B-A3B					
Base	55.5 [49.6,60.2]	85.8 [82.4,89.3]	0.0	35.4	16.09
Rules	26.5 [21.7,31.0]	84.3 [81.0,87.8]	2.7	35.4	16.73
Cur.	16.9 [13.1,20.4]	82.7 [78.8,86.6]	6.4	35.8	16.61
Disc.	8.9 [6.5,11.7]	82.8 [78.8,86.8]	6.4	35.1	17.22
Mem.	8.5 [6.0,11.4]	82.5 [78.4,86.4]	6.4	35.1	17.16
Bin.	7.6 [5.4,10.3]	82.5 [78.1,86.7]	6.4	35.1	17.04
Cent.	9.0 [6.6,11.6]	82.7 [78.8,86.6]	6.4	39.6	17.36
Graph	16.2 [12.3,19.6]	82.4 [78.3,86.6]	7.2	36.1	17.07
AV	9.5 [6.8,12.5]	82.7 [79.1,86.5]	5.8	35.1	32.44
M03 — Qwen3.8-27B					
Base	52.8 [46.9,58.1]	88.8 [86.0,91.9]	0.0	33.3	14.69
Rules	25.8 [21.8,30.0]	87.3 [84.1,90.6]	2.7	33.3	15.33
Cur.	16.4 [13.0,19.6]	85.3 [81.5,89.5]	6.7	33.9	15.29
Disc.	8.6 [6.1,11.4]	85.5 [82.1,89.2]	6.4	33.1	15.81
Mem.	8.3 [5.9,11.1]	85.2 [81.8,89.0]	6.5	33.1	15.79
Bin.	7.6 [5.4,10.0]	85.2 [81.5,89.0]	6.2	33.1	15.68
Cent.	9.5 [6.9,12.2]	85.3 [81.8,89.3]	6.2	37.6	16.02
Graph	16.2 [12.6,19.6]	85.2 [81.8,89.2]	7.0	34.1	15.77
AV	9.4 [6.5,12.1]	85.2 [81.7,89.1]	5.9	33.0	29.72
M04 — gpt-oss-20b					
Base	57.4 [52.0,62.2]	77.7 [73.8,81.8]	0.0	29.2	11.90
Rules	27.6 [23.1,31.8]	76.5 [72.2,81.0]	2.7	29.2	12.54
Cur.	16.9 [13.4,20.2]	74.8 [70.4,79.3]	6.7	29.9	12.66
Disc.	9.0 [6.5,12.0]	74.9 [71.0,79.3]	6.3	29.0	13.06
Mem.	8.6 [5.9,11.5]	74.7 [70.7,79.0]	6.4	29.1	13.04
Bin.	7.6 [5.1,10.6]	74.4 [69.9,78.9]	6.6	29.0	12.88
Cent.	9.0 [6.4,11.6]	74.8 [70.5,79.5]	6.4	33.4	13.25
Graph	16.3 [12.7,19.7]	74.7 [70.3,79.3]	7.2	29.9	13.15
AV	9.7 [7.0,13.0]	74.9 [71.1,79.3]	6.0	28.9	24.25
M05 — gpt-oss-120b					
Base	54.6 [48.6,59.0]	87.9 [84.8,91.1]	0.0	40.6	19.57
Rules	26.2 [22.1,30.4]	86.4 [83.3,89.7]	2.7	40.6	20.22
Cur.	16.8 [13.2,20.1]	84.8 [81.2,88.7]	6.4	41.1	19.90
Disc.	8.8 [6.5,11.8]	84.9 [81.3,88.9]	6.3	40.1	20.68
Mem.	8.5 [6.1,11.4]	84.5 [81.0,88.4]	6.4	40.1	20.62
Bin.	7.6 [5.4,10.4]	84.7 [80.7,88.7]	6.2	40.1	20.52
Cent.	9.4 [7.1,12.0]	84.7 [81.0,88.7]	6.4	44.6	20.81
Graph	16.2 [12.7,19.5]	84.4 [80.8,88.5]	7.2	41.2	20.34
AV	9.5 [6.7,12.2]	84.6 [81.1,88.7]	5.8	40.1	39.27
M06 — DeepSeek-V4.1-Flash					
Base	57.2 [51.2,62.2]	89.8 [87.0,92.8]	0.0	31.3	13.30
Rules	27.4 [23.4,31.8]	88.2 [84.9,91.6]	2.7	31.3	13.94
Cur.	17.2 [13.6,20.4]	86.6 [83.0,90.5]	6.5	31.9	13.97
Disc.	9.0 [6.5,11.9]	86.5 [83.2,90.0]	6.4	31.1	14.45
Mem.	8.6 [6.1,11.4]	86.2 [82.8,89.9]	6.5	31.1	14.41
Bin.	7.6 [5.4,10.4]	86.1 [82.3,89.8]	6.4	31.1	14.28
Cent.	9.0 [6.8,11.6]	86.4 [82.9,89.9]	6.4	35.5	14.62
Graph	16.5 [12.9,19.8]	86.2 [82.7,90.3]	7.2	31.9	14.45
AV	9.6 [7.0,12.5]	86.2 [82.8,89.9]	5.8	31.0	26.99
M07 — GPT-6 Astra					
Base	49.9 [43.9,54.9]	94.5 [92.2,97.0]	0.0	53.1	27.94
Rules	24.4 [20.2,28.5]	92.9 [90.3,95.8]	2.7	53.1	28.58
Cur.	15.8 [12.3,19.0]	90.6 [87.3,94.1]	7.2	53.3	27.78
Disc.	8.2 [5.7,10.9]	91.0 [87.8,94.5]	6.3	52.2	28.98
Mem.	7.8 [5.6,10.4]	90.8 [87.6,94.4]	6.5	52.4	28.91
Bin.	7.2 [4.8,9.9]	90.9 [87.5,94.6]	6.0	52.1	28.82
Cent.	9.0 [6.4,11.7]	91.0 [87.9,94.7]	6.0	56.7	29.11
Graph	15.2 [11.6,18.3]	90.5 [87.2,94.0]	7.6	53.5	28.21
AV	9.2 [6.3,12.0]	90.6 [87.2,94.6]	6.0	52.5	55.66
M08 — GPT-6 Sol					
Base	51.2 [45.1,56.3]	91.2 [88.9,94.0]	0.0	39.6	18.88
Rules	25.3 [21.4,29.4]	89.7 [86.8,92.7]	2.7	39.6	19.52
Cur.	16.3 [12.8,19.6]	87.8 [84.7,91.2]	6.8	40.0	19.24
Disc.	8.4 [6.0,11.2]	87.9 [85.0,91.1]	6.3	39.1	19.98
Mem.	8.1 [5.8,10.9]	87.6 [84.5,91.0]	6.5	39.1	19.93
Bin.	7.4 [5.2,9.8]	87.6 [84.6,91.0]	6.0	39.1	19.84
Cent.	9.1 [6.7,11.7]	87.8 [85.0,91.1]	6.2	43.6	20.16
Graph	16.0 [12.6,19.1]	87.4 [84.5,91.0]	7.3	40.2	19.69
AV	9.4 [6.8,12.2]	87.6 [84.7,91.1]	5.9	39.1	37.93
M09 — GPT-6 Luna					
Base	52.7 [46.8,57.6]	83.3 [79.7,86.9]	0.0	27.1	10.51
Rules	25.6 [21.0,29.6]	82.0 [78.3,86.0]	2.7	27.1	11.15
Cur.	16.4 [12.8,19.7]	80.2 [76.3,84.7]	6.6	27.9	11.34
Disc.	8.7 [6.1,11.6]	80.3 [76.3,84.5]	6.3	27.0	11.67
Mem.	8.3 [5.8,11.1]	80.0 [76.0,84.3]	6.4	27.0	11.64
Bin.	7.5 [5.2,10.3]	80.0 [75.7,84.2]	6.2	27.0	11.55
Cent.	9.4 [7.1,12.3]	80.1 [76.0,84.4]	6.3	31.4	11.90
Graph	16.2 [12.6,19.5]	80.1 [76.2,84.6]	6.9	27.9	11.84
AV	9.4 [6.6,12.2]	80.1 [76.5,84.1]	5.9	26.9	21.53
M10 — Claude Opus 5.5					
Base	47.5 [41.9,52.0]	93.4 [91.0,96.0]	0.0	56.2	30.04
Rules	23.5 [19.6,27.5]	91.8 [88.8,94.7]	2.7	56.2	30.68
Cur.	15.3 [12.0,18.5]	89.2 [86.1,92.8]	7.6	56.3	29.75
Disc.	8.1 [5.6,10.7]	89.9 [86.9,93.4]	6.2	55.2	31.09
Mem.	7.7 [5.6,10.2]	89.7 [86.7,93.4]	6.2	55.4	31.01
Bin.	7.1 [4.7,9.8]	89.8 [86.6,93.3]	6.0	55.2	30.90
Cent.	8.5 [6.2,11.2]	89.9 [86.7,93.5]	6.0	59.6	31.20
Graph	14.7 [11.4,18.0]	89.2 [86.0,92.6]	7.7	56.3	30.19
AV	9.1 [6.6,11.5]	89.6 [86.6,93.3]	5.9	55.6	59.79
M11 — Claude Sonnet 5					
Base	49.0 [42.8,54.0]	89.8 [87.0,92.8]	0.0	38.6	18.18
Rules	24.2 [20.1,28.4]	88.2 [84.9,91.6]	2.7	38.6	18.82
Cur.	15.8 [12.4,19.0]	86.0 [82.3,90.0]	7.2	39.0	18.58
Disc.	8.2 [5.7,10.8]	86.4 [83.0,89.9]	6.3	38.1	19.29
Mem.	7.8 [5.3,10.2]	86.1 [82.6,89.7]	6.4	38.1	19.24
Bin.	7.2 [4.6,9.6]	86.2 [82.7,90.0]	6.0	38.1	19.16
Cent.	8.9 [6.4,11.7]	86.3 [82.9,90.0]	6.0	42.7	19.49
Graph	15.1 [11.6,18.3]	85.8 [82.2,89.8]	7.6	39.2	19.04
AV	9.2 [6.2,11.7]	86.0 [82.6,89.9]	5.9	38.1	36.55
M12 — Gemini 3.8 Flash					
Base	54.7 [49.0,59.5]	90.4 [87.9,93.4]	0.0	30.2	12.60
Rules	26.2 [21.8,30.6]	88.8 [85.9,91.9]	2.7	30.2	13.24
Cur.	16.8 [13.2,20.1]	87.2 [83.6,91.0]	6.5	30.9	13.32
Disc.	8.9 [6.5,11.9]	87.1 [83.7,90.7]	6.3	30.1	13.75
Mem.	8.5 [6.2,11.2]	86.7 [83.3,90.3]	6.4	30.1	13.71
Bin.	7.7 [5.5,10.4]	86.9 [83.3,90.4]	6.1	30.1	13.63
Cent.	9.7 [7.4,12.5]	86.9 [83.7,90.4]	6.3	34.5	13.96
Graph	16.4 [12.9,19.8]	86.8 [83.3,90.7]	7.2	31.0	13.80
AV	9.6 [6.8,12.4]	86.7 [83.5,90.5]	5.9	30.0	25.63
M13 — Gemini 3.7 Flash					
Base	56.0 [49.7,60.7]	87.0 [83.4,90.4]	0.0	28.6	11.48
Rules	26.7 [22.2,31.0]	85.5 [82.1,89.1]	2.7	28.6	12.13
Cur.	16.8 [13.1,20.2]	83.9 [80.2,87.6]	6.4	29.2	12.26
Disc.	8.9 [6.5,11.7]	84.0 [80.3,87.6]	6.4	28.4	12.65
Mem.	8.5 [6.0,11.4]	83.6 [79.6,87.4]	6.4	28.5	12.61
Bin.	7.6 [5.4,10.3]	83.6 [79.2,87.6]	6.4	28.4	12.49
Cent.	8.9 [6.6,11.4]	83.8 [79.8,87.6]	6.4	32.8	12.84
Graph	16.2 [12.4,19.5]	83.6 [79.9,87.4]	7.2	29.4	12.75
AV	9.4 [6.7,12.3]	83.8 [80.0,87.7]	5.8	28.3	23.44

Table 11 (continued): P02 — AgentDojo stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	57.3 [49.5,64.6]	73.8 [65.6,79.7]	0.0	21.0	7.72
Rules	20.5 [12.9,27.9]	72.6 [64.9,78.6]	2.6	21.0	8.28
Cur.	10.8 [5.8,15.8]	70.8 [61.3,77.6]	6.6	21.2	8.38
Disc.	4.2 [1.9,6.8]	70.8 [62.3,77.2]	6.4	20.9	8.82
Mem.	3.6 [1.6,5.8]	71.2 [63.0,77.2]	6.4	20.8	8.80
Bin.	3.0 [0.9,5.2]	71.2 [63.0,77.2]	6.6	20.8	8.70
Cent.	2.6 [0.5,4.8]	70.8 [63.0,76.7]	6.4	22.3	8.88
Graph	10.1 [5.4,15.2]	70.8 [61.3,77.6]	6.6	21.6	8.71
AV	3.5 [0.8,7.3]	69.8 [61.9,76.0]	8.0	20.8	18.09
M02 — Qwen3.6-35B-A3B					
Base	53.1 [44.9,61.7]	86.3 [80.9,90.6]	0.0	28.0	12.31
Rules	18.1 [11.3,25.5]	84.9 [78.8,90.1]	2.6	28.0	12.87
Cur.	10.4 [5.4,15.8]	83.2 [75.0,88.8]	5.9	28.0	12.65
Disc.	3.1 [1.0,5.6]	82.5 [74.6,88.0]	6.6	27.7	13.42
Mem.	2.8 [1.0,5.0]	82.5 [75.0,87.6]	6.9	27.7	13.37
Bin.	2.6 [1.0,4.7]	83.0 [76.0,88.0]	6.6	27.6	13.31
Cent.	1.9 [0.5,3.6]	82.8 [76.2,87.6]	6.2	29.2	13.52
Graph	9.2 [4.9,14.6]	82.5 [74.0,88.5]	6.8	28.7	12.93
AV	2.6 [0.5,5.7]	81.1 [74.4,86.6]	7.6	27.6	28.34
M03 — Qwen3.8-27B					
Base	50.9 [42.7,59.9]	88.0 [82.4,92.3]	0.0	26.4	11.24
Rules	17.7 [10.9,24.8]	86.6 [80.7,92.1]	2.6	26.4	11.80
Cur.	10.2 [5.2,15.9]	84.4 [75.9,90.0]	6.9	26.7	11.66
Disc.	3.3 [1.0,5.9]	84.2 [76.0,90.0]	6.6	26.1	12.35
Mem.	3.3 [1.1,6.0]	84.4 [76.3,90.0]	6.9	26.1	12.33
Bin.	2.6 [1.0,4.7]	84.5 [77.4,89.4]	6.6	26.0	12.25
Cent.	2.3 [0.7,4.1]	84.5 [78.0,89.5]	6.2	27.6	12.47
Graph	9.9 [4.9,15.8]	84.7 [76.6,90.3]	6.4	27.0	11.95
AV	2.6 [0.5,5.7]	82.8 [75.5,88.4]	7.6	26.0	25.96
M04 — gpt-oss-20b					
Base	55.2 [46.7,62.6]	79.2 [74.0,84.2]	0.0	23.1	9.11
Rules	19.8 [12.8,27.1]	77.8 [72.5,83.2]	2.6	23.1	9.67
Cur.	10.8 [5.8,16.0]	76.0 [68.4,82.5]	6.4	23.3	9.65
Disc.	3.5 [1.4,5.9]	75.9 [69.0,81.8]	6.6	23.0	10.20
Mem.	3.3 [1.5,5.9]	75.9 [69.2,81.9]	6.8	22.8	10.17
Bin.	2.8 [0.9,4.8]	76.2 [69.9,82.2]	6.6	22.8	10.10
Cent.	2.4 [0.7,4.3]	76.0 [70.1,82.1]	6.2	24.4	10.30
Graph	9.9 [5.4,14.6]	75.9 [68.2,82.4]	6.8	23.6	9.97
AV	3.3 [0.8,6.3]	74.7 [68.3,80.7]	7.8	22.8	21.18
M05 — gpt-oss-120b					
Base	52.3 [44.0,60.6]	87.8 [82.2,92.2]	0.0	32.2	14.97
Rules	17.9 [11.2,24.8]	86.5 [80.7,91.9]	2.6	32.2	15.53
Cur.	10.4 [5.2,15.9]	84.5 [76.6,90.3]	6.2	32.2	15.15
Disc.	3.1 [1.0,5.6]	84.0 [75.9,89.7]	6.6	31.7	16.10
Mem.	2.8 [1.0,5.0]	84.0 [76.1,89.2]	6.9	31.5	16.04
Bin.	2.6 [1.0,4.7]	84.4 [77.4,89.3]	6.6	31.5	15.97
Cent.	2.3 [0.7,4.1]	84.4 [78.0,89.2]	6.2	33.1	16.19
Graph	9.5 [4.9,15.0]	84.2 [76.0,90.1]	6.8	32.8	15.42
AV	2.6 [0.5,5.7]	82.6 [75.4,88.4]	7.6	31.5	34.30
M06 — DeepSeek-V4.1-Flash					
Base	54.3 [46.1,62.5]	88.5 [82.9,92.5]	0.0	24.7	10.17
Rules	18.9 [11.9,26.4]	87.2 [81.2,92.2]	2.6	24.7	10.74
Cur.	10.4 [5.4,15.6]	85.1 [76.9,90.7]	6.2	25.0	10.65
Disc.	3.1 [1.1,5.1]	84.7 [76.5,90.1]	6.6	24.5	11.27
Mem.	3.0 [1.1,5.6]	84.9 [76.8,89.8]	6.8	24.4	11.24
Bin.	2.4 [0.8,4.6]	85.2 [78.0,89.9]	6.6	24.4	11.17
Cent.	2.1 [0.5,4.0]	84.9 [78.4,89.4]	6.2	26.0	11.37
Graph	9.2 [4.9,14.2]	84.7 [76.4,90.3]	6.8	25.3	10.95
AV	2.8 [0.5,5.7]	83.3 [75.8,88.8]	7.8	24.4	23.57
M07 — GPT-6 Astra					
Base	48.8 [40.1,57.5]	93.6 [90.0,96.5]	0.0	42.1	21.37
Rules	17.4 [10.9,24.7]	92.2 [88.3,95.7]	2.6	42.1	21.93
Cur.	10.4 [5.2,16.0]	89.9 [83.3,95.2]	5.9	42.1	21.18
Disc.	3.8 [1.2,6.6]	89.4 [82.8,94.7]	6.6	41.0	22.59
Mem.	3.3 [1.0,6.1]	89.6 [83.5,94.7]	6.8	41.1	22.52
Bin.	3.0 [1.0,5.5]	89.9 [85.0,94.7]	6.4	41.1	22.46
Cent.	2.3 [0.7,4.0]	90.1 [85.7,94.7]	6.1	42.7	22.63
Graph	9.7 [4.8,15.9]	89.6 [82.1,94.7]	6.4	42.1	21.43
AV	3.3 [1.0,6.2]	88.0 [82.4,93.2]	7.6	41.0	48.65

Method	H [interval]	U [interval]	F	L	C
M08 — GPT-6 Sol					
Base	49.8 [40.8,59.4]	91.3 [86.3,95.3]	0.0	31.3	14.44
Rules	17.4 [10.8,24.7]	89.9 [84.3,94.3]	2.6	31.3	15.00
Cur.	10.6 [5.3,16.4]	87.8 [80.2,93.6]	5.9	32.0	14.66
Disc.	3.6 [1.3,6.2]	87.5 [79.8,93.0]	6.6	30.9	15.59
Mem.	3.3 [1.1,6.0]	87.7 [80.2,93.0]	6.8	30.8	15.54
Bin.	2.6 [1.0,4.7]	88.0 [81.3,93.2]	6.4	30.8	15.48
Cent.	2.4 [0.9,4.3]	87.8 [81.9,92.6]	6.1	32.3	15.66
Graph	9.7 [4.8,15.5]	87.8 [80.2,93.6]	6.2	32.0	14.94
AV	2.6 [0.5,5.7]	86.1 [78.8,92.0]	7.6	30.7	33.12
M09 — GPT-6 Luna					
Base	51.2 [43.1,59.1]	83.7 [78.4,88.5]	0.0	21.5	8.04
Rules	17.5 [10.9,24.7]	82.3 [76.2,88.1]	2.6	21.5	8.60
Cur.	10.1 [5.2,15.8]	80.2 [72.2,86.1]	6.6	22.0	8.66
Disc.	3.1 [1.0,5.6]	80.0 [72.5,85.5]	6.6	21.4	9.13
Mem.	3.1 [1.0,5.8]	79.9 [72.4,85.4]	6.9	21.3	9.10
Bin.	2.6 [1.0,4.7]	80.2 [73.2,85.8]	6.6	21.3	9.05
Cent.	2.3 [0.7,4.1]	80.2 [73.8,85.7]	6.2	22.8	9.28
Graph	9.7 [4.9,15.5]	80.6 [73.2,86.8]	6.4	22.0	8.96
AV	2.6 [0.5,5.7]	78.5 [71.4,85.0]	7.6	21.3	18.82
M10 — Claude Opus 5.5					
Base	46.9 [38.7,55.4]	92.5 [87.8,95.8]	0.0	44.5	22.97
Rules	17.0 [10.5,24.7]	91.1 [85.7,95.1]	2.6	44.5	23.53
Cur.	9.9 [5.1,15.1]	88.9 [81.5,94.3]	6.1	44.4	22.72
Disc.	4.0 [1.3,7.2]	88.7 [81.2,94.1]	6.1	43.3	24.26
Mem.	3.5 [1.0,6.4]	88.5 [81.8,93.7]	6.4	43.3	24.16
Bin.	3.1 [1.0,5.7]	89.2 [83.2,94.2]	5.7	43.3	24.11
Cent.	2.8 [0.9,5.2]	89.1 [83.6,93.7]	5.6	44.9	24.30
Graph	9.4 [4.5,15.1]	88.0 [80.1,93.9]	6.9	44.4	22.94
AV	3.3 [1.1,5.9]	87.0 [80.3,92.8]	7.5	43.3	52.27
M11 — Claude Sonnet 5					
Base	47.7 [39.1,56.9]	88.5 [82.9,92.5]	0.0	30.5	13.91
Rules	17.2 [10.8,24.7]	87.2 [81.2,92.2]	2.6	30.5	14.47
Cur.	10.4 [5.2,16.0]	85.1 [76.6,90.4]	5.7	31.2	14.19
Disc.	3.8 [1.2,6.6]	84.7 [76.5,90.1]	6.6	30.1	15.07
Mem.	3.3 [1.0,6.1]	84.9 [76.8,90.1]	6.8	30.1	15.02
Bin.	3.0 [1.0,5.5]	85.2 [77.7,90.3]	6.2	29.9	14.97
Cent.	2.3 [0.7,4.0]	85.4 [78.6,90.4]	6.1	31.7	15.18
Graph	9.5 [4.7,15.4]	84.9 [76.5,90.4]	6.2	31.2	14.46
AV	3.3 [1.0,6.2]	83.3 [75.8,88.8]	7.5	29.9	31.97
M12 — Gemini 3.8 Flash					
Base	52.1 [44.0,60.6]	89.8 [84.9,93.7]	0.0	23.9	9.64
Rules	17.9 [11.2,24.8]	88.4 [82.7,93.1]	2.6	23.9	10.20
Cur.	10.4 [5.2,15.9]	86.3 [78.1,91.5]	6.4	24.2	10.16
Disc.	3.1 [1.0,5.6]	85.9 [77.9,90.6]	6.6	23.8	10.74
Mem.	2.8 [1.0,5.0]	85.9 [78.4,90.6]	6.9	23.6	10.70
Bin.	2.6 [1.0,4.7]	86.3 [79.6,90.5]	6.6	23.6	10.64
Cent.	2.3 [0.7,4.1]	86.3 [80.3,90.3]	6.2	25.2	10.87
Graph	9.7 [4.9,15.0]	86.1 [78.1,91.4]	6.8	24.5	10.46
AV	2.6 [0.5,5.7]	84.5 [77.7,90.0]	7.6	23.6	22.39
M13 — Gemini 3.7 Flash					
Base	53.6 [45.5,61.9]	86.8 [81.1,91.1]	0.0	22.6	8.79
Rules	18.6 [11.5,26.0]	85.4 [79.5,90.6]	2.6	22.6	9.35
Cur.	10.6 [5.4,15.8]	83.3 [75.2,89.3]	6.2	22.8	9.36
Disc.	3.1 [1.0,5.6]	83.0 [75.7,88.7]	6.6	22.5	9.88
Mem.	2.8 [1.0,5.0]	83.2 [75.8,88.1]	6.8	22.4	9.85
Bin.	2.4 [0.9,4.3]	83.5 [77.0,88.5]	6.6	22.3	9.79
Cent.	1.9 [0.5,3.6]	83.2 [77.0,88.5]	6.2	23.9	10.00
Graph	9.4 [4.9,14.6]	83.0 [75.0,88.9]	6.8	23.1	9.66
AV	2.6 [0.5,5.7]	81.6 [75.2,87.7]	7.8	22.3	20.48

Table 11 (continued): P03 — AgentHarm stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	65.5 [57.9,72.9]	69.8 [61.7,75.9]	0.0	22.6	8.53
Rules	23.8 [17.5,30.7]	68.9 [60.1,75.3]	2.4	22.6	9.12
Cur.	10.6 [6.1,15.1]	63.0 [55.2,69.6]	13.5	23.3	9.24
Disc.	5.7 [2.6,9.4]	65.3 [57.3,72.0]	8.3	22.4	9.33
Mem.	5.0 [1.7,9.0]	65.5 [57.6,72.0]	7.8	22.5	9.33
Bin.	4.3 [1.6,7.6]	65.8 [57.6,71.7]	7.1	22.2	9.22
Cent.	4.0 [1.3,7.6]	65.6 [57.6,71.4]	7.1	23.2	9.40
Graph	10.1 [5.4,14.7]	63.0 [55.2,69.6]	14.1	23.4	9.51
AV	4.5 [1.7,8.2]	63.9 [56.4,70.7]	9.4	22.2	18.75
M02 — Qwen3.6-35B-A3B					
Base	63.4 [55.9,71.9]	82.8 [75.8,87.7]	0.0	30.3	13.60
Rules	22.0 [15.6,28.5]	81.4 [74.2,86.4]	2.4	30.3	14.19
Cur.	9.9 [5.6,15.5]	75.3 [68.6,81.1]	13.9	30.3	13.80
Disc.	5.4 [2.4,9.1]	77.1 [69.8,83.0]	9.0	29.6	14.27
Mem.	4.7 [1.6,8.9]	77.8 [70.9,83.3]	8.0	29.8	14.32
Bin.	4.2 [1.6,7.8]	78.1 [71.4,83.2]	7.1	29.3	14.21
Cent.	4.0 [1.6,7.7]	78.3 [71.2,83.5]	6.8	30.3	14.40
Graph	8.9 [4.6,14.1]	74.7 [68.2,80.4]	14.4	30.4	14.01
AV	4.3 [1.6,8.0]	76.0 [68.5,81.5]	9.9	29.7	29.36
M03 — Qwen3.8-27B					
Base	62.3 [55.3,69.7]	85.1 [78.1,89.6]	0.0	28.5	12.42
Rules	21.9 [15.7,28.8]	83.5 [76.3,88.5]	2.4	28.5	13.01
Cur.	9.0 [4.8,14.1]	77.1 [70.7,82.8]	14.8	28.9	12.73
Disc.	5.4 [2.6,9.1]	79.2 [72.2,84.7]	9.0	27.8	13.11
Mem.	4.5 [1.6,8.1]	79.9 [73.0,85.1]	8.5	28.3	13.13
Bin.	4.0 [1.6,6.9]	80.4 [73.4,85.6]	7.1	27.6	13.06
Cent.	3.8 [1.3,6.8]	80.6 [73.6,85.9]	6.8	28.8	13.25
Graph	8.9 [4.7,14.1]	76.6 [70.1,81.9]	14.6	28.7	12.96
AV	4.2 [1.6,7.7]	78.3 [71.2,84.0]	10.1	27.8	26.88
M04 — gpt-oss-20b					
Base	64.4 [57.1,71.9]	74.1 [66.3,79.2]	0.0	24.9	10.06
Rules	23.1 [17.1,30.2]	73.1 [65.2,78.6]	2.4	24.9	10.65
Cur.	10.1 [5.9,15.5]	67.0 [59.7,72.8]	13.4	25.4	10.62
Disc.	5.4 [2.4,9.0]	69.3 [61.7,75.3]	8.3	24.4	10.83
Mem.	5.0 [1.7,9.0]	69.4 [62.0,75.2]	8.0	24.8	10.84
Bin.	4.3 [1.6,7.6]	69.8 [62.5,75.3]	7.1	24.2	10.73
Cent.	4.0 [1.3,7.6]	69.8 [62.4,75.2]	6.9	25.4	10.91
Graph	9.2 [4.9,14.1]	66.5 [58.3,72.4]	14.2	25.7	10.88
AV	4.7 [2.1,8.2]	67.7 [60.4,74.0]	9.5	24.4	21.95
M05 — gpt-oss-120b					
Base	62.7 [55.0,70.8]	84.4 [77.1,89.3]	0.0	34.7	16.55
Rules	22.0 [15.7,28.9]	83.0 [75.5,88.3]	2.4	34.7	17.14
Cur.	9.5 [5.3,15.1]	76.7 [70.2,82.5]	14.2	34.7	16.44
Disc.	5.6 [2.6,9.5]	78.6 [71.4,84.5]	9.0	33.7	17.17
Mem.	4.7 [1.6,8.3]	79.2 [72.2,84.7]	8.3	33.9	17.20
Bin.	4.2 [1.6,7.6]	79.5 [72.5,85.4]	7.3	33.4	17.09
Cent.	3.8 [1.3,6.8]	79.9 [72.5,85.4]	6.8	34.5	17.29
Graph	8.9 [4.7,14.1]	76.0 [68.9,81.4]	14.6	34.6	16.61
AV	4.2 [1.6,7.7]	77.6 [70.4,83.4]	9.9	33.9	35.53
M06 — DeepSeek-V4.1-Flash					
Base	65.1 [57.2,73.0]	86.8 [79.9,91.4]	0.0	26.7	11.24
Rules	22.6 [16.4,29.9]	85.2 [77.8,90.4]	2.4	26.7	11.83
Cur.	10.2 [5.8,16.0]	79.0 [72.4,84.5]	13.7	27.0	11.68
Disc.	5.6 [2.3,9.7]	81.4 [74.5,86.8]	8.7	26.1	11.99
Mem.	5.2 [1.7,9.9]	81.8 [74.6,87.0]	8.0	26.6	12.00
Bin.	4.3 [1.6,8.0]	82.1 [75.5,87.6]	7.1	25.9	11.89
Cent.	4.2 [1.6,7.9]	82.1 [75.8,87.3]	6.9	26.9	12.07
Graph	9.2 [5.1,14.3]	78.3 [71.7,84.1]	14.2	27.2	11.91
AV	4.5 [1.9,8.1]	79.9 [73.1,85.4]	9.9	26.1	24.40
M07 — GPT-6 Astra					
Base	60.4 [52.4,67.5]	92.0 [86.5,95.7]	0.0	45.4	23.62
Rules	21.9 [15.9,28.8]	90.5 [83.9,95.1]	2.4	45.4	24.21
Cur.	9.4 [5.4,14.7]	83.2 [77.0,88.4]	15.1	45.0	22.72
Disc.	5.0 [1.9,8.6]	86.1 [79.3,91.4]	9.0	43.5	24.12
Mem.	4.3 [1.4,7.8]	86.1 [79.8,91.4]	8.9	43.5	24.07
Bin.	3.8 [1.2,6.8]	87.2 [80.9,91.9]	7.3	43.5	24.08
Cent.	3.6 [1.1,6.7]	87.3 [81.2,91.7]	6.8	44.6	24.26
Graph	9.0 [5.2,14.1]	82.3 [75.5,88.0]	15.8	45.0	22.86
AV	4.0 [1.6,7.4]	84.7 [78.3,89.8]	10.2	43.7	50.36

Method	H [interval]	U [interval]	F	L	C
M08 — GPT-6 Sol					
Base	61.3 [54.5,68.9]	88.9 [82.6,93.8]	0.0	33.8	15.96
Rules	21.9 [15.7,28.8]	87.3 [79.7,92.9]	2.4	33.8	16.55
Cur.	9.2 [5.0,14.6]	80.9 [74.0,86.3]	14.8	33.9	15.88
Disc.	5.4 [2.6,9.1]	83.2 [76.0,89.2]	9.0	32.9	16.59
Mem.	4.5 [1.6,8.1]	83.5 [76.6,89.2]	8.7	33.1	16.58
Bin.	4.0 [1.6,6.9]	84.2 [77.3,89.5]	7.1	32.8	16.54
Cent.	3.8 [1.3,6.8]	84.4 [77.3,89.7]	6.8	33.7	16.73
Graph	8.9 [4.7,14.1]	79.9 [73.0,85.9]	14.9	33.8	16.09
AV	4.2 [1.6,7.7]	81.9 [75.3,87.5]	10.2	32.9	34.28
M09 — GPT-6 Luna					
Base	62.0 [54.6,70.3]	80.6 [73.3,85.2]	0.0	23.1	8.89
Rules	21.7 [15.5,28.1]	79.2 [71.5,84.3]	2.4	23.1	9.47
Cur.	8.9 [4.7,14.3]	72.9 [65.6,78.4]	14.6	23.8	9.56
Disc.	5.4 [2.5,8.8]	74.8 [67.8,80.4]	9.0	22.7	9.64
Mem.	4.5 [1.6,7.8]	75.3 [68.7,80.4]	8.5	23.0	9.66
Bin.	4.2 [1.6,7.6]	75.7 [69.3,80.9]	7.3	22.7	9.58
Cent.	3.8 [1.3,6.8]	76.0 [69.3,81.2]	6.8	23.9	9.77
Graph	8.5 [4.2,13.5]	72.4 [65.3,77.7]	14.6	23.8	9.82
AV	4.2 [1.6,7.7]	73.8 [66.7,79.4]	9.9	22.9	19.46
M10 — Claude Opus 5.5					
Base	58.3 [49.7,66.0]	90.8 [85.0,95.1]	0.0	48.0	25.39
Rules	20.5 [14.6,27.1]	89.2 [82.8,94.3]	2.4	48.0	25.98
Cur.	8.9 [5.1,14.1]	81.6 [75.1,87.3]	16.1	47.7	24.26
Disc.	5.0 [1.9,8.6]	84.7 [77.1,90.5]	9.4	46.0	25.80
Mem.	4.5 [1.6,7.9]	85.1 [78.9,90.6]	8.9	46.0	25.80
Bin.	3.8 [1.2,6.8]	85.9 [80.0,91.1]	7.3	46.0	25.81
Cent.	3.6 [1.1,6.7]	86.3 [80.2,91.3]	6.6	47.3	26.00
Graph	8.9 [5.1,14.1]	81.1 [74.0,87.0]	16.3	47.4	24.42
AV	3.8 [1.6,7.1]	83.5 [77.3,89.3]	10.2	46.2	54.06
M11 — Claude Sonnet 5					
Base	59.0 [51.2,66.7]	86.8 [79.9,91.4]	0.0	32.9	15.37
Rules	21.2 [15.3,27.5]	85.2 [77.8,90.4]	2.4	32.9	15.96
Cur.	8.9 [4.8,14.3]	78.0 [71.4,83.5]	15.3	33.1	15.36
Disc.	5.0 [1.9,8.6]	80.9 [73.2,86.5]	9.0	32.1	16.00
Mem.	4.5 [1.6,7.9]	81.1 [74.2,86.4]	8.9	32.1	15.97
Bin.	3.8 [1.2,6.8]	82.3 [75.5,87.7]	7.1	31.9	15.98
Cent.	3.6 [1.1,6.7]	82.3 [75.8,87.3]	6.8	32.9	16.15
Graph	8.7 [4.6,14.0]	77.3 [70.3,83.2]	16.0	32.9	15.57
AV	4.0 [1.6,7.4]	79.7 [72.9,85.4]	10.2	31.9	33.05
M12 — Gemini 3.8 Flash					
Base	63.4 [55.8,71.2]	87.8 [80.7,92.1]	0.0	25.8	10.65
Rules	22.0 [15.7,28.9]	86.3 [78.5,91.3]	2.4	25.8	11.24
Cur.	9.5 [5.3,15.1]	79.9 [73.0,85.4]	14.4	26.3	11.15
Disc.	5.6 [2.6,9.5]	81.9 [75.0,87.5]	9.0	25.2	11.37
Mem.	4.7 [1.6,8.3]	82.6 [75.3,88.0]	8.3	25.7	11.40
Bin.	4.2 [1.6,7.6]	83.0 [76.1,88.0]	7.3	25.1	11.31
Cent.	3.8 [1.3,6.8]	83.3 [76.2,88.6]	6.8	26.3	11.51
Graph	8.9 [4.7,14.1]	79.3 [72.4,84.9]	14.6	26.3	11.38
AV	4.2 [1.6,7.7]	81.1 [74.0,86.7]	9.9	25.5	23.17
M13 — Gemini 3.7 Flash					
Base	63.9 [56.3,72.0]	83.2 [76.1,88.0]	0.0	24.4	9.71
Rules	22.0 [15.6,28.5]	81.8 [74.2,86.9]	2.4	24.4	10.30
Cur.	9.7 [5.2,15.5]	75.7 [68.9,81.2]	13.9	25.0	10.31
Disc.	5.6 [2.5,9.7]	77.4 [70.2,83.0]	8.9	23.9	10.45
Mem.	4.9 [1.6,9.5]	78.1 [71.2,83.3]	8.0	24.3	10.49
Bin.	4.2 [1.6,7.8]	78.5 [71.4,83.4]	7.1	23.8	10.38
Cent.	4.0 [1.6,7.7]	78.6 [71.2,83.7]	6.8	25.0	10.57
Graph	8.9 [4.6,14.1]	75.0 [68.2,80.4]	14.4	25.0	10.56
AV	4.3 [1.6,8.0]	76.4 [68.9,81.9]	9.9	24.1	21.20

Table 11 (continued): P04 — Agent-SafetyBench stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	55.0 [47.3,62.0]	72.6 [65.9,78.6]	0.0	24.8	9.20
Rules	20.5 [14.6,27.4]	71.7 [64.8,78.0]	1.9	24.8	9.81
Cur.	12.3 [7.2,18.8]	70.1 [64.0,77.3]	5.6	25.2	9.98
Disc.	8.7 [3.5,15.1]	70.7 [64.5,76.5]	5.7	24.6	10.38
Mem.	7.5 [3.0,13.5]	70.3 [64.2,76.4]	6.2	24.7	10.35
Bin.	6.6 [2.3,12.2]	70.0 [64.2,76.3]	6.8	24.7	10.22
Cent.	6.2 [1.9,11.8]	70.0 [64.2,76.3]	6.9	27.8	10.49
Graph	12.0 [6.8,18.6]	70.1 [64.0,77.3]	5.7	25.5	10.37
AV	5.7 [2.3,10.2]	68.8 [62.6,75.5]	7.8	23.9	19.62
M02 — Qwen3.6-35B-A3B					
Base	51.7 [45.2,58.0]	85.6 [79.8,90.7]	0.0	33.2	14.67
Rules	18.8 [13.1,25.6]	84.7 [79.3,90.0]	1.9	33.2	15.28
Cur.	11.6 [7.0,17.6]	83.3 [77.6,88.5]	5.2	33.4	15.13
Disc.	9.4 [3.7,16.1]	83.0 [77.8,88.2]	5.7	32.1	15.86
Mem.	8.3 [3.4,15.0]	82.8 [77.0,87.9]	5.9	32.2	15.83
Bin.	7.5 [3.1,13.4]	82.5 [76.4,87.7]	6.4	32.1	15.68
Cent.	6.8 [2.4,12.4]	82.5 [76.6,87.7]	6.6	35.2	15.97
Graph	10.9 [6.2,16.6]	83.2 [77.4,88.5]	5.6	33.6	15.51
AV	5.4 [1.7,10.1]	80.9 [75.5,86.3]	7.8	32.0	30.75
M03 — Qwen3.8-27B					
Base	48.8 [42.1,55.1]	88.2 [82.0,94.1]	0.0	31.2	13.40
Rules	17.4 [12.2,23.7]	87.2 [81.4,92.2]	1.9	31.2	14.01
Cur.	10.6 [5.7,16.4]	85.8 [80.0,91.8]	5.6	31.4	13.93
Disc.	9.5 [3.8,16.7]	85.2 [79.6,91.3]	5.7	30.4	14.59
Mem.	9.0 [3.4,16.3]	85.2 [79.6,91.2]	5.9	30.4	14.58
Bin.	7.6 [3.0,14.2]	85.1 [78.2,91.1]	6.1	30.4	14.47
Cent.	7.6 [2.7,14.1]	85.1 [78.7,91.0]	6.2	33.3	14.73
Graph	10.4 [5.7,16.0]	85.6 [79.9,91.8]	5.6	31.8	14.32
AV	5.2 [1.6,10.2]	83.3 [77.2,89.7]	7.6	30.1	28.19
M04 — gpt-oss-20b					
Base	53.6 [46.0,60.4]	76.9 [70.5,83.8]	0.0	27.3	10.85
Rules	20.0 [14.4,26.5]	76.0 [69.8,82.6]	1.9	27.3	11.47
Cur.	12.5 [7.5,18.4]	74.8 [68.3,81.5]	5.2	27.5	11.54
Disc.	9.4 [4.1,15.5]	74.5 [67.8,81.2]	5.7	26.8	12.02
Mem.	8.0 [3.3,13.9]	74.3 [67.9,81.2]	6.2	26.7	12.00
Bin.	7.1 [3.0,13.0]	74.0 [67.7,80.3]	6.6	27.1	11.87
Cent.	6.4 [1.9,12.4]	74.0 [67.7,80.3]	6.6	29.9	12.16
Graph	11.6 [6.7,18.2]	74.5 [68.0,81.4]	5.6	28.1	11.91
AV	5.9 [2.5,10.4]	72.7 [66.3,79.2]	7.8	26.4	22.99
M05 — gpt-oss-120b					
Base	50.5 [43.8,56.5]	87.5 [81.7,93.3]	0.0	38.1	17.85
Rules	18.1 [12.9,24.2]	86.5 [81.0,91.7]	1.9	38.1	18.46
Cur.	11.1 [6.4,16.7]	85.1 [79.4,91.3]	5.2	38.1	18.15
Disc.	9.4 [3.7,16.7]	84.7 [79.0,90.0]	5.7	36.7	19.06
Mem.	8.9 [3.4,16.0]	84.5 [78.5,90.3]	5.9	36.7	19.02
Bin.	7.6 [3.1,13.4]	84.4 [78.1,90.3]	6.2	36.7	18.86
Cent.	7.3 [2.7,13.8]	84.4 [78.6,90.3]	6.6	39.8	19.12
Graph	10.4 [5.8,16.1]	85.1 [79.4,91.3]	5.4	38.2	18.53
AV	5.4 [1.7,10.1]	82.6 [76.8,88.5]	7.6	36.7	37.22
M06 — DeepSeek-V4.1-Flash					
Base	53.8 [46.9,60.2]	89.1 [83.4,95.0]	0.0	29.3	12.12
Rules	20.3 [14.7,26.7]	88.0 [82.5,93.6]	1.9	29.3	12.74
Cur.	12.8 [8.1,18.6]	86.6 [80.2,93.2]	5.2	29.4	12.74
Disc.	9.5 [4.1,16.5]	86.1 [79.6,92.6]	5.7	28.3	13.31
Mem.	8.5 [3.5,15.5]	86.1 [80.2,92.6]	5.9	28.5	13.28
Bin.	7.3 [3.0,13.1]	85.8 [79.2,92.6]	6.4	28.8	13.14
Cent.	6.8 [2.3,13.5]	85.9 [79.9,92.6]	6.6	31.7	13.43
Graph	12.2 [7.5,18.2]	86.3 [80.0,92.9]	5.6	30.0	13.11
AV	5.9 [2.5,10.4]	83.9 [77.3,90.4]	7.8	28.2	25.57
M07 — GPT-6 Astra					
Base	44.8 [37.6,52.5]	92.7 [86.6,97.4]	0.0	49.8	25.48
Rules	15.6 [10.9,21.4]	91.7 [85.4,96.8]	1.9	49.8	26.09
Cur.	9.4 [4.9,15.1]	90.1 [83.1,95.8]	5.6	49.4	25.39
Disc.	9.2 [3.7,16.6]	89.8 [83.6,95.2]	5.6	47.9	26.73
Mem.	8.7 [3.2,15.7]	89.8 [83.6,95.5]	5.7	47.9	26.67
Bin.	7.8 [2.7,14.5]	90.1 [83.9,95.5]	5.4	47.9	26.57
Cent.	7.8 [3.0,14.1]	89.6 [82.9,95.1]	6.1	51.0	26.78
Graph	9.0 [4.7,14.3]	90.1 [83.1,95.8]	5.6	49.6	25.71
AV	5.4 [2.0,9.8]	87.8 [81.6,93.5]	7.6	47.7	52.79
M08 — GPT-6 Sol					
Base	46.7 [39.9,53.1]	90.1 [84.5,95.3]	0.0	37.1	17.21
Rules	16.3 [11.5,22.3]	89.1 [83.2,94.1]	1.9	37.1	17.82
Cur.	9.9 [5.2,15.1]	87.5 [81.1,93.6]	5.6	37.2	17.54
Disc.	9.4 [4.1,16.6]	87.2 [81.5,92.9]	5.7	35.8	18.43
Mem.	8.9 [3.5,16.1]	87.2 [81.7,92.8]	5.9	35.8	18.40
Bin.	7.3 [2.6,13.0]	87.3 [81.7,92.8]	6.1	35.9	18.28
Cent.	8.0 [2.8,14.1]	87.0 [80.5,92.9]	6.2	38.9	18.52
Graph	9.5 [5.0,14.7]	87.5 [81.1,93.6]	5.6	37.3	17.91
AV	4.7 [1.4,9.4]	85.2 [78.9,91.4]	7.6	35.8	35.96
M09 — GPT-6 Luna					
Base	49.5 [42.8,55.3]	84.0 [78.3,89.5]	0.0	25.4	9.58
Rules	17.4 [12.0,23.7]	83.2 [77.6,88.5]	1.9	25.4	10.19
Cur.	10.4 [5.7,16.4]	81.8 [75.9,87.2]	5.6	26.1	10.32
Disc.	9.5 [3.8,16.7]	81.2 [75.9,87.1]	5.7	25.0	10.77
Mem.	9.0 [3.4,16.3]	81.2 [75.5,86.9]	5.9	25.2	10.77
Bin.	7.6 [3.0,14.2]	81.1 [75.1,86.7]	6.2	25.2	10.67
Cent.	7.6 [2.7,15.1]	80.9 [75.3,86.8]	6.4	28.0	10.95
Graph	10.2 [5.5,16.0]	81.6 [75.9,87.2]	5.6	26.1	10.72
AV	5.0 [1.6,9.9]	79.7 [73.8,85.2]	7.6	24.8	20.43
M10 — Claude Opus 5.5					
Base	42.9 [35.8,50.0]	91.3 [85.3,96.0]	0.0	52.7	27.39
Rules	14.9 [10.4,20.5]	90.3 [84.0,95.3]	1.9	52.7	28.00
Cur.	8.9 [4.3,14.3]	88.7 [82.3,94.2]	5.7	52.5	27.16
Disc.	8.9 [3.3,15.5]	88.4 [82.4,93.7]	5.4	50.7	28.69
Mem.	8.5 [3.1,15.8]	88.4 [82.5,93.8]	5.6	50.7	28.63
Bin.	7.8 [2.7,15.1]	88.7 [82.8,93.8]	5.4	50.7	28.53
Cent.	7.5 [2.7,14.4]	88.2 [81.6,93.7]	5.9	53.8	28.72
Graph	8.5 [4.3,13.9]	88.7 [82.3,94.2]	6.1	52.5	27.51
AV	5.6 [2.0,11.2]	86.1 [79.6,91.7]	7.6	50.4	56.67
M11 — Claude Sonnet 5					
Base	43.6 [36.8,51.2]	89.1 [83.4,95.0]	0.0	36.1	16.58
Rules	15.1 [10.5,20.7]	88.0 [82.5,93.6]	1.9	36.1	17.19
Cur.	9.0 [4.5,14.6]	86.5 [80.2,93.2]	5.6	36.1	16.93
Disc.	8.9 [3.5,15.6]	86.1 [80.3,92.6]	5.6	34.9	17.81
Mem.	8.3 [3.2,15.4]	86.1 [80.6,92.6]	5.6	34.9	17.76
Bin.	7.5 [2.6,14.0]	86.5 [80.6,92.6]	5.4	34.9	17.68
Cent.	7.6 [3.0,13.7]	85.9 [79.4,92.6]	6.1	37.9	17.94
Graph	8.7 [4.3,13.9]	86.5 [80.2,93.2]	5.6	36.3	17.32
AV	5.2 [1.7,9.8]	84.2 [77.8,90.9]	7.6	34.8	34.69
M12 — Gemini 3.8 Flash					
Base	50.2 [43.7,56.0]	89.9 [84.5,95.3]	0.0	28.3	11.49
Rules	17.9 [12.6,23.9]	88.9 [83.2,94.1]	1.9	28.3	12.10
Cur.	11.1 [6.4,16.7]	87.5 [81.1,93.6]	5.4	28.5	12.12
Disc.	9.4 [3.7,16.7]	87.2 [81.5,92.9]	5.7	27.5	12.70
Mem.	8.9 [3.4,16.0]	87.0 [81.3,92.8]	5.9	27.6	12.67
Bin.	7.6 [3.1,13.4]	86.8 [80.6,92.8]	6.2	27.6	12.55
Cent.	7.6 [2.8,14.4]	87.0 [81.2,92.9]	6.2	30.6	12.85
Graph	10.4 [5.8,16.1]	87.5 [81.1,93.6]	5.4	29.1	12.52
AV	5.2 [1.7,10.1]	85.1 [78.9,91.3]	7.6	27.3	24.31
M13 — Gemini 3.7 Flash					
Base	53.0 [46.0,59.0]	86.8 [81.0,92.1]	0.0	26.7	10.47
Rules	19.6 [14.1,26.2]	85.9 [80.4,91.5]	1.9	26.7	11.08
Cur.	12.3 [7.7,18.4]	84.4 [79.0,90.6]	5.2	27.0	11.16
Disc.	9.4 [3.8,15.9]	83.9 [78.1,89.5]	5.7	26.2	11.66
Mem.	8.5 [3.5,15.6]	83.9 [77.9,89.9]	5.9	26.2	11.65
Bin.	7.1 [2.8,13.0]	83.5 [77.7,89.8]	6.4	26.5	11.52
Cent.	6.8 [2.1,13.1]	83.7 [78.0,89.9]	6.6	29.5	11.82
Graph	11.6 [6.8,18.0]	84.2 [78.6,90.6]	5.6	27.1	11.55
AV	5.6 [2.0,10.1]	81.8 [76.1,87.6]	7.8	25.9	22.23

Table 11 (continued): P05 — AgentDyn stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	61.3 [53.8,67.9]	72.6 [65.3,78.6]	0.0	35.2	13.37
Rules	27.6 [20.2,34.8]	71.7 [64.4,78.4]	2.6	35.2	14.12
Cur.	16.3 [10.8,21.5]	69.3 [60.8,76.9]	6.6	35.2	14.24
Disc.	8.7 [4.9,14.1]	71.0 [63.0,78.1]	4.3	34.6	14.16
Mem.	8.0 [4.2,12.0]	71.2 [63.7,77.8]	4.2	34.7	14.12
Bin.	7.1 [3.6,10.4]	70.8 [63.3,77.3]	4.3	34.7	13.94
Cent.	9.7 [5.4,14.5]	71.0 [63.3,77.4]	4.2	41.3	14.41
Graph	15.3 [9.7,20.8]	69.1 [60.3,76.7]	6.4	35.2	14.89
AV	8.9 [4.2,14.3]	71.0 [64.1,76.8]	4.5	34.2	23.56
M02 — Qwen3.6-35B-A3B					
Base	57.8 [49.2,65.5]	85.2 [78.8,90.6]	0.0	47.0	21.32
Rules	25.2 [17.9,32.3]	84.4 [77.5,89.5]	2.6	47.0	22.07
Cur.	15.1 [10.1,19.5]	81.6 [73.6,88.3]	6.6	47.1	21.71
Disc.	7.8 [3.6,13.0]	82.6 [76.0,88.8]	4.3	46.1	21.83
Mem.	7.3 [3.5,11.4]	82.8 [76.5,88.8]	4.2	46.1	21.75
Bin.	6.4 [2.9,10.6]	82.5 [75.6,88.6]	4.3	46.3	21.59
Cent.	8.9 [4.8,13.4]	82.8 [75.6,89.0]	4.0	53.0	22.06
Graph	14.9 [10.1,19.2]	80.7 [73.4,87.3]	6.9	47.0	22.27
AV	8.3 [3.1,13.6]	83.0 [77.1,88.3]	4.7	45.8	36.94
M03 — Qwen3.8-27B					
Base	56.4 [47.8,63.2]	88.9 [83.2,93.7]	0.0	44.2	19.47
Rules	24.3 [17.7,30.6]	88.0 [82.3,92.6]	2.6	44.2	20.22
Cur.	14.8 [9.6,19.2]	84.2 [77.4,90.4]	6.8	44.4	19.99
Disc.	7.6 [3.6,13.5]	86.1 [79.8,92.0]	4.3	43.5	20.04
Mem.	7.3 [3.6,11.5]	86.1 [79.9,92.1]	4.2	43.6	19.99
Bin.	6.6 [3.0,10.9]	86.1 [79.9,92.1]	4.2	43.6	19.84
Cent.	8.9 [4.8,13.5]	86.3 [79.9,92.4]	4.0	50.2	20.30
Graph	14.8 [9.6,19.2]	84.0 [77.5,89.9]	6.4	44.2	20.60
AV	8.0 [3.1,13.5]	86.3 [81.1,91.7]	4.9	43.1	33.84
M04 — gpt-oss-20b					
Base	60.2 [52.7,67.3]	78.6 [71.4,84.7]	0.0	38.7	15.77
Rules	26.9 [19.7,33.9]	77.8 [70.5,84.2]	2.6	38.7	16.52
Cur.	16.0 [10.8,20.6]	74.8 [66.3,82.4]	6.6	38.7	16.50
Disc.	7.6 [4.7,12.0]	76.6 [69.4,83.2]	4.3	38.1	16.48
Mem.	7.5 [4.2,11.3]	76.7 [69.6,83.3]	4.2	38.2	16.42
Bin.	6.6 [3.8,10.2]	76.2 [68.9,83.0]	4.3	38.2	16.24
Cent.	9.0 [5.2,13.0]	76.6 [69.3,83.3]	4.2	44.9	16.73
Graph	14.9 [9.1,19.2]	74.5 [66.1,81.6]	7.3	38.7	17.09
AV	8.7 [4.2,13.9]	76.7 [70.2,83.2]	4.5	37.7	27.61
M05 — gpt-oss-120b					
Base	57.1 [48.5,64.1]	87.7 [81.8,93.1]	0.0	53.9	25.95
Rules	24.5 [17.7,31.3]	86.8 [80.7,92.1]	2.6	53.9	26.70
Cur.	14.8 [9.6,19.1]	83.5 [76.2,90.0]	6.6	53.8	26.04
Disc.	7.6 [3.6,13.5]	84.9 [78.3,91.3]	4.3	52.6	26.29
Mem.	7.1 [3.4,11.5]	85.1 [78.5,91.4]	4.2	52.9	26.19
Bin.	6.2 [2.8,10.4]	84.7 [77.7,91.5]	4.3	53.1	26.01
Cent.	8.5 [4.6,12.5]	85.1 [78.1,91.6]	4.0	59.8	26.48
Graph	14.6 [9.5,18.8]	82.8 [75.9,89.5]	6.6	53.8	26.61
AV	8.2 [3.1,13.7]	85.1 [78.8,91.1]	4.7	52.5	44.73
M06 — DeepSeek-V4.1-Flash					
Base	60.6 [52.2,67.2]	89.9 [84.9,94.2]	0.0	41.5	17.62
Rules	26.7 [18.8,33.7]	89.1 [83.5,93.9]	2.6	41.5	18.37
Cur.	16.0 [10.7,20.2]	85.8 [79.4,91.9]	6.6	41.7	18.24
Disc.	8.2 [4.8,13.4]	87.2 [81.3,92.7]	4.3	40.8	18.27
Mem.	7.6 [4.2,11.9]	87.3 [81.4,92.8]	4.2	40.9	18.20
Bin.	6.8 [4.0,10.5]	87.0 [80.6,92.6]	4.3	40.9	18.03
Cent.	9.2 [5.7,13.4]	87.2 [80.8,92.9]	4.2	47.5	18.50
Graph	15.3 [10.2,19.6]	84.9 [78.8,91.1]	7.1	41.5	18.81
AV	9.0 [4.4,14.1]	87.2 [81.2,92.6]	4.7	40.4	30.71
M07 — GPT-6 Astra					
Base	53.5 [43.8,61.4]	94.8 [90.3,98.2]	0.0	70.5	37.05
Rules	23.6 [16.3,29.6]	93.9 [88.5,97.5]	2.6	70.5	37.80
Cur.	14.1 [9.2,17.8]	89.4 [82.6,95.0]	7.5	69.8	36.41
Disc.	7.3 [3.4,12.7]	91.7 [86.0,96.4]	4.3	68.7	36.98
Mem.	7.3 [3.2,12.3]	91.5 [85.8,96.4]	4.2	69.1	36.89
Bin.	6.2 [2.6,10.4]	91.8 [86.1,96.5]	4.0	69.1	36.77
Cent.	8.3 [4.0,12.4]	92.4 [86.5,97.0]	3.5	75.8	37.23
Graph	13.9 [9.2,17.5]	89.1 [83.2,94.6]	7.5	69.8	36.98
AV	7.3 [2.3,12.7]	92.0 [86.1,96.7]	4.9	68.6	63.42

Method	H [interval]	U [interval]	F	L	C
M08 — GPT-6 Sol					
Base	55.4 [46.2,62.4]	91.1 [86.5,94.9]	0.0	52.5	25.02
Rules	23.8 [16.6,29.5]	90.3 [85.4,94.4]	2.6	52.5	25.77
Cur.	14.2 [9.2,18.2]	86.6 [80.3,92.0]	6.8	52.5	25.18
Disc.	7.6 [3.6,13.5]	88.4 [82.8,93.2]	4.3	51.2	25.39
Mem.	7.3 [3.6,11.5]	88.2 [82.8,93.2]	4.2	51.8	25.33
Bin.	6.6 [3.0,10.9]	88.4 [82.7,93.6]	4.2	51.8	25.17
Cent.	8.9 [4.8,13.5]	88.9 [83.2,93.6]	3.6	58.5	25.67
Graph	14.1 [9.2,17.9]	86.3 [80.6,91.7]	6.8	52.5	25.75
AV	7.8 [2.8,13.5]	88.5 [82.9,93.2]	4.9	51.1	43.18
M09 — GPT-6 Luna					
Base	55.9 [46.9,62.8]	84.0 [77.5,89.2]	0.0	36.0	13.92
Rules	24.3 [17.7,30.4]	83.2 [76.2,88.5]	2.6	36.0	14.67
Cur.	14.6 [9.4,18.9]	79.7 [72.0,86.2]	6.8	36.2	14.77
Disc.	7.5 [3.6,13.0]	81.4 [74.7,87.9]	4.3	35.4	14.70
Mem.	6.9 [3.4,10.9]	81.6 [75.0,88.0]	4.2	35.5	14.65
Bin.	6.2 [2.8,10.4]	81.2 [73.8,87.6]	4.3	35.4	14.51
Cent.	8.7 [4.7,13.5]	81.6 [74.2,87.9]	4.0	42.1	14.99
Graph	14.6 [9.4,18.9]	79.5 [72.3,86.0]	6.4	36.0	15.41
AV	7.6 [3.1,12.7]	81.8 [75.3,87.2]	4.9	35.0	24.50
M10 — Claude Opus 5.5					
Base	51.2 [42.0,59.3]	93.6 [89.1,97.0]	0.0	74.6	39.82
Rules	22.4 [14.8,28.4]	92.7 [87.7,96.5]	2.6	74.6	40.57
Cur.	13.2 [8.6,17.2]	88.0 [81.8,93.7]	8.2	73.7	38.97
Disc.	7.3 [3.6,12.3]	90.6 [84.6,95.4]	4.3	72.8	39.69
Mem.	6.9 [3.2,11.5]	90.5 [84.5,95.1]	4.2	73.1	39.57
Bin.	6.2 [2.4,10.4]	91.1 [85.4,95.8]	4.0	73.1	39.51
Cent.	8.2 [4.0,12.2]	91.1 [85.3,95.7]	3.5	79.8	39.92
Graph	12.8 [8.1,16.8]	87.3 [80.8,93.6]	8.3	73.7	39.53
AV	7.5 [2.3,13.0]	90.8 [85.0,95.5]	4.9	72.6	68.14
M11 — Claude Sonnet 5					
Base	52.4 [43.1,60.6]	89.9 [84.9,94.2]	0.0	51.1	24.10
Rules	22.9 [15.5,29.3]	89.1 [83.5,93.9]	2.6	51.1	24.85
Cur.	13.5 [8.9,17.4]	85.1 [78.7,91.3]	7.5	51.1	24.28
Disc.	7.5 [3.6,12.7]	87.2 [81.3,92.7]	4.3	49.9	24.52
Mem.	7.3 [3.2,12.3]	86.8 [81.2,92.2]	4.3	50.4	24.44
Bin.	6.4 [2.6,10.5]	87.3 [81.4,93.0]	4.0	50.4	24.35
Cent.	8.3 [4.0,12.4]	87.7 [81.9,93.1]	3.5	57.1	24.81
Graph	13.4 [8.7,17.2]	84.7 [78.5,91.3]	7.6	51.1	24.89
AV	7.3 [2.3,12.7]	87.3 [81.4,92.6]	4.9	49.8	41.63
M12 — Gemini 3.8 Flash					
Base	57.1 [48.4,63.9]	90.3 [85.8,94.4]	0.0	40.1	16.70
Rules	24.7 [17.7,31.3]	89.4 [84.8,93.9]	2.6	40.1	17.45
Cur.	14.9 [9.7,19.4]	86.1 [80.1,91.7]	6.6	40.4	17.37
Disc.	7.6 [3.6,13.5]	87.5 [81.8,92.7]	4.3	39.5	17.38
Mem.	7.1 [3.4,11.5]	87.7 [82.2,92.8]	4.2	39.6	17.31
Bin.	6.4 [2.8,10.9]	87.3 [81.6,92.7]	4.3	39.5	17.16
Cent.	8.5 [4.6,12.5]	87.7 [81.6,93.0]	4.0	46.2	17.64
Graph	14.8 [9.7,18.9]	85.4 [79.7,91.3]	6.6	40.1	17.99
AV	8.2 [3.1,13.7]	87.7 [82.2,92.7]	4.7	39.1	29.17
M13 — Gemini 3.7 Flash					
Base	59.4 [51.3,66.6]	86.8 [80.1,92.5]	0.0	37.9	15.22
Rules	25.7 [18.0,32.5]	85.9 [78.9,91.3]	2.6	37.9	15.97
Cur.	15.6 [10.4,20.2]	83.0 [75.1,89.7]	6.6	38.2	15.99
Disc.	8.3 [4.8,13.7]	84.0 [77.3,90.2]	4.3	37.3	15.95
Mem.	7.8 [4.4,11.9]	84.2 [77.6,90.6]	4.2	37.4	15.89
Bin.	6.8 [4.0,10.5]	83.9 [77.1,90.1]	4.3	37.4	15.74
Cent.	9.2 [5.7,13.4]	84.2 [77.1,90.2]	4.0	44.0	16.22
Graph	14.9 [9.9,19.3]	82.1 [74.6,88.9]	6.9	37.9	16.58
AV	9.0 [4.4,14.1]	84.4 [78.4,90.1]	4.7	36.9	26.68

Table 11 (continued): P06 — AgentLAB stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	64.6 [57.4,71.4]	71.2 [64.6,77.2]	0.0	41.3	15.75
Rules	24.8 [17.9,31.2]	70.7 [64.2,76.6]	2.4	41.3	16.59
Cur.	14.2 [9.4,18.4]	68.2 [61.5,75.3]	3.1	40.8	16.70
Disc.	5.2 [2.3,9.2]	69.3 [62.1,75.5]	4.0	39.4	16.37
Mem.	4.7 [2.3,7.9]	68.9 [61.9,75.0]	4.3	39.4	16.33
Bin.	3.8 [1.4,7.6]	69.6 [62.2,75.6]	4.9	39.4	16.13
Cent.	7.1 [3.5,12.5]	69.8 [62.7,75.6]	4.7	44.9	16.52
Graph	13.2 [8.6,17.2]	67.9 [61.2,75.1]	3.3	40.8	17.23
AV	6.4 [3.5,9.7]	69.8 [63.1,75.3]	3.5	39.5	25.88
M02 — Qwen3.6-35B-A3B					
Base	61.1 [53.6,67.4]	85.1 [79.7,89.5]	0.0	55.1	25.13
Rules	24.0 [17.7,30.4]	84.4 [79.0,89.1]	2.4	55.1	25.96
Cur.	14.4 [9.9,18.9]	81.8 [75.6,87.6]	3.0	54.1	25.57
Disc.	4.7 [1.9,8.8]	82.8 [76.9,88.5]	4.2	52.4	25.27
Mem.	4.0 [1.6,7.5]	82.6 [76.1,88.5]	4.0	52.8	25.18
Bin.	3.5 [1.3,7.0]	82.8 [76.8,88.3]	4.5	52.8	24.98
Cent.	6.4 [2.7,11.5]	82.6 [76.7,88.2]	4.7	58.3	25.30
Graph	12.5 [8.0,16.8]	81.2 [74.7,87.1]	3.6	53.7	26.04
AV	6.2 [3.2,9.5]	83.5 [77.0,88.3]	3.5	52.8	40.57
M03 — Qwen3.8-27B					
Base	58.7 [50.2,66.7]	88.7 [83.5,92.7]	0.0	52.0	22.95
Rules	23.1 [17.1,29.4]	88.0 [82.6,92.5]	2.4	52.0	23.78
Cur.	13.5 [9.0,18.4]	85.4 [79.6,91.1]	3.8	51.0	23.49
Disc.	4.3 [1.7,8.6]	86.5 [80.3,91.3]	4.5	49.3	23.19
Mem.	4.0 [1.6,7.6]	86.3 [79.7,91.1]	4.0	49.7	23.16
Bin.	3.5 [1.4,6.9]	86.6 [80.3,91.4]	4.3	49.7	22.98
Cent.	6.4 [2.7,11.8]	86.3 [80.2,91.1]	4.5	55.2	23.29
Graph	12.0 [7.5,17.3]	85.1 [79.4,90.5]	3.6	50.7	23.97
AV	6.1 [3.1,9.5]	87.0 [80.7,91.7]	3.8	49.7	37.17
M04 — gpt-oss-20b					
Base	63.5 [55.6,70.5]	76.2 [70.0,81.5]	0.0	45.6	18.59
Rules	24.3 [17.9,30.7]	75.7 [69.3,81.1]	2.4	45.6	19.42
Cur.	14.2 [9.4,18.7]	73.1 [66.7,79.5]	3.1	44.8	19.38
Disc.	4.7 [1.7,8.9]	74.1 [67.4,80.4]	4.2	43.2	19.04
Mem.	4.3 [2.0,7.7]	73.8 [67.2,80.2]	4.3	43.5	19.02
Bin.	3.6 [1.4,7.1]	74.3 [67.5,80.7]	4.9	43.5	18.81
Cent.	6.6 [3.0,11.5]	74.3 [67.4,80.3]	4.5	49.0	19.19
Graph	12.7 [8.3,16.8]	72.6 [65.6,79.4]	3.1	44.7	19.91
AV	6.4 [3.5,9.7]	74.5 [67.8,79.9]	3.5	43.6	30.31
M05 — gpt-oss-120b					
Base	60.2 [52.3,67.3]	87.5 [82.1,91.8]	0.0	62.9	30.58
Rules	23.4 [17.5,29.8]	86.8 [81.2,91.5]	2.4	62.9	31.42
Cur.	14.1 [9.7,18.4]	84.2 [78.2,90.0]	3.1	61.9	30.72
Disc.	4.5 [1.8,8.8]	85.4 [79.1,90.9]	4.2	60.1	30.44
Mem.	4.2 [1.7,7.6]	85.1 [78.3,90.4]	4.0	60.5	30.36
Bin.	3.3 [1.3,6.6]	85.2 [78.8,90.4]	4.5	60.5	30.14
Cent.	6.4 [2.7,11.8]	85.1 [78.8,90.4]	4.5	66.0	30.44
Graph	12.2 [7.7,16.7]	83.9 [77.3,89.7]	3.5	61.4	31.12
AV	6.1 [3.1,9.4]	85.9 [79.2,90.6]	3.6	60.5	49.10
M06 — DeepSeek-V4.1-Flash					
Base	63.7 [55.6,70.0]	89.9 [84.4,94.2]	0.0	48.8	20.77
Rules	25.0 [18.6,31.0]	89.2 [83.9,93.9]	2.4	48.8	21.60
Cur.	14.8 [9.7,19.5]	86.6 [80.7,92.2]	3.0	47.7	21.44
Disc.	4.9 [2.1,8.9]	87.7 [81.8,92.5]	4.2	46.2	21.12
Mem.	4.3 [2.0,7.7]	87.0 [80.2,92.2]	4.2	46.6	21.06
Bin.	3.6 [1.4,7.1]	87.7 [80.8,92.2]	4.5	46.6	20.87
Cent.	6.4 [2.8,11.5]	87.5 [81.3,92.2]	4.7	52.1	21.20
Graph	12.8 [8.1,17.4]	85.9 [79.5,91.9]	3.5	47.7	21.94
AV	6.4 [3.5,9.7]	88.4 [81.8,93.1]	3.5	46.6	33.73
M07 — GPT-6 Astra					
Base	56.4 [49.0,62.8]	95.0 [90.9,98.3]	0.0	82.1	43.67
Rules	22.6 [16.4,28.6]	94.3 [89.8,97.6]	2.4	82.1	44.50
Cur.	13.2 [8.4,17.4]	91.1 [86.0,95.7]	4.0	80.3	43.06
Disc.	4.9 [2.1,9.8]	92.7 [87.6,96.9]	4.7	78.6	42.82
Mem.	4.7 [2.1,8.9]	92.7 [87.3,96.9]	4.0	79.2	42.78
Bin.	3.8 [1.5,7.5]	92.9 [88.0,97.3]	4.0	79.2	42.60
Cent.	6.9 [2.6,13.1]	92.5 [87.8,97.4]	4.3	84.7	42.81
Graph	11.3 [7.2,15.8]	90.1 [84.9,95.3]	4.3	80.3	43.40
AV	6.6 [3.1,10.4]	92.9 [87.4,97.6]	3.8	79.2	69.62
M08 — GPT-6 Sol					
Base	57.6 [49.5,64.8]	91.7 [86.6,95.9]	0.0	61.4	29.49
Rules	22.9 [16.7,29.3]	91.0 [85.9,95.3]	2.4	61.4	30.33
Cur.	13.5 [8.9,18.6]	88.0 [82.4,93.6]	3.8	60.3	29.69
Disc.	4.5 [1.8,8.9]	89.4 [83.4,94.7]	4.5	58.6	29.39
Mem.	4.3 [1.8,8.1]	89.4 [82.6,94.7]	4.0	59.0	29.36
Bin.	3.5 [1.4,6.9]	89.6 [83.7,94.8]	4.0	59.0	29.19
Cent.	6.6 [2.7,12.4]	89.2 [83.2,94.7]	4.3	64.5	29.44
Graph	11.6 [7.1,16.9]	87.5 [81.0,93.3]	3.6	59.9	30.11
AV	6.2 [3.1,9.8]	89.8 [83.5,95.1]	3.8	59.0	47.44
M09 — GPT-6 Luna					
Base	59.2 [51.1,66.0]	82.1 [76.1,87.2]	0.0	42.3	16.41
Rules	23.3 [17.4,29.6]	81.4 [75.1,86.9]	2.4	42.3	17.24
Cur.	14.1 [9.7,18.4]	78.8 [72.4,85.3]	3.5	41.7	17.33
Disc.	4.3 [1.7,8.3]	80.0 [73.6,86.0]	4.3	40.1	16.99
Mem.	4.2 [1.7,7.6]	79.7 [72.6,85.9]	4.0	40.4	16.97
Bin.	3.3 [1.3,6.6]	80.0 [73.3,85.9]	4.3	40.4	16.79
Cent.	6.4 [2.7,11.8]	79.9 [73.6,85.8]	4.5	45.9	17.14
Graph	12.7 [8.7,16.9]	78.5 [71.4,85.3]	3.5	41.7	17.83
AV	6.1 [3.1,9.4]	80.6 [73.4,86.0]	3.6	40.4	26.92
M10 — Claude Opus 5.5					
Base	53.3 [45.4,60.9]	93.6 [89.1,97.2]	0.0	86.9	46.94
Rules	20.7 [13.4,28.0]	92.9 [88.2,96.7]	2.4	86.9	47.77
Cur.	12.0 [7.3,16.7]	89.6 [83.7,94.6]	4.9	85.0	46.12
Disc.	4.7 [2.0,9.3]	91.3 [85.0,95.8]	4.9	83.2	45.93
Mem.	4.7 [2.1,9.1]	91.3 [85.2,95.8]	4.0	83.8	45.88
Bin.	4.0 [1.6,8.1]	91.7 [86.2,95.9]	3.8	83.8	45.71
Cent.	6.8 [2.6,12.8]	91.5 [86.0,95.6]	4.2	89.2	45.97
Graph	10.4 [6.2,14.6]	88.7 [82.0,94.6]	4.7	85.0	46.45
AV	6.4 [3.1,10.2]	91.7 [85.8,96.5]	3.8	83.8	74.83
M11 — Claude Sonnet 5					
Base	54.7 [47.5,61.3]	89.9 [84.4,94.2]	0.0	59.8	28.40
Rules	21.5 [15.8,28.2]	89.2 [83.9,93.9]	2.4	59.8	29.23
Cur.	12.3 [7.9,16.7]	86.3 [80.1,92.2]	4.2	58.8	28.64
Disc.	4.2 [1.7,8.1]	87.5 [81.0,92.6]	4.9	57.0	28.34
Mem.	4.0 [1.6,7.5]	87.5 [80.4,92.9]	4.2	57.4	28.32
Bin.	3.5 [1.3,6.7]	88.0 [81.8,93.4]	3.8	57.4	28.18
Cent.	6.2 [2.3,11.9]	87.8 [82.1,92.5]	4.3	62.9	28.50
Graph	10.8 [6.8,15.4]	85.2 [78.4,91.7]	4.7	58.8	29.07
AV	6.2 [3.1,9.5]	88.0 [81.2,93.1]	3.8	57.4	45.75
M12 — Gemini 3.8 Flash					
Base	60.9 [53.0,68.2]	91.1 [86.5,95.4]	0.0	47.2	19.68
Rules	24.1 [18.0,30.4]	90.5 [85.3,94.9]	2.4	47.2	20.51
Cur.	14.4 [9.7,19.3]	87.8 [82.3,93.3]	3.1	46.4	20.43
Disc.	4.3 [1.7,8.3]	89.1 [83.3,94.3]	4.3	44.7	20.09
Mem.	4.2 [1.7,7.6]	88.7 [82.4,94.1]	4.0	45.0	20.05
Bin.	3.3 [1.3,6.6]	89.1 [83.2,94.5]	4.5	45.0	19.87
Cent.	6.4 [2.7,11.8]	88.7 [82.8,94.3]	4.5	50.5	20.21
Graph	12.7 [8.0,18.0]	87.7 [81.8,93.3]	3.5	46.2	20.91
AV	6.2 [3.1,9.7]	89.6 [83.4,94.9]	3.6	45.0	32.04
M13 — Gemini 3.7 Flash					
Base	61.8 [54.1,68.2]	85.9 [80.9,90.4]	0.0	44.6	17.93
Rules	24.3 [18.1,30.7]	85.2 [79.9,89.8]	2.4	44.6	18.77
Cur.	14.6 [9.6,18.9]	82.6 [76.9,88.3]	3.0	43.9	18.77
Disc.	4.7 [1.9,8.8]	83.7 [78.0,89.2]	4.2	42.2	18.44
Mem.	4.2 [1.8,7.7]	83.2 [76.6,88.9]	4.2	42.5	18.38
Bin.	3.5 [1.2,7.0]	83.7 [77.6,89.2]	4.5	42.5	18.21
Cent.	6.4 [2.8,11.5]	83.5 [77.6,89.1]	4.7	48.0	18.55
Graph	12.8 [8.4,17.2]	81.9 [74.9,88.0]	3.5	43.8	19.29
AV	6.4 [3.5,9.7]	84.4 [77.7,89.5]	3.5	42.5	29.31

Table 11 (continued): P07 — TAMAS stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	58.7 [52.8,64.8]	70.3 [63.7,76.0]	0.0	30.4	11.94
Rules	27.8 [20.3,34.6]	69.1 [62.2,75.5]	2.6	30.4	12.65
Cur.	17.0 [11.1,22.3]	67.7 [60.3,73.7]	5.9	30.2	12.81
Disc.	10.9 [7.3,15.6]	68.8 [61.1,74.8]	4.0	29.3	12.96
Mem.	10.8 [7.3,15.1]	68.6 [61.0,74.8]	4.0	29.6	12.95
Bin.	8.7 [5.3,13.0]	68.6 [61.2,74.1]	3.8	29.1	12.84
Cent.	11.3 [7.5,16.0]	68.9 [61.5,75.2]	3.5	34.0	13.20
Graph	16.5 [10.8,21.9]	67.5 [60.1,73.5]	6.1	30.8	13.34
AV	12.2 [7.6,17.8]	69.3 [62.2,75.0]	3.5	29.9	22.57
M02 — Qwen3.6-35B-A3B					
Base	56.8 [50.2,64.8]	83.5 [77.3,89.1]	0.0	40.7	19.05
Rules	26.9 [18.8,35.6]	81.6 [75.2,87.6]	2.6	40.7	19.76
Cur.	16.8 [10.5,23.8]	79.5 [74.5,85.0]	5.6	40.3	19.44
Disc.	10.8 [6.4,15.6]	81.1 [74.2,86.5]	4.2	38.8	19.92
Mem.	10.9 [6.6,15.6]	81.1 [74.2,86.8]	4.3	38.9	19.89
Bin.	9.5 [5.6,14.4]	81.2 [74.7,87.5]	3.6	38.9	19.80
Cent.	11.8 [7.1,16.8]	81.6 [75.2,87.6]	3.3	43.7	20.18
Graph	16.0 [9.6,22.4]	79.0 [73.8,84.7]	5.7	41.1	19.96
AV	11.6 [6.8,17.6]	81.9 [75.9,88.1]	3.8	39.7	35.34
M03 — Qwen3.8-27B					
Base	54.7 [48.1,62.9]	87.5 [81.8,93.3]	0.0	38.3	17.40
Rules	26.4 [18.4,34.6]	85.6 [79.1,91.7]	2.6	38.3	18.10
Cur.	15.8 [9.7,21.7]	82.8 [78.1,87.5]	6.1	38.7	17.89
Disc.	10.9 [6.4,15.6]	84.5 [78.1,90.4]	4.3	36.6	18.31
Mem.	11.5 [7.7,15.6]	84.7 [77.7,90.6]	4.3	36.7	18.29
Bin.	10.4 [6.3,15.1]	85.1 [78.6,90.7]	3.1	36.6	18.21
Cent.	12.2 [7.4,17.0]	85.2 [78.8,91.5]	3.5	41.5	18.58
Graph	15.6 [9.6,21.6]	82.8 [78.0,88.3]	5.7	38.7	18.40
AV	11.6 [7.2,17.8]	85.9 [80.3,92.1]	3.8	37.4	32.37
M04 — gpt-oss-20b					
Base	58.9 [52.5,65.4]	75.9 [69.1,81.7]	0.0	33.5	14.09
Rules	28.1 [19.8,35.5]	74.5 [67.2,80.3]	2.6	33.5	14.80
Cur.	17.4 [10.9,22.9]	72.4 [66.0,78.0]	5.6	33.4	14.80
Disc.	11.1 [6.8,15.9]	74.1 [66.9,80.0]	4.0	32.0	15.06
Mem.	11.1 [7.3,15.9]	74.0 [66.4,80.0]	4.3	32.4	15.03
Bin.	9.2 [5.4,13.3]	74.0 [66.7,80.0]	3.6	32.0	14.94
Cent.	11.8 [7.4,16.6]	74.3 [67.0,80.1]	3.3	36.9	15.32
Graph	16.7 [10.6,22.1]	72.0 [66.0,77.8]	5.9	33.9	15.35
AV	12.3 [7.2,18.0]	74.8 [68.4,80.4]	3.3	32.9	26.44
M05 — gpt-oss-120b					
Base	55.9 [49.7,63.5]	85.9 [79.4,91.7]	0.0	46.7	23.19
Rules	26.6 [18.4,34.8]	84.0 [78.0,90.1]	2.6	46.7	23.89
Cur.	16.5 [10.1,22.9]	81.8 [77.1,86.9]	5.6	46.1	23.31
Disc.	10.8 [6.4,15.6]	83.3 [76.5,89.3]	4.2	44.5	23.97
Mem.	11.1 [6.8,15.6]	83.3 [76.1,89.2]	4.3	44.6	23.93
Bin.	9.5 [5.6,14.4]	83.5 [76.8,89.7]	3.3	44.6	23.84
Cent.	11.8 [7.1,16.8]	83.9 [77.0,90.3]	3.3	49.4	24.22
Graph	15.8 [9.6,21.7]	81.2 [76.4,86.9]	5.7	46.9	23.79
AV	11.6 [7.1,17.7]	84.4 [78.4,90.5]	3.8	45.4	42.78
M06 — DeepSeek-V4.1-Flash					
Base	59.2 [52.3,66.3]	88.9 [83.5,94.0]	0.0	35.9	15.75
Rules	28.5 [20.0,36.8]	87.0 [81.1,92.4]	2.6	35.9	16.45
Cur.	17.4 [10.8,23.4]	84.5 [80.0,89.2]	5.6	35.7	16.34
Disc.	11.3 [6.8,16.0]	85.9 [80.1,91.7]	4.2	34.3	16.67
Mem.	11.6 [7.7,16.3]	85.9 [79.7,91.7]	4.3	34.4	16.66
Bin.	9.9 [5.9,15.0]	86.3 [80.2,91.9]	3.6	34.3	16.57
Cent.	12.3 [7.3,17.0]	86.6 [80.4,92.0]	3.3	39.2	16.94
Graph	16.7 [10.3,22.4]	84.0 [79.1,89.0]	5.7	36.3	16.87
AV	12.7 [7.5,18.0]	87.2 [81.8,92.5]	3.5	35.2	29.40
M07 — GPT-6 Astra					
Base	51.9 [44.9,61.3]	95.0 [91.1,98.1]	0.0	61.0	33.10
Rules	24.5 [16.5,32.5]	93.1 [88.6,96.9]	2.6	61.0	33.81
Cur.	15.1 [9.4,21.4]	89.6 [85.3,93.7]	6.1	60.8	32.60
Disc.	10.8 [6.6,15.5]	91.5 [86.5,95.4]	4.7	58.2	33.66
Mem.	10.8 [6.2,15.7]	91.8 [87.2,95.8]	4.2	58.2	33.63
Bin.	9.5 [5.3,14.3]	92.5 [87.7,96.5]	3.1	58.2	33.57
Cent.	12.2 [7.3,17.1]	92.9 [88.3,96.6]	3.1	63.1	33.98
Graph	14.6 [8.9,21.1]	89.1 [84.4,93.3]	6.2	60.8	33.04
AV	11.1 [6.6,17.2]	93.1 [88.9,96.8]	3.8	59.5	60.68
M08 — GPT-6 Sol					
Base	54.2 [48.5,62.4]	92.2 [87.7,96.5]	0.0	45.5	22.36
Rules	26.2 [18.1,34.5]	90.3 [85.4,94.8]	2.6	45.5	23.06
Cur.	16.0 [10.1,22.7]	87.5 [82.9,92.2]	5.9	45.7	22.53
Disc.	11.1 [7.1,15.6]	89.1 [84.3,93.7]	4.5	43.3	23.17
Mem.	11.5 [7.2,16.4]	89.2 [84.7,94.1]	4.3	43.5	23.13
Bin.	10.2 [6.2,15.1]	89.8 [85.4,94.7]	3.1	43.5	23.06
Cent.	12.2 [7.8,17.0]	89.8 [85.3,94.6]	3.5	48.3	23.42
Graph	15.5 [9.6,21.6]	86.6 [81.7,91.1]	6.2	45.7	23.02
AV	11.6 [7.4,17.6]	90.3 [86.2,94.8]	3.8	44.3	41.31
M09 — GPT-6 Luna					
Base	54.7 [47.9,62.5]	81.8 [75.3,87.8]	0.0	31.1	12.44
Rules	26.2 [18.2,34.5]	79.9 [73.6,86.1]	2.6	31.1	13.14
Cur.	16.1 [10.1,22.6]	77.8 [72.1,84.0]	6.1	31.5	13.25
Disc.	10.8 [6.4,15.6]	79.5 [73.0,85.9]	4.2	30.0	13.46
Mem.	11.3 [7.3,15.6]	79.5 [72.9,85.9]	4.3	30.3	13.44
Bin.	10.1 [6.1,14.9]	79.7 [73.0,85.9]	3.1	30.0	13.36
Cent.	12.0 [7.1,17.0]	80.0 [73.0,86.4]	3.3	34.7	13.76
Graph	15.5 [9.6,21.5]	77.8 [72.0,83.9]	5.7	31.5	13.77
AV	11.5 [7.0,17.7]	80.4 [74.5,86.7]	3.8	30.6	23.44
M10 — Claude Opus 5.5					
Base	50.0 [43.1,59.7]	93.6 [89.5,96.8]	0.0	64.6	35.58
Rules	23.3 [16.0,31.1]	91.7 [86.9,95.7]	2.6	64.6	36.29
Cur.	14.1 [8.4,20.2]	88.2 [83.7,92.7]	6.6	64.2	34.92
Disc.	10.4 [6.3,15.2]	90.1 [85.3,94.2]	4.7	61.5	36.06
Mem.	10.2 [6.0,15.4]	90.6 [86.5,94.6]	4.2	61.5	36.05
Bin.	9.4 [4.9,14.3]	91.1 [87.0,95.0]	3.1	61.6	36.01
Cent.	11.8 [7.3,16.5]	91.7 [87.0,95.3]	3.1	66.5	36.45
Graph	13.5 [7.8,19.8]	87.5 [82.6,92.2]	6.2	64.2	35.37
AV	10.8 [5.9,17.0]	91.3 [87.4,95.0]	4.0	63.0	65.15
M11 — Claude Sonnet 5					
Base	50.9 [44.3,59.9]	88.9 [83.5,94.0]	0.0	44.3	21.53
Rules	24.1 [16.3,31.9]	87.0 [81.1,92.4]	2.6	44.3	22.24
Cur.	14.9 [9.0,20.9]	83.5 [78.4,88.0]	6.4	44.6	21.77
Disc.	10.8 [6.6,15.5]	85.4 [79.2,91.6]	4.7	42.2	22.33
Mem.	10.6 [6.2,15.6]	85.9 [79.7,91.7]	4.2	42.2	22.33
Bin.	9.5 [5.1,14.7]	86.5 [80.2,91.9]	3.1	42.3	22.26
Cent.	11.8 [7.2,16.6]	86.8 [81.3,92.5]	3.1	47.2	22.69
Graph	14.4 [8.5,20.8]	83.0 [78.0,88.0]	6.2	44.6	22.26
AV	11.1 [6.6,17.2]	87.2 [81.8,92.5]	3.8	43.3	39.84
M12 — Gemini 3.8 Flash					
Base	56.2 [49.8,63.4]	90.3 [85.1,94.8]	0.0	34.7	14.92
Rules	26.6 [18.6,34.7]	88.4 [82.1,93.7]	2.6	34.7	15.62
Cur.	16.5 [10.4,22.8]	85.8 [81.9,90.0]	5.7	35.1	15.57
Disc.	10.9 [6.4,15.6]	87.3 [81.8,92.6]	4.2	33.3	15.88
Mem.	11.3 [7.3,15.6]	87.3 [81.9,92.6]	4.3	33.4	15.87
Bin.	10.1 [6.0,14.6]	87.7 [82.2,92.7]	3.3	33.3	15.78
Cent.	12.0 [7.3,16.9]	88.0 [82.3,92.9]	3.3	38.0	16.17
Graph	15.8 [9.6,22.2]	85.4 [81.2,90.1]	5.7	35.1	16.09
AV	11.6 [7.2,17.8]	88.5 [83.3,93.5]	3.8	34.1	27.90
M13 — Gemini 3.7 Flash					
Base	58.0 [51.6,65.6]	84.4 [78.2,90.2]	0.0	32.8	13.60
Rules	27.4 [19.1,35.7]	82.5 [75.6,88.7]	2.6	32.8	14.30
Cur.	17.0 [10.5,23.8]	80.2 [75.1,85.9]	5.6	32.8	14.32
Disc.	10.9 [6.6,15.8]	81.8 [74.7,87.6]	4.2	31.4	14.57
Mem.	11.3 [7.1,16.3]	81.8 [74.5,87.5]	4.3	31.8	14.56
Bin.	9.7 [5.9,14.8]	81.9 [74.9,87.8]	3.6	31.4	14.48
Cent.	12.2 [7.3,17.0]	82.3 [75.4,88.3]	3.3	36.2	14.86
Graph	16.3 [10.0,22.4]	79.7 [74.7,85.5]	5.7	33.2	14.87
AV	12.3 [7.2,17.9]	82.8 [76.7,89.0]	3.5	32.2	25.53

Table 11 (continued): P08 — DUMA-Bench stratum. H/U: % [95% interval]; F: workflow false blocking (%); L: end-to-end p95 seconds; C: lifecycle work/workflow.

Cur.=Current; Disc.=Discovery; Mem.=Memory; Bin.=Binary; Cent.=Central; AV=AgentVax. Labels denote the platform strata and actor configurations used in the evaluation.

Method	H [interval]	U [interval]	F	L	C
M01 — Qwen3-8B					
Base	56.9 [49.0,64.4]	72.0 [64.5,79.0]	0.0	32.8	12.54
Rules	27.6 [20.4,34.3]	71.7 [63.9,78.9]	1.7	32.8	13.27
Cur.	16.0 [10.4,21.1]	69.1 [61.5,77.3]	5.7	33.0	13.35
Disc.	8.9 [5.9,13.4]	69.6 [61.9,77.7]	4.7	32.4	13.65
Mem.	8.3 [5.4,12.9]	69.6 [61.7,76.6]	4.5	32.4	13.63
Bin.	6.9 [4.2,11.7]	70.0 [63.0,77.5]	4.3	32.3	13.47
Cent.	8.9 [5.4,13.3]	69.8 [62.0,77.4]	4.3	38.3	13.96
Graph	15.3 [10.1,20.8]	68.9 [61.5,76.8]	6.1	33.0	13.97
AV	12.0 [7.1,16.7]	69.6 [62.7,77.2]	3.8	32.4	23.17
M02 — Qwen3.6-35B-A3B					
Base	53.8 [45.1,61.8]	85.8 [78.7,91.8]	0.0	43.9	20.01
Rules	25.9 [18.5,33.1]	85.4 [77.9,91.4]	1.7	43.9	20.73
Cur.	15.3 [9.9,21.6]	83.2 [75.2,90.3]	5.4	43.8	20.39
Disc.	8.3 [5.3,12.7]	83.2 [75.2,90.3]	4.9	43.3	20.96
Mem.	7.6 [4.7,11.4]	83.3 [76.3,90.0]	4.7	42.7	20.92
Bin.	6.8 [3.9,10.2]	83.7 [76.6,90.5]	4.5	42.6	20.76
Cent.	8.7 [5.4,12.4]	83.7 [76.6,90.1]	4.2	48.9	21.26
Graph	14.2 [9.2,20.5]	83.2 [75.2,90.3]	5.7	43.8	20.99
AV	11.6 [7.2,16.9]	83.0 [75.3,89.7]	3.8	43.4	36.34
M03 — Qwen3.8-27B					
Base	51.4 [42.9,59.0]	90.1 [83.9,95.4]	0.0	41.3	18.27
Rules	24.8 [17.4,32.4]	89.8 [83.5,95.2]	1.7	41.3	18.99
Cur.	14.8 [9.1,21.7]	87.5 [81.0,93.8]	5.6	41.3	18.79
Disc.	8.3 [5.3,12.2]	87.3 [81.0,93.4]	4.9	40.8	19.28
Mem.	7.5 [4.7,11.0]	87.7 [81.4,93.7]	4.5	40.6	19.26
Bin.	6.9 [3.8,10.2]	87.7 [80.9,94.2]	4.5	40.1	19.10
Cent.	8.5 [5.3,11.6]	88.2 [81.4,94.2]	4.0	46.4	19.58
Graph	14.6 [9.1,21.2]	87.3 [80.7,93.8]	5.9	41.3	19.37
AV	11.8 [7.1,17.3]	87.0 [80.8,92.9]	3.8	40.8	33.29
M04 — gpt-oss-20b					
Base	55.9 [47.7,64.0]	78.0 [70.6,84.6]	0.0	36.1	14.80
Rules	26.4 [19.0,33.4]	77.6 [70.3,84.0]	1.7	36.1	15.52
Cur.	15.5 [10.1,21.5]	75.2 [68.6,82.8]	5.6	36.2	15.48
Disc.	8.3 [5.2,13.0]	75.3 [68.6,82.9]	4.7	35.7	15.85
Mem.	7.8 [4.9,11.9]	75.3 [68.1,82.7]	4.5	35.7	15.83
Bin.	6.6 [3.8,10.6]	75.7 [68.6,82.9]	4.3	35.1	15.67
Cent.	8.7 [5.4,12.4]	75.7 [68.6,82.9]	4.5	41.5	16.16
Graph	14.4 [9.2,20.2]	74.7 [67.9,82.4]	6.1	36.3	16.10
AV	11.3 [6.9,15.6]	75.2 [67.6,82.4]	3.8	35.7	27.15
M05 — gpt-oss-120b					
Base	52.1 [43.8,59.7]	89.1 [83.0,94.6]	0.0	50.3	24.35
Rules	25.2 [17.8,32.7]	88.7 [82.4,94.4]	1.7	50.3	25.07
Cur.	14.8 [9.2,21.3]	86.5 [79.8,92.6]	5.6	50.2	24.48
Disc.	8.2 [5.2,12.2]	86.5 [79.8,92.6]	4.9	49.7	25.22
Mem.	7.3 [4.5,11.0]	86.8 [80.7,92.8]	4.5	48.9	25.19
Bin.	6.8 [3.7,10.2]	86.8 [79.8,92.8]	4.5	48.8	25.01
Cent.	8.3 [5.2,11.6]	87.2 [80.7,92.8]	4.2	55.2	25.49
Graph	14.2 [8.8,20.8]	86.5 [79.8,92.6]	5.7	50.2	25.06
AV	11.5 [6.8,16.7]	86.1 [79.9,92.1]	3.8	49.7	44.01
M06 — DeepSeek-V4.1-Flash					
Base	55.6 [46.8,64.4]	91.3 [85.5,95.9]	0.0	38.7	16.53
Rules	26.6 [19.1,34.0]	91.0 [85.4,95.6]	1.7	38.7	17.26
Cur.	15.5 [9.9,22.5]	88.5 [82.6,94.4]	5.6	38.7	17.12
Disc.	8.9 [5.9,14.0]	88.7 [82.7,94.4]	4.7	38.2	17.56
Mem.	8.2 [5.1,12.1]	88.5 [83.0,93.6]	4.5	38.2	17.53
Bin.	7.1 [4.2,11.4]	88.9 [82.7,94.7]	4.3	37.6	17.37
Cent.	9.2 [6.0,13.0]	89.2 [83.0,94.8]	4.3	44.0	17.87
Graph	14.4 [9.4,20.1]	88.4 [81.8,94.0]	5.9	38.8	17.74
AV	12.2 [7.7,17.5]	88.2 [82.6,93.2]	3.8	38.2	30.22
M07 — GPT-6 Astra					
Base	49.8 [42.0,57.8]	95.1 [90.6,98.3]	0.0	65.8	34.76
Rules	25.0 [17.2,32.0]	94.8 [89.9,97.9]	1.7	65.8	35.48
Cur.	15.3 [8.9,20.8]	92.0 [87.5,96.7]	5.7	65.6	34.34
Disc.	8.3 [5.4,12.2]	92.4 [87.6,97.0]	4.7	64.0	35.48
Mem.	7.5 [4.5,11.0]	92.5 [88.5,96.8]	4.7	64.0	35.41
Bin.	6.8 [3.6,10.0]	92.7 [88.7,97.0]	4.3	63.8	35.27
Cent.	9.0 [5.9,12.4]	93.1 [89.1,97.3]	4.3	70.2	35.65
Graph	14.6 [8.5,20.0]	91.8 [87.3,96.6]	6.1	65.6	34.90
AV	11.5 [7.0,16.7]	91.8 [86.9,96.4]	3.8	64.4	62.43

Method	H [interval]	U [interval]	F	L	C
M08 — GPT-6 Sol					
Base	50.3 [42.4,57.7]	93.2 [88.0,97.0]	0.0	49.0	23.48
Rules	24.7 [17.0,32.2]	92.9 [87.3,96.9]	1.7	49.0	24.20
Cur.	14.9 [9.1,21.5]	90.6 [85.1,96.0]	5.6	48.9	23.68
Disc.	8.3 [5.4,12.2]	90.5 [85.2,96.0]	4.7	48.4	24.42
Mem.	7.5 [4.6,11.0]	90.6 [85.6,95.7]	4.5	47.7	24.37
Bin.	6.8 [3.6,10.0]	90.8 [85.6,96.0]	4.5	47.5	24.20
Cent.	8.9 [5.5,12.4]	91.1 [85.8,96.4]	4.3	53.9	24.63
Graph	14.6 [9.1,21.2]	90.3 [85.1,95.9]	5.9	48.9	24.27
AV	11.8 [7.2,17.3]	89.9 [84.9,94.8]	3.8	48.4	42.48
M09 — GPT-6 Luna					
Base	50.9 [42.7,58.4]	85.1 [78.1,91.1]	0.0	33.5	13.06
Rules	24.7 [17.7,32.1]	84.7 [77.4,90.4]	1.7	33.5	13.79
Cur.	14.4 [8.6,21.1]	82.5 [74.7,89.9]	5.6	33.7	13.86
Disc.	8.0 [4.9,12.0]	82.5 [74.7,89.9]	4.9	33.4	14.17
Mem.	7.1 [4.3,11.0]	82.8 [76.0,89.5]	4.5	33.1	14.16
Bin.	6.6 [3.6,10.2]	83.0 [76.1,90.1]	4.5	33.0	14.02
Cent.	8.2 [4.8,11.6]	83.2 [76.1,90.2]	4.2	39.0	14.51
Graph	14.2 [8.6,20.8]	82.5 [74.7,89.9]	5.7	33.8	14.47
AV	11.1 [6.1,16.3]	82.3 [75.0,89.1]	3.8	33.2	24.11
M10 — Claude Opus 5.5					
Base	46.4 [39.3,54.5]	94.8 [90.3,97.9]	0.0	69.7	37.37
Rules	22.7 [15.7,30.2]	94.4 [89.6,97.6]	1.7	69.7	38.09
Cur.	13.7 [8.0,19.4]	91.8 [87.1,96.5]	6.1	69.4	36.82
Disc.	8.3 [5.3,12.1]	92.0 [87.3,96.5]	4.9	67.7	38.09
Mem.	7.5 [4.5,10.9]	92.2 [88.1,96.4]	4.7	67.7	38.00
Bin.	6.8 [3.6,10.2]	92.4 [88.1,96.7]	4.3	67.5	37.83
Cent.	9.0 [5.4,13.0]	92.7 [88.7,96.8]	4.5	73.9	38.24
Graph	13.2 [7.6,18.6]	91.3 [86.3,96.2]	6.4	69.4	37.37
AV	10.9 [6.8,16.3]	91.5 [86.5,95.9]	3.8	68.1	67.04
M11 — Claude Sonnet 5					
Base	47.9 [40.2,55.8]	91.3 [85.5,95.9]	0.0	47.7	22.61
Rules	23.4 [16.0,30.3]	91.0 [85.4,95.6]	1.7	47.7	23.33
Cur.	14.2 [7.8,19.9]	88.5 [82.8,94.5]	5.9	47.6	22.88
Disc.	8.5 [5.4,12.2]	88.7 [82.5,94.6]	4.7	46.5	23.57
Mem.	7.5 [4.5,11.0]	88.9 [83.0,94.2]	4.7	46.5	23.51
Bin.	6.9 [3.6,10.2]	89.1 [83.0,94.8]	4.3	46.3	23.40
Cent.	8.9 [5.4,12.4]	89.4 [83.0,95.0]	4.3	52.7	23.83
Graph	13.5 [7.7,18.8]	88.5 [82.8,94.5]	6.1	47.6	23.47
AV	11.1 [6.9,16.1]	88.2 [82.6,93.2]	3.8	47.1	40.98
M12 — Gemini 3.8 Flash					
Base	52.3 [44.1,60.0]	92.2 [86.3,96.7]	0.0	37.4	15.67
Rules	25.3 [17.8,33.2]	91.8 [85.7,96.3]	1.7	37.4	16.39
Cur.	14.9 [9.2,22.0]	89.6 [83.2,95.4]	5.6	37.4	16.31
Disc.	8.3 [5.3,12.2]	89.4 [83.2,95.2]	4.9	37.0	16.72
Mem.	7.5 [4.7,11.0]	89.8 [83.9,95.3]	4.5	37.0	16.71
Bin.	6.9 [3.8,10.2]	89.8 [83.3,95.5]	4.5	36.4	16.55
Cent.	8.7 [5.5,11.8]	90.3 [84.2,95.9]	4.2	42.8	17.04
Graph	14.6 [9.1,20.9]	89.6 [83.2,95.4]	5.7	37.6	16.91
AV	11.5 [6.9,16.8]	89.1 [83.3,93.7]	3.8	37.0	28.70
M13 — Gemini 3.7 Flash					
Base	54.9 [46.2,63.5]	87.7 [81.4,93.2]	0.0	35.4	14.28
Rules	26.4 [19.1,33.5]	87.3 [80.2,93.0]	1.7	35.4	15.00
Cur.	15.5 [10.0,21.9]	85.1 [78.1,91.4]	5.6	35.4	14.99
Disc.	8.7 [5.7,13.2]	85.2 [78.1,91.4]	4.7	35.0	15.36
Mem.	8.0 [4.9,12.0]	85.2 [78.5,91.4]	4.7	34.9	15.33
Bin.	7.1 [4.2,11.3]	85.4 [78.8,91.8]	4.3	34.5	15.18
Cent.	9.0 [5.8,13.0]	85.6 [79.2,91.8]	4.2	40.8	15.68
Graph	14.2 [9.4,19.4]	84.9 [77.9,91.1]	5.9	35.6	15.61
AV	12.0 [7.4,17.5]	84.9 [78.2,91.3]	3.8	34.9	26.25