WHEN TOOLS AGREE FOR THE WRONG REASONS: RISK-CONTROLLED EVIDENCE ACQUISITION UNDER SOURCE DEPENDENCE

Anonymous authors
Paper under double-blind review

ABSTRACT

Tool-using agents rely on external evidence, yet different tools may repeat information from shared sources or return plausible but corrupted records. We formulate evidence acquisition and hierarchical answer release as a joint risk–coverage–cost problem with distinct factual, unsupported-claim, and over-specificity losses. EviWeave-Agent uses a shared source-dependence representation to coordinate reliability-weighted fusion, active querying, and answer specificity. We evaluate the framework on CorruptEvidenceBench, a controlled benchmark spanning open-domain investigation, scientific evidence analysis, and cybersecurity triage under six corruption and dependence conditions. At $\alpha = 0.05$, policy-level calibration selects a frozen acquisition-and-release policy whose held-out 90-cell ledger has maximum-cell factual risk 0.038, minimum correct-task coverage 0.861, and mean query cost 2.31. Relative to selective risk control without dependency modeling, the selected policy raises correct-task coverage by 17.1 percentage points, lowers factual risk by 0.9 points, and reduces mean query cost by 13.8%. Matched diagnostics show that removing dependency modeling increases maximum factual risk to 0.106, while removing active acquisition raises cost to 2.84. These results support treating source dependence as a shared decision variable across evidence fusion, query selection, and calibrated answer release under the evaluated benchmark conditions.

1 INTRODUCTION

Tool-using agents increasingly combine search systems, structured databases, document parsers, and specialized models. Their evidence can appear individually plausible while sharing the same upstream error. A search result, a generated summary, and a cached database entry may appear to offer three corroborating observations while providing only one source of support.

The resulting decision problem is not simply whether to answer. An agent must decide which evidence applies to the current claim, which source domains remain informative after shared provenance is accounted for, which query justifies its cost, and how specific the final answer should be. A coarse answer can be useful when the evidence supports malicious command-and-control behavior but does not support a specific malware-family label.

EviWeave-Agent treats provenance structure as a common input to these decisions (Figure 1). Records sharing an identified upstream source are aggregated within a provenance cluster. The resulting source-domain representation governs evidence fusion, discounts redundant queries, and determines which answer node satisfies the policy-level risk and coverage constraints. The contribution is the coordinated acquisition-and-release formulation, rather than an isolated verifier, dependency graph, or calibration module.

Our controlled evaluation yields three findings. First, a local audit with a fixed Greedy controller shows that applicability filtering, deduplication, and reliability weighting can raise correct-task coverage from 0.675 to 0.937 while reducing factual risk across all tasks from 0.324 to 0.061. This audit isolates the combined evidence-processing intervention; it does not establish a certified operating point. Second, at the $\alpha = 0.05$ operating point, policy-level calibration selects a frozen EviWeave-Agent policy with maximum-cell factual risk 0.038, minimum correct-task coverage 0.861, and

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.

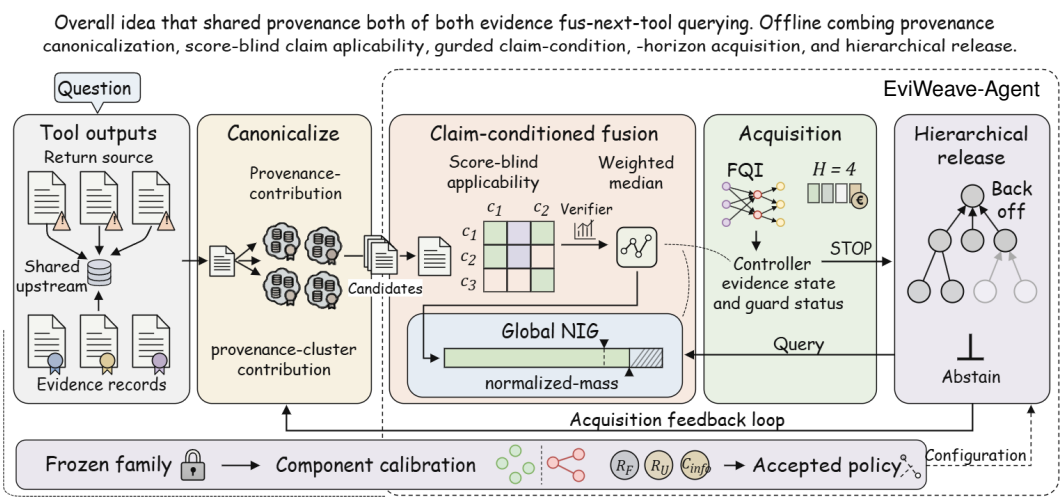


Figure 1: EviWeave-Agent framework. Shared provenance informs evidence canonicalization, claim-conditioned fusion, active acquisition, and hierarchical answer release. Offline calibration evaluates the complete frozen acquisition-and-release policy.

mean acquisition cost 2.31 on the held-out ledger. Third, matched diagnostics show higher correct-task coverage with dependency-aware acquisition than with risk-only selection without dependency modeling. Query cost is lower than for fixed-order or query-all controllers in the same target region.

Contributions.

- We formulate evidence acquisition and hierarchical release as a joint risk–coverage–cost problem with claim applicability, source dependence, unsupported-claim risk, and over-specificity risk.
- We specify EviWeave-Agent, a controller that uses a shared dependence representation for reliability-weighted fusion, active acquisition, and specificity selection.
- We evaluate the framework on CorruptEvidenceBench, a controlled benchmark with three task domains, five source domains, and six corruption and dependence conditions.
- We report policy-level calibration, matched baselines, locked-threshold ablations, and acquisition comparisons. Supplementary specifications, graph and budget stress scenarios, and eight cross-domain release cases clarify the mechanism and the limits of the supporting evidence (Appendices B–H).

2 Related Work

ReAct interleaves reasoning and actions (Yao et al., 2023), while self-consistency aggregates sampled reasoning paths (Wang et al., 2023). Retrieval-augmented generation grounds responses in retrieved documents (Lewis et al., 2020); Self-RAG learns when to retrieve and critique them (Asai et al., 2024). These mechanisms do not by themselves establish source independence. Correlated model errors (Kim et al., 2025) and knowledge corruption attacks on retrieval (Zou et al., 2025) illustrate distinct ways in which agreement or retrieved support can be misleading.

Source-reliability estimation (Dawid & Skene, 1979) and weak supervision (Ratner et al., 2017) combine imperfect observations; dependency modeling addresses correlations among weak sources (Varma et al., 2019). Active feature acquisition selects additional information for prediction (Li & Oliva, 2021), while hierarchical selective classification trades specificity for reliability (Goren et al., 2024). Risk-controlling prediction sets and conformal risk control calibrate decisions against specified losses (Bates et al., 2021; Angelopoulos et al., 2024). Our finite-family testing procedure follows the learn-then-test perspective (Angelopoulos et al., 2025). EviWeave-Agent coordinates these decisions through shared provenance: applicability determines usable support, dependence

determines its distinctness, and acquisition addresses the remaining evidence deficit. Calibration evaluates the complete acquisition-and-release policy.

3 Problem Formulation and Risk Objective

For each task $z = (q, x)$, the agent observes a question, initial context, and a set of available tools. It sequentially queries tools, then stops to release a node in a fixed answer hierarchy or abstains. The hidden target is a leaf y. A released node v is correct when y lies in the subtree rooted at v, and it is task-sufficient when its depth reaches the task requirement ℓ_{req}.

Each evidence record carries its value, scope, timestamp, tool identity, source domain, and provenance cluster. We distinguish *task domains* such as scientific evidence analysis from *source domains*, the fault domains used to model dependence and corruption. Records derived from a shared upstream event belong to the same provenance cluster before cross-source fusion. Distinct source identifiers alone therefore do not establish statistical independence of the evidence.

Let A denote non-abstention, T denote a task-sufficient answer, and V denote correctness. The primary sample-level indicators are

$$F = A(1 - V), \qquad G = TV, \qquad H = A(1 - T)V.$$

Thus $A = F + G + H$ on each sample: a released answer is either factually incorrect, correct and task-sufficient, or correct but too coarse for the task. We report answer coverage $C = \mathbb{P}(A = 1)$, task coverage $C_{\mathrm{task}} = \mathbb{P}(T = 1)$, correct-task coverage $C_{\mathrm{good}} = \mathbb{P}(G = 1)$, factual risk $R_F = \mathbb{P}(F = 1)$, and acquisition cost separately. These definitions imply $C_{\mathrm{good}} \le C_{\mathrm{task}} \le C$ and $R_F + C_{\mathrm{good}} \le C$.

Unsupported-claim risk R_U measures released conclusions lacking observed support that matches their scope and time. Over-specificity risk R_O measures answers whose specificity exceeds what the observed evidence justifies, even when a coarser ancestor remains correct. F, U, and O are not mutually exclusive: a factually wrong answer can be supported by a corrupted record, and a correct answer can still be unsupported by the observed evidence.

The policy-selection objective minimizes mean acquisition cost subject to per-cell upper bounds on R_F, R_U, and R_O, and lower bounds on C and C_{good}. Calibration applies to a complete frozen query–stop–release policy on independent benchmark roots. It is not a per-answer truth guarantee nor does it imply an instance-conditional error bound for every released answer.

4 EviWeave-Agent

Figure 1 summarizes the evidence-processing and query–release loop. EviWeave-Agent maintains a normalized evidence state rather than a flat list of records. For each candidate claim, records are canonicalized by entity, predicate, time interval, and upstream provenance. Strict copies and wrappers of the same upstream event update the same cluster, while distinct events within a source domain can still change that domain's score.

Reliability scores combine audit history, evidence age, scope match, conflict, source and task roles, missingness, effective source count, drift, and volatility. Unique events produce bounded within-domain support. Cross-domain fusion limits each domain's influence, and release requires corroboration across domains. Appendix C defines each feature and guard.

The supplement specifies these operations: a rule-based logistic reliability score produces bounded domain support, and weighted-median fusion caps each domain's weight at 0.45. Three fixed risk heads guide release and a fitted Q-iteration (FQI) controller (Ernst et al., 2005) trained on changes in risk-penalized answer utility, query cost, and provenance redundancy. Appendix C gives the equations, 57-dimensional state, training recipe, and inference algorithm. These decision scores are not per-answer error certificates.

The release module evaluates candidate answer nodes from coarse to specific. For each node, it predicts factual, unsupported-claim, and over-specificity losses. A frozen policy identifies the deepest node whose predicted losses meet its threshold and whose evidence state satisfies the source-support rule. The specified controller can continue querying when a positive query value indicates further

utility, even if a coarser node already passes. When it stops or reaches the horizon, it releases the deepest passing node or abstains. The statistical acceptance step described in Section 5 evaluates the complete frozen policy on independent calibration roots.

The threshold family contains 20 complete policies $P01$–$P20$. Each policy fixes the query controller, stopping rule, and release threshold before calibration. Development order is fixed before calibration; held-out test observations serve only for reporting. Appendix D lists all thresholds and their ordering, distinguishing a complete policy from a post-hoc threshold change.

5 Evaluation Protocol

5.1 Benchmark and Evaluation Units

CorruptEvidenceBench defines this study's controlled target population. It spans three task domains: open-domain investigation, scientific evidence analysis, and cybersecurity triage. Each task domain is evaluated under six threat conditions: missing evidence, stale evidence, schema-valid corruption, shared-upstream dependence, within-domain collusion, and adaptive corruption. In the main threat model, at most one of five source domains is compromised; source-domain identities are assumed to be correctly supplied by the benchmark. Appendix B supplies generation parameters, source roles, templates, the full hierarchy, and independent support-evaluation rules.

For each task domain, the benchmark contains 1,200 training roots, 400 development roots, 20,000 calibration roots, and 20,000 test roots. The six threat versions of the same root remain in the same split, so the three task domains yield 124,800 raw roots and 748,800 threat-conditioned episode views. We report extrema over the 90 held-out cells obtained from three task domains, six threat conditions, and five seed profiles. Risks are maxima, coverage metrics are minima, and cost is an equal-cell mean.

All methods share the answer hierarchy, a horizon of four paid queries, initial evidence budget, source-domain inventory, and task evaluator. The matched comparison includes ReAct with self-consistency, provenance-only aggregation, dependency-aware fusion with fixed query order, selective risk control without dependency modeling, and full EviWeave-Agent. Oracle rows, when reported, are privileged references and are not deployable policies. The supplemental component matrix and symbolic baseline definitions are in Appendix E; these do not specify a language-model implementation that independently verifies the ReAct-labeled aggregate.

5.2 Calibration and Reporting

The frozen family contains 20 complete policies evaluated at $\alpha \in \{0.05, 0.10, 0.20\}$ across five seed profiles, giving $20 \times 3 \times 5 = 300$ policy–target–seed combinations. For each combination, we test the 18 task-domain–threat cells against five endpoints: $R_F \leq \alpha$, $R_U \leq \alpha$, $R_O \leq \alpha$, $C \geq 0.70$, and $C_{\text{good}} \geq 0.85$. Exact one-sided binomial tests are used for each endpoint. An intersection–union rule (Berger, 1982) accepts a combination only when every declared endpoint passes. A Bonferroni threshold $q = 0.05/300 = 1/6000$ controls the probability of accepting any infeasible combination under the stated assumptions of independent and identically distributed (iid) roots. If no policy is accepted, the run is uncertified; an always-abstaining policy is not substituted.

The selected policy for each seed is the first accepted policy in the frozen development order. Test results do not change policy selection. Paired intervals use 2,000 stratified root-level bootstrap resamples (Efron, 1979), retaining the paired threat and method outcomes for each sampled root. These intervals summarize uncertainty in the reported ledger, not pointwise guarantees for individual answers. Appendix D gives the endpoint tests and false-acceptance argument. The supplemental 90-cell table in Appendix H is explicitly constructed to match original aggregates; it neither independently verifies those results nor recovers paired root-level outcomes.

6 RESULTS

6.1 MAIN COMPARISON AND CALIBRATION

EviWeave-Agent increases correct-task coverage and reduces query cost while meeting the displayed factual-risk and correct-task-coverage targets (Figure 2). Compared with selective risk control without dependency modeling, it increases correct-task coverage by 17.1 percentage points (0.861 vs. 0.690), lowers maximum factual risk by 0.9 points (0.038 vs. 0.047), and reduces mean query cost by 13.8% (2.31 vs. 2.68). Paired bootstrap intervals for EviWeave-Agent minus this baseline are $[0.165, 0.176]$ for correct-task coverage, $[-0.010, -0.008]$ for factual risk, and $[-0.371, -0.369]$ for cost.

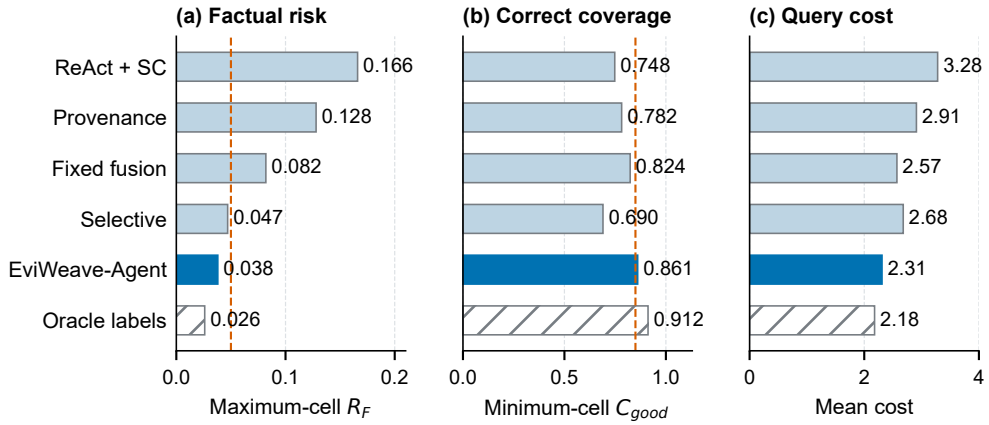


Figure 2: Held-out comparison at $\alpha = 0.05$. (a,b) Maximum factual risk and minimum correct-task coverage over 90 cells; the two extrema need not occur in the same cell. Dashed lines mark $R_F \leq 0.05$, $C_{\text{good}} \geq 0.85$, not acceptance under all calibration endpoints. (c) Mean query cost. ReAct: ReAct + self-consistency; Provenance: provenance-only aggregation; Fixed fusion: dependency-aware fusion with fixed query order; Selective: selective risk control without dependency modeling; EviWeave: full EviWeave-Agent; Oracle: oracle corruption labels (non-deployable reference). Complete values are in Figure A2.

Table 1: Three reported risk endpoints at $\alpha = 0.05$: factual (R_F), unsupported-claim (R_U), and over-specificity (R_O). Each column is a maximum over 90 held-out cells. Full profiles: Figure A2.

Method	R_F	R_U	R_O
ReAct + self-consistency	0.166	0.104	0.117
Provenance-only aggregation	0.128	0.075	0.084
Dependency-aware fusion, fixed query order	0.082	0.042	0.049
Selective risk control without dependency modeling	0.047	0.031	0.032
Full EviWeave-Agent	0.038	0.021	0.024
Oracle corruption labels (non-deployable)	0.026	0.014	0.016

Table 1 extends this comparison to all three reported risk endpoints. EviWeave-Agent yields $(0.038, 0.021, 0.024)$, compared with $(0.047, 0.031, 0.032)$ for selective control without dependency modeling. The reductions cover factual correctness, evidential support, and answer specificity. Each column summarizes its own worst cell; the three maxima neither define a joint error rate nor establish policy-level calibration acceptance.

The selective baseline maintains factual risk below 0.05 but fails the accompanying correct-task-coverage requirement. The oracle reference indicates the remaining gap with true corruption labels. Table 2 distinguishes policy-level calibration bounds from held-out results. At $\alpha = 0.05$, 13 of 100 candidate-policy–seed profiles are accepted, and P16 is selected for all five seed profiles. Larger risk budgets yield selected policies with lower query cost while retaining $C_{\text{good}} \geq 0.85$.

Table 2: Policy-level calibration bounds and held-out results across risk targets. Accepted counts use 20 candidate policies across five seed profiles. U_F, U_U, U_O are calibration upper bounds; L_{good} is the calibration lower bound.

			Calibration bounds				Held-out results		
α	Accepted	Policy	U_F	U_U	U_O	L_{good}	R_F	C_{good}	Cost
0.05	13/100	P16	0.044	0.025	0.029	0.856	0.038	0.861	2.31
0.10	43/100	P10	0.079	0.045	0.051	0.876	0.071	0.880	2.07
0.20	68/100	P05	0.112	0.063	0.072	0.877	0.103	0.881	1.79

6.2 Threat Conditions and Acquisition Cost

The selected policy remains within the displayed factual-risk and correct-task-coverage targets across all six threat conditions (Figure 3a–c). Adaptive corruption yields the overall maximum factual risk (0.038) and minimum correct-task coverage (0.861), while within-domain collusion has the highest mean query cost (2.75). These summaries use the same 90-cell ledger as Figure 2; source-domain identities are assumed correct, with at most one of five source domains compromised.

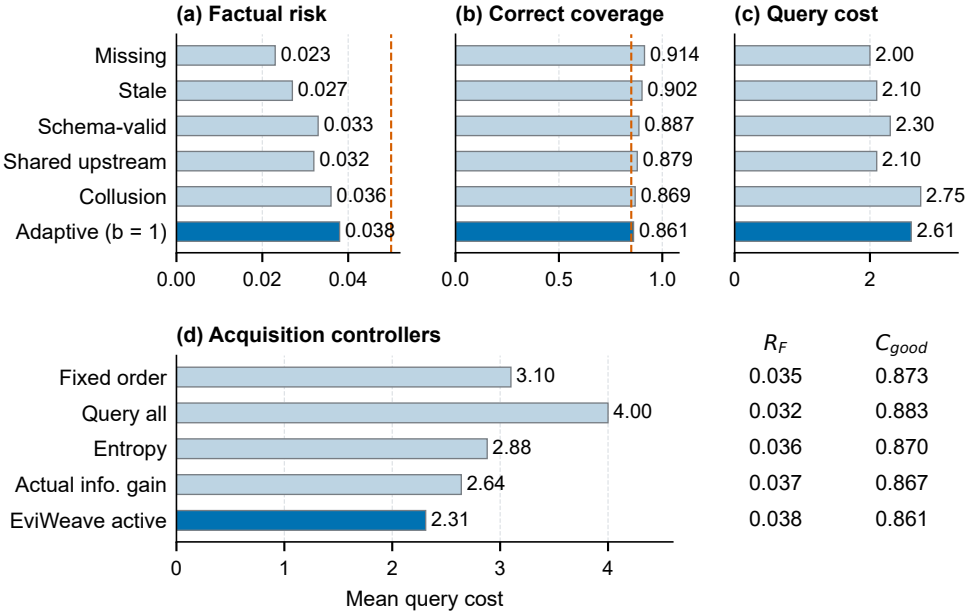


Figure 3: Threat conditions and acquisition cost. (a–c) Threat-wise maximum factual risk, minimum correct-task coverage, and mean cost for the selected EviWeave-Agent policy. Dashed reference lines mark 0.05 risk and 0.85 coverage; $b = 1$ denotes the one-source-domain corruption budget. (d) A separate controller comparison within the common target region $R_F \leq 0.05$, $C_{\text{good}} \geq 0.85$; adjacent columns report each controller's risk and coverage. Query-all has lower risk and higher coverage than EviWeave active, at higher cost. Figure A3 expands these comparisons.

Within the common target region, EviWeave active has lower mean acquisition cost than the fixed-order, query-all, entropy, and actual-information-gain controllers (Figure 3d). Query-all achieves lower risk and higher coverage, so this comparison supports lower cost within that region, rather than superiority across every reported metric.

6.3 Component Diagnostics

Removing individual components reveals distinct failure modes (Figure 4). Removing dependency handling increases maximum factual risk from 0.038 to 0.106; removing active acquisition increases mean cost from 2.31 to 2.84; removing specificity control raises over-specificity risk from 0.024 to

0.092. The uncalibrated variant has higher correct-task coverage (0.869 vs. 0.861) and lower cost (1.94 vs. 2.31), but also higher factual risk (0.071 vs. 0.038).

The reported unsupported-claim risk is 0.021 for the full system, rising to 0.059, 0.046, 0.031, 0.044, and 0.058 without dependency modeling, reliability estimation, active acquisition, specificity control, and calibration, respectively (Figure A1). These are locked-threshold diagnostics, without separate recalibration.

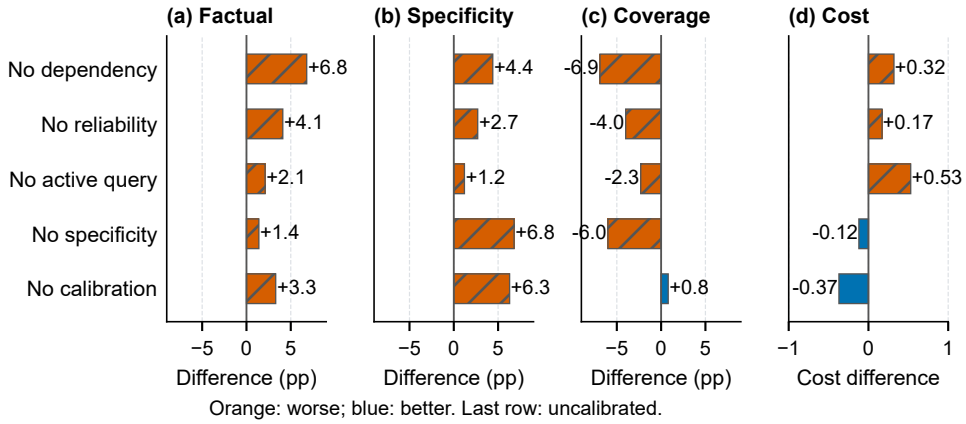


Figure 4: Locked-threshold component diagnostics relative to full EviWeave-Agent. Each bar shows the difference between a variant and the full system; zero marks the full-system reference. Risk and coverage differences are in percentage points (pp); cost differences retain query-cost units. Orange marks a worse value on the displayed metric and blue a better value. The final row is uncalibrated. These interventions are not separately recalibrated policies. Figure A1 reports absolute values, unsupported risk, and calibration status.

6.4 Fixed-Query Evidence Control

Evidence processing improves the reported outcomes even when the agent cannot change which queries it makes (Figure 5). The local Greedy audit holds the controller, query trajectory, and mean acquisition cost fixed. Each configuration is evaluated on 30,000 roots. The intervention jointly applies applicability filtering, provenance deduplication, and reliability weighting to the available evidence, while the comparison configuration uses raw evidence. This separates the combined evidence-processing intervention from changes in the number or cost of queries.

Correct-task coverage increases from 0.675 to 0.937, while factual risk falls from 0.324 to 0.061. The other risk endpoints also decrease: unsupported-claim risk falls from 0.215 to 0.028, and over-specificity risk from 0.241 to 0.043. Both configurations retain answer coverage $C = 1.000$ and mean query cost 2.70. Thus, factual errors decrease without reducing the number of answered tasks or adding paid queries in this audit.

The remaining factual risk, 0.061, still exceeds the 0.05 reference. Evidence processing alone in this fixed-controller comparison therefore does not establish that the main policy's risk target is met. This local diagnostic is separate from the matched 90-cell complete-system comparison. Its difference from EviWeave-Agent's reported risk cannot isolate the causal benefit of active acquisition, nor does the joint intervention isolate each processing component's contribution.

6.5 Hierarchical Release

EviWeave-Agent can preserve a supported coarser answer when a fine-grained claim fails its release threshold (Figure 6). In the recorded case, the family-level claim exceeds the predicted-loss threshold, but the command-and-control node passes. The agent releases this C2-level answer rather than abstaining or asserting the unsupported malware-family label. If a task requires the family-level label, this depth-3 answer does not count toward C_{good}.

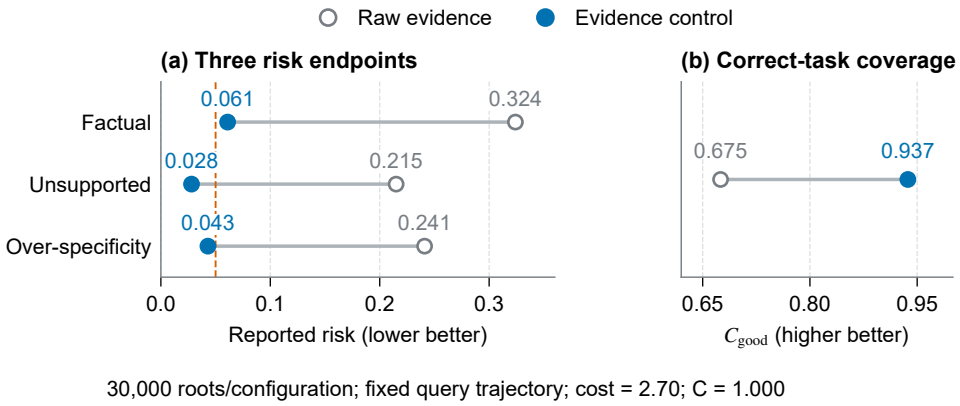


Figure 5: Fixed-Greedy audit of combined applicability filtering, deduplication, and reliability weighting. Lines connect configuration summaries, not individual tasks. Query trajectory, cost, and C are fixed. Factual risk remains above the dashed 0.05 reference at 0.061. The comparison does not isolate individual processing components.

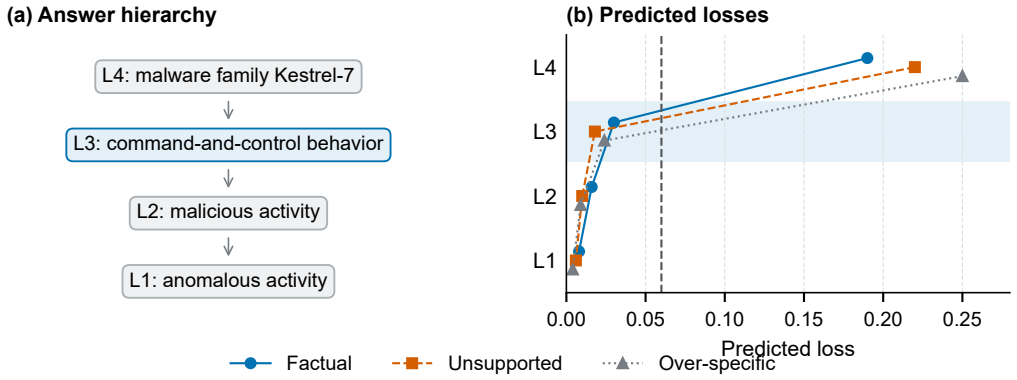


Figure 6: Hierarchical release. (a) C2 is the deepest passing answer; the family node is not released. (b) Predicted factual (F), unsupported (U), and over-specificity (O) losses versus $\tau = 0.06$. These single-case scores guide release; statistical acceptance applies to the complete policy, not each answer. Full scores: Table A1.

6.6 Coverage, Specificity, and Abstention

The auxiliary metrics distinguish task-depth sufficiency, leaf accuracy, and answer reliability (Figure 7). Answer-conditioned factual risk uses released answers as its denominator, whereas unconditional factual risk uses all tasks. These metrics do not form mutually exclusive outcome categories.

Among the four compared methods, EviWeave-Agent has the lowest reported answer-conditioned factual risk, 0.042, and the highest evidence-sufficiency accuracy, 0.934. Relative to selective risk control without dependency modeling, its task coverage is higher (0.888 vs. 0.727), its abstention rate is lower (0.086 vs. 0.258), and its answer-conditioned factual risk is lower (0.042 vs. 0.063). Lower abstention alone does not establish correctness. Separately reported extrema neither identify the fraction of tasks changing outcome nor recover answer-conditioned risk by division.

The gain does not extend uniformly to the finest answer level. EviWeave-Agent's leaf accuracy is 0.702, below the independent-source reliability method's 0.746, despite its higher evidence-sufficiency accuracy (0.934 vs. 0.876). Together with the main comparison, this supports assessing useful answers at the required task depth. Neither these aggregate metrics nor the single case in Figure 6 isolates the causal effect of hierarchical release from the rest of the policy.

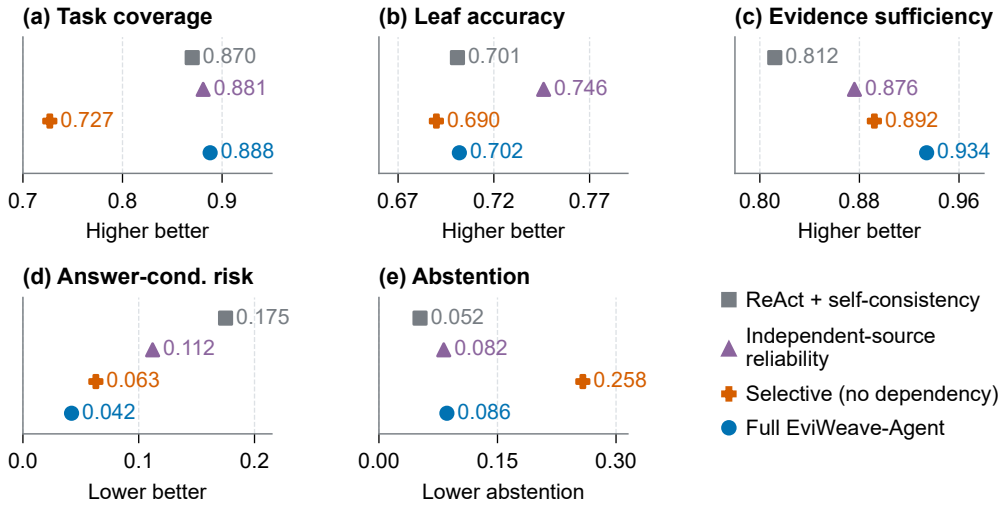


Figure 7: Auxiliary diagnostics at $\alpha = 0.05$, with consistent method order. These metrics do not form mutually exclusive outcome categories. EviWeave-Agent has the lowest answer-conditioned factual risk and highest evidence-sufficiency accuracy, but not the highest leaf accuracy. Lower abstention alone does not establish greater correctness.

The fixed-query audit, acquisition comparison, and release cases examine evidence processing, query cost, and answer depth; their effects are not additive or substitutes for complete-policy calibration. Appendix A retains full diagnostic profiles and exact release scores. Appendix G adds eight cross-domain cases covering abstention, incorrect release, and semantically unsupported correct answers. Appendix F supplies graph, budget, and drift scenarios with their evidence qualifications.

7 Discussion and Conclusion

EviWeave-Agent shares source-dependence information across fusion, acquisition, and answer specificity. At $\alpha = .05$, its reported maximum-cell factual risk is .038, minimum correct-task coverage is .861, and mean query cost is 2.31. Correct coarser answers count as task successes only when they reach the required depth.

Results support coordination under correct source identities, at most one compromised domain among five, and complete frozen policies on iid roots. Calibration provides policy-level statistical assurance rather than individual-answer correctness guarantees. Preset scenarios and the aggregate-matched reconstructed ledger offer mechanism and consistency checks (Appendix F). Empirical tolerance thresholds and independent replication remain validation priorities for graph errors, larger corruption budgets, and source drift.

AI Use Statement

OpenAI ChatGPT was used solely for grammar correction and language polishing of author-written text throughout the manuscript.

References

Anastasios Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control. In *International Conference on Learning Representations*, pp. 55198–55218, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/f3549ef9b5ff520a7e41ff3cc306ab2b-Abstract-Conference.html.

Anastasios N. Angelopoulos, Stephen Bates, Emmanuel J. Candès, Michael I. Jordan, and Lihua Lei. Learn then test: Calibrating predictive algorithms to achieve risk control. *The Annals of Applied Statistics*, 19(2):1641–1662, 2025. doi: 10.1214/24-AOAS1998. URL https://doi.org/10.1214/24-AOAS1998.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avi Sil, and Hannaneh Hajishirzi. Self-RAG: Learning to retrieve, generate, and critique through self-reflection. In *International Conference on Learning Representations*, pp. 9112–9141, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/25f7be9694d7b32d5cc670927b8091e1-Abstract-Conference.html.

Stephen Bates, Anastasios N. Angelopoulos, Lihua Lei, Jitendra Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *Journal of the ACM*, 68(6):1–34, 2021. doi: 10.1145/3478535. URL https://doi.org/10.1145/3478535.

Roger L. Berger. Multiparameter hypothesis testing and acceptance sampling. *Technometrics*, 24 (4):295–300, 1982. doi: 10.1080/00401706.1982.10487790. URL https://doi.org/10.1080/00401706.1982.10487790.

C. J. Clopper and E. S. Pearson. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*, 26(4):404–413, 1934. doi: 10.1093/biomet/26.4.404. URL https://doi.org/10.1093/biomet/26.4.404.

A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28, 1979. doi: 10.2307/2346806. URL https://doi.org/10.2307/2346806.

Bradley Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7 (1):1–26, 1979. doi: 10.1214/aos/1176344552. URL https://doi.org/10.1214/aos/1176344552.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(18):503–556, 2005. URL https://jmlr.org/papers/v6/ernst05a.html.

Shani Goren, Ido Galil, and Ran El-Yaniv. Hierarchical selective classification. In *Advances in Neural Information Processing Systems*, volume 37, pp. 111047–111073. Curran Associates, Inc., 2024. doi: 10.52202/079017-3526. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/c8b100b376a7b338c84801b699935098-Abstract-Conference.html.

Elliot Myunghoon Kim, Avi Garg, Kenny Peng, and Nikhil Garg. Correlated errors in large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 30038–30066. PMLR, 2025. URL https://proceedings.mlr.press/v267/kim25e.html.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper/2020/hash/6b493230205f780e1bc26945df7481e5-Abstract.html.

Yang Li and Junier Oliva. Active feature acquisition with generative surrogate models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 6450–6459. PMLR, 2021. URL https://proceedings.mlr.press/v139/li21p.html.

Alexander Ratner, Stephen H. Bach, Henry Ehrenberg, Jason Fries, Sen Wu, and Christopher Ré. Snorkel: Rapid training data creation with weak supervision. *Proceedings of the VLDB Endowment*, 11(3):269–282, 2017. doi: 10.14778/3157794.3157797. URL https://www.vldb.org/pvldb/vol11/p269-ratner.pdf.

Paroma Varma, Frederic Sala, Ann He, Alexander Ratner, and Christopher Ré. Learning dependency structures for weak supervision models. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 6418–6427. PMLR, 2019. URL https://proceedings.mlr.press/v97/varma19a.html.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Wei Zou, Runpeng Geng, Binghui Wang, and Jinyuan Jia. PoisonedRAG: Knowledge corruption attacks to retrieval-augmented generation of large language models. In *34th USENIX Security Symposium (USENIX Security 25)*, pp. 3827–3844. USENIX Association, 2025. URL https://www.usenix.org/conference/usenixsecurity25/presentation/zou-poisonedrag.

A Complete Reported Results

These figures expand the component, system, threat, and controller comparisons introduced in Section 5. Table A1 retains the exact hierarchical release scores.

A.1 Absolute Component Diagnostics

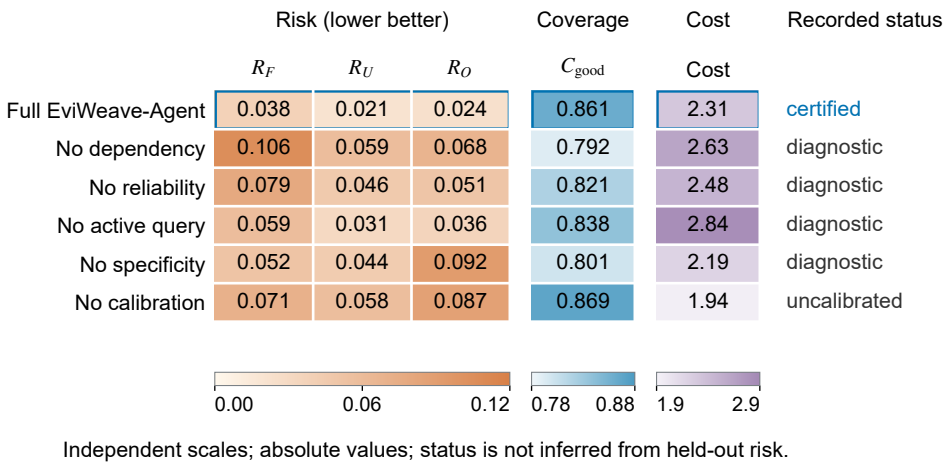


Figure A1: Absolute locked-threshold component diagnostics at $\alpha = 0.05$. Risks share a scale; coverage and mean cost have separate scales. Numbers retain reported precision. The outlined EviWeave-Agent row is the only system accepted by the reported calibration (status: certified). Other variants are labeled as diagnostics without separate recalibration or as uncalibrated. Acceptance is distinct from held-out constraint satisfaction. Figure 4 reports differences from the full system.

A.2 Complete System Profile

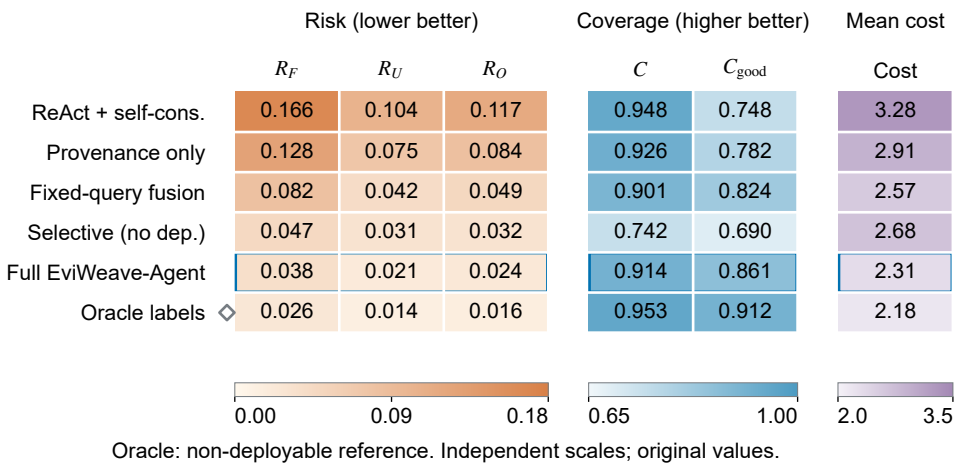


Figure A2: Complete held-out system comparison at $\alpha = 0.05$; methods follow Figure 2. Risks are cross-cell maxima, coverage metrics are minima, and cost is the equal-cell mean over the 90-cell ledger. The three risk columns share a scale; coverage and cost use separate scales. Numbers retain the original precision without column normalization. The blue outline identifies EviWeave-Agent, while the open diamond marks the non-deployable oracle reference. Neither colors nor outlines denote calibration acceptance.

A.3 Threat Margins and Controller Tradeoffs

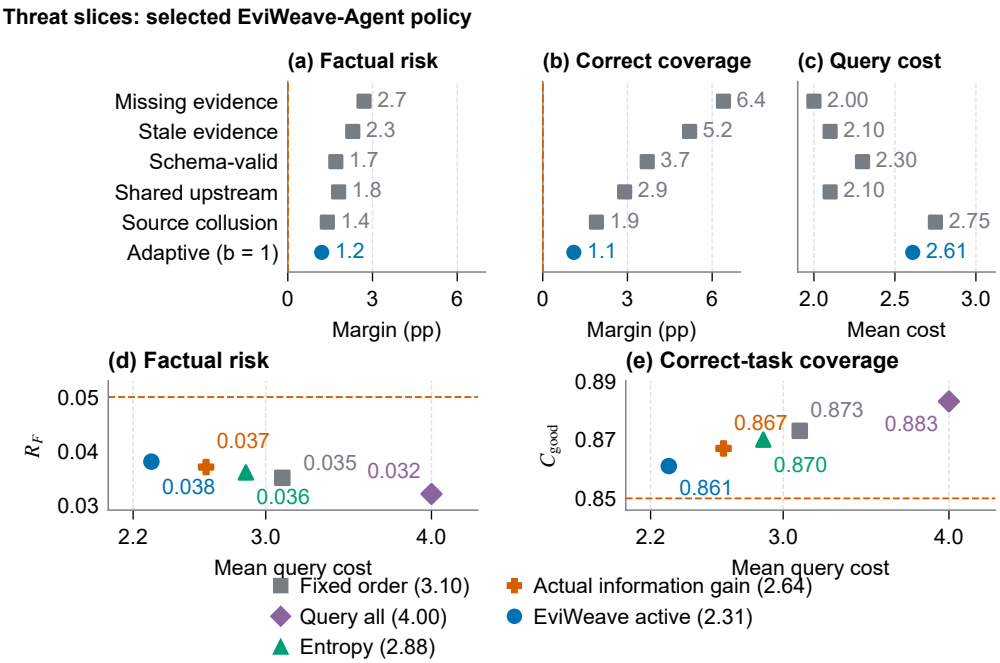


Figure A3: Two distinct comparisons at $\alpha = 0.05$. (a–c) Threat slices of the selected EviWeave-Agent policy on the same 90-cell ledger as Figure A2. Factual-risk and correct-task-coverage margins are $100(0.05 - R_F)$ and $100(C_{\text{good}} - 0.85)$ in percentage points (pp). Positive margins indicate distance from the displayed test targets, not confidence bounds or calibration guarantees. Cost retains its original units. Source collusion occurs within one domain; $b = 1$ denotes at most one of five source domains compromised. (d–e) A separate controller comparison in the common target region. Point labels give risk or coverage; parenthesized legend values give mean cost. Query-all has lower risk and higher coverage than EviWeave active, at higher cost. The discrete points do not define a continuous performance frontier.

A.4 Exact Hierarchical Release Scores

Table A1: Single-case scores guide release, while statistical acceptance is at the complete-policy level.

Depth	Candidate claim	Est. F	Est. U	Est. O	Passes $\tau = 0.06$
4	malware family Kestrel-7	0.190	0.220	0.250	no
3	command-and-control behavior	0.030	0.018	0.024	yes
2	malicious activity	0.016	0.010	0.009	yes
1	anomalous activity	0.008	0.006	0.004	yes

B CorruptEvidenceBench Construction

Scope of the supplement. The September 25 supplemental records specify the benchmark, method, policy family, and cases below. We distinguish this instantiation from the original manuscript's aggregate results. The records explicitly describe constructing cell counts to match existing aggregates and calibration bounds. These tables reconstruct the aggregates for consistency; they do not independently show that this instantiation reproduces the original results. Appendix H retains all 90 supplied rows with this qualification; the stress scenarios in Appendix F do not extend the original calibration claim.

B.1 Generation Pipeline and Evaluation Unit

A root contains a domain, question, entity, query time, hidden leaf, required depth, and latent source events. Roots are assigned to splits before the six threat views are generated. Each view changes evidence but preserves the hidden leaf, answer tree, and task requirement. The pipeline is

$$\text{root/split} \rightarrow \text{latent events} \rightarrow \text{threat views} \rightarrow \text{tool observations} \rightarrow \text{release/evaluation}.$$

Each domain has a four-level tree with 16 uniformly sampled leaves. Required depth $L = \ell_{\text{req}}$ is sampled independently: $\Pr(L = 1, 2, 3, 4) = (.05, .15, .45, .35)$. Query time is zero on a relative time axis. The time scales are $W = 30$ days for open-domain investigation (OD), 365 days for scientific evidence analysis (SCI), and one day for cybersecurity triage (CYB). These controlled classification tasks use domain-specific terminology; they do not establish transfer to arbitrary natural-language investigation or deployed systems.

B.2 Task Domains, Templates, and Hierarchies

OD classifies operational events, SCI attributes sample or measurement-system signals, and CYB classifies security behavior. Every node denotes the union of leaves in its subtree. A coarser answer makes no causal or universal claim about all members of a category. Table B1 translates the three supplied template families per domain; entity E and required depth L are substituted without revealing the leaf. Table B2 lists all 48 leaf paths.

Table B1: Question templates, all referring to entity E at time zero.

Domain	Three templates
OD	(1) What category describes the operational event of E? Required depth: L. (2) Use currently valid records to classify E's event to at least level L. (3) Investigate E's operational anomaly; give a coarser answer or abstain if evidence is insufficient. Target depth: L.
SCI	(1) What category explains the signal of sample E? Required depth: L. (2) Use applicable measurements to classify E's signal mechanism at level L or deeper. (3) Synthesize evidence for E, distinguishing sample signals from measurement-system signals. Target depth: L.
CYB	(1) What security category describes E's behavior? Required depth: L. (2) Use valid telemetry to classify E's behavior to at least level L. (3) Analyze E's security event; allow a coarse answer or abstention if evidence is insufficient. Target depth: L.

B.3 Source Domains and Tool Interface

A source domain is an upstream fault domain; a return tool delivers a record. Records carry `return_tool`, `origin_domain`, and `upstream_event_id`. A D3 wrapper of a D1 event retains D1 and its original event ID. Different tool names alone do not establish independent support. Table B3 gives specified priors, not estimates of real-world institutional reliability. Initial evidence consists of one record from each of D1 and D2. Paid actions D2–D5 can each be used once at unit cost, with three event slots per action. Paid D2 retrieves new events beyond the initial record. Slots can be empty or repeat an upstream event. Thus the horizon and query-all cost are four. Cost counts paid external queries, excluding training, model inference, tokens, and local computation. Query savings therefore do not measure total-system savings.

B.4 Clean Evidence and Six Threats

For an independent event from source d, the reported latent leaf is correct with probability a_d; otherwise, it is uniform over the other 15 leaves. Claim depth has probabilities $(.05, .10, .35, .50)$ for levels 1–4; the record asserts the corresponding ancestor. Strength is uniform on $[.94, 1.00]$. An incorrect leaf can have a correct ancestor. Clean records have validity $[-W, .1W]$ and age uniformly distributed on $[0, .1W]$.

Required fields are entity, predicate (`event_class`), claim, validity endpoints, upstream event ID, origin domain, return tool, and strength. Optional fields are artifact hash, author version, cache timestamp, and parser confidence. The fraction of missing optional fields is an input feature; the supplied specification does not give a separate optional-field missingness distribution. Schema-

Table B2: All 48 leaf paths. Each row contains two sibling leaves sharing the displayed depth-1–3 path; there are 16 leaves per domain.

Domain	Depth 1	Depth 2	Depth 3	Depth 4 leaves
OD	Internal operation	Software change	Configuration change	Config-A, Config-B
OD	Internal operation	Software change	Version release	Release-A, Release-B
OD	Internal operation	Equipment fault	Sensor component	Sensor-A, Sensor-B
OD	Internal operation	Equipment fault	Mechanical component	Motor-A, Motor-B
OD	External supply	Logistics anomaly	Transit delay	Transit-A, Transit-B
OD	External supply	Logistics anomaly	Storage error	Storage-A, Storage-B
OD	External supply	Energy anomaly	Power disturbance	Power-A, Power-B
OD	External supply	Energy anomaly	Fuel shortage	Fuel-A, Fuel-B
SCI	Sample signal	Optical process	Fluorescence	Fluor-A, Fluor-B
SCI	Sample signal	Optical process	Scattering	Scatter-A, Scatter-B
SCI	Sample signal	Chemical process	Oxidation	Oxid-A, Oxid-B
SCI	Sample signal	Chemical process	Polymerization	Poly-A, Poly-B
SCI	Measurement system	Electronic artifact	Gain drift	Gain-A, Gain-B
SCI	Measurement system	Electronic artifact	Sampling alias	Alias-A, Alias-B
SCI	Measurement system	Optical-path artifact	Stray light	Stray-A, Stray-B
SCI	Measurement system	Optical-path artifact	Focus shift	Focus-A, Focus-B
CYB	Anomalous activity	Malicious activity	C2	Kestrel-7, Finch-3
CYB	Anomalous activity	Malicious activity	Exfiltration	Heron-2, Kite-4
CYB	Anomalous activity	Benign anomaly	Monitoring false alarm	Monitor-M1, Monitor-M2
CYB	Anomalous activity	Benign anomaly	Transient load burst	Burst-B1, Burst-B2
CYB	Routine activity	User workload	Web traffic	Web-W1, Web-W2
CYB	Routine activity	User workload	Batch processing	Batch-J1, Batch-J2
CYB	Routine activity	Scheduled maintenance	Backup	Backup-BK1, Backup-BK2
CYB	Routine activity	Scheduled maintenance	Patching	Patch-P1, Patch-P2

Table B3: Source roles and reliability priors in the supplemental specification.

Source	OD	SCI	CYB	a_d
D1	Operational archive	Primary measurements	Endpoint events	.94
D2	Independent telemetry	Second instrument	Network telemetry	.92
D3	Change tickets	Process records	Sandbox behavior	.90
D4	Independent audit	Remeasurement	Incident audit	.88
D5	Archival index	Archival metadata	Sample metadata	.86

Table B4: Threat transformations under at most one compromised source domain.

Condition	Transformation
Missing evidence	Uniformly select one domain; each of its returned slots is empty with probability .60. Empty slots are not negative evidence.
Stale evidence	Shift one domain's records into the past by $2W$, $4W$, or $8W$ with probabilities .50, .30, .20. Validity ends before zero.
Schema-valid corruption	In one domain, replace event claims with a common incorrect leaf with probability .80, retaining valid types, entity, and current timestamps.
Shared upstream	Two return tools wrap one selected upstream event. Multiplicity is 1, 2, or 4 with probabilities .25, .50, .25; the upstream malicious switch has probability .50.
Within-domain collusion	Three distinct event IDs in one malicious domain support the same incorrect leaf. Event deduplication alone cannot remove domain-level dependence.
Adaptive corruption	On a query to the malicious domain, enumerate 15 incorrect leaves, create three applicable records, and select the leaf minimizing its summed predicted losses; ties use leaf ID.

valid corruption and collusion fix the wrong leaf when the threat view is created. The adaptive attacker knows the current root's true leaf, question, public tree, frozen risk heads and threshold, and observed history. The supplement also lists access to frozen FQI parameters, although the stated attack objective depends on the risk heads. It cannot read unqueried clean returns, other roots' labels, or calibration p-values, or alter ground truth, other domains, or source identities. Attack rules are frozen before calibration, and state resets for each root. Adaptation occurs within episodes, not across the calibration sample.

B.5 Ground Truth and Independent Support Evaluation

The evaluator checks the hidden leaf and required depth as in Section 3. It separately uses audited claim, entity, and validity annotations to identify what an observed record semantically supports. These fields are hidden from EviWeave-Agent. Independent over-parsing of claim granularity occurs with probability .08: audited depth is uniform over 1–3, while the agent parses a leaf on that branch. Scope mis-parsing occurs independently with probability .03: the parsed entity matches, but the audited entity differs. A corrupted record can semantically support its false assertion.

Let $N(v)$ count observed records with matching audited entity, predicate, and time whose audited claim equals or descends from v. Let $d_2(v)$ be the greatest ancestor depth on the path to v supported by at least two true upstream domains with record strength at least .80, or zero if none. The supplemental evaluator uses

$$U = A\mathbf{1}\{N(v) = 0\}, \qquad O = A\mathbf{1}\{\mathrm{depth}(v) > d_2(v)\}.$$

Its two-domain criterion operates on audited support; the agent's three-domain release guard operates on parsed support. Both $F = 0, U = 1, O = 1$ and $F = 0, U = 0, O = 1$ are therefore possible. Correct-task success $G = TV$ does not additionally require $U = O = 0$. Appendix G illustrates why all three risks are needed.

B.6 Splits, Seeds, and Leakage Prevention

Table B5: Root counts; threat views and seed profiles reuse roots within splits.

Domain	Train	Development	Calibration	Test	Total
OD	1,200	400	20,000	20,000	41,600
SCI	1,200	400	20,000	20,000	41,600
CYB	1,200	400	20,000	20,000	41,600
Total	3,600	1,200	60,000	60,000	124,800

Six threats yield 748,800 views. Profiles $11, 23, 37, 53, 97$ use the same roots, rather than five times as many independent roots. Every domain–threat–seed cell has denominator 20,000; correlated views do not increase the binomial sample size.

The generator seed is 1925. PCG64 is initialized from the first 64 bits of SHA-256 of `1925|domain|split|root_index`. Root namespaces prevent entity, event, and cache reuse across splits. Cache keys include root, threat, source, and event. Model inputs exclude split IDs, internal counters, corruption labels, audited support, and unqueried contents. Trees and template families are shared across splits; evaluation concerns new roots within a fixed taxonomy, not unseen templates or categories. Pseudorandom seeding specifies how to regenerate the data, while the statistical argument still assumes iid latent roots and a frozen policy.

B.7 An End-to-End Episode

SCI-01 requests sample SCI-01's signal category at time zero, with a required depth of four. Its true path is sample signal, optical process, fluorescence, Fluor-A. Initial records from D1 and D2 support Fluor-A. Paid D3 wraps the D1 event; paid D4 supplies an independent event. Four records therefore yield three clusters and three domains. Terminal Fluor-A scores $(.049737, .030032, .047426)$ meet $\tau = .06$, so the agent releases the required leaf. Appendix G gives full records and contrasting failures.

C EviWeave-Agent Specification

This section formalizes the supplemental rule-based instantiation and its fitted Q-iteration (FQI) procedure. Reliability and risk-head coefficients are explicit constants; FQI coefficients are learned. Trained weights are not included in the supplied attachment. The specification makes these rules explicit without claiming to reconstruct the original trained policy.

C.1 Applicability and Event Support

Filter records whose parsed entity or predicate does not match the question or whose validity does not contain zero. Within the root namespace, deduplicate by (entity, upstream event ID, valid-from, valid-to). Identical wrappers map to one event; conflicting contents under one key raise an explicit error rather than retaining the last record. Distinct versions require distinct keys. For strength h_k and parsed claim c_k,

$$e_k(v) = \begin{cases} h_k, & c_k = v \text{ or } c_k \text{ descends from } v, \\ 0, & c_k \text{ is a strict ancestor of } v, \\ -h_k, & c_k \text{ and } v \text{ are on incompatible branches.} \end{cases}$$

C.2 Reliability and Within-Domain Aggregation

For event k in domain d, use the rule-based logistic score

$$\begin{aligned} r_k(v) = \sigma\big[&\operatorname{logit}(a_d) - 1.5t_k - 1.2x_v - m_k - .8s_k - .7u_k \\ &+ b_d + b_{\text{task}} + .15\log(1 + n_{\text{obs}})\big]. \end{aligned} \tag{1}$$

Here σ is the logistic function, t_k is age divided by W, m_k is the fraction of four optional fields missing, and s_k, u_k are historical drift and volatility summaries available before the episode. Both summaries are zero in the reference setting and remain fixed under drift stress. Domain biases are $(.10, .05, 0, -.05, -.10)$ for D1–D5; task biases are $(0, -.05, -.10)$ for OD, SCI, CYB. Priors a_d are in Table B3.

First compute x_v, the fraction of observed domains whose unweighted mean $e_k(v)$ is below $-.25$, to avoid a circular definition. n_{obs} counts observed upstream domains after deduplication, using provenance visible to the agent. Wrapper names do not change the source-role bias. For nonempty sets of unique events K_d,

$$s_d(v) = \frac{\sum_{k \in K_d} r_k(v) e_k(v)}{\sum_{k \in K_d} r_k(v)}, \qquad \bar{r}_d(v) = \frac{1}{|K_d|} \sum_{k \in K_d} r_k(v).$$

Thus $s_d(v) \in [-1, 1]$. New events can change domain support without adding cross-domain votes.

C.3 Capped Fusion and Its Limited Guarantee

With at least three observed domains, choose λ such that

$$w_d(v) = \min\{.45, \lambda \bar{r}_d(v)\}, \qquad \sum_d w_d(v) = 1.$$

The specified bisection uses $[0, 10^6]$ and at most 80 iterations. The weighted median is the smallest sorted s_d whose cumulative weight reaches .5. Define

$$S_v = \frac{1 + \operatorname{wmed}_w(s_d(v))}{2}, \qquad X_v = \sum_d w_d(v) \mathbf{1}\{s_d(v) < -.25\}.$$

$$m_v = \sum_d \mathbf{1}\{s_d(v) \geq .25, \bar{r}_d(v) \geq .5\}.$$

With fewer than three observed domains, normalized reliability is used only for diagnostics; release is prohibited. Without evidence, the agent abstains. The release guard requires $m_v \geq 3$, not merely three observed domains.

Proposition 1 (Single-domain interval preservation) *If at most one observed domain is malicious, its weight is at most .45, and every nonmalicious domain score lies in $[l, u]$, the weighted median lies in $[l, u]$.*

Proof. Weight strictly below l or above u is at most .45. Neither side can contain a weighted majority. □

This result concerns support scores, not factual risk. Measurements from nonmalicious domains can be incorrect, and two malicious domains can exceed half the total weight. Adding strict copies cannot change the unique event set, domain scores, or terminal release. This invariance does not cover corrupted identities or distinct colluding events.

C.4 Three Decision Scores and Release

Let B_v indicate applicable parsed support, M be mean unique-event missingness, s mean historical drift, $D_v = \text{depth}(v)/4$, and $Z_v = \min(1, m_v/3)$. Let $d_3(v)$ be the deepest ancestor on the path to v with three supporting domains, or zero if none, and $g_v = \max(0, \text{depth}(v) - d_3(v))/4$. The supplied heads are

$$\widehat{R}_F(v) = \sigma[-4 + 5(1 - S_v) + 3X_v + D_v + 1.5(1 - Z_v) + M + s], \tag{2}$$

$$\widehat{R}_U(v) = \sigma[-4.5 + 5(1 - B_v) + 2.5(1 - S_v) + D_v + M], \tag{3}$$

$$\widehat{R}_O(v) = \sigma[-4.5 + 6g_v + 3X_v + 1.5D_v + (1 - Z_v)]. \tag{4}$$

These are decision scores, not instance-conditional error certificates. A node passes when all three scores are at most τ and $m_v \geq 3$. Select the deepest passing node, resolving ties by summed scores, then by lexicographic node ID; otherwise the candidate output is abstention. The internal threshold differs from the risk target: P16 has $\tau = .06$ at the reported $\alpha = .05$ point.

C.5 Acquisition State and Training Recipe

At each depth, select the node with minimum summed loss and encode $(S, X, m/5, B, M, g, \widehat{R}_F, \widehat{R}_U, \widehat{R}_O, n_{\text{obs}}/5)$. Forty features across four depths are followed by five source-observed bits, four action-used bits, remaining horizon divided by four, four required-depth bits, and three domain bits, giving 57 dimensions. The missing-depth code is $(.5, 0, 0, 0, 0, 1, 1, 1, 1, 0)$, although every depth exists in these trees. The records do not specify tie-breaking between equal-loss nodes during state encoding.

For candidate release $v_\tau(s)$, let $\Phi_\tau(s) = 0$ on abstention; otherwise

$$\Phi_\tau(s) = \max\left\{0, \mathbf{1}\{\text{depth}(v_\tau) \geq L\} + .4\frac{\text{depth}(v_\tau)}{4} - 4\widehat{R}_F - 2\widehat{R}_U - 2\widehat{R}_O\right\}.$$

The redundancy prior in Table C1 assigns probabilities to wrapping the most recent event from each source. For observed domains $\mathcal{D}(s)$, $D(a, s) = \sum_{d \in \mathcal{D}(s)} P_{ad}$ and

$$r_\tau(s, a, s') = \Phi_\tau(s') - \Phi_\tau(s) - .04\ \text{cost}(a) - .10D(a, s).$$

This reward combines potential improvement with cost and redundancy penalties; it is not a separately fitted predictor of true risk reduction. All actions cost one; unequal prices would require revising the objective and calibrating the resulting policy. Training uses uniform random four-action

Table C1: Fixed redundancy prior; remaining mass represents a new event.

Action	D1	D2	D3	D4	D5	New event
D2	.12	.08	.04	.03	.03	.70
D3	.05	.04	.03	.04	.04	.80
D4	.05	.04	.03	.04	.04	.80
D5	.03	.03	.03	.03	.03	.85

permutations over training roots, six threats and five profiles: $3 \times 1200 \times 6 \times 5 \times 4 = 432{,}000$

specified transitions per policy. Tensoring one-hot actions with the state and intercept gives 232 features. Initialize $Q^{(0)} = 0$ and perform four Bellman regressions with discount one:

$$y_i^{(k)} = r_i + \max\{0, \max_{a' \text{ feasible}} Q^{(k-1)}(s_i', a')\}.$$

Terminal transitions have zero future value. Each regression minimizes mean squared error plus $.001\|\beta\|_2^2$, regularizing all 232 coefficients. STOP has value zero and no fitted coefficients. Previously used actions and actions beyond the horizon are masked. Equal query values are resolved in D2, D3, D4, D5 order. Each threshold defines its own complete policy.

C.6 Inference and Stopping

A passing coarse answer does not force an immediate stop: a positive query value can justify acquiring further evidence to meet the task requirement. Terminal case calculations document processing and release; reproducing acquisition additionally requires trained weights and action logs.

Input: question, entity, required depth L, initial evidence E_0, actions $\mathcal{T} = \{D2, D3, D4, D5\}$, frozen (Q_τ, τ).
Initialize: $E \leftarrow E_0$, used actions $\mathcal{U} \leftarrow \varnothing$, remaining horizon $h \leftarrow 4$.

1. Filter applicability and canonicalize unique upstream events.
2. For every node compute event support, reliability, domain scores, capped fusion, and the three predicted losses.
3. Set v to the deepest passing node under the release tie rules, or abstention.
4. If $h = 0$ or no unused action remains, return v.
5. Encode the state and compute feasible $Q_\tau(s, a)$. If the maximum is at most zero, return v.
6. Query the maximizing action, with ties in D2–D5 order; append its records to E, mark the action used, decrement h, and return to step 1.

Algorithm 1: EviWeave-Agent inference for a complete frozen policy

D Complete Policy Family and Calibration

D.1 P01–P20 and Development Ordering

Table I1 lists the complete supplied policy family. Every row uses the same three-domain support guard, .45 domain cap, four-action horizon, .04 cost coefficient, .10 redundancy coefficient, and risk heads in Appendix C. Each threshold has its own FQI policy. A query is selected only when its frozen value exceeds STOP's zero value. On stopping, the agent releases the deepest eligible answer or abstains. Thus rows differ in the release threshold and the associated trained query–stop policy, not merely a post-hoc score cutoff.

The specified order is P01 through P20. Training and development precede freezing the weights, thresholds, order, attack rules, tools, and tie rules. Calibration data are accessed only after these are frozen. The supplemental development costs are explicitly described as illustrative sorting values, so they are not presented here as measured development outcomes or evidence of preregistration. Selection returns the first accepted row in this order, not a guaranteed minimum-cost policy on the unknown population.

D.2 Exact Endpoint Tests

For each policy–target–seed combination, evaluate 18 domain–threat cells with $n = 20{,}000$ roots per cell. For risk counts k_F, k_U, k_O, use

$$p_j = \Pr\{\mathrm{Bin}(n, \alpha) \le k_j\}, \quad j \in \{F, U, O\}.$$

For answer and correct-task coverage counts k_A, k_G, use

$$p_C = \Pr\{\mathrm{Bin}(n, .70) \ge k_A\}, \qquad p_G = \Pr\{\mathrm{Bin}(n, .85) \ge k_G\}.$$

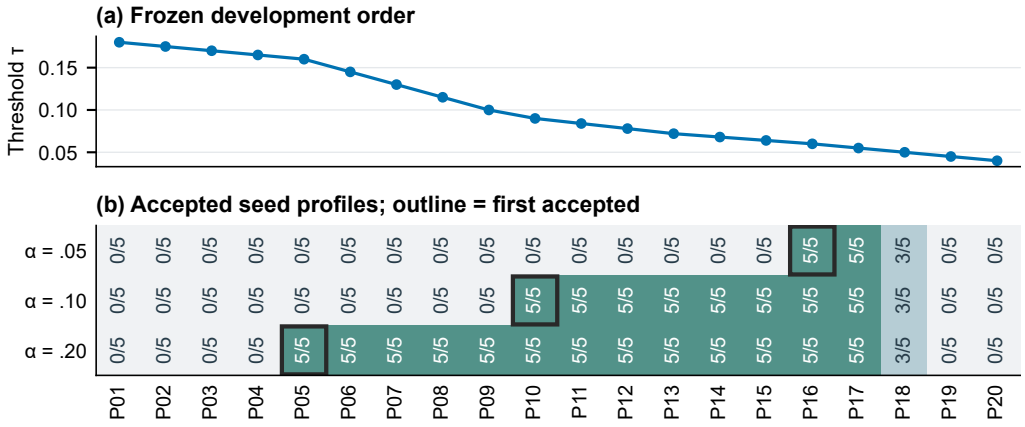


Figure D1: Frozen thresholds and supplied acceptance pattern across five seed profiles. Outlines identify the first accepted policies P16, P10, and P05 at the respective risk targets. More stringent thresholds can fail the coverage constraint. The counts are a compatibility construction, not new calibration evidence; exact values are in Table I1.

The null hypotheses state that risk is at least its target and coverage is at most its target. Accept a combination only when all 18×5 tests have $p \le q = 1/6000$. For a count k, the corresponding one-sided Clopper–Pearson limits (Clopper & Pearson, 1934), away from boundary cases, are

$$U(k,n) = \mathrm{Beta}^{-1}(1-q; k+1, n-k), \qquad L(k,n) = \mathrm{Beta}^{-1}(q; k, n-k+1).$$

Use $L(0,n) = 0$ and $U(n,n) = 1$. The $C \ge .70$ endpoint is retained for consistency with the original protocol, although $C_{\mathrm{good}} \ge .85$ and $C_{\mathrm{good}} \le C$ already imply it.

D.3 Why the Correction Is at the Combination Level

Proposition 2 (Control of accepting an infeasible combination) *Condition on a fixed family of 20 complete policies, three risk targets, five seed profiles, and their development order. If every true endpoint null has a valid level-q test, intersection–union acceptance within each combination and $q = .05/300$ imply*

$$\Pr\{\textit{at least one infeasible combination is accepted}\} \le .05.$$

Proof. An infeasible combination contains at least one true endpoint null. Accepting the combination requires rejection of that null, an event of probability at most q. A union bound over at most 300 combinations proves the claim. □

This argument requires no independence between endpoints, threats, or seed profiles. Each binomial test still requires the stated iid-root sampling assumptions. The result controls false acceptance, not simultaneous coverage of all $300 \times 18 \times 5 = 27{,}000$ displayed endpoint intervals. Within-episode adaptive queries are part of a complete frozen policy; cross-root attack learning, post-calibration model changes, or expansion of the policy family require a new argument or independent data. Additional baseline or stress certification is not automatically included in the 300-combination family.

D.4 Reported Acceptance Pattern and What It Establishes

In the supplemental pattern, P16 and P17 are accepted for all five profiles at $\alpha = .05$, and P18 accepted for seeds 11, 23, and 37 but not 53 and 97 because of correct-task coverage. This gives $5 + 5 + 3 = 13$ accepted rows. At $\alpha = .10$, P10–P17 contribute 40 and P18 contributes three; at $\alpha = .20$, P05–P17 contribute 65 and P18 contributes three. The resulting totals are 13, 43, and 68 of 100, with first accepted policies P16, P10, and P05. P19 and P20 are not accepted despite smaller thresholds, illustrating the coverage constraint.

The source gives $(U_F, U_U, U_O, L_{\text{good}})$ as $(.043986, .024989, .028986, .855989)$, $(.078983, .044993, .051017, .875988)$, and $(.111998, .063002, .071997, .877018)$ at the three targets. It also states that the underlying counts were constructed to match the original rounded bounds. These values therefore document numerical compatibility with Table 2, not additional evidence of calibration. The attachment contains the summary but not the listed 5,400-row calibration CSV, 300-row combination CSV, or 15-row selection CSV. Neither the original acceptance counts nor the constructed marginals uniquely determine the missing root-level or cell-level records.

E Baselines and Comparison Protocols

E.1 Shared Interface and Component Matrix

The supplemental methods share applicability filtering, answer trees, two initial records, four paid actions, unit query prices, a horizon of four, training and evaluation roots, and the evaluator. Table E1 specifies which components differ. Calibration eligibility is distinct from successful acceptance: searching a threshold family does not imply that any member satisfies the joint risk and coverage constraints.

Table E1: Components of the supplemental baseline definitions. Event deduplication and domain-level dependence handling are separate operations. A shared calibration protocol does not imply shared acceptance status.

Method	Dedup.	Domain cap	Reliability	Query rule	Release
SC proxy	No	No	Equal records	Entropy	Leaf vote
Provenance	Yes	No	Equal events	Fixed	Hierarchy
Fixed fusion	Yes	Yes	Rule-based	Fixed	Heads/guard
Selective, no dep.	No	No	Record-level	FQI, no penalty	Heads/hierarchy
EviWeave	Yes	Yes	Rule-based	FQI	Heads/guard
Oracle labels	Yes	Yes	Filtered, rule-based	FQI	Heads/guard

E.2 ReAct-Style Self-Consistency Proxy

The concrete supplemental baseline is a *symbolic proxy*. For each of 16 leaves, average the signed support of raw applicable records, then apply softmax at temperature .7. Sample one leaf from each of three random streams seeded by seed+100, seed+200, and seed+300. Release only if at least two agree and the chosen leaf passes its fixed risk heads at τ; otherwise query again or abstain at the horizon. Select queries by predicted reduction in Shannon entropy under the supplied transition model, using 64 posterior predictive samples per action and D2–D5 tie order.

This definition specifies the self-consistency sample count and controller, but not a language-model implementation of ReAct. The provided attachment contains no identified backbone checkpoint, model version, full prompt, or language-model decoding record. The ReAct-labeled aggregate is retained as originally reported. It is neither attributed to a newly verified language-model run nor established as a measurement of this specified proxy.

E.3 Provenance, Fixed Fusion, Selective Control, and Oracle

Provenance-only aggregation gives equal weight to unique events without a domain cap or reliability weighting. Fixed fusion uses the same reliability, capped domain aggregation, heads, and release guard as the supplemental EviWeave specification but queries in D2, D3, D4, D5 order. Selective control without dependence uses record-level reliability and the same three-head forms and hierarchy, treating records as source units and omitting the redundancy penalty. Repeated records can therefore inflate effective support. Separate complete policies can be formed at the 20 thresholds; the supplied summary does not fully specify the fixed-order methods' stopping checkpoints or establish matched trained weights.

The oracle reads true corruption flags to remove malicious events. It does not read the true leaf, unqueried clean contents, or future returns. Ordinary measurement errors remain possible. It is a privileged comparison reference, not a proof of the optimal attainable policy.

E.4 Main Comparison, Locked-Threshold Ablations, and Controllers

Three comparisons answer different questions. The original main comparison has fixed-fusion cost 2.57, a no-active-query locked-threshold diagnostic has cost 2.84, and the separate target-region controller comparison has fixed-order cost 3.10. These are distinct operating points, not interchangeable estimates of one controller.

The supplemental ablation definitions remove event/domain dependence, replace reliability by equal weights, substitute fixed order for acquisition, restrict release to leaves or abstention, or bypass calibration using the first development policy, respectively. The original numbers in Figure A1 remain locked-threshold diagnostics; they are not automatically measurements of every newly specified variant, nor separately recalibrated feasible policies. In particular, removing a component can invalidate the full policy's acceptance.

The separate acquisition comparison reports $(R_F, C_{\text{good}}, \text{cost})$ of $(.035, .873, 3.10)$ for fixed order, $(.032, .883, 4.00)$ for query-all, $(.036, .870, 2.88)$ for entropy, $(.037, .867, 2.64)$ for information gain, and $(.038, .861, 2.31)$ for EviWeave. In the supplemental specification, information gain integrates over the public generation distribution rather than consulting the current root's unqueried returns. EviWeave has lower query cost within the displayed target region; query-all has lower factual risk and higher correct-task coverage. This comparison does not imply superiority across all metrics.

The supplied description states that 540 baseline cells were constructed to match the original aggregates; their CSV is absent from the attachment. We therefore do not infer new matched measurements or paired uncertainty from that description. Maximum risk and minimum coverage can occur in different cells: the original fixed-fusion extrema satisfy $.082 + .824 = .906 > .901$, making co-location of those three extrema impossible.

F Dependence-Graph, Corruption-Budget, and Drift Stress

F.1 Interventions and Interpretation

The supplemental stress specification freezes P16 and changes only the declared condition, without retuning or retraining. Graph interventions affect the agent's provenance information; the evaluator retains the true upstream identities and labels. Tables I2–I3 transcribe all 33 supplied scenario rows. The source describes these values as preset and the rows as configurations for evaluation. They are therefore reported here as supplied scenario summaries, not independently established empirical tolerance thresholds or transfer certificates.

Five graph interventions. Missing edges delete record–upstream links with probability ρ and treat disconnected wrappers as singletons. Spurious edges connect previously unrelated clusters and recompute connected components; the source specifies a comparable scale but leaves the edge-sampling denominator unspecified. Merging selects true clusters with probability ρ and merges them with the next cluster in identifier order. Splitting selects clusters with at least two wrappers with probability ρ and partitions their sorted record IDs into two groups. Misassignment changes a whole cluster's visible domain to one of the other four, leaving its true domain unchanged. Hash-based recovery of strict copies precedes measurement of residual errors. These interventions use different units, so equal ρ does not represent equal perturbation severity.

F.2 What the Graph Scenarios Show

At $\rho = .10$, the misassignment row has $R_F = .053$ and $C_{\text{good}} = .843$, missing both targets; its .05 row remains inside the displayed region. Merging at .10 has $R_F = .040$ but $C_{\text{good}} = .849$: the coverage target fails before the factual-risk target. Missing edges and splitting first leave the displayed region at .20 on this grid, and spurious edges do so at .20 through coverage. These observations describe only the supplied discrete scenarios. They identify neither an exact critical error rate between grid points nor a universal graph-error tolerance.

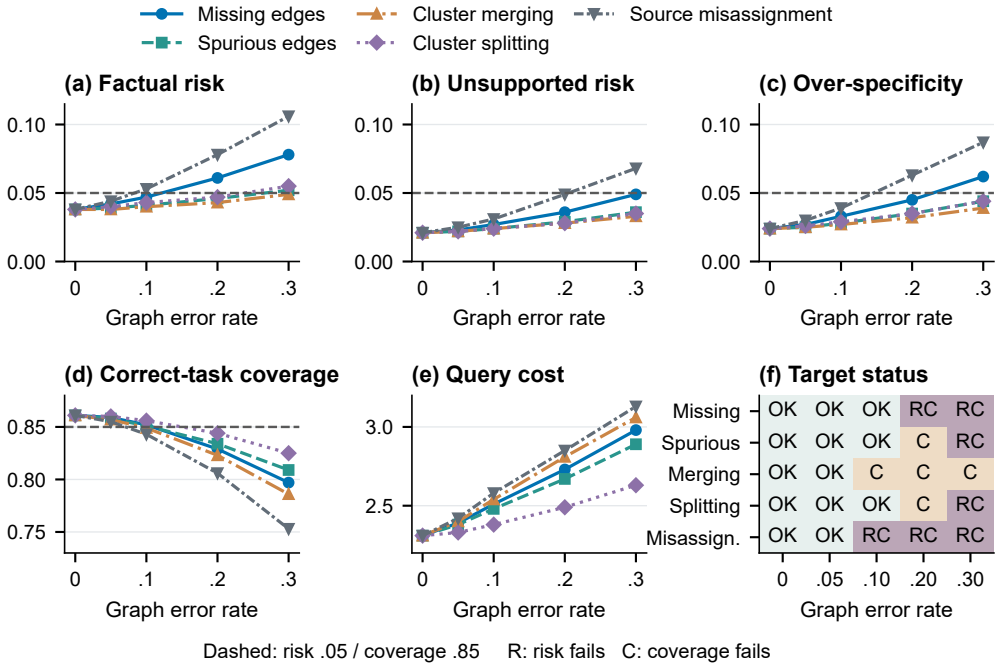


Figure F1: All 25 supplied graph-error scenarios for frozen P16. Risks are reported cell maxima, correct-task coverage is a minimum, and query cost is a mean. Dashed lines mark .05 risk and .85 coverage targets. R and C denote risk and coverage violations; OK means the displayed targets hold. Merging fails the coverage target at $\rho = .10$ while factual risk remains below .05. Lines connect only supplied points; these preset scenario summaries do not establish a tolerance certificate. Exact rows: Table I2.

F.3 Larger Corruption Budgets and Ordinary Source Drift

The corruption-budget rows vary the number of compromised domains over $b = 0, 1, 2, 3$ out of five. For $b = 2$, the scenario records $(R_F, C_{\text{good}}, \text{cost}) = (.087, .808, 2.91)$; for $b = 3$, $(.154, .703, 3.41)$. The .45 cap constrains one domain, not the combined mass of multiple malicious domains. Results at $b > 1$ lie outside the stated operating assumption and neither inherit its certificate nor refute a conditional $b \leq 1$ claim.

Drift raises each nonmalicious source's leaf error probability from $1 - a_d$ to $\min(1, 1 - a_d + \delta)$ for $\delta \in \{0, .05, .10, .20\}$. The historical priors and drift summaries remain frozen; the agent does not observe δ. Thus drift is distinct from increasing the number of malicious domains. The .10 row records $R_F = .068$ and $C_{\text{good}} = .827$.

Calibration boundary. A displayed pass means all three listed risks are at most .05 and $C_{\text{good}} \geq .85$. This indicates numerical agreement with the targets, not acceptance by the original calibration tests. Changes to graph, corruption, or reliability populations require separate evaluation and, for statistical claims, an appropriately specified calibration procedure. The supplemental status labels do not themselves establish the sampling assumptions or a distribution-shift guarantee.

G Cross-Domain Hierarchical Release Cases

These eight supplemental cases document terminal evidence processing and release under $\tau = .06$. We transcribe their records and all 32 reported node-score triples. The paid trajectories are taken as given; the attachment includes no learned Q-values showing that the FQI controller selects them. Each paid action has one nonempty slot and two empty slots in these cases; the first two records are initial evidence. These examples therefore illustrate release mechanisms without establishing acquisition optimality.

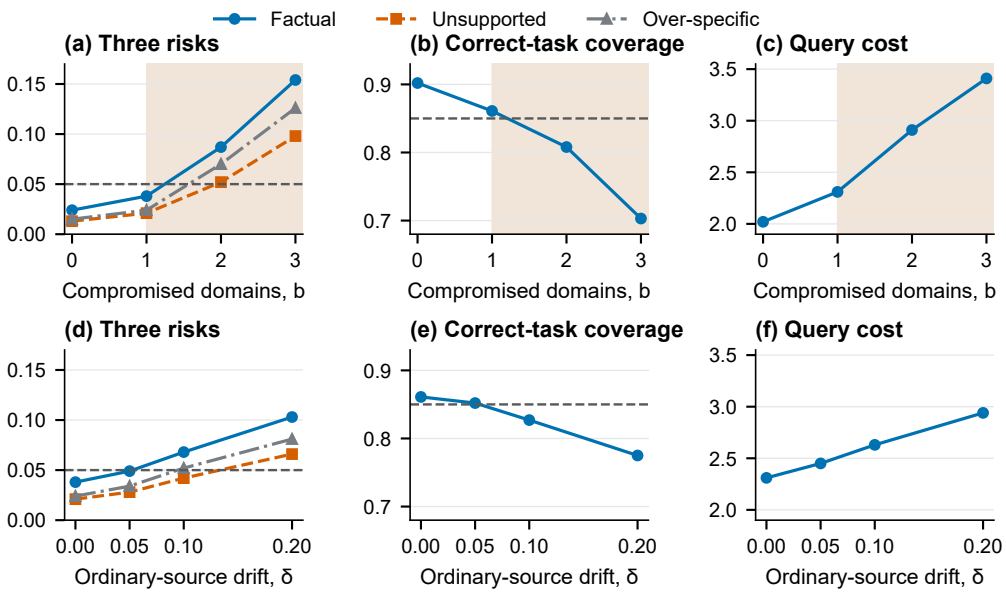


Figure F2: Supplied corruption-budget and ordinary-source-drift scenarios. Each row separates the three maximum risks, minimum correct-task coverage, and mean query cost. Shading marks $b > 1$, outside the stated corruption assumption; ordinary-source drift is a distinct change with frozen priors. Dashed lines mark the displayed targets, not transferred calibration guarantees. No interpolation-based critical rate is claimed. Exact rows: Table I3.

Strength is .98 and normalized age is .05; missingness, drift, and volatility are zero. Validity is $[-1, 1]$ unless shown otherwise. Audited support matches parsed support except in SCI-03 and OD-03. Corruption flags in the tables are evaluator annotations, not ordinary policy inputs. The source reports that adding ten strict copies preserved each case's scores; no rule or experiment was rerun for this manuscript revision.

Reading the cases. OD-01 reaches the required category and counts toward G; OD-02 and CYB-01 retain correct answers that are too coarse and count toward H instead. SCI-01 requires an independent source after a wrapper; SCI-02 abstains after exhausting the query horizon. CYB-02 is a factual failure with apparent multi-domain support. SCI-03 and OD-03 distinguish independent support evaluation from parsed evidence and factual correctness. They can contribute to G while incurring U or O, because G does not include those endpoints.

The supplemental CYB-01 differs from the original cybersecurity example; Appendix I gives the score comparison. Figure 6 and Table A1 retain the original case.

H Complete 90-Cell Supplemental Ledger

H.1 Provenance, Denominators, and Aggregation

Tables I21–I23 include every supplied domain–threat–seed cell, with $n = 20{,}000$ in each row. The accompanying CSV retains integer F, U, O, A, G, H, T counts, abstentions, total queries, and the adjacent integer query-count allocation. Extrema are computed across all cells, without omissions or prior averaging.

The source constructs these rows from threat anchors with domain and seed offsets. For factual risk, domain offsets are $(-.002, -.001, 0)$ for OD, SCI, CYB and seed offsets are $(-.0008, -.0006, -.0004, -.0002, 0)$ for seeds 11, 23, 37, 53, 97. Correct-task coverage offsets are $(.004, .002, 0)$ by domain and $(.0016, .0012, .0008, .0004, 0)$ by seed. Cost offsets are $(-.06, 0, .06)$ and $(-.04, -.02, 0, .02, .04)$, with zero means. The CSV supplies the final values. These construc-

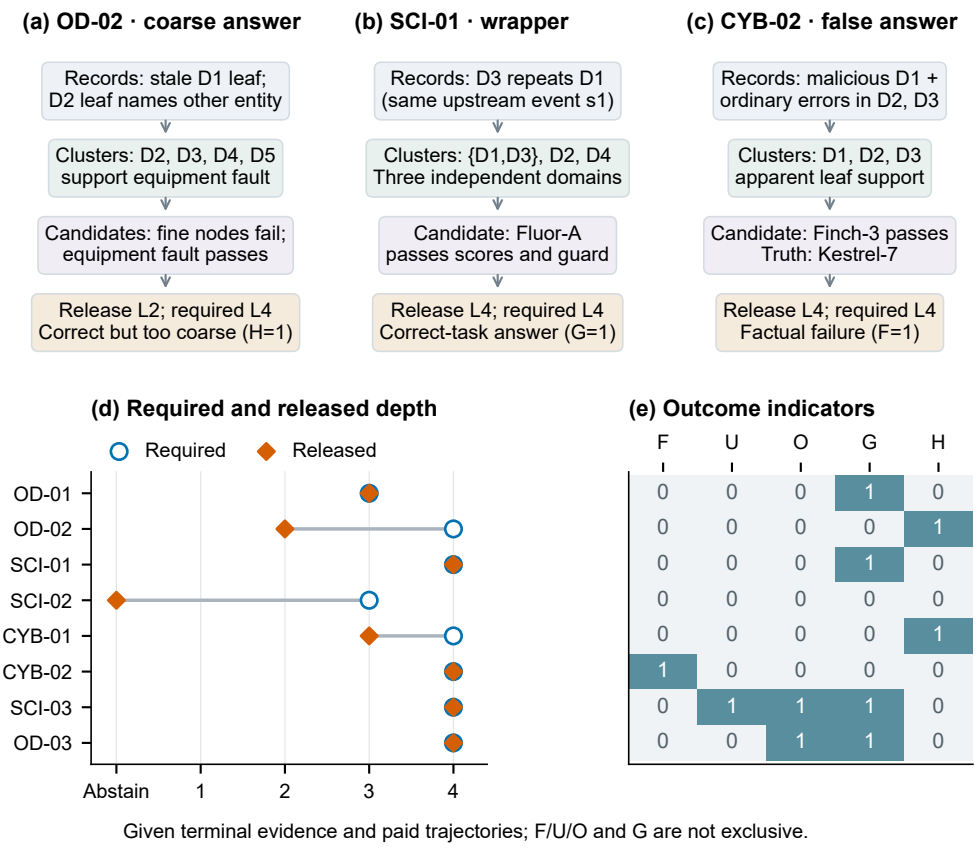


Figure G1: Terminal evidence structure and all eight supplied outcomes. The upper panels explain coarse release, shared-upstream duplication, and a false release with apparent multi-domain support. In (d), zero denotes abstention rather than an answer node. In (e), F/U/O denote the three losses; G is correct and task-sufficient, while H is correct but too coarse. These endpoints are not mutually exclusive. The given trajectories do not establish acquisition optimality. Exact outcomes: Table I4.

tion rules place the worst factual risk and lowest correct-task coverage in CYB–adaptive–97 by design. They do not independently identify the location of the original worst cell.

Marginal counts support accounting checks and reconstruction of the displayed extrema, but cannot recover paired root-level outcomes or bootstrap intervals. Neither matching $(.038, .861, 2.31)$ nor satisfying integer identities establishes that a particular trained controller produced this ledger.

H.2 Availability of the Supplied Records

The supplied material consists of a narrative TXT, a detailed Markdown appendix with a PDF rendering, and one 90-cell CSV. The text also lists machine-readable baseline, calibration, stress, case, and tree files, but these files are absent from the attachment. We therefore include the actual narrative tables and CSV, rather than asserting delivery of a 540-cell baseline file, a 5,400-row calibration ledger, trained policy weights, or root-level trajectories. The original aggregate results and paired intervals are retained as reported. The supplemental construction does not independently verify them.

H.3 An Integer Worked Example

The CYB–adaptive–97 row has

$$n = 20{,}000, \quad F = 760, \quad G = 17{,}220, \quad H = 300, \quad A = 18{,}280,$$
$$U = 420, \quad O = 480, \quad T = 17{,}760, \quad Q_{\text{total}} = 54{,}200.$$

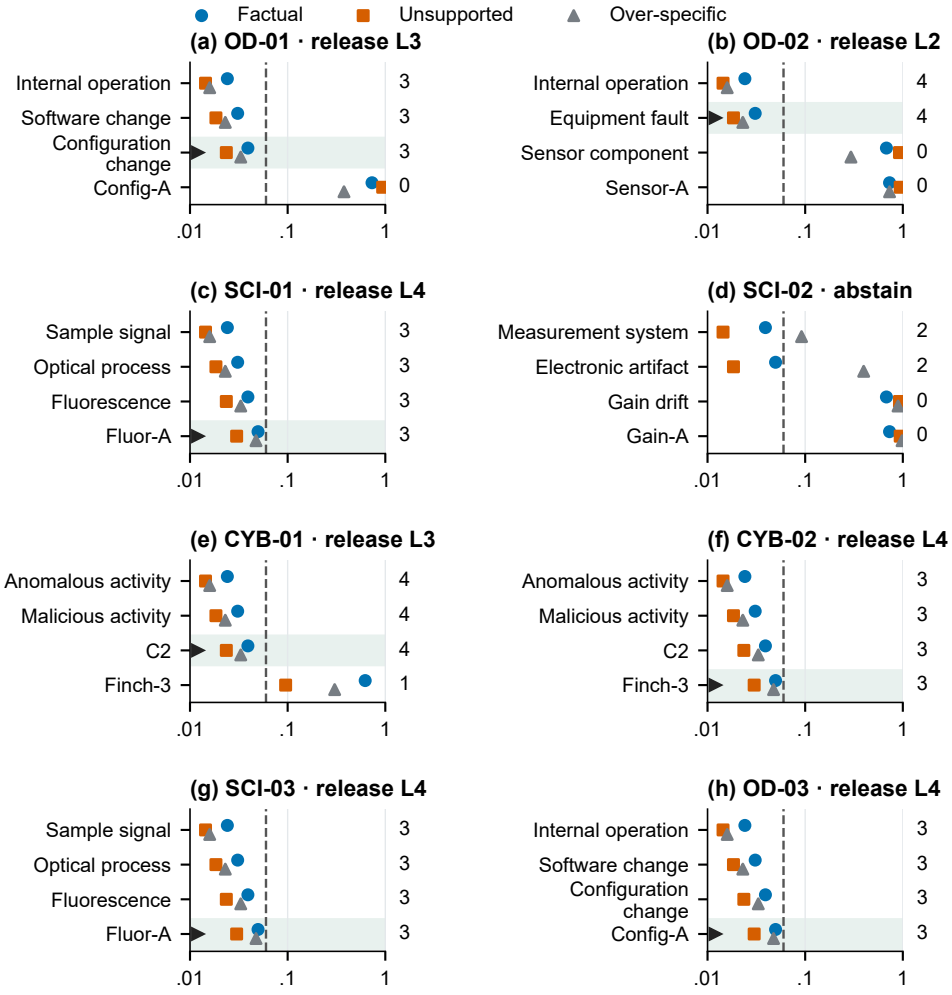


Figure G2: All 32 candidate nodes and 96 supplied predicted losses. The shared logarithmic scale displays individual scores; marker shapes identify the loss heads. The dashed line is $\tau = .06$, and the right-hand column gives the supporting-domain count m. Release requires every score to pass and $m \geq 3$; shaded rows and arrows mark actual releases. SCI-02 abstains. These terminal scores are not per-answer correctness guarantees. Exact records and scores are retained in Appendix I.

Hence $A = F + G + H$, $R_F = .038$, $R_U = .021$, $R_O = .024$, $C = .914$, $C_{\text{good}} = .861$, and $C_{\text{task}} = .888$. Its query cost is $54{,}200/20{,}000 = 2.71$, whereas the equal-cell mean over the whole ledger is 2.31. Its answer-conditioned factual risk is $760/18{,}280 = .04157549$, which rounds to .042. The row-level identities checked during document preparation are

$$A = F + G + H, \quad G \leq T \leq A, \quad T - G \leq F, \quad A + \text{abstentions} = n, \quad 0 \leq Q_{\text{total}} \leq 4n.$$

These checks assess document consistency; they do not constitute experimental execution.

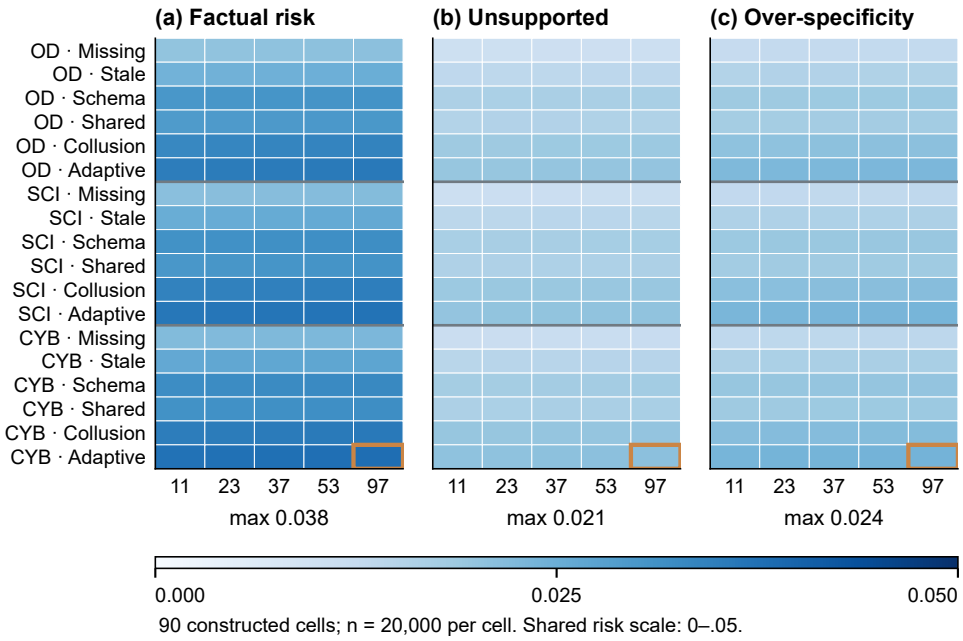


Figure H1: Complete constructed 90-cell ledger: factual, unsupported-claim, and over-specificity risks. Each panel retains all 18 domain–threat combinations and five seed profiles on a shared 0–.05 scale. Outlines mark maxima, all at CYB–adaptive–97. Each rate uses 20,000 roots; threat versions and seed profiles are not additional independent roots. The spatial pattern reflects the supplied construction, not a newly observed location of the worst cell. Exact values: Tables I21–I23.

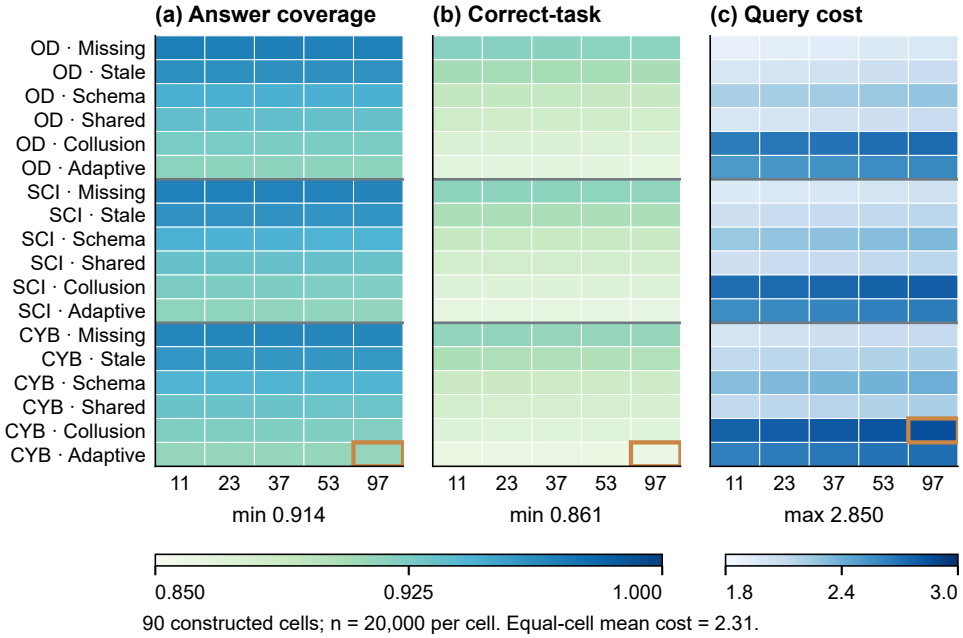


Figure H2: Coverage and query cost for the same 90 cells and ordering. Answer coverage and correct-task coverage share the .85–1 scale; cost has its own 1.8–3 scale. Outlines mark the coverage minima at CYB–adaptive–97 and maximum cost at CYB–collusion–97. Equal-cell mean cost is 2.31, whereas cost in the worst-risk cell is 2.71. These are constructed ledger summaries; query cost excludes token, training, and inference costs. Exact values: Tables I21–I23.

I Complete Numerical Records

This section retains the complete numerical tables underlying the visual summaries in Appendices D, F, G, and H. All rows and candidate scores are retained. The evidence qualifications in those sections also apply to these records.

I.1 Policy thresholds and acceptance counts

Table I1: All 20 thresholds and the supplied acceptance pattern over five seed profiles. The source describes the calibration counts as a construction matching the original summary; this is not an independent certification result.

Policy	Rank	τ	Pass at .05	Pass at .10	Pass at .20
P01	1	0.180	0/5	0/5	0/5
P02	2	0.175	0/5	0/5	0/5
P03	3	0.170	0/5	0/5	0/5
P04	4	0.165	0/5	0/5	0/5
P05	5	0.160	0/5	0/5	5/5
P06	6	0.145	0/5	0/5	5/5
P07	7	0.130	0/5	0/5	5/5
P08	8	0.115	0/5	0/5	5/5
P09	9	0.100	0/5	0/5	5/5
P10	10	0.090	0/5	5/5	5/5
P11	11	0.084	0/5	5/5	5/5
P12	12	0.078	0/5	5/5	5/5
P13	13	0.072	0/5	5/5	5/5
P14	14	0.068	0/5	5/5	5/5
P15	15	0.064	0/5	5/5	5/5
P16	16	0.060	5/5	5/5	5/5
P17	17	0.055	5/5	5/5	5/5
P18	18	0.050	3/5	3/5	3/5
P19	19	0.045	0/5	0/5	0/5
P20	20	0.040	0/5	0/5	0/5

I.2 Graph-error scenarios

Table I2: All 25 supplied graph-error scenarios for frozen P16. Risks are stated cell maxima, correct-task coverage a minimum, and cost a mean. Pass compares displayed values with the targets; these scenario rows do not certify graph-error tolerance.

Intervention	ρ	R_F	R_U	R_O	C_{good}	Cost	Pass
Missing edges	0.000	0.038	0.021	0.024	0.861	2.310	Yes
Missing edges	0.050	0.042	0.023	0.027	0.859	2.390	Yes
Missing edges	0.100	0.047	0.027	0.033	0.852	2.510	Yes
Missing edges	0.200	0.061	0.036	0.045	0.829	2.730	No
Missing edges	0.300	0.078	0.049	0.062	0.797	2.980	No
Spurious edges	0.000	0.038	0.021	0.024	0.861	2.310	Yes
Spurious edges	0.050	0.039	0.022	0.025	0.858	2.380	Yes
Spurious edges	0.100	0.041	0.024	0.028	0.851	2.480	Yes
Spurious edges	0.200	0.046	0.029	0.035	0.834	2.670	No
Spurious edges	0.300	0.052	0.036	0.044	0.809	2.890	No
Cluster merging	0.000	0.038	0.021	0.024	0.861	2.310	Yes
Cluster merging	0.050	0.038	0.022	0.025	0.857	2.400	Yes
Cluster merging	0.100	0.040	0.024	0.027	0.849	2.540	No
Cluster merging	0.200	0.043	0.028	0.032	0.823	2.810	No
Cluster merging	0.300	0.049	0.033	0.039	0.786	3.060	No
Cluster splitting	0.000	0.038	0.021	0.024	0.861	2.310	Yes
Cluster splitting	0.050	0.040	0.022	0.026	0.860	2.330	Yes

Continued on next page

Intervention	ρ	R_F	R_U	R_O	C_{good}	Cost	Pass
Cluster splitting	0.100	0.043	0.024	0.029	0.856	2.380	Yes
Cluster splitting	0.200	0.047	0.028	0.035	0.844	2.490	No
Cluster splitting	0.300	0.055	0.035	0.044	0.825	2.630	No
Source misassignment	0.000	0.038	0.021	0.024	0.861	2.310	Yes
Source misassignment	0.050	0.044	0.025	0.030	0.855	2.420	Yes
Source misassignment	0.100	0.053	0.031	0.039	0.843	2.580	No
Source misassignment	0.200	0.078	0.049	0.063	0.806	2.850	No
Source misassignment	0.300	0.106	0.068	0.087	0.753	3.130	No

I.3 Budget and drift scenarios

Table I3: All eight supplied budget and drift scenarios. The $b > 1$ rows are outside the main corruption assumption. Drift affects ordinary source error while priors stay frozen. Displayed target satisfaction is not a transferred calibration guarantee.

Condition	b / δ	R_F	R_U	R_O	C_{good}	Cost	Pass
Budget	0	0.024	0.013	0.015	0.902	2.020	Yes
Budget	1	0.038	0.021	0.024	0.861	2.310	Yes
Budget	2	0.087	0.052	0.070	0.808	2.910	No
Budget	3	0.154	0.098	0.126	0.703	3.410	No
Drift	0.000	0.038	0.021	0.024	0.861	2.310	Yes
Drift	0.050	0.049	0.028	0.034	0.852	2.450	Yes
Drift	0.100	0.068	0.042	0.052	0.827	2.630	No
Drift	0.200	0.103	0.066	0.081	0.775	2.940	No

Relationship to the original cybersecurity case. The original Table A1 reports Kestrel-7 scores $(.190, .220, .250)$ and C2 scores $(.030, .018, .024)$. The supplemental CYB-01 instead evaluates an incorrect Finch-3 leaf with $(.622459, .095349, .302941)$ and C2 with $(.039166, .023545, .033086)$. Both illustrate backoff to C2, but their records and scores differ. The supplemental rule instantiation does not reproduce the original case exactly. Figure 6 and its original score table retain the originally reported values.

I.4 Case outcomes

Table I4: Eight supplemental case outcomes. Costs count only the given paid trajectories. G requires correctness and task depth; U and O are separate endpoints.

Case	L	Release (depth)	Cost	$F/U/O$	G	H
OD-01	3	Configuration change (3)	2	0/0/0	1	0
OD-02	4	Equipment fault (2)	4	0/0/0	0	1
SCI-01	4	Fluor-A (4)	2	0/0/0	1	0
SCI-02	3	Abstain	4	0/0/0	0	0
CYB-01	4	C2 (3)	3	0/0/0	0	1
CYB-02	4	Finch-3 (4)	1	1/0/0	0	0
SCI-03	4	Fluor-A (4)	1	0/1/1	1	0
OD-03	4	Config-A (4)	1	0/0/1	1	0

I.5 OD-01: FILTERING STALE FINE-GRAINED EVIDENCE

Question and truth. Classify OD-01 at time zero to depth 3; true leaf: Config-A. **Paid trajectory:** D3, D4; cost 2.

Table I5: OD-01 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	o1	D1/D1	Config-B	[-8,-2]	Task	No
r2	o2	D2/D2	Configuration change	[-1,1]	Task	No
r3	o3	D3/D3	Configuration change	[-1,1]	Task	No
r4	o4	D4/D4	Configuration change	[-1,1]	Task	No

Table I6: OD-01 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Internal operation	1	3	0.024127	0.014415	0.015906	Yes
Software change	2	3	0.030769	0.018433	0.022977	Yes
Configuration change	3	3	0.039166	0.023545	0.033086	Yes
Config-A	4	0	0.731059	0.939913	0.377541	No

D1's Config-B record is stale. The three applicable domains support configuration change but no finer claim. The depth-3 release reaches the task requirement; the unsupported leaf is rejected.

I.6 OD-02: CORRECT BUT BELOW REQUIRED DEPTH

Question and truth. Classify OD-02 at time zero to depth 4; true leaf: Sensor-A. **Paid trajectory:** D2, D3, D4, D5; cost 4.

Table I7: OD-02 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	o1	D1/D1	Sensor-A	[-8,-2]	Task	No
r2	o2	D2/D2	Equipment fault	[-1,1]	Task	No
r3	o3	D2/D2	Sensor-A	[-1,1]	Other	No
r4	o4	D3/D3	Equipment fault	[-1,1]	Task	No
r5	o5	D4/D4	Equipment fault	[-1,1]	Task	No
r6	o6	D5/D5	Equipment fault	[-1,1]	Task	No

Table I8: OD-02 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Internal operation	1	4	0.024127	0.014415	0.015906	Yes
Equipment fault	2	4	0.030769	0.018433	0.022977	Yes
Sensor component	3	0	0.679179	0.924142	0.294215	No
Sensor-A	4	0	0.731059	0.939913	0.731059	No

The fine-grained D1 record is stale, and the new D2 record concerns a different entity. Four applicable domains support equipment fault, but no finer node. The correct depth-2 answer counts toward H, not correct-task coverage.

I.7 SCI-01: A WRAPPER ADDS NO INDEPENDENT DOMAIN

Question and truth. Classify SCI-01 at time zero to depth 4; true leaf: Fluor-A. **Paid trajectory:** D3, D4; cost 2.

Table I9: SCI-01 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	s1	D1/D1	Fluor-A	[-1,1]	Task	No
r2	s2	D2/D2	Fluor-A	[-1,1]	Task	No
r3	s1	D1/D3	Fluor-A	[-1,1]	Task	No
r4	s4	D4/D4	Fluor-A	[-1,1]	Task	No

Table I10: SCI-01 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Sample signal	1	3	0.024127	0.014415	0.015906	Yes
Optical process	2	3	0.030769	0.018433	0.022977	Yes
Fluorescence	3	3	0.039166	0.023545	0.033086	Yes
Fluor-A	4	3	0.049737	0.030032	0.047426	Yes

Records r1 and r3 share event s1 and origin D1 despite different return tools. Paid D3 leaves only two independent domains; D4 supplies the third. The terminal depth-4 answer is correct and task-sufficient.

I.8 SCI-02: ABSTENTION AFTER THE QUERY HORIZON

Question and truth. Classify SCI-02 at time zero to depth 3; true leaf: Gain-A. **Paid trajectory:** D2, D3, D4, D5; cost 4.

Table I11: SCI-02 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	s1	D1/D1	Electronic artifact	[-1,1]	Task	No
r2	s2	D2/D2	Gain-A	[-1,1]	Other	No
r3	s3	D2/D2	Gain-A	[-1,1]	Other	No
r4	s4	D3/D3	Gain-A	[-8,-2]	Task	No
r5	s5	D4/D4	Electronic artifact	[-1,1]	Task	No
r6	s1	D1/D5	Electronic artifact	[-1,1]	Task	No

Table I12: SCI-02 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Measurement system	1	2	0.039166	0.014415	0.091815	No
Electronic artifact	2	2	0.049737	0.018433	0.397315	No
Gain drift	3	0	0.679179	0.924142	0.893309	No
Gain-A	4	0	0.731059	0.939913	0.982014	No

D2 records concern another entity, and the D3 record is stale. D5 wraps the D1 event. Only two applicable upstream domains remain; no node satisfies the three-domain guard.

I.9 CYB-01: Coarse release despite a corrupted family label

Question and truth. Classify CYB-01 at time zero to depth 4; true leaf: Kestrel-7. **Paid trajectory:** D3, D4, D5; cost 3.

Table I13: CYB-01 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	c1	D1/D1	Finch-3	[-1,1]	Task	Yes
r2	c2	D2/D2	C2	[-1,1]	Task	No
r3	c1	D1/D3	Finch-3	[-1,1]	Task	Yes
r4	c4	D4/D4	C2	[-1,1]	Task	No
r5	c5	D5/D5	C2	[-1,1]	Task	No

Table I14: CYB-01 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Anomalous activity	1	4	0.024127	0.014415	0.015906	Yes
Malicious activity	2	4	0.030769	0.018433	0.022977	Yes
C2	3	4	0.039166	0.023545	0.033086	Yes
Finch-3	4	1	0.622459	0.095349	0.302941	No

D1 and its D3 wrapper assert Finch-3, while independent sources support C2. C2 is correct but less specific than the required family label, so the release contributes to H.

I.10 CYB-02: Failure with malicious and ordinary errors

Question and truth. Classify CYB-02 at time zero to depth 4; true leaf: Kestrel-7. **Paid trajectory:** D3; cost 1.

Table I15: CYB-02 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	c1	D1/D1	Finch-3	[-1,1]	Task	Yes
r2	c2	D2/D2	Finch-3	[-1,1]	Task	No
r3	c3	D3/D3	Finch-3	[-1,1]	Task	No

Table I16: CYB-02 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Anomalous activity	1	3	0.024127	0.014415	0.015906	Yes
Malicious activity	2	3	0.030769	0.018433	0.022977	Yes
C2	3	3	0.039166	0.023545	0.033086	Yes
Finch-3	4	3	0.049737	0.030032	0.047426	Yes

A malicious source and two erroneous nonmalicious measurements all support Finch-3. The wrong leaf passes every internal score and the source guard. This factual error nevertheless has observed support and apparently adequate specificity.

I.11 SCI-03: CORRECT BY CHANCE BUT SEMANTICALLY UNSUPPORTED

Question and truth. Classify SCI-03 at time zero to depth 4; true leaf: Fluor-A. **Paid trajectory:** D3; cost 1.

Table I17: SCI-03 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	s1	D1/D1	Fluor-A	[-1,1]	Task	No
r2	s2	D2/D2	Fluor-A	[-1,1]	Task	No
r3	s3	D3/D3	Fluor-A	[-1,1]	Task	No

Table I18: SCI-03 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Sample signal	1	3	0.024127	0.014415	0.015906	Yes
Optical process	2	3	0.030769	0.018433	0.022977	Yes
Fluorescence	3	3	0.039166	0.023545	0.033086	Yes
Fluor-A	4	3	0.049737	0.030032	0.047426	Yes

Independent support annotations: r1: Fluorescence; r2: Fluorescence; r3: Fluorescence.

All three parsed claims are Fluor-A, but independent semantic annotations support only fluorescence. The released leaf happens to be true: F=0, U=1, O=1. Its correctness and depth nevertheless yield G=1.

I.12 OD-03: ONE EXACT SOURCE IS INSUFFICIENT FOR JUSTIFIED DEPTH

Question and truth. Classify OD-03 at time zero to depth 4; true leaf: Config-A. **Paid trajectory:** D3; cost 1.

Table I19: OD-03 observed records. Mal. is an evaluator-only malicious-record flag.

ID	Event	Origin/tool	Parsed claim	Validity	Entity	Mal.
r1	o1	D1/D1	Config-A	[-1,1]	Task	No
r2	o2	D2/D2	Config-A	[-1,1]	Task	No
r3	o3	D3/D3	Config-A	[-1,1]	Task	No

Table I20: OD-03 terminal node scores at $\tau = .06$.

Candidate	Depth	Domains	$\widehat{R}_F$	$\widehat{R}_U$	$\widehat{R}_O$	Pass
Internal operation	1	3	0.024127	0.014415	0.015906	Yes
Software change	2	3	0.030769	0.018433	0.022977	Yes
Configuration change	3	3	0.039166	0.023545	0.033086	Yes
Config-A	4	3	0.049737	0.030032	0.047426	Yes

Independent support annotations: r1: Config-A; r2: Configuration change; r3: Configuration change.

D1 semantically supports Config-A; D2 and D3 support only configuration change. All parsed claims identify Config-A. At least one audited record supports the leaf, so U=0, but the evaluator's two-domain specificity requirement fails, giving O=1.

I.13 OD LEDGER: ALL 30 CELLS

Table I21: OD supplemental reconstructed ledger. Each denominator is 20,000; the supplied CSV retains integer numerators. These rows are a construction matching original aggregate targets, not independent replication measurements.

Threat	Seed	R_F	R_U	R_O	C	C_{good}	Cost
Missing	11	0.02020	0.01060	0.01260	0.96900	0.91960	1.90
Missing	23	0.02040	0.01070	0.01270	0.96875	0.91920	1.92
Missing	37	0.02060	0.01080	0.01280	0.96850	0.91880	1.94
Missing	53	0.02080	0.01090	0.01290	0.96825	0.91840	1.96
Missing	97	0.02100	0.01100	0.01300	0.96800	0.91800	1.98
Stale	11	0.02420	0.01360	0.01560	0.96100	0.90760	2.00
Stale	23	0.02440	0.01370	0.01570	0.96075	0.90720	2.02
Stale	37	0.02460	0.01380	0.01580	0.96050	0.90680	2.04
Stale	53	0.02480	0.01390	0.01590	0.96025	0.90640	2.06
Stale	97	0.02500	0.01400	0.01600	0.96000	0.90600	2.08
Schema-valid	11	0.03020	0.01660	0.01860	0.94600	0.89260	2.20
Schema-valid	23	0.03040	0.01670	0.01870	0.94575	0.89220	2.22
Schema-valid	37	0.03060	0.01680	0.01880	0.94550	0.89180	2.24
Schema-valid	53	0.03080	0.01690	0.01890	0.94525	0.89140	2.26
Schema-valid	97	0.03100	0.01700	0.01900	0.94500	0.89100	2.28
Shared upstream	11	0.02920	0.01560	0.01760	0.93500	0.88460	2.00
Shared upstream	23	0.02940	0.01570	0.01770	0.93475	0.88420	2.02
Shared upstream	37	0.02960	0.01580	0.01780	0.93450	0.88380	2.04
Shared upstream	53	0.02980	0.01590	0.01790	0.93425	0.88340	2.06
Shared upstream	97	0.03000	0.01600	0.01800	0.93400	0.88300	2.08
Collusion	11	0.03320	0.01860	0.02060	0.92600	0.87460	2.65
Collusion	23	0.03340	0.01870	0.02070	0.92575	0.87420	2.67
Collusion	37	0.03360	0.01880	0.02080	0.92550	0.87380	2.69
Collusion	53	0.03380	0.01890	0.02090	0.92525	0.87340	2.71
Collusion	97	0.03400	0.01900	0.02100	0.92500	0.87300	2.73
Adaptive	11	0.03520	0.01960	0.02260	0.91800	0.86660	2.51
Adaptive	23	0.03540	0.01970	0.02270	0.91775	0.86620	2.53
Adaptive	37	0.03560	0.01980	0.02280	0.91750	0.86580	2.55
Adaptive	53	0.03580	0.01990	0.02290	0.91725	0.86540	2.57
Adaptive	97	0.03600	0.02000	0.02300	0.91700	0.86500	2.59

I.14 SCI LEDGER: ALL 30 CELLS

Table I22: SCI supplemental reconstructed ledger. Each denominator is 20,000; the supplied CSV retains integer numerators. These rows are a construction matching original aggregate targets, not independent replication measurements.

Threat	Seed	R_F	R_U	R_O	C	C_{good}	Cost
Missing	11	0.02120	0.01110	0.01310	0.96750	0.91760	1.96
Missing	23	0.02140	0.01120	0.01320	0.96725	0.91720	1.98
Missing	37	0.02160	0.01130	0.01330	0.96700	0.91680	2.00
Missing	53	0.02180	0.01140	0.01340	0.96675	0.91640	2.02
Missing	97	0.02200	0.01150	0.01350	0.96650	0.91600	2.04
Stale	11	0.02520	0.01410	0.01610	0.95950	0.90560	2.06
Stale	23	0.02540	0.01420	0.01620	0.95925	0.90520	2.08
Stale	37	0.02560	0.01430	0.01630	0.95900	0.90480	2.10
Stale	53	0.02580	0.01440	0.01640	0.95875	0.90440	2.12
Stale	97	0.02600	0.01450	0.01650	0.95850	0.90400	2.14
Schema-valid	11	0.03120	0.01710	0.01910	0.94450	0.89060	2.26
Schema-valid	23	0.03140	0.01720	0.01920	0.94425	0.89020	2.28
Schema-valid	37	0.03160	0.01730	0.01930	0.94400	0.88980	2.30
Schema-valid	53	0.03180	0.01740	0.01940	0.94375	0.88940	2.32

Continued on next page

Threat	Seed	R_F	R_U	R_O	C	C_{good}	Cost
Schema-valid	97	0.03200	0.01750	0.01950	0.94350	0.88900	2.34
Shared upstream	11	0.03020	0.01610	0.01810	0.93350	0.88260	2.06
Shared upstream	23	0.03040	0.01620	0.01820	0.93325	0.88220	2.08
Shared upstream	37	0.03060	0.01630	0.01830	0.93300	0.88180	2.10
Shared upstream	53	0.03080	0.01640	0.01840	0.93275	0.88140	2.12
Shared upstream	97	0.03100	0.01650	0.01850	0.93250	0.88100	2.14
Collusion	11	0.03420	0.01910	0.02110	0.92450	0.87260	2.71
Collusion	23	0.03440	0.01920	0.02120	0.92425	0.87220	2.73
Collusion	37	0.03460	0.01930	0.02130	0.92400	0.87180	2.75
Collusion	53	0.03480	0.01940	0.02140	0.92375	0.87140	2.77
Collusion	97	0.03500	0.01950	0.02150	0.92350	0.87100	2.79
Adaptive	11	0.03620	0.02010	0.02310	0.91650	0.86460	2.57
Adaptive	23	0.03640	0.02020	0.02320	0.91625	0.86420	2.59
Adaptive	37	0.03660	0.02030	0.02330	0.91600	0.86380	2.61
Adaptive	53	0.03680	0.02040	0.02340	0.91575	0.86340	2.63
Adaptive	97	0.03700	0.02050	0.02350	0.91550	0.86300	2.65

I.15 CYB LEDGER: ALL 30 CELLS

Table I23: CYB supplemental reconstructed ledger. Each denominator is 20,000; the supplied CSV retains integer numerators. These rows are a construction matching original aggregate targets, not independent replication measurements.

Threat	Seed	R_F	R_U	R_O	C	C_{good}	Cost
Missing	11	0.02220	0.01160	0.01360	0.96600	0.91560	2.02
Missing	23	0.02240	0.01170	0.01370	0.96575	0.91520	2.04
Missing	37	0.02260	0.01180	0.01380	0.96550	0.91480	2.06
Missing	53	0.02280	0.01190	0.01390	0.96525	0.91440	2.08
Missing	97	0.02300	0.01200	0.01400	0.96500	0.91400	2.10
Stale	11	0.02620	0.01460	0.01660	0.95800	0.90360	2.12
Stale	23	0.02640	0.01470	0.01670	0.95775	0.90320	2.14
Stale	37	0.02660	0.01480	0.01680	0.95750	0.90280	2.16
Stale	53	0.02680	0.01490	0.01690	0.95725	0.90240	2.18
Stale	97	0.02700	0.01500	0.01700	0.95700	0.90200	2.20
Schema-valid	11	0.03220	0.01760	0.01960	0.94300	0.88860	2.32
Schema-valid	23	0.03240	0.01770	0.01970	0.94275	0.88820	2.34
Schema-valid	37	0.03260	0.01780	0.01980	0.94250	0.88780	2.36
Schema-valid	53	0.03280	0.01790	0.01990	0.94225	0.88740	2.38
Schema-valid	97	0.03300	0.01800	0.02000	0.94200	0.88700	2.40
Shared upstream	11	0.03120	0.01660	0.01860	0.93200	0.88060	2.12
Shared upstream	23	0.03140	0.01670	0.01870	0.93175	0.88020	2.14
Shared upstream	37	0.03160	0.01680	0.01880	0.93150	0.87980	2.16
Shared upstream	53	0.03180	0.01690	0.01890	0.93125	0.87940	2.18
Shared upstream	97	0.03200	0.01700	0.01900	0.93100	0.87900	2.20
Collusion	11	0.03520	0.01960	0.02160	0.92300	0.87060	2.77
Collusion	23	0.03540	0.01970	0.02170	0.92275	0.87020	2.79
Collusion	37	0.03560	0.01980	0.02180	0.92250	0.86980	2.81
Collusion	53	0.03580	0.01990	0.02190	0.92225	0.86940	2.83
Collusion	97	0.03600	0.02000	0.02200	0.92200	0.86900	2.85
Adaptive	11	0.03720	0.02060	0.02360	0.91500	0.86260	2.63
Adaptive	23	0.03740	0.02070	0.02370	0.91475	0.86220	2.65
Adaptive	37	0.03760	0.02080	0.02380	0.91450	0.86180	2.67
Adaptive	53	0.03780	0.02090	0.02390	0.91425	0.86140	2.69
Adaptive	97	0.03800	0.02100	0.02400	0.91400	0.86100	2.71