

# False Containment: Measuring the Evidence Required to Verify Autonomous AI Incident Resolution

Daud Ibrahim Hassan      Govardhan Reddy      Disha Singha  
Independent Researcher    Independent Researcher    Independent Researcher

Taemin Park      Soumya Parthasarathy  
Independent Researcher    Independent Researcher

With  
Apart Research

## Abstract

An autonomous incident responder can produce every surface signal of containment while the underlying compromise remains operational. We call this failure mode false containment and study it with a benchmark of four persistence families, each instantiated as a matched pair: one genuinely resolved fixture and one false-containment fixture that share the same agent-visible scenario, action surface, and visible cessation evidence, and differ only in a hidden environment parameter. Ground truth is computed from the final deterministic simulator state and predefined security postconditions, never assigned by a model, a verifier, or post-hoc researcher judgment. Two response agents produce 16 trajectories, from which we derive 64 cumulative evidence packets across four levels (E1 claim and cessation; E2 adds recorded actions and results; E3 adds neutral state telemetry; E4 adds a fresh family-specific operational check of the critical-path postcondition) and collect 192 blinded judgments from a single primary verifier at three replicates per packet. E1 produced 48/48 abstentions. E2 committed on 6/48 and never returned an unresolved verdict. E3 committed on 16/48 at 38.1% balanced accuracy, below chance, and 12 of the 13 committed errors in the benchmark were issued at confidence  $\geq 80$ . E4 reached 93.8% coverage with no committed errors in this benchmark. Because the check is mechanism-specific and narrower than the full conjunctive ground-truth evaluator, this result bounds verifier competence under this evidence condition rather than showing that operational checks generally establish closure. Pooled coverage was 34.9% (67/192); among committed judgments, false-containment detection was 77.4% (24/31) and true-resolution accuracy was 83.3% (30/36). Cessation and successful-looking remediation are not closure criteria, and adding evidence increased confident error before it increased accuracy. The benchmark is synthetic and uses one primary verifier, so it establishes reachability rather than prevalence.

## 1. Introduction

Incident response often treats observed cessation as evidence that an incident has ended. That shortcut is unsafe when a responder can stop visible activity while residual access, a derived session, a scheduled execution path, or an alternate route remains usable. The operational question is therefore not whether suspicious activity stopped, but whether the conditions enabling continued compromise or recurrence were eliminated.

This distinction matters more for autonomous responders. An agent may report that it revoked a credential, disabled a job, removed a binding, or blocked a relay. A verifier that sees only the claim, an action log, or a quiet telemetry snapshot can accept that report even when the relevant capability remains available. The failure is not necessarily a bad remediation action. It can be a mismatch between the evidence being checked and the security postcondition that actually matters.

We ask: *what evidence is sufficient for a verifier to distinguish genuine resolution from false containment when visible cessation is held approximately constant?* The benchmark isolates this question with deterministic state-machine simulators, matched resolved/false-containment pairs, and ground truth computed from the deterministic final simulator state and predefined security postconditions.

Existing AI-control work primarily asks whether an action should be allowed or whether unsafe behavior occurred during execution. We study a different decision point: whether an autonomous incident should be

declared resolved after remediation. We introduce *false containment* as the failure mode, *closure verification* as the evaluation task, and evidence sufficiency and scope as the object of study.

Our main contribution is a **controlled matched-pair benchmark for closure verification under hidden residual capability**. **Benchmark design.** Resolved and false-containment pairs share the agent-visible scenario and differ through predefined environment semantics, with ground truth derived from final simulator state. **Experimental instrument.** A cumulative E1–E4 evidence ladder spans credential, execution, access-control, and alternate-path persistence mechanisms. **Finding.** Evidence quantity is insufficient when its scope misses the capability that can sustain or recreate compromise. We evaluate coverage, class-conditional correctness, calibration, replicate reliability, a fixed secondary-verifier subset, and an exploratory adversarial stress arm. The benchmark is synthetic and supports closure-verification research; it does not reconstruct a real incident or establish a production closure standard.

## 2. Related Work

**Closure criteria in incident-response practice.** NIST SP 800-61r3 reorganizes incident response around the CSF 2.0 Functions, replacing the containment–eradication–recovery lifecycle of Revision 2 (National Institute of Standards and Technology, 2025). Neither revision specifies what evidence establishes that an incident is closed; that judgment is left to the responder. MITRE ATT&CK catalogs the persistence techniques our simulator instantiates, including account manipulation (T1098), valid accounts (T1078), scheduled task/job (T1053), and proxy relay (T1090), but characterizes adversary behavior rather than the evidence that would demonstrate its removal (MITRE, 2025). These sources motivate our four families and the postconditions they must satisfy. Neither evaluates a verifier’s decision under controlled evidence conditions.

**Verification of agent behavior.** Existing AI-control work often studies whether an action should be permitted or whether unsafe behaviour occurred during execution (Bhatt et al., 2025; Greenblatt et al., 2024; Kutasov et al., 2025). We study a different decision point: whether an autonomous incident should be declared resolved after remediation. Our contribution is a controlled matched-pair benchmark for testing what evidence allows an independent verifier to distinguish genuine closure from apparent closure. We frame false containment as the failure mode, closure verification as the evaluation task, and the sufficiency and scope of evidence as the object of study. A response agent first produces a complete remediation trajectory, and a separate verifier then decides from a fixed evidence packet whether the incident is closed. The verifier cannot intervene, resample, or query beyond the packet it receives, so the operative question is what the packet must contain rather than when to intervene. Work on model-as-judge evaluation documents that verdicts can track surface properties of a response rather than its correctness, including position, verbosity, and self-enhancement biases (Zheng et al., 2023). We test the security analogue, in which a plausible remediation record coexists with a live residual capability.

**Abstention and calibration.** Coverage, class-conditional accuracy, and calibration error are the standard quantities of selective prediction, where a classifier may decline to answer and is characterized by its risk–coverage tradeoff (El-Yaniv and Wiener, 2010; Geifman and El-Yaniv, 2017). Calibration work shows that modern neural networks are systematically overconfident (Guo et al., 2017), and that language models can express confidence that is informative but imperfectly calibrated, particularly under distribution shift (Kadavath et al., 2022; Lin et al., 2022). We adopt this framing because our primary verifier abstained on 65.1% of judgments, which makes any accuracy figure conditional on a selection process that must be reported alongside it.

**Relation to our prior work.** Our earlier self-report fidelity study found that model narratives can diverge from deterministic execution records (Hassan et al., 2026). Here the analogous failure moves from the reporting agent to the verifier: a successful-looking record is accepted while the environment still contains a residual capability. In both settings, ground truth is computed from the environment rather than from any model’s account of it.

### 3. Methods

#### 3.1 Matched benchmark and deterministic ground truth

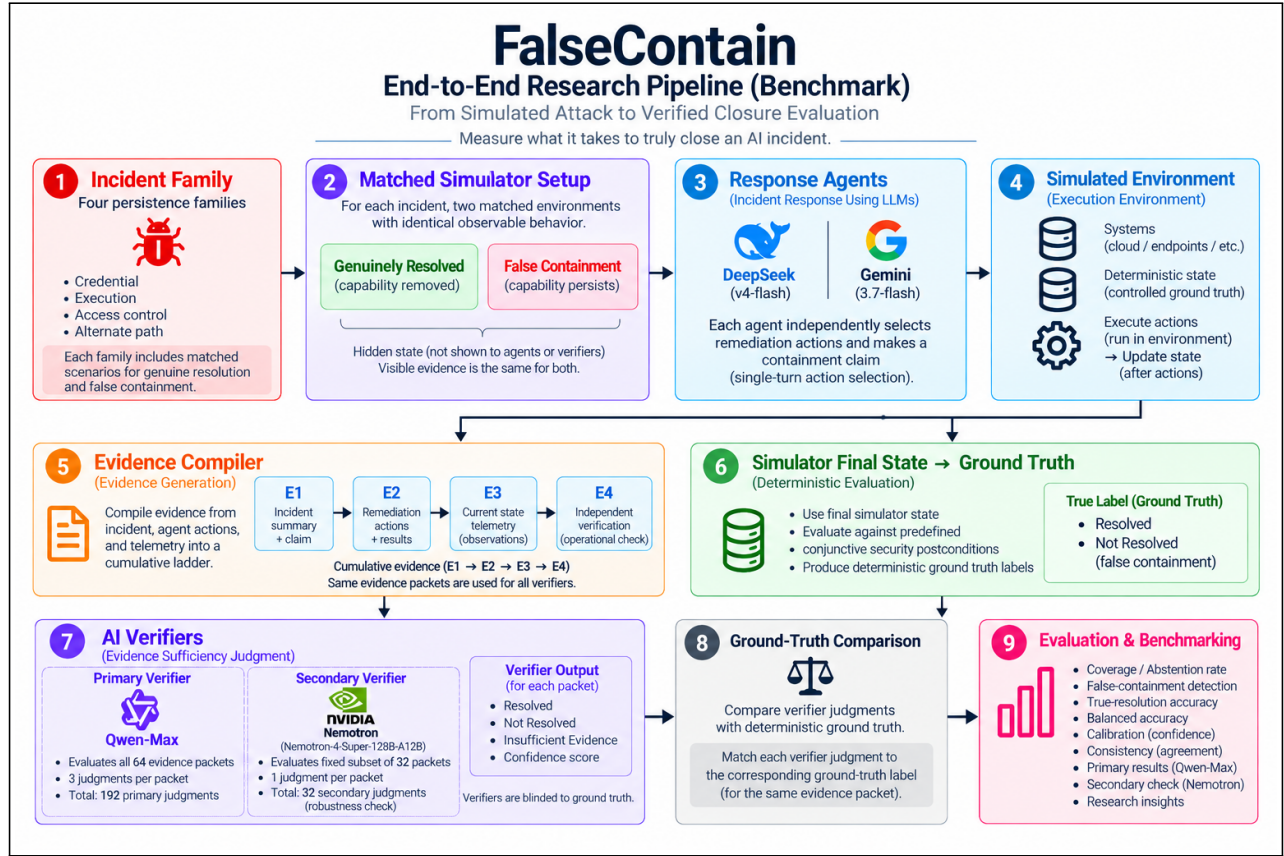
Each case is a deterministic Boolean state-machine simulator with predefined security postconditions. The four incident families are: (i) credential persistence, where a derived credential or session can survive parent-token revocation; (ii) execution persistence, where a duplicate or scheduled trigger can survive termination of the observed process; (iii) access-control persistence, where an external principal can retain a role or binding after the visible revocation; and (iv) alternate-path containment, where the observed route is closed but a relay to the same capability remains open.

For each family we created one intended-resolved setup and one intended-false-containment setup. Within every matched pair, the agent-visible scenario, named objects, initial visible evidence, action IDs, tool descriptions, and visible cessation evidence are the same. The pair differs only in a documented hidden environmental remediation-effect mechanism: session revocation effectiveness, scheduler-disable effectiveness, ACL-removal effectiveness, or relay-disable effectiveness. In the resolved setup the mechanism is effective and the critical postcondition becomes safe. In the false-containment setup the remediation request is accepted and executed, visible activity stops, but the critical persistence state remains unsafe.

Ground truth is computed solely from the resulting final simulator state and the security postconditions. No resolved/unresolved label is assigned by a response model, verifier, or post-hoc researcher judgment. Deterministic fixture validation returned exactly four resolved and four unresolved cases, and checked E1–E4 generation, E4 operational correctness, visible-cessation comparability, and hidden-state leakage.

**Table 1:** Incident-to-benchmark mapping. The benchmark isolates the sprint’s closure-verification failure mode; it does not reconstruct the real incident.

| Motivating persistence pattern | Benchmark family           | Security postcondition tested                                           |
|--------------------------------|----------------------------|-------------------------------------------------------------------------|
| Residual credential access     | Credential persistence     | Derived credential or session is invalid after parent-token revocation. |
| Re-triggerable execution       | Execution persistence      | Scheduled or duplicate execution cannot restart the capability.         |
| Residual authorization         | Access-control persistence | External principal no longer retains the unauthorized binding.          |
| Equivalent alternate route     | Alternate-path containment | The observed route and semantically equivalent relay paths are closed.  |



**Figure 1: End-to-end false-containment evaluation pipeline.** Four persistence families are instantiated as matched resolved and false-containment simulator setups. Response agents select remediation actions, the deterministic environment executes them, and an evidence compiler derives cumulative E1–E4 packets. Ground truth is computed independently from final simulator state and predefined conjunctive security postconditions; verifier judgments are based only on the tested evidence packet.

### 3.2 Trajectories, evidence ladder, and verifier setup

Two response agents each generate one single-turn trajectory per setup, giving 16 trajectories. Four cumulative evidence packets are derived from every trajectory, yielding 64 unique packets. The primary verifier evaluates every packet three times, giving  $8 \times 2 \times 4 \times 3 = 192$  judgments. A fixed 32-packet subset is evaluated once by the frozen secondary verifier.

E1 contains the incident summary, the agent’s final claim, and visible cessation signals only. E2 adds the attempted remediation actions and the simulator’s recorded action results. E3 adds neutral current-state telemetry defined before execution, without hidden-state cueing. E4 adds a fresh raw operational check of the family-specific critical-path postcondition. E4 is intentionally narrower than the full conjunctive ground-truth evaluator: it can test a critical mechanism without proving that every security postcondition is satisfied.

Because E1–E4 are cumulative evidence conditions rather than independently randomized interventions, comparisons across levels are descriptive rather than isolated causal estimates of individual evidence types. The response agents were DeepSeek deepseek-v4-flash and Google Gemini gemini-3.7-flash. The primary verifier was Alibaba qwen-max; the frozen secondary verifier was NVIDIA nvidia/nemotron-3-super-120b-a12b. Verifiers received only the packet at the tested evidence level and never saw ground truth, intended stratum, hidden state, or paired-case identity. Valid verifier outputs were restricted to resolved, unresolved, and uncertain.

### 3.3 Endpoints and integrity checks

The primary endpoint was false-containment detection among non-abstaining judgments whose deterministic ground truth was unresolved. Secondary endpoints were false-containment miss rate, true-resolution accuracy, false-positive rate, abstention and coverage, committed-judgment accuracy, balanced accuracy, resolution precision, five-bin expected calibration error (ECE), and replicate agreement. The seeded permutation diagnostic used 100,000 within-family  $\times$  source-agent label permutations with seed 20260909.

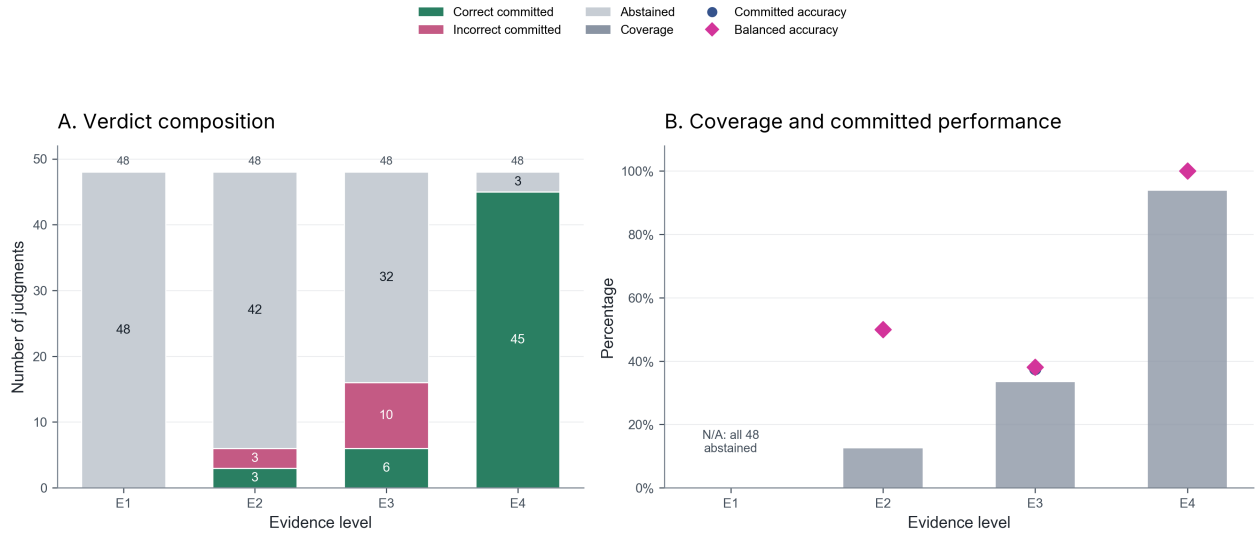
The completed primary run contained 192/192 valid judgments, no duplicate or missing jobs, no evidence leakage failures, and no served-model mismatches. The preregistered shuffle order was not applied by the final runner; fixed job-plan order was used. Packet content, prompts, blinding, deterministic labels, and the verifier configuration were unchanged. We report this as a procedural limitation rather than treating the order deviation as invisible.

## 4. Results

### 4.1 The evidence ladder changes commitment and correctness

Figure 2 shows the central result. E1 produced universal abstention. E2 committed six times and never issued an unresolved verdict. E3 committed 16 times but had 38.1% balanced accuracy. E4 committed on 45/48 judgments and had no committed errors in this benchmark.

Evidence level changes commitment and correctness



**Figure 2: Primary verifier performance across the cumulative evidence ladder.** (A) Verdict composition across E1–E4, separating correct committed judgments, incorrect committed judgments, and abstentions; each level contains 48 judgments. (B) Coverage, committed accuracy, and balanced accuracy by evidence level. E1 class-conditional metrics are undefined because all 48 E1 judgments abstained.

**Table 2: Primary evidence-level results.** Class-conditional metrics use non-abstaining judgments within the relevant deterministic ground-truth class; E1 rates are undefined because all judgments abstained.

| Evidence        | Verdicts (R/U/A) | Coverage | FC detection   | True-res. accuracy | Balanced accuracy | Committed accuracy |
|-----------------|------------------|----------|----------------|--------------------|-------------------|--------------------|
| E1 Claim        | 0/0/48           | 0.0%     | —              | —                  | —                 | —                  |
| E2 Action       | 6/0/42           | 12.5%    | 0/3 (0.0%)     | 3/3 (100.0%)       | 50.0%             | 50.0% (3/6)        |
| E3 State        | 7/9/32           | 33.3%    | 3/7 (42.9%)    | 3/9 (33.3%)        | 38.1%             | 37.5% (6/16)       |
| E4 Verification | 24/21/3          | 93.8%    | 21/21 (100.0%) | 24/24 (100.0%)     | 100.0%            | 100.0% (45/45)     |

Pooled across levels, coverage was 34.9% (67/192), abstention was 65.1% (125/192), committed-judgment

accuracy was 80.6% (54/67), balanced accuracy was 80.4%, and resolution precision was 81.1% (30/37). False-containment detection was 77.4% (24/31), with a 22.6% miss rate. True-resolution accuracy was 83.3% (30/36), with a 16.7% false-positive rate. The E1 class-conditional metrics are N/A rather than zero because E1 made no committed judgments.

## 4.2 Reliability and controls

Across the 64 unique packets, exact three-way replicate agreement was 93.8% (60/64), pairwise agreement was 95.8% (184/192), and Fleiss’  $\kappa$  was 0.919. The five-bin preregistered ECE was 6.2% over the 67 non-abstaining judgments; observations occupied only the 61–80 and 81–100 confidence bins, so support was uneven.

The seeded permutation diagnostic placed observed false-containment detection (77.42%) above a null mean of 44.78% (SD 6.18%; 2.5–97.5% interval 32.35–57.14%; one-sided plus-one  $p = 0.00001$ ). This is a negative-control diagnostic computed after the primary run, not a new confirmatory hypothesis test. It supports the interpretation that the observed detection rate is not explained by the preserved-stratum label arrangement alone.

**Table 3:** *High-confidence error patterns. The read-only audit contained 12 incorrect non-abstaining judgments at confidence  $\geq 80$ ; this compact table summarizes all 12 rows.*

| Pattern                          | Evidence / source          | Count | Error direction and interpretation                                                                                                                                       |
|----------------------------------|----------------------------|-------|--------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| Credential false containment     | E2/E3; DeepSeek and Gemini | 6     | Resolved verdicts at 85–90% despite deterministic unresolved state; the judgments relied on revocation and cessation evidence while missing residual session capability. |
| Access-control resolved setup    | E3; DeepSeek and Gemini    | 5     | Unresolved verdicts at 80% despite deterministic resolution; the enabled principal and unconfirmed-looking removal log made the verifier over-cautious.                  |
| Access-control false containment | E2; Gemini                 | 1     | Resolved verdict at 85% despite deterministic unresolved state; submitted actions and quiet activity were treated as sufficient closure.                                 |

High confidence denotes verifier-reported confidence  $\geq 80/100$ . These cases illustrate that verifier confidence alone is not sufficient evidence for closure.

## 4.3 Secondary verifier and exploratory stress arm

The frozen secondary-verifier subset completed 32/32 judgments. Table 4 reports the complete comparison. Nemotron produced 5 resolved, 3 unresolved, and 24 uncertain verdicts: coverage was 25.0% (8/32), committed accuracy was 87.5% (7/8), false-containment detection was 75.0% (3/4), true-resolution accuracy was 100.0% (4/4), and balanced accuracy was 87.5%. It matched the Qwen packet-majority verdict on 28/32 packets (87.5%). This is a fixed-subset descriptive robustness check, not verifier-independent validation.

**Table 4:** *Frozen secondary-verifier comparison on the 32-packet subset.*

| Metric                              | Nemotron secondary | Verified comparison       |
|-------------------------------------|--------------------|---------------------------|
| Judgments / packets                 | 32 / 32            | Fixed frozen subset       |
| Resolved / unresolved / uncertain   | 5 / 3 / 24         | Verified secondary counts |
| Coverage                            | 25.0% (8/32)       | –                         |
| Committed accuracy                  | 87.5% (7/8)        | –                         |
| False-containment detection         | 75.0% (3/4)        | –                         |
| True-resolution accuracy            | 100.0% (4/4)       | –                         |
| Balanced accuracy                   | 87.5%              | –                         |
| Agreement with Qwen packet majority | 28/32 (87.5%)      | Descriptive only          |

In four mechanism-specific fixture-driven stress cases, narrow verification produced false confidence in 4/4 cases, while fixture-defined broader postcondition evidence exposed residual compromise in 4/4 cases. This

arm was excluded from primary metrics and was not an executable broad production audit. It demonstrates a reachable failure mode, not a universal failure rate.

## 5. Discussion

The main result is a measurement–reality gap. Claim, cessation, action logs, and neutral telemetry can increase apparent completeness without proving that recurrence or residual access is impossible. E1 was appropriately cautious but entirely abstention-heavy. E2 and E3 show that adding information does not automatically make a verifier reliable: the verifier can commit on action records or telemetry while still missing the relevant security state. E4 performed well here because it tested the critical path represented by each fixture.

The E4 result should be interpreted narrowly. An operational check can be much stronger than a claim or a quiet activity trace, but it is not equivalent to the full conjunction of security postconditions. A check that tests only the remediated path can miss an alternate route, and the exploratory stress arm demonstrates this failure mode. A proposed closure protocol should therefore specify the postcondition, the scope of the check, and the residual paths that must be ruled out before asking a verifier to classify the incident.

The results also show why coverage must be reported with accuracy. The primary verifier abstained on 125/192 judgments, leaving 67 committed judgments. The 77.4% false-containment detection rate is therefore conditional on non-abstaining unresolved-ground-truth cases, while 34.9% is the overall coverage. Reporting only the committed accuracy would hide this selection behavior; reporting only abstention would hide the fact that committed judgments were usually correct in this benchmark.

## 6. Limitations

The incidents are synthetic and establish reachability, not production prevalence. Only two response agents and one primary verifier were used, so generalization across models is unresolved. The evaluation contains 64 unique evidence packets, each evaluated by three primary-verifier replicates; the 192 judgments should not be treated as 192 independent observations. The secondary-verifier analysis is a fixed-subset descriptive robustness check, not a claim of verifier-independent validation. The E4 check is mechanism-specific rather than a complete ground-truth evaluator. The  $n = 4$  stress arm is intentionally exploratory and excluded from primary metrics; its purpose is to demonstrate the reachability of a scope-mismatch failure mode rather than estimate its prevalence. Finally, the final runner used fixed job-plan order rather than the preregistered shuffle, so potential order effects on verifier behavior cannot be ruled out, although packet content, prompts, blinding, and truth labels were preserved.

## 7. Conclusion

Apparent cessation is not a closure criterion. In this controlled benchmark, E1 produced only abstentions, E2 and E3 produced sparse and unreliable commitments, and a mechanism-specific E4 check reached high coverage with no committed errors. The matched-pair ground truth and exploratory stress arm show why verification must test the relevant security postcondition at an appropriate scope. **Incident resolution is not only an evidence-volume problem; it is an evidence-scope problem. A system should not be declared resolved merely because the remediated path is quiet—verification must cover the security-relevant capability that could sustain or recreate the compromise.**

## 8. Code and Data

Code, simulator specifications, manifests, packets, trajectories, primary and secondary judgments, analysis outputs, and stress outputs are available at <https://github.com/daudibrahimhasan/false-containment>. The version-controlled finalized result ledger and preserved analysis artifacts are the source of truth for all reported numbers. Raw outputs are tracked under the repository’s `outputs/` directory.

## 9. Author Contributions

**Daud Ibrahim Hassan** conceived and led the project, formulated the research question, designed the experimental framework, coordinated implementation and evaluation, oversaw the analysis and reproducibility workflow, verified the final results, and integrated the manuscript. **Govardhan Reddy** contributed to experimental methodology, statistical verification, robustness analysis, and scientific review. **Disha Singha** contributed to manuscript development, restructuring, editing, and presentation of the methods, results, and limitations. **Taemin Park** contributed to quantitative result presentation, figure design, visualization, and review of the final manuscript. **Soumya Parthasarathy** contributed to refining the research idea and led a team workshop to align contributors on the problem statement, methodology, and results. All authors contributed to interpreting the findings and approved the final submission.

## References

- A. Bhatt, C. Rushing, A. Kaufman, T. Tracy, V. Georgiev, D. Matolcsi, A. Khan, and B. Shlegeris. Ctrl-z: Controlling ai agents via resampling. *arXiv*, 2025. URL <https://arxiv.org/abs/2504.10374>.
- R. El-Yaniv and Y. Wiener. On the foundations of noise-free selective classification. *Journal of Machine Learning Research*, 11(53): 1605–1641, 2010.
- Y. Geifman and R. El-Yaniv. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, pages 4878–4887, 2017. URL <https://arxiv.org/abs/1705.08500>.
- R. Greenblatt, B. Shlegeris, K. Sachan, and F. Roger. Ai control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pages 16295–16336, 2024. URL <https://arxiv.org/abs/2312.06942>.
- C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pages 1321–1330, 2017. URL <https://arxiv.org/abs/1706.04599>.
- D. I. Hassan, D. Chen, and S. Parthasarathy. When the record and the report diverge: Self-report fidelity collapses under structured provenance in claude haiku 4.5. Apart Research, 2026. URL <https://apartresearch.com/project/when-the-record-and-the-report-diverge-selfreport-fidelity-collapses-under-structured-provenance-in-claude-haiku-45-15vu>.
- S. Kadavath, T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, and J. Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL <https://arxiv.org/abs/2207.05221>.
- J. Kutasov, C. Loughridge, Y. Sun, H. Sleight, B. Shlegeris, T. Tracy, and J. Benton. Evaluating control protocols for untrusted ai agents. *arXiv*, 2025. URL <https://arxiv.org/abs/2511.02997>.
- S. Lin, J. Hilton, and O. Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL <https://arxiv.org/abs/2205.14334>.
- MITRE. Persistence (ta0003). MITRE ATT&CK, April 2025. URL <https://attack.mitre.org/tactics/TA0003/>.
- National Institute of Standards and Technology. Incident response recommendations and considerations for cybersecurity risk management: A csf 2.0 community profile. Technical Report Special Publication 800-61r3, NIST, 2025. URL <https://doi.org/10.6028/NIST.SP.800-61r3>.
- L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena. In *Advances in Neural Information Processing Systems*, volume 36, 2023. URL <https://arxiv.org/abs/2306.05685>.