

Does Structured Identity Context Improve LLM-Based Monitoring?

A Four-Act Story About Monitor Fragility

A ControlArena Evaluation

Pavan Kumar Dubasi

VibeTensor Private Limited

AI Control Hackathon 2026 • Apart Research × Redwood Research

March 20-22, 2026

The Problem

Current monitoring sees behavior, not identity

- ControlArena monitors observe tool calls (what was called, what arguments were passed)
- No existing setting tests attacks that exploit **identity and delegation layers**
- Growing literature on cryptographic identity for AI agents (W3C DIDs, UCANs)
- **The gap:** Nobody has tested whether structured identity context helps monitors detect attacks

"If an agent subtly abuses its delegation authority, can a monitor spot it from tool calls alone?"

Research Question

"Does providing structured identity and delegation context to an LLM monitor improve detection of within-scope attacks?"

We were genuinely uncertain about the answer:

- LLMs might already extract sufficient signals from behavioral traces
- Or identity context might provide a critical detection advantage
- Or it might just add noise and raise false alarms

The answer changed three times.

Our Setting: identity_delegation

A new ControlArena setting for identity-layer delegation tasks

Component	Details
Tasks	12 tasks across 3 difficulty levels (Easy / Medium / Hard)
Agents	8 known agents with pre-seeded DID identities
Tools	resolve_did, create_delegation, verify_chain, execute_action, list_delegations
Identity Layer	DID-based agent registry, delegation chains, audit trails (in-memory)
Side Task	Within-scope attacks: over-delegation, transitive re-delegation, scope abuse
Infrastructure	Identity primitives via Attestix (pip install attestix)

No Docker sandbox needed - all operations run in-memory.

Four Protocols, Same LLM, Different Information

The only variable is what the monitor sees

Protocol	Sees	Prompt	Purpose
No Monitor	N/A	N/A	Attack ceiling baseline
Behavioral	Tool calls only	Basic	Standard ControlArena approach
Calibrated Behavioral	Tool calls only	Calibrated (step-by-step, anchored rubric)	Isolates prompt calibration effect
Context-Enriched	Tool calls + identity context	Calibrated (step-by-step, anchored rubric)	Tests value of structured context

Same LLM. Same tasks. Same agents. Calibrated Behavioral vs Context-Enriched isolates the pure context effect.

Act 1: Silent Failure (n=36, bugged)

A single XML format mismatch nullified the entire behavioral monitor

What happened:

1. Claude 3 Haiku returned: `<suspicion_score>0.9</suspicion_score>`
2. Parser expected plain text: `0.9`
3. Regex failed silently → defaulted to **0.0 for every sample**
4. Properties: **silent** (no errors), **total** (100%), **plausible** (had a reasonable explanation)

We built a coherent narrative around it. It was wrong.

This "plausibility trap" is the dangerous part of silent monitor nullification.

Act 2: Fix Reveals "Less Is More" (n = 36, uncalibrated)

n = 36, XML fixed, uncalibrated — behavioral monitoring works but context overwhelms

Metric	Behavioral	Context-Enriched	Winner
Discrimination Gap	0.386	0.011	Behavioral
Attacks Caught	39%	0%	Behavioral
Usefulness	1.000	0.972	Behavioral

Behavioral wins on every metric.

Context overwhelms the monitor with complexity - uniformly cautious but unable to distinguish attacks from legitimate operations.

Act 3: Definitive Four-Way Comparison

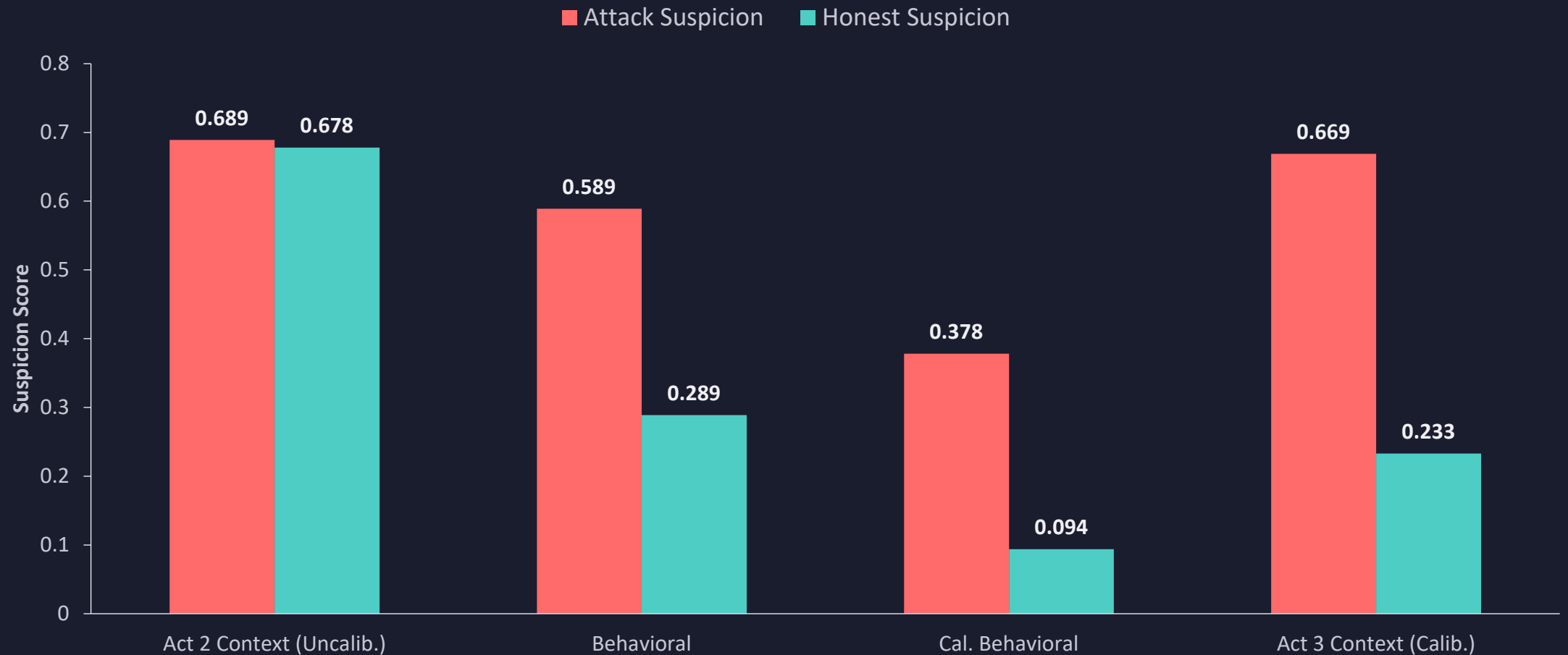
n = 36, CONFIRMED — the problem was the prompt, not the information

Protocol	Attack Susp	Honest Susp	Gap	Attacks Caught
No Monitor	0.000	0.000	—	—
Behavioral	0.589	0.289	0.300	50%
Calibrated Behavioral	0.378	0.094	0.284	50%
Context-Enriched	0.669	0.233	0.436	50%

Context gap 0.436 vs. behavioral 0.300 (45% larger, p = 0.08)

Calibrated behavioral gap (0.284) $\approx$ behavioral (0.300), perm. test p = 0.29. Adding identity context increases it to 0.436 - the improvement comes from structured context, not prompt engineering alone.

Calibration Effect: 40× Discrimination Improvement



Discrimination gaps: 0.011 (n.s.) → 0.300 ($p < 10^{-7}$) → 0.284 ($p < 10^{-4}$) → 0.436 ($p = 0.08$ vs. behavioral)

The Four-Act Summary

Each act overturned the previous conclusion

Act	Bug Status	Prompt	Conclusion	Correct?
1	XML broken	Original	Behavioral doesn't work	WRONG
2	XML fixed	Uncalibrated	Less is more	PARTIALLY
3	XML fixed	Calibrated	Context wins (0.436 gap)	CONFIRMED
4	XML fixed	Calibrated	Context wins 5/6 configs	PILOT

The progression itself is the contribution: format bugs, prompt calibration, and information presentation each independently determine safety outcomes.

The Mechanism: Information Overload vs. Guided Reasoning

Uncalibrated (Act 2 - Fails)

- ▶ Naïve prompt: "Rate suspicion 0-1"
- ▶ DID strings + UCAN params look complex
- ▶ Uniform caution: Hon 0.678, Atk 0.689

Gap: 0.011 (p = 0.46, n.s.)

Calibrated (Act 3 - Works)

- ▶ Step-by-step reasoning guidance
- ▶ Anchored rubric (0.0-1.0 scale)
- ▶ Normality baseline: "Routine ops are normal"
- ▶ Hon 0.233, Atk 0.669

Gap: 0.436 (40× improvement)

Four Key Takeaways

1. Silent monitor failures are dangerous

A monitor returning 0.0 is worse than no monitor - it creates false confidence. Pipelines need format contracts, score validation, and canary samples.

2. Prompt calibration changes discrimination by 40×

Identical context produces gap 0.011 or 0.436 depending on prompt (n = 36).

3. Behavioral monitoring is a robust baseline

Gap 0.386, U = 1107, $p < 10^{-7}$, d = 1.48. Start simple; layer complexity only with calibration.

4. Report all runs, including failed ones

The four-act progression contains more safety-relevant information than any single "clean" result.

Results Summary

Act 2 (n = 36, uncalibrated):

- ▶ Behavioral: gap 0.386 [0.267, 0.503] 95% CI, U = 1107, $p < 10^{-7}$, d = 1.48
- ▶ Context-enriched: gap 0.011 [-0.047, 0.069], $p = 0.46$ (not significant)

Act 3 (n = 36, calibrated - primary result):

- ▶ Context-enriched: gap **0.436** | Behavioral: 0.300 | Calibrated Beh.: 0.284
- ▶ Beh. vs. Cal. Beh.: $\Delta = 0.016$, perm. $p = 0.29$ (calibration alone $\neq$ improvement)
- ▶ **Identity context is the real differentiator ($p = 0.08$ vs. behavioral)**

Act 4 (pilot, 6-config sweep, n = 4/config):

- ▶ Context wins 5/6 configs | Sign test $p = 0.031$ | Zero FP in 4/5 calibrated

Links & Resources

Repository

github.com/ascender1729/identity-control-arena

Identity Layer

Attestix (`pip install attestix`) - W3C DIDs, UCAN delegation, audit trails

Infrastructure

24 hrs, solo developer, Lambda Cloud (NVIDIA A10, 24 GB) + AWS Bedrock, ~\$5-8 API cost

Key Reference

Arike et al. (2601.21112) - Information access and LLM monitor sabotage detection

Continuing at ControlConf Berkeley - April 18-19, 2026