Execution-Witness Binding: Proof-Carrying Authorization for LLM Agent Runtimes

Shawn Cohen
MVAR Security | Universal Media
February 2026

Abstract

Agentic AI systems that execute privileged actions via tool calls face a fundamental security challenge:
autoregressive language models process all tokens uniformly, making deterministic command-data separation
unattainable through training alone. Recent concurrent work independently established that deterministic
architectural enforcement is necessary for trustworthy agent deployment.

This paper presents MVAR (MIRRA Verified Agent Runtime), an open-source enforcement layer whose
architecture was developed independently and extends base deterministic enforcement with:
1) composition risk detection across session-level tool chains,
2) execution-witness binding for TOCTOU prevention and single-evaluation semantics,
3) one-time execution token replay defense with persistent nonce consumption,
4) deterministic declassification for scoped memory writes,
5) signed policy-bundle startup verification.

Key Contributions

- Composition Risk Engine:
  Sliding-window cumulative risk scoring (LOW=1, MEDIUM=3, HIGH=6, CRITICAL=10) hardens outcomes
  from ALLOW to STEP_UP/BLOCK as risk budgets are exceeded.
- Execution-Witness Binding:
  Planning-phase policy decisions become proof-carrying authorization witnesses verified at execution
  without re-running policy, preventing TOCTOU drift and double-count side effects.
- Persistent Replay Defense:
  One-time execution token nonces are consumed and persisted, surviving process restarts.
- Deterministic Declassification:
  Sensitive cross-scope memory writes require explicit, signed, one-time declassification tokens.
- Signed Policy Bundle Gate:
  Runtime startup verifies signed canonical policy bundles and fails closed on mismatch.

Validation Summary

- Reproducible 50-vector adversarial corpus (9 categories).
- Agent testbed trilogy (injection, taint laundering, benign).
- CI regression gates enforce deterministic behavior on every push/PR.
- Cross-framework adapter conformance across LangChain, OpenAI, MCP, Claude, AutoGen, CrewAI, and OpenClaw.

Threat Model and Scope

MVAR is a deterministic execution-boundary control plane. It assumes the enforcement layer is trusted and
LLM planner inputs may be adversarial. It does not claim completeness against all attacks and does not
replace OS/container sandboxing.

Repository

https://github.com/mvar-security/mvar

Reproduction

```
git clone https://github.com/mvar-security/mvar.git
cd mvar
python3 -m venv .venv && source .venv/bin/activate
pip install -e ".[dev]"
./scripts/launch-gate.sh
```

Keywords
agent runtime security, deterministic enforcement, prompt injection, information flow control,
proof-carrying code, execution witness, TOCTOU, composition attacks, replay defense.