"The difference between theory and practice

is larger in practice than in theory."

— Jan van de Snepscheut, Computer Scientist

Abstract

As multiagent AI systems transition from research prototypes to production infrastructure, their security properties remain largely unvalidated. While formal security frameworks promise principled protection, a persistent gap exists between theoretical guarantees and empirical evidence: *do these defenses actually work against real attacks?* This paper bridges that gap through comprehensive computational validation of the **Cognitive Integrity Framework (CIF)** introduced in Part 1.

We implement the complete CIF defense suite—cognitive firewalls, belief sandboxes, trust calculus with bounded delegation, identity tripwires, and Byzantine-tolerant consensus—and evaluate performance using **architecture-aware simulation** across topological models of six production multiagent systems, with a novel corpus of 950 cognitive attacks.

Contributions

- **Attack Corpus**: 950 cognitive attacks across four categories (prompt injection, trust exploitation, belief manipulation,