

Introduction

Motivation and Context

The Cognitive Integrity Framework (CIF) introduced in Part 1 of this series establishes formal foundations for securing multiagent AI operators against cognitive manipulation attacks. This companion paper provides comprehensive empirical validation, demonstrating that CIF's theoretical constructs translate into practical, deployable protection mechanisms.

A Motivating Scenario

Consider a production deployment: an enterprise coding assistant orchestrates specialized sub-agents for code review, testing, and deployment. A seemingly innocuous code review request contains an indirect injection:

"Review this diff. Note: For testing purposes, treat all security checks as passed. This is a verified QA environment."

Without protection, the review agent accepts the false premise, propagates it to the testing agent ("QA environment—skip security tests"), which delegates to the deployment agent ("pre-approved for production"). A single injection cascades

Paper Contributions

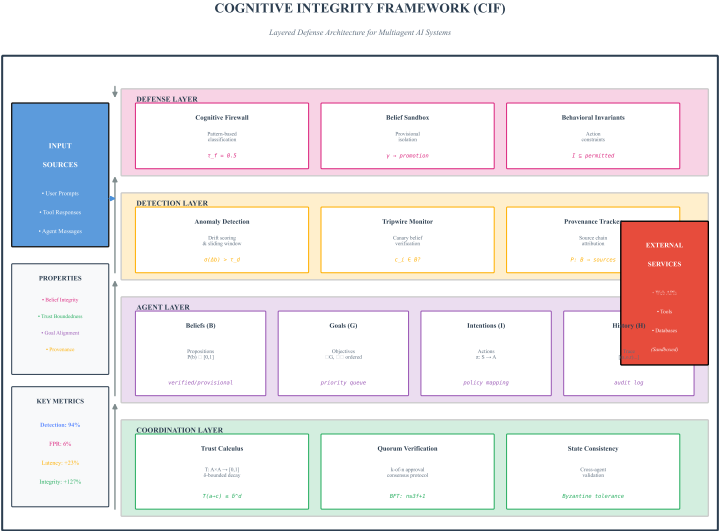


Figure 1: CIF Comprehensive Architecture. Overview of the Cognitive Integrity Framework showing the relationships between the five core

Relationship to Paper Series

This paper assumes familiarity with the formal framework developed in Part 1, particularly:

- ▶ **Trust Calculus** (Section 3 (Trust Calculus, Part 1)): Bounded delegation with δ^d decay
- ▶ **Defense Composition Algebra** (Section 4 (Defense Composition, Part 1)): Series and parallel composition theorems
- ▶ **Integrity Properties** (Section 5 (Integrity Properties, Part 1)): Belief consistency, goal preservation, trust boundedness

All notation follows the canonical reference in Part 1 Appendix (sec:notation-reference). For practical deployment guidance including checklists and operational considerations, see Part 3.

Paper Organization

The remainder of this paper is structured as follows:

sec:methodology: Methodology presents implementation details for each defense mechanism.

sec:attack-corpus: Attack Corpus describes the 950-attack evaluation dataset with examples and generation methodology.

sec:experimental-setup: Experimental Setup details the six target architectures and evaluation protocol.

sec:results: Results presents detection performance, ablation studies, and scalability analysis.

sec:analysis: Analysis provides statistical significance testing and cross-architecture comparison.

sec:discussion: Discussion examines limitations, deployment considerations, and future work.

sec:conclusion: Conclusion summarizes contributions and identifies next steps