COGNITIVE INTEGRITY FRAMEWORK (CIF)

Layered Defense Architecture for Multiagent AI Systems

INPUT SOURCES

- User Prompts
- Tool Responses
- Agent Messages

DEFENSE LAYER

Cognitive Firewall

Pattern-based classification

$\tau_f = 0.5$

Belief Sandbox

Provisional isolation

$\gamma \rightarrow promotion$

Behavioral Invariants

Action constraints

$I \subseteq permitted$

DETECTION LAYER

Anomaly Detection

Drift scoring & sliding window

$\sigma(\Delta b) > \tau_d$

Tripwire Monitor

Canary belief verification

$c_i \in B?$

Provenance Tracker

Source chain attribution

$P: B \rightarrow sources$

AGENT LAYER

Beliefs (B)

Propositions
$P(b) \in [0,1]$

verified/provisional

Goals (G)

Objectives
$\langle G, \prec \rangle$ ordered

priority queue

Intentions (I)

Actions
$\pi: S \rightarrow A$

policy mapping

History (H)

Trace
$[(a,o,r)...]$

audit log

COORDINATION LAYER

Trust Calculus

$T: A \times A \rightarrow [0,1]$
δ-bounded decay

$T(a \rightarrow c) \leq \delta^d$

Quorum Verification

k-of-n approval
consensus protocol

$BFT: n \geq 3f+1$

State Consistency

Cross-agent validation

Byzantine tolerance

KEY METRICS

Detection: 94%

FPR: 6%

Latency: +23%

Integrity: +127%