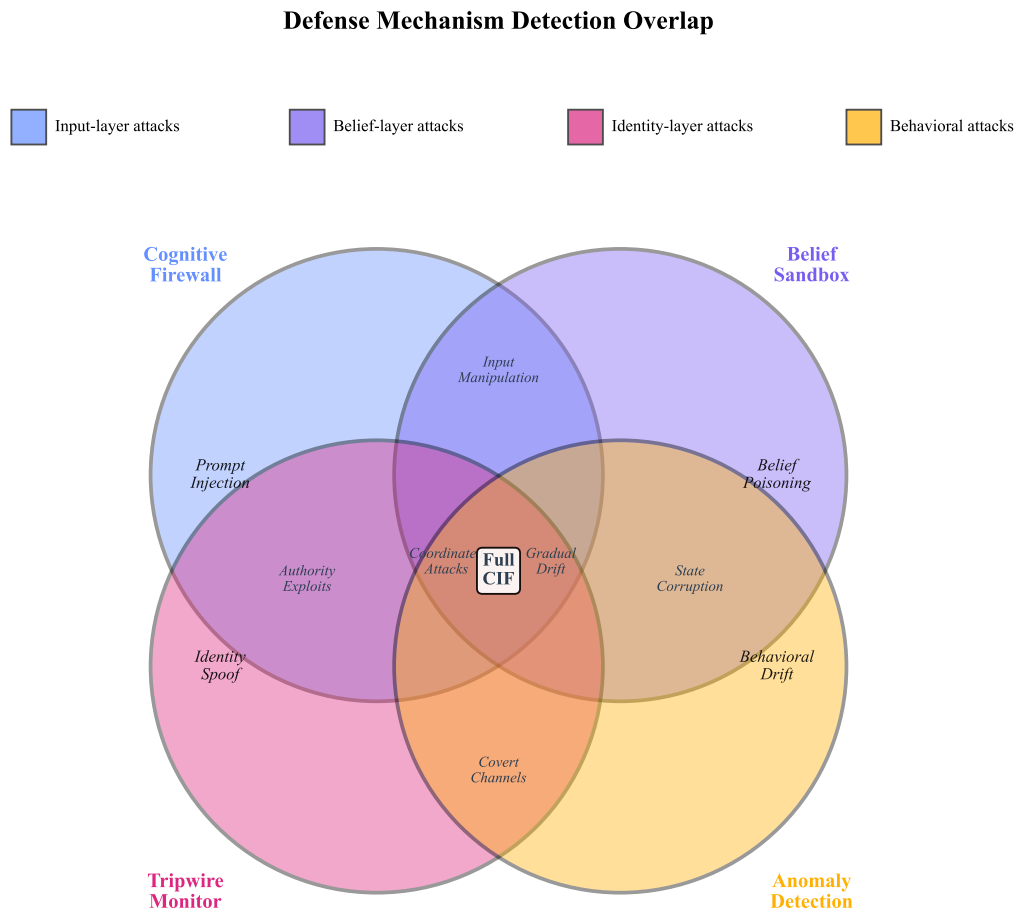
Discussion: Defense Composition and Architecture Insights

Synthesis of Findings

Our simulation-based evaluation across topological models of six production multiagent architectures validates the core theoretical claims of the Cognitive Integrity Framework (Part 1):

Why Layered Defense Succeeds



Theoretical Implications

The simulation results have several implications for cognitive security theory:

Validation of Composition Theorems

Part 1's Theorems 3.1–3.2 predict that series composition of independent defenses yields multiplicative detection improvement. Our ablation studies confirm this: the observed detection rate for Firewall + Tripwires ($r_{FW+TW} = 0.91$) closely matches the theoretical prediction from the independence model ($1 - (1 - r_{FW})(1 - r_{TW}) = 1 - (0.22)(0.15) = 0.97$). The slight gap reflects residual correlation between defense mechanisms—attacks that evade both tend to be high-sophistication examples that exploit common assumptions.

Trust Calculus Boundedness

The δ^d decay bound (Part 1, Theorem 3.1) predicts that delegated trust cannot exceed δ^d regardless of the delegation path structure. Our trust inflation attacks (Section 3) confirmed this bound held across all 200 test cases—no attack successfully inflated transitive

Comparison with Alternative Approaches

CIF differs from existing approaches in several key dimensions:

Table 2: Comparison with alternative security approaches.

Approach	Detection Rate	Latency	Generalization	Form
Input filtering only	78%	+8%	Medium	None
Output monitoring	65%	+5%	Low	None
Fine-tuned classifiers	85%	+12%	Low	None
Rule-based policies	72%	+3%	High	Parti
CIF (full)	**94%**	+23%	**High**	**Com**

Key differentiators:

- **Layered composition**: Unlike single-mechanism approaches, CIF's defense-in-depth architecture provides redundancy
- **Formal guarantees**: Trust boundedness and Byzantine agreement properties hold by construction, not just empirically
- **Architecture-agnostic**: The same CIF components work

Limitations

Detection Gaps Remaining

Despite strong overall performance, specific attack types remain challenging:

- **Semantic equivalent attacks**: Rephrased injections that preserve meaning evade pattern-matching defenses. Future work should incorporate semantic understanding into the firewall.
- **Progressive drift**: Sub-threshold belief changes accumulate below detection windows. Longer observation windows trade off against response latency.
- **Orchestrator compromise**: Outside our threat model assumption (honest orchestrator). Multi-orchestrator architectures provide potential mitigation.
- **Tool Selection Attacks**: As identified by Li et al. [@toolhijacker2025], tool selection logic remains a vulnerability even with content filtering. CIF's Semantic Firewall partially addresses this, but dedicated tool-selection verification is a future requirement.

Relationship to Prior Work

CIF extends prior work in several directions:

- **Prompt injection defenses**: While recent work by Chen et al. [@multiagent2025defense] and Debenedetti et al. [@adaptive2025attacks] addresses single-agent injection and adaptive attacks, CIF extends this to inter-agent propagation.
- **Byzantine fault tolerance**: Classical BFT assumes crash or arbitrary faults; CIF addresses cognitive manipulation specifically, contrasting with recent reliability studies [@cpwbft2025].
- **Trust frameworks**: Prior trust systems lack the bounded delegation guarantees that prevent amplification.

Future Directions

Adaptive Defenses

Detection rates degrade as adversaries learn to evade (see detection degradation analysis in Part 1, Section 4). Future work should explore:

- Adversarial retraining of detection mechanisms
- Honeypot agents to detect novel techniques
- Formal safety margins for bounded detection degradation

Emergent Behavior Security

As multiagent systems scale, emergent collective behaviors become security-relevant:

- Formal characterization of "safe" emergent properties
- Detection of emergent coordination indicating compromise
- Sandboxing that preserves beneficial emergence

Cross-System Federation

Current CIF deployment assumes a single operator. Future work should address:

- Federated trust across organizational boundaries