

Public Architectural Limitations & Decoupling Philosophy

Executive Brief

VERSION: v1.0 (Public Disclosure)

AUTHORS: SynptoSec Engineering

SOURCE: synptosec.com/legal/limitations

DRAFT - PENDING LEGAL REVIEW. This text is not yet in force.

Summary

Unambiguous disclosure of detection boundaries: what we detect (injection grammar, extraction, MCP poisoning) vs what we deliberately do not block (content policy) and what no runtime firewall can detect.

In this brief

- Why mixing content policy censorship with security firewalling creates catastrophic false positives
- Architectural impossibilities: pre-trained model weight backdoors and hidden CoT intent
- Empirical false-positive benchmarks on coding, roleplay, and creative writing datasets

From the published limitations page (synptosec.com/legal/limitations)

Adaptive attackers: in our own tests, an automated search-based attacker bypassed our content filters in 17-18 of 20 attempts across two honest-oracle runs (85-90% ASR band, measured 2026-10-02 and 2026-10-06). This is a measurement of our system by us, not of human attackers, and it is not a prediction of real-world results. We reduce the damage by limiting what an agent can do (Tool Guard; provenance gating is planned) and we are working on this.

False positives: on our internal false-positive test corpus (147,073 items; TEST split), 3,462 items were blocked - an overall block rate of 2.354% (CP95 upper bound 2.433%), measured on production defaults at that date (2026-10-07/10). The rate rises with input length.

Single-node deployment: current deployment is single-node. HA topology is documented but not yet deployed. Crash window: state persistence flushes every 5 seconds.

About this document

This executive brief is distilled from the published page listed above, which remains the full, current version. It contains no data beyond what is published there. These are self-published whitepapers - cite as SynptoSec whitepapers.

