

Defense-in-Depth Security Architecture

Eight complementary layers protecting autonomous systems









Threat Vectors

Instruction-Level
Prompt injection, reward hacking

Data-Level
Poisoning, indirect injection

Execution-Level
Tool manipulation, evasion

Defense Layers (Maturity Level)

-  **1. Input Validation**
Lexical filtering, semantic validation (85% effective)
-  **2. Instruction Isolation**
XML tags, system/user separation (85-90% effective)
-  **3. Retrieval Augmentation Security**
Source attestation, content verification (Emerging)
-  **4. Tool Execution Controls**
Approval gates, least privilege (95%+ effective)
-  **5. Watermarking & Attribution**
zkDL++ cryptographic proofs (Rapidly evolving)
-  **6. Privacy-Preserving Learning**
Federated learning, differential privacy (Proven at scale)
-  **7. Behavioral Monitoring**
Anomaly detection, decision auditing (Improving)
-  **8. Model Risk Management**
Fed Reserve SR 11-7, governance (Established)

2026 Security Landscape




Threat Evolution:

- 58% of attacks in 2026 will be agentic-driven (first time majority non-human)
- OWASP Top 10 LLM 2025: Prompt injection remains #1 critical vulnerability
- \$10B+ annual AI model IP theft; 98% fidelity model extraction possible

Defense Maturity:

- Tool execution controls: 95%+ effective (most mature layer)
- RAG security & watermarking: Rapidly evolving (monthly breakthroughs)

Maturity:

-  Established
-  Moderate
-  Emerging