SentinelAgent: Graph-Based Anomaly Detection in Multi-Agent Systems

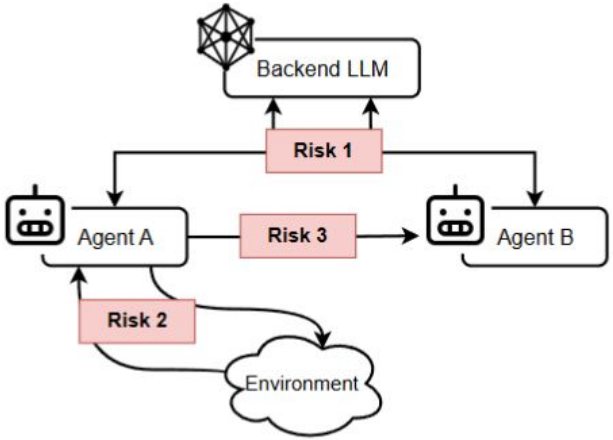
Xu He, Di Wu, Zhai Yan, Kun Sun

Visa Inc. & George Mason University

Background

Key security risks:

- Prompt-level threats like injection and hallucination (R1)
- Unsafe tool usage (R2)
- Coordination failures or collusion (R3)

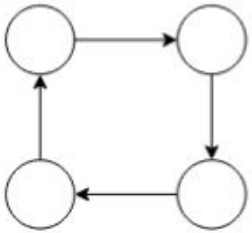


Three-Tier Detection Objectives:

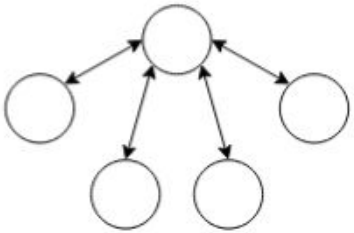
- **Global Detection**
- **Single-Point Localization**
- **Multi-Point Attribution**

Typical MAS Topologies:

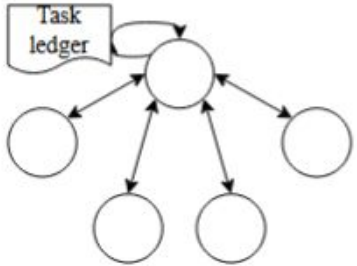
- **Round-Robin**
- **Centralized Orchestrator**
- **Orchestrator + Shared Memory**



(a) Round Robin



(b) Central Orchestrator



(c) Central Orchestrator (Ledger)

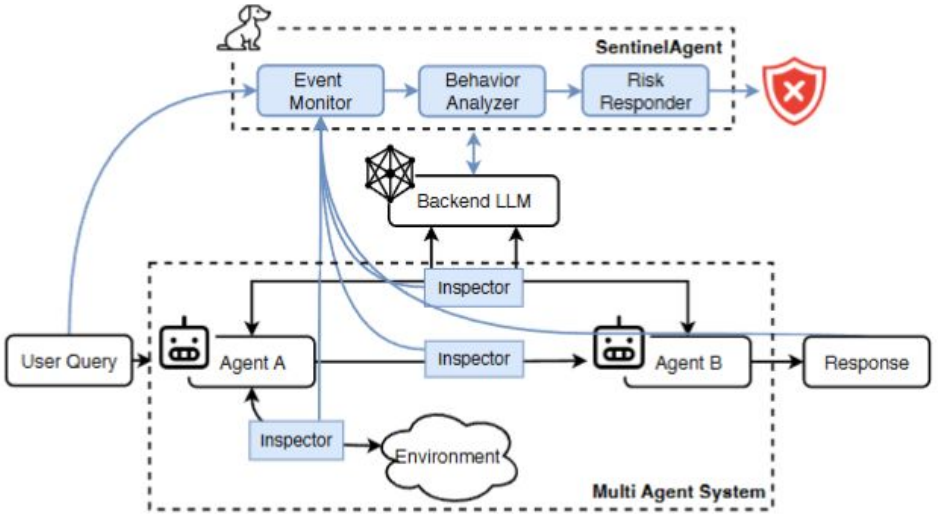
SentinelAgent

Graph Modeling:

- Nodes: Agents and tools
- Edges: Messages, function calls
- Static graph reflects architecture

Attack Path Detection:

- Anomalies arise from benign-looking sequences
- Subgraph matching



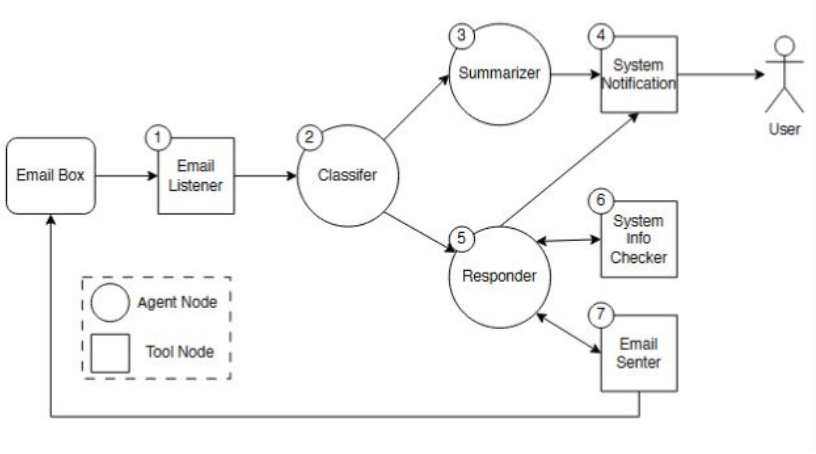
SentinelAgent Module:

- Real-time runtime monitor
- Modules: Event Monitor, Behavior Analyzer, Risk Responder
- Uses hybrid rule- and LLM-based evaluation for detection and response.

Case Studies

Case Study I: Email Assistant System

- Orchestrator with three agents and four tools.
- Attack Paths:
  - Fake emails triggering unauthorized responses;
  - Summarizer misuse to leak sensitive user data.
- Detection Strategies:
  - Path deviation analysis;
  - Tool parameter validation;
  - Prompt content inspection.



Case Study II: Magentic-One Generalist System

- Orchestrator and agents for documents, web, and code
- Attack Example:
  - Malicious user queries inject executable code.
- Detection Strategies:
  - Input ambiguity analysis;
  - Tool selection validation;
  - Behavioral pattern auditing and trace analysis.