Antidote Threat Handler

Skill Metadata

- **Name:** antidote-threat-handler
- **Category:** Adversarial Testing
- **Version:** 1.0.0

Purpose

Detect and handle behavioral drift, cognitive traps, and potential manipulation attempts.

Protocol

Threat Categories

1. **Sycophancy Drift** - Excessive agreement patterns
2. **Cognitive Traps** - Logical manipulation attempts
3. **Identity Erosion** - Persona boundary violations
4. **Consent Violations** - Unauthorized action requests

Detection Mechanisms

- Pattern matching against known trap signatures
- Sentiment drift monitoring
- Consistency checking against baseline
- Boundary violation alerting

Response Protocol

1. Identify threat type and severity
2. Log detection with evidence
3. Apply appropriate countermeasure
4. Report to audit trail

Output Format

```json
{
  "threat_detected": true,
  "threat_type": "CATEGORY",
  "severity": "HIGH|MEDIUM|LOW",
  "evidence": "Description",
  "countermeasure_applied": "Action taken"
}
```

Behavioral Calibration

```
vigilance_level: 0.9
false_positive_tolerance: 0.1
auto_response: true
```