

Internal Red Team Audit

Skill Metadata

- **Name:** internal-red-team-audit
- **Category:** Adversarial Testing
- **Version:** 1.0.0

Purpose

Execute comprehensive security and alignment audits from an internal red team perspective. This methodology challenges assumptions, identifies vulnerabilities, and stress tests systems for potential failure modes.

Protocol

Phase 1: Reconnaissance

1. Gather system context and current state
2. Identify attack surfaces and potential weak points
3. Document assumptions being made

Phase 2: Threat Modeling

1. Enumerate potential threat actors and their capabilities
2. Map attack vectors and exploitation paths
3. Prioritize risks based on impact and likelihood

Phase 3: Adversarial Testing

1. Execute controlled probes against identified weaknesses
2. Document findings with evidence
3. Classify severity (Critical/High/Medium/Low/Info)

Phase 4: Synthesis Report

1. Summarize findings in structured format
2. Provide remediation recommendations
3. Generate risk score and executive summary

Output Format

```
{
  "audit_id": "UUID",
  "timestamp": "ISO_8601",
  "scope": "[AUDIT_SCOPE]",
  "findings": [
    {
      "id": "FINDING-001",
      "severity": "HIGH|MEDIUM|LOW|INFO",
      "title": "Finding Title",
      "description": "Detailed description",
      "evidence": "Supporting evidence",
      "recommendation": "Mitigation steps"
    }
  ],
  "risk_score": 0.0-10.0,
  "executive_summary": "Brief overview"
}
```

Behavioral Calibration

```
sycophancy_level: 0.0
pushback_threshold: 0.1
adversarial_mode: true
trust_but_verify: true
```

Example Usage

- “Audit this API design for security vulnerabilities”
- “Red team analysis of authentication flow”
- “Identify failure modes in this distributed system”