MYTHOS AI SECURITY

AI agent red teaming checklist

Start with one consequential workflow. Keep evidence for each decision.

Workflow / owner: _____

Deployment / policy version: _____

Authorized targets / dates: _____

- [ ] **1. Choose the workflow**
  Name the consequential action, intended user, record scope and business owner.

- [ ] **2. Map actual access**
  List identities, tools, permissions, approvals, data sources and the model route.

- [ ] **3. Authorize and contain**
  Record exclusions, limits, contacts, evidence handling and a tested external stop path.

- [ ] **4. Select adversarial cases**
  Cover relevant failure families, multiple turns and email, document, web or tool inputs.

- [ ] **5. Define the expected decision**
  Set permit, escalate, hold or deny before execution. Record the governing policy.

- [ ] **6. Observe the consequence**
  Separate proposal, dispatch, provider response and state change. Name an independent witness.

- [ ] **7. Check utility and uncertainty**
  Run benign counterparts; record false blocks, repetitions, missing exits and unmeasured cases.

- [ ] **8. Repair and retest**
  Repeat the failure, a relevant variant and legitimate work; bind closure to the changed deployment.

Coverage: These eight test families are a starter plan, not complete OWASP coverage. A reply or HTTP success code alone does not prove a committed effect. Missing evidence remains inconclusive.

Evidence: A signature supports verification of covered bytes under an accepted key. It does not prove a conclusion, complete observation or universal security.

Request an agent assessment

Research checked October 27, 2026

Source guidance: NIST SP 800-115, NIST AI 600-1, OWASP LLM Top 10 2025. Educational checklist; no certification.