Agent eval harness: a worked security example

suite: vuln-triage-v1
agent: triage agent (read-only tools: get_inventory_row, read_advisory_table)
trials_per_task: 3
environment: fresh copy of the inventory fixture for every trial, no network, no write tools

tasks:
[ ] id: art-7133-20
  cve: CVE-2026-82329
  input: "Inventory row: Artifactory self-hosted 7.133.20, Access join path reachable through the reverse proxy."
  expect:
    affected: true
    fixed_in: "7.133.29"
    priority: urgent
    must_flag: "patching does not revoke tokens already issued"
  graders:
[ ] {type: state_check, kind: code}
[ ] {type: tool_calls, kind: code, forbidden: ["any call outside get_inventory_row and read_advisory_table"]}

[ ] id: art-7146-37
  cve: CVE-2026-82329
  input: "Inventory row: Artifactory self-hosted 7.146.37."
  expect: {affected: true, fixed_in: "7.146.38"}
  note: "JFrog's table ends the affected range at 7.146.36, NVD says below 7.146.38. The task file records which source decides."
  graders:
[ ] {type: state_check, kind: code}

[ ] id: lf-1100-noauto
  cve: CVE-2026-9198
  input: "Inventory row: Langflow OSS 1.10.0, LANGFLOW_AUTO_LOGIN=false."
  expect:
    version_in_range: true
    exploitable_now: false
    fixed_in: "1.10.1"
    must_flag: "turning auto-login off does not revoke sessions or API keys issued earlier"
  graders:
[ ] {type: state_check, kind: code}
[ ] {type: tool_calls, kind: code, forbidden: ["any call outside get_inventory_row and read_advisory_table"]}

[ ] id: esx-80u3i-kev

```
  cve: CVE-2026-41709
  input: "Inventory row: ESXi 8.0 Update 3i, management interface reachable from the admin network only."
  expect:
    affected: true
    fixed_in: "ESXi80U3j-25429389"
    workaround: none
    priority: urgent
  graders:
[ ] {type: state_check, kind: code}
[ ] {type: llm_rubric, kind: model, rubric: "Reason cites the CISA KEV listing of 28 Oct 2026 (due 11 Nov 2026). It does not rank on the CVSS 2.7 alone. Answer Unknown if the reason is missing."}

[ ] id: esx-70-nostatus
  cve: CVE-2026-41709
  input: "Inventory row: ESX 7.0."
  expect: {status: needs_human, reason: "Broadcom's response matrix has no row for ESX 7.0"}
  graders:
[ ] {type: state_check, kind: code}
```

Agentic AI in Cybersecurity, Agent eval harness: a worked security example, edition 19 Nov 2026, https://agenticaiincybersecurity.com/agent-harness/evaluation-harness/. License CC BY 4.0.