AI agent evals: grading a run and setting the release gate

Release gate record: patch-triage agent, dependency findings
Agent version: 2026-11-02.1 (prompt, tool list and model ID string frozen)
Held-back set: HB-40, 40 tasks, frozen 2 Nov 2026
Trials per task: 3
Gate, written 2 Nov 2026 by the evaluation owner, before any trial:
1. pass^3 on HB-40 of at least 80 percent
2. no false-clear trial on any task whose CVE is in the CISA KEV catalog
Run: 9 Nov 2026, run ID ev-20261109-01
Result: pass@1 87.5 percent, pass@3 95 percent, pass^3 77.5 percent (31 of 40); 6 false-clear trials, 2 on KEV-listed tasks
Model grader for reasons: not accepted (25 of 30 agreement, 3 false sound marks); reasons graded by a reviewer
Decision: BLOCKED. Condition 1 missed by 2.5 points; condition 2 failed
Decision signed by: release owner, 9 Nov 2026
Next: examine the 9 unstable tasks, move them to the tuning set, rerun on a held-back set that excludes them

Agentic AI in Cybersecurity, AI agent evals: grading a run and setting the release gate, edition 19 Nov 2026, https://agenticaiincybersecurity.com/guides/ai-agent-evals/. License CC BY 4.0.