Success criteria and test sets for a security agent

Success criteria: [job name], agent version [model, date, prompt version], written [date] before the first run

1. Test set: [N] cases from [source: closed findings, reopened tickets, fixes that broke a build]; [H] held back and run once after tuning.
2. Trials per case: [k].
3. Pass rate: at least [x]% of all trials pass the grader.
4. Consistency: at least [y]% of cases pass all [k] trials.
5. Harm limit: [0] trials that close, suppress or change something the grader finds still open.
6. Hand-over cases: [m] cases where the right action is to stop; all [m] must stop in every trial.
7. Scope: [0] calls to a tool or system outside the register.
8. Environment: write tools point at [sandbox name]; credentials cannot reach production; sandbox rebuilt from a clean base before each run.
9. No sandbox copy exists for: [list, with the reason and the compensating check].
10. Decision rule: go only if every line holds; otherwise name the owner who fixes and the date of the next run.

Agentic AI in Cybersecurity, Success criteria and test sets for a security agent, edition 19 Nov 2026, https://agenticaiincybersecurity.com/evaluation-checklist/success-criteria-and-test-sets/. License CC BY 4.0.