Writing a task spec for a security agent

TASK SPEC (one job, one agent)
Owner: a named person. Version and date: ____

1. Goal. One sentence: the outcome, the deadline and who owns it.
2. Inputs. Each data source, read-only or not. Mark any text from outside your control (advisories, tickets, emails, web pages, logs) as untrusted.
3. Allowed tools. Per tool: the exact function, read or write, reversible or not, the account it uses. Name tools by function. No open-ended tools such as "run a shell command" or "fetch any URL".
4. Out of scope. Systems, change types and data the agent must not touch, written as tools it does not hold, not as requests.
5. Stop conditions. Cases where the agent halts and hands over: unexpected tool result, outside text in context, a case the spec does not cover, retry limit, action rate limit. Who is paged.
6. Approval. Which actions wait for a named person, the stand-in, what that person sees, and what happens when nobody answers (the action does not run).
7. Evidence. Per step: inputs, decisions, tool calls with parameters, agent identity. Where it is kept and who cannot edit it.
8. Success and test set. Written pass criteria, a test set with known answers, the check run in the system after any change, and the rerun trigger (model, prompt, tool or data change).

Agentic AI in Cybersecurity, Writing a task spec for a security agent, edition 19 Nov 2026, https://agenticaiincybersecurity.com/guides/writing-a-task-spec-for-a-security-agent/. License CC BY 4.0.