

Homework — Design a Sandbox for a Ticket-to-PR Agent

~25 minutes · sandbox design exercise

This homework is a design exercise, not a coding exercise. You will write a sandbox specification for a realistic autonomous agent and then check it against the [reference answer](#). The point is to force yourself to make every policy explicit before you look at how someone else answered it.

Do not open [answer.html](#) until you have written your own five policies. The worked reference is there to check your reasoning, not to seed it.

Scenario & design questions

You are designing an autonomous AI agent that reads issues from a ticket system and opens PRs to fix them. The agent runs the model, git, and a test runner. Design its sandbox. Specify: (1) network policy — what it must reach and what it must not; (2) filesystem policy — what's read-only, what's read-write, where secrets live; (3) capability policy — what caps, what user; (4) secret scoping — what tokens, what scopes, how rotated; (5) escape budget — what's the worst case if the agent is fully prompt-injected?

Write out each of the five policies in one or two paragraphs. Be specific: name the paths, name the caps, name the token scopes. "It should be locked down" is not an answer.

✓ **Self-check: you have written all five policies, every policy names concrete mechanisms (not aspirations), and your escape-budget answer describes the**

worst-case outcome under the assumption that the model is fully compromised — not the best-case outcome under the assumption that it isn't.