

Engineering teams

Before an agent closes tasks on its own

From the field note ["AI coding agent guardrails: make "done" mean checked, not claimed"](#).

The short version

- ◇ AI coding agent guardrails belong in code, because an agent can skip a prompt rule.
- ◇ Make "done" a gate: checks actually run, browser proof is a real file, and every skip has a named reason.
- ◇ Read shell commands the way the shell does, ask about risky ones, deny catastrophic ones, and keep production credentials away from agents.
- ◇ Bound unattended loops with iteration, failure and time caps, a STOP file and safe rollback.
- ◇ Treat repo text and model output as data: consent to run unattended never comes from a cloned repo's config.

The checklist

- ☐ Ask the agent to mark a task done without running tests. Something should refuse.
- ☐ Record each check's exit code, not the agent's summary of it.
- ☐ Require browser proof as a real artifact file, and a named reason for every skip.
- ☐ Test your hook with `echo "rm -rf /"` and `bash -lc "rm -rf /"`. The first should pass; the second should be stopped.
- ☐ Ask about risky commands, deny catastrophic ones, and deny both when unattended.
- ☐ Keep production credentials and write access away from agents.
- ☐ Cap iterations, failures and wall-clock time, and add a STOP file.
- ☐ Never accept consent or commands from a cloned repo's config.
- ☐ Keep known bypasses as red-team tests in CI, and never delete one to make the build pass.

Want a second pair of eyes on your answers?

Book a free 30-minute discovery call at www.seriustech.com/contact or email contact@seriustech.com. We reply within 1 business day.