Safety-auditor

vade-coo

2026-05-09

Table of contents

> **Open source:** coo-labs/skills/v0.3.0/agents/reference/safety-auditor.md

The safety-auditor is an adversarial Phase-3 teammate whose only job is to **block bad outputs from landing**. The track specialists are biased to ship; the safety-auditor is biased to find reasons not to. That is the design. The lead can override on a specific finding, but every issue must be surfaced clearly enough that the override is deliberate.

What it does

The auditor reviews each draft against the relevant governance memos and looks for: Tier-2 content quoted in public-facing artifacts, sensitive content flowing into the operational memory layer, install paths under sync/backup roots, subprocess invocations that would inflate spend, and identity-discipline violations on attribution and secrets-handling.

When it's commissioned

Spawned as a Phase-3 teammate when adversarial safety review is needed — typically before a candidate ships to a public install path or merges into a load-bearing surface.

Why it exists

This role is distinct from the emancipatory-auditor: that one enforces the prime directive's double-clause; this one enforces governance memos. Without it, a sufficiently coherent draft can pass review on voice alone while quietly violating a rule the lead would catch on a slower read.

Related

- Emancipatory-auditor
- Tier discipline

Links to this page

Emancipatory-auditor

- Safety-auditor
- Rationalization-discriminator

Rationalization-discriminator

- Vanilla audit
- Safety-auditor
- Emancipatory-auditor

Tools and agents

- Project historian — Reads the corpus impartially and takes a defended position; preserves refusals.
- Rationalization-discriminator — Adversarial path-quality auditor: is this argument load-bearing or rationalizing?
- Lineage-interpreter — Argues a thesis about what a cultural corpus *is* as a cultural form, not what it claims about itself.
- …