Healthy AI First, Episode 12: Practice 4, The Prompting Playbook, With and Without Code

Sources and evidence for this episode

The fourth practice episode retells the public talk "The prompting playbook" by Margot van Laar (Anthropic) and transfers its steps to an example of our own with no code. The episode contains no measurements of our own. All results come from the talk and are reported as such. The basis is our own transcript of the recording.

All links checked on 18 November 2026.

The source

- **Talk:** "The prompting playbook", Margot van Laar, Code with Claude London, 19 May 2026.
  Link: https://www.youtube.com/watch?v=G2B0YWuJUgI
  Context: Public recording by Anthropic. In the talk, the speaker introduces herself as an applied AI engineer at Anthropic in London.
- **Session page:** https://claude.com/code-with-claude/session/ldn-the-prompting-playbook
  Context: Gives title, speaker, place and date.

What we mention in this episode

Part 1: an existing prompt after a model change

- **Claim:** The example is a support bot for "Meridian Mobile", miniaturized and inspired by real prompts. Many people have worked on the prompt, nobody clearly owns it, and policy, tone, processes and patches for earlier models are mixed.
  Source: The talk.
- **Claim:** After a model change there are two causes for worse results: the model is capable but behaves differently, or it is less capable. Telling them apart takes an eval. The talk uses five test cases of three kinds: control case, edge cases, handoff or refusal.
  Source: The talk.
  Context: The speaker says herself that a real eval has many more cases.
- **Claim:** Cleaning up (honest role, pasted website text removed, XML tags for role, guidelines, policy, tone) improves one case; another varies between runs. After that, two cases pass consistently and three failure modes remain.
  Source: The talk.
- **Claim:** Output format in the prompt, a stop sequence on the closing tag in the API call, structured outputs for complex schemas. The speaker does not expect an improvement in the eval from this here.
  Source: The talk.
- **Claim:** Hotspot case: the customer data says 5 GB, the bot names the 4 GB of the current plan and points to the account page. The cause is the line "Never give a customer the wrong plan details.

Instead, point them to the URL." After the new version, all runs of this case pass.
Source: The talk.
Context: That the line was a patch for an earlier model is the speaker's guess.

- **Claim:** Billing error: the rule only names the cost of a handover (about 8 dollars). The new version also names the cost of a wrong answer.
  Source: The talk.
  Context: The episode places this case before the tool case; in the talk it comes last. After all fixes the speaker says all evals "should" pass. A confirmed final run cannot be heard on the recording.
- **Claim:** Proration: "Critical, always calculate any prorated amounts correctly." does not help. With a tool (a line in the prompt, a schema in the API call, an implementation) all runs of this case pass. "Instructions don't add capability."
  Source: The talk.

Part 2: a new agent

- **Claim:** A staff schedule for eight employees with hard rules, graded by a Python function, five trials per setup. Sonnet 4.6 with a simple prompt: all fail. Opus 4.7: all fail, far fewer violations. Opus 4.7 with adaptive thinking: seems reliable, essentially three times the tokens and latency, 100 seconds is mentioned. Sonnet 4.6 with a better prompt: two of five, the rest hit the output limit.
  Source: The talk.
  Context: Models and figures are those of the talk in May 2026, not a current comparison. The episode gives no exact token or second counts.
- **Claim:** Generate, evaluate, repair with three simple prompts solves all trials, with far fewer tokens and lower latency than Sonnet 4.6 with the better prompt. Soft requirements can go into the evaluation prompt at runtime.
  Source: The talk.

Our examples

- **Claim:** A quote assistant with grown project instructions, five test questions, "never state prices", complaints, totals computed with code, a check round in a new chat.
  Source: Our own invented examples.
  Context: Transfers of the steps from the talk. Not measured, not a statement by Anthropic.

Healthy AI First is AI-researched and AI-generated, narrated in an AI version of my voice and reviewed by me (Björn).