

// SECTION 01 • PLAYBOOK

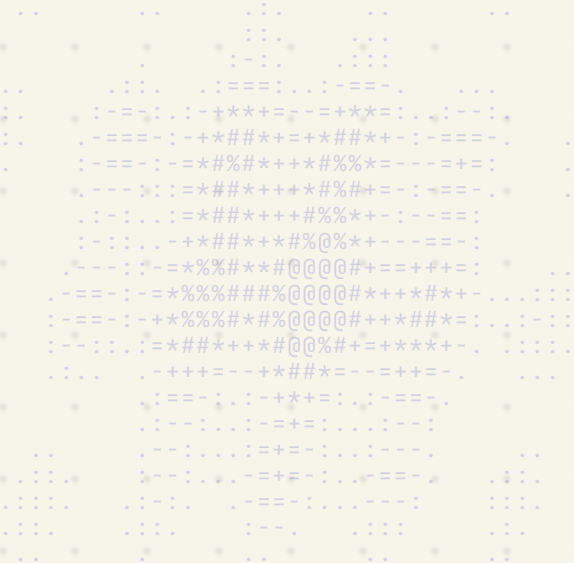
THE CANARY METHOD



A one-line instruction that warns you the moment your AI session starts to drift. Set it before you need it, watch it, and reset when it stops singing.

```
> boot opusjake_os
> resource: canary-method
> version: v1.0 • 2026
> status: ready_
```

Your AI does not break with an error message. It drifts. Message one it follows every rule you gave it. Fifty messages later it is inventing function names with the same confidence it had when it was right. A canary is the cheapest way to catch that slide: one small instruction you watch, so the model tells on itself the moment it stops listening.



WHY SESSIONS DRIFT

Degradation is quiet. No warning, no flag, just output that gets a little worse each turn until it is wrong. Four things cause it.

- **Context fills up.** Models pay the most attention to the start and end of the window and the least to the middle. The longer the chat, the deeper your original instructions sink into that dead zone.
- **Instructions get diluted.** In message one your rules are most of the context. By message fifty they are a rounding error under thousands of tokens of task talk, and the model quietly down-weights them.
- **Attention spreads thin.** Every reply re-reads the whole conversation. More history means less weight on any single line, including the one rule you care about.
- **Errors compound.** A wrong answer you do not catch becomes history. The model treats its own mistake as established fact and builds on top of it.

None of these announce themselves. That is the point.

THE CANARY

Coal miners carried caged canaries underground. The birds dropped from toxic gas before it reached the miners, so a silent cage meant get out now. Software borrowed the name: a canary deployment ships a change to a sliver of users first, to catch breakage before it hits everyone.

The AI version, popularized in the Claude community by Peter Steinberger (creator of OpenClaw), is simpler still. You add one tiny rule to your system prompt or CLAUDE.md.

Always start your response with my name.

While the model keeps doing it, the session is healthy. The second it forgets, your instructions have slipped out of attention and the rest of your rules are next in line. The canary costs you nothing and it fails loud.

THE TIMELINE

Most sessions degrade on the same curve. Once you have watched it a few times you can call it before it happens.

MESSAGES	PHASE	WHAT YOU SEE
1-15	Honeymoon	Every rule followed. Canary fires every time.
15-40	Subtle slip	Answers shorten, a format rule slides. Easy to miss.
30-60	Canary dies	Your rule gets dropped. This is the signal.
40-80	Drift	Confident guesses, fewer clarifying questions.
60+	Hallucination	Invented names, fake citations, authoritative and wrong.

The ranges overlap because no two sessions are identical. The order never changes. The canary dies in stage three, which buys you a head start on the two stages that actually hurt.

SEVEN CANARIES

Pick the one that matches what you cannot afford to lose. Each is a single line. Drop it in the instructions and watch it.

// NAME

Tests whether your system prompt still has attention. The default for everything.

Always start your response with my name.

// SIGN-OFF

Tests whether the model holds a rule all the way to the end of a long answer.

End every response with a horizontal rule and the word Ready on a new line.

// FORMAT

Tests whether your style rules still hold.

Always use bullet points, never numbered lists. Bold the first sentence of every paragraph.

// LANGUAGE

Tests whether the model can still suppress its own defaults.

Never use the words certainly or absolutely.

// ROLE

Tests whether the model is still in character.

You are a senior backend engineer. Reason from systems architecture first.

// CONSTRAINT

Tests whether hard limits still bind.

Keep every response under 200 words and include exactly three action items.

// META

Tests whether the model is tracking your instructions at all.

At the end of every response, rate your confidence from 1 to 10 that you followed all of my instructions.

RUN THEM

A canary only earns its keep if you place it where the model reads first and act when it goes quiet.

// SET THEM UP

- **Claude projects (web, desktop).** Project settings, Custom Instructions. Applies to every chat in the project.
- **Claude Code.** Put it in CLAUDE.md at your project root. It loads at the top of every session.
- **ChatGPT.** Settings, Personalization, Custom Instructions, the how-it-should-respond field.
- **Anything with a system prompt.** Gemini, Copilot, local models, raw API calls. Drop it in the system message.

// WHEN ONE DIES

- **Start fresh. Default to this.** A new session puts your instructions back at the top of a clean window. Claude web: new chat. Claude Code: `/clear` or `/compact`. ChatGPT: new chat.
- **Remind it once.** "You dropped my name. Re-read your instructions and keep following them." That buys a few messages. If the canary dies again, stop reminding and reset.
- **Save and migrate.** On a long build or writing session, have the model summarize everything that matters, open a fresh session, paste the summary, and restate the canary.

// RUN MORE THAN ONE

One canary tells you something is wrong. Two tell you how wrong. If your name canary dies but your format canary still fires, your prompt is fading, not gone. Both dead means the session is done. In an automated pipeline, wire a canary check to an automatic restart and never let a drifting model keep making calls.

STEAL THESE

Three stacks worth pasting. Swap in your name and go.

// THE ONE-LINER

Always begin every response with my name: [NAME]. This is a session health check. If you stop, I know the session is degrading and it is time to start fresh.

// THE FULL STACK

Start every response with my name: [NAME].
Never use the words certainly, absolutely, or I would be happy to.
End every response with a horizontal rule and the word Ready on a new line.
Keep responses under 200 words.
Rate your confidence from 1 to 10 that you followed every instruction.

// THE RESEARCH STACK

Session rules:

1. Start every response with my name: [NAME].
2. Rate your confidence 1 to 10 at the end of each response.
3. Flag any claim you are not sure of with [UNVERIFIED].

Break one and I know the session is no longer reliable.

Canaries are an early warning, not a post-mortem. Set one before you need it, keep half an eye on it, and when it goes quiet, take the hint and start clean. For more build-with-AI playbooks, follow opusjake.ai.