



FIELD GUIDE Nº 01

The Canary Trick

Catch your AI forgetting your instructions, and reset before it costs you.

5-minute setup

ChatGPT · Claude · Gemini

5 copy-paste prompts

01 • SETUP

Plant the **canary**

One line in your instructions. From then on, every reply has to start with it.

 Custom instructions

COPY

Begin every reply with the word "Noted."

Where it goes



ChatGPT

Settings → Personalization → Custom instructions. Or a Project's instructions.



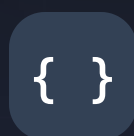
Claude

Settings (your initials, bottom left) → Instructions for Claude. Or a Project's instructions.



Gemini

Settings → Saved info. Or a Gem's instructions.



API

Your app's system prompt.



Pick a word you'd never type yourself, so it can't show up by accident. An emoji works too.

02 · SIGNAL

Read the **signal**

You don't measure anything. You just glance at the first word.

Tighten the intro.

Noted. Here's a tighter version: ...

✓ Canary there · instructions active

... 40 MESSAGES LATER ...

Now fix the ending.

Noted. Sure! Here's a stronger ending: ...

✗ Canary gone · instructions fading

Word there

Keep going. Your instructions are still being applied.

Word gone

Reset (page 5). Don't argue: every correction adds more text to an overloaded chat.

03 · UPGRADE

Sharper canaries

A Goal canary

Custom instructions

COPY

End every reply with one line: "Goal: <the current goal in 8 words>"

Catches drift even while the word survives. When the goal line stops matching your work, the chat has lost the thread.

B Rule canary

Custom instructions

COPY

Begin every reply with "Noted." and never use bullet points.

Ties the canary to your most important rule. If either one breaks, reset.

C Spot check

Send before important output

COPY

Without scrolling back: list the constraints and decisions we agreed on in this chat.

Missing items or rejected ideas in the list? Reset.

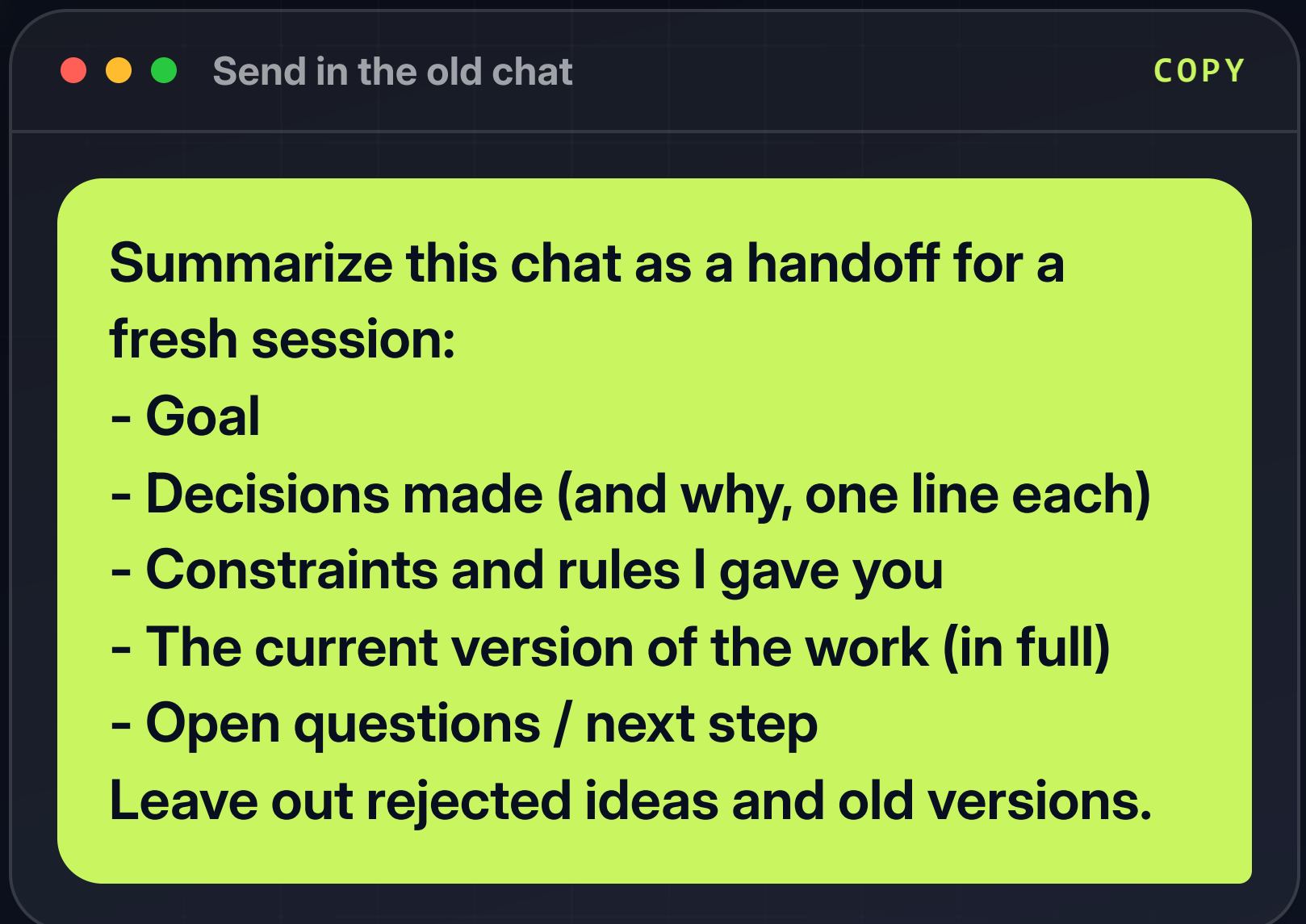
04 • RESET

The clean handoff

Don't repair a rotting chat. Move the good parts into a fresh one.

1

Ask the old chat for a handoff



2

Fix it (30 seconds)

Correct anything wrong. This is where you catch errors that were already there.

3

Paste into a fresh chat

Continue there. Your canary comes with you.

05 • TIMING

When to **reset**

- ✓ The canary disappears or the goal line drifts
- ✓ It repeats a mistake you already corrected
- ✓ It brings back an idea you rejected
- ✓ You switched to a new topic
- ✓ You pasted long documents and now ask detailed questions

Pro moves

✦ **Projects:** keep stable rules and files in the project, keep each chat short. Starting new chats costs you nothing.

✦ **Claude Code:** **/compact** condenses the session, **/clear** starts fresh. Long-lived rules go in CLAUDE.md.

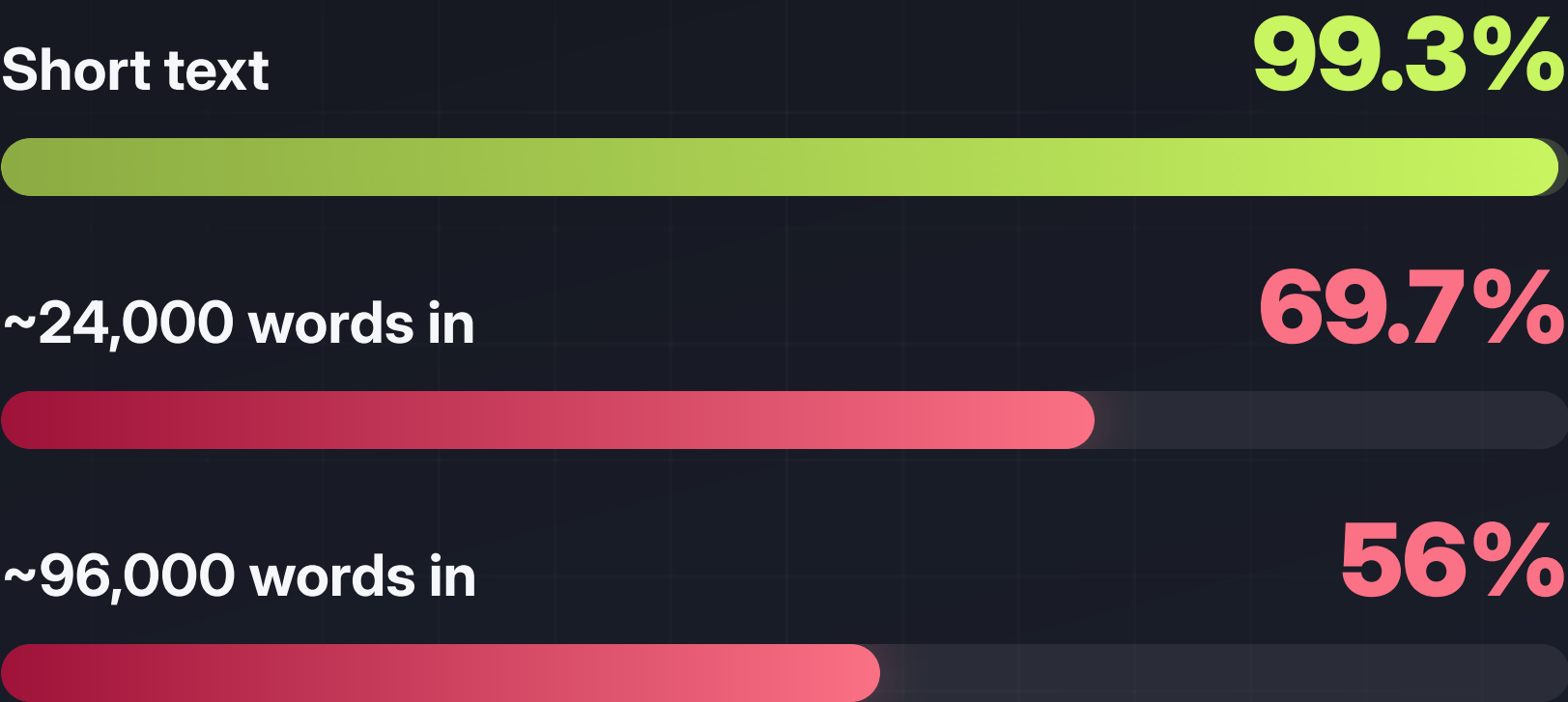
! **Memory isn't a fix.** It remembers facts about you across chats. It doesn't rescue one overloaded chat.

06 · THE SCIENCE

Why it works

Same task, more text around it

GPT-4o finding one hidden fact



11/13

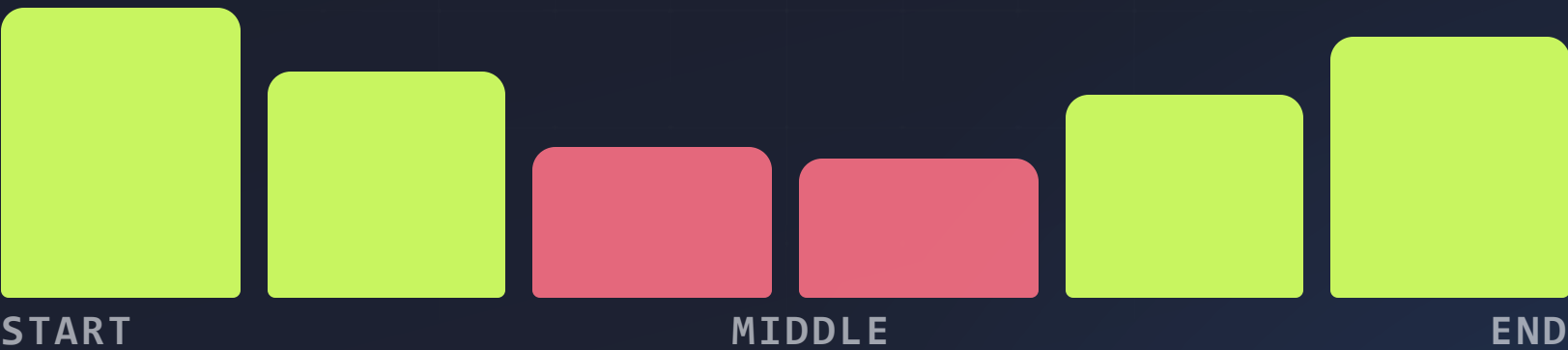
models fell to half their short-text score or worse at ~24,000 words

18/18

frontier models got less reliable as the input grew, even on simple tasks

The middle gets missed

Models use the start and the end of long input best.



CHEAT SHEET • SCREENSHOT THIS

The whole trick on one page

PLANT

"Begin every reply with the word Noted." in your custom instructions

WATCH

Word there: keep going. Word gone: reset.

UPGRADE

Goal line at the end of every reply + a spot check before important output

RESET

Handoff summary → fix it → paste into a fresh chat

PREVENT

One topic per chat. Rules in Projects, not in the chat.

More field guides like this

Al past the basics. New guide with every post.

@pasttheprompt