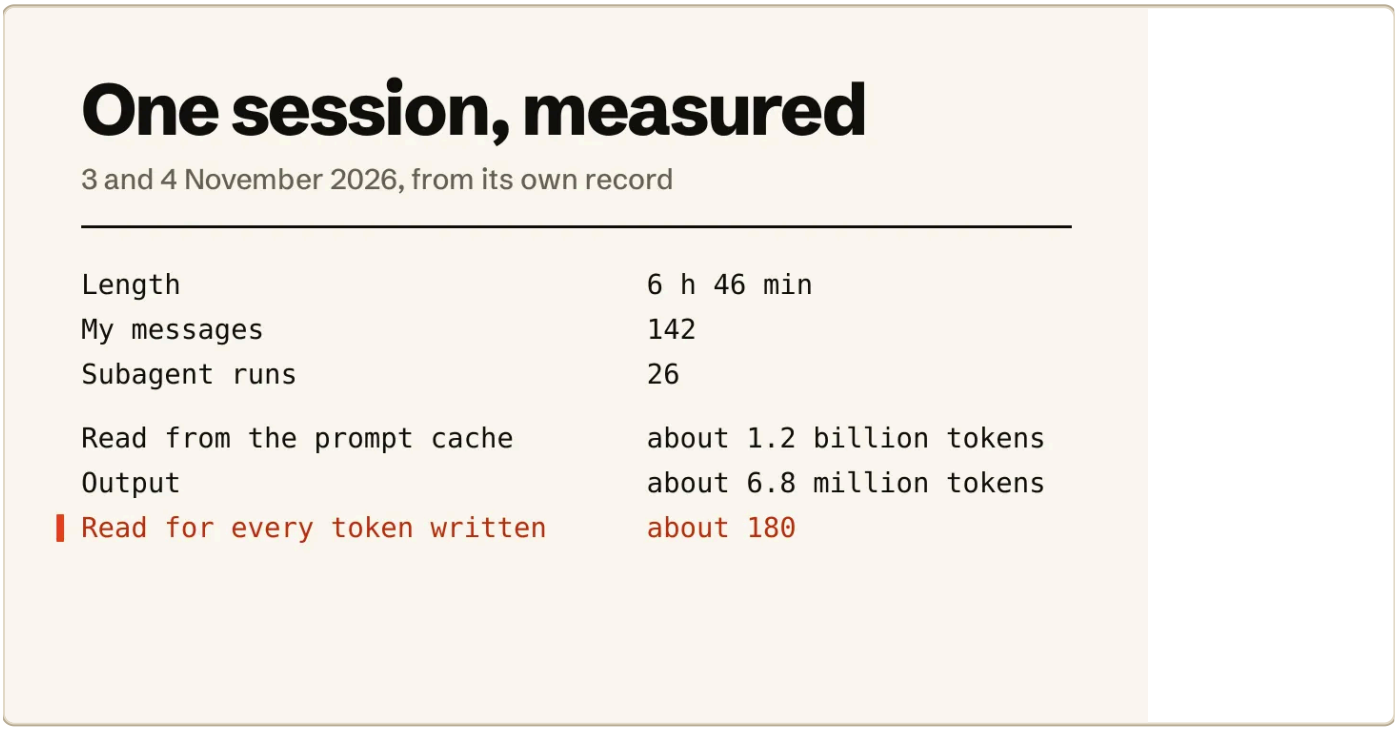
One session, about seven hours, 26 subagent runs: what I measured

I measured a long working session from its own record: about seven hours on 3 and 4 November 2026, 26 subagent runs, 142 messages from me. Re-reading context made up nearly all of the tokens.



Published
4 November 2026

Reading time
1 minute

Tags
AI

I like to measure my own setup before I try to improve it. So here is one long working session, measured from its own record on 4 November 2026.

- It ran for about seven hours: 6 hours 46 minutes from the first message to the last, across 3 and 4 November.
- I typed 142 messages. This was not an unattended job. I was there for it.
- Claude's main thread produced 1,520 messages, and ran 371 shell commands, 127 edits and 52 file writes.
- There were 26 subagent runs alongside it.

The number that matters

About 1.2 billion tokens were read from the prompt cache, across the main thread and the agents. The output from both came to about 6.8 million tokens. That is roughly 180 tokens read for every token written. The ratio is my arithmetic from those two totals.

Every turn re-reads the whole conversation so far, so the cost of a long session grows with how big the context is and how many turns it takes. A session that does many different jobs pays for all of them on every turn.

So the biggest lever isn't the price of a model. It's how much gets re-read. I'll measure each change I make against these figures, and I'll report what moves.

What this doesn't say

It says nothing about what the session produced, and I haven't turned tokens into money. It's one session, and a different kind of day would look different.