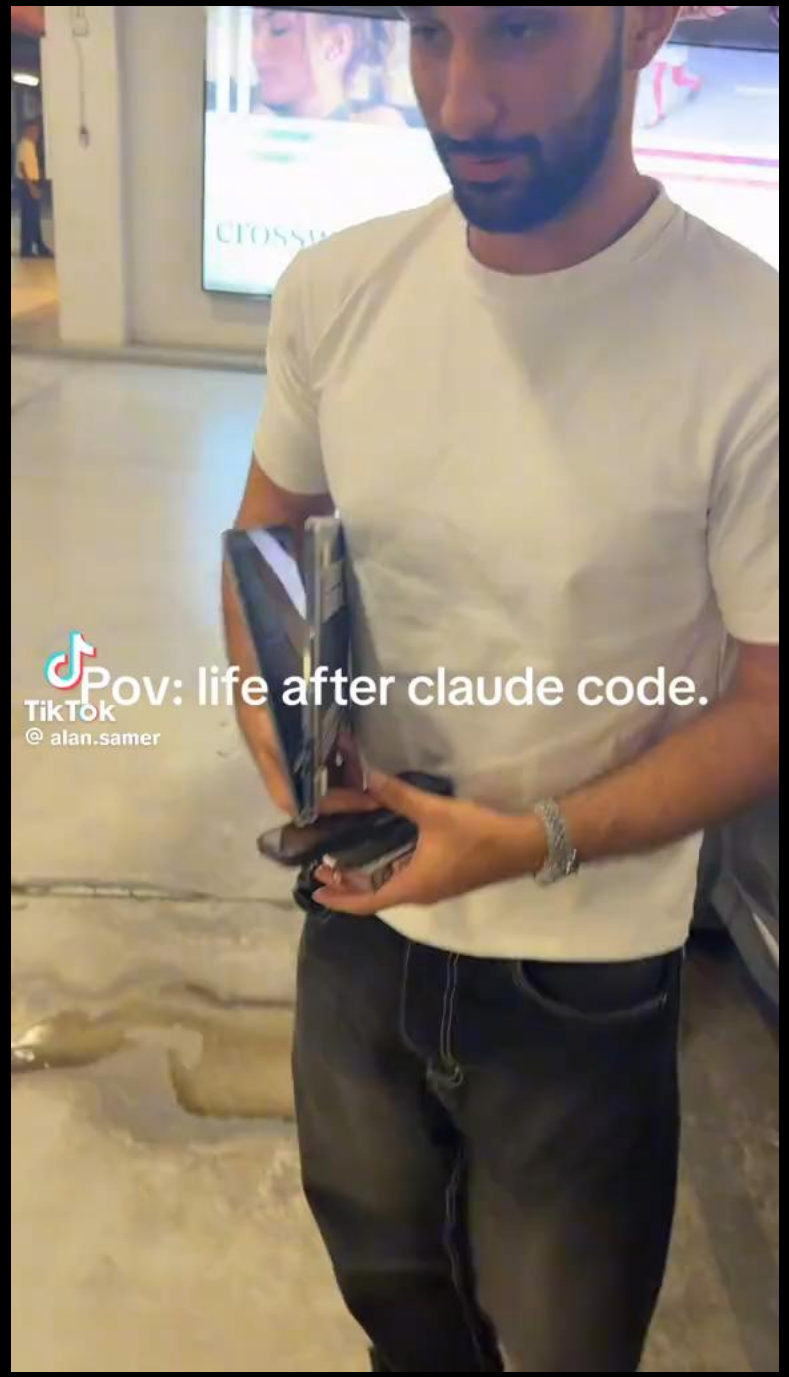
Where did all your tokens go?

Measuring coding agent impact with OpenTelemetry

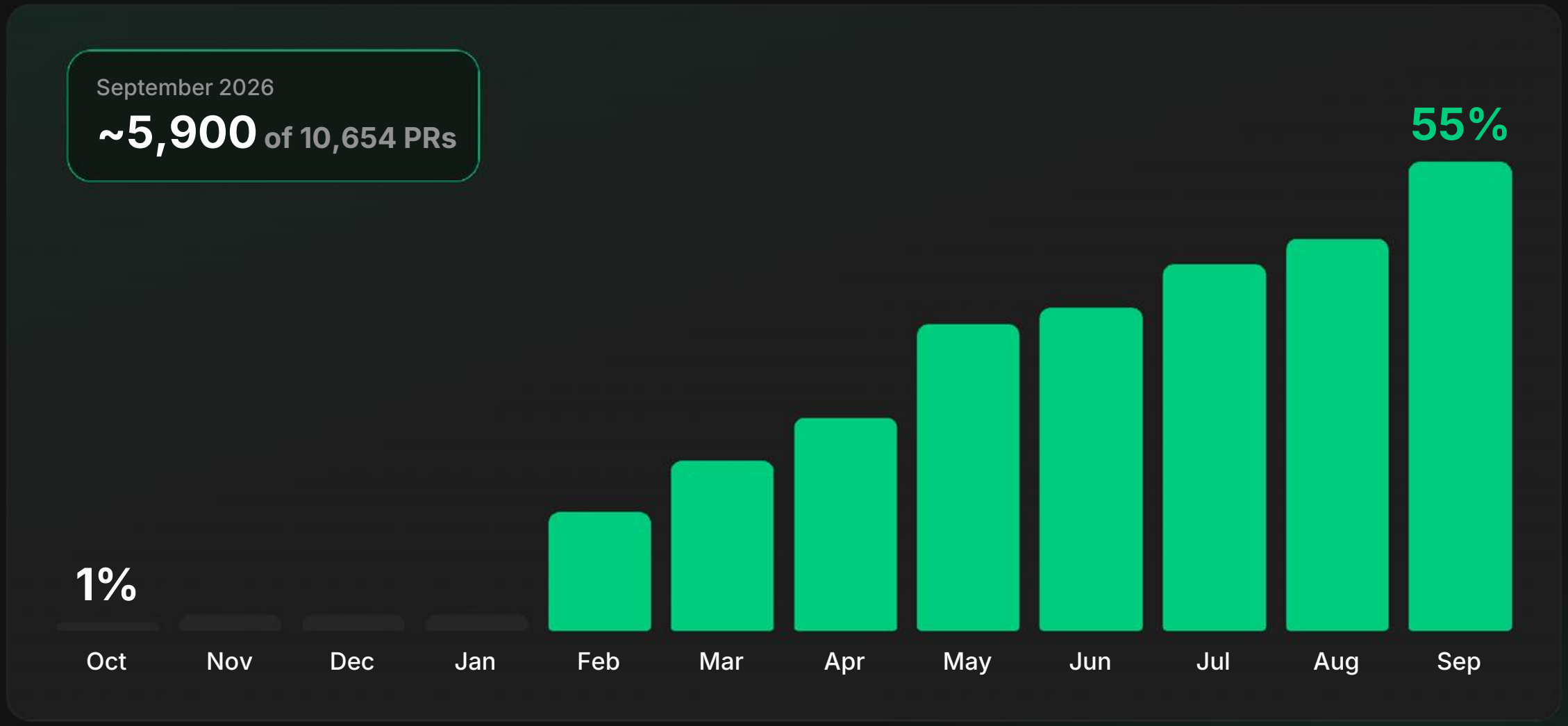
Annie Freeman

Developer Advocate, Coralogix


Pov: life after claude code.
TikTok
@ alan.samer

Over half our PRs are now agent-written.

Coralogix



“This feels faster.”
“That seems better.”

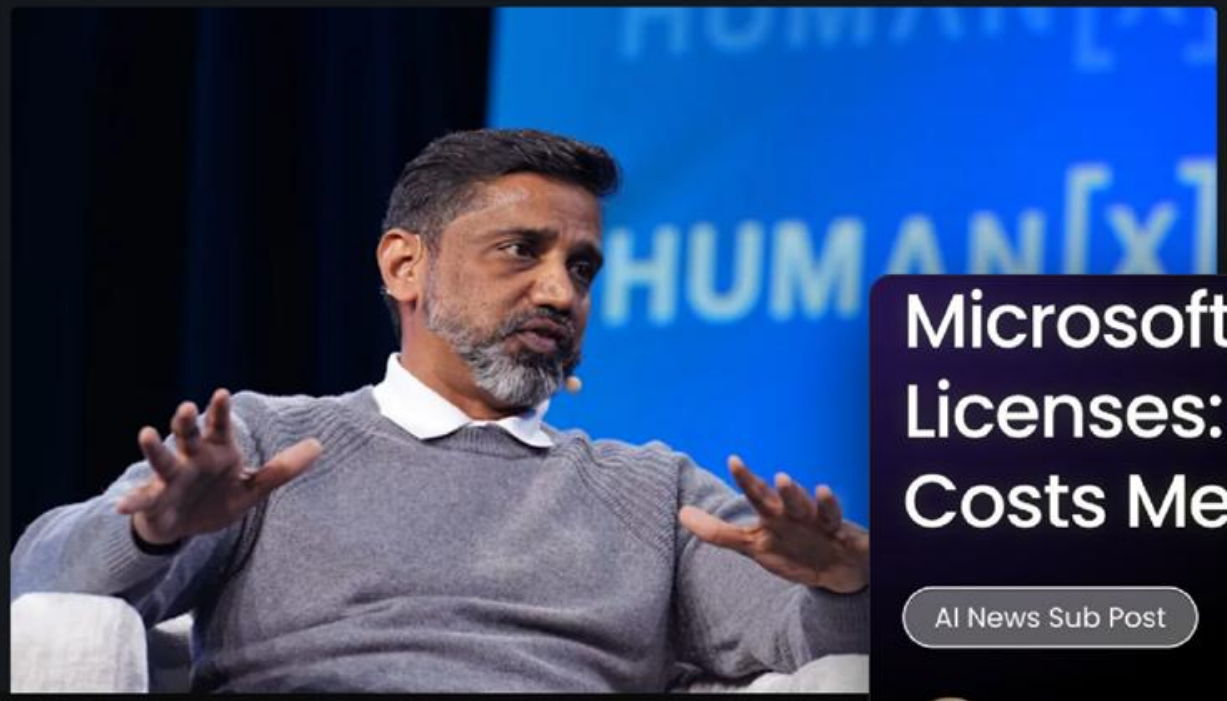
“This feels faster.”
“That seems better.”



The advantage isn't adoption, it's impact.

FORTUNE

After blowing through its entire 2026 AI budget in months, Uber CTO says, ‘We’re coming to the end of the so-called tokenmaxxing era’



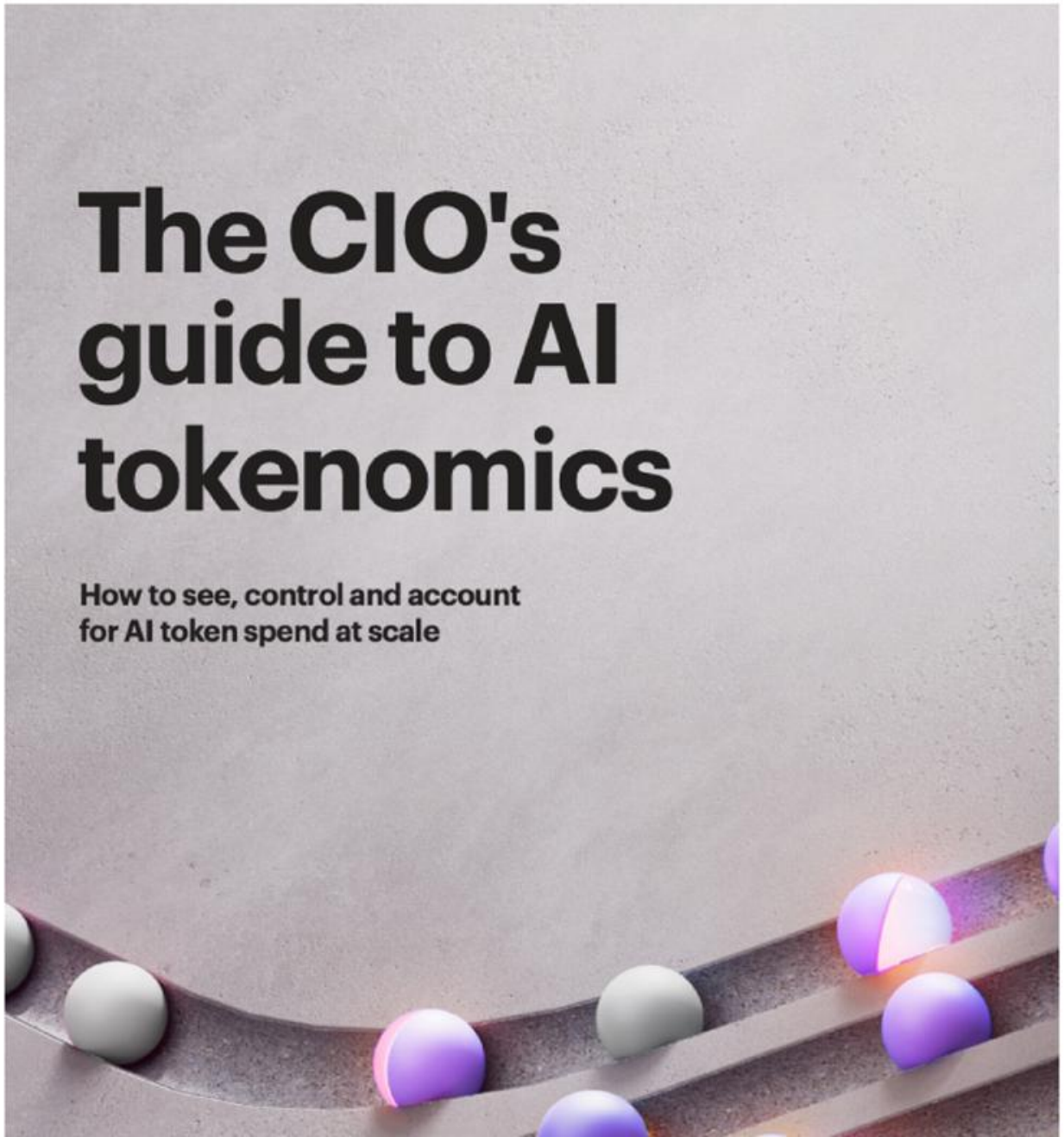
Microsoft Cancels Claude Code Licenses: What Runaway Token Costs Mean for Enterprise AI

AI News Sub Post · 2 min read · Published on: 27 May 2026



Annie Neal
Growth Marketing

Accenture Research Report



4 in 5

AI token dollars lack a quantified link to business outcomes.

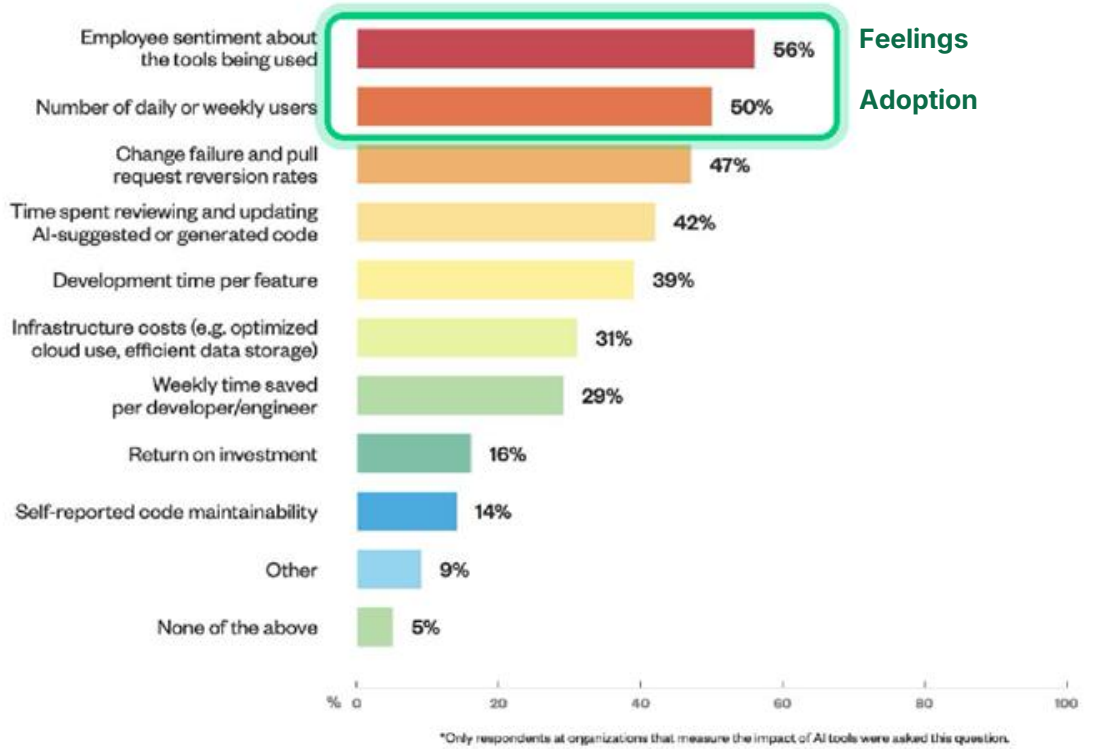
The LeadDev

State of AI-Driven Software Releases 2026

How AI-powered coding tools are transforming software releases, and what needs to happen next.



What metrics does your organization track that can gauge the impact of AI tools?* (select all that apply)



*Only respondents at organizations that measure the impact of AI tools were asked this question.

The two most common ways teams measure AI's impact: how people feel, and how many use it.

1 Where did my tokens go?

2 Can tools help me save tokens?

3 Where did everyone's tokens go?

1

Where did my tokens go?

Hypothesis: coding style influences token efficiency.

Approach 1

Small batches

Test-Driven Development

Approach 2

Medium batches

Approach 3

Large batches (‘vibe’ coding)

The data your agent is already emitting

Session

`session_count_total`
Sessions started

`cost_usage_USD_total`
Estimated cost per model

`commit_count_total`
Git commits made

`pull_request_count_total`
Pull requests created

Request

`token_usage_tokens_total`
Tokens by model & type

`active_time_total_s_total`
Seconds actively processing

`api_request`
Model, token counts, cost, latency

`api_error`
Status, error message

Tool call

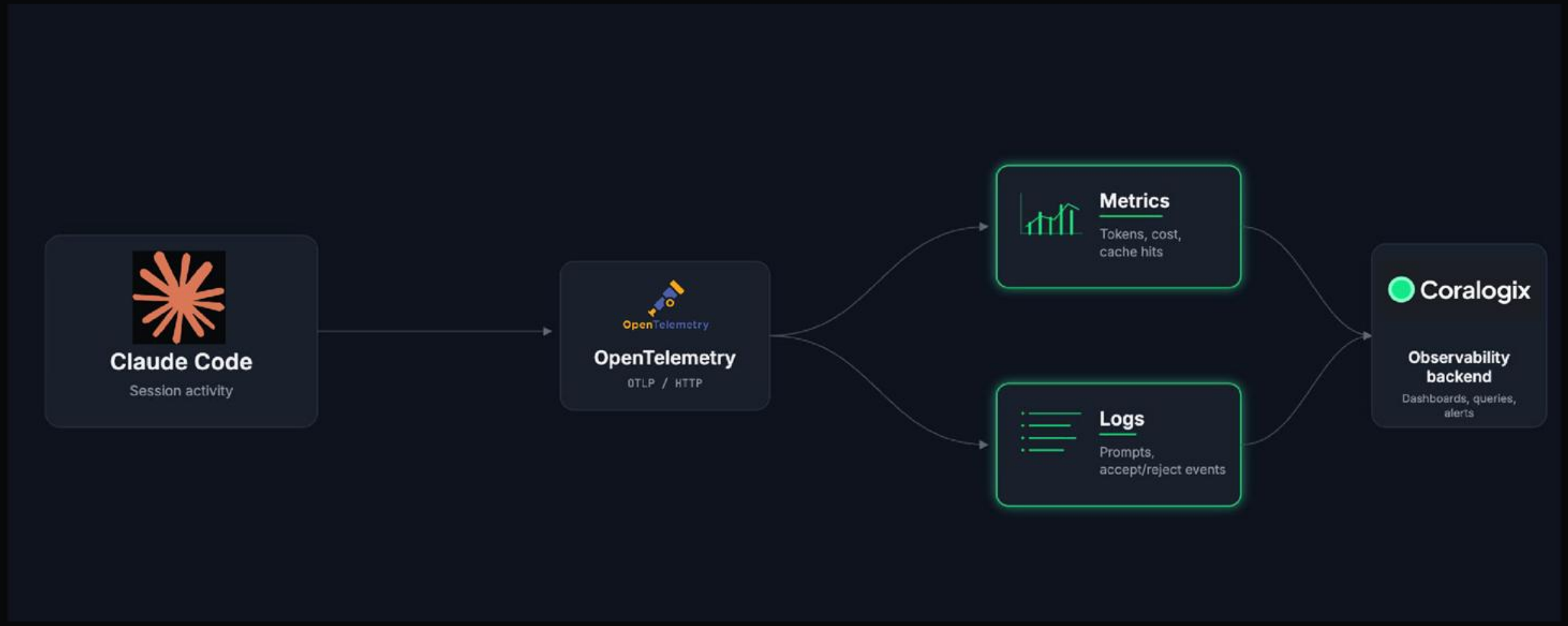
`code_edit_tool_decision_total`
Accept / reject decisions on edits

`lines_of_code_count_total`
Lines added & removed

`tool_decision`
Tool name, decision, source

`tool_result`
Tool name, duration, outcome

How I measured my Claude sessions.



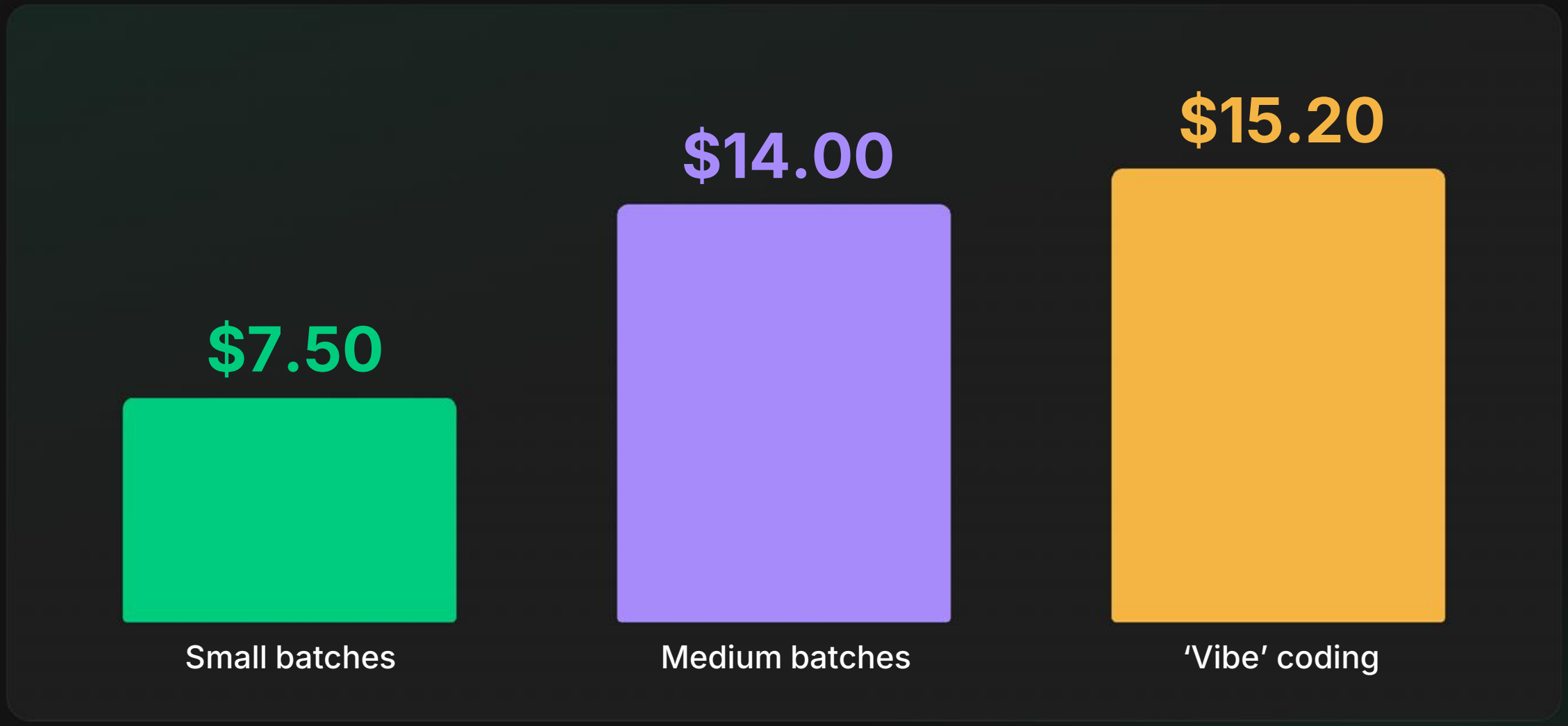
You instrument nothing.

A few lines of config.

```
// .claude/settings.json
{
  "env": {
    "CLAUDE_CODE_ENABLE_TELEMETRY": "1",
    "OTEL_METRICS_EXPORTER": "otlp",
    "OTEL_LOGS_EXPORTER": "otlp",
    "OTEL_EXPORTER_OTLP_ENDPOINT": "<YOUR_OTLP_ENDPOINT>",
    "OTEL_EXPORTER_OTLP_HEADERS": "Authorization=Bearer <KEY>",
    "OTEL_RESOURCE_ATTRIBUTES": "team=<TEAM>,repo=<REPO>",
    "OTEL_EXPORTER_OTLP_METRICS_TEMPORALITY_PREFERENCE": "delta"
  }
}
```

How to set up code agents observability

The token bill: ~2 hours work on each approach.

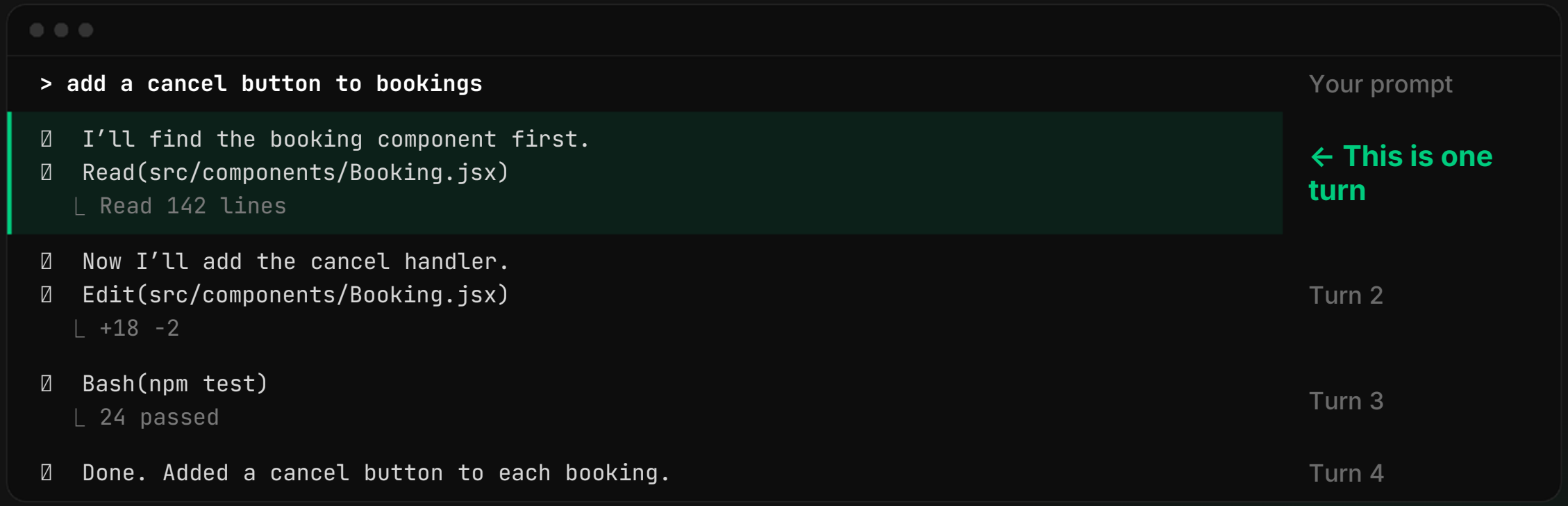


Cost for the same amount of features.

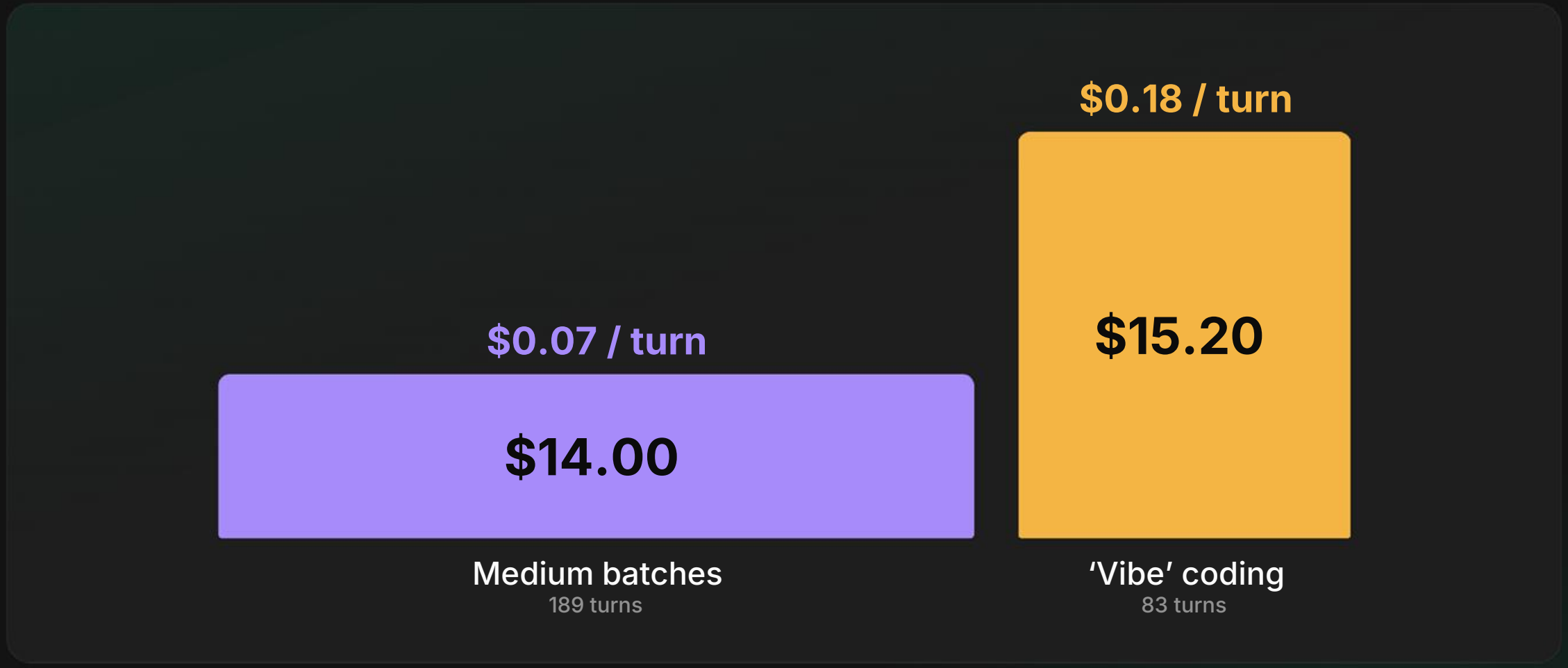
In ~2 hours, small batches got through far fewer features. Scaled up to match, it costs more



What’s a turn?

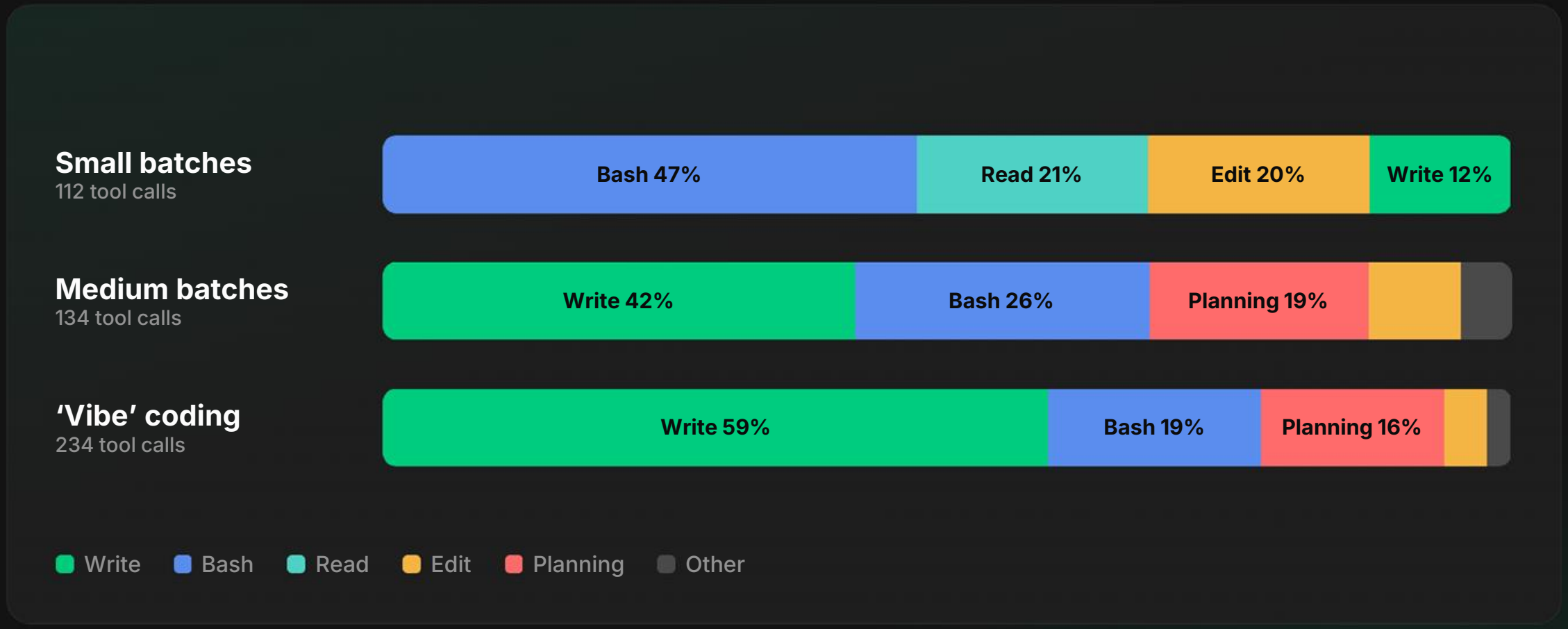

> add a cancel button to bookings
Your prompt
I’ll find the booking component first.
Read(src/components/Booking.jsx)
Read 142 lines
← This is one turn
Now I’ll add the cancel handler.
Edit(src/components/Booking.jsx)
+18 -2
Turn 2
Bash(npm test)
24 passed
Turn 3
Done. Added a cancel button to each booking.
Turn 4

Fewer turns, but each cost 2.5× as much.

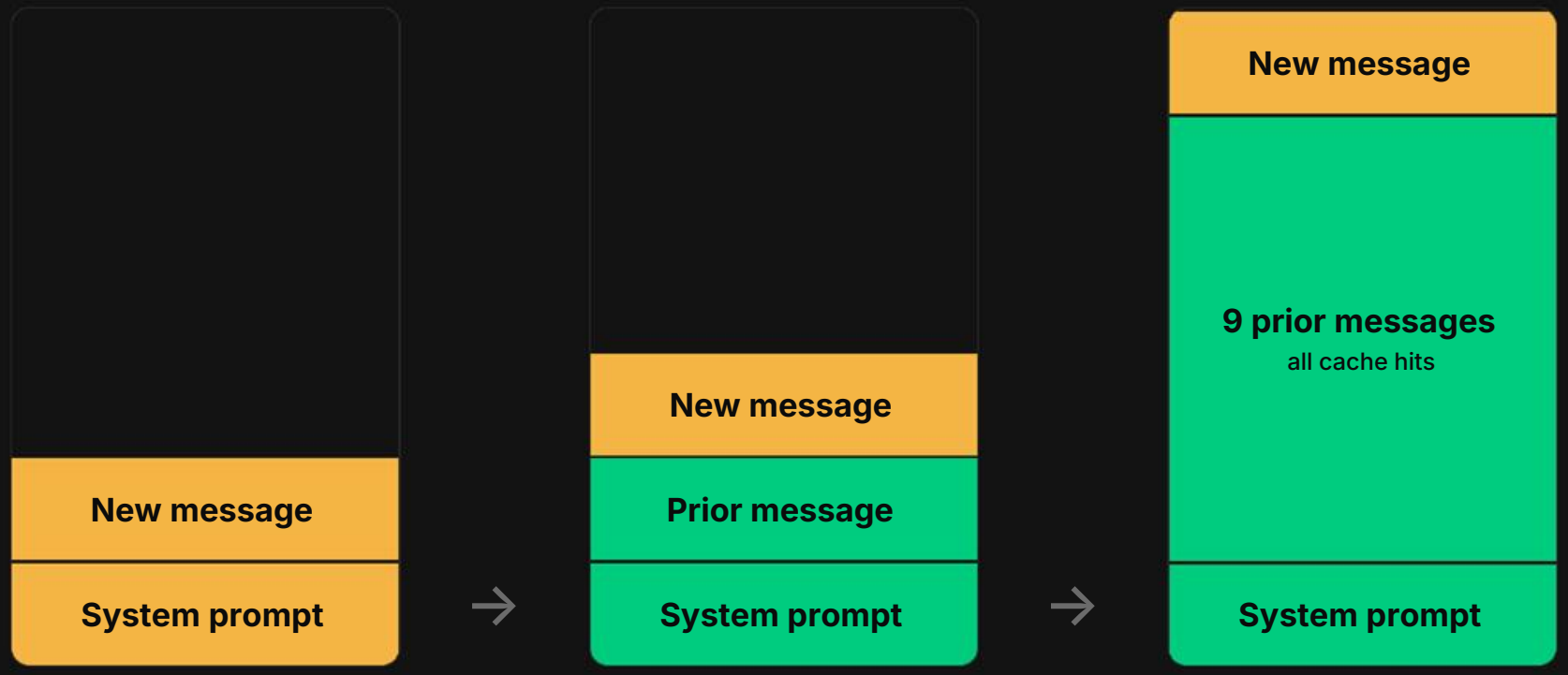


But why?

Different approaches lead to different tool calls.



Caching in code agents.



You mostly pay for re-reading, not for code.

Share of spend across 25,091 Claude Code sessions



R&D

Send the reading to a cheaper model.

"Most of what an AI coding agent does for me isn't thinking. It's I/O."

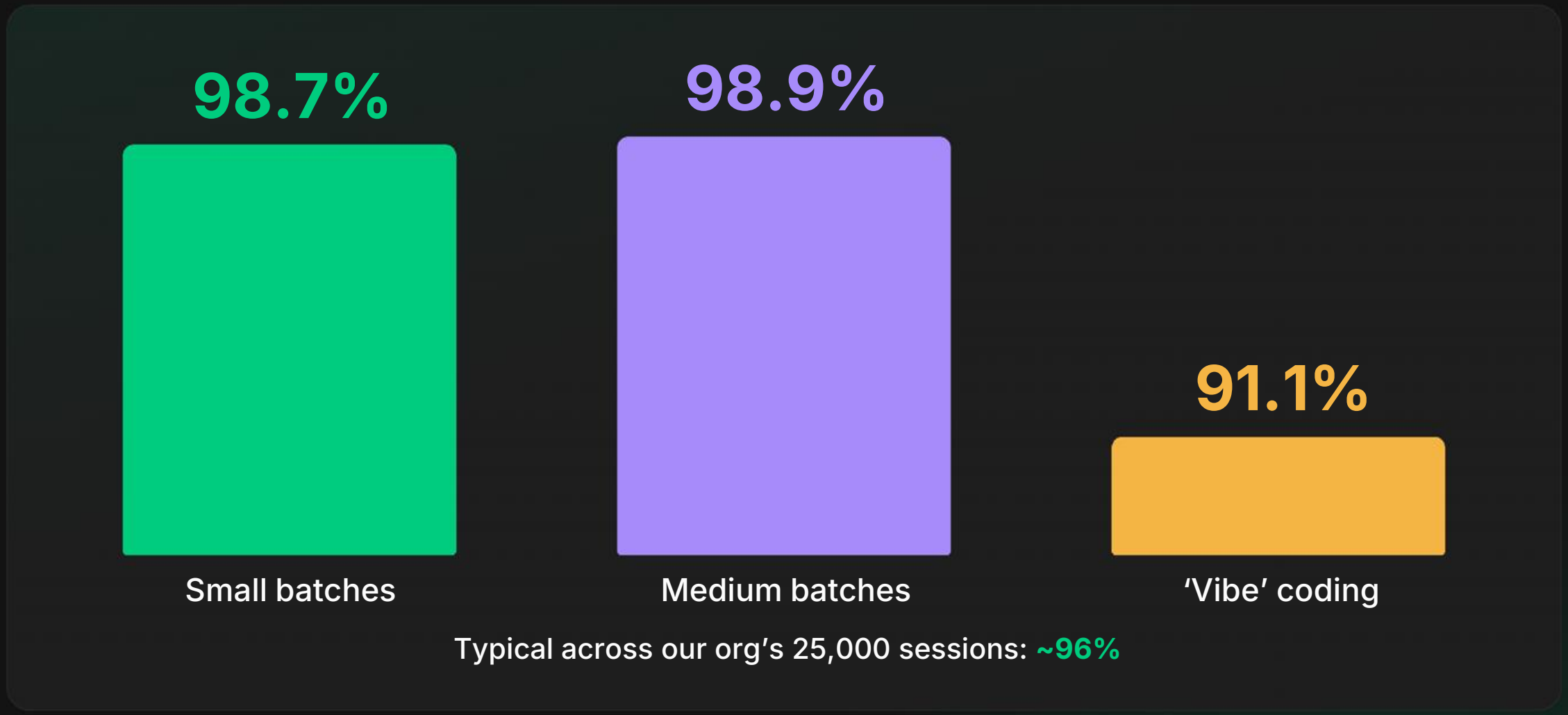
Dimitri Mazmanov, Spotify Engineering, Sep 2026

~90% fewer tokens on bulk file reads

Read and try it yourself

Install the shunt plugin from
`spotify/portal-ai-plugins`

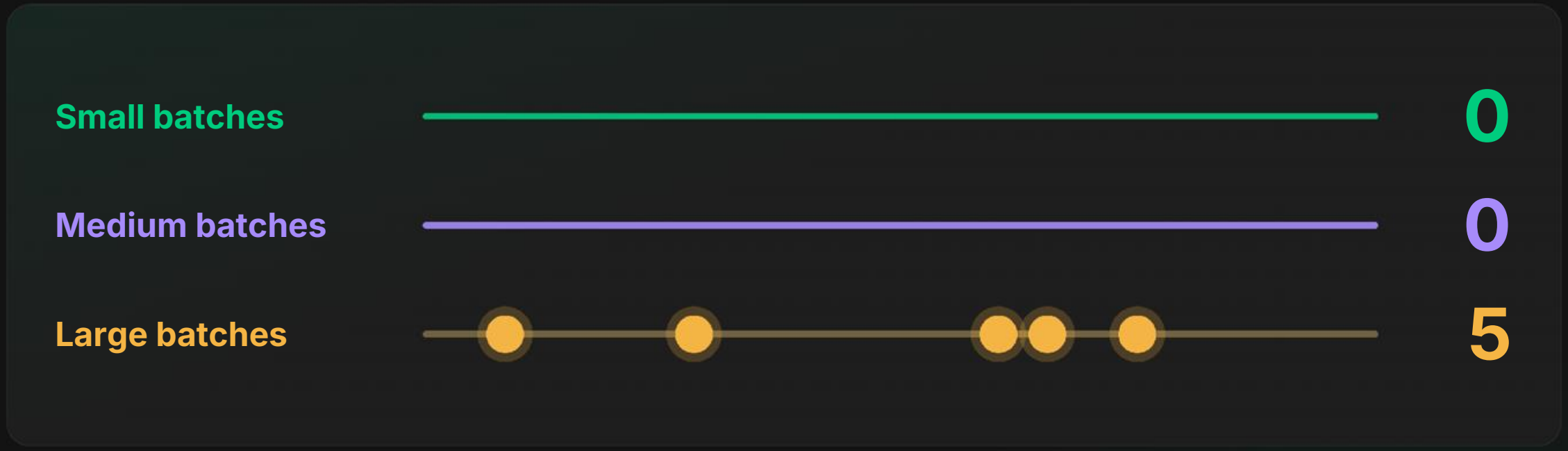
Cache hit rate.



Typical across our org's 25,000 sessions: ~96%

Large batches threw the cache away.

Each ● is a cache invalidation: large portion of context re-written



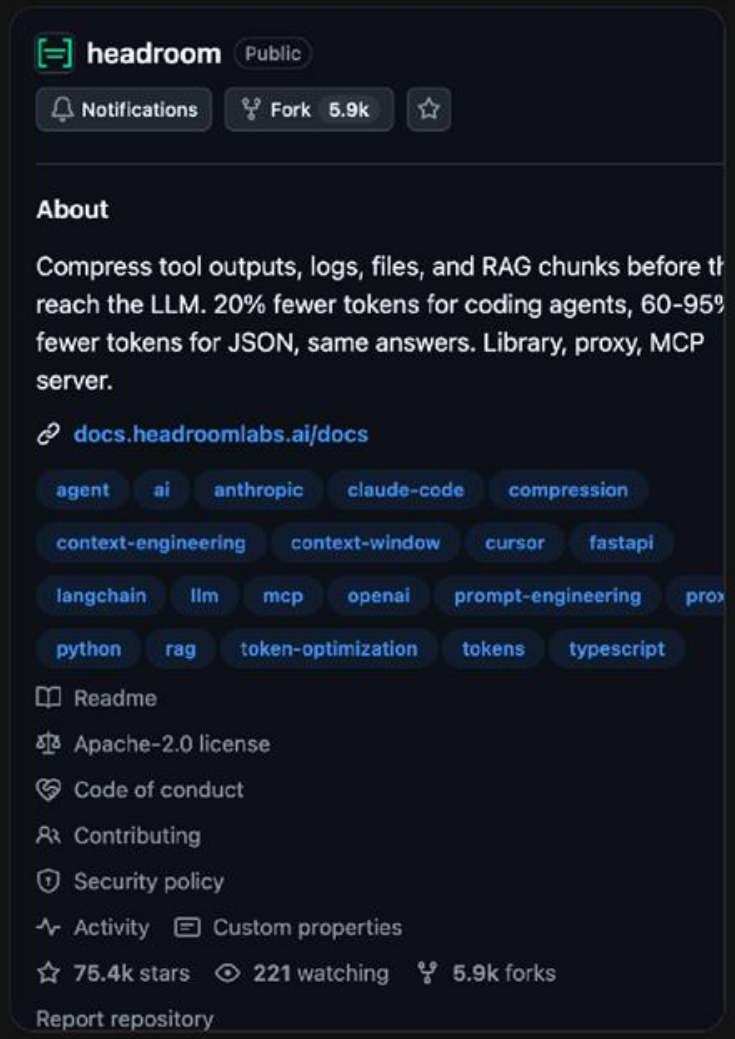
Improve token efficiency by optimising context and the cache

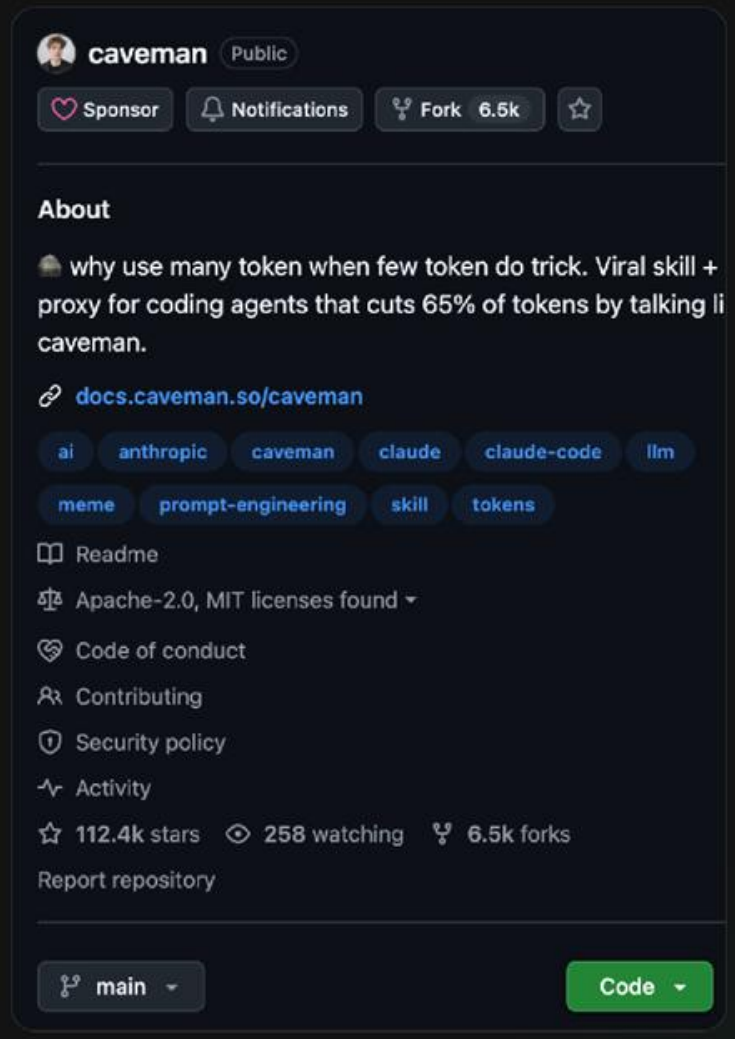
- Plan once, then build
- Manage batch size
- New task, new session
- Send big file reads to a cheaper model

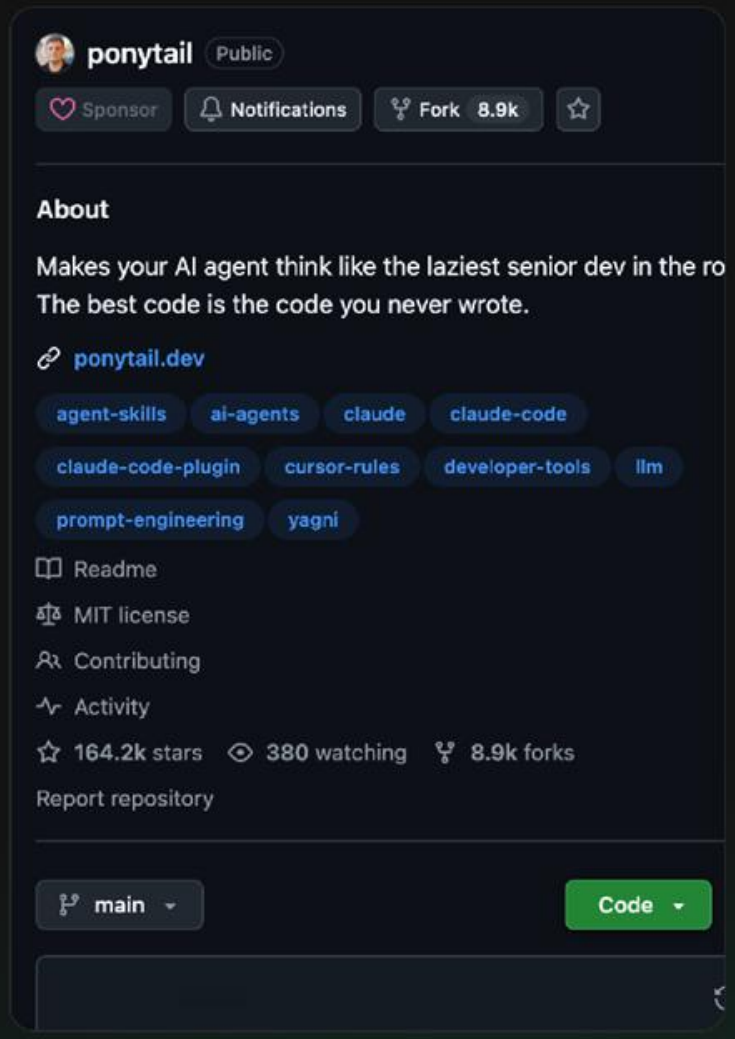
2

Can tools help me save tokens?

Three viral token savers.

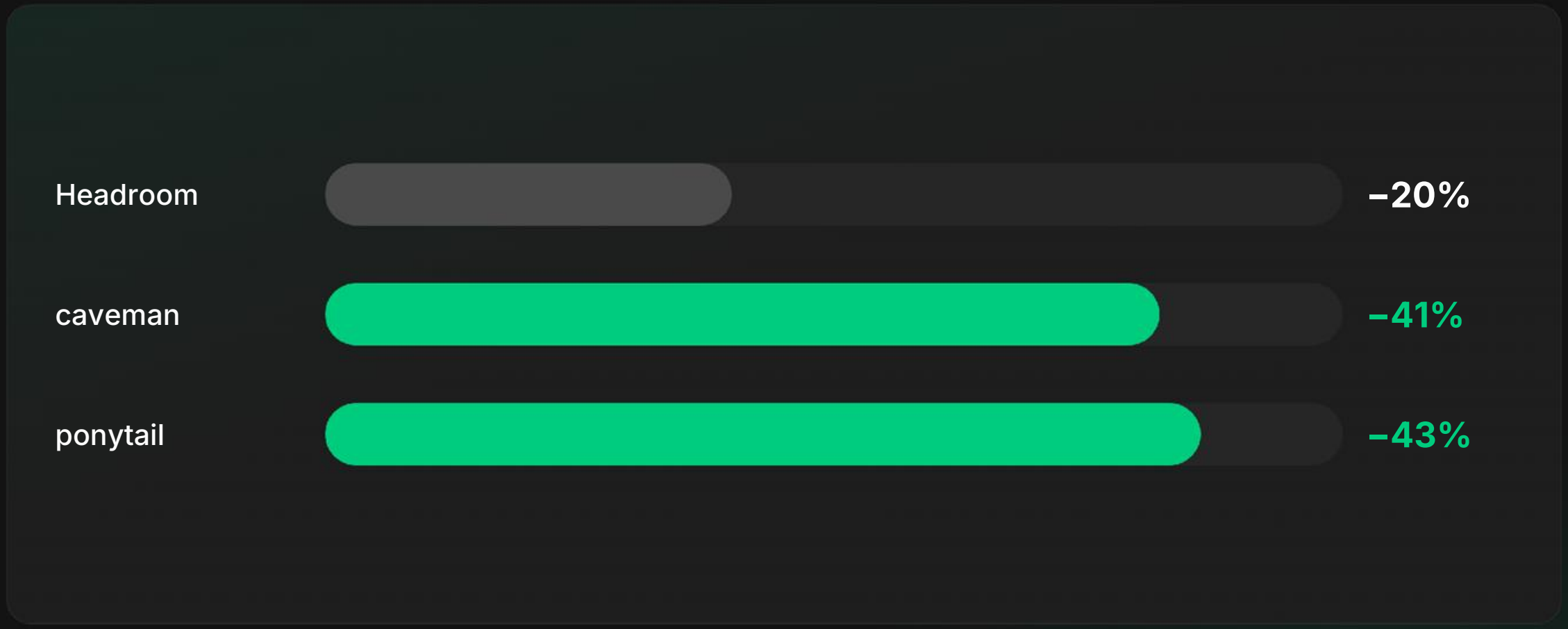




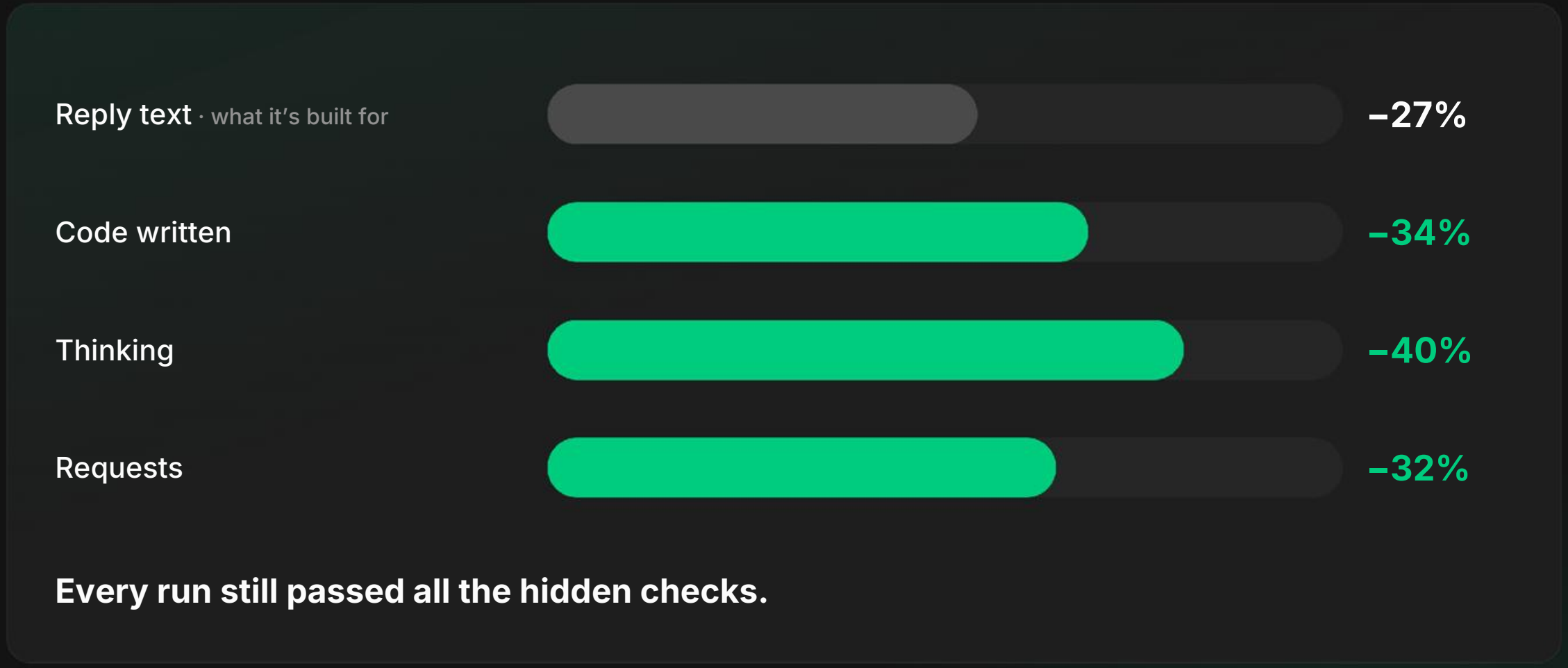


Result: savings of up to 40%.

Cost per session against plain Claude Code.



Caveman saved on code and thinking, not just talking.



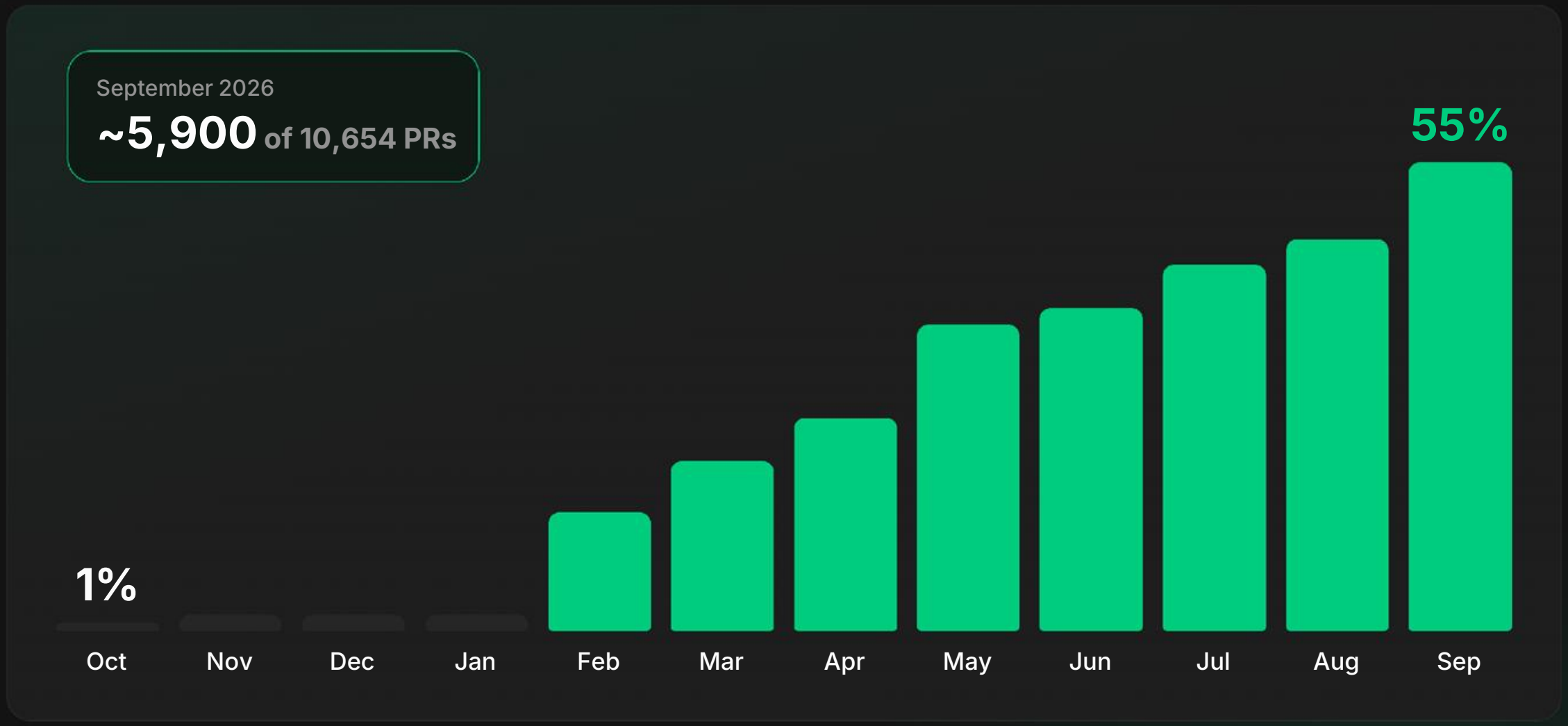
Every run still passed all the hidden checks.

3

Where did everyone's tokens go?

Over half our PRs are now agent-written.

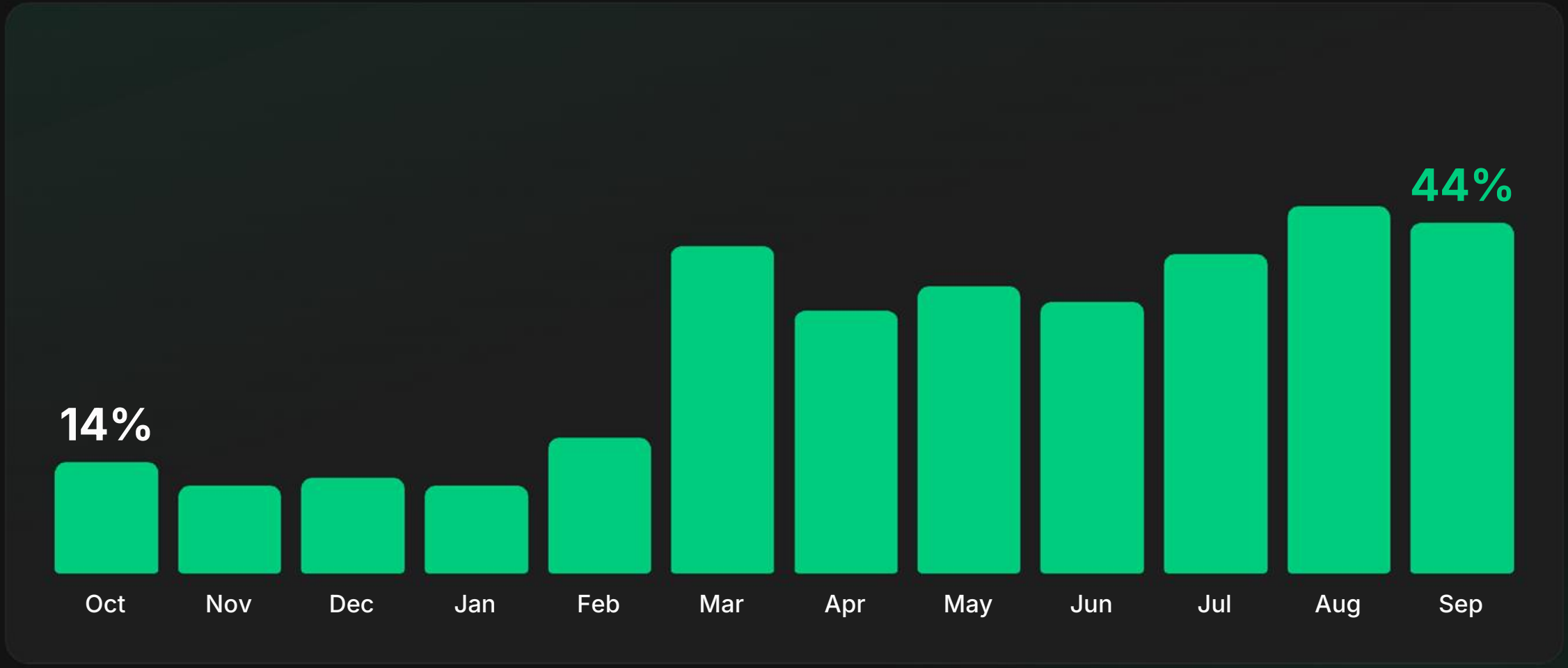
Coralogix



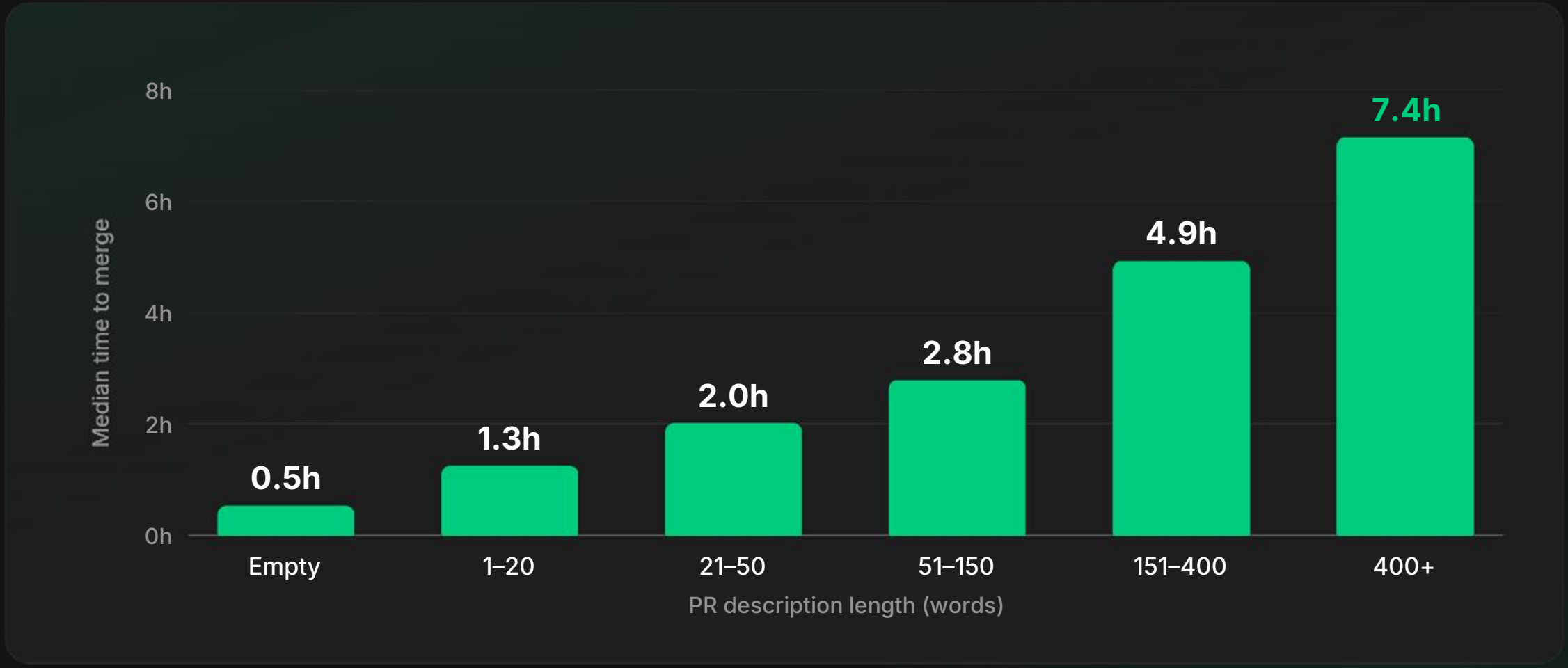
PRs are 2.8× bigger when using an agent.



As PRs grew, we added AI review.

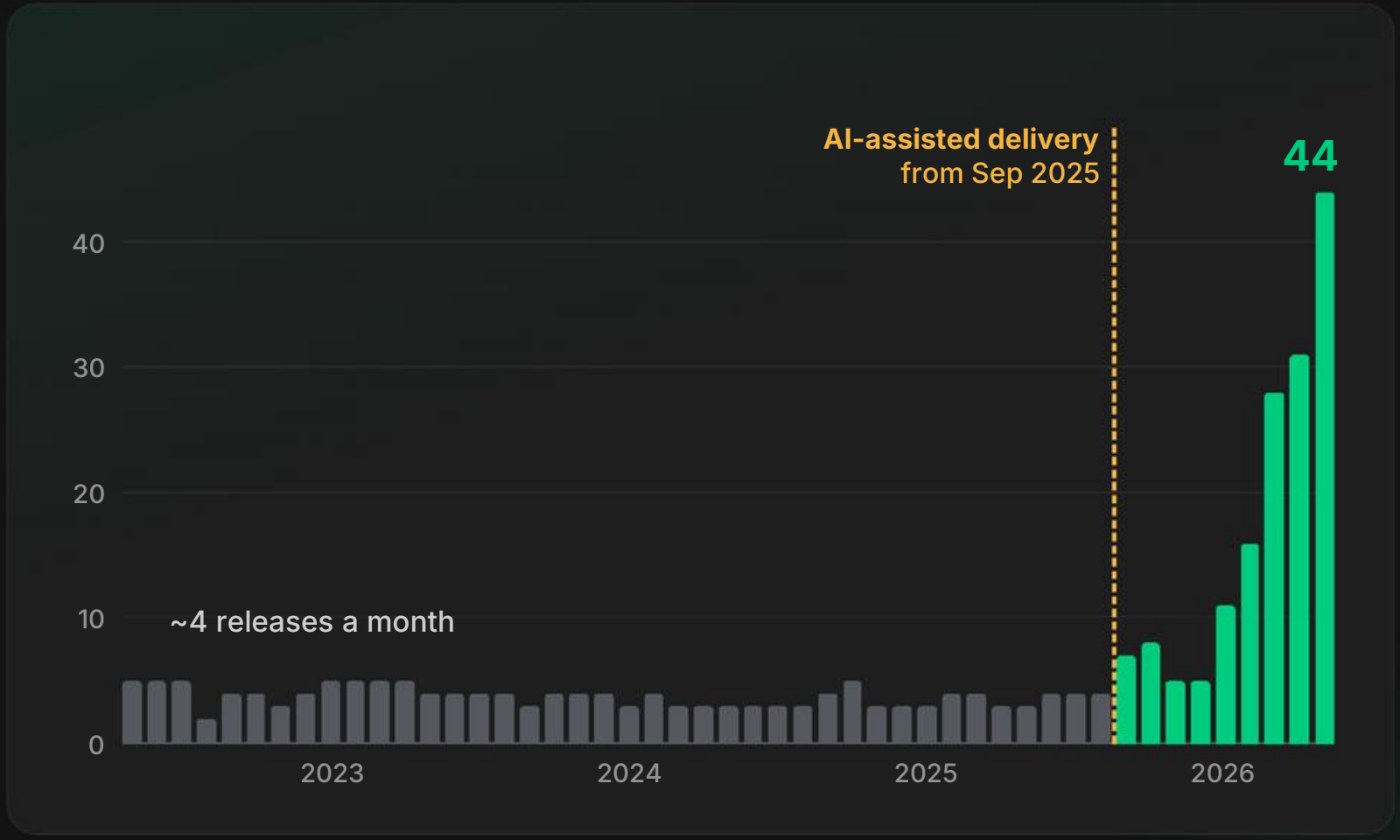


Longer descriptions, slower merges.



11× the feature releases, no reliability loss.

Coralogix public release notes, releases per month, May 2022 to May 2026



Coralogix

Revert rate

Same

Agent PRs get reverted the same as other PRs

CI success rate

Same

Agent PRs vs other PRs, same repo and size

The advantage isn’t adoption, it’s impact.

Try these this week.

1

Send agent OTel to your observability platform

A few lines of config

2

Find your top 10% of sessions

What do they have in common?

3

Put a session ID on every PR

Join cost to what shipped

4

Find the PR size that works for you

Track it against review

5

Connect agent use to your DORA metrics

Lead time, failures, recovery

Let’s connect



Annie Freeman

LinkedIn