

LORONG AI · THE ONE ABOUT AI × SWE · 14 OCTOBER 2026

Agentic engineering at AI Singapore

Learning at the edge, bringing it home

Najib Ninaba

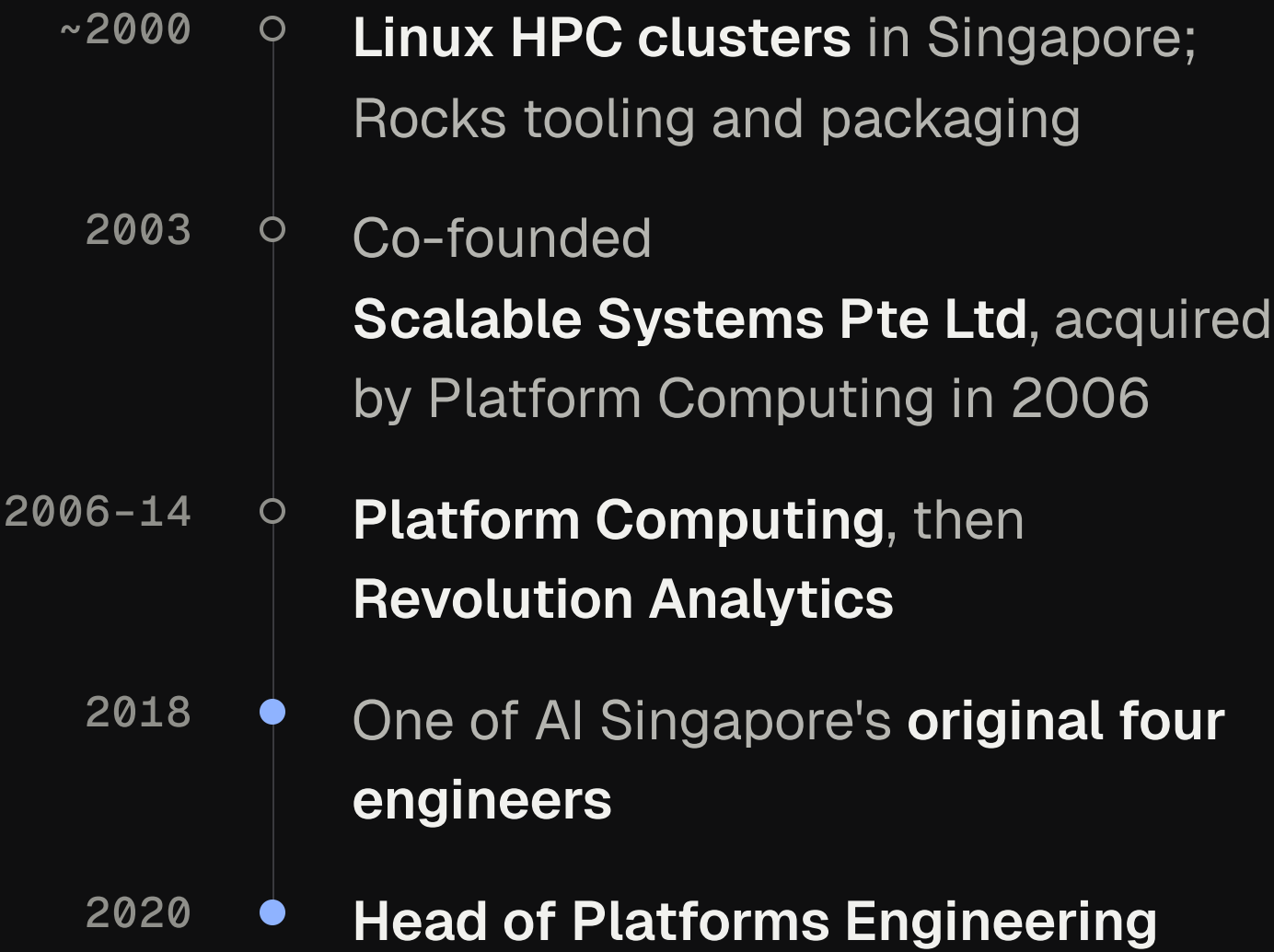
Head of Platforms Engineering, AI Singapore

najibninaba.com

github.com/najibninaba · x.com/najibninaba

WHERE I COME FROM

Old HPC habits still apply



- ⚡ **Performance and efficiency**, while keeping results accurate
- 🔄 **Scale and automate** so infrastructure and tools do the heavy lifting
- 📦 **Split work into independent pieces** that run in parallel, with **map-reduce** as a default pattern

Building AI platforms and tools with AI agents. Old HPC habits still apply.

[My X bio ↗](#)

A small team supporting many builders

9

people in Platforms Engineering

7+

teams we support

Platforms · Infrastructure · Operations

with mentoring and education running across all of it



AIAP apprentices

AI Apprenticeship Programme.
Engineers trained through real
project work.



AIEH engineers

Delivering 100E and SIP projects
with industry sponsors.



And more

LADP batches, AIAP Core, AlxTech,
and our PM and Admin teams.

TODAY

Four parts

01

How my practice changed

From watching, to evaluating, to running several agents in parallel

02

Patterns I learned

Building Orchard and contributing to RepoPrompt CE

03

Bringing it back to AISG

Access, environments, plugins, mentoring and assessment

04

Where AISG is now

Uneven adoption, and what we're still working out

01

How my practice changed

2023 to today: from watching, to learning to evaluate, to running agents in parallel.

Before coding agents

“ If Prompt Engineering is the new hype, how long before PromptOps becomes a thing...

X · 7 January 2023 ↗

- Mar 2023 ○ Azure OpenAI experiments
- Sep 2023 ○ Semantic Kernel and Prompt Flow sandboxes
- 2024 ● My outlook: small fine-tuned models, open-source LLMs, **OpenAI-compatible adapters**, LLM infrastructure platforms

EARLY TO MID 2025 · FIRST REAL USE

Learning to evaluate, not vibe check



Before, I was doing ‘vibe checks’: code looked right, tests passed, ship it. Now I understand why that fails.

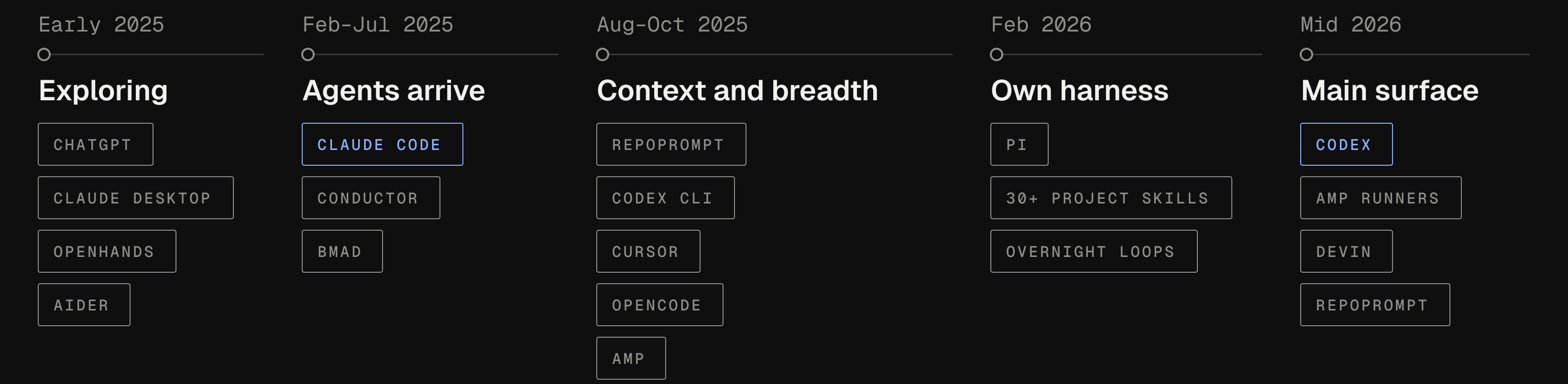
LinkedIn · 21 August 2025 ↗

Stop obsessing over perfect prompts. Start by systematically analyzing WHERE and WHY they fail.

X thread · 21 August 2025 ↗

2025 TO 2026 · TRYING EACH WAVE

Trying the field to see what holds up



I put my own time and money into each wave, so I could see what holds up before bringing it to AISG.

NOVEMBER 2024 TO OCTOBER 2026 · FROM MY RECEIPTS

Paying to learn early

My own money on AI tools, compute and courses.

US\$11.9k

≈ S\$16,000



Models and agents

~\$7.9k

ChatGPT Pro since Aug 2025 · Claude Max, then API credits · Devin, xAI, Warp

Infra, dev tools

~\$1.9k

RepoPrompt, Exa, Greptile, others

Learning

~\$2.1k

AI Evals course, a Maven AI course

Claude, before and after AISG's plan:

Max \$200

→

Max \$100

→

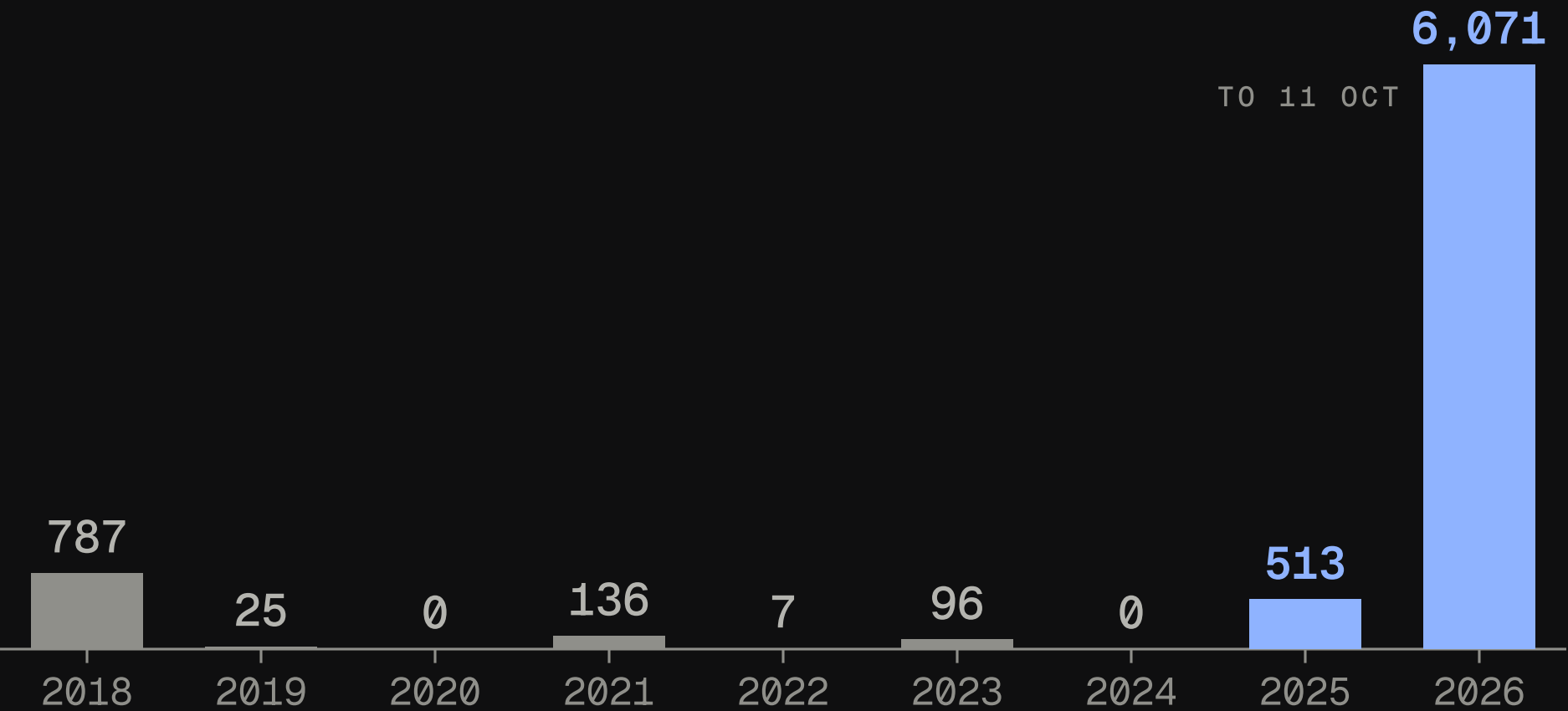
Pro \$20

once AISG's plan covered my work

DECEMBER 2025 TO MARCH 2026 · ORCHESTRATION

Scale, once the workflow held

GITHUB CONTRIBUTIONS BY YEAR



2.6B

tokens in 30 days, across 574 sessions

I plan with one model, implement with another, review with a third.

X · 10 January 2026 ↗

Model quality is improving quickly.
Results still depend on workflow quality.

LinkedIn · 22 February 2026 ↗

MARCH TO JUNE 2026 · STRUCTURE

Adding structure around the agents



SPEC.md first

Written before any code



AGENTS.md by default

In every new repo



Workbench repos

Plans and saved exports
beside each project



OpenSpec

Change packages with spec
deltas

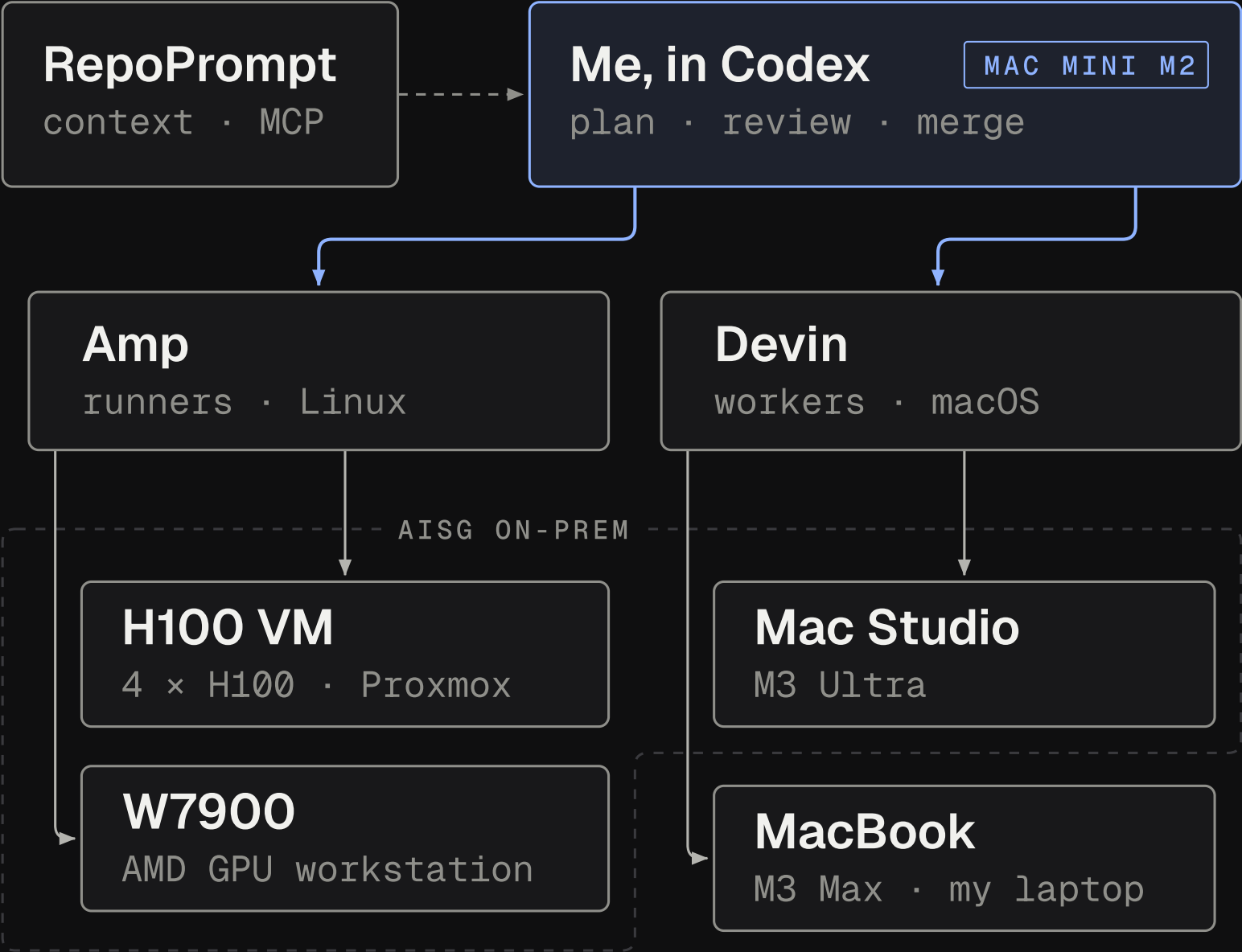


Validation gates

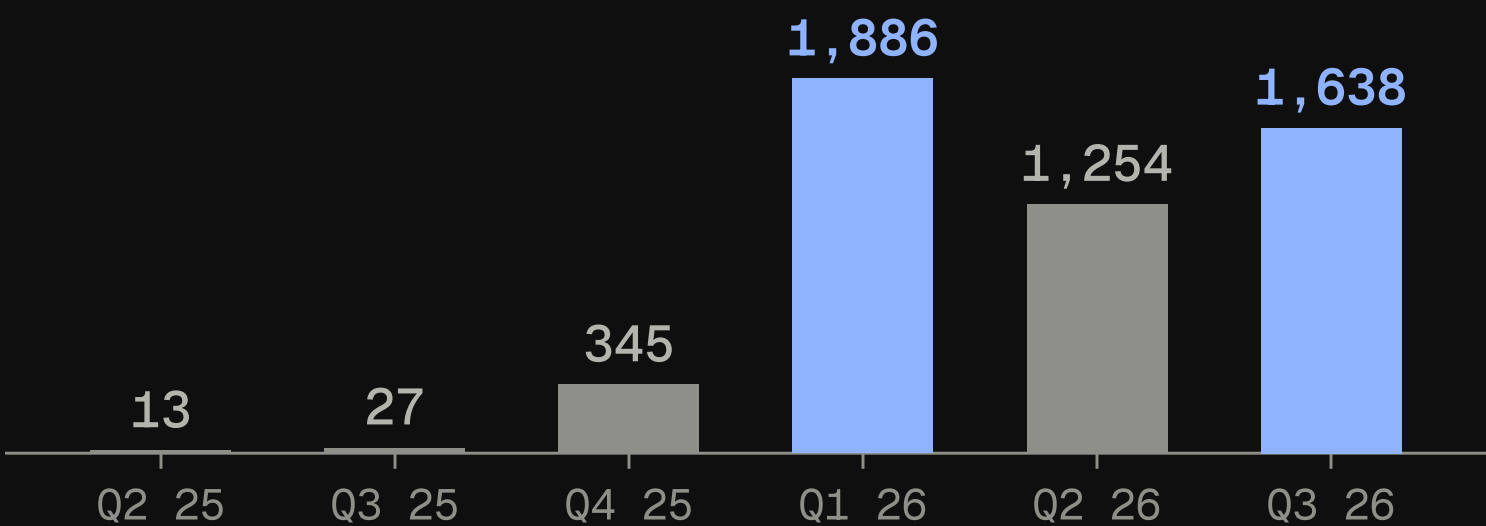
Review and checks before
pushing, added in June
2026

AUGUST 2026 TO NOW · FAN-OUT

Fan-out across runners



COMMITTS IN MY REPOS, BY QUARTER



Commits with agent attribution, 2026

Q2 1%

Q3 13%

Q4 28%

AI agents have helped me spend less time writing code directly and more time guiding the work.

X · 4 October 2026 ↗

02

Patterns I learned building with agents

Orchard is where I build. RepoPrompt CE is the tool I use and help build.

Local AI on our own hardware

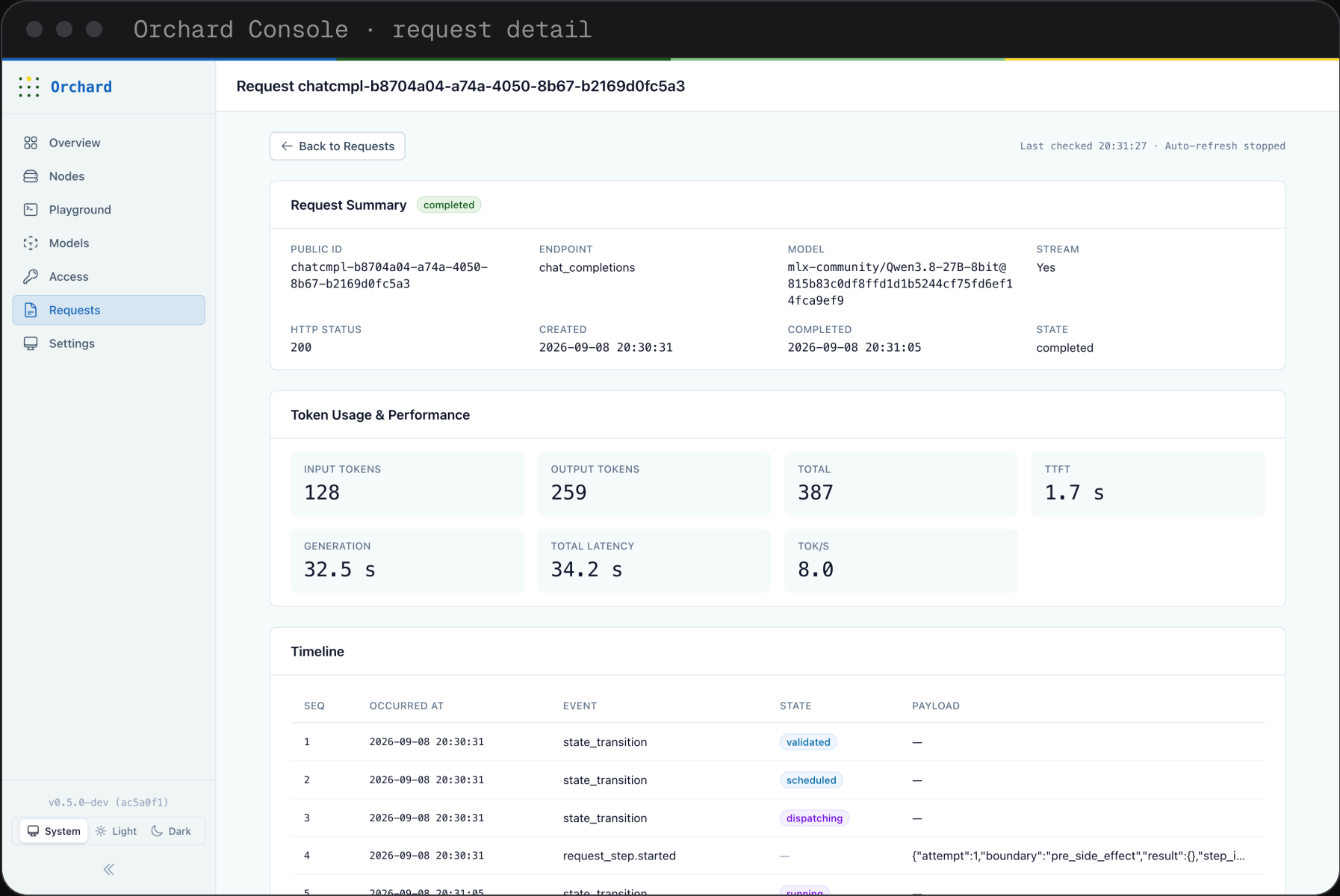
It started as a request to run local AI on-prem, on **eight Mac Studio M3 Ultras, 512 GB each**.

- 🌐 One OpenAI-compatible endpoint on Apple Silicon
- 🔑 Workspace tokens; every model explicitly granted
- ⚙️ Elixir/OTP control plane, MLX workers
- </> Source-only pilot, Apache-2.0

~1,880
commits

7_{mo}
since the first
commit

93%
of commits mine,
written with agents



PATTERN 1 · REPOPROMPT · REPOPROMPT.COM ↗ · GITHUB.COM/REPOPROMPT/REPOPROMPT-CE ↗

Context before action

With **RepoPrompt**, an open-source macOS app that builds context for coding agents.



I ask the coding agent to explain its assumptions before it starts writing code.

X · 1 October 2026 ↗

[REPOPROMPT CE](#) ↗ · ~950 ★

I was the **first contributor besides the author**: [PR #2](#), on 31 May 2026, the day it went open source.

Since then: Cursor agent support ([#902](#), [#960](#)) and the [Homebrew cask](#).

PATTERN 2

Spec and plan first

```
$ git show --stat bb432055
bb432055 2026-03-08
init: orchard v2 – spec + agents guide

SPEC.md      | 2855 ++++++
AGENTS.md    |   ++
.gitignore   |   +

# no code yet
```

- 📄 Orchard's **first commit** held only the spec and the agent guide
- ☰ Since June, changes trace to the spec through **OpenSpec** change packages
- ⚖️ **36 ADRs** record the decisions and their reasons
- 📖 The spec is now **6,000+ lines**, revised in 180+ commits

PATTERN 3

Every codebase has collaborators, and agents are among them

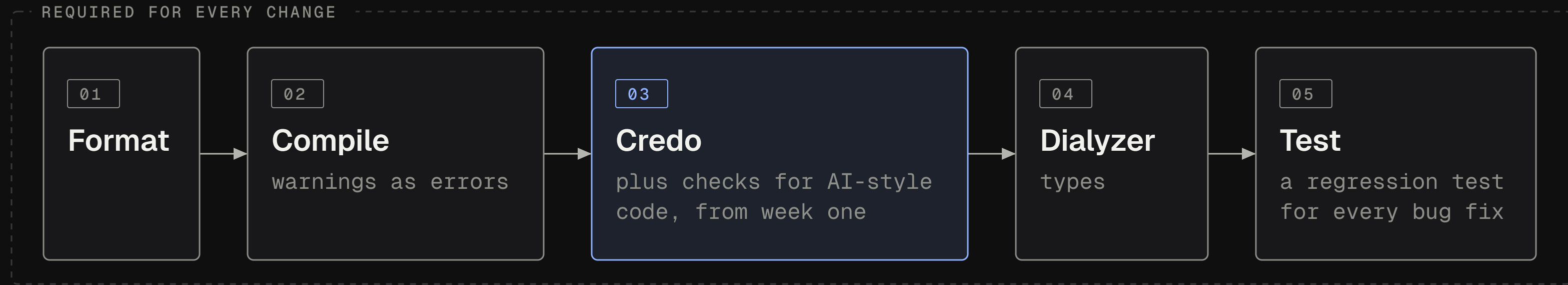


Orchard started as a solo project, built with my agents. Since July, it's two of us, each with our own agents.

- 📖 **One onboarding guide.** Orchard's CLAUDE.md just points to AGENTS.md, so every agent reads the same thing.
- 🧩 **Conventions agents can follow.** RepoPrompt CE ships repo skills and a preflight every agent runs.
- # **A shared vocabulary.** Glossary, ADRs and a spec both people and agents read.
- ✅ **Tests state the contract,** so an agent's change is checked like anyone's.

PATTERN 4

Gates and evidence



181k

lines of Elixir test code

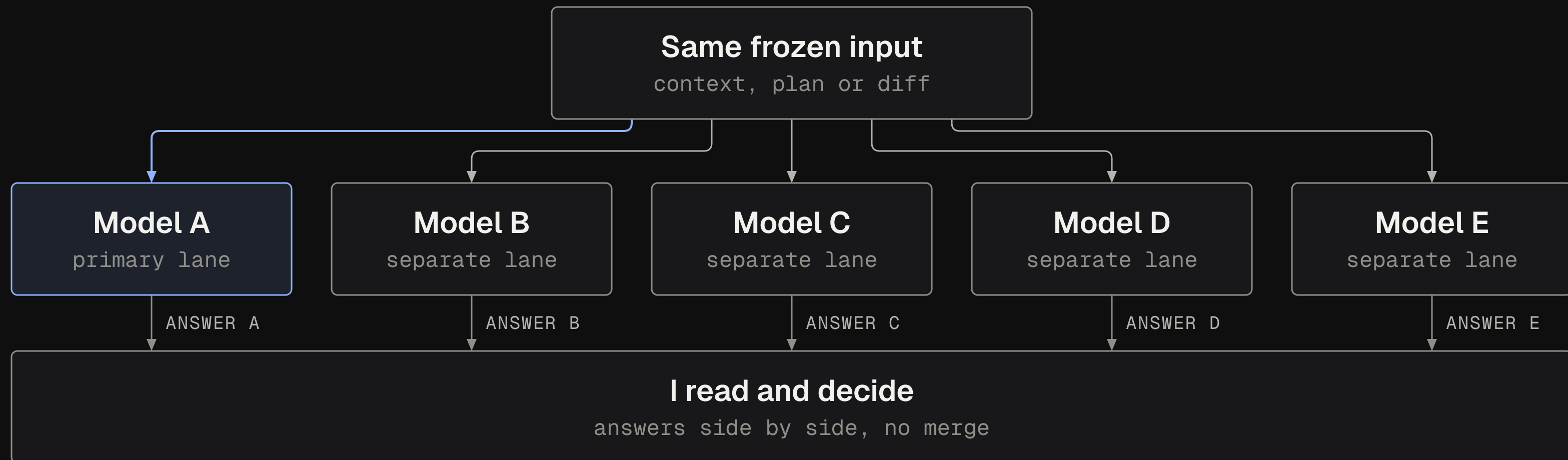
127k

lines of Elixir application code (lib/)

The 2025 evals lesson,
applied to code.

PATTERN 5

Second opinions from other models



December 2025: a review on a second model found **critical issues** in code the first model wrote. Answers stay side by side and aren't merged. [X · 4 December 2025 ↗](#)

PATTERN 6

Autonomy is earned, and a human approves the merge

“ Human in the loop first, then human on the loop. The autonomy is earned.

X · 19 February 2026 ↗

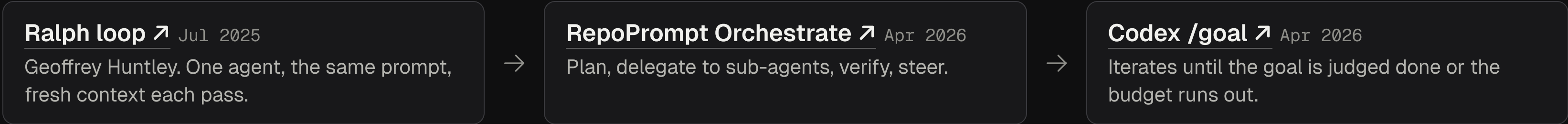
“ The agent can draft anything. I approve merges.

X Article · 21 February 2026 ↗

ORCHARD PR READINESS PACKET

- ✓ Exact head and base SHAs
- ✓ Test results and CI status
- ✓ Independent review, with counterexamples
- ✓ “An accountable human contributor has reviewed and stands behind any AI-assisted work.”

UNATTENDED LOOPS



PATTERN 7

Scale out like a cluster

HPC CLUSTER Job scheduler	↔	AGENTS · REPOPROMPT CE conductor queues builds and tests, with one heavy slot per machine
HPC CLUSTER Node isolation	↔	AGENTS A cheap worktree for each agent, so parallel work stays isolated
HPC CLUSTER Accounting	↔	AGENTS A journal of protected changes agents make through MCP

The same problems we solved on HPC clusters (scheduling, isolation and accounting), now for agents.

PATTERN 8

Plan for whoever runs it

“AI made software cheap to build, not cheap to own.

LinkedIn · 24 February 2026 ↗



Off by default

A new experimental backend shipped behind a switch, off by default.



Fail closed

Worker errors cross the boundary only as stable codes, never raw text.

PATTERN 9

Stay model- and harness-agnostic

“Bet on models and methodology, not wrappers.”

X Article · 21 February 2026 ↗

MODELS

GPT

Claude Opus / Sonnet

GLM

DeepSeek

Qwen (local)

APIS

OpenAI-compatible (Orchard)

HARNESSES

Codex

Claude Code

Amp

Devin

Pi

RepoPrompt (context)

My main surface so far: Claude Code, then Pi, now Codex.

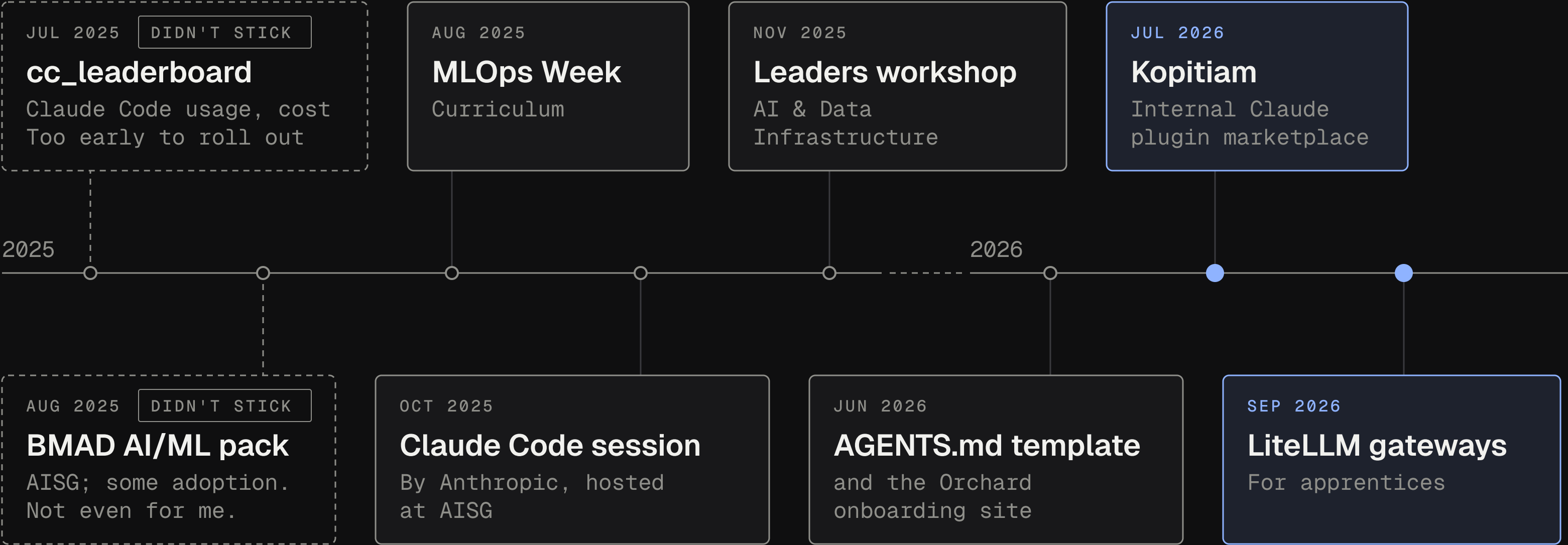
03

Bringing it back to AISG

Access, environments, plugins, mentoring and assessment. Including what didn't stick.

HOW IT FLOWED BACK

Not everything was ready for AISG



ACCESS TO CODING AGENTS AND MODELS

Each cohort, engineer or project gets its own model access

1

Own

Foundry account per cohort or team, programme batch, engineer sandbox or engagement

keys, capacity and cost tracking for each account

Running on AISG's Azure since June 2026.

- GPT-5.x

GPT-6

DeepSeek

Claude Haiku

GLM
- embeddings

 Claude Code through AISG's organisational plan

KSS (KAPITAN STEADY SELFSERVE) MAPS PROJECT ROLES TO AZURE ROLES

PROJECT ROLE	AZURE ROLE
Mentor	Owner
Apprentice	Contributor
Project manager	Reader

PROJECT ENVIRONMENTS

How a project gets started



-----▶ Operator handoff, done by hand. The tools don't call each other.

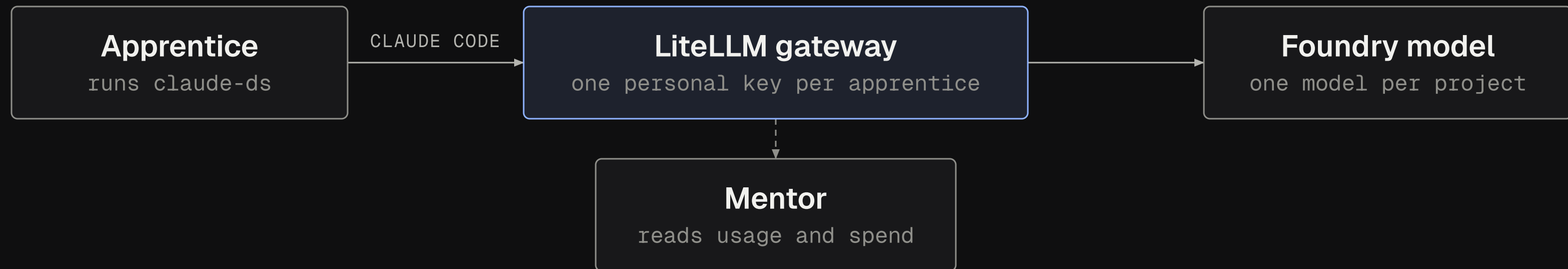


Sponsor's infrastructure first; **our GCP** if they can't provide it. Otherside provides the tooling either way.

LITELLM GATEWAYS FOR APPRENTICES

Models for apprentices, with a view for mentors

Otherside gateway design



STATUS

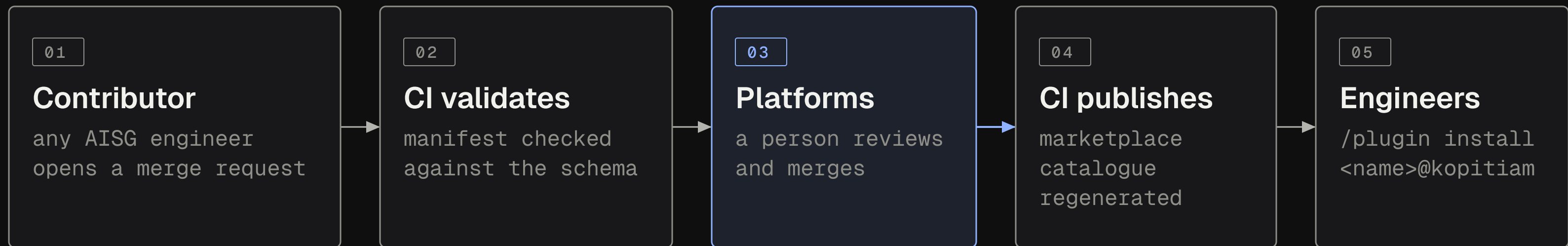
The Otherside gateway is in a **small rollout**. Separately, a LiteLLM gateway for AIAP23 has run on Azure since September.

BEYOND TRAINING

We also use LiteLLM in our sponsor project work.

KOPITIAM

Our internal Claude plugin marketplace



```
# in Claude Code
/plugin marketplace add git@.../aisg/kopitiam/kopitiam.git
/plugin install howmuchisit@kopitiam
```

WHAT ENGINEERS HAVE PACKAGED

Organisational know-how, as plugins



event-speaker-panel

Triage speaking invitations



howmuchisit

Infrastructure cost estimates



pay-me-now

Expense claims from receipts



sponsor-dd

Sponsor due diligence and meeting prep



scopeskills

Scoping AI and ML projects

HOW HOWMUCHISIT IS BUILT

 Python does the arithmetic. A script checks the output document.

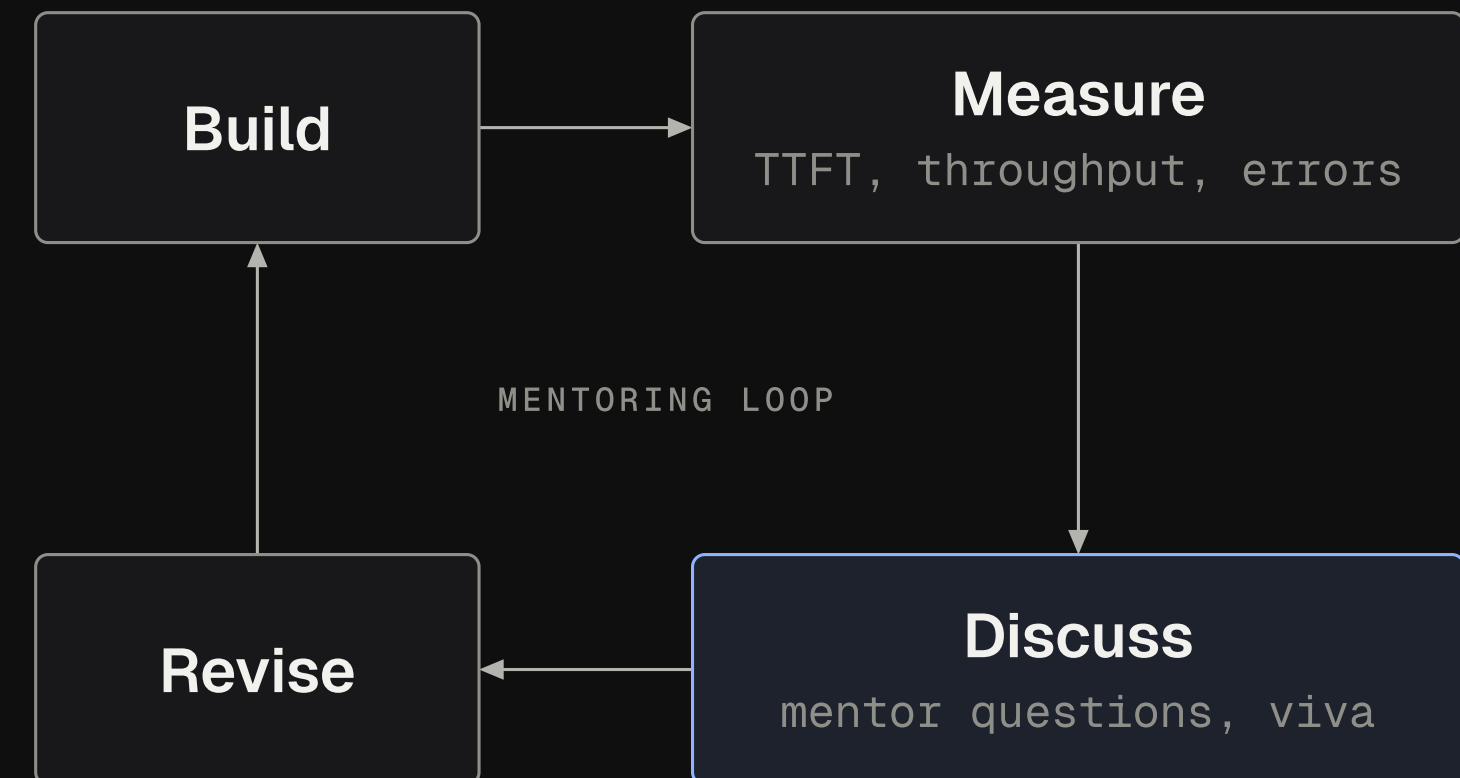
 A reviewer sub-agent sees only the report and the form; at most two fix cycles.

 The author checked it against three human estimates. A fix for the miss landed the same day.

MENTORING WITH MEASUREMENT

Gauntlet: an LLM serving exercise

- Each apprentice serves a model on **one L4 GPU** with vLLM
- Locust measures **time to first token**, throughput, latency and errors
- Results go to a **leaderboard**
- Checks against gaming: optional prompt fuzzing to defeat response caches, reported model IDs checked, and entries at 25% errors or more unranked



The numbers give the mentor and the apprentice something concrete to talk about.

AGENTS HELPING US ASSESS MLOPS WORK

Map-reduce, applied to assessment



Agents gather evidence and do the first reading. Mentors bring the results together and make the call.

GETTING OUR PLATFORMS READY FOR AGENTS

Agent-first where we can, with guardrails

FOUNDATION	WHAT WE HAVE SO FAR	STILL IN PROGRESS
MCP endpoints and agent skills	Kopitiam plugins and skills; Orchard's OpenAI-compatible API	MCP endpoints on KSS, Otherside and Concierge
Machine-readable docs	AGENTS.md template; Orchard onboarding site; AGENTS.md in our repos	Platform docs agents can use directly
Guardrails for people and agents	KSS role mapping; per-apprentice gateway keys; Orchard's deny-by-default model grants	Gateway budgets and rate limits
Explicit ownership	Platforms reviews every Kopitiam plugin	Ownership after publish; keeping plugins in sync

If an agent has to fight its way through your wiki, the platform isn't ready.

LinkedIn · 4 March 2026 ↗

LLM RED-TEAMING IN 100E PROJECTS

Red-teamed before handover

Led by **SAGE**, our AI Safety & Governance Engineering team



What it tests

AI safety: harmful or off-policy answers, jailbreaks. **Security:** prompt injection, data leaks, tool misuse. **Governance:** mapped to the EU AI Act, NIST AI RMF, ISO 42001, OWASP LLM Top 10 and MITRE ATLAS.



How it works

An attacker model, the project's app as target, and a grader model. `/discover` extracts the prompts; `/attack` generates tests, grades them, fixes failures and re-tests the fixes.



What it changes

Each finding is documented and walked through with the project team. Fixes go at the highest layer that can control the problem: architecture, then application controls, then prompt hardening.

04

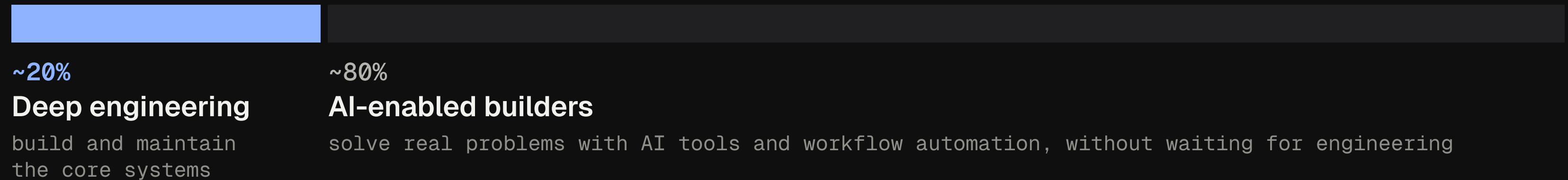
Where AISG is now

People across AISG are picking up these tools at different paces. The platform team's job is to support each of them.

ADOPTION HAPPENS AT DIFFERENT PACES

Two tracks, both doing real work

A working model of the support people need, from my 2 March post



“ Agents don't replace engineers.
They create more builders.

LinkedIn · 2 March 2026 ↗

1865

Jevons' paradox ↗. In *The Coal Question*, Jevons observed that more efficient use of coal led to **more** coal being used, not less. I expect something similar as building software gets cheaper.

WHAT WE'RE STILL WORKING OUT

Open problems we're working on

-  **Scaling past one person's dev environment.** When anyone and their agents can implement anything, coordinating a small team is the open question.
-  **Gateway budgets and rate limits** for apprentices and projects.
-  **Agent access to our platforms:** MCP endpoints and machine-readable docs.

WHY I STAY AT THE EDGE

I invest in myself to work with the newest tools and models, **learn the patterns early**, and bring them back to my team and AISG.

Talent development, project delivery, and platform and infrastructure practices must be built together.

LinkedIn · 25 February 2026 ↗

WHAT CARRIES ACROSS

Four things I'd take anywhere



Context and a plan first

before the agent acts



Write for every collaborator

people and agents



Evidence before acceptance

evaluate, don't vibe check



An accountable owner

for everything that ships

Q&A

Questions and discussion

My practice · Orchard and RepoPrompt CE · how AISG supports its builders

THANK YOU

najibninaba.com

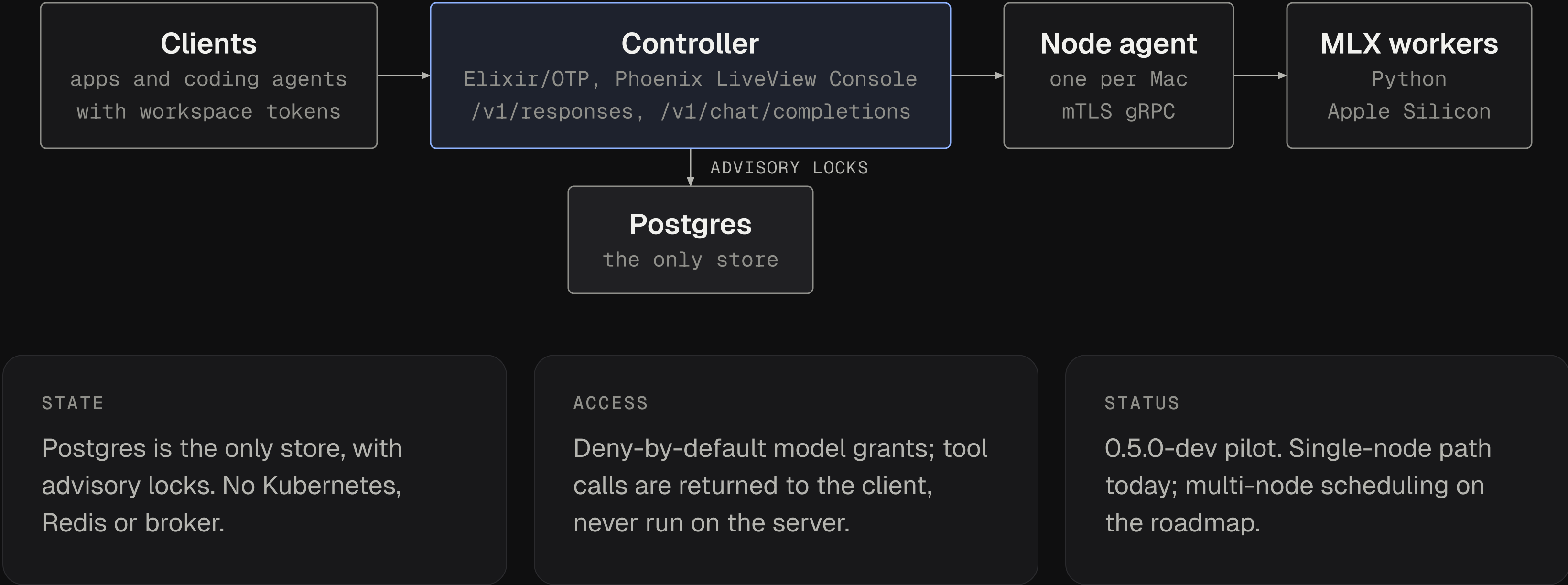
Najib Ninaba

Head of Platforms Engineering, AI
Singapore

github.com/najibninaba
x.com/najibninaba

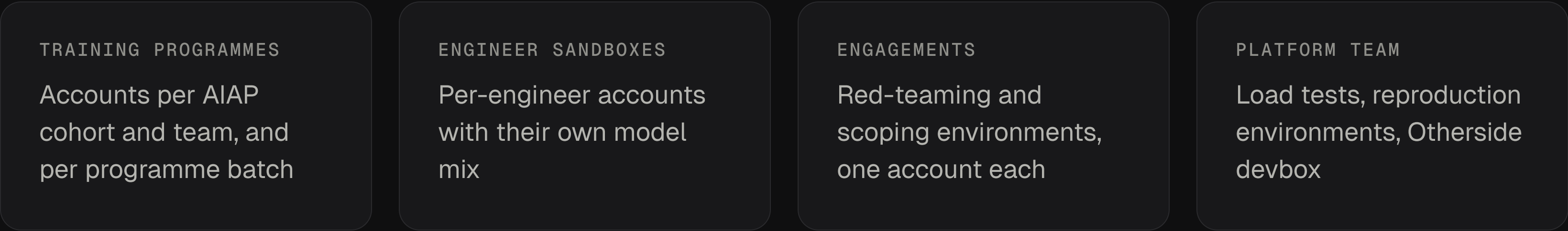
github.com/kapitan-ai/orchard
github.com/repoprompt/repoprompt-ce

Orchard, in one picture



ANNEX · MODEL ACCESS ON AZURE, JUNE TO OCTOBER 2026

Where model access sits



- GPT-5.x
- GPT-4.x / 4o
- embeddings
- DeepSeek
- GPT-6
- o-series
- GLM
- Claude Haiku

ANNEX · THE INFRASTRUCTURE WE PROVIDE

Our own fleet, plus three clouds

H100

A few dozen H100 GPUs on a Proxmox cluster

8

Mac Studio M3 Ultras, 512 GB unified memory each

100s_{TB}

storage

Azure · AWS · GCP

For 100E and SIP, the sponsor's infrastructure comes first



EARLIER, AT NUS

We once ran **close to 300 nodes** from a small OCP server farm, mainly for compute and storage. We decommissioned it when AISG moved to NTU.