UPSIDE PODCAST — #Episode 115

Revolut says it cut its AI coding costs by about 90% by self-hosting a Chinese model

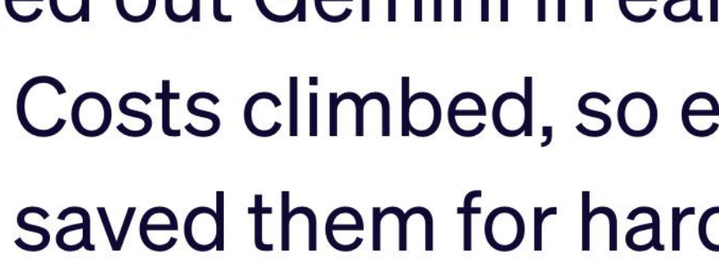
About 2,000 engineers now default to GLM, an open-weight model from China’s Z.ai, running on Revolut’s own Nvidia chips.



THE PROBLEM

Revolut’s engineers had to ration AI tokens

Revolut rolled out Gemini in early 2025 and Claude Code in early 2026. Costs climbed, so engineers were given limited tokens and saved them for hard tasks. In April 2026 Revolut began self-hosting open models, and it now runs GLM 5.3.



THE SETUP

56 Nvidia B200 GPUs serve around 2,000 engineers

Revolut puts the blended cost at about $0.30 per million tokens, roughly a tenth of what it paid Anthropic and OpenAI. Open-weight means the model’s parameters are published, so a company can download it and run it on its own hardware, which keeps its code in-house.

WHAT CHANGED

Without token limits, engineers used AI for everyday tasks

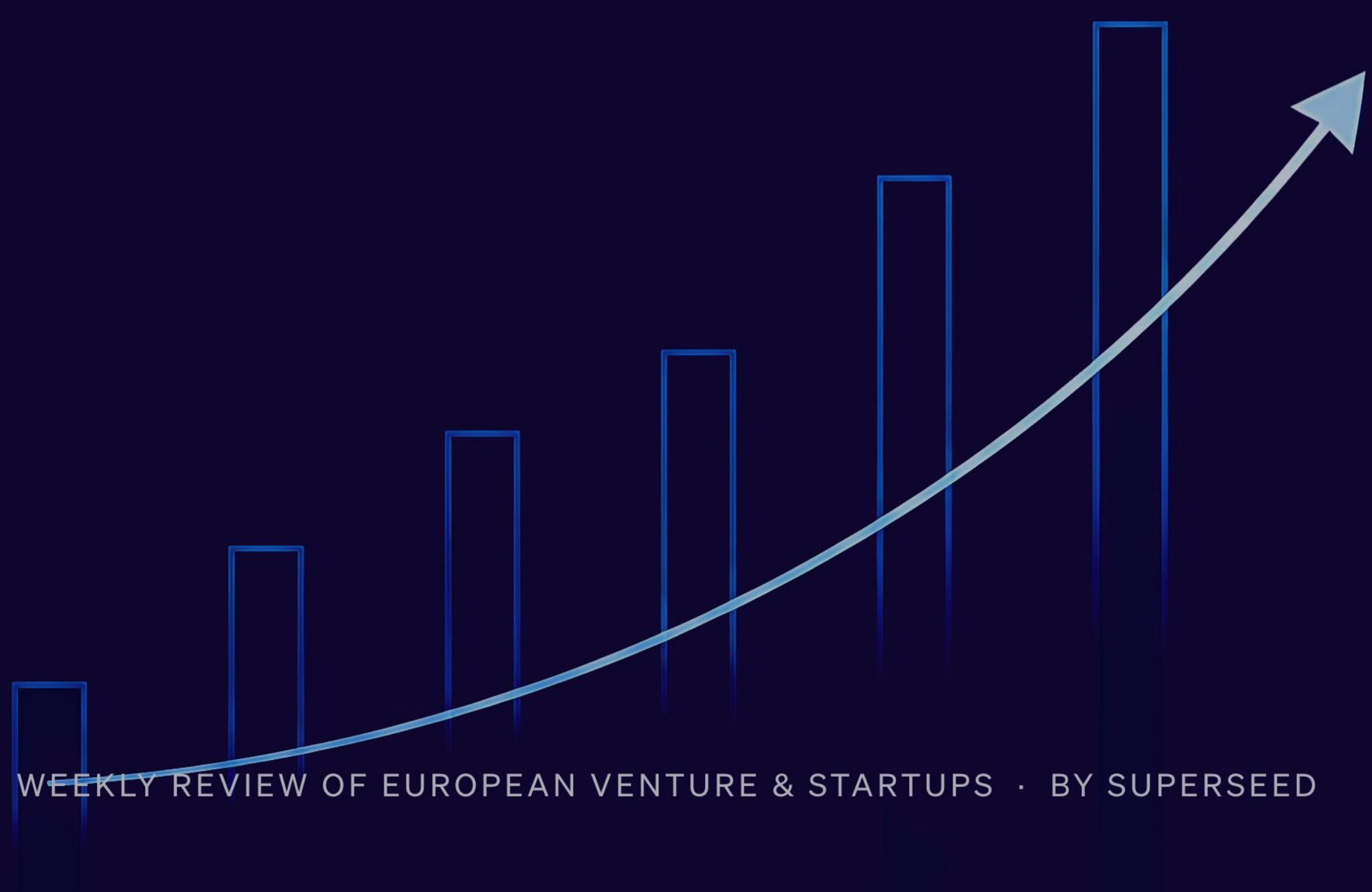
They used it for debugging, Jira tickets and documentation. Revolut reports coding time about 60% better and lead time about 70% better after the switch.



THE CAVEATS

The figures are Revolut's own and haven't been audited

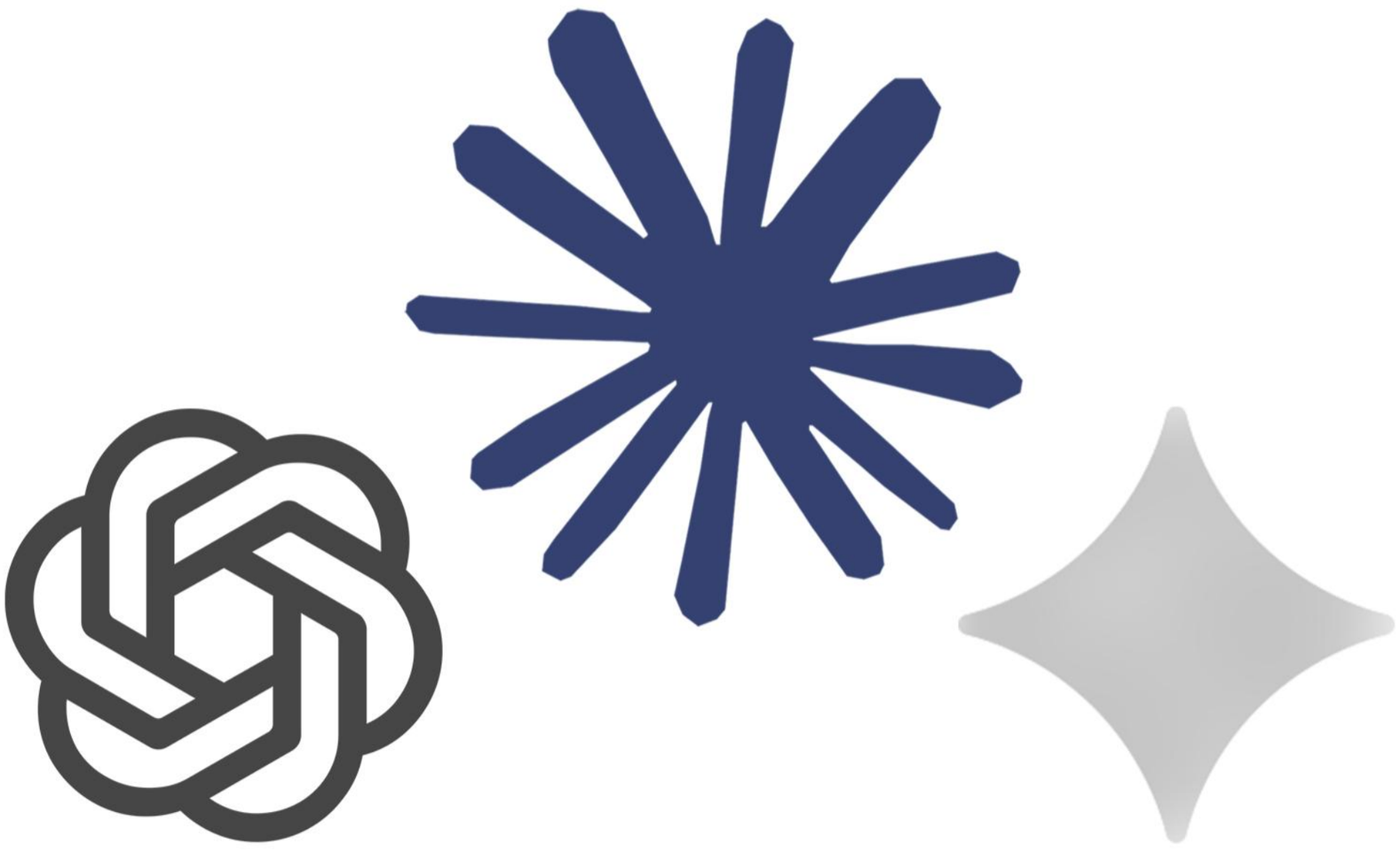
The $0.30 excludes cloud spend, staffing and monitoring costs. The conference write-up also says the productivity comparison doesn't separate the model change from other factors, so some of the gain may come from lifting the token caps.



FRONTIER MODELS

Frontier models still handle some of the work

GLM only works with text, so image and UI tasks go to Gemini or similar models. Claude Code and Codex remain available for harder tasks.



THE OTHER SIDE

In SuperSeed's tests, GLM wasn't the best value for any workload

Across tens of thousands of test cases, OpenAI and Anthropic models gave the best price and performance on almost every workload when tokens were bought from a provider. Self-hosting changes the sums only if you run your own GPUs. Mads Jensen compares it to the old debate over running your own data centre, where most companies ended up using the cloud, and expects self-hosting to be a false economy for most firms today.

CLOSER

More on AI costs, the US midterms and EU Inc in the full episode 👇

Full episode out now ↓

Find us wherever you get your podcasts.

European venture, every week

Listen / Watch

▶ PLAY EPISODE

