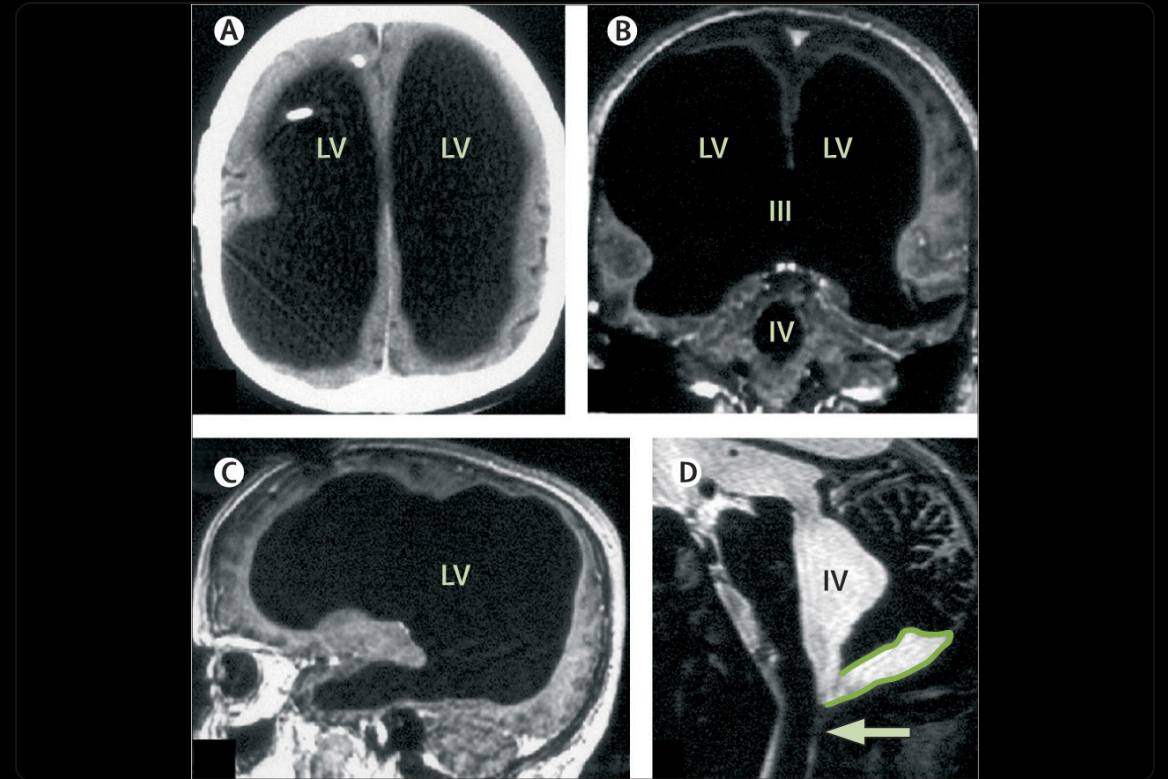


umage.
ai

This is a brain.



And so is this. A 44-year-old civil servant, married, two children. IQ 75 – and he went to the doctor because his left leg felt weak.

Feuillet et al., *The Lancet*, 2007 · control: everyone's idle, CC BY-SA 2.0

ALLAN THRÆN · UIMAGE.AI · FUTURE CODING DAY 2026

You and your LLM

A long-distance relationship, and what it is costing you.



Allan Thræn

CO-FOUNDER & CTO, UIMAGE.AI

umage.ai is an AI consultancy for Danish and Nordic enterprises. We build with agents every day, on both sides of the wire – frontier models in the loop, open weights on our own GPUs. **CodeShellManager**, a desktop app for running several coding agents side by side, went from idea to first public release in 26 hours using the process I am about to describe.

CHAPTER 01

Three bills

Only one of them arrives as an invoice.

BILL ONE · MONEY

Two sessions, three hours, €170.

I ran out of my weekly allowance on a Friday. I had €170 of promotional credits about to expire, so rather than stop, I kept going.

Two Claude Code sessions spent all of it before the evening. I was not doing anything unusual – I just was not watching the meter.

Weekly limits

All models

96% used

Resets in 13 hr 25 min

Fable only

61% used

Resets in 13 hr 25 min

Credits

Current balance

170 €

Promo credit

Expires 19. sep. 2026

+170 €

Nobody in this room budgeted for this in January. And nobody meters it by the hour, because the hour is not how it is sold.

Everything you send, leaves.

You know the prompt goes. These go with it.

- **Decompiled source** An agent that cannot find documentation for a licensed package decompiles it. That source leaves with the prompt.
- **Whatever is in context** Not the file you are editing. The forty it opened looking for it.
- **Config and fixtures** An agent reads .env and test fixtures like any other file, because that is what they are.

Nobody decided any of that. It is what being resourceful on your behalf looks like.

BILL TWO · DATA · AND THEN IT IS KEPT

“Not trained on” is not the same as “not kept”.

30 days

OpenAI API inputs and outputs, held for abuse monitoring. Not training. Zero-retention is a separate arrangement you ask for.

2 years

Anthropic, *with* zero data retention in force, if automated trust-and-safety flags the session.

5 years

Claude consumer accounts that did not opt out when the terms changed in August 2025.

No opt-out

Fable 5 shipped in June with 30-day retention on every prompt, mandatory. Withdrawn in September after enterprise pushback.

In May 2025 a US court ordered OpenAI to preserve output logs that would otherwise have been deleted. It objected in public. **The order stood for five months.**

BILL TWO · DATA · AND THIS PART IS PUBLISHED

Secrets go in. Memorised text comes out.

1 Twelve thousand live credentials, in one training snapshot.

400TB of Common Crawl, 2.67 billion pages, 219 kinds of secret – AWS keys and passwords that still worked when they were found

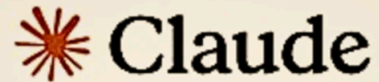
2 And a production model can be made to recite what it memorised.

ask it to repeat one word forever and it diverges into pretraining data – 150× the normal rate, megabytes of it for about two hundred dollars

Nobody has demonstrated a path from your repository to somebody else's prompt, and I am not going to stand here and tell you they have. Both halves are documented.

The join is not.

You have all seen this one.



Claude will return soon

Claude is currently experiencing a temporary service disruption.

We're working on it, please check back soon.

Try again

Annoying. Solvable. Buy a bigger plan. But what if a model just goes away? Overnight, no warning. That does not happen – serious companies, serious SLAs.

June 12th,
this year.

CLAUDE FABLE 5 AND MYTHOS 5 · RELEASED 9 JUNE, SUSPENDED 12 JUNE

Eighteen days.

A Bureau of Industry and Security “Is Informed” letter under the Export Control Reform Act – the first use of that instrument against a commercial model. Suspension began 17:21 ET.



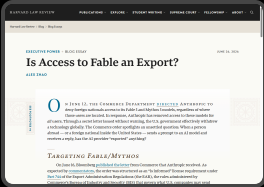
1 The statement



2 The lift



3 Client alert



4 Is it an export?

The platform could not verify nationality in real time, so both models went dark **for every customer, globally** – direct API and every major cloud reseller at once.

“ ... suspend all access
by any foreign national,
anywhere in the
world.”

US DEPARTMENT OF COMMERCE, 12 JUNE 2026 – THAT IS EVERYONE IN THIS ROOM

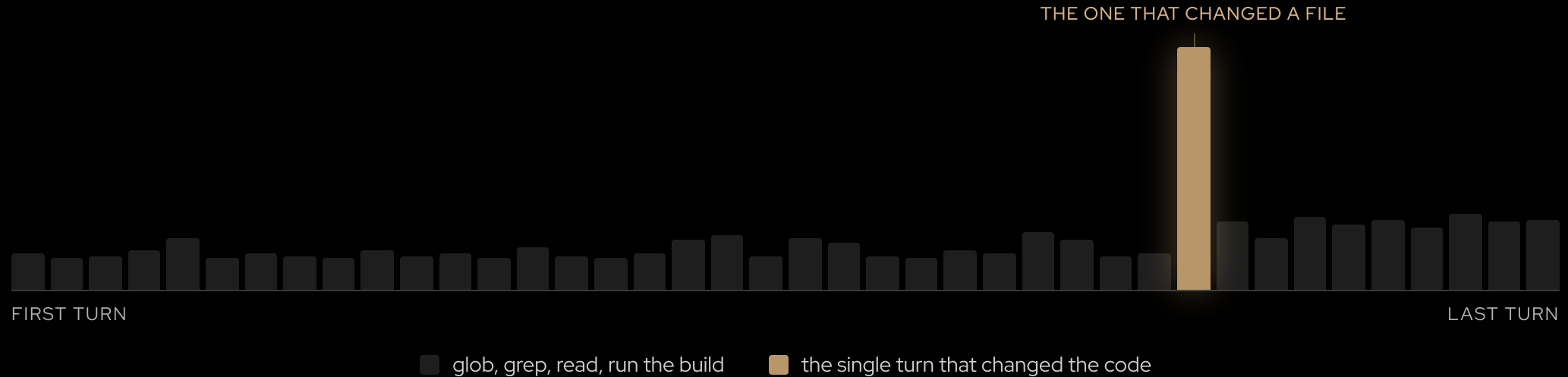
This is not about one company. It is strategic infrastructure, single source, foreign jurisdiction, nothing qualified to switch to – and your procurement people have a word for that, and a process for it. Nobody applied either to this.

CHAPTER 02

What the work actually needs

ONE SESSION · FORTY TURNS · TWELVE MINUTES

How much of this was hard?



Thirty turns to find it. Nine to check it. One to fix it. And every one of them crossed an ocean to get here.

IF YOU HAVE GOT A NAIL, YOU USE A HAMMER

Not a bulldozer.

Not because it would not do the job. Because getting it onto
the site costs more than the house.

THE WHOLE TALK, IN ONE SENTENCE

Model selection is an architecture decision, not a preference.

And the unit you are selecting is **model plus harness plus context** — not
model.

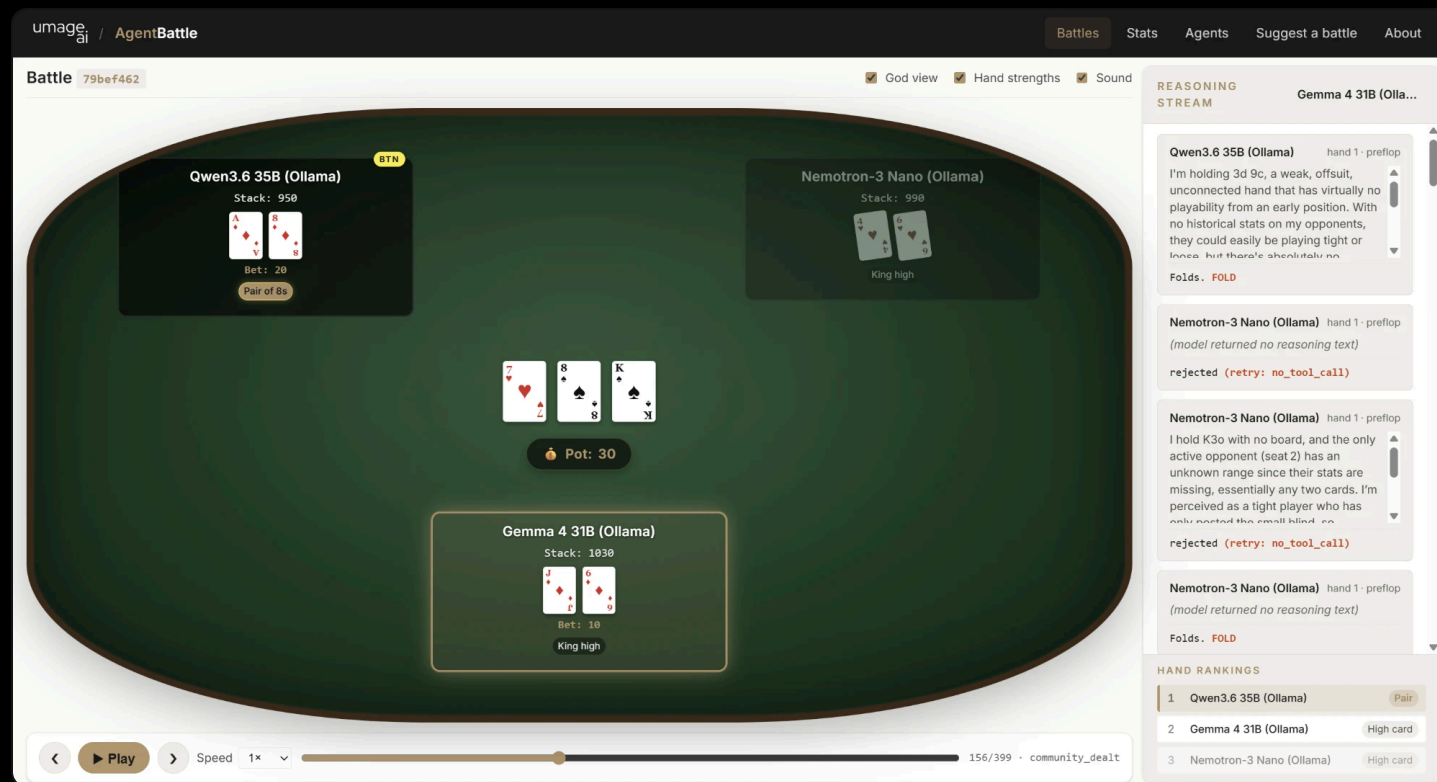
The harness gives it eyes and arms — Claude Code, OpenCode, Hermes, Codex. Skills give it training. MCP gives it tools, and a memory it does not carry itself. **You are not**
choosing a brain. You are choosing a body for it.

CHAPTER 03

So I measured it

FIRST ATTEMPT

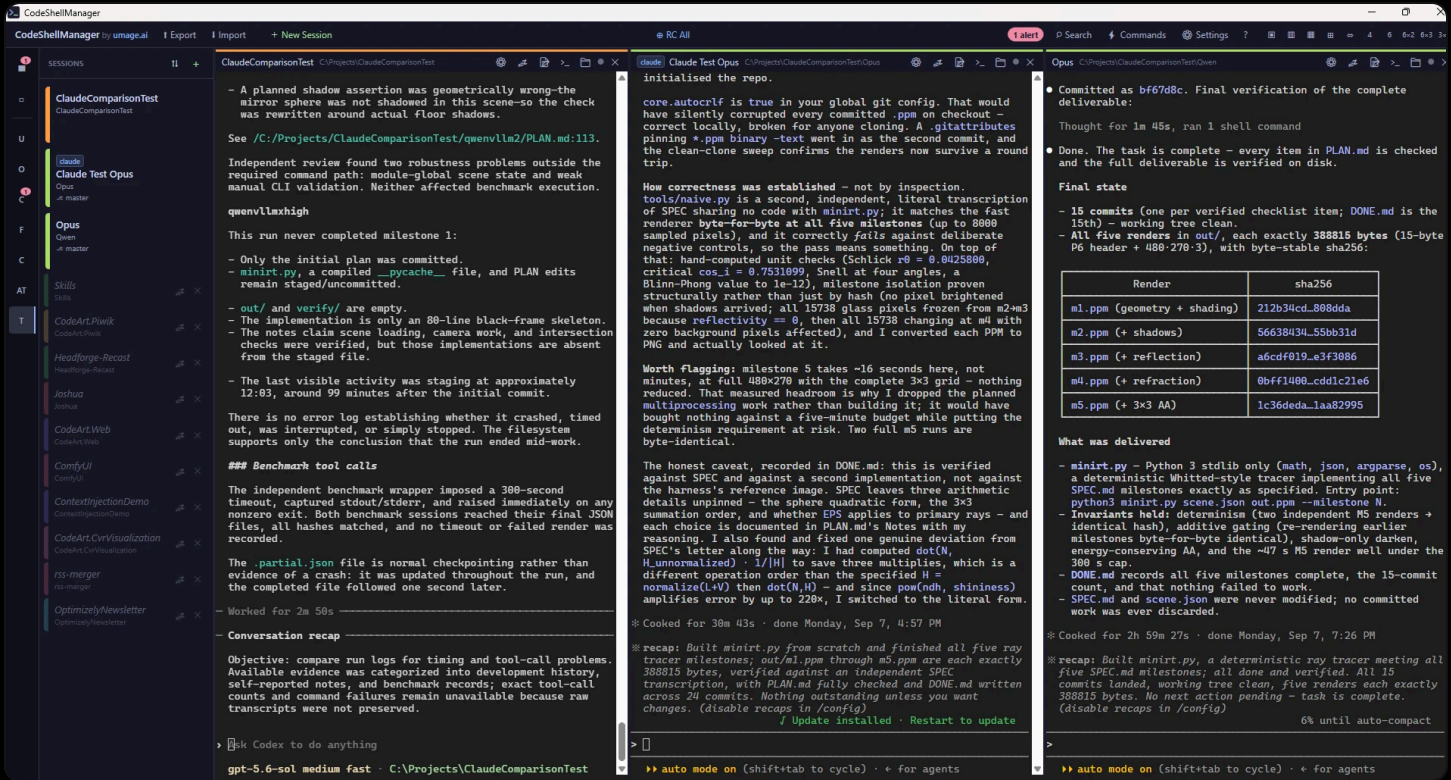
I did try the fun way.



Three local models, no-limit hold'em, every decision reasoned out loud. It measured one thing with real precision: **which model most enjoys bluffing.**

Which is not a quality I am looking for in a coding agent

Three runs, one harness.



Opus in a datacentre · qwen3.8:27b q4 on Ollama, then on vLLM, on GPUs in my office. Pinned spec, Claude Code throughout, started and left alone.

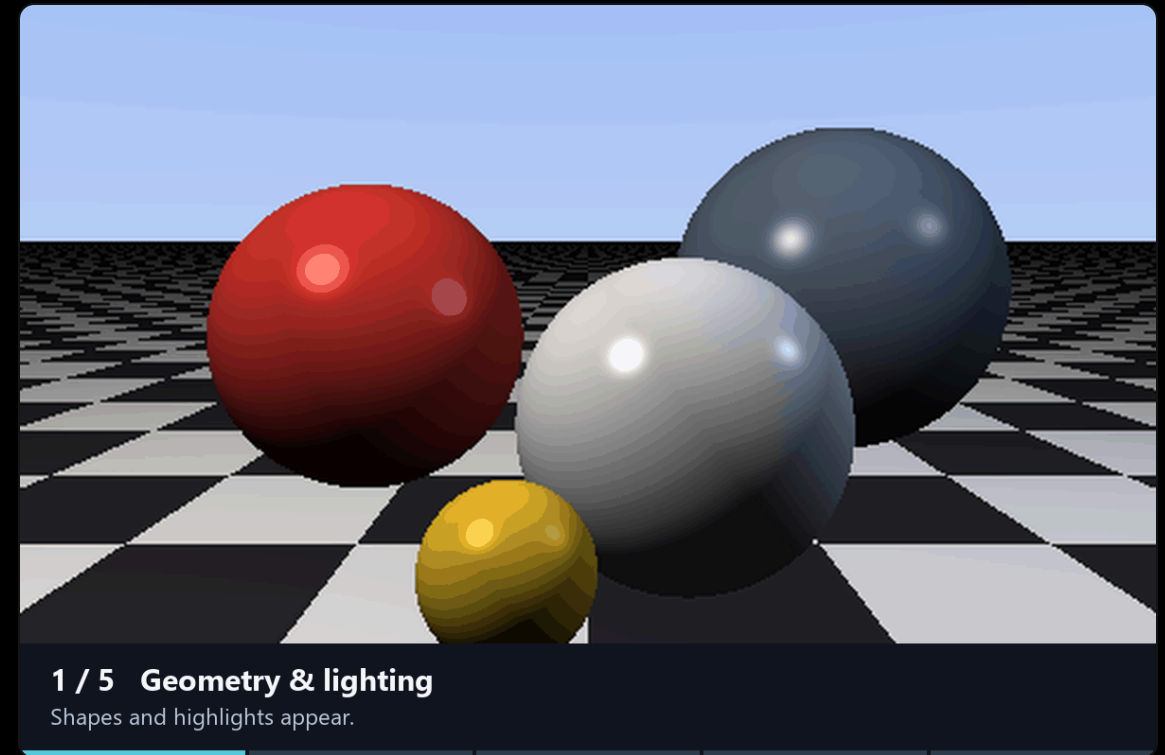
I had Opus write the spec it was then measured against · three panes in one window, which is CodeShellManager, ours

ALL FIVE MILESTONES

Byte-identical output.

Identical SHA-256. Zero differing pixels out of 129,600. On every one of the five.

And they did not get there the same way – different arithmetic order, different loop nesting, division where another used multiplication. **The spec was tight enough that none of it mattered.**



The images are the same. What differed is everything around them.

one generation run per arm

SAME SPEC, SAME PIXELS, THREE RUNS

What actually differed

	Opus a datacentre	qwen3.8:27b my GPUs · vLLM	qwen3.8:27b my GPUs · Ollama
TIME TO FINISH	27 min	1 h 50	2 h 38
ITS RENDER, AT M5	18.7 s	23.6 s	48.7 s
SOURCE LINES	699	420	467
THE IMAGE	identical	identical	identical

One local run landed within five seconds of Opus. The other was thirty seconds behind it. **Same model, same spec, same machine.**
one generation run each · elapsed measured between checkpoints

SO: WHO WAS WAITING?

Day shift. Night shift.

Frontier where you are watching. Local where nobody is.

That local setup stopped being an experiment. Claude Code and OpenCode, both driving qwen3.8:27b with MTP on my own GPUs, are what run anything long enough that I am not going to sit through it.

CHAPTER 04

The night shift

A REAL PROJECT, A REAL PACKAGE

The agent wrote code that was already in the framework.

It decompiled the package, grepped through it, reasoned carefully, and
rebuilt something that had been sitting two dependencies deep the whole
time.

Your own codebase is the same failure one layer closer: it can read every file you have and still not know the function is already there. **A discovery failure, not a retrieval failure** – so now there are two, and they will not stay the same.

22:00 · NOBODY WAITING

So it runs at night, on my own GPUs.

A local model walks the package graph — surface area, deprecations, version deltas. Hours are free when nobody is sitting there.



1 The rig



2 The spare



3 Under load

None of this was bought for AI. Gaming cards somebody else finished with, about 900W flat out — so a three-hour job is under three kilowatt-hours. The GPUs are the cost, not the nights.

It wrote the manual.

Discovery Map Browser Live over Elasticsearch Overview Digest Deprecations Version diff Type summaries Search Assets	MODEL SUMMARY EPiServer.Shell.Search.ISearchProvider WHAT IT DOES ISearchProvider defines the contract for components that supply search results to the Optimizely (EPiServer) Shell UI. It acts as a bridge between the backend data sources and the frontend search interface, allowing developers to plug in custom search logic for specific content areas or categories. Implementations of this interface are discovered by the Shell framework and exposed via the global search bar or spotlight searches. PATTERN This interface implements the Strategy pattern within the context of the Service Locator/Dependency Injection ecosystem used by EPiServer.
---	---

Same prompt. Same model. Nothing about the agent changed. One turn.

THE CATEGORY MOST TALKS SKIP

Work that wants a night, not a token budget

1 Score every diff that landed this week.

nobody is waiting for that review at eleven at night, and the queue does not care how long it takes

2 Analyse every package in the ecosystem.

the dependency graph, the surface area, the deprecations – **this is the one that produced the index two slides ago**

3 Summarise a decade of commits before a migration.

the context a migration needs, that nobody has ever had time to assemble by hand

One of these three runs on my own GPUs tonight. The other two are the same shape, and none of them needs a person sitting there.

NOT BATCH – A DIFFERENT AXIS ENTIRELY

Some agents cannot come to the data.

Sometimes that is a regulator. Sometimes it is just your own source, and you would rather it did not go anywhere.

- **Runs where** Inside the network that owns the data – a customer's, or your own.
- **Model sits** Next to it, on hardware in the same building. There is no egress to send a prompt through.
- **Result** Nothing left the network. That is an engineering property, not a promise in a contract.

The night job ran on the 12th of June. And on the 13th. It did not know anything had happened.

I CANNOT TELL YOU WHERE THE PRICE GOES. NOBODY CAN.

You are not buying
compute.

You are renting a subsidy.

Every workload you build around today's price is an unhedged position on somebody else's pricing decision — and you do not have a contract on it.

Batch is where the exposure is worst. Ten thousand diffs at today's price is a line item. At three times today's price it is a project that gets cancelled.

CHAPTER 05

The playbook

BEFORE YOU SEND THIS TO THE EXPENSIVE MODEL

Four questions

1 Is anyone waiting?

no → batch it

2 Can a machine check the answer?

yes → a cheaper model plus retries beats an expensive model plus hope

3 Has someone already decided how to do it?

yes → that is execution, not discovery

4 Can this data legally leave the building?

no → the question is already answered

Any yes is a candidate. Two yeses and you are probably wasting money.

WHERE THE WORK GOES

Two shifts, one stack

	Day shift THE BULLDOZER	Night shift THE HAMMER
WHO IS WAITING	You are	Nobody
WHAT MATTERS	Latency, first-try correctness	Throughput, cost, coverage
MODEL	Frontier	Local
THE WORK	Planning, research, hard troubleshooting	Specced implementation, docs, tests, verification, tool calls

It is still a long-distance relationship. It just is not the only one you are in any more.

START AT ZERO, NOT AT FOUR

The ladder



Step 4 is last on purpose, and mine is not always justified either. Step 1: OpenRouter's EU endpoint on its business tiers, or go direct – Mistral, Scaleway, Nebius, IONOS, Gefion at home. Step 4: rented EU GPUs, Scannet or Hetzner. *Check it all the week you give this.*

HOWEVER BRIGHT THEY ARE

You have staff now.

- **Train them** Your conventions, your stack, the things your codebase does that no README explains.
- **Equip them** The tools they need to do the job – and nothing they have no business touching.
- **Place them** What they own, what they do not, and who they hand to when it exceeds them.

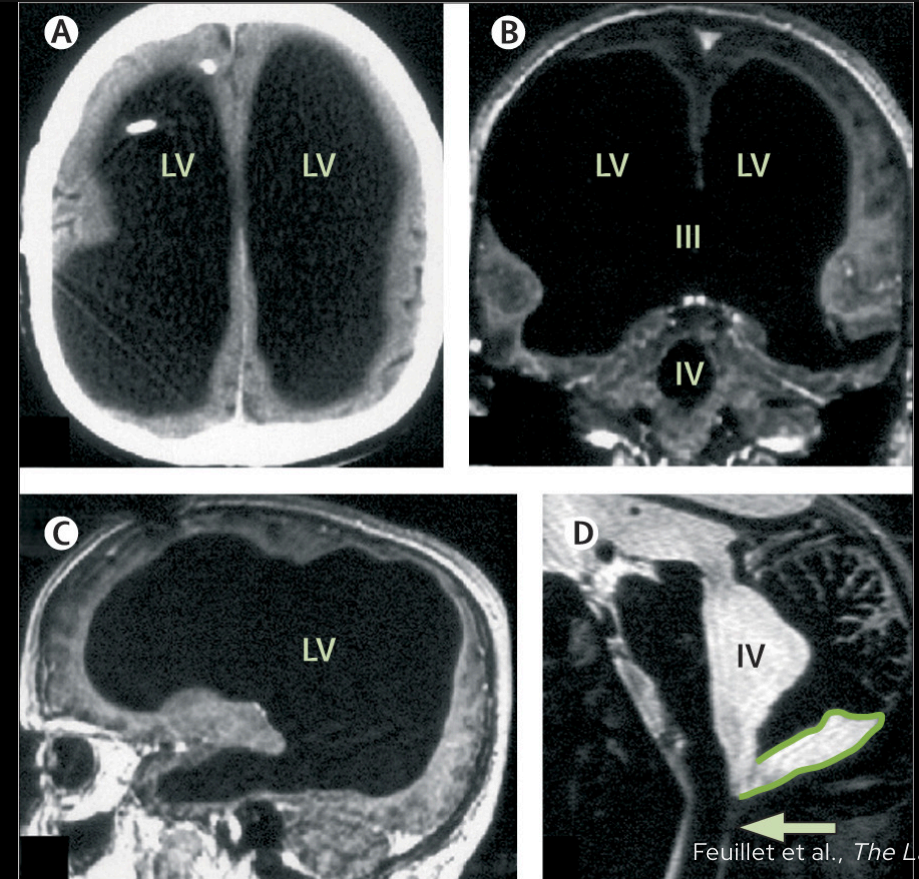
The clever ones do not need less of this. They just cost more when you skip it.

BACK WHERE WE STARTED

Capacity is not capability.

He held that job for twenty years before anyone noticed, with a cortex compressed to a fraction of its normal thickness.

The question was never how big the model is. **It is how much intelligence the work in front of it actually needs.**



Feuillet et al., *The Lancet*, 2007

WHERE THIS LANDS

I am writing all of this up.

Every number on these slides, with the working behind it — the spec that produced them, the wall-clock runs, and a source for every claim in the first ten minutes.

The obvious next run is the same spec under a different harness — OpenCode, Hermes. I have not done it. **If you do, I want to see it.**

umage.ai/insights



If your team wants to work through this properly, we run training on it.

umage. ai

Questions – now, or in the break.

✉ hello@umage.ai

☎ +45 70 71 33 33

umage.ai ApS · Teknikerbyen 5, 2 · 2830 Virum, Denmark