



Inference for agents that work for hours, not seconds.

• In stealth · Pre-Seed

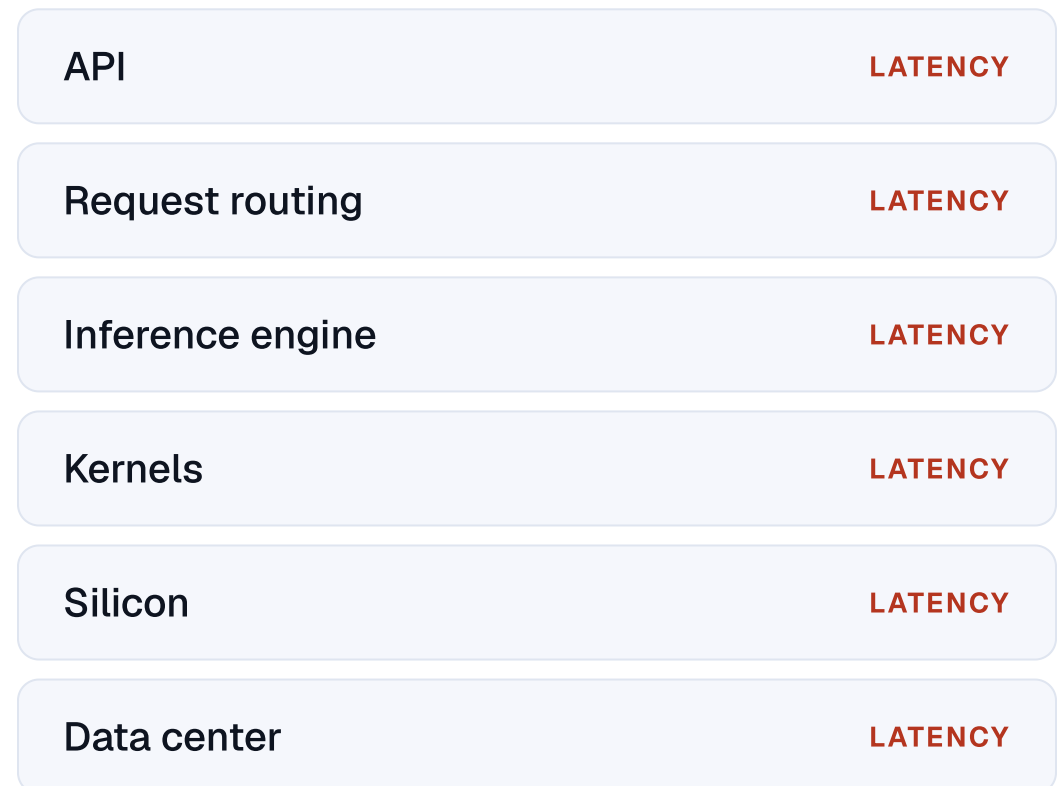
Jonathan Rosado, Founder

weinfer.ai

THE PROBLEM

The stack that serves today's models was built for **chat**, not for **agents**.

Providers sell to humans, and humans cannot wait, so every layer buys speed. Agents can wait. The opposite end of that trade fits a future where agents, not people, consume most inference.



Six layers. Every one bought speed.

From latency-first to throughput-first.

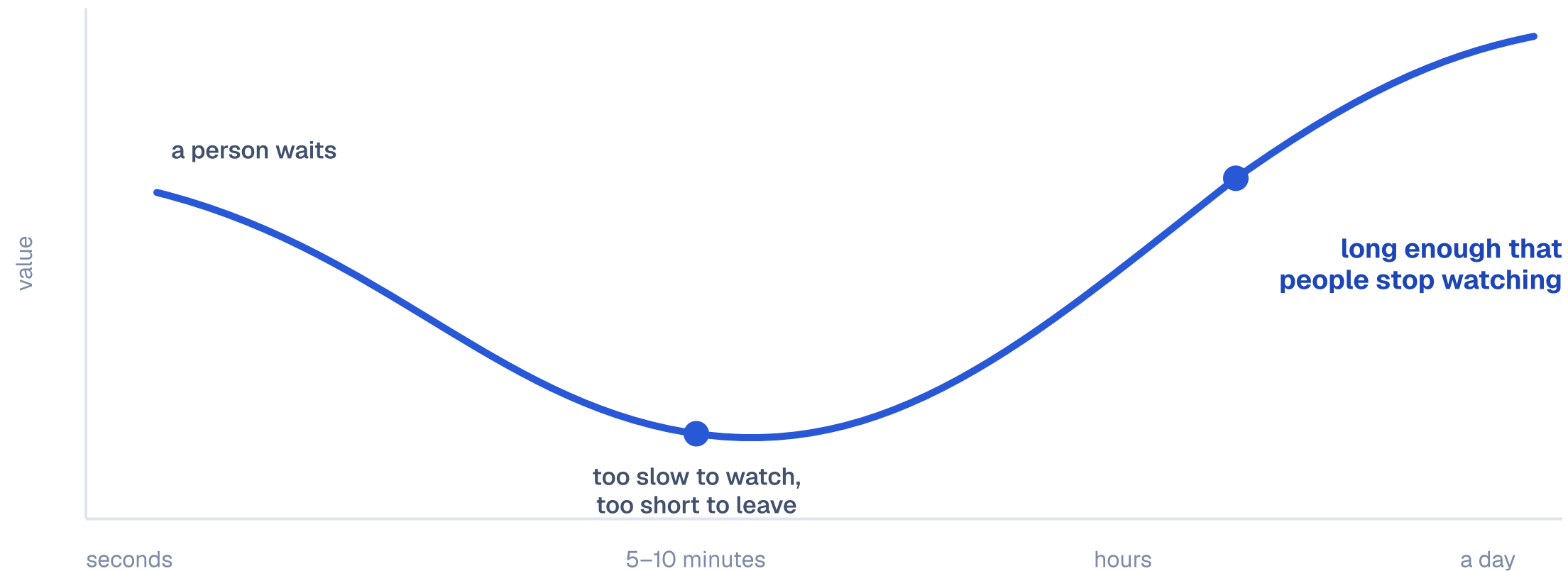
Customers tell us how long they're willing to wait for each answer. The longer they can wait, the cheaper we make it.

- idle capacity
- deeper batches
- the most efficient chip

API	THROUGHPUT
Request routing	THROUGHPUT
Inference engine	THROUGHPUT
Kernels	THROUGHPUT
Silicon	THROUGHPUT
Data center	THROUGHPUT

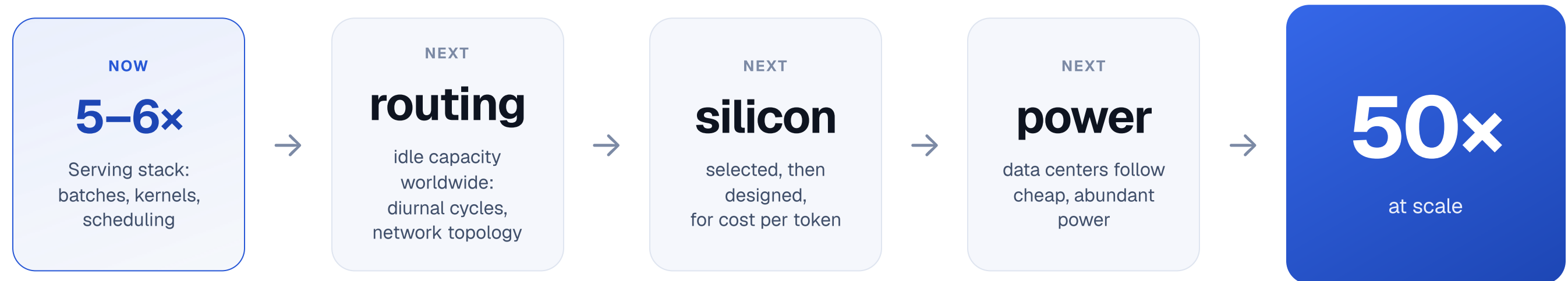
WeInfer

Agents just crossed into work that **can wait**.



When someone is waiting, less time is better. When no one is waiting, more time means lower cost.

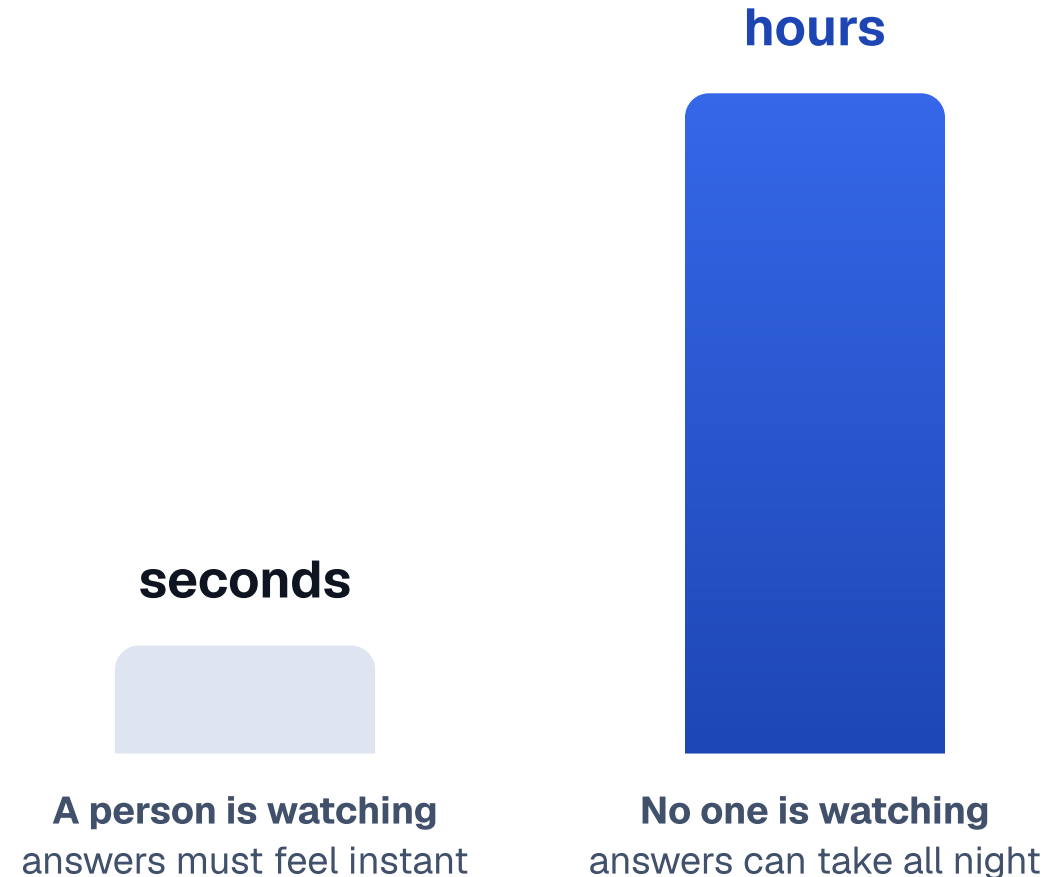
Inference **fifty times** cheaper.



Five to six times is what the serving stack gets today, and that vertical still has headroom: pushing past six is the immediate goal. Routing, silicon and power take it the rest of the way.

MARKET

With a person in the loop, only speed matters. Take the person out, and **only cost does.**



Everything on the left pays for speed. Everything on the right doesn't have to.

The only ceiling left is the bill. **We lower it.**

FIRST CUSTOMERS

Agent startups running **open models**

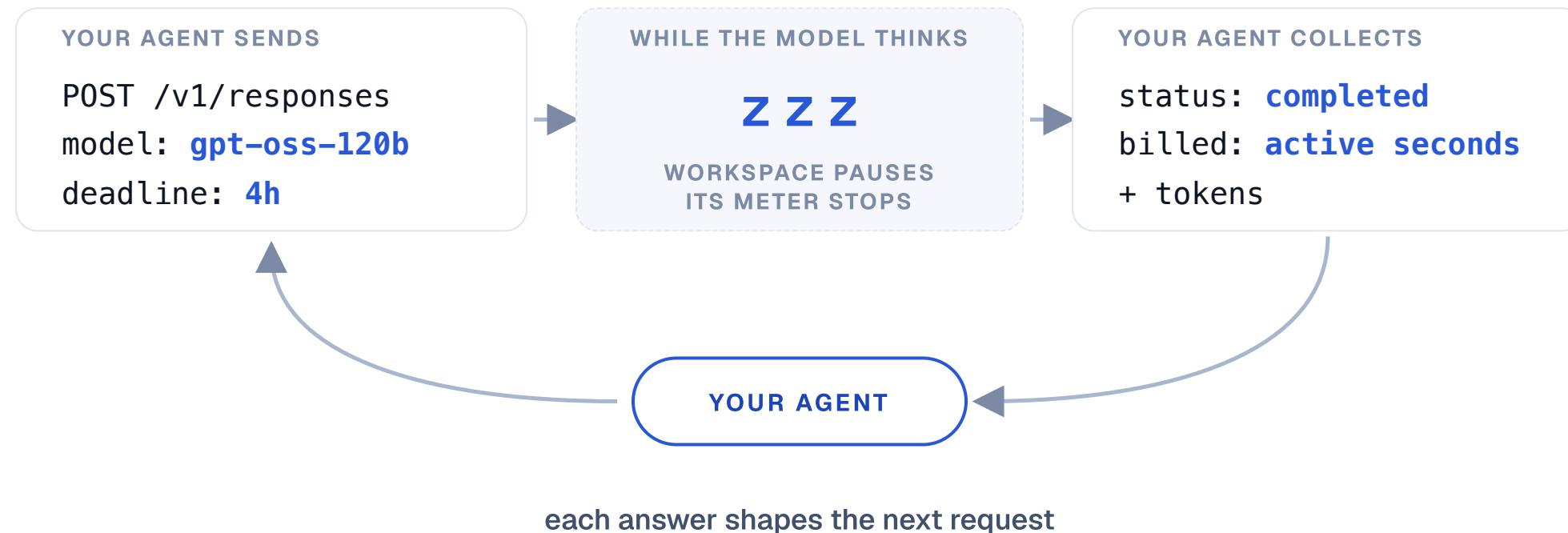
coding

code review

research

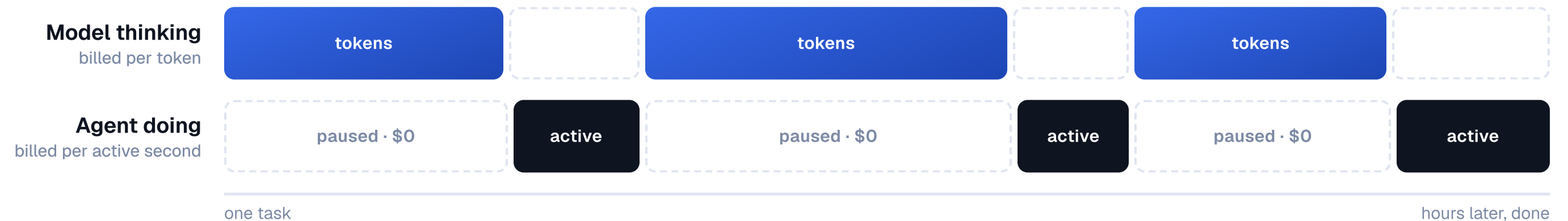
data pipelines

One API, **built for agents**. You say when you need each answer.



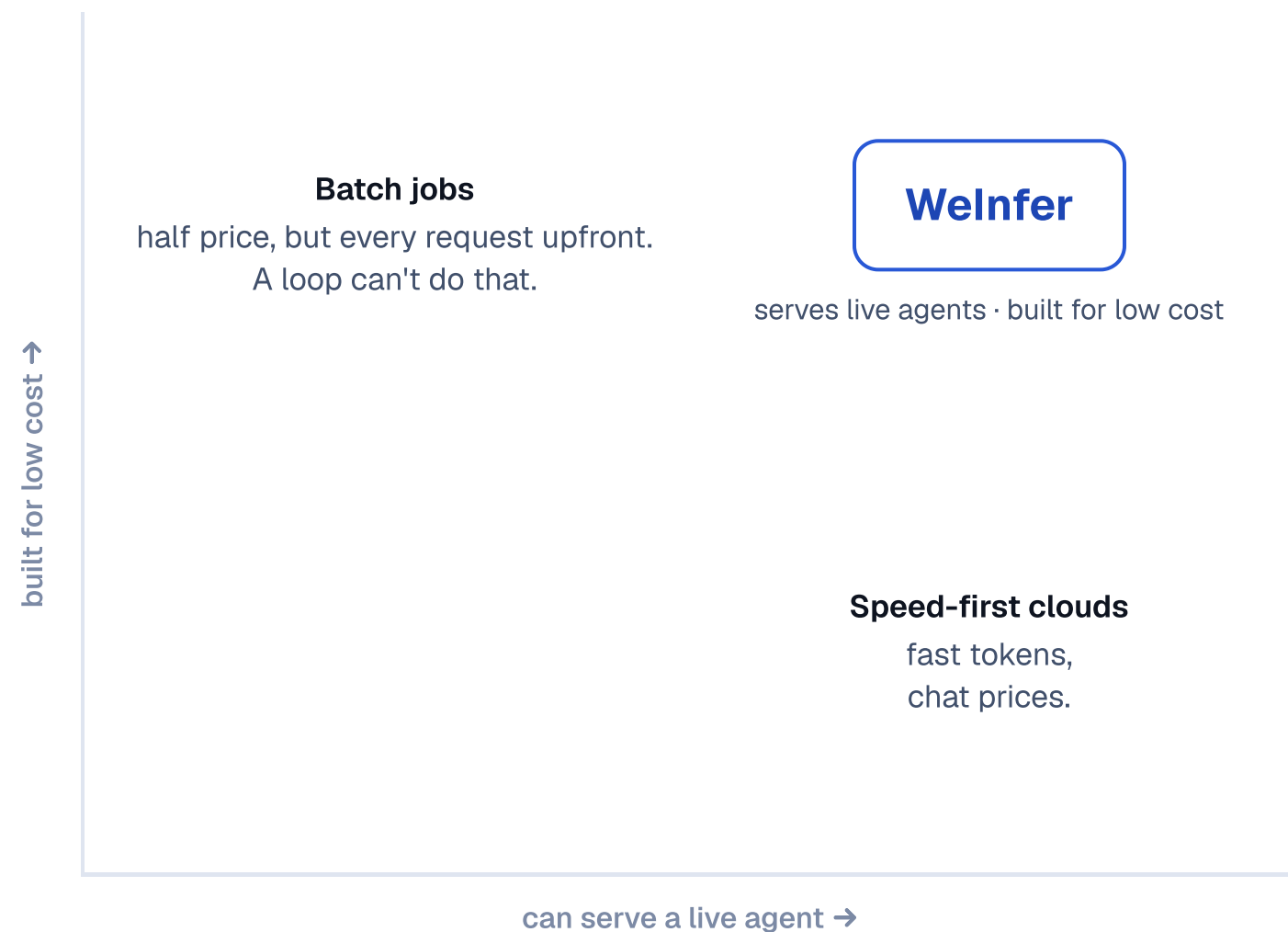
The deadline is the price dial. **The longer you give us, the less you pay.**

You pay for thinking and for doing. **Never for waiting.**



While the model thinks, the agent's machine pauses and its meter stops. We manage the business on one number: **cost per successfully completed task.**

Everyone optimizes for a human waiting. Nobody optimizes for an agent that can wait.



Batch jobs cannot follow us.

They need the whole task upfront, and a live agent writes its next request only after the last answer arrives.

Speed-first clouds will not follow us. Their margins come from selling speed.

Agents can wait. Waiting is cheaper. **Nobody else sells that.**

I have run this trade before, inside Meta.



Jonathan Rosado

Founder
Staff ML Infrastructure Engineer, Meta 2022–2026

Ran inference efficiency for **FBPredictor**, the serving platform behind Meta's ads business, across **MTIA, NVIDIA and AMD** simultaneously.

3x

cheaper inference, with the same
cross-stack strategies this deck describes

Make inference radically cheaper, and intelligence becomes infrastructure. Priced like power. Spent on everything.



jon@weinfer.ai weinfer.ai