

AGENT LIBRE

FREE WEBINAR · 45 MIN · ONLINE

MANAGED AI INFERENCE SERVICES

Understand your AI bill before it starts deciding for you

Cost structure, margin impact, five-year outlook, and the risk of depending on a single vendor.

🍁 Hosted and operated in Canada | Montréal, Québec

AGENDA

- What your AI bill is really made of, and how it moves
- Its impact on your margins and your competitiveness
- Five-year outlook: costs and usage both rising — and you?
- Betting everything on one vendor: the concrete risks
- AI is now part of your operating margin — and between geopolitics, the US dollar and tariffs, who controls it?

STARTING POINT

Agent demos are impressive. They are not free.

You see the videos: one agent handling support, another writing code, another doing the bookkeeping. Nobody shows the meter running the whole time.

ONE HUMAN

~5,000

tokens generated per day. He sleeps, he eats, he takes vacation.

A FLEET OF AGENTS

~100,000,000

tokens per day: 20 agents at 50 tokens/second, around the clock. It never sleeps, and every turn re-reads the whole history.

THE GAP

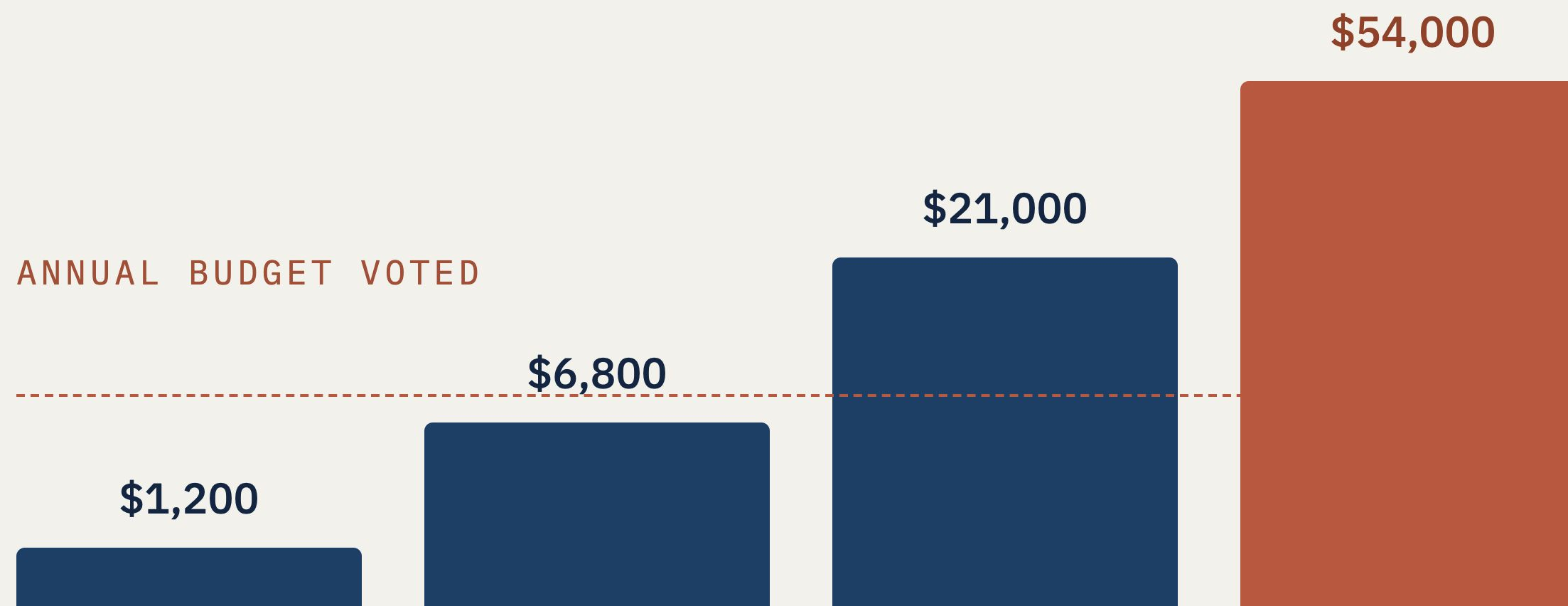
20,000×

Your bill no longer tracks your headcount.

Math: an autonomous coding agent sustains roughly 50 tokens/second, or 4.3M tokens a day – 20 agents running continuously gives ~86M. On the human side, 2,000 words written a day is about 3,000 tokens.

THE DYNAMIC

The more we adopt it, the more it costs



Typical progression observed over 12 to 18 months of real usage.

An assistant drafting her emails, then customer service answering with an AI agent, then invoicing and quotes, then prospecting around the clock. Every step is good news for the business.

And every step is bad news for the bill. So you hit the brakes — quotas per person, limited tools, projects pushed back.

You end up optimizing the bill instead of the process.

WHERE DO YOU SIT

Three profiles, one horizon

COMPANIES OF
30 TO 60 EMPLOYEES

01 **\$2,000 / month**

“That's enough.”

WHAT WE SEE

AI adoption still low. The tool is there, the processes are not.

The transformation is coming in the next 16 months.

02 **\$15,000 / month**

“That's enough.”

WHAT WE SEE

AI adoption in line with current trends — except you may be paying up to 5× the real cost.

Competition will pull your volumes up, and your bill with them.

03 **\$15,000 / month**

“And I'm capped.”

WHAT WE SEE

AI adoption ahead of current trends. You are already throttling your teams.

The ceiling is no longer technical, it is budgetary.

For profiles 2 and 3, sovereignty and cutting out the middlemen come up on their own.

Everyone at their own pace. But it is a certainty.

WHAT YOU ARE REALLY PAYING FOR

Between the GPU and you, there is a crowd

01

The data centre

Building, power, cooling, GPUs. This is where the work actually happens.

+

02

The hyperscaler

Rents the hardware by the hour, in US dollars, with its margin.

+

03

The model lab

Amortizes its training, its R&D and its global marketing over every token you generate.

+

04

The usage-based API, then the aggregator

Bills per token, output at 3 to 5× input, re-bills every re-read of the history — and the reseller adds its own layer.

=

You

Your bill

You pay for the same GPU four times — and you do not know what it costs.

Real marginal cost of one million input tokens on a dedicated server:
~\$0.09 CAD

Billed at **20 to 45×** that price.

THE BLIND SPOT

Watch for your indirect AI spend

You buy a great AI-powered tool — it handles your customer support, it does your prospecting. What the contract does not say out loud: it passes its usage bill straight to you, with its margin on top.

It is a token bill you pay without seeing it, without measuring it and without any way to negotiate it.

THE HABIT TO BUILD

Ask to run that software on your own AI infrastructure. Serious vendors let you plug in your own keys.

THREE REASONS TO INSIST

01 **Your costs**

That volume runs on throughput you have already reserved, instead of being re-billed per token by a third party.

02 **Your data governance**

Your client files and sensitive data stay on your inference layer, with your own audit logs.

03 **Your AI culture**

Your own internal guardrails apply everywhere: same rules, same filtering, whichever tool is calling.

The question to ask before signing: **“Can I use my own AI inference API keys?”**

The two reflexes that do not survive the long term

“We'll just run it on our laptops”

- A laptop is nowhere near the 50 tokens/s a single agent needs.
- GLM 5.2 does not fit on a laptop: the shrunken model that does is not the one you tested.
- Memory blows up well before 256K tokens.
- The machine is tied up, and nothing is pooled.

A 1,500-TOKEN RESPONSE

Laptop, shrunken model: ~8 tokens/s **3 min 07 s**



Reserved throughput, ~99 tokens/s per agent **15 s**

Fine for a demo, never for production.

“We'll buy a personal seat for every employee”

The equivalent of letting them work from their personal Gmail.

- No usage monitoring: it gets used to plan vacations and everything else.
- Leakage: your capacity goes to personal side projects.
- Client data sits in a personal history, out of your control.
- A seat is built for a human: your agents do not stop at 5 p.m.

A tooling expense, not an inference strategy.

“The minute” was invented too

Telecoms billed by the minute. Their real cost was infrastructure: towers, copper, fibre — paid once, then amortized. The minute appeared nowhere on their balance sheet.

01

A billing unit, not a cost

The minute was invented to turn fixed infrastructure into variable revenue. The more you talked, the more it earned — without costing any more.

02

And everyone held back

You counted your calls, called back after 6 p.m., saved long distance for Christmas. Exactly what your teams do today with their tokens.

03

The flat plan settled it

Unlimited plans did not arrive out of generosity but out of competition: someone finally billed what actually cost money — network access.

The token is the minute of 2026. **The real cost is the GPU — and a GPU can be reserved.**

THE FLIP

You no longer buy tokens. You reserve throughput.

A managed service provider runs your email, your website or your security system. A managed inference provider does the same for your AI: capacity sized, operations monitored, bill fixed. Neither a model lab nor a reseller.

THE USAGE-BASED MODEL

You buy tokens by the unit

You absorb every swing in load

You watch the quotas and the rate limits

The vendor decides the model and the context

You take the price hikes and the exchange rate

MANAGED INFERENCE SERVICES

You reserve throughput, not tokens

Swings are measured; the flat fee does not move

Capacity and uptime are managed for you

Load is syndicated across members

You cut out the middlemen between the GPU and you

On usage billing, what you mostly pay for is the right to consume anytime

USAGE - BASED

The right to show up whenever

Zero commitment: you can consume nothing this month, or ten times more tomorrow morning

The vendor has to hold capacity in reserve to absorb that freedom

That reserve is paid for by everyone, on every token

That is a flexibility premium.

RESERVED THROUGHPUT

Tokens per second, guaranteed

You commit to a capacity, and it is reserved for you at all times

No reserve to finance: the capacity is already yours

No premium to pay, so a price that reflects real compute

You pay for a server, not an option.

Why pay the "anytime" premium? **Your agents run every day, at a rate you already know.**

The syndicate

Your reserved throughput on shared infrastructure. What the flat fee covers:

01

Technical operations

Guaranteed throughput, reserved in advance

Continuous model optimization on the GPU

Same hardware, rising throughput over time

02

Administration

Matching members, arrivals and departures

Freed capacity takes on more members

Data centre leases and partners

03

Cost negotiation

Arbitrage as combined volume grows

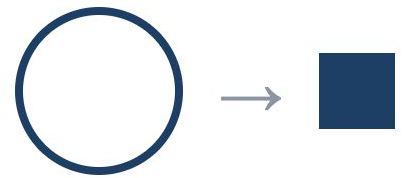
Continuous GPU market watch

Agreements renegotiated every 6 to 12 months

All you do is point your volume at the shared gateway.

THE ANALOGY, CONTINUED

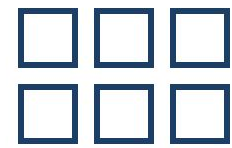
The meter, or your own panels



THE METER

Hydro-Québec

You plug in, you consume, you pay per kilowatt-hour. No capital outlay — and no say on the rate. That is the big AI vendor.



PANELS, ON YOUR OWN

On your own roof

You buy, you install, you maintain, you handle the overloads and the outages. Real autonomy — but an expertise you do not have.



PANELS, OPERATED

The managed provider

A managed provider installs a system that belongs to you — or is leased to you — and carries the expertise. You still cut out every middleman between you and the electricity.

The installation is yours, the expertise comes with it — and between you and the electricity, there is nobody left.

THE NUMBERS

What the same capacity costs, depending on the billing unit

FLAT MONTHLY PLAN — MANAGED SERVICE

\$18,000

Reserved throughput: **1,980 tok/s**

≈ **40** agents at 50 tok/s — or 20 agents at **99 tok/s** each

The amount does not change if your agents run around the clock.

THE SAME CAPACITY, BILLED BY USAGE

50% utilization

\$52,700

per month

70% utilization

\$73,800

per month

100% utilization

\$105,500

per month

Math: tokens generated by the reserved throughput over a full month, at Claude Sonnet output pricing (US\$15/M), converted to CAD.

What if your competitor pays 600% less for the same task?

This has nothing to do with technology: it is a cost advantage on every quote, every email, every file processed.

OPTION A

They cut their prices

Their processing cost per file is six times lower. They can cut prices and stay profitable — you cannot. You lose bids without ever seeing why.

OPTION B

They keep their margin and do six times more

Same price, same margin — but six times more prospecting, client follow-ups, content, new markets tested. On an equal budget, they move six times faster.

Inference cost is no longer an IT line item. **It is built into your price structure.**

PERFORMANCE

“Sure, but is the model any good?”

GLM 5.2, Z.ai's open-weights model, on public and independent evaluations.

EVALUATION	GLM 5.2	Claude Sonnet 5	Claude Opus 4.8	GPT-5.5
SWE-bench Pro	62.1	63.2	69.2	58.6
Terminal-Bench 2.1	81.0	80.4	85.0	84.0
FrontierSWE	74.4	—	75.1	72.6
IDOR detection (Semgrep)	39% F1	—	32% F1	—

For the bulk of day-to-day agentic work, the quality gap is thin. The cost gap is not — and that is what makes the flat fee possible.

And the trend is on your side: the lag between open weights and the private labs has gone from 6-9 months to 3-5 months.

Sources: Emergent, Digital Applied, Semgrep, Eden AI, Interconnects.

THE NUANCE THAT MATTERS

What you experience day to day comes from the harness, not the model

The agent harness is all the code around the model: the tools it can call, the memory, the retries when something fails, the guardrails, the way a task gets broken down.

THE HARNESS

What makes the felt difference

The right tools wired in at the right moment

The context you give it — and the context you spare it

Retries and verification when a step fails

Your internal rules, applied on every call

AND ON COST

The harness is a fixed cost

Written once, owned by you, improved over time

Never re-billed per token

Works with whichever model you choose

Switching models does not throw it away

You invest in the harness once. **The model becomes interchangeable — which is exactly the point.**

WHERE IS THE CATCH

How does an open model even exist?

The closest equivalent: a scientific paper published in open access. The method is public, anyone can reuse it, verify it, build on it.

01

Published, not given away

What is open is the model parameters: the result of the training, published like a research paper. You can download it, audit it, run it wherever you want.

02

Why they publish

A huge community of scientists and developers publishes out of contribution and reciprocity. In exchange, whoever publishes keeps the lead on their community — and collects the reputation that comes with it.

03

The catch is elsewhere

The model is open; running it is not. It takes GPUs, operations, optimization and uptime — the trade of a managed inference provider.

The models are open. **The hardware that runs them is not — and that is the only bill left.**

Inference has become a line in your margin. Who controls it?

THE EXCHANGE RATE

Your rates are in US dollars. Your revenue is not. Every move in the rate goes straight into your margin.

TARIFFS AND GEOPOLITICS

Hardware, energy, model access: three political files you have no leverage on.

THE SINGLE VENDOR

They change the price, the quota, the model or the terms: you find out by email, and you absorb it.

RISING USAGE

Your agentic volumes will grow by an order of magnitude. Billed by usage, your bill follows the same curve.

Reserved throughput, under contract, hosted and operated in Canada: that is a margin decision, not a tooling decision.

It is also the only version of this math your finance team can commit to over three years.

Your critical processes. Your key AIs now sit alongside your key employees.

An employee, you know how to manage

- Salary bands give you cost predictability over the year.
- If they are swamped, they come to you: before hiring, you talk optimization.
- Staff turnover you have already lived through: people can be replaced and trained.

It is predictable.

A single AI vendor, less so

- They change the price, the quota, the model or the terms: you find out by email.
- Your critical processes are built on its behaviour: its answers, its latency, its limits.
- Switching vendors is not a swap: it is retesting every process, every prompt, every guardrail.

It stays very abstract.

The more you build on it, the less you can leave it. **Reserved throughput and open-weights models are what keep your sovereignty.**

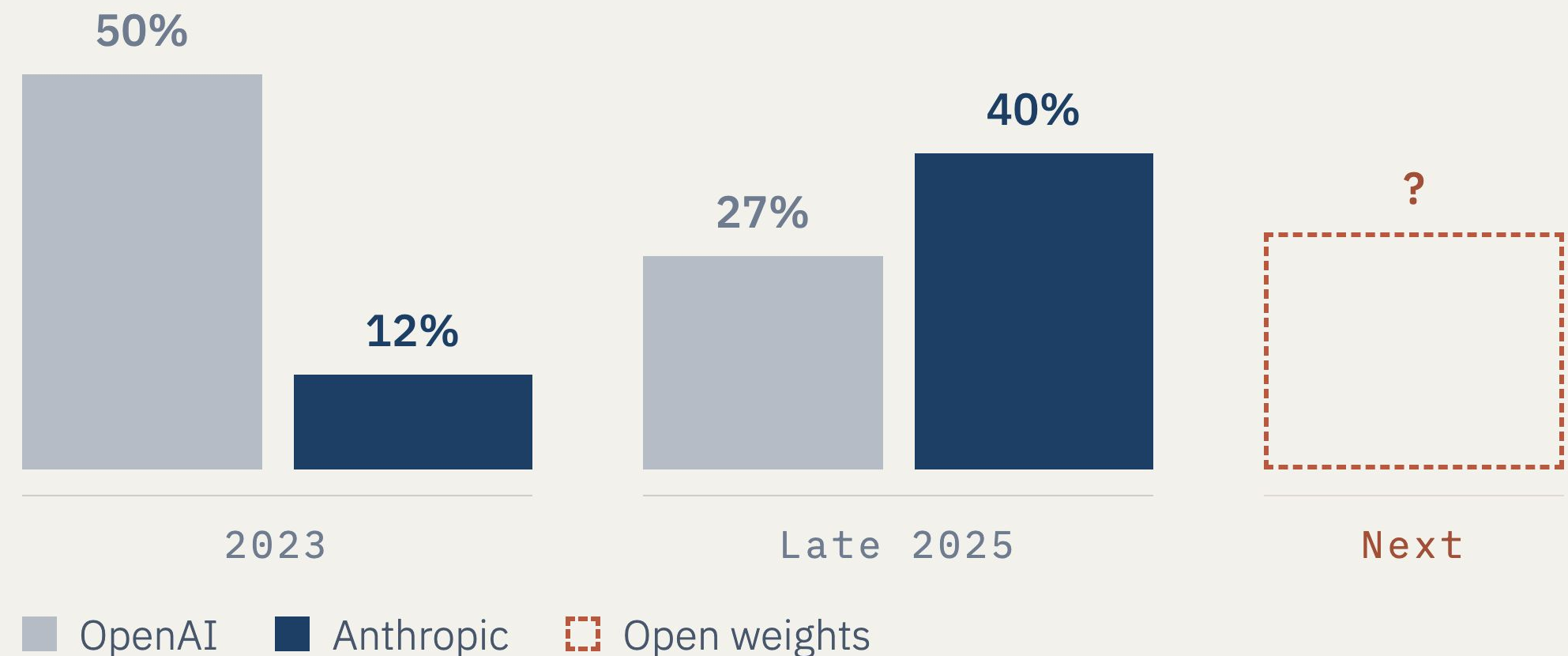
A BIT OF HUMILITY

In 2023, everyone swore by ChatGPT

“Just put it on ChatGPT” was the answer to everything. Three years later, in the enterprise, Claude leads — and nobody apologizes for it.

The question that arises: why would Claude be the end of the story?

Open-weights models may be living their own “ChatGPT” moment — the one Claude handed ChatGPT.



Enterprise LLM API market share, Menlo Ventures (2023 → late 2025). On code alone, Claude reached 54% versus 21% for OpenAI.

AGENT LIBRE

WEBINAR · 45 MIN

Thank you

Questions? Now is the time.

Free the agents. Fix the bill.

 Hosted and operated in Canada · Montréal, Québec