THE VARIABLE COST THAT'S EATING ENTERPRISE AI BUDGETS

Pay the Token Tax. Better yet, stop paying for tokens.

If tokens are 98% cheaper, why are AI budgets 6× bigger?

Here is why the meter that never stops is the wrong way to pay for intelligence at scale — and what replaces it.

Intelligence Unshared: The AI Sovereignty Papers — this is the second in a six-paper series on Private AI.
Find the series at **iterate.ai/resources/white-papers**

PREPARED BY

Iterate.ai · Engineering & Research
With the Lifeboat inference team

EDITION

Economics Brief
The cost of renting intelligence

SUBJECT

Per-Token Pricing
Where the meter breaks at scale

WHEN THE METER NEVER STOPS

Twelve companies. One pattern.

AI budgets built for chatbots — **burned by agents.** These aren't projections. They happened in 2026, and they're why the token meter is now a board-level line item.

UBER

Budget gone by April

Put Claude Code in front of its 5,000-engineer workforce; adoption surged **32% → 84%**, 70%+ of committed code now AI-generated. Costs spiked to $500–$2,000 per engineer per month, exhausting the annual budget by April.

Fortune, Forbes · May 2026

UBER CTO

$1,200 in 2 hours

Burned $1,200 in a single 2-hour demo session — **$600/hour**. Agentic mode eats 5–20× more tokens than a simple prompt, with no matching output.

Forbes · May 2026

OPENCLAW

$1.3M in 30 days

A three-person team running 100 autonomous Codex instances ran up a **$1,305,088 retail API bill in 30 days** (603B tokens, 7.6M requests). Even without "Fast Mode," the bill would still top ~$300K/month.

Business Insider, The Next Web · May 2026

MICROSOFT

Licenses revoked

Revoked Claude Code across its Experiences & Devices division (Windows, M365, Teams, Outlook) as token costs ballooned — migrating those engineers to the cheaper in-house Copilot CLI.

Fortune, Tom's Hardware · May 2026

AMAZON

Leaderboard killed

Employees gamed an internal leaderboard ("KiroRank") with meaningless, repetitive tasks to boost scores. Amazon killed it May 29, replacing raw token counts with **"normalized deployments"** — actual functional output.

The Street, Business Insider · May 2026

DISNEY

16.4B tokens / week

Warned staff against "tokenmaxxing." In one week teams burned **16.4B tokens** (3.1B Claude, 13.3B Cursor); one user's agent swarms hit Claude 51,000×/day — 460,000 calls in nine days.

Business Insider, India Today · 2026

META

60 trillion tokens / month

Shut its "Claudeonomics" dashboard after finding **85,000 employees consumed 60 trillion tokens in 30 days** — about $9B if retail API list prices are applied to that raw volume.

Tom's Hardware · May 2026

GITHUB COPILOT

A week's budget in one sitting

On day one of GitHub Copilot's usage-based billing (June 1), **"a few questions" burned 14% of the whole month's quota** — the monthly allotment gone in about a week. Developers hit billing shock and cancelled.

Visual Studio Magazine, Reddit · Jun 2026

REPLIT

$50 / day / user

Power users spend $100–$1,000+/month; one reported **$50/day** on Replit Agent. Heavy teams reach six figures a year — 10–20× traditional SaaS pricing.

Replit Community Reports · 2026

OPENAI ENTERPRISE

Year's budget gone in Q1

Altman says companies are spending a full year's AI budget in the first quarter. A **$500K/yr plan running at $150K/mo** is a $1.8M pace — 3.6× over budget.

Business Insider · Jun 2026

CURSOR

$2.4M for 1,000 developers

What looked like $10/month "unlimited" AI coding turned into $100–$200/month per heavy user — **20× traditional dev-tool pricing**. A 1,000-engineer org pays $2.4M/year for one coding assistant. The meter tracks every autocomplete.

Reddit, Hacker News · 2026

SALESFORCE

$800M Agentforce ARR

Hit $800M ARR by end of FY2026 — **29,000 deals in Q4**, 2.4B "Agentic Work Units" processed. The meter didn't disappear — it just measures outcomes instead of tokens. Enterprises pay every time an agent acts.

Salesforce Q4 FY2026 Earnings · Feb 2026

Twelve companies learned the same lesson the hard way. The fix isn't less AI — it's the right AI, on the right infrastructure.

THE REFRAME

Why fly a $101M fighter jet when a $5M (or $500) drone wins the war?

Ukraine showed that lower-cost drones — from $500 FPVs to ~$5M Bayraktars — can reshape battlefield economics. But they do not replace advanced aircraft in every mission. The lesson is fit: which tool wins *your* mission at a price you can sustain, sortie after sortie.

Enterprise AI has reached the same fork. Frontier shared models are the fighter jets: extraordinary, and billed by the sortie — a token meter that runs on every query, every agent, every document. Right-sized private models are the drones: owned outright, mission-fit, and free to fly as often as you want.

THE SAME WORKLOAD · 100 BILLION TOKENS A MONTH

That number sounds like JP Morgan. It isn't. Any company that puts AI in front of its *customers* gets there fast: a mid-cap retailer—a national shoe chain, a beauty brand—with 10–20 million monthly site visitors running conversational search hits 50–100 billion tokens a month. The math: if 25–50% of visitors engage in one 10,000-token conversation (4–5 back-and-forth exchanges), that's 50–100 billion tokens monthly. Customer-facing AI can make the token meter explode.

CLOUD / SHARED AI

Pay-per-token

$4.8M–$9.6M / year

in token fees alone, at ~$4–$8 per million tokens

- The meter runs on every query, forever
- Costs scale linearly with usage — success is penalized
- You rent intelligence; you never own it

PRIVATE / CONTROLLED AI

No token cost

$0 / year in token fees

one-time infrastructure $20K–$6M, plus energy

- Flat cost at any volume — run it as hard as you like
- You pay for hardware and energy. That's it
- The model — and what it learns — is yours

Figures developed in full on pages 11–17, including break-even math and what it costs to own the hardware.

Cost is only half the case. Private / Controlled AI also carries **intelligence, security, and IP benefits** — what your model learns from your operations stays yours, not a shared vendor's. We make that argument in a companion Iterate.ai white paper, *Private AI: Why Model Isolation Matters More Than Data Privacy* — this and other companion papers are at **iterate.ai/resources/white-papers**.

So why is everyone still paying the meter? Because token prices keep falling — and that hides what's happening to the bill. Start with the paradox →

WHY THIS PAPER, WHY NOW

Here is the paradox of 2026: LLM token prices fell 98%, but enterprise AI budgets grew nearly 6×.

Organizations aren't using less AI — they're using exponentially more. Every query, every agent interaction, every document analysis runs through a token meter that never stops ticking.

This brief is written for the engineering and finance leaders who have to forecast that meter — and decide whether per-token pricing still makes sense once the volume is real.

> **Tokens are fine for experiments. They may even make sense when you need frontier-scale intelligence. But should your daily operations run on a meter?**

HOW TO READ IT

The argument is about *scale*, not unit price. A token is cheap. A billion tokens a month is not. The ladder on page 12 and the side-by-side on page 14 are where the per-token model stops working — and where Private AI takes over.

WRITTEN FOR

CFOs, CTOs, platform and infrastructure leaders, and the architects deciding where enterprise inference should run.

PUBLISHED BY

Iterate.ai — San Jose, CA & Denver, CO. Private AI infrastructure for the enterprise.

ASSUMPTIONS USED IN THIS ANALYSIS

This document models AI deployment economics using illustrative assumptions for token pricing, average tokens per task, retry rates, routing overhead, infrastructure amortization, and operating overhead. Results are scenario-based, not audited costs, and should be interpreted as directional rather than definitive.

CONTENTS

What's inside.

WHEN THE METER NEVER STOPS · 12 COMPANIES, ONE PATTERN 02

THE PARADOX · CHEAPER TOKENS, BIGGER BILLS 06

FIND YOUR SIZE · WHO USES 100 BILLION TOKENS? 07

PART ONE · DEATH BY A THOUSAND QUERIES 08 — 12

01 **What a Token Is** 08
The unit, the million-token yardstick, and why you're billed twice.

02 **The Anatomy of a Token Bill** 09
One support ticket, 3,550 tokens — and what it costs.

03 **Every Interaction Is a Ticket** 10
It isn't just support — and "enterprise scale" is bigger than you think.

04 **The Costs Compound** 11
One use case is cheap. Eight running at once is a budget.

05 **What a Million Tokens Costs** 12
From $4 a million to $8M a year — the meter at full volume.

PART TWO · THE VARIABLE COST GOES TO ZERO 13 — 29

· **Same Tokens, Two Machines** 13
The metered, water-cooled cluster vs. the small set you own — one picture.

06 **Private AI: No Meter, No Surprise Bills** 14
You pay for infrastructure and energy. That's it.

07 **Proof: It Already Works** 15
Two live deployments — 95% lower inference cost, 16× cheaper images.

08 **What It Costs to Own It** 16
From embedded edge devices to a $6M cluster — without GPT-4's hardware.

· **Why the Cloud Pushes Back** 17 – 19
The repatriation wave · follow the money · and why a hyperscaler's "cloud-local" appliance still isn't private AI.

11 **The $6.7 Trillion Buildout** 20
Where data-center capital goes — and why ~$4.3T is servers & storage.

· **The Strategic Case for Distributed AI** 21 – 25
Two futures for AI infrastructure · no token costs · train big, infer small · the Lifeboat runtime · edge AI — the architecture and economics of owning inference.

12 **The China Lesson** 26
Why constraint — DeepSeek, Qwen, Manus — out-engineers abundance.

· **Making the Move** 27 – 29
The control point · governing spend with AgentWatch · and the questions leaders ask.

· **Private AI in Practice** 30
The DGX Spark example — a $5,000 desk appliance that ends token fees for a team.

THE BOTTOM LINE · BUILD. RUN. GOVERN. · ABOUT 31 — 33

THE PARADOX

Tokens have never been cheaper. AI has never cost more.

GPT-4-equivalent performance now costs roughly **$0.40 per million tokens**, down from **$20** in late 2022. Yet the average enterprise AI budget grew from **$1.2M a year** in 2024 to **$7M** in 2026.

Source: AI to ROI newsletter (Ray Rike & Peter Buchanan, 2026).

Why? Because organizations aren't using less AI — they're using exponentially more. Every query, every agent interaction, every document analysis, every customer support ticket runs through a token meter that never stops ticking.

The unit price collapsed. The unit count exploded faster. **Net of both, the bill grew nearly 6×** — and the steepest part of the curve is still ahead.

Falling prices didn't save money. They removed the friction that was holding usage down — and usage is the thing the meter charges for.

−98%	$0.40/M	$7M	5.8×
TOKEN PRICE, 2022 → 2026	GPT-4-EQUIVALENT, PER MILLION TOKENS	AVG ENTERPRISE AI BUDGET, 2026	BUDGET GROWTH, 2024 → 2026

> **A token is cheap. A billion tokens a month is a line item the board will ask about.**

FIND YOUR SIZE

Who actually uses 100 billion tokens — and what happens at 600× that?

Meta employees burned **73.7 trillion tokens** in 30 days — tracked on an internal leaderboard called "Claudeonomics." *Yes, a trillion.* At Claude's API rates that's roughly **$1.1 billion a month**. *Source: Business Insider, 2026.*

Against that, 100 billion looks small — and Meta is the barometer of where this is heading. Volume scales with your *customers and employees*, not how "technical" you are. Find the profile that sounds like yours.

SPECIALTY RETAILER · HUMAN-LED

~$1–2B revenue · a few hundred stores · 10–20M monthly visits

Companies this size: **Shoe Carnival, Hanna Andersson**

~8B

tokens / month

Support automation · internal knowledge search · a conversational product-finder pilot.

SHARED APIS	PRIVATE, OWNED
$400–770K/yr	**$30–75K**/yr

GLOBAL CONSUMER BRAND · HUMAN-LED

Heritage brand · image-rich catalog · global DTC + wholesale

Companies this size: **Samsonite**

~45B

tokens / month

Product customization & image generation · marketing content at scale · multilingual support.

SHARED APIS	PRIVATE, OWNED
$2–4M/yr	**$150–400K**/yr

GLOBAL ENTERPRISE & BANKING · HUMAN-LED

Tens of thousands of employees · regulated · agents across every function

Companies this size: **MUFG, JPMorgan**

~250B

tokens / month

Org-wide agents (fraud, claims, research) · employee copilots · customer chat across millions of accounts.

SHARED APIS	PRIVATE, OWNED
$12–24M/yr	**$0.5–1.5M**/yr

These are working enterprise AI tiers — all **human-led**. Meta-class systems operate at a different order of magnitude entirely. *Private /yr figures are illustrative all-in estimates (amortized hardware + energy); shared figures at ~$4–$8 per million tokens. Named companies are size references only.*

FRONTIER AI-NATIVE · MACHINE-LED

~60–90T tokens/mo

Internal *agent ecosystems*, not chat — continuous background inference, org-wide copilots embedded in workflows, and massive prompt-chaining loops. **Human-led scale ends here; machine-led scale begins.**

~800–1,000×
the specialty retailer

~150–300×
the global bank

Whatever your size, the question is the same: at your volume, does a per-token meter still make sense? What if your token costs dropped to near-zero?

And at frontier scale, token usage stops being a cost metric — it becomes an organizational **throughput** metric.

01

START HERE · THE BASICS

First — what are you actually paying for?

Before the numbers get big, the unit is small. Every charge on an AI bill traces back to one thing: the token. Here is what it is, how many of them a real workload burns, and why you pay twice for each one.

WHAT IS A TOKEN?

A token is a chunk of text the model reads or writes — roughly a short word or a word-piece. On average, a token is about three-quarters of a word.

A short word like “cat”	**1 token**
A longer word like “understand”	**~2 tokens**
A phrase like “customer support”	**2 tokens**

Shared-model providers charge for every token the model **reads** — your question plus all the context — and every token it **writes**. The meter counts both directions.

HOW MUCH IS 1 MILLION TOKENS?

About **750,000 words** — roughly:

- 5–7 average-length novels
- The entire *Harry Potter* series (all 7 books)
- 1.5× the *Lord of the Rings* trilogy
- 1.3 copies of *War and Peace*

At enterprise scale you process millions of tokens a *day* — every ticket, every document, every agent query adds to the meter.

> Now scale that up: Meta burned **73.7 trillion tokens in 30 days** — about 55 trillion words, on the order of **12,000 English Wikipedias**, or the printed Library of Congress many times over. In a single month.

WHY THE METER CHARGES TWICE

AI pricing has two rates. Output costs **5× more** than input — generating text takes more computation than reading it.

INPUT · YOU SEND	OUTPUT · AI WRITES
$3.00/M	**$15.00**/M

In a typical support ticket the response is only **11% of the tokens** — but **39% of the cost**. The more the AI writes, the faster the meter spins.

Rates shown for Claude 4 Sonnet, 2026 published pricing.

02 THE ANATOMY OF A TOKEN BILL

One support ticket. 3,550 tokens.

Take enterprise customer-support automation — a single AI-handled ticket. Most of the cost isn't the customer's question. It's the context you feed the model to answer it well.

USAGE-ONLY EXAMPLE

TOKENS CONSUMED BY ONE TICKET IN AN ENTERPRISE

500
SYSTEM PROMPT — AGENT PERSONA & WORKFLOW LOGIC

2,500
RETRIEVED DOCUMENTS FROM THE KNOWLEDGE BASE

150
THE CUSTOMER'S SUPPORT MESSAGE

400
THE AI-GENERATED RESPONSE

WHAT ONE TICKET LOOKS LIKE

CUSTOMER WRITES

"I ordered a laptop 5 days ago (Order #47382) but tracking shows it's stuck in Memphis. Can you help?"

THE AI READS

The message (150) + system instructions (500) + retrieved order, shipping policy & carrier data (2,500).

THE AI WRITES

A 400-token reply: delay explained, FedEx contacted, a $15 credit applied, a tracking update promised.

3,550 tokens **$0.0155 / ticket**

SAME TICKET, SCALED BY VOLUME

MONTHLY TICKETS	PER MONTH	PER YEAR
10,000 Small deployment	**$154.50**	**$1,854**
100,000 Mid-sized enterprise	**$1,545**	**$18,540**
1,000,000 Large enterprise	**$15,450**	**$185,400**

Important: this is a **model-fee estimate**, not the full system cost — hardware, orchestration, monitoring, and labor are accounted for separately in the operating-cost breakdown.

One use case, one model. $185,400 a year to answer support tickets — before any other agent runs.

03 IT ISN'T JUST SUPPORT

Every interaction is a "ticket."

When we say "support ticket," most readers picture customer service. In practice a ticket is *any* AI interaction — and enterprises run far more of them than they realize. The same token math applies to all of it.

THE SAME METER RUNS ON ALL OF IT

Document analysis
contracts, reports, research papers

Code generation & review
pull requests, debugging, documentation

Internal knowledge search
HR policies, IT procedures, onboarding

Data extraction
invoice processing, form parsing, records

Sales intelligence
lead scoring, account & competitive research

Compliance monitoring
policy checks, audit trails, risk assessments

Strategic research
market analysis, trend & executive briefings

These are usage-based estimates — they don't yet include infrastructure amortization or enterprise overhead. *Add IT help-desk, HR and sales queries to 10,000 support tickets and a mid-sized firm easily clears 100,000+ interactions a month.*

WHAT WE MEAN BY "ENTERPRISE SCALE"

ORGANIZATION	TYPICAL PROFILE	MONTHLY VOLUME	ANNUAL TOKEN COST
Small	Startup, single-department pilot, or SMB (under 500 employees)	10,000 tickets ~35.5M tokens/month	**$1.9K–$180K**
Mid-sized	500–2,000 employees, multiple departments using AI	100,000 tickets ~355M tokens/month	**$18.5K–$1.8M**
Large	2,000+ employees, Fortune 1000, org-wide AI deployment	1,000,000 tickets ~3.55B tokens/month	**$185K–$18M**

"Tickets" = any AI interaction. Upper bounds reflect multi-step agentic loop overhead (5–10× token multipliers per ticket) combined with premium model routing. At mid-sized and large volumes, token costs run $180K–$18M a year — where Private AI's fixed-cost model pulls dramatically ahead.

$4.8M–$9.6M
PER YEAR · ONE FEATURE

Now make it customer-facing: conversational commerce.

A major beauty retailer with a heavily trafficked e-commerce site — ~20 million monthly shoppers — adds an LLM to its search box. Search becomes conversation: multi-turn, with product-catalog context, ~10,000 tokens per session. If just half of those shoppers have one conversation a month, that's **100 billion tokens** — **$400K–$800K a month** in token fees for a single customer-facing feature. And the better it works, the more they use it.

04 THE COSTS COMPOUND

And that was just one use case.

The support agent is one meter. Now run the others. Each department spins up its own agents; each agent reads context, reasons, and writes — and every span of that runs through the same per-token charge.

EIGHT METERS RUNNING AT ONCE

Customer support agents	**Sales intelligence tools**	**Document analysis pipelines**	**Code generation & review**
Strategic research assistants	**Internal knowledge search**	**Compliance monitoring**	**Data extraction workflows**

The token costs compound fast. Organizations running multiple AI agents across departments are burning through **millions of dollars annually** on token fees alone — and the meter never stops.

There is no efficiency lever inside this model. You can prune prompts and cache context at the margins, but the structure is fixed: **more value from AI means more tokens, and more tokens means a larger bill**, forever.

THE STRUCTURAL PROBLEM

Per-token pricing ties your cost *directly* to your success with AI. The more useful it becomes, the more you adopt it — and the faster the meter runs.

Usage is the one variable you actually want to grow. It's also the one the meter punishes.

> **Every query, every agent, every document — all of it metered, none of it slowing down.**

FINANCE

Scrambling to explain why the AI bill doubled quarter-over-quarter — again.

ENGINEERING

Throttling usage to stay in budget: "we can't turn that feature on until next quarter."

THE CFO

"If AI makes us more efficient, why is it the fastest-growing line on the P&L?"

The meter doesn't just charge you for success. It punishes you for it.

05 THE COST, AT FULL VOLUME

So what does a million tokens actually cost?

A token is fractions of a cent. A *million* of them is the unit your bill is actually written in — and at enterprise volume, the meter runs into the millions.

COST OF 1 MILLION TOKENS

INPUT · PROCESSING	OUTPUT · GENERATING
$3.00/M	**$15.00**/M

MIXED · TYPICAL ENTERPRISE
A blend of input-heavy and output-heavy work — **~$4–$8**/M

Claude 4 Sonnet, 2026 published pricing. The "mixed" rate already includes multi-turn agent context and the output-token premium — the real-world blend, not the input-only sticker price.

AT ENTERPRISE VOLUME (MIXED)

TOKENS PER MONTH	ANNUAL TOKEN COST
1 billion Large enterprise	**$48K–$96K**
10 billion Org-wide deployment	**$480K–$960K**
100 billion Agent-driven scale	**$4.8M–$9.6M**

Token fees alone — before infrastructure, tooling, or any second use case.

BUT TOKEN FEES ARE ONLY ONE LINE — THE FULL COST PICTURE

This section separates usage-based model fees from the non-usage costs of running the system, so the break-even math reflects full business-case economics.

CATEGORY	INCLUDED ITEMS
Model fees	Input tokens, output tokens, context, routing premiums
Operating costs	Hardware amortization, energy, storage, monitoring, orchestration, admin labor
One-time costs	Integration, evaluation, setup, migration
Total cost of ownership	**Model fees + operating costs + one-time costs**

```
Annual Model Fees = annual interactions × blended token cost
Annual Operating Costs = infrastructure + labor + overhead
Annual TCO = Model Fees + Operating Costs
```

THE NUMBER THAT LANDS

A single enterprise running **100 billion tokens a month** pays **$4.8M–$9.6M a year** in token fees alone. The meter never reads zero — it only climbs.

$0/token
WITH PRIVATE AI
Infrastructure + energy only

THE SAME TOKENS, TWO MACHINES

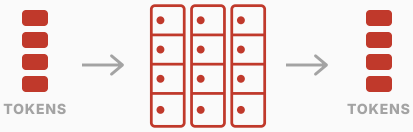
The tokens don't change. The machine that runs them does.

Every token costs power to process. What changes from one deployment to the next is **how much power, whether it needs water to stay cool** — and whether someone bills you for each one on the way through.

This scenario analysis shows how the economics change when usage volume, routing, or overhead assumptions move up or down.

SHARED CLOUD

Large GPU cluster · metered



$ · $ · $
CHARGED PER TOKEN

POWER	Very high — a rack can draw as much as ~80 homes
COOLING	Water-cooled — millions of litres a year
PER TOKEN	Metered — you pay on every pass

PRIVATE / OWNED

Small GPU set · no meter

$0
NO METER, EVER

POWER	Low — runs on hardware you already power
COOLING	Air-cooled — no water needed
PER TOKEN	$0 — you own the machine, not a meter

Same tokens in, same answers out. One machine bills you for every one and burns water to stay cool; the other runs them on hardware you own and never sends a bill. **That difference is the whole paper.**

06 PRIVATE AI

The token pricing model disappears entirely.

When you run AI on private infrastructure with Iterate.ai, there are no per-token charges, no usage meters, no surprise bills. The variable cost structure simply vanishes. You pay for two fixed things instead.

YOU PAY FOR — 01

Infrastructure

GPUs and servers you own, lease, or run in a private cloud — a fixed allocation, not a per-call charge.

YOU PAY FOR — 02

Energy

Electricity to run inference — low on efficient chips: **$1,800–$3,000 per multi-GPU node a year** at $0.12–$0.14/kWh. Nothing else scales with usage.

YOU ALSO GET — 03

2× more efficient · Lifeboat

Our inference engine delivers **2× more efficient compute** than compiled, world-class open-source benchmarks like SGLang and vLLM — each GPU does more work, cutting cost-per-query further.

THE SAME 1,000,000 SUPPORT TICKETS A MONTH

SHARED API · CLAUDE 4 SONNET

Pay per token

$185,400/yr

$15,450 / month in token costs

- × Costs scale linearly with usage
- × Bills rise every time AI gets more useful
- × Shared base model — it can learn for competitors

PRIVATE AI · ITERATE.AI

Pay for infrastructure

$0/yr in tokens

Infrastructure + energy only — fixed

- Costs stay flat regardless of volume
- Predictable budget, no surprise bills
- Model isolation — queries train only your model

Performance is a cost multiplier. If Lifeboat runs **3× faster**, you need **a third of the GPUs** for the same load — so even the fixed cost falls. And whether you own the hardware or run in a shared-tenant private cloud (e.g. Equinix Metal), you pay for **allocation, not tokens** — fixed and predictable no matter the volume. At scale, private AI isn't just cheaper — it's **orders of magnitude cheaper**.

07 PROOF, NOT THEORY

This isn’t a model. It’s in production.

The economics above aren’t a projection. Two live deployments — on the *same* Google Cloud the meter runs on — show what happens when an optimized runtime replaces brute-force token spend: inference cost collapses by an order of magnitude.

A NATIONAL RETAILER

LLM-powered conversational search

BRUTE FORCE	RUNTIME-OPTIMIZED
$5–16M	**~$0.5M**
added cost · ~1,000 GCP pods	added cost · 100+ GCP pods

95% lower

inference cost — with **47× less memory** and 16× more efficiency. What once took 10 servers now runs on 1.

A PREMIUM OUTDOOR-GEAR BRAND

AI image generation for product customization

GOOGLE VERTEX AI	RUNTIME-OPTIMIZED
80¢	**5¢**
per image · 12s / 4 images	per image · 2s / 4 images

16× cheaper

and **6× faster** than Vertex AI — so shoppers can generate far more custom images without the bill following.

BACK TO THE METER

Every one of these wins is a **token-cost win**. A conversational-search session burns roughly 10,000 tokens; an image prompt, thousands more. On shared APIs you pay for each of those tokens, on every request, forever — which is how that retailer’s search ran up **$5–16M**. Runtime optimization collapses the tokens-per-task *and* the hardware they run on, so the same workload bills **~90% less** — and on owned infrastructure, the per-token charge goes to zero entirely.

Customer-facing AI is exactly where the meter runs hottest — and exactly where optimization pays back most. Same workload, same cloud, one-tenth the inference cost.

08 THE PRICE OF OWNERSHIP

What it costs to own it.

The token meter has no ceiling. Owned infrastructure does — and the entry point is far lower than most assume. Here is the real range, from a chip in a handheld device to a mission-critical cluster.

WHAT PRIVATE AI INFRASTRUCTURE ACTUALLY COSTS

DEPLOYMENT MODEL	ENTRY COST	WHAT YOU GET
Edge devices	**Embedded**	Runs on Qualcomm chips inside handheld devices — no power, cooling, or connectivity required.
Standard servers	**$3K–$5K**	A single Nvidia RTX Pro GPU — no water cooling, no specialized datacenter.
Small cluster	**$20K–$100K**	5B–30B-parameter custom models on a small GPU cluster.
Enterprise scale	**$100K–$6M**	Multi-GPU deployment for high-volume, mission-critical workloads.

Compare to: **$185,400 a year** in perpetual token costs for 1M support tickets — with no hardware, no control, and no ceiling.

"WHY NOT JUST RUN GPT-4 YOURSELF?"

You can't — unless you're the US Government or some other gargantuan entity. A trillion-parameter model like GPT-4 needs **thousands of water-cooled GPUs** and **$100M+** in infrastructure.

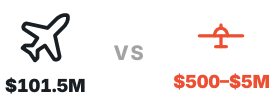
Trillion-parameter frontier models hit diminishing returns on standard corporate tasks — **8B–35B custom models are the enterprise sweet spot.** Iterate.ai's custom models run at **5B–30B parameters** — deployable on a single GPU or small cluster — and match or beat frontier models on domain-specific work.

THE POWER REALITY · ONE SERVER RACK

DRAWS THE POWER OF...

1999 — **1 dishwasher**

2026 — **80 households** — and GPT-4-scale models need **thousands of racks**, running continuously.



Why fly a $101.5M F-35 when a $5M (or $500) drone wins the battle?

Ukraine showed that lower-cost drones — from $500 FPVs to ~$5M Bayraktars — can reshape battlefield economics, without replacing advanced aircraft in every mission. Same logic here: a **$100M+ trillion-parameter model** is the fighter jet — a right-sized **$20K–$6M custom model** is the drone that wins most day-to-day missions. (F-35 Lot 18, Air & Space Forces Magazine, 2025; TB2 export cost, Atlantic Council.)

Private AI isn't just cheaper at your scale — it's part of a broader industry shift. Enterprises across every sector are asking the same question: **which workloads belong in public cloud, and which belong on infrastructure they control?** AI inference is simply the newest, fastest-growing workload forcing that decision.

09 THE CLOUD REPATRIATION WAVE

Cloud-first is becoming cloud-selective.

EQUINIX
PRIVATE-CLOUD & COLOCATION

The first wave of cloud adoption moved workloads quickly into public infrastructure. **The next wave is more surgical.**

Enterprises are deciding which workloads belong in public cloud, which belong in private cloud, and which belong closer to the data. **Broadcom's 2026 Private Cloud Outlook reports that 83% of enterprises are considering moving some workloads from public cloud to private infrastructure.** This is not a mass exit from cloud — it is workload placement becoming more disciplined.

THE SAME PATTERN, EVERY TIME

1 Start small on shared APIs
At low volume, variable cost looks attractive — no upfront investment, fast to start.

2 Scale up
Adoption spreads across teams. The token bills explode.

3 Realize the math doesn't work
The variable cost that was fine at low volume becomes unsustainable at scale.

4 Move to private infrastructure
Cost becomes fixed and predictable. The meter is gone.

THE 2026 DATA

83% of CIOs plan to move some workloads from public cloud to private infrastructure.
Broadcom 2026 Private Cloud Outlook

50% of enterprises have already repatriated some workloads — a 15-point jump in one year.
Broadcom 2026 Private Cloud Outlook

39% cite cost predictability as a top driver — now the second-biggest reason for repatriation.
Across surveyed organizations

43% of repatriating orgs are moving AI training, LLMs, and inference — a category that didn't exist in 2025.
Newly tracked in 2026

The logic mirrors the token meter: **variable costs that looked attractive at low volume become harder to justify at scale.** AI inference is now the newest, fastest-growing workload forcing enterprises to ask where each workload truly belongs.

10 FOLLOW THE MONEY

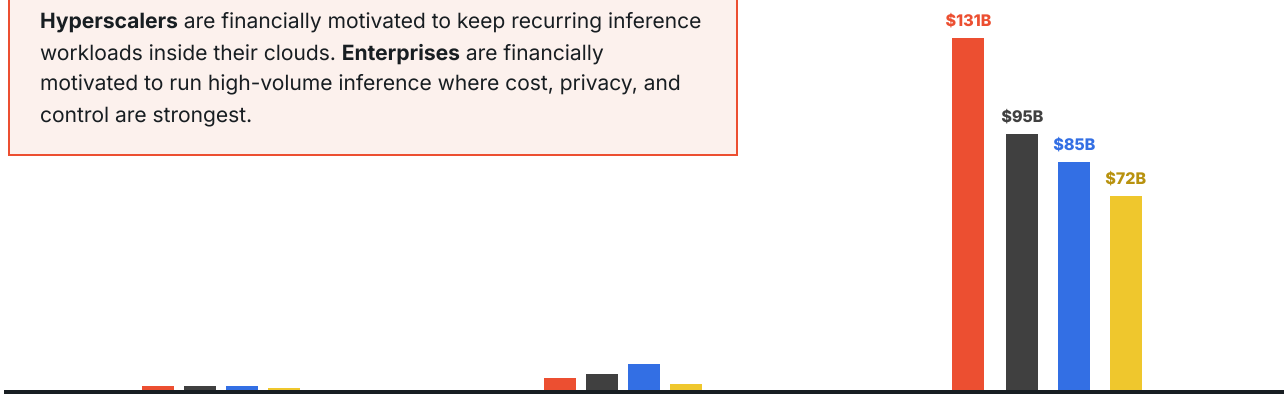
If private inference is cheaper, why doesn't the cloud say so?

Their incentives are different. The cause is business-model design — it's anchored in the cloud. Hyperscalers have committed historic capital to centralized AI infrastructure, and that capital earns a return when workloads keep running, scaling, and metering inside their clouds. **Private inference changes the equation** — it shifts high-volume AI from recurring cloud usage to enterprise-controlled infrastructure.

ANNUAL CAPEX · AWS, MICROSOFT, ALPHABET, META · 2005 → 2025

AWS Microsoft Alphabet Meta

Hyperscalers are financially motivated to keep recurring inference workloads inside their clouds. **Enterprises** are financially motivated to run high-volume inference where cost, privacy, and control are strongest.



From roughly **$2 billion to $410 billion in twenty years** — a ~200× run-up — and the five biggest plan to add **~$2 trillion** in AI assets to their balance sheets by 2030. That capital only earns a return if your workloads keep metering in their clouds. (Houlihan Lokey, 2025; FT/company filings, 2026; Goldman Sachs, 2026. Figures approximate.)

TWO DIFFERENT MACHINES FOR TWO DIFFERENT JOBS

Cloud / Shared AI → built for training

Training is massive, centralized, and intermittent — huge bursts of compute a handful of times. Renting an enormous shared cluster for those bursts is exactly right.

Private / Controlled AI → built for inference

Inference is the opposite — continuous, distributed, latency- and cost-sensitive, running forever. Metering a workload that never stops is exactly where owning the hardware wins.

The cloud was built for the workload that happens a few times. Private AI was built for the one that runs forever.

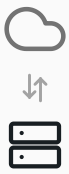
READ THE FINE PRINT

Cloud-local is not the same as private AI.

The hyperscalers have heard the private-AI demand — so they're offering "local" versions of their cloud. **But local hardware is not the same as private AI.**

If the stack is still controlled, synchronized, governed, and priced by the cloud provider — and designed as a bridge back to cloud — the enterprise hasn't escaped the meter. It has only moved the meter closer to home. True private AI means the enterprise controls the **infrastructure, the runtime, the models, the data boundary, and the long-term economics.**

CLOUD-LOCAL APPLIANCE



A local extension of the cloud — synced, governed, and priced by the provider.

PROTECTS **cloud revenue**

PRIVATE AI INFRASTRUCTURE

FlexPod · NetApp AI Pod · Equinix · IBM Cloud Private

An owned inference layer — the enterprise controls runtime, models, data, and economics.

PROTECTS **enterprise economics**

If the goal is to bring workloads back to the cloud, it isn't private AI. It's cloud retention.

11 WHERE THE CAPITAL IS GOING

The $6.7 trillion buildout.

That's roughly **1% of global GDP, every year**. Nearly two-thirds of it — **about $4.3 trillion** — is servers and storage.

CUMULATIVE CAPEX BY VALUE-CHAIN SEGMENT · 2025–30 · $ TRILLION **$6.7T total**

BUILDERS		ENERGIZERS		DESIGNERS & MANUFACTURERS	
Labor	$0.6T	Electrical & mechanical	$0.8T	**Servers**	**$3.5T**
Shell & site	$0.3T	Power generation	$0.4T	**Storage**	**$0.8T**
Land acquisition	$0.1T	Network infrastructure	$0.1T		
Total	**$1.0T**	**Total**	**$1.3T**	**Total**	**$4.4T**

The question is not whether the infrastructure gets built. The question is who pays for it.

THE HIDDEN ASSUMPTION

Training created the arms race.

Frontier models demanded huge GPU clusters, power contracts, cooling, and new data-center capacity.

Inference is supposed to pay for it.

Every prompt, search, agent step, document, and image becomes recurring metered usage.

But inference may not stay centralized.

Small models, private pods, edge devices, and optimized runtimes can move high-volume work off the meter.

The world is spending trillions to build AI infrastructure — and that bet assumes inference revenue flows back through centralized clouds. But if enterprises run high-volume inference privately, locally, and cheaply, the economics shift. The winners may not be the owners of the biggest data centers.

The winners may be the owners of the runtime layer.

Source: McKinsey Data Center CAPEX TAM & Demand model; Goldman Sachs; S&P Capital IQ. Cumulative 2025–30. Global GDP $106T, 2023.

TWO FUTURES FOR AI INFRASTRUCTURE

The AI buildout has two possible futures.

Training built the AI cloud. **Inference will decide who gets paid.**

What if the day-to-day token meter moves off cloud?

SCENARIO 1

The Tokendom

Most inference flows through large public-cloud models. Every prompt, search, document, image, and agent step pays the token toll — monetized through metered usage, subscriptions, and platform lock-in. This is the world the current capex curve appears to assume.

DEFAULT VENUE	**Public cloud**
COST MODEL	**Variable token meter**
INFRASTRUCTURE	**Massive centralized clusters**
DATA MOVEMENT	**Data moves to the AI**
WINNER	**The owner of the token meter**

SCENARIO 2

Distributed Inference

Training stays centralized, but routine enterprise inference moves closer to the data — edge devices, private clouds, colocation, FlexPods, AI Pods, storage-adjacent systems, and optimized runtimes. Enterprises stop renting intelligence for every repetitive task, and own more of the inference layer.

DEFAULT VENUE	**Private, local, or colo**
COST MODEL	**Fixed-cost infrastructure**
INFRASTRUCTURE	**Small GPUs, AI pods, edge chips**
DATA MOVEMENT	**AI moves to the data**
WINNER	**The owner of the runtime layer**

IF INFERENCE MOVES PRIVATE, THE WINNERS CHANGE

Hyperscalers want recurring cloud usage. Hardware and colocation providers are more neutral — **Dell, Cisco, NetApp, Equinix, Flexential, and neo-clouds** win whether compute runs in a mega-cloud, private cloud, colo, or enterprise-controlled environment. They sell the picks and shovels. **The strategic battleground is the runtime layer.**

OLD **Hyperscaler cloud** → **token meter** → **recurring revenue**

NEW **Hardware + colo + private cloud + optimized runtime** → **owned inference economics**

NO TOKEN COSTS

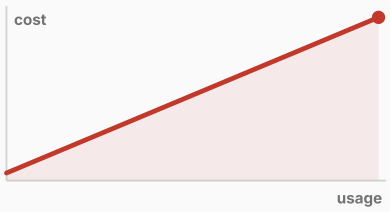
No token costs.

Or, more precisely: **no recurring public-cloud token toll** — and near-zero marginal cost once the relatively inexpensive private infrastructure is deployed.

Public-cloud AI turns every unit of usage into a meter. Private inference turns high-volume usage into a **fixed-cost asset**. At low volume the difference looks small. At enterprise scale, it changes the entire AI adoption curve.

PUBLIC-CLOUD AI

Cost climbs with every query

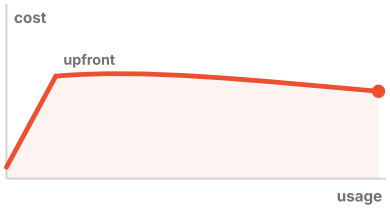


tokens · API calls · agents · documents · images · searches

Every unit of usage becomes a toll.

PRIVATE INFERENCE

Pay once, then run it flat



edge devices · small models · RTX PRO · FlexPod · NetApp AI Pod Mini · Equinix

High-volume usage becomes a fixed-cost asset.

UNIT ECONOMICS	PUBLIC-CLOUD PATH	PRIVATE ITERATE PATH	REDUCTION
One AI image generated	**$0.80** 12 seconds	**$0.05** 2 seconds	**16×** 6× faster

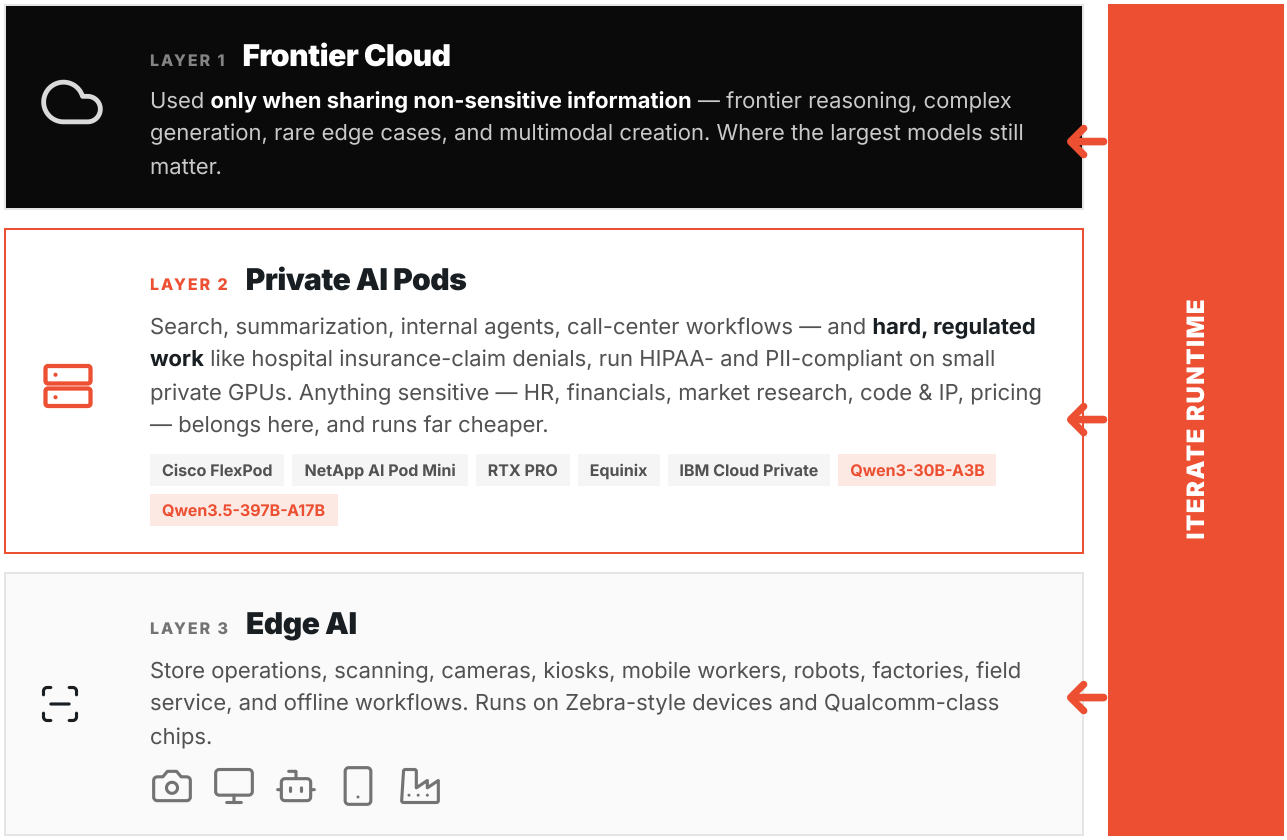
The AI boom assumes usage becomes revenue. But if enterprises own more of the inference layer, less of that usage becomes public-cloud revenue. The infrastructure is still needed — **but the profit pool moves.**

The capex boom assumes the token meter stays in the cloud. Private inference breaks that assumption.

TRAIN BIG. INFER SMALL. ESCALATE SELECTIVELY.

Train big. Infer small. Escalate selectively.

The winning architecture may not be one giant brain in the cloud. It may be **millions of smaller brains, close to the work** — calling the giant brain only when needed.



Iterate's runtime is called **Lifeboat**. As of June 2026 it benchmarks **~2× more efficient than open-source runtimes like vLLM and SGLang**, turning AI from a token-rental into an owned operating layer.

Simple task → small model	**Sensitive or regulated** → private AI	**Non-sensitive, shareable** → frontier cloud	**High-volume task** → optimized runtime

The future is not cloud versus private. It is the right model, on the right infrastructure, at the right cost, for the right task. That is how enterprises escape the token toll.

Bring the AI to the data. Don't always move the data to the AI.

THE RUNTIME LAYER

Open-source runtimes are a starting point, not the finish line.

Open-source runtimes like vLLM and SGLang are a good place to begin. **But at enterprise scale, the gains you can squeeze out change the plan entirely.**

Iterate's **Lifeboat** delivers meaningful runtime and inference improvements.

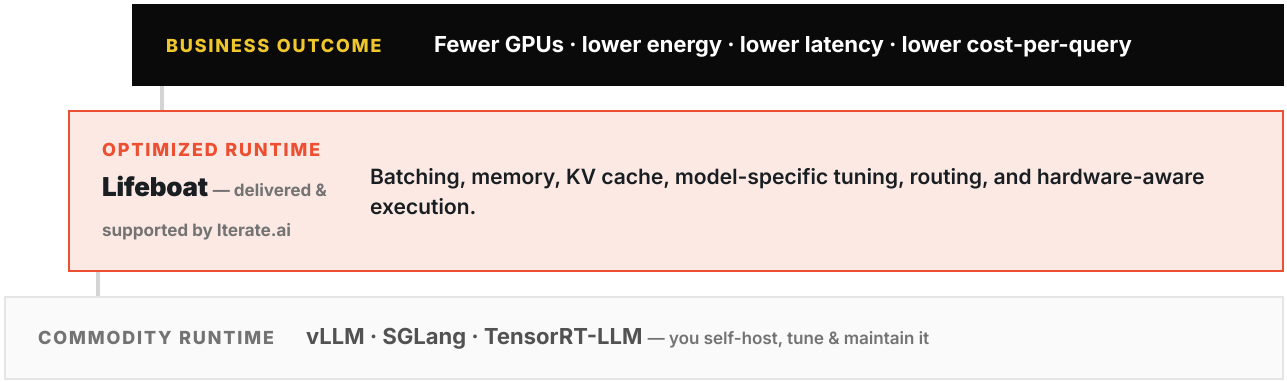
2×

more inference throughput per GPU than vLLM & SGLang — each chip serves twice the requests.

½

the hardware, energy & cost for the same workload — the flip side of running 2× faster.

Lifeboat is **not a wrapper around commodity inference software** — it is an optimization layer that makes smaller infrastructure do more work, **delivered, maintained, and supported by Iterate.ai** (you don't self-host and babysit open source). It delivers **2× more efficient compute** than compiled, world-class benchmarks like SGLang and vLLM — **halving the hardware, energy, and cost-per-query.** It's why infrastructure leaders roll Lifeboat out to bare-metal private-cloud customers, and enterprise networks rely on Iterate for high-throughput, zero-token inference.



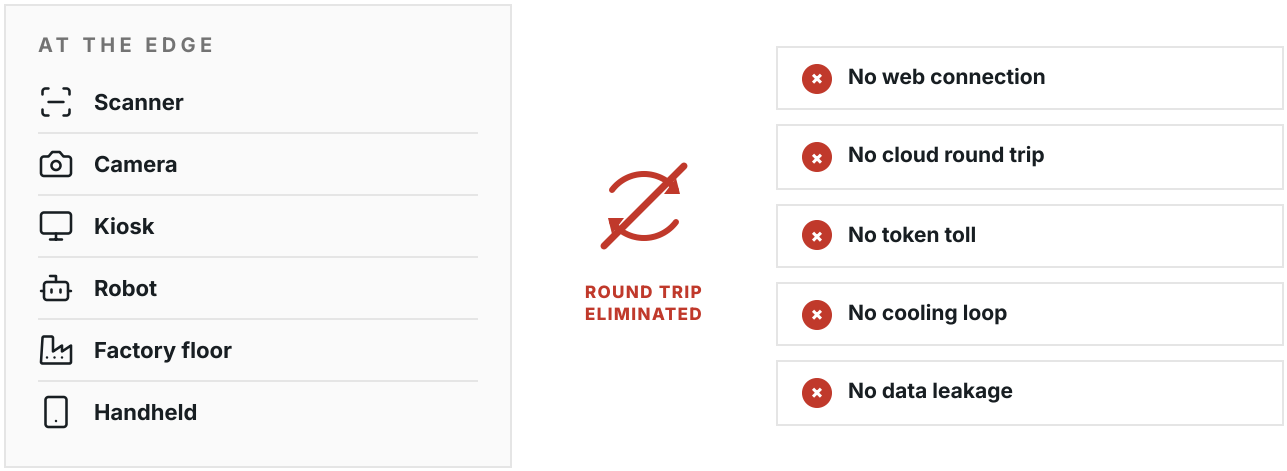
A 20% gain is nice. Lifeboat's 2× — half the hardware, half the cost — rewrites the business case, and that is the work that happens above the commodity runtime.

EDGE AI

Tiny devices matter more than people think.

Edge AI is not trying to replace frontier models. **It is replacing unnecessary round trips.**

A Zebra-style handheld computer with a **Qualcomm QCS6490-class chip** can make local AI decisions close to the worker, camera, scanner, kiosk, robot, handheld, or factory floor — with nothing leaving the environment.



Local decision = **lower cost** + **lower latency** + **better privacy**

That is not just a technical shift. **It is an economic shift.**

The smallest AI infrastructure may remove the largest amount of waste.

12

CONSTRAINTS BREED INVENTION

The most efficient AI is built by those who can’t buy their way out.

Cut off from the best GPUs by American export controls, China’s leading labs couldn’t win by spending more — so they won by spending *smarter*, pushing the envelope on both training and inference efficiency. Scarcity, not abundance, is producing the cutting edge.

DOING MORE WITH LESS

CHINA’S LABS

DeepSeek

Trained frontier-class models at a fraction of the usual cost — mixture-of-experts and FP8 efficiency on restricted hardware. Proof capability isn’t purely a function of GPU budget.

Alibaba · Qwen

Open-weight models tuned to run efficiently and deploy widely — now among the most-used foundations on earth, because they’re light enough to own and run yourself.

Moonshot · Kimi

Long-context, open-weight models that wring frontier-level reasoning out of lean compute — efficiency engineering, not raw hardware, doing the heavy lifting.

Manus

An agentic system pushing autonomous, multi-step workflows — squeezing real agent performance out of constrained compute rather than waiting for unlimited capacity.

ITERATE.AI

WHY ITERATE ACTS LIKE A CHINESE COMPANY

Iterate.ai isn’t a hyperscaler with a $2-trillion balance sheet to defend. It is **cash-constrained** — and that constraint is the engine. The same pressure that forces China’s labs to invent forced Iterate to make LLMs and RAG run on CPUs, edge devices, and small chipsets — **100× smaller, no water cooling, on hardware enterprises already own** — with many of its runtime and memory-optimization (KV-cache) inventions **patented or patent-pending**. Abundance builds bigger data centers; scarcity builds better engineering.

The hyperscalers are racing to spend the most. The future of private AI belongs to whoever needs the least.

13 THE CONTROL POINT

Private AI is not just a cost decision. It is a control decision.

At scale, the question is not only *when* the infrastructure pays for itself. The bigger question is what you stop exposing, renting, and metering.

WHAT YOU STOP

- You stop **exposing sensitive data** to shared AI systems.
- You stop **renting every unit of intelligence** by the token.
- You stop **depending on public APIs** — availability, rate limits, and pricing changes.
- You stop **routing high-volume workflows** through infrastructure you don't control.

WHAT YOU START

You start owning the AI layer that runs your daily operations.

The savings can be dramatic. **But the strategic value is bigger:**

Safer data | **Stronger privacy** | **Lower attack surface** | **Predictable cost** | **Model isolation** | **Control over where inference runs**

Private AI is the point where cost control becomes strategic control.

14

GOVERNANCE FOR A HYBRID WORLD

AgentWatch
By Iterate.ai

You can't cap a bill — or a leak — you can't see.

Few enterprises will abandon shared models entirely — the future is **hybrid**. Either way, governance is non-negotiable: your board and security team need to see what employees do with AI and to monitor a fast-growing fleet of agents — your new digital employees.

AgentWatch, Iterate.ai's governance gateway, routes every app and agent through one OpenAI-compatible endpoint — so usage is visible, attributable, capped, and shadow AI is caught. It matters: **nearly 40% of employee AI interactions (prompts, pastes, and uploads) involve sensitive corporate data.** *Source: Cyberhaven, 2025.*

COST GOVERNANCE, NOT JUST OBSERVABILITY

Token-level tracking
every prompt counted across OpenAI, Anthropic, Google, and private models

Per-team & per-person budgets & alerts
hard limits by org, team, or user — before the overage, not after

Chargeback & attribution
100% cost attribution — know exactly which workload spends what

Multi-provider routing
send each task to the cheapest capable model — including your private one

Shadow-AI discovery
surface ungoverned ChatGPT/Claude use, even on remote VPN-enabled devices — uncounted spend and data risk

WHAT GOVERNANCE BUYS YOU

100%
cost-attribution visibility — no more unallocated AI line item

1 endpoint
every app and agent on one policy plane — OpenAI-compatible, minimal code change

7 yr
audit-log retention — who used which model, when, and under which policy

AGENTS AS EMPLOYEES

AgentWatch treats every agent like a human employee on the payroll — each gets an identity, and **every query it makes is logged, attributed, and policy-checked.** When agents spin up and call other agents, the whole chain stays on the record.

SHADOW AI AT THE GATEWAY

VPN-linked gateway discovery puts every managed device behind one policy plane — activating the VPN requires AgentWatch. So when someone tries to paste a contract or source code into public ChatGPT, AgentWatch automatically identifies and **blocks the upload**. That's how Shadow AI gets caught before data leaves.

The endgame is private AI with no meter. The bridge is governing the meter you still have — budgets, attribution, and routing on one gateway, so spend stops being a surprise.

15 OBJECTIONS, ANSWERED

The questions leaders ask.

Two questions come up in nearly every evaluation — the lift of moving, and the fear of falling behind the frontier. Here are the straight answers.

→ What about migration?

Worried about the lift to move off API calls? Here is what it actually looks like with Iterate.ai.

API-compatible endpoints

Your existing code doesn't change. Swap the base URL, keep everything else.

Deployment in days, not quarters

Standard servers run on Nvidia RTX Pro hardware — no specialized cooling or datacenter.

We handle the infrastructure

Iterate.ai builds, runs, and governs the deployment. You don't need a new team.

The transition isn't the risk. Staying on the meter is.

→ What if I need the latest frontier model?

Private AI doesn't lock you into outdated models — here is the reality.

Small models, frontier results

Custom 5B–30B models match or beat GPT-4 (1.7T) on domain-specific tasks — trained on your data, not generic internet text.

The weights are yours

Update, fine-tune, or swap models as new architectures emerge.

Frontier is a moving target

Most enterprises don't need trillion-parameter models for internal workflows.

Not "Can I run GPT-5?" but "Do I need a $100M model when a $100K one wins on my use cases?"

PRIVATE AI IN PRACTICE

You can start with a $5,000 box on a desk.

Private AI doesn't require a data center. A single **NVIDIA DGX Spark** — a desktop appliance at $4,000–$6,500 — runs models up to 200B parameters locally and serves 5–6 concurrent users — which could mean up to 10–25 attorneys, or ~50 employees who use it less frequently, in total. For a small law firm, finance team, or department, that's enough for document Q&A, contract review, report generation, and knowledge search — without sending a single document to the cloud.

WHAT IT IS

Chip	GB10 Blackwell
Memory	128GB
Storage	4TB
Network	200GbE
Footprint	150×150×50
Power	240W
Price	**$4K–$6.5K**

WHAT IT CAN DO

70–200B
up to 70B native; up to 200B via FP4 quantization

4–5 users
concurrent, at 128K context each

9–14 tok/s
Qwen3-30B-A3B (MoE), running local

Document Q&A, contract review, deposition summaries, report generation, knowledge search.

THE ECONOMICS

ONE-TIME	ENERGY
$4K–$6.5K	**~$50–75**/mo

No token fees. No API charges. No per-user licensing.

BREAK-EVEN VS. CLOUD (5 USERS)

$3.6K over 3 yr — light use ($20/user/mo)

$36K over 3 yr — heavy use ($200/user/mo)

For under ~$6K/year, you can power up the intelligence of 50 people — privately.

WHO IT'S FOR

- **Small law firms · 10–25 attorneys**
- **Finance & compliance teams**
- **HIPAA-bound healthcare offices**
- **Department-level deployments**
- **Edge / low-connectivity branches**

ENTERPRISE SCALE · KLARNA: replaced Salesforce CRM with internally built AI data tools — saving **$2M in licensing** and retiring 1,200+ microservices. When you own the AI infrastructure, agents don't just cut costs — they eliminate entire software categories. *LinkedIn · 2026*

THE FLIP SIDE: global contact-center operating costs are projected to fall **$80B in 2026** as AI handles 60–70% of interaction volume — the highest-ROI AI category available, *when you own the infrastructure*.

This isn't a replacement for enterprise GPU infrastructure — it's proof that private AI can start small, start fast, and scale. A $5,000 appliance ends token fees for a team; a rack ends them for a company. Same principle: own the infrastructure, control the cost.

THE BOTTOM LINE

Token pricing made sense when AI was experimental.

It lowered the barrier to entry and let organizations test AI without upfront infrastructure investment. But at scale, the economics flip — what looked attractive at 10,000 queries a month becomes unsustainable at a million.

Private AI eliminates the token tax entirely. You pay for infrastructure and energy — both fixed, predictable, and dramatically cheaper than cumulative token costs at enterprise scale.

The question isn't whether you can afford private AI. It's whether you can afford to keep paying per token while every team you have leans on it harder each quarter.

THE TWO QUESTIONS

~~"Can we afford private AI?"~~

"Can we afford to keep paying per token?"

Any comparison that merges model fees with operating costs will understate the true cost of ownership. The proper business case **separates usage charges, infrastructure, and enterprise overhead** before comparing options.

COMPANION WHITE PAPER

Why Haven't We Seen Private AI Until 2026?

The four technological convergences that made private AI suddenly feasible — and dramatically more cost-effective than shared APIs.

FIND IT AT

iterate.ai/resources/white-papers

PRIVATE AI: BUILD. RUN. GOVERN.

Models run on your infrastructure. Data never leaves.

Iterate.ai builds, runs, and governs private AI infrastructure for banks, hospitals, insurance companies, retailers, big tech, and datacenters — designed for the strictest requirements, with zero external connectivity.

Data privacy

Data never leaves your environment. Nothing is sent to a third-party API to be metered, logged, or retained.

Model isolation

Your model is yours alone. Competitors don't learn from your queries; your prompts train no one else.

Hardware control

Own, lease, or run in a private cloud. Fixed capacity, fixed cost — on infrastructure you control.

BUILT FOR THE STRICTEST REQUIREMENTS

HIPAA · SOC 2 · ISO 27001 · GDPR · FEDRAMP-READY

Deployments meet government standards, with models running air-gapped from external connectivity. All three pillars on infrastructure you control.

POWERED BY LIFEBOAT

2× more efficient inference

Our proprietary inference engine delivers **2× more efficient compute** than world-class open-source benchmarks (SGLang, vLLM) — the reason private inference now runs cheaper than the API call.

SOURCES & REFERENCES

01 *Broadcom 2026 Private Cloud Outlook* — 83% of CIOs plan to move workloads to private infrastructure; 50% have already repatriated; 43% are moving AI training & inference.

02 *Salesforce Q4 FY2026 Earnings (Feb 2026)* — $800M Agentforce ARR, 29,000 deals, 2.4B Agentic Work Units.

03 *Equinix* — 60%+ of public cloud workloads expected to migrate to private cloud.

COLOPHON

Set in **Inter**. Figures reflect 2026 published pricing for Claude 4 Sonnet ($3.00 input / $15.00 output per million tokens) and are illustrative of relative scale.

ABOUT ITERATE.AI

Private AI infrastructure for the enterprise.

Iterate.ai builds, runs, and governs private AI infrastructure for banks, hospitals, insurance companies, retailers, big tech, and datacenters. Founded in Silicon Valley and Colorado in 2013 by one team that helped invent the iPhone and another that sold $1.65 billion worth of travel bags before its exit to Samsonite.

We serve enterprises that have decided the meter should stop — that their AI workloads, and the bill those workloads generate, belong on infrastructure they own at a cost they can forecast.

SOME OF ITERATE'S PRODUCT FAMILY

Lifeboat

Inference acceleration — 2× more efficient than SGLang & vLLM. The same model, at a fraction of the cost-per-token.

AgentWatch

Governance and observability of employee LLM usage and agent behavior — so token spend is seen before it is billed.

Generate

A no-code platform for building and running AI agents privately — on hardware you control, with no metered API in the loop.

RECOGNITION

01	**20 Hottest AI Software Companies** Channel Reseller News	2025 · 2026	
02	**Best Innovation in AI for Healthcare** AI Tech Awards · *Generate for Healthcare*	2026	
03	**Best AI Edge Deployment** Pinnacle Awards	2026	
04	**Best Use of AI in Healthcare** Pinnacle Awards	2026	
05	**AI 100** KM World	2023 · 2024 · 2025	
06	**Best Workplaces for Innovators** Fast Company · *AI + Robotics*	2024	
07	**Technology of the Year** InfoWorld · *AI/ML Models*	2024	
08	**Best in Business** Inc. Magazine · *AI + Data*	2023 · 2024	

SERIES

Intelligence Unshared: The AI Sovereignty

Paper: 2 of 6

June 2026

HEADQUARTERS

San Jose, CA
Denver, CO

WEBSITE

Iterate.ai

hello@iterate.ai

PUBLICATIONS

iterate.ai/resources/white-papers
iterate.ai/resources/books