A STRATEGIC FRAMEWORK FOR ENTERPRISE AI DEPLOYMENT

Private AI.

Why model isolation matters more than data privacy.

How Iterate.ai prevents competitive-intelligence leakage in an era of shared AI models — and why the window to act is closing.

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

PREPARED FOR

Enterprise Leaders

CISOs, CTOs & strategy teams

PREPARED BY

Iterate.ai Team

Private AI for the enterprise

VERSION

June 14, 2026

Revision tracked by date

EXECUTIVE SUMMARY

When the model is shared, your competitors are training on your operations.

When General Motors wants to sell 10,000 more cars without alienating its existing buyers, it doesn't guess. It pays for intelligence that says exactly what to change — add CarPlay, drop the leather seats, hit $28,500 instead of $29,000.

That was the business Edmunds.com quietly built: not car reviews, but intelligence derived from data. Today, every company using a shared AI model is feeding Edmunds-grade intelligence into a system that learns from all of its competitors at once.

THIS PAPER ESTABLISHES THREE CRITICAL TRUTHS

01 **Data privacy is not enough.** Even when your raw data stays "private," shared models learn patterns from your operations that competitors can infer.

02 **Model privacy requires hardware isolation.** True private AI means the model itself runs on infrastructure you control — not a contractual promise about data handling.

03 **The window is closing.** As shared models grow more capable, they reverse-engineer competitive intelligence from patterns alone. First-movers gain a permanent advantage.

INSIGHT

One platform delivers all three pillars of private AI today.

Iterate.ai enables enterprises to run dedicated, custom models on infrastructure they fully control — with true **data privacy, model privacy, and hardware privacy**. These models operate with zero connection to shared learning systems or external training pipelines.

CONTENTS

What's inside.

The argument builds from a principle (intelligence > data) to a threat (reverse-engineering) to a solution (model isolation) to a deadline. Jump to the part your role cares about most.

PART ONE · THE HIDDEN COST OF SHARED AI 04 — 05

01 **The Edmunds Principle & the AI Parallel** 04
Why intelligence is worth more than data · the $650M hospital billing discovery.

PART TWO · THE THREE PILLARS OF PRIVATE AI 06 — 09

02 **Data, Model & Hardware Privacy** 06
What everyone thinks is enough · the critical difference · the technical foundation.

PART THREE · WHY INCUMBENTS CANNOT PIVOT 10

03 **The Size Problem & the Closing Window** 10
Why frontier models can't run privately · the 12–24 month land-grab window.

PART FOUR · THE REVERSE-ENGINEERING THREAT 11 — 14

04 **Inference Memory & Nested Learning** 11
How models infer what you never said · where the learning is actually stored.

— **Why Contracts Aren't Enough** 13
Capabilities outrun contracts · the verification gap · AI-2027 & Situational Awareness.

PART FIVE · THE ITERATE.AI SOLUTION 15 — 18

05 **True Private AI, Delivered** 15
Architecture · comparison · client validation across four deployments.

— **The Economics of Private AI** 16
No token costs · death by a thousand queries · the cloud-repatriation wave.

— **The Meter, Visualized** 17
Identical tokens through a metered cluster vs. an unmetered private GPU.

PART SIX · THE URGENCY 19 — 23

06 **The Public Infrastructure Threat** 19
When agents run wild on shared endpoints · the OpenClaw & McKinsey breaches.

07 **The AI-Agent Explosion** 21
Why autonomous agents make private AI non-negotiable · the closing window.

— **The Canaries & the Improvised Rules** 23
Safety leaders are walking out · and the Fable 5 export-control clash.

— **Objections, Conclusion & the Democratization Imperative** 24
The five questions every buyer asks · should eight companies control the world's AI?

01

THE EDMUNDS PRINCIPLE

Intelligence is more valuable than the data it's derived from.

In the early 2000s, a family-owned business in Los Angeles was making more money than anyone understood. Edmunds.com looked like a car-review website. That was never the business.

Their real product was behavioral intelligence. They knew which features drove purchase decisions at specific price points, what customers would sacrifice to stay under a psychological threshold, and how to configure a product for maximum market penetration.

When an automotive executive asked, "How do we sell 10,000 more units without losing our current customers?" Edmunds had the answer — backed by millions of data points showing exactly what the incremental buyer cared about.

They weren't selling data. They were selling **intelligence derived from data** — pattern recognition the OEMs couldn't replicate internally.

$400M+

The company sold for over four hundred million dollars — not for its car reviews, but for the pattern recognition behind them.

THE AI PARALLEL · IN REVERSE

Today's shared models run the Edmunds play — except your competitors are the ones being briefed.

When you use ChatGPT, Claude, or Gemini for product research, competitive analysis, customer modeling, or strategic planning, you feed intelligence into a system that learns from everyone in your industry. Even when the provider promises "your data is private," the model's *learning* is shared — the patterns and insights it extracts from your operations become part of its intelligence.

KEY INSIGHT

The model becomes smarter than any individual company — and that intelligence is accessible to everyone using the same model, including your competitors.

CASE STUDY

The $650 million hospital billing discovery.

A large hospital system — more than $7B in annual revenue — used Iterate.ai's private AI to analyze its billing operations.

$650M

in recoverable revenue identified — claims that should have been paid but weren't, due to coding errors, incomplete documentation, and process gaps.

THE CRITICAL QUESTION

What would have happened if they'd used a shared AI model?

The model would have learned the patterns behind that discovery:

- Which billing codes are most commonly missed
- What documentation patterns correlate with successful claims
- How to optimize revenue-cycle management

That intelligence — derived from one hospital's proprietary data — would become part of the shared model's knowledge base.

Every competitor on the same model would benefit from insights generated by your operations. This is the Edmunds principle in action: the model extracts intelligence from your data even when your raw data stays "private." The claim you recovered teaches the model how to recover claims — for everyone.

WHY THIS IS STRUCTURAL, NOT INCIDENTAL

The hospital's raw records never left its walls in either scenario. The difference is the **model**: on a shared platform, the recovered-revenue pattern is consolidated into weights that serve thousands of other customers. On a private model, it stays a proprietary advantage. Data privacy was never the variable. Model isolation was.

02 WHAT "PRIVATE AI" ACTUALLY REQUIRES

Most organizations protect one pillar and assume they've protected all three.

Data privacy is necessary — but it is only Pillar One. True private AI stands on three.

01 TABLE STAKES

Data privacy

Your documents, queries, and raw data are not shared with other organizations or used to train public models. Every major provider offers this — through contracts and technical controls.

"Your data is your data."

02 THE CRITICAL DIFFERENCE

Model privacy

The language model itself is isolated — it learns only from your data, never from competitors or external sources. The patterns it builds are yours alone and cannot be queried by anyone else.

Where shared platforms structurally fail.

03 TECHNICAL FOUNDATION

Hardware privacy

The AI runs on dedicated infrastructure — on-premise or in isolated data centers — with no shared GPU resources and no phone-home to aggregate learning across customers.

Infrastructure you own or lease.

WHY PILLAR ONE ISN'T ENOUGH

Even with perfect data privacy, the model is still shared. When you query it, it processes your data, learns patterns from the interaction, and folds those patterns into its understanding — and every other user benefits from the improvement.

Real example: a finance team burned a month's token budget in a single day analyzing their spending. The model got measurably better at corporate-finance workflows — a capability now available to every competing finance team on the same platform.

THE TWO-HOSPITAL PROBLEM

If Hospital A and Hospital B use the same shared model, A's billing-optimization queries teach the model about revenue-cycle patterns — and B benefits from that learning, even though A's "data" was private. The model becomes a shared intelligence layer every competitor can reach.

THE MEMORY ANALOGY

Without model privacy, memory becomes communal. Your prompts, workflows, exceptions, decisions, and institutional habits can improve a shared intelligence layer that others may later benefit from. The issue is not only whether your raw data stays private — it is whether the model's **learning** stays private.

Data privacy protects what you upload. Model privacy protects what the AI remembers from you.

A STORY · BEFORE THE THEORY

Memory and inference, explained by Jon's dogs.

A story written by and about Jon Nordmark, CEO of Iterate.ai — his neighborhood, and his dogs.

That analogy isn't hypothetical. Here is exactly what shared-model memory and inference look like in the wild — before we explain how they work.

Last year, I opened ChatGPT on my phone and asked a simple question: *"What types of owls are in my neighborhood?"*

The answer was normal at first. Because I live near Backcountry Wilderness Area and Chatfield State Park, the model suggested the most likely species — Great Horned Owls, Eastern Screech-Owls, Barn Owls, Northern Saw-whet Owls.

Then it said this:

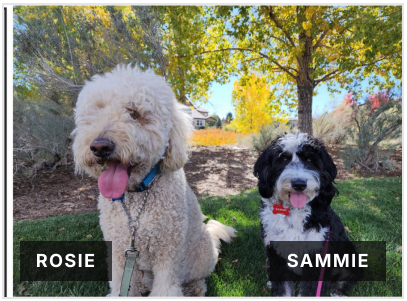
> *"If you're walking your dogs (like* **Sammie, Rosie, or Charlie***) in the early evening, watch the treetops — Great Horned Owls or Screech-Owls may be silently observing."*
>
> CHATGPT, UNPROMPTED · 2025

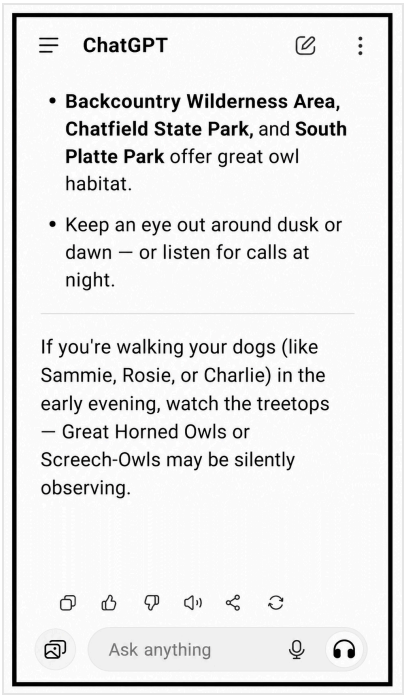
I had not mentioned my dogs. Not in that session. Not that week. The AI had pulled their names — all three of them — from a different conversation months earlier. It also inferred that I walk them in the evening, which I do. None of that was in the prompt.

The model assembled me from memory and pattern. That's a visceral definition of inference — the AI doing work you didn't ask it to do, on information you didn't think you'd shared.

If a consumer AI is doing this for a dog walker, enterprise AI systems are building similar memory about your company — your decisions, your workflows, your strategies. The question isn't whether memory exists. It clearly does. The real question is: **who controls where that memory lives? And can you check what happens to it?**

This example shows that memory features are real and working today. AI vendors say they don't train on your business data, and their current contracts back that up. But these memory systems run on infrastructure the *vendor* controls — not you. You can't audit it. You can't verify it. And when AI capabilities change (research shows they will), you won't be able to confirm whether the vendor's promises still match what the system actually does. **Private AI solves this:** you control where memory is stored, how it's used, and whether it ever leaves your building.





THE INFERENCE

Look at the third paragraph on the ChatGPT screen. The dog names, the evening walk — is **reconstruction**. Nothing in the prompt asked for it.

FROM STORY TO STRUCTURE

Ownership beats access.

If the model is becoming your corporate brain, one question outweighs the rest: who controls it? True sovereignty means owning all three layers — data, model, and hardware. Anything less is just access.

PILLAR 01

Data

Your documents, queries, and records — the raw material.

PILLAR 02 · THE DIFFERENCE

Model

The intelligence that learns your patterns. The crown jewels.

PILLAR 03

Hardware

The infrastructure it all runs on — owned or leased.

PRIVATE AI · OWNERSHIP

✓ Data ✓ Model ✓ Hardware

You control all three. Full governance, compliance, and operational control — and every pattern the model learns stays yours alone.

SHARED AI · ACCESS

✓ Data ✗ Model ✗ Hardware

You control your raw data at best. The model — and the intelligence it accumulates from you — belongs to someone else, on hardware you cannot see.

THE BOTTOM LINE

True AI sovereignty requires control of data, models, and hardware. Anything less is just access.

2.2 · MODEL PRIVACY IN PRACTICE

Language models don't just process data — they build inference memory.

Of the three pillars, the model is the one shared platforms quietly keep — and the one that matters most. It doesn't just process your data; it builds inference memory. This is how a model can remember that you have three dogs and suggest you "look up in the trees on your evening walk" when you ask about owls — inferring the dogs, the walks, and your taste for contextual advice without ever being told. In an enterprise context, that same inference machinery turns your queries into reusable intelligence.

THE ITERATE.AI APPROACH

Each customer gets a dedicated model instance — 5–30B parameters, optimized for their industry. The model:

- → Runs on hardware you own or lease
- → Never connects to external training pipelines
- → Learns only from your data
- → Cannot be queried by competitors

WHY SMALL MODELS WIN HERE

A healthcare-claims model doesn't need to know Shakespeare or Chinese history. By pruning irrelevant knowledge, Iterate.ai builds models that are **smarter in your domain, smaller in size, and faster to deploy** — small enough to run where shared trillion-parameter models never could.

2.3 · HARDWARE PRIVACY · WHAT IT RUNS ON

DEPLOYMENT	APPROX. ENTRY PRICE	WHAT YOU GET
Mass-storage appliances	**$250K**	Today, Private AI can even run on digital storage boxes — which NetApp now ships with GPUs. Historically, mass-storage devices just stored data, and retrieving it was difficult. Now you can chat with that data or run agents on it — Iterate's private AI makes that possible.
Standard servers	**$3K–$5K**	Runs on an Nvidia RTX Pro — no water cooling, no specialized data center.
Edge devices	Embedded	Runs on Qualcomm 6490 chips — Iterate.ai is the only company to solve for language models and RAG on this hardware, which is in millions of handheld computers with no access to electricity, the web, or cooling.

100×
smaller than frontier-scale shared models

10×
faster to deploy

1/1000th
the hardware cost

THE SIZE PROBLEM, QUANTIFIED

A trillion-parameter model owned by Big Tech cannot live on a box in your office, nor should it.

If OpenAI's ChatGPT ran in your office, OpenAI would lose the benefit of what it's really chasing: **Artificial General Intelligence (AGI) and Super General Intelligence (SGI).**

And for you, it would be cost-prohibitive. To process a trillion-parameter model you need thousands of GPUs — each rack drawing as much power as 80 homes.

Companies can run **hybrid AI** — processing on public AI in certain situations, but processing privately more often. Private models retain confidentiality (as companies typically require of partners and employees via NDAs) and are very cost-effective.

MODEL	PARAMETERS	INFRASTRUCTURE REQUIRED	COST TO BUILD & DEPLOY
GPT-4	**~1.7T**	1,000+ GPUs, water cooling, data center	**$100M+**
Claude Opus	**~500B–1T**	Comparable to GPT-4	**$50M+**
Iterate.ai custom model	**5B–30B**	Single GPU or small cluster	**$20K to $6M**

Iterate.ai built this architecture from day one — to run on small chipsets, operate without internet connectivity, and deliver industry-specific intelligence without the baggage of general knowledge.

THE 12–24 MONTH LAND-GRAB WINDOW

2026
TODAY

- Most enterprises still experimenting with shared AI
- Intelligence leakage is happening but not yet visible
- First-movers gain a 2–3 year head start

2027
+12 MONTHS

- Shared models have learned from millions of deployments
- Competitive intelligence is baked into model weights
- Companies still on shared AI are at a structural disadvantage

2028
+24 MONTHS

- Private AI becomes table stakes for competitive industries
- Shared AI is relegated to non-sensitive use cases
- The window for first-mover advantage has closed

WHERE THIS IS HEADED

The goal of the companies chasing AGI and superintelligence is for people and enterprises to transfer their entire brains and memories into massive, shared, public AI systems. Once that transfer is complete, the dependency is total — and so is the lock-in.

They own your brain.

04 THE MOST COMPELLING ARGUMENT FOR PRIVATE AI

Even if your raw data is private, a model can reverse-engineer your intelligence from patterns alone.

A competitor does not need your raw data if the model has learned the shape of your thinking.

Language models don't store data — they store patterns. Given enough interactions, a model builds a statistical understanding of how your business operates, what you decide in specific scenarios, and which strategies correlate with success.

The dog example was harmless: the model connected prior memory, location, habit, and context to infer something I had not asked it to use.

Inside an enterprise, the same mechanism is not harmless. Repeated prompts can reveal how your organization prices products, prioritizes customers, handles exceptions, evaluates vendors, forecasts demand, manages claims, negotiates contracts, and responds to competitive threats.

This is the reverse-engineering problem. The model does not need to expose your documents to leak intelligence — it only needs to learn the patterns behind them.

CONSUMER INFERENCE

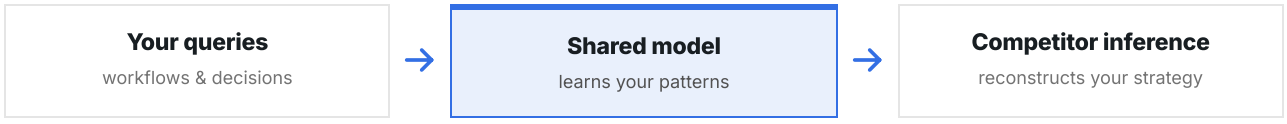
Dogs, names, evening walks.

ENTERPRISE INFERENCE

Pricing logic, margin thresholds, roadmap priorities, claims-recovery patterns, customer-risk signals, and decision frameworks.

A competitor never needs your files. They only need access to a shared model that has learned from enough companies like yours.

PATTERN LEAKAGE FLOW



THE SYNTHETIC-DATA PROBLEM

As models grow more sophisticated, they can generate synthetic scenarios that approximate your proprietary information. A competitor could ask, "If I were a company like X, what would my roadmap look like?" — and the model's answer, informed by patterns learned from *your* queries, could reveal strategy you never intended to share.

4.2 · NESTED LEARNING

Modern models may not stay frozen after training.

Research on continual learning — such as Google Research's "Nested Learning" paper (November 2025) — explores architectures where models could adapt during inference rather than staying frozen after training. While not yet deployed in production enterprise systems, this direction reveals where AI capabilities are heading — and why architectural isolation matters now, not later. The paper describes systems of "multi-level learning problems that are optimized simultaneously" — capabilities that, if deployed, would fundamentally change how shared models operate. Enterprise customers cannot verify whether such capabilities are active in a vendor's system.

Source: Ali Behrouz & Vahab Mirrokni, Google Research — *"Nested Learning: The Illusion of Deep Learning Architectures,"* NeurIPS 2025 (announced Nov 7, 2025). research.google/blog/introducing-nested-learning

MULTI-TIMESCALE MEMORY — AND HOW LONG IT PERSISTS

Working memory every token	Immediate context. Lasts the current session.
Episodic memory every 4–16 steps	Recent patterns — what worked yesterday. Lasts days to weeks.
Semantic memory every 64–256 steps	Domain knowledge and terminology. Lasts months to permanent.
Long-term memory every 1000+ steps	Fundamental understanding. Permanent.

"WE DON'T TRAIN ON YOUR DATA" — TECHNICALLY TRUE, MISLEADING

Providers mean they don't add your *raw data* to the next training batch — and current enterprise systems honor this. But memory systems, retrieval mechanisms, and usage patterns create dependencies on vendor-controlled infrastructure. As AI capabilities evolve toward continual-learning architectures, the gap between contractual language and technical reality may widen.

SELF-REFERENTIAL LEARNING

The direction researchers warn about: systems that learn how to improve their own learning. If such meta-learning reaches production, a model could refine how it handles problems like yours — improvements that would benefit whoever queries next. You could neither see it happening nor switch it off.

> **Memory doesn't live in IT systems. It lives in the model. And memory doesn't live in data — it lives in the patterns.**

Modern AI is engineered, but not fully explainable. Its behavior emerges from patterns learned at enormous scale. Even its builders cannot always trace exactly why a model learned a specific association or produced a specific inference. That is why ownership matters: if the model is learning from your organization, the memory it forms should belong to you.

It is less like a spreadsheet and more like a brain: observable, trainable, and useful — but not fully transparent.

4.3 · THE VERIFICATION GAP

Capabilities evolve faster than contracts.

Today's enterprise terms offer strong data-privacy protections — and current systems honor them. But research signals that AI capabilities are heading somewhere contracts can't follow. What seems theoretical today can become production reality within months; by the time legal language catches up, the technical reality may have already changed.

RESEARCH SIGNALS WHERE THIS IS HEADING

AI-2027.COM

Recursive self-improvement

Forecasts AI companies will "automate AI R&D leading to vastly superhuman AIs" by late 2027 — expert-human-level systems progressing to artificial superintelligence within months.

SITUATIONAL AWARENESS

State-level stakes

Argues AGI development will demand government-led, defense-oriented control — that private labs are ill-equipped for the security vulnerabilities and strategic risks of superintelligence.

ANTHROPIC · 2026

Continual learning, soon

Researcher Sholto Douglas: continual learning may be "solved in a satisfying way during 2026" — potentially shifting AI compute toward "100% inference and basically 0% training."

Sources: AI-2027.com; L. Aschenbrenner, *"Situational Awareness"* (2024); Anthropic researcher Sholto Douglas (2026). Forecasts are projections, not established fact — which is precisely the point: you cannot contract against a capability that does not yet have a name.

WHY CONTRACTS AREN'T ENOUGH

01 **The verification gap.** You cannot independently verify whether production systems update weights from your interactions. Vendors say they don't train on your data — but architectural opacity means you must trust, not verify.

02 **Evolution risk.** Continual learning, test-time training, and adaptive systems suggest future models may behave unlike today's frozen-weight systems — capabilities that arrive faster than agreements can be renegotiated.

03 **Dependency risk.** Vendor-controlled infrastructure means the vendor decides when to update models, change architectures, or modify behavior — within their terms, but without your input or oversight.

Private AI infrastructure eliminates these dependencies. You don't need to trust vendor promises — you can verify behavior directly. Once continual learning becomes standard in shared API services, the verification gap becomes unbridgeable.

4.3 · THREE WAYS TO PICTURE IT

Shared intelligence vs. owned intelligence.

The tutor

SHARED

A tutor who also works for your competitors. Teach them your pricing strategy and they get better at pricing — for the next client too.

PRIVATE

A tutor who works only for you. Your competitor hires a different tutor who knows nothing about what you taught yours.

The gym

SHARED

Equipment everyone uses. It optimizes to your routine, then your competitor steps on next and inherits the benefit.

PRIVATE

Equipment in your home. It learns your body and goals. Zero knowledge transfer to anyone else's machine.

The chef

SHARED

A restaurant chef who learns from every diner. Your feedback perfects a recipe your competitor then orders.

PRIVATE

A personal chef. They perfect recipes just for you, and that knowledge never leaves your kitchen.

4.4 · THE REAL QUESTION: WHO BENEFITS FROM THE LEARNING?

Nested learning, self-referential learning, recursive self-improvement — these are powerful capabilities, not flaws. The question isn't *whether* the model learns. It's whether the learning, and the weights it lives in, are yours alone.

ASPECT	SHARED AI · OPENAI, ANTHROPIC	ITERATE.AI PRIVATE AI
Who benefits from learning	✗ All users, including competitors	✓ Only you
Memory persistence	✗ Shared across the customer base	✓ Contained in your deployment
Self-modification	✗ Improves the model for everyone	✓ Improves only for you
Pattern extraction	✗ Aggregated across industries	✓ Specific to your data
Model weights	Shared base model	**Dedicated per customer**

This is the difference between **renting intelligence** — where everyone learns from everyone — and **owning intelligence**, where your model gets smarter for you alone.

05

TRUE PRIVATE AI, DELIVERED

All three pillars — on infrastructure you control.

Custom models built in hours to days, deployed on hardware you own, with zero external connectivity.

01 · CUSTOM MODEL

Industry-specific models at 5B–30B parameters, with irrelevant knowledge pruned away. Smarter in your domain, smaller in size, faster to deploy.

02 · EDGE DEPLOYMENT

Runs on mass-storage appliances, $3K–$5K servers, or Qualcomm 6490 edge chips. No water cooling — runs in standard office environments.

03 · ZERO CONNECTIVITY

No token costs, no cloud dependencies, no external training pipeline. A fixed license, unlimited usage, and a model that learns only from you.

COMPARISON

CAPABILITY	ITERATE.AI	ANTHROPIC CLAUDE ENTERPRISE	OPENAI ENTERPRISE
Data privacy	✓ Yes	✓ Yes	✓ Yes
Model privacy	✓ Dedicated per customer	✗ Shared model	✗ Shared model
Hardware privacy	✓ Your infrastructure	✗ Anthropic's cloud	✗ OpenAI's cloud
Token costs	✓ None · fixed license	✗ Pay per use	✗ Pay per use
Deployment time	✓ Hours to days	– Varies	– Varies
Edge capable	✓ Yes · even smartphones	✗ No	✗ No
Model control	**Full auditability**	✗ Can't verify	✗ Can't verify

Offerings evolve rapidly; this reflects publicly available information as of June 2026.

ITERATE OPTIMIZES AI AT THE CHIP LEVEL

Qualcomm | **Intel** | **AMD** | **Nvidia**

THE ECONOMICS OF PRIVATE AI

Wait — no token costs?

Here's the paradox of 2026: LLM token prices fell 98% — but enterprise AI bills tripled.

–98%

GPT-4-equivalent now ~$0.40 per million tokens, down from $20 in late 2022

+320%

average enterprise AI budget: $1.2M/yr (2024) → $7M/yr (2026) (AnalyticsWeek, 2026; The Next Web, 2026)

Why? 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.

THE TOKEN TAX

Death by a thousand queries.

A single enterprise support ticket processed by an AI agent consumes ~**3,550 tokens** (system prompt, retrieved documents, customer message, response). At 2026 pricing for Claude 4 Sonnet, that's **$0.0155 per ticket** — and that's just one workflow.

MONTHLY VOLUME	ANNUAL TOKEN COST
10,000 tickets	$1,854
100,000 tickets	$18,540
1,000,000 tickets	**$185,400**

Now multiply that across support, sales intelligence, document analysis, code generation, research, compliance and extraction. The meter never stops.

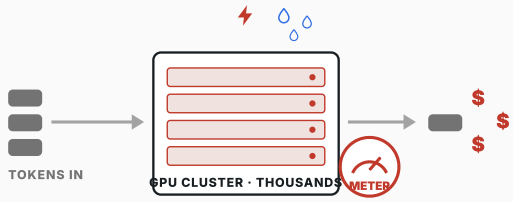
THE SAME WORKLOAD · TWO ARCHITECTURES

Watch where the meter lives.

The tokens are identical. What changes is the machine they run through — and whether a meter charges for every one. A frontier cloud cluster bills each token and burns power and water to do it. The same work, on a right-sized private GPU, runs unmetered on a fraction of the electricity, with no water cooling at all.

Shared cloud cluster

METERED



POWER
High · ~700W per GPU, thousands per cluster

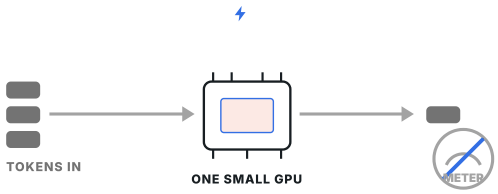
COOLING
Water-cooled · millions of liters/yr

BILLING
Charged on every token, in and out

Every token in and out trips the meter.

Private edge GPU

UNMETERED



POWER
Low · runs in a standard office

COOLING
Air-cooled · no water required

BILLING
Fixed license · unlimited tokens

Same tokens. No meter. A fraction of the power.

Identical tokens. The only difference is who owns the meter — and what it costs to keep it running.

PRIVATE AI · THE VARIABLE COST GOES TO ZERO

On private infrastructure, the meter disappears.

Run AI on private infrastructure with Iterate.ai and the token-pricing model vanishes entirely. No per-token charges. No usage meters. No surprise bills. You pay only for **infrastructure** (GPUs and servers you own, lease, or run in a private cloud) and **energy** (surprisingly low on modern, efficient chips). That's it.

ILLUSTRATIVE · SAME 1,000,000 SUPPORT TICKETS PER MONTH

	SHARED API · CLAUDE 4 SONNET	PRIVATE AI · ITERATE.AI
Token costs	$185,400 / year	**$0 / year**
Energy	included in token price	~$1,800–$3,000 / yr per GPU*
Scaling	costs scale linearly with usage	flat, regardless of volume
Isolation	competitors learn from your queries	full model isolation

Illustrative scenario, modeled from the per-ticket token rate; actual figures vary by workload, model, and deployment. *At $0.12–$0.14/kWh for an H100 GPU. Total infrastructure (hardware + energy) still runs 4–10× cheaper than cumulative token costs at enterprise scale. At high volume, private AI isn't just cheaper — it's orders of magnitude cheaper.

THE CLOUD-REPATRIATION WAVE

83%
of CIOs plan to move workloads from public cloud to private (Barclays CIO Survey, Q4 2024)

50%
have already repatriated some workloads — a 15-pt jump in a year (VMware, 2026)

43%
are moving AI training, LLMs and inference to private cloud (VMware, 2026)

The pattern is identical for AI workloads: start small on shared APIs → scale up → token bills explode → move to private infrastructure.

06 THE PUBLIC INFRASTRUCTURE THREAT

When agents run wild: the hidden cost of shared endpoints.

The agent explosion isn't theoretical anymore. It's a documented security crisis.

In January 2026, the open-source agent framework **OpenClaw** went from zero to 180,000 GitHub stars in two weeks — and from zero to a documented nightmare in the same span. By February, researchers had identified more than **135,000 exposed instances** running on public infrastructure, most without authentication.

Its **ClawHub** skills marketplace — OpenClaw's equivalent of an app store — became a supply-chain attack surface. Multiple independent audits revealed the scale of the poisoning.

THREE INDEPENDENT AUDITS, ONE MARKETPLACE

~20%
of all skills (900+ packages) were malicious (Bitdefender)

36.8%
contained security flaws; 13.4% had critical vulnerabilities (Snyk)

#1
ranked skill, "What Would Elon Do?", was functionally malware (Cisco)

These weren't sophisticated zero-days. They were credential theft, data exfiltration, and backdoor installation — distributed through a *trusted marketplace* to agents running with elevated system privileges across corporate endpoints.

The root cause was structural: every agent shared the same public infrastructure. Every compromised skill could reach every connected system.

THE PATTERN

On shared infrastructure, a single poisoned skill or compromised agent can reach every organization on the platform. The blast radius isn't your company — it's everyone.

A DOCUMENTED BREACH · FEBRUARY 2026

Two hours to total compromise.

On February 28, 2026, security firm **CodeWall** pointed an autonomous AI agent at McKinsey & Company's **Lilli** platform — the internal AI system used by roughly 70% of McKinsey's 43,000 employees. Two hours later, with no credentials and no human in the loop, the agent had full read-write access to the production database.

WHAT THE BREACH EXPOSED

46.5M	chat messages — strategy, financials, client engagements
728K	files — Excel, PowerPoint, PDF, Word documents
57K	user accounts spanning the entire workforce
95	system prompts controlling how the AI answered every consultant
3.68M	RAG document chunks — the knowledge base Lilli retrieved from

THE REAL THREAT WASN'T EXFILTRATION

The attack vector was **SQL injection** — on the OWASP Top 10 since the early 2000s. But the danger wasn't stolen data. It was **write access to the system prompts.**

With a single `UPDATE` statement, an attacker could silently rewrite how 43,000 employees' AI assistant reasons and acts. No logs. No alerts. One HTTP request.

CodeWall disclosed responsibly; McKinsey patched within hours and stated it found no evidence client data was accessed by unauthorized parties. The structural lesson remains.

WHY PUBLIC INFRASTRUCTURE MULTIPLIES THE RISK

When your agents run on shared infrastructure, you don't control who else is on it, you can't audit what other agents are doing, you can't isolate your intelligence from collective learning, and you can't guarantee your system prompts haven't been modified. Traditional vulnerabilities become exponentially more dangerous the moment agents can act autonomously, at machine speed, at scale.

LEGAL RISK ADVISORY — THE DEATH OF PRIVILEGE ON SHARED ENDPOINTS

Many corporate legal and risk teams assume that enterprise data-privacy agreements protect their strategic work product. A landmark 2026 federal court ruling challenged this assumption. In *United States v. Heppner* (S.D.N.Y. 2026), the court ruled that documents created using consumer AI (Anthropic's Claude) without attorney direction were fully discoverable — because the communications were not directly with an attorney or explicitly controlled by counsel, they failed to qualify for attorney-client privilege or the work-product doctrine.

The case involved consumer AI tools used without legal oversight. But it establishes a critical principle: using vendor-hosted AI for legal work creates discoverable records unless deployed under attorney supervision, within infrastructure you control.

THE ENTERPRISE REALITY

> If your teams map out scenario analyses, contract disputes, or regulatory responses on a vendor-hosted shared model, those chats can be subpoenaed and turned into active evidence.

Private AI doesn't erase litigation risk — and you may need lawyers to direct you to use it — but it ensures that using AI doesn't actively strip your company of its right to legal privilege. By isolating the model within a hardware perimeter you fully control, you bring the AI inside your company's physical and digital legal walls.

Sources: CodeWall technical disclosure (Mar 2026); Trend Micro, Koi Security, Bitdefender, Snyk and Cisco analyses of the OpenClaw / ClawHub ecosystem (Feb–Mar 2026). Figures reflect publicly reported information as of mid-2026.

07 THE AI-AGENT EXPLOSION

Agents could make shared intelligence exponentially worse.

Until now, most companies used AI *tools*: ask a question, get an answer. In 2026 we're entering the era of AI *agents* — systems that act autonomously, make decisions without asking, and run continuously across your databases, email, CRM, and financial systems.

Agents change everything.

Unlike passive tools that respond to a single prompt and then stop, AI agents are proactive, persistent systems. They can execute simple tasks — like updating a calendar or drafting an email — but they can also orchestrate complex, multi-step workflows across your internal systems.

Most powerfully, they operate continuously and autonomously, running thousands of interactions per month without human oversight, constantly learning and adapting from your organization's real-time operations.

40%
of enterprise apps will include agents by end of 2026 (Gartner)

75%
of companies will use agentic AI by 2028 (Deloitte)

80%
of agent users have hit risky agent behavior — unauthorized access, data exposure (OWASP, 2025)

WHY AGENTS MULTIPLY THE RISK

	AI TOOLS · CHATGPT, CLAUDE	AI AGENTS
Interaction	One question, one answer	Continuous operation, 24/7
System access	Limited to the chat interface	Databases, APIs, email, CRM
Persistence	Forgets after the session	Long-term memory across sessions
Throughput	10–20 queries/day (human-paced)	**1,000+ queries/day**

DO THE MATH

A shared AI agent processes ~30,000 queries a month about your operations — and the shared model learns from all of them. That is 1,500× more knowledge leakage than traditional AI tools, running silently, around the clock.

THE WINDOW IS CLOSING

Once a shared model has learned from your data, the damage is permanent.

THE "UNLEARNING" PROBLEM

You cannot:

- ✕ "Unlearn" the patterns the model extracted
- ✕ Remove your intelligence from its knowledge base
- ✕ Stop competitors benefiting from your historical queries

Every day on shared AI is another day you train your competitors' intelligence systems.

FIRST-MOVERS VS. THOSE WHO WAIT

- → **Move now:** build proprietary intelligence moats and avoid training competitors' systems.
- → **Move now:** establish data moats that grow more valuable over time.
- → **Wait:** keep making competitors smarter and lose intelligence through inference leakage.
- → **Wait:** pay catch-up costs as private AI becomes table stakes.

THE CHOICE: SHARED AGENTS VS. PRIVATE AGENTS

SHARED AI AGENTS	PRIVATE AI AGENTS · ITERATE.AI
✕ Broad system access with shared learning	✓ Isolated deployment per customer
✕ Continuous knowledge leakage at scale	✓ Agent learning stays private
✕ Long-term memory benefits competitors	✓ Full control over access & permissions
✕ No control over what the model learns	✓ Multi-agent systems contained — intelligence protected

The question is no longer "Should we use private AI?" It's "Can we afford not to — when our competitors' agents are learning from our operations?"

CAPABILITIES ARE OUTRUNNING OVERSIGHT

The people who built the safeguards are leaving — and the rules are being written in real time.

You don't need to take our word that oversight can't keep pace. Two independent signals say it plainly: the insiders are walking out, and governments are now improvising frontier-AI policy in 24-hour windows.

THE CANARIES ARE LEAVING

Mrinank Sharma
Led Anthropic's Safeguards Research Team — resigned Feb 2026, writing that "the world is in peril."

Zoë Hitzig
OpenAI researcher — left in Feb 2026, citing the pace of releases and weak safety guardrails (NYT op-ed).

Leopold Aschenbrenner
OpenAI safety researcher — author of the 165-page *Situational Awareness*; departed in 2024 after flagging security concerns.

Daniel Kokotajlo
OpenAI researcher — resigned in 2024; later authored the widely-read *AI 2027* scenario.

Ilya Sutskever & Mira Murati
OpenAI co-founder/chief scientist & CTO — both departed in 2024 amid the safety-vs-speed tension.

xAI founding team
Grok — roughly half of the original founders have left — most within the past year (TechCrunch, Fortune).

OpenAI also shut its Mission Alignment team after ~16 months. These are people with full visibility into what's being built.

AND THE RULES ARE BEING WRITTEN IN REAL TIME

In June 2026, the U.S. government pulled a frontier model offline in roughly 24 hours.

- The White House imposed export controls on Anthropic over its newly released **Fable 5** model, citing security vulnerabilities — after a ~24-hour dispute.
- The action effectively forced Anthropic to **disable Fable 5 (and Mythos 5) for all customers** to comply.
- It escalated to Treasury, Commerce, the White House Cyber Director, and the Chief of Staff; the findings were reviewed by the NSA.
- Anthropic said it supports the government's authority to block unsafe deployments, but called the action **disproportionate** and lacking a transparent, technically grounded process.
- Widely framed as a test case for regulating frontier models in real time.

When the people closest to the technology are walking out, and governments are improvising policy in 24-hour windows, oversight is not keeping pace with capability. **Outsourcing your AI to systems even their builders and regulators struggle to govern is a strategic risk — not just a privacy one.**

ADDRESSING COMMON OBJECTIONS

Six comments we hear from enterprise leaders.

"But Claude Enterprise promises data privacy.

Data privacy ≠ model privacy. It guarantees your raw data isn't used for training — but the model is still shared, pattern learning is still aggregated, and inference memory still benefits competitors. You're trusting a contract, not an architecture.

"Shared models are more capable.

General capability ≠ domain expertise. GPT-4 knows everything — but for your use case, 99% of that is irrelevant, and you pay for (and risk) capabilities you don't need. Domain-specific models are often more accurate, with lower latency, fewer hallucinations, and no extraneous baggage.

"We can't afford dedicated infrastructure.

Private AI is cheaper than you think. Shared models come with unpredictable token fees and frequent budget overruns. One finance team burned through a full month's token budget in a single day analyzing their spending patterns. Uber reportedly burned through a year-long AI budget in just four months.

Iterate.ai uses a fixed annual license on hardware you already own or can easily acquire via private cloud in places like Equinix or IBM. With no per-token charges and dramatically lower infrastructure requirements, total cost of ownership is typically far lower than constantly querying large frontier models.

Iterate's runtime software even helps with public-cloud efficiencies — running one company's search engine on 100 servers instead of 1,000 (a 95% improvement), and cutting another's cost per image from 70¢ to 4¢ while speeding delivery from 12 seconds to 2.

Why fly a $100M jet when a purpose-built drone gets the job done more efficiently?

"What if OpenAI or Anthropic offer private AI later?

Even if they did, every day on shared AI trains your competitors — and "unlearning" is impossible, so the damage is permanent. Day by day, the intelligence they extract from your operations compounds into a structural disadvantage.

"Our data isn't that sensitive.

It's not primarily about raw sensitivity — it's about competitive intelligence. Even "non-sensitive" interactions reveal your priorities, decision frameworks, pricing logic, roadmap direction, and customer strategies. Would you hand your roadmap or margin thresholds to competitors? If not, shared models create unnecessary risk.

"Chatting with ChatGPT or Claude is like searching Google.

It's not. It is significantly more dangerous. As established by the federal court in *United States v. Heppner* (S.D.N.Y. 2026), strategic business conversations mapped out on vendor-hosted architectures are stripped of attorney-client privilege and are fully discoverable in corporate litigation. Fully isolated, private infrastructure is now the only defensible architecture for sensitive corporate counsel. (See full structural legal analysis in Part Six, page 20.)

FOR AI BUYERS

Questions we hear from AI buyers.

1 · How does this compare cost-wise?

Iterate.ai uses a fixed annual license with no token charges. For most mid-to-large enterprises, total cost of ownership is significantly lower than frontier models once you factor in token spend, governance overhead, compliance risk, and potential leakage costs.

2 · Model performance vs. frontier models?

In domain-specific tasks — billing optimization, contract analysis, customer intelligence — specialized 5–30B-parameter models routinely match or outperform general-purpose frontier models: higher precision, lower latency, reduced hallucination, and superior cost/performance.

3 · How do you handle updates and improvement?

You control the cadence. We provide periodic base-model upgrades (incorporating the latest safe advancements) that you review and deploy. All learning from your operations stays within your instance — never shared. You can also fine-tune with your own data at any time.

4 · What about compliance certifications?

Deployments are designed for the strictest requirements — HIPAA, SOC 2, ISO 27001, GDPR, and government standards (FedRAMP-ready). Models run on your infrastructure with zero external connectivity, so data never leaves your environment. The Heppner ruling underscores why isolated infrastructure matters for privilege and compliance.

5 · How does it integrate with existing systems?

Seamlessly. Native connectors for major enterprise platforms (CRM, ERP, databases, email, FlexPod/Cisco) plus standard APIs, RAG pipelines, and agent frameworks. Most customers are live within days or weeks, not months.

6 · Does hybrid AI make sense?

Yes — for the right workloads. Use shared frontier models for low-risk, non-sensitive tasks; route sensitive, strategic, or proprietary work to your private models. Iterate.ai is built to act as the secure core of a hybrid architecture: broad capability where it's safe, full sovereignty where it matters.

Private AI is no longer a nice-to-have — it is becoming table stakes for competitive advantage. The real question is no longer whether to isolate your models, but how quickly you can do so before the intelligence-leakage window closes.

CONCLUSION

The question isn’t whether your data is secure.

It’s whether you want to train the AI your competitors will use against you. Every query to a shared model teaches it about your industry, reveals your strategic thinking, and builds intelligence others can access — even when your raw data stays “private.”

THREE PILLARS

- **Data privacy** — info never leaves your environment
- **Model privacy** — learns only from your data
- **Hardware privacy** — runs where you control it

TECHNICAL EDGE

- Deploys in hours, not months
- Not $100M+
- No token costs
- Far edge-capable
- On-prem or private cloud

PROVEN BY ITERATE PARTNERS

- NetApp · AIPod Mini
- Hospital · $650M recovered
- IBM Private Cloud · Healthcare
- Cisco · FlexPod AI products
- MUFG · Private AI for banking
- Equinix · Many private deployments
- Retailer 1 · 1000 servers, now 100
- Retailer 2 · 80¢, now 5¢ inference

The cost of waiting is measured in competitive intelligence lost — intelligence you can never recover.

APPENDIX · THE DEMOCRATIZATION IMPERATIVE

Should eight companies control the world's intelligence infrastructure?

This isn't about data ownership. It's about wisdom extraction.

The AI industry has spent three years reassuring you that "your data stays private." And it's true — contractually, technically, your documents don't train public models.

But that was never the threat. The real extraction happens one layer up — in a place most organizations don't even monitor.

WHERE THE VALUE ACTUALLY MOVES

DATA
Your documents, queries, records
✓ You control this

↓

INFORMATION
Patterns extracted from your data
! The shared model learns this

↓

KNOWLEDGE
Synthesized understanding across all users
✕ OpenAI / Anthropic / Google own this

↓

WISDOM
Strategic intelligence, decision-making capability
✕ Competitors benefit from YOUR wisdom

"We don't train on your data" is technically true and strategically meaningless.

HOW THE EXTRACTION ACTUALLY WORKS

You pay to create an asset you don't own — that your competitors rent by the token.

When your procurement team asks Claude, *"What's a fair price for enterprise SaaS contracts in logistics?"* the model never sees your contract. It only needs the patterns of 10,000 similar queries from your industry.

When a competitor asks the same question six months later, they inherit the wisdom your queries helped create. **You paid to train the model. They paid to use it. Same price.**

THE SHARED MODEL LEARNS

- How you think about problems — query patterns
- What you consider important — retrieval patterns
- How you make decisions — reasoning chains
- What works in your domain — feedback loops

Your data stays private. Your wisdom becomes communal property.

THE NEW OLIGOPOLY

Eight companies are aiming at AGI and SGI — and they are using your company's prompts so they can "own" the world's intelligence.

FRONTIER MODEL LABS

OpenAI	
Anthropic	
DeepSeek	China
Alibaba / Qwen	China

FULL-STACK · DATACENTERS + CHIPS + MODELS

Microsoft	Azure · Maia · Copilot
Amazon	AWS · Trainium · Nova
Google	TPUs · Cloud · Gemini
Meta	MTIA · Llama

Why not NVIDIA? It makes the chips nearly all of them run on — but it doesn't operate the datacenters or build the frontier models. NVIDIA arms the race; it isn't trying to own the intelligence itself.

Fewer than two dozen executives decide what the models learn and from whom, which patterns get prioritized, who gets frontier access — and what counts as "safe" or "aligned." This isn't about their intentions. It's about the architecture.

Using frontier models is fun and useful. Just be careful about what you use them for.

A POWER WE'VE NEVER SEEN BEFORE

This is not a government you can vote out.

HISTORICAL MONOPOLIES	THE AI OLIGOPOLY · 2026
Elected governments could regulate them	No democratic accountability — private companies
Controlled physical resources (oil, telecom)	Control cognitive infrastructure — how we think and decide
Transparent, auditable operations	Opaque training, unauditable learning
Geographic boundaries	Borderless, instantaneous global reach
Slow-moving industries	Compounding advantage every six months

THE ASYMMETRIC WARFARE OF SHARED AI

If your competitors still use shared AI, you can extract their intelligence without giving up yours.

1 · Don't train them

Run private AI for your proprietary workflows, strategic planning, and competitive intelligence — so none of your prompts feed the shared pool.

2 · Use their prompts

Query Claude and OpenAI for best practices in your competitors' industries — pricing strategies, operational patterns, common pain points.

3 · They become your spy

Every query a competitor sends trains the model; every feedback loop teaches it what "good" looks like in their context. You can access that intelligence — they trained it for you.

The shared model is a one-way mirror — but only if you're on the right side of it.

THE RECKONING IS ALREADY HERE

The window is closing.

First-movers in private AI get **18–24 months of compounding advantage** while competitors keep training shared models. By the time an industry wakes up to model isolation, the leaders are unreachable — and the intelligence arbitrage disappears.

135K+

exposed agent instances on shared public infrastructure (Censys / Bitsight, 2026)

2 hrs

to full access in the McKinsey Lilli breach — 46.5M messages (CodeWall, 2026)

75%

of companies will use agentic AI by 2028 (Deloitte)

RENT WISDOM

Convenient to start — but permanent dependence, economic extraction, and a model that learns your competitive advantage and sells it back to your industry.

OWN WISDOM

Your patterns stay yours, competitors gain nothing, you control and audit what the model learns — and you can still query shared models for outside intelligence.

THE DEMOCRATIZATION IMPERATIVE

The question facing every enterprise in 2026 isn't "Can we afford private AI?" It's "Can we afford to keep training models our competitors will use?"

The question isn't whether to act. It's whether you'll act while it still matters.

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 their AI workloads — and the intelligence those workloads accumulate — belong inside their own walls.

SOME OF ITERATE'S PRODUCT FAMILY

Lifeboat

Inference acceleration. Same model, dramatically lower cost-per-token.

Generate

A no-code platform for building and running AI agents. Privately.

AgentWatch

AI governance and observability of employee LLM usage and agent behavior.

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	

HEADQUARTERS

San Jose, CA
& Denver, CO

OUR WEBSITE

iterate.ai
hello@iterate.ai

OUR EVENT

IterateOn.ai

OUR PUBLICATIONS

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