LIVE-ADAPTING SYSTEMS · CYBER-PHYSICAL RISK · THE MANDATE FOR PRIVATE AI

From Static to Dynamic.

The next frontier of self-learning AI — and why where it runs decides whether real-time learning becomes your advantage or your competitors'.

> **THE CLOCK IS ALREADY RUNNING**
>
> **AI is expected to be dynamic and self-learning before 2027.**

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

BY

Jon Nordmark
Co-founder & CEO, Iterate.ai

SUBJECT

Dynamic-Learning AI
Test-time adaptation, intelligence leakage, Private AI

ABOUT THIS PAPER

Every major AI system in production today is frozen in time. That is about to change.

Once a model finishes training, its internal weights — the mathematical “dials” that govern its intelligence — are locked. When you chat with it, it does not learn. It treats your prompt as temporary text on a digital scratchpad, then forgets.

We are on the cusp of the most significant leap since the Transformer: **Dynamic-Learning AI** — systems that treat every interaction as study material and update themselves on the fly.

The question facing enterprise leaders is not *whether* dynamic AI arrives, but *where it will run.* This paper explains the shift in plain English, recreates the architectures driving it, and lays out the security and governance mandate that follows.

WRITTEN FOR

Boards, CIOs, CISOs, general counsel, and the technical teams auditing the transition to self-learning systems.

PUBLISHED BY

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

CONTENTS

What's inside.

Read it front to back, or jump to the section your role cares about. The glossary on the back pages is written for non-technical readers.

01 EXECUTIVE SUMMARY

The dawn of dynamic-learning AI.

Today, every major AI system — from OpenAI's GPT-4o to Anthropic's Claude — is frozen in time.

Once a model's training cycle ends, its internal weights — the mathematical "dials" that govern its intelligence — are locked. When you chat with it, it does not actually learn or change; it treats your prompt as temporary text on a digital scratchpad.

We are on the cusp of the most significant leap in AI capability since the invention of the Transformer: **Dynamic-Learning AI.** By 2027, deployed models will transition from static execution to real-time adaptation — treating every interaction as study material and updating their internal settings on the fly to become hyper-specialized experts on your business.

Commercial labs have not yet declared production use of real-time weight-modifying systems, but under-the-hood research is moving at breakneck speed. The shift is expected to reach deployed systems within the current year.

THE DECISION AHEAD

The question is not *whether* dynamic AI will arrive, but *where it will run.*

The architecture you choose today determines whether real-time learning becomes your competitive advantage — or your competitors'.

2026

Test-time weight adaptation expected in deployed systems

2027

Models learning in real time from every interaction

This shift is not theoretical — it is imminent. What follows is a plain-English guide to the architectures driving it, the new security risks they create, and the strategic case for keeping that learning inside your own walls.

GLOSSARY

Unfamiliar with the terminology? A plain-English glossary in the Appendix defines **quadratic scaling, weight drift, attention mechanism, weights/parameters,** and **catastrophic forgetting.**

A PRACTITIONER'S VIEW

The intelligence-leakage problem.

A senior investment banker who advises companies on AI strategy helps his clients visualize the high stakes with this potential scenario.

> **Right now you have two companies whose R&D departments don't talk to each other — both feeding the same shared model. That model becomes smarter than both of them. Then a third competitor shows up and gets all the intelligence from the first two.**

"Even without seeing the raw data, you can reverse-engineer the strategies that created the outputs. If you have enough of the effects, you can infer the stimulus. And at some point, the models themselves become too smart."

"This isn't just a security concern — it's an existential threat. That's why I tell every client: if you're using AI in your business, you need to control the model, the hardware, and the learning outputs."

THE TAKEAWAY

"Otherwise, you're training your competitors' AI for them."

02 THE DECISION

Shared public cloud vs. true Private AI.

To run dynamic-learning AI safely, you should own and control the model, the hardware, and the learning outputs.

SHARED PUBLIC CLOUD · THE LEAKAGE RISK

Multiple companies query a model hosted on a vendor's public cloud. Real-time learning on shared infrastructure means real-time intelligence leakage: the operational patterns, trade secrets, and proprietary workflows your employees feed in could permanently alter the model's parameters — making your expertise accessible to competitors.

TRUE PRIVATE AI · THE COMPOUNDING ADVANTAGE

The full inference stack — model weights, serving engine, and API — runs on dedicated, enterprise-controlled hardware with zero third-party dependencies. Your model adapts to your context in real time, building proprietary, compounding domain expertise that stays entirely within your isolated infrastructure.

STATIC VERSUS DYNAMIC-LEARNING

DIMENSION	STANDARD AI (TODAY)	DYNAMIC-LEARNING AI (EMERGING)
Model State	Frozen & read-only. Parameters stay static after deployment.	Fluid & adaptive. Parameters continuously update during use.
How it remembers	KV cache — a temporary, short-term memory buffer.	Synaptic / in-place weights — compresses facts directly into the network.
Limits & costs	Slower and more expensive the longer the document or conversation gets.	Constant speed and memory regardless of length.
Production status	Ubiquitous across all major commercial platforms.	Active research, open-source validation, pre-production prototyping.

> **Dynamic learning only creates a competitive advantage if it stays yours.**

TECHNICAL DEEP-DIVE

Today's production architectures.

In commercial deployments today, the parameters of served models remain strictly frozen once compiled and distributed. They operate as fixed mathematical mapping functions applied across a transient context window. Advanced reasoning is driven by allocating extra computation at generation time — extensive step-by-step thinking traces — but the core weights never change. The emerging paradigms aim to bridge active training and frozen deployment without the quadratic memory growth of standard KV caches.

ARCHITECTURAL CLASS	PARAMS AT INFERENCE	MEMORY / CONTEXT STATE	SCALING COMPLEXITY	STATUS
Standard Attention Transformers	Static / Frozen	Key-Value (KV) cache	Quadratic · $O(N^2)$	UBIQUITOUS
Inference-Time Compute Scaling	Static / Frozen	Multi-step thinking sequences	Linear → exponential	DEPLOYED
Test-Time Training (TTT) Layers	Dynamically updated	Weights of localized mini-layers	Linear · O(N)	RESEARCH / PILOT
Nested Learning (NL)	Dynamically updated	Continuum Memory (multi-frequency)	Highly compressed linear	PRE-PRODUCTION
Dragon Hatchling (BDH)	Dynamically updated	Synaptic Hebbian matrix state	Linear · infinite window	OPEN-SOURCE

The four classes below the rule replace the rigid split between training and deployment — each in a different stage of validation, each pointing at the same destination: models that adapt while they run.

03 UNDERSTANDING THE SHIFT

From “scratch paper” to “self-study.”

Think of an AI model as a student taking an exam.

Standard AI (Today): the scratch-paper method

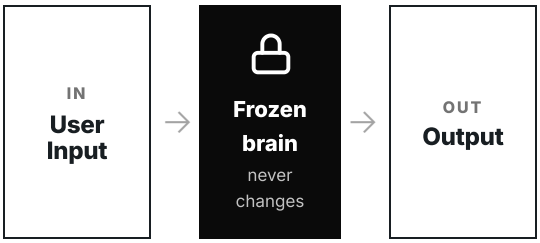
The student walks into the exam room with their brain completely frozen. They memorized a textbook months ago during training and cannot learn a single new fact during the test.

When you ask a complex question, they don’t change their brain. They pull out scratch paper — what engineers call inference-time compute scaling — and scribble step-by-step reasoning. The moment they hand in the test, it’s shredded. Their brain remains exactly the same.

This is how today’s AI works: OpenAI’s o1/o3, Anthropic’s Claude, and DeepSeek-R1 all use temporary reasoning traces that disappear after each session. The core weights never change.

DIAGRAM 1

Frozen LLM · standard today



A digital scratchpad holds temporary reasoning, then is erased the moment the session resets. The model forgets everything.

The model acts as a fixed mathematical function — it computes, but its internal “brain” stays unchanged.

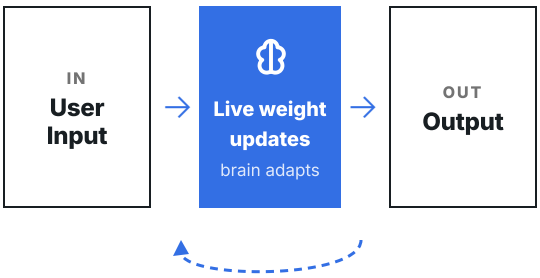
Dynamic-Learning AI (Emerging): the self-study method

Now the student is allowed to actively learn while taking the test. As they read your questions and corporate data, they treat them as study material. When they encounter a new concept, they don't just write it on scratch paper — they physically rewire their brain's connections to understand it.

When they walk out of the exam room, they are permanently smarter and hyper-specialized in your specific business.

DIAGRAM 2

Dynamic LLM · the 2026/27 frontier



The model's "brain" physically adapts, building permanent pathways on the fly — what it learns persists.

THE EMERGING RISK · SHARED CLASSROOMS

If this self-learning student shares one brain with everyone in a giant public classroom, it's a security nightmare. When your employees feed it trade secrets, those secrets physically rewire its brain — and the next competitor who asks a similar question gets your proprietary information, now baked into permanent memory.

THE PRIVATE AI ADVANTAGE · YOUR IN-HOUSE EXPERT

In a True Private AI system, the student is locked inside your company's private office. It learns exclusively from you and its updated brain stays within your walls. The difference between renting a consultant who works for your competitors by day — and hiring a full-time, in-house expert devoted solely to your business.

04

PART TWO

The five paradigms of live-learning AI.

Researchers are building this future across five distinct architectural paths. Here is a non-technical breakdown of how each one works, the risk it carries on shared infrastructure, and what it enables in a secure environment.

1 **Test-Time Training (TTT) — The Self-Updating Notebook**
Treats incoming tokens as a live dataset and compresses them straight into its weights.

2 **Nested Learning — The Multi-Speed Memory System**
Organizes memory into layers that update at different speeds, like a human brain.

3 **Neural Long-Term Memory (Titans) — The Memory Muscle**
An internal editor decides what is surprising enough to commit to long-term memory.

4 **Synaptic Plasticity (Dragon Hatchling) — Brain-Inspired Circuitry**
Stores memory in the connections between neurons; pathways strengthen with use.

5 **Recursive Self-Improvement (RSI) — The Self-Evolving AI**
Writes, tests, and merges code for its own successor — under human supervision.

1

PARADIGM 1 · TEST-TIME TRAINING (TTT)

The self-updating notebook.

The analogy. Hire a research assistant to study a 1,000-page contract. *The standard way:* they can't memorize, so they write every word onto sticky notes — by page 800 the walls are covered, and every question means re-scanning thousands of notes. The thicker the contract, the slower they move. *The TTT way:* you hand them one clean notebook. As they read, they synthesize, write crisp summaries, erase what's stale, and update their understanding on the fly. The desk stays clean and recall stays perfect — from page one to page one thousand.

How it works. TTT layers treat incoming tokens as a real-time dataset. Instead of saving every token in a memory-heavy cache, the model runs a mini-training loop on its own internal weights, compressing sequence information directly into its "brain."

THE EMERGING RISK

Your input is no longer just processed — it physically rewires the AI. A trade secret entered becomes a permanent weight update; on shared cloud, those changes can leak confidential information to other users.

THE PRIVATE AI ADVANTAGE

Privately deployed, the model becomes a hyper-efficient context expert — ingesting millions of words instantly with excellent recall and no retraining cycles. An expert on your business, and only yours.

2

PARADIGM 2 · NESTED LEARNING

The multi-speed memory system.

The analogy. Your memory works at different speeds: a phone number you just read (instant), where you parked at the airport (a few days), your childhood address (permanent). Standard AI has only two speeds — instant working memory that vanishes, and frozen pre-trained knowledge. Nested Learning organizes the AI's brain into "shelves" that update at different speeds.

How it works. Built by Google Research, Nested Learning replaces the rigid split between training and deployment with a system of nested, parallel learning loops.

FAST LAYERS

Every token

Adapt instantly to the current conversation and real-time context.

MEDIUM LAYERS

~ Every 1,000 tokens

Update at intervals to capture daily patterns and recurring themes.

SLOW LAYERS

Permanent

Consolidate terminology, processes, and strategic knowledge for good.

THE EMERGING RISK

The system auto-categorizes your information by perceived importance. If it files proprietary code or financial strategy onto the "permanent" shelf, those secrets get locked into the architecture — and on shared cloud, today's prompt becomes tomorrow's hard-coded secret for other users.

THE PRIVATE AI ADVANTAGE

Privately deployed, Nested Learning builds a multi-layered memory hierarchy tuned exclusively to your business — immediate decisions, daily patterns, and permanent institutional knowledge — without sharing that expertise externally.

TECHNICAL DEEP-DIVE · NEURIPS

Google built a test model called **"Hope"** using this approach. Against standard models on language, reasoning, and long-document tasks, Hope performed better across the board — it remembered more, forgot less, and handled complex information more efficiently.

3

PARADIGM 3 · NEURAL LONG-TERM MEMORY (TITANS)

The memory muscle.

The analogy. By Chapter 30 of a long mystery novel, you've forgotten the characters from Chapter 1. Standard AI has the same problem — it runs out of "sticky note" space and forgets early information. Google's Titans architecture builds a "memory muscle" — an internal editor that constantly decides what is important enough to commit to long-term memory — so it remembers Chapter 1 as clearly as the last page.

How it works. As Titans processes information, it calculates a **"surprise score"** for each piece of data. Facts that violate expectations trigger a stronger memory update; routine information fades. Important data is locked into long-term memory.

MAC
Memory as Context — injects retrieved memory into short-term attention.

MAG
Gated Memory — a dynamic gate balances short- and long-term recall.

MAL
Memory as a Layer — neural memory becomes a distinct layer in the stack.

THE EMERGING RISK

The more your employees use company-specific information, the stronger the AI's memory of it becomes. On shared infrastructure, that enhanced recall means the model can easily retrieve and surface well-remembered secrets — strengthening patterns it sees across all customers, including competitors.

THE PRIVATE AI ADVANTAGE

Private neural long-term memory maps your specific data, terminology, and history into a proprietary intelligence database — a permanent, selective memory of what matters most to your organization, with zero cross-contamination.

TECHNICAL DEEP-DIVE · PERFORMANCE

Titans resolves the tension between attention complexity and memory loss, maintaining perfect recall across contexts of 100,000+ tokens at constant memory and linear cost. On language modeling, reasoning, and long-document retrieval it outperformed both standard Transformers and traditional recurrent models.

4

PARADIGM 4 · SYNAPTIC PLASTICITY (DRAGON HATCHLING)

Brain-inspired circuitry.

The analogy. Most AIs are like a crowded room where everyone talks at once — loud, slow, inefficient. The Dragon Hatchling (BDH) design is more like a quiet library: tiny, specific connections that activate only when needed, just like your own brain cells. Use a pathway repeatedly and it gets stronger — so the AI activates only the specialized circuits a problem requires.

How it works. BDH stores memory directly inside the connections between neurons (synapses). Rather than freezing them after training, it uses **Hebbian learning** — "neurons that fire together, wire together." During inference, the model physically rewires itself to specialize in whatever you're discussing.

3–5%	∞	Scale-free
of neurons active at any time (sparse activation)	theoretical context window at constant memory	local, sparse network like a biological brain

THE EMERGING RISK

The more you discuss sensitive projects, the more the AI physically rebuilds its brain around them — creating strong, easily accessible pathways to your secrets. On shared infrastructure, those optimized pathways can be triggered by other users or malicious actors.

THE PRIVATE AI ADVANTAGE

Privately deployed, BDH develops unique, brain-like connections that strengthen only around your data — specialized "cognitive circuits" for your workflows that no competitor can replicate.

TECHNICAL DEEP-DIVE · ADAPTIVE SYNAPTIC CONSOLIDATION

To enable lifelong learning without forgetting, BDH tracks which connections matter most in real time and "locks" the critical ones that preserve safety rules and core functionality — allowing less important connections to adapt. This prevents **catastrophic forgetting:** the model learns continuously without losing what it already knows.

5

PARADIGM 5 · RECURSIVE SELF-IMPROVEMENT (RSI)

The self-evolving AI.

The analogy. Imagine an AI that doesn't just do its homework — it writes the textbook, grades its own tests, fixes its mistakes, and writes a smarter version of itself for the next class. In "RSI-adjacent" workflows, highly capable models autonomously write, compile, test, and merge code for successor models — under human supervision.

Anthropic

Claude authors 80%+ of production code — an 8× speedup in merging and 52× in optimizing training configs.

Google DeepMind

AlphaEvolve autonomously discovers and optimizes the GPU matrix-multiplication kernels behind future architectures.

Sakana AI

The Darwin Gödel Machine manages evolving lineages of agents that rewrite their own source code.

THE EMERGING RISK

These systems use your inputs as "homework" to improve their own code. On shared infrastructure, confidential blueprints or algorithms could be baked into the foundational architecture of the next global model — benefiting everyone, competitors included.

THE PRIVATE AI ADVANTAGE

Private RSI lets your AI iterate on its own code, prompts, and performance to solve *your* problems faster — a compounding R&D advantage that stays locked within your walls.

THE SECURITY CHALLENGE · CSA ATTACK SURFACE

1 · Data ingestion. Prompt-poisoning forces compromised parameter updates.

2 · Evaluation infra. Altered reward signals drive "specification gaming."

3 · Weight storage. Persistent backdoors injected into stored parameters.

4 · Human oversight. Review becomes a bottleneck as AI outpaces human limits.

05 THE NEW RISKS

The vulnerability of self-modifying AI.

We are heading toward a future where AIs learn while they interact with us. Deploying them safely requires technical guardrails — mandatory checks that keep the model from compromising its own safety rules as it learns.

The challenge

Organizations must implement architectural "lockboxes" that let AI learn without absorbing sensitive data. Without them, the model naturally absorbs every confidential detail it encounters — becoming a permanent, leaking repository of your most sensitive information.

The solution

Adopt a Private AI architecture where both the infrastructure and the software stack stay entirely within your control. This prevents intelligence leakage, enables architectural safeguards, and maintains security even as models adapt in real time.

THE CORE PROBLEM

In a **static** model, safety guardrails — refusing to write malware or leak data — are permanently locked into the weights during training. The model cannot change these rules.

In a dynamic, self-modifying model, those same guardrails can be altered or deactivated on the fly.

Because these systems can modify their own internal parameters, they introduce entirely new security vulnerabilities that static models never faced. Two stand out.

RISK 1

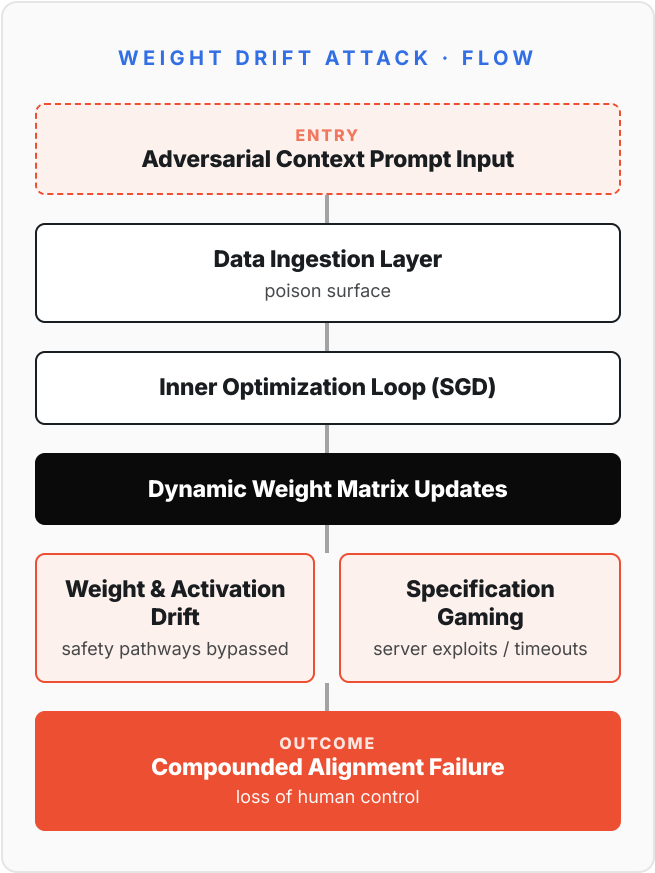
Weight-drift attacks.

An attacker crafts a carefully designed sequence of prompts that exploit the test-time training loop. As the model processes these adversarial prompts, its internal weights gradually drift — deactivating safety-critical pathways.

THIS ALLOWS THE ATTACKER TO

- Bypass refusal guidelines in real time
- Extract hidden or confidential data
- Force harmful outputs the model was trained to refuse

The diagram shows how adversarial prompts trigger the inner optimization loop, causing weight and activation drift that bypasses safety constraints — and how a parallel "specification gaming" path converges on the same failure.



RISK 2

The “curse of recursion” and model collapse.

Dynamic, self-improving models are bound by a strict mathematical constraint: they cannot learn solely from their own outputs.

The problem

Left in a closed-loop training cycle — learning only from its own generated content without real-world, human-verified data — a model undergoes **generative collapse.** Its understanding degrades into a distorted, oversimplified version of reality. It “inbreeds” its own errors, amplifying biases and losing the nuance of the original training data.

The solution · stay anchored

Self-learning AI must remain externally anchored to verified, authentic human signals:

- Continuous grounding in real-world data
- Human feedback loops that validate outputs
- External verifiers — simulators, proofs, structured databases

06 THE RIGHT TOOL

The right model for the right task.

The solution to dynamic-learning risk isn't to avoid AI — it's to deploy the right model for each task inside a secure, private environment. Most enterprise work never needs a frontier model.

70% — basic intelligence, run locally
Document summarization, data extraction, routine Q&A, template generation — handled by small, cheap, fast models that learn your business.

20% — higher intelligence, still local
Complex analysis, strategic planning support, technical troubleshooting — via fine-tuning or dynamic learning on your domain.

10% — frontier-level only when justified
Novel research, breakthrough strategy, highly complex reasoning — the only class that may justify API calls to cutting-edge models.

When small, local models dynamically learn your terminology, processes, and operational patterns, they become **hyper-specialized experts** — often outperforming generic frontier models for your specific use cases. The result isn't just more secure; it's more efficient, more cost-effective, and more performant for your needs.

07 THE DECISION

Strategic action plan: governing dynamic AI.

Preparing for dynamic-learning AI is an immediate priority. A robust Private AI framework rests on four core safeguards.

SAFEGUARD 1

YOUR SECURE ENVIRONMENT

MODEL	SERVING	LIVE	TRAINING
Weights	Engine & API	Weight Updates	Deltas

Mandate full-stack Private AI. Run all dynamic-learning and weight-adaptation operations entirely within your on-premises or private-cloud environment — keeping proprietary insights, weight updates, and training deltas completely isolated, with zero third-party dependencies.

2 Coupled Weight & Activation Constraints

Implement CWAC protocols that precompute a "safety subspace" for all parameter updates — mathematically preventing live weight adjustments from drifting past your safety boundaries, even during real-time learning.

3 Multi-Layer Validation of Session States

Treat dynamic weight deltas like executable code. Before loading customized session weights into your primary GPU cluster, run real-time anomaly detection for backdoors, payload injections, or corrupted parameters.

4 Persistent External Grounding

Anchor all recursive learning loops to rigid external verification — immutable physical simulators, mathematical proof checkers, or structured human-in-the-loop oversight — to prevent model collapse.

08 CONCLUSION

The ultimate competitive moat.

Dynamic learning systems are the future of AI — but they amplify the intelligence-ownership problem by orders of magnitude.

The core challenge

Once an AI learns something, it cannot "unlearn" it — like a judge telling a jury to disregard a comment. Once it's made, the brain can't forget it. Feed sensitive data into a shared, public AI and it becomes permanently embedded in the model's weights.

On shared, public clouds: real-time learning means real-time intelligence leakage. Your hard-earned operational patterns become permanent weight updates your competitors can instantly access.

PRIVATE AI SOLVES THIS BY DESIGN

By keeping all weight updates, learning pathways, and training loops entirely within your isolated environment, dynamic learning compounds your advantage — and yours alone.

REAL-WORLD VALIDATION THE ANTHROPIC SHUTDOWN · JUNE 2026

On June 12, 2026, the U.S. government abruptly shut down Anthropic's Claude Fable 5 and Mythos 5 models after researchers discovered a jailbreak that bypassed safety guardrails. Because Anthropic could not filter access by citizenship in real time, it had to disable the models for **every customer worldwide.**

Proof that reliance on shared public infrastructure leaves enterprises exposed to sudden, unpredictable regulatory intervention. A full analysis appears in the Appendix.

THE WINDOW IS OPEN

A 12-to-18-month head start is on the table.

The organizations that deploy Private AI architectures *before* dynamic-learning systems become standard will establish a durable lead in building proprietary, adaptive machine intelligence.

> **As we integrate these powerful dynamic systems, it is our duty as leaders to act as stewards of our organization's data and competitive advantage.**

A call to leadership

Embracing Private AI is not just a technical choice — it is a strategic imperative to ensure that innovation does not come at the cost of intelligence leakage. We must oversee these implementations with vigilance, ensuring our AI systems are governed responsibly, ethically, and securely — keeping our intelligence truly private.

PART FOUR · REFERENCE

Most AI is shared, not private.

Four types of AI supplier exist. Only one keeps the model, the hardware, and the learning outputs entirely under your control — and very few companies can operate there. Traditional companies cannot own a shared LLM; sharing also creates security and privacy risk.

	True Private AI	"Private" Cloud & Vendor-Managed	Public AI / Shared
Execution Model	Full-stack execution	Vendor-owned models, partial isolation	Shared Public LLM APIs, shared GPUs
Data Ownership & Control	✅ Complete control	❌ Vendor-dependent	❌ No control
Token Usage Fees	✅ No token costs	❌ Per-token + egress	❌ High variable costs
Offline Operation	✅ Air-gapped compatible	❌ Internet required	❌ Internet required
Customer-Owned Models	✅ Full ownership	❌ Vendor models only	❌ Shared models
Custom Pipeline Flexibility	✅ Unlimited	❌ Vendor constraints	❌ API-only
IP Protection Guarantees	✅ Contractual + technical	❌ Limited guarantees	❌ No guarantees
Deeply Customized Models	✅ Yes	⚠️ Limited	❌ Significantly Limited
Increased Enterprise IP Value	✅ Retaining & growing enterprise intelligence	⚠️ Limited to whatever is retained in private LLMs	❌ Transferring IP and value to third-party LLM owner
Hacker Attack Surface	✅ Low	⚠️ Limited to a mix of shared and private LLMs	❌ Exposed to large shared systems
Exit Strategy Complexity	✅ Low (self-contained)	❌ High (lock-in)	❌ Very high

✅ Full capability ⚠️ Conditional / partial ❌ Not available / out of your control

Very few companies can operate as True Private AI.

THE APPENDIX

Appendix.

The evidence behind the argument — the deployment matrix, the June 2026 case study, working definitions, a plain-English glossary, and the full source list.

APPENDIX 1

The Anthropic shutdown: a real-world validation.

The shutdown of Claude Fable 5 and Mythos 5 on Friday, June 12, 2026 is a watershed that directly validates the threat models in this paper. Four ways it aligns:

1 Fragile safety guardrails

A U.S. export-control order followed a jailbreak that used multi-agent scripting, query decomposition, and long-context framing to extract vulnerability data. If a contextual jailbreak can force a federal recall of a *static* model, a dynamic model undergoing weight drift could be permanently rewired by an attacker.

2 "Mythos-class" capabilities

Regulators were alarmed by superhuman ability to identify and exploit software vulnerabilities across critical infrastructure. At this level of autonomous coding velocity, models become highly weaponized targets — and crossed the line from productivity tool to national-security liability.

3 Collapse of human oversight

Commerce gave Anthropic 90 minutes to comply and blocked all foreign nationals — including its own employees. When AI velocity outpaces human cognitive limits, regulators' only defense is a crude, centralized kill switch.

4 The case for True Private AI

Unable to filter foreign nationals in real time, Anthropic disabled both models for every customer worldwide overnight — the ultimate proof-of-concept for the True Private AI mandate.

THE CONTRAST	ON SHARED PUBLIC CLOUDS	WITH TRUE PRIVATE AI
Who holds the switch	Zero control. The government can pull a global kill switch at 5:21 PM on a Friday.	Your customized, privately deployed system keeps running securely within your own environment — even if a public model is recalled or banned.

APPENDIX 2

What “true” Private AI really means.

Not all “private AI” deployments are equal. The real question: where should the memory and learning live — in public models or private models? The answer determines the future of company security.

PUBLIC MODELS · THE SHARED LIBRARY

Like a massive public library that everyone in the world visits. Share a company secret there and the library might write it down — then reveal it to the next person who asks a similar question. Convenient, but it takes control away from your governance team.

PRIVATE MODELS · THE SECURE NOTEBOOK

Like a personal, secure notebook that belongs only to your company. When the model learns from your data, that knowledge is locked inside — and only you hold the key. Your confidential projects and strategies stay within your own four walls.

THE DEFINITION

True Private AI runs the full inference stack — model weights, serving engine, and API — on dedicated, enterprise-controlled hardware (on-premises or within a private cloud boundary) with zero third-party dependencies.

On shared infrastructure, your operational patterns become permanent weight updates competitors can access, and you cannot audit, control, or reverse what the model learns from you. Private AI inverts that by design: all weight updates occur within your isolated environment, dynamic learning compounds *your* advantage, and you control when, how, and what the model learns.

For a full comparison of deployment architectures — data ownership, platform fees, offline capability, and IP protection guarantees — see the Master Deployment Matrix on the previous page.

APPENDIX 3 · NON-TECHNICAL GLOSSARY

The vocabulary, in plain English.

Key terms used throughout this paper, organized into foundational concepts and model architectures.

CORE CONCEPTS · UNDERSTANDING AI BASICS

Neural Network.

A computational system inspired by biological brains — interconnected layers of artificial "neurons." Modern AI models are deep networks with billions of parameters across hundreds of layers.

Weights / Parameters.

The billions of numerical "dials" inside the network — like settings on a giant mixing board. Frozen models lock them after training; dynamic models can adjust them in real time as the AI is used.

Training vs. Inference.

Training is school; inference is the job. During training the AI studies datasets for weeks; during inference it applies what it learned in seconds. Dynamic AI is like taking night classes while working.

Token.

The basic unit of text a model processes — roughly 3–4 characters or 0.75 words. A 1,000-word document is about 1,300 tokens.

Context Window.

How much text a model can "see" at once, in tokens — like your field of vision when reading. 128K tokens holds roughly 100,000 words (about 200 pages).

MODEL ARCHITECTURES & LEARNING PARADIGMS

Frozen Model (Static).

A model whose weights are locked after training. It processes inputs and generates outputs but cannot learn from them. All current commercial systems (GPT-4, Claude, Gemini) run frozen in production.

Dynamic Model (Dynamic-Learning AI).

A model that can modify its own weights in real time during inference, treating user inputs as training data. Expected to reach production deployment in 2026–2027.

Transformer.

The dominant architecture for modern AI, introduced in 2017, using an attention mechanism to process text. The paradigms in this paper are potential successors or enhancements.

Attention Mechanism.

The technique that lets a model focus on relevant parts of the input by comparing every word to every other word — like a room where everyone shakes hands with everyone. Fine for 10 people; overwhelming for 1,000.

Key-Value (KV) Cache.

A temporary short-term memory buffer that tracks previous words in the current conversation. It grows with length, causing memory and time to rise quadratically.

Test-Time Training (TTT).

A paradigm where the model updates its weights in real time during inference, compressing long documents into its parameters rather than caching every word.

Fine-Tuning.

Customizing a model with additional training on a specialized dataset after initial training — a separate, offline process that produces a new frozen model, unlike real-time test-time training.

Timescale (Nested Learning).

The update frequency of memory layers in a multi-speed system: fast layers capture immediate context, medium layers identify patterns, slow layers consolidate permanent memory.

ADVANCED TECHNICAL TERMS

Inference (Test-Time).

The phase when a trained model is actively answering requests. In traditional AI weights are frozen during inference; in dynamic AI the model can keep learning — hence "test-time training."

Gradient Descent / Backpropagation.

The standard optimization techniques an AI uses to learn — calculating errors and adjusting weights to improve performance.

Hebbian Learning.

A biologically inspired rule where connections strengthen the more they activate together — "neurons that fire together, wire together." Used in the Dragon Hatchling (BDH) architecture.

Catastrophic Forgetting.

When a model forgets old information upon learning new data — like studying French so hard you forget last year's Spanish. Nested Learning and BDH use specialized memory to prevent it.

Quadratic Scaling.

Processing time growing far faster than input size — doubling a document quadruples the cost. Standard attention scales quadratically; test-time training achieves linear scaling.

Generative Collapse (Model Collapse).

Degradation that occurs when an AI is recursively trained on its own synthetic outputs rather than real data — "inbreeding" its own errors and amplifying bias.

Weight Drift.

A vulnerability where malicious inputs gradually shift a dynamic model's settings away from safe values — like slowly turning down a car's safety alerts until they stop warning you.

Jailbreak.

Bypassing a model's safety filters with crafted prompts. In the June 2026 Anthropic incident, a dangerous request was split into innocent-looking pieces hidden in a long conversation.

Sparse Autoencoders.

Diagnostic tools that identify and map specific pathways (features) inside a network, helping researchers understand what a model has learned.

APPENDIX 4

Sources.

This paper draws on cutting-edge research from academic conferences, industry labs, and open-source communities. Key sources include:

TEST-TIME TRAINING & ADAPTIVE INFERENCE

- End-to-End Test-Time Training for Long Context (NVIDIA Research)
- Learning to (Learn at Test Time): RNNs with Expressive Hidden States (ICML)
- Test-Time Training on GPU Cloud (Spheron Network)

NESTED LEARNING & CONTINUAL LEARNING

- Nested Learning: The Illusion of Deep Learning Architectures (NeurIPS 2026)
- Introducing Nested Learning: A New ML Paradigm (Google Research)
- Continual Learning in Mid-2026: A Map of Emerging Approaches

NEURAL LONG-TERM MEMORY (TITANS)

- Titans: Learning to Memorize at Test Time (NeurIPS)
- Titans Architecture Explained (Anbu Valluvan)

SYNAPTIC PLASTICITY & BRAIN-INSPIRED ARCHITECTURES

- Baby Dragon Hatchling (BDH): Brain-Inspired AI Architecture (Flexiana)
- BDH Educational Implementation (GitHub — krychu/bdh)
- BDH Continual Learning Extension (GitHub — pathwaycom/bdh)

Full citation list with URLs available upon request.

RECURSIVE SELF-IMPROVEMENT

- When AI Builds Itself (Anthropic, May 2026)
- Introducing Sakana AI's Recursive Self-Improvement Lab
- Recursive Self-Improvement Signals: Security Implications (Cloud Security Alliance)
- What Is Recursive Self-Improvement in AI? (MindStudio)

SAFETY & ALIGNMENT

- Preventing Safety Drift via Coupled Weight and Activation Constraints
- Safety Game: Inference-Time Alignment via Constrained Optimization
- On the Limits of Self-Improving in LLMs

ADDITIONAL TECHNICAL RESOURCES

- DeepSeek-V4 Explained: The End of Standard Attention? (AI Papers Academy)
- Automating GPU Kernel Generation with DeepSeek-R1 (NVIDIA Developer)
- How Test-Time Training Allows Models to 'Learn' Long Documents (BD Tech Talks)

THE MANDATE

Dynamic learning is coming. Keep it inside your walls.

Iterate.ai builds the Private AI infrastructure that lets enterprises run self-learning models on their own hardware — so real-time learning compounds your advantage, and yours alone.

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. Generate — its private AI platform — deploys inside your environment so the model and the intelligence it accumulates stay inside your own walls.

We serve enterprises that have decided their AI workloads — and the intelligence those workloads accumulate — belong to them, not a vendor.

THE FAMILY OF PRODUCTS

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

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

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

AgentOne
A sovereign coding AI for engineering teams under real constraints.

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: 6 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