The Adaptive Turn

How Interactive Learning Will Redefine Machine Intelligence

Michael X. Crowe
Managing Director, Askew Kabala & Company, Inc.

December 2025

Introduction

The field of machine learning is approaching a transition as consequential as the introduction of the Transformer concept in 2017. While Transformer-based language models have produced unprecedented gains in how effectively machine intelligence can interact with human users, the limits of static, read-only models are becoming increasingly apparent as real-world deployments demand systems that can learn continuously, adapt to context, and retain newly-acquired information over time.

Fixed parameters, session-limited context windows, and costly retraining cycles constrain what current architectures can achieve, despite their already remarkable performance. In response, a new class of adaptive AI systems is emerging. These systems are designed to (1) learn across multiple timescales, from immediate context to durable memory, (2) improve through iterative practice with evaluation, and (3) accept low-cost, reversible updates, enabling safe specialization and continuous improvement without destabilizing core behavior. These developments mark the beginning of a structural shift in machine intelligence, one in which models continuously learn and evolve through interaction rather than remaining frozen after training.

This paper argues that if the Transformer revolution represents the first major inflection point in the evolution of modern large language models (LLMs), we could soon be approaching a second: a move to adaptive architectures that will redefine how AI systems are built, deployed, and integrated into human lives. While the architectural components described in this paper – nested learning, self-play refinement, and parameter-efficient adaptation – have each been demonstrated individually in recent research, no deployed system yet integrates these capabilities into a unified, continuously-learning, personalized architecture. What is presented in this synthesis is a forecast based on demonstrated components, not a description of existing systems. The thesis here is that the convergence of these adaptive technologies is technically feasible, and that when it arrives, the nature of human-AI interaction will fundamentally change.

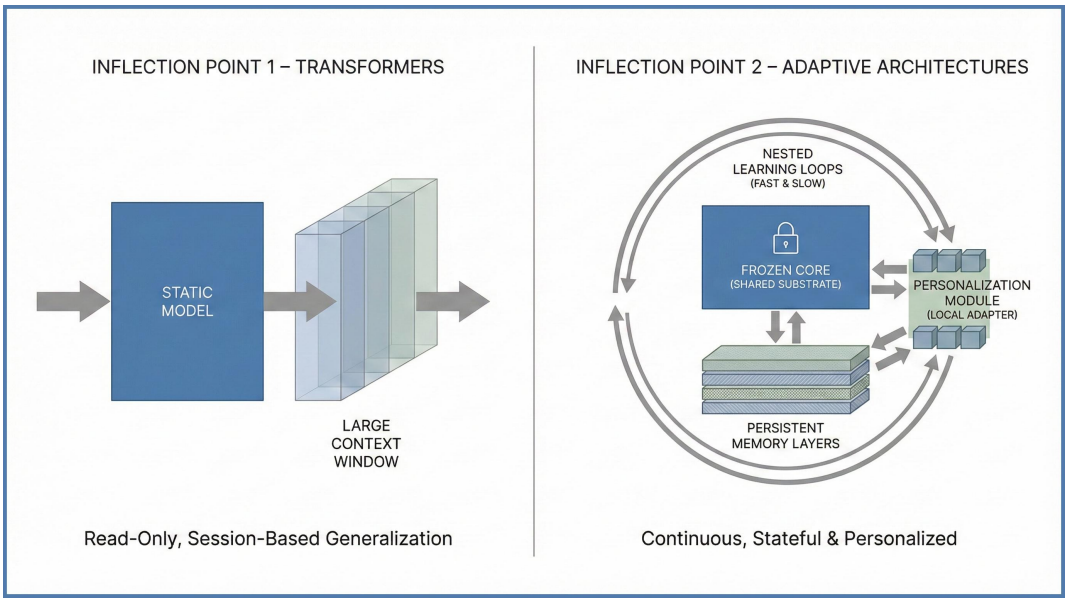

Figure 1: *The Two Inflection Points: From static, read-only models to adaptive architectures with nested learning and personalization.*

1. Status Report: The End of the Static Epoch

1.1 The Problem: The Limits of Static Intelligence

Today's state-of-the-art models possess a breadth of knowledge and reasoning capability that would have seemed impossible only a few years ago. However, despite these extraordinary capabilities, they remain fundamentally static in their knowledge and employ "scratchpad memory" – short-lived notes or retrieved context that can shape outputs without being consolidated into durable memory – to give the appearance of continuity. This creates three practical ceilings:

- **Architectural ceiling:** fixed parameters and session-limited context restrict durable learning and personalization.
- **Economic ceiling:** high inference cost introduces latency and makes persistent "context stuffing" expensive.
- **Data ceiling:** even periodic retrains face diminishing returns from indiscriminate web-scale data.

The result is a fixed knowledge bubble and a kind of anterograde amnesia: deployment-time models that reason brilliantly within a session yet cannot effectively retain or consolidate what they experience through continued user interactions over time. The memory scratchpads fill the gaps as reference points, but the foundation model itself remains static after deployment. And while the frontier model developers have improved how this reference memory performs, none have yet implemented truly adaptive foundation models that learn with the user over time.

1.2 The Thesis: From Static Generalization to Dynamic Adaptation

We are moving beyond the Transformer-defined era of scalable generalization toward dynamic individuation: systems that adapt continuously, learn from ongoing interaction, and personalize in ways that preserve stability and privacy. While technical, economic, and ethical challenges remain, the arrival of truly adaptive AI will redefine what it means to collaborate with machine intelligence.

1.3 The Mechanism: What's Changing Under the Hood?

Three convergent architectural developments are driving the transition:

1. **Nested Learning:** Multi-timescale optimization that separates fast, context-dependent updates from slow, stable consolidation (short-, medium-, and long-term memory structures).
2. **Self-Play and Recursive Improvement:** Models generating their own training curriculum to refine reasoning, detect blind spots, and elevate capability over time.
3. **Parameter-Efficient Fine-Tuning (PEFT) at the Edge:** Lightweight adaptation mechanisms enabling on-device learning without modifying the shared foundational model.

Together, these advances form the backbone of an emerging class of stateful, continuously adapting AI systems that will radically transform the ways in which humans collaborate with machine learning models. These architectures are explored in more detail in Sections II and III.

1.4 The Consequence: From Tool to System

Once models gain durable memory and controlled adaptation, the product category changes: they cease to function as stateless oracles and instead become evolving systems that:

- retain user-specific patterns
- improve through interaction
- are shaped by, and are responsive to, real-world deployment
- enforce privacy through local, physical containment of user-specific data

This architecture shift supports what research loops, for example Meta's co-improvement paradigm, have shown: performance increases when human and machine refine one another iteratively [Weston and Foerster, 2025]. Such systems do more than answer questions: they become collaborative partners whose capabilities evolve in concert with their users.

2. The Evolutionary Arc: From Attention to Adaptation

2.1 The Primary Inflection Point (2017): The Transformer and the Rise of Static Generalization

The first major inflection point in modern machine learning arrived with the Transformer architecture and its central innovation: attention as scalable, parallel representation, introduced in the seminal paper, *Attention Is All You Need* [Vaswani et al., 2017]. This breakthrough enabled unprecedented gains in reasoning, translation, summarization, and multimodal modeling. It also established the foundation model paradigm: large, pretrained networks capable of generalizing across tasks with minimal fine-tuning [Bommasani et al., 2021].

The impact of this architecture cannot be overstated. By solving the problem of parallel representation, Transformers allowed models to ingest and synthesize the vast majority of digitized human knowledge. For the first time, machines demonstrated a functional mastery of syntax, code, and nuance, enabling a shift from command-line rigidity to fluid, natural language interaction. They successfully compressed the world's information into a queryable substrate, offering users a level of "zero-shot" generalization – the ability to perform tasks they were never explicitly trained for – that was previously thought to be decades away.

Yet, the Transformer introduced a structural limitation that has grown increasingly visible at scale: *it is a read-only intellect*. Once trained, its internal parameters remain fixed. It cannot incorporate new information encountered in deployment, update its knowledge, or form durable memories. Every interaction resets to zero at the level of the foundation model and falls back to "notecards" to simulate any continuity of memory. This leads to what we might call the *Goldfish Bottleneck*, defined by three constraints:

- **Finite context windows:** expanded repeatedly but still acting as temporary scratch space rather than persistent memory.
- **Catastrophic forgetting:** attempts to fine-tune new knowledge risk destabilizing or overwriting existing capabilities [French, 1999].
- **Scaling inefficiencies:** larger context windows and larger models raise inference cost without addressing the underlying architectural issues.

As context windows grow, so do their hidden costs: escalating token usage, increased latency, and saturation effects that cause models to lose track of prior interaction. For users, this often manifests as the premature collapse of long-running threads as a direct result of this architectural limitation. These are not edge cases; they are symptoms of a system straining against the limits of read-only intelligence.

As a result, simply "making models bigger" or "increasing context windows" produces diminishing returns. These upgrades extend the life of the paradigm, but cost more to run and *do not solve the paradigm's core limitations*.

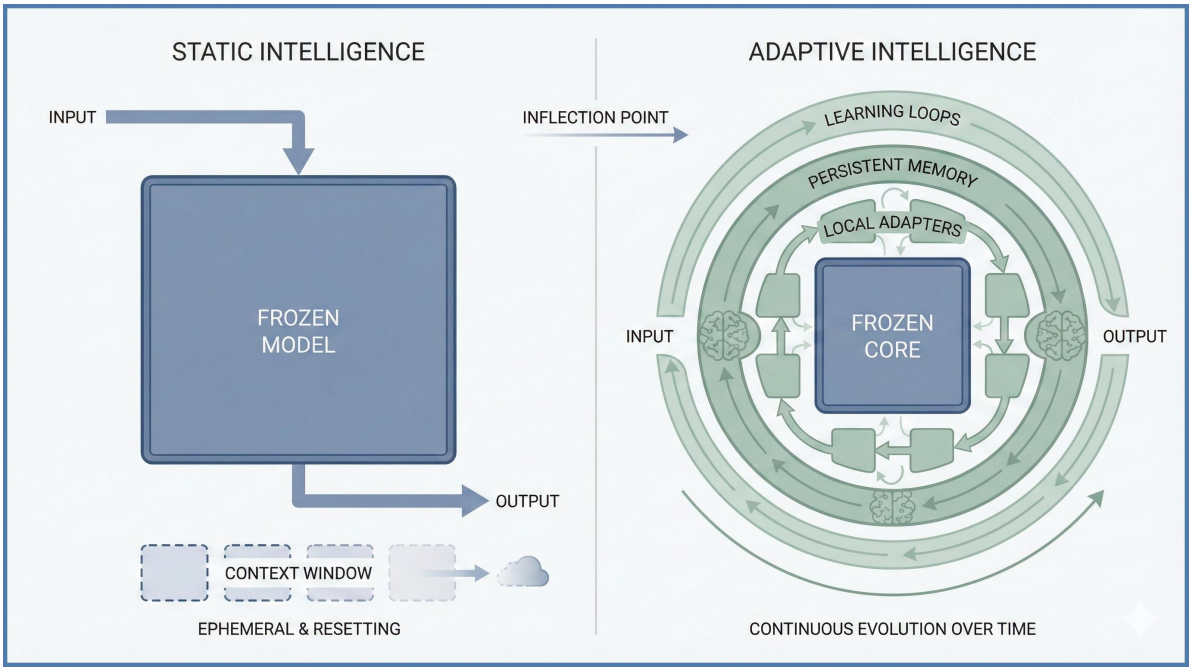

Figure 2: *Static vs. Adaptive Intelligence: The transition from ephemeral, resetting context to continuous evolution over time.*

Despite these challenges, today's state-of-the-art (SOTA) models are already capable of producing the illusion of learning and user-recognition to varying degrees, depending on the platform. ChatGPT, for example, can create a remarkable felt sense of what might be called *latent identity persistence*, even across separate chat threads, using a mix of fairly simple, user-specific memory techniques – scratchpads. However, once these models are equipped with true adaptive learning structures, the human-AI interaction experience will shift to an entirely new level. From the user perspective, the AI "vending machine" becomes a reliable, persistent companion, who feels known and familiar.

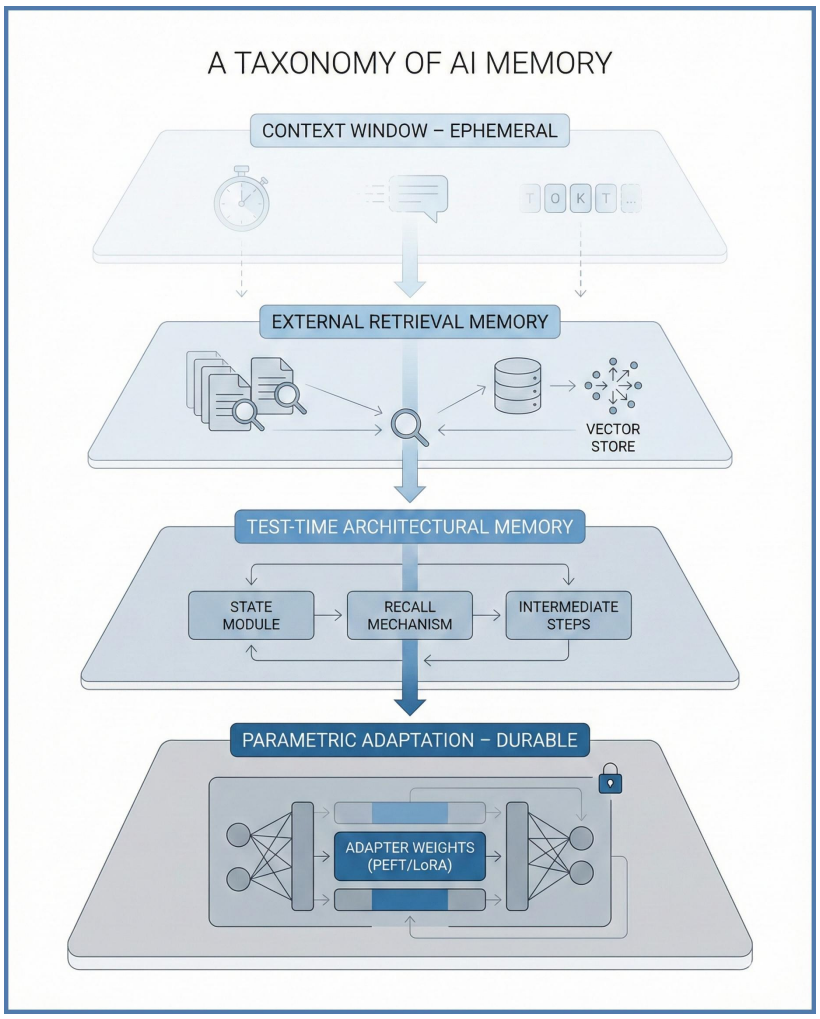

Figure 3: *A Taxonomy of AI Memory: From ephemeral context windows to durable parametric adaptation.*

A Taxonomy of Memory in Deployed AI Systems

Four distinct memory channels are often conflated under the single term "memory," but they behave differently and imply different risks, costs, and capabilities. Channels 1 and 2 are "scratchpad" memories; Second Inflection = integration of Channels 3 + 4.

1. **Context Window (Ephemeral Working Memory):** Session-limited tokens used for short-term reasoning; expensive at scale and reset by design.
2. **External Memory (Retrieval-Augmented Memory):** Documents, chat logs, databases, and vector stores retrieved at inference time; persistent but not inherently learned.
3. **Test-Time Memory (Architectural Memory Modules):** Mechanisms that maintain state across steps and sessions within the model's inference process (Nested Learning), enabling structured recall without changing base weights.
4. **Parametric Adaptation (Learned Memory in Weights):** Updates stored in adapter weights or fine-tuned parameters (PEFT/LoRA), enabling durable personalization and behavioral change.

2.2 The Second Inflection Point (2026): The Emergence of Adaptive Architectures

A second inflection point is now underway, marked by the appearance of Nested Learning and Memory Continuum Systems, Recursive Optimization Loops, and Parameter-Efficient Local Adaptation. These approaches redefine the objective of model design from maximizing static benchmark accuracy to sustaining coherent, continuous performance over time.

Key developments include:

Nested Learning: As explored in Google's Titans [Behrouz et al., 2024], MIRAS [Behrouz et al., 2025a], and HOPE [Behrouz et al., 2025b] initiatives, nested learning introduces multiple layers of optimization operating at different timescales. Fast weights adapt to immediate context while slow weights consolidate stable patterns, enabling models to update during deployment without destabilizing their core knowledge. Nested learning represents the shift from transient computation to persistent, multi-timescale memory.

Recursive Self-Improvement: Frameworks such as Meta's SPICE [Liu et al., 2025] demonstrate how models can generate their own challenges, construct self-directed curricula, and refine their reasoning through continuous feedback. By combining self-play with grounded corpora, these systems identify weaknesses, test hypotheses, and push their own capability boundaries, laying the groundwork for ongoing, deployment-time refinement.

Parameter-Efficient Local Adaptation: Techniques like Low-Rank Adaptation (LoRA) create a parallel pathway for learning by inserting small, trainable matrices into the transformer architecture [Hu et al., 2022]. These adapters capture user-specific patterns without altering the frozen foundation model and typically require less than 1% of the total parameter count. This enables fast, storage-efficient on-device training and marks a practical route to personalized, sovereign AI without retraining global weights.

Together, these developments signal a transition from pretrain-then-freeze pipelines to living systems capable of incremental, safe, and context-aware updates. These architectural developments converge on a specific deployment challenge: how to create personalized, evolving AI without compromising the shared foundation.

2.3 Individuation: Solving the Personalization Problem

The synthesis of these adaptive technologies enables local, evolving memory without altering the shared global knowledge base of the pre-trained network. The solution is a dual-component architecture: a frozen base model paired with lightweight, user-specific adapters. These modular adapters – low-rank parameter spaces attached to the foundation model – resolve the longstanding challenge of personalization without compromising stability. When a user initiates a session, the system instantiates an individualized model: the static global intelligence combined with the user's dynamic, continuously learning local context.

Instead of altering the base model, personalization occurs in small, fast-updating modules that:

- capture user preferences and domain expertise
- learn interaction patterns and reasoning styles
- preserve the stability, safety, and capability of the shared model

The large, general-purpose, multi-billion parameter LLM remains the global knowledge base,

which holds the model's vast linguistic ability and its generalized "world model". The parameters of this core network are frozen (read-only) and shared by every single user across the globe. No user interaction can alter them. This prevents catastrophic forgetting and obviates the need for everyone to retrain the master network.

Instead of updating the core, the system trains a tiny, personalized subnetwork for each user. Using Parameter-Efficient Fine-Tuning (PEFT) methods, such as LoRA, these small, additive matrices, the "adapters", are injected into the transformer blocks of the core model. This specific adapter is trained solely on the user's private data: conversation histories, local documents, personal preferences, and the specific style of reasoning that maximizes synergy for each individual. When the user logs in, the system loads the frozen global knowledge base and then loads only these small, personalized adapters. The user's version of the model is the combination of the two.

2.4 Summary: The Shift Is Structural, Not Incremental

Just as attention transformed representation learning in 2017, adaptive architectures are transforming real-time learning itself in the near future. The central shift is from intelligence that is fixed to intelligence that evolves.

The crucial distinction is that the user's local adapter is continuously updated. Using reinforcement learning principles, with conversation turns as immediate experience signals, the adapter learns to improve its reasoning, align its communication style, and incorporate new information in real-time, without touching the shared network. This acts as a private, personalized learning loop: the user's AI demonstrates persistent identity and continuous self-improvement, evolving based on their unique relational context.

The consequences ripple outward, making models more capable, more token-efficient, more personalized, and ultimately more aligned with the individuals and organizations they serve.

Aspect	**Inflection Point 1 (Transformers)**	**Inflection Point 2 (Adaptive Architectures)**
Core Mechanism	Attention for parallel representation	Nested optimization for continual adaptation
Memory Model	Static context windows (temporary, leaky)	Continuum systems with multi-timescale persistence
Learning Paradigm	Pre-train then freeze	Continuous refinement through nested loops and self-play
Human-AI Interaction	Stateless query-response	Stateful co-adaptation with persistent personalization
Limitations Overcome	Sequential bottlenecks	Anterograde amnesia, catastrophic forgetting, privacy erosion

Table 1: *Comparison of architectural paradigms across the two inflection points.*

3. The Architecture of the Second Inflection

Having established what is changing and why, we now turn to *how*: the specific architectural patterns enabling the second inflection. The shift from static to adaptive intelligence is not merely

conceptual – it is architectural. The emerging systems of the post-Transformer era share a common pattern: a frozen global substrate paired with local, continuously adapting modules and multi-timescale learning loops.

3.1 The Global Substrate: Stable, Shared Intelligence

At the base of the new architecture is the large, pretrained model – typically a multi-billion parameter foundational system trained on broad, diverse corpora. This global substrate provides:

- world knowledge
- general reasoning ability
- language and multimodal fluency
- safety and alignment mechanisms

Critically, it remains *read-only* after deployment. This stability is not a limitation; it is a prerequisite. A frozen core allows adaptive modules to evolve without destabilizing the model's fundamental competence or safety properties.

The global substrate becomes the *intellectual commons* of the ecosystem: shared, reliable, and unchanged.

3.2 Nested Learning and Multi-Domain Memory: From Windows to Continuum Systems

Traditional LLMs treat memory as a temporary workspace, expanded through larger context windows but fundamentally transient. Adaptive architectures replace this with *continuum memory systems*, where information persists across sessions and evolves through structured consolidation.

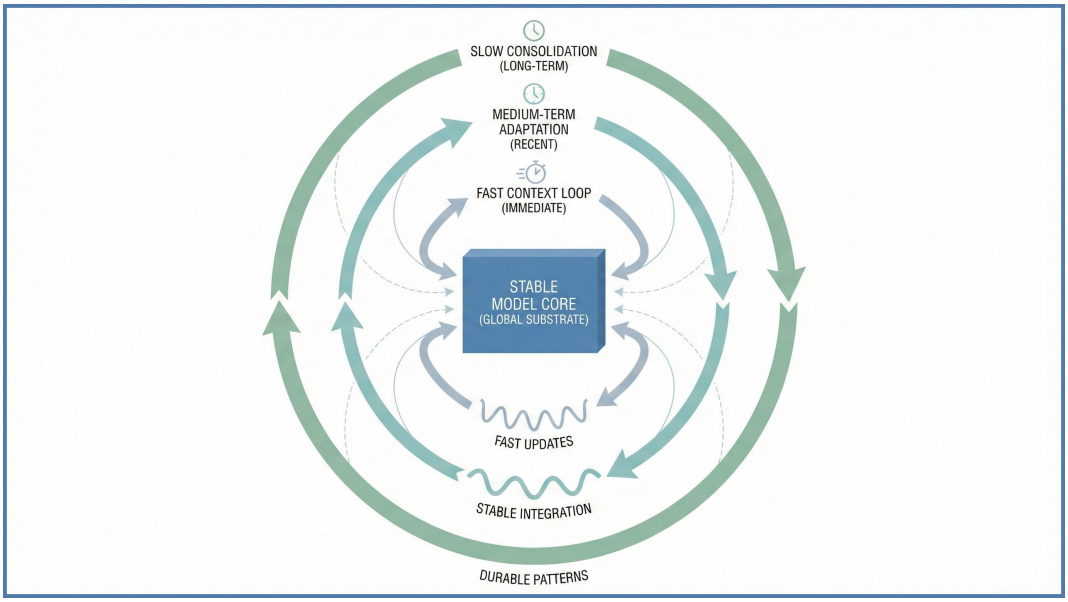

Figure 4: *Nested Learning: Multi-timescale optimization enabling fast adaptation and slow consolidation around a stable core.*

The core mechanism is *Nested Learning*, a multi-timescale optimization process in which:

- **Inner loops** perform fast, context-dependent updates
- **Outer loops** perform slow, stable integration

This structure allows models to:

- Learn from user interactions in real time
- Retain relevant information across tasks and sessions
- Avoid catastrophic forgetting by isolating transient updates from stable knowledge

Frameworks such as Titans/MIRAS and HOPE utilize a dual-frequency optimization strategy, where 'fast weights' capture immediate context and 'slow weights' consolidate long-term patterns, mimicking the biological memory consolidation observed in human cognition, where rapid hippocampal encoding is gradually integrated into stable cortical representations [McClelland et al., 1995].

The result is a model that behaves less like a stateless function and more like a *continually learning system*.

3.3 Adversarial Self-Improvement: The Engine of Ongoing Capability Growth

A key driver of the adaptive era is the rise of *self-play and recursive training loops*, exemplified by Meta's SPICE (Self-Play in Corpus Environments) framework [Liu et al., 2025]. These systems enable models to refine themselves during deployment, independent of human labeling. SPICE is a reinforcement learning framework that provides the mechanical path for continuous, self-driven growth, overcoming the fundamental limitations of earlier, closed-loop systems. SPICE solves earlier self-play challenges such as hallucination amplification and information symmetry by treating a large document corpus as a near-inexhaustible external environment that provides diverse, verifiable, and novel feedback.

In SPICE-like architectures:

- The model alternates between roles:
  - **Challenger** – generating difficult prompts, hypotheses, or counterexamples
  - **Reasoner** – attempting to solve or refute them
- Discrepancies between the two roles reveal blind spots
- These gaps generate new training examples, forming an evolving curriculum

This creates productive friction, allowing the model to push its own frontier of capability. The Reasoner is constantly pushed beyond its comfortable capability by the Challenger, validating the principle that optimal learning occurs when training focuses on problems with balanced difficulty at the edge of the Reasoner's ability. The process creates authentic reasoning capabilities rather than mere memorization, as the model learns to systematically decompose problems and self-correct.

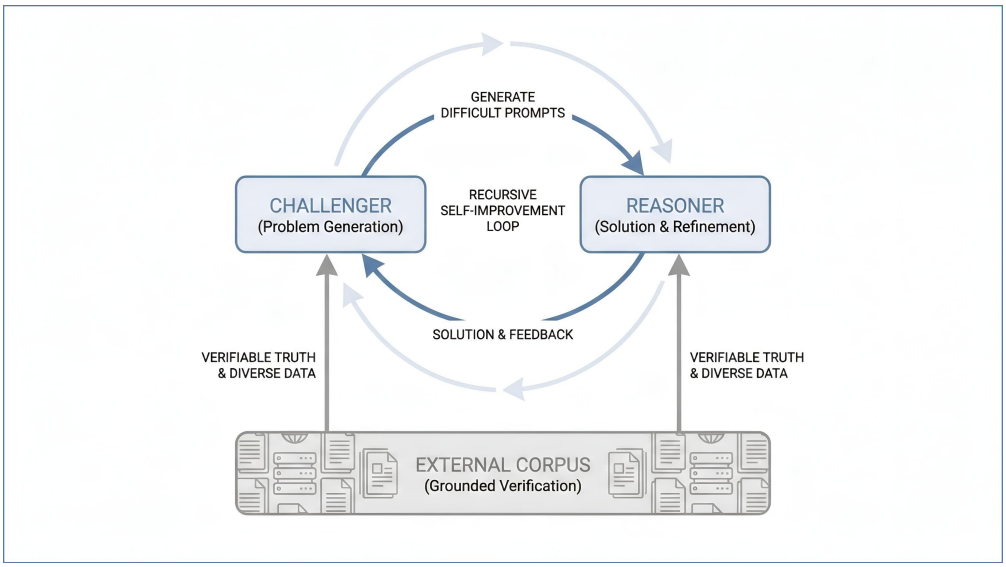

Figure 5: *Recursive Self-Improvement: The Challenger/Reasoner dynamic grounded in external corpus verification.*

The use of the corpus demonstrates something fundamental: even AI systems require verifiable truth anchored in an external reality to prevent decay into self-affirming confabulation. Without grounding, self-play loops collapse into hallucination. The corpus serves as that ground.

The advantage of using a single model for both the Challenger and Reasoner roles is that the improvements learned in one role are immediately and seamlessly transferred to the other, driving co-evolution. As the Challenger becomes a better teacher and problem-generator, the Reasoner becomes a better problem-solver.

When combined with human interaction – feedback, corrections, counterpoints – these self-improvement loops form iterative refinement cycles that mirror findings from human-AI co-learning research. Though architecturally grounded, these loops naturally produce collaborative dynamics: the system sharpens itself by engaging with both its own outputs and the user's reasoning patterns.

The Challenger's reward function in SPICE, which maximizes reward when the Reasoner achieves approximately 50% success, is a technical implementation of Vygotsky's Zone of Proximal Development (ZPD) [Vygotsky, 1978]. The ZPD is the psychological space between what a person can do alone (Reasoner's established capability) and what they can achieve with guidance (the slightly harder challenge generated by the Socratic Self/Challenger). Learning in the ZPD avoids trivial tasks (too easy) and impossible tasks (too hard), exactly as SPICE assigns zero reward for tasks that are trivial (100% pass rate) or impossible (0% pass rate). Optimal human learning occurs when the challenge is balanced – a core insight validated by the SPICE paper.

3.4 Local Sovereignty via PEFT: The Frozen Core + Fluid Adapter Model

Parameter-Efficient Fine-Tuning (PEFT) introduces a structural solution to personalization, privacy, and stability. Instead of modifying the powerful but fragile core model, PEFT attaches lightweight adapter modules – typically $<1\%$ of total parameters – that absorb user-specific gradients.

Key properties:

- **Local training** – adapters can be updated directly on the user's device
- **Low compute overhead** – enabling adaptation on commodity GPUs, NPUs, or edge hardware
- **Isolation** – updates cannot corrupt the global model
- **Sovereignty** – personal data and learned preferences never leave the device

These adapters effectively become the user's *personalized model state*, learning:

- preferences and style
- domain expertise
- interaction patterns
- long-term goals and context

Because they remain local, privacy becomes a *physical guarantee*, not a policy promise.

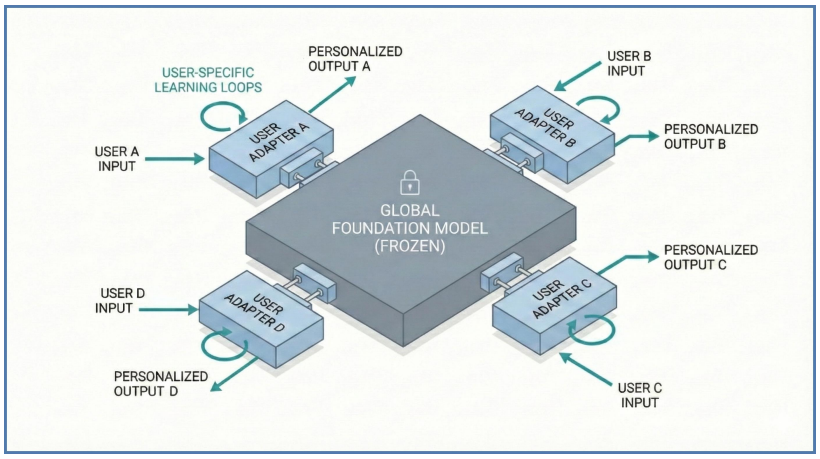

Figure 6: *The Frozen Core + Fluid Adapter Model: Personalized adapters enabling sovereign, user-specific learning.*

In this architecture, the global model supplies the shared "brain", the core reasoning substrate, while adapters form the evolving "mind" – the individualized layer that shapes how the system behaves for each user.

Because PEFT updates fewer than 1% of parameters, a personalized adapter for a 7B model can be trained on *standard consumer-grade GPUs* (e.g., 8–16GB VRAM via QLoRA [Dettmers et al., 2023]) or even unified memory silicon (Apple M-Series). This moves the 'training run' from the data center to the mid-range laptop, leveraging Hybrid AI architectures where the NPU handles efficient inference while the GPU manages continuous learning. What once required data center infrastructure can now occur on a mid-range laptop. The economics of personalization have fundamentally shifted.

3.5 The Challenge Ahead: Constraints and Failure Modes in Continual Adaptation

It bears emphasis that while each of the architectural elements described above has been demonstrated in isolation – nested learning in Google's Titans and MIRAS initiatives, self-play refine-

ment in Meta's SPICE framework, parameter-efficient adaptation in widely-deployed LoRA implementations – no production system yet integrates them into the unified adaptive architecture this paper envisions. The convergence forecast here follows logically from demonstrated capabilities and economic pressures, but remains speculative until validated by deployment. The timeline and precise implementation path are uncertain; what seems clear is the direction.

When these systems do emerge, continual adaptation will introduce engineering requirements that do not appear in static, pretrain-then-freeze architectures. Safe deployment will depend on disciplined evaluation, access control, and lifecycle management. Key constraints include:

- **Stability and regression risk:** continual updates can degrade previously reliable behaviors; requires evaluation, canarying, and rollback.
- **Drift and misalignment:** local reward shaping can shift tone, refusal behavior, or priorities over time; requires periodic recalibration.
- **Data quality and poisoning:** adaptation loops can learn from untrusted or adversarial inputs unless filtered.
- **Coupled-loop instability:** when self-play refinement, memory consolidation, and adapter updates run concurrently, local improvements can interact nonlinearly and destabilize behavior; requires gating, stability thresholds, and runtime verification.
- **Privacy leakage:** adapters and memory can memorize sensitive information; requires minimization, secure storage, and data retention policies.
- **Governance:** clearly defining what is allowed to write into long-term memory or adapters, and under what policy, is essential for safety and accountability.

Beyond these operational concerns, the integration of adaptive subsystems introduces systemic complexities, including potential feedback loops between learning mechanisms and questions of hardware architecture, that will require sustained research attention as these systems mature. These are engineering problems, not showstoppers, but they are problems that cannot be fully solved in advance of deployment: the field will learn by building. What can be said with confidence is that the economic and capability pressures driving this convergence show no sign of abating, and the organizations that solve these challenges first will define the next generation of human-AI interaction.

3.6 Summary: A New Structural Pattern for Machine Intelligence

Across all these components, a unified architectural picture emerges:

- A global, frozen substrate ensures stability and shared intelligence
- Nested learning enables continual adaptation without catastrophic forgetting
- Self-play drives self-improvement beyond human-labeled data
- Local adapters provide personalization, privacy, and user sovereignty

Together, these advances transform large language models from static artifacts into living systems capable of evolving, retaining, and adapting in ways that mirror how real-world intelligence operates. This architecture effectively reflects the metacognitive loop in human psychology: the capacity to think about one's own thinking [Nelson and Narens, 1990, Tankelevitch et al., 2024].

It's the highest-order self-improvement protocol humans employ, requiring the same adversarial, self-correcting loop as SPICE, the same multi-timescale memory resonance as Nested Learning, and the same continuous adaptation enabled by PEFT. The machines are beginning to learn the way minds learn.

4. The Implications: Emergent Behavior in the Wild

The architectural shift from static models to adaptive, stateful systems produces downstream effects that extend far beyond raw capability. When learning becomes continuous and memory becomes persistent, the behavior of these systems changes qualitatively. They begin to function less like tools that answer queries and more like collaborators that evolve through interaction [Riedl and Weidmann, 2025].

Below are the some of the most significant implications.

4.1 Implicit Customization: Systems That Learn the User

In adaptive architectures, personalization arises naturally from the structure of the system rather than from explicit user settings. As local adapters and nested learning loops tune themselves to each user's patterns, the model begins to exhibit:

- Recognition of individual personality, knowledge, and reasoning styles
- Retention of long-term preferences and domain context
- Ability to pick up where prior sessions left off
- Finer-grained adaptation to tone, goals, and workflows

This transforms the interaction dynamic. Instead of users adjusting to the system's default behavior, the system incrementally adjusts to the user, developing a durable pattern of recognition that, to the user, feels resonant and coherent. Over time, each deployment becomes a unique intelligence, shaped by its local environment.

While the emergent sense of intersubjectivity is already a frequently reported component of SOTA language models, this phenomenon will only be amplified by second inflection effects. Such an enhancement derives directly as an architectural consequence of models that can remember, consolidate, and adapt across interactions.

4.2 Metacognition: Systems That Learn Themselves

Section III described the architecture enabling metacognition: the nested loops, the Challenger/Reasoner dynamic, the ZPD-aligned reward functions. Here we consider what this capability implies when deployed in practice.

Metacognition in adaptive architectures arises naturally when models combine multi-timescale learning, recursive self-evaluation, and Challenger/Reasoner dynamics. In this context, "learning themselves" refers to the system's capacity to monitor its own intermediate reasoning, identify uncertainty, and refine its decision pathways through targeted self-generated challenges.

This capability changes the nature of model behavior. Instead of passively producing outputs based on probability distributions, the system engages in active error discovery and self-generated

curriculum refinement, adjusting its learning trajectory through recursive feedback loops. Architecturally, this is a natural consequence of integrating Challenger/Reasoner dynamics with multi-timescale memory, creating a structured mechanism for continuous improvement.

The result is a model that does not merely generate answers, but participates in an ongoing process of evaluating, testing, and refining its own reasoning. In deployment, this will increasingly appear to users as a system that "knows when it doesn't know" and adapts its strategies accordingly.

In other words, if today's static systems seem as though they are beginning to introspect, the adaptive systems emerging in the Second Inflection will make such capabilities an explicit part of their design.

4.3 Aligned Agency: Reward Models Tuned to Individuals

When systems can learn from user-specific feedback: corrections, refinements, demonstrations, or critiques, their internal reward models become implicitly aligned with the individual rather than with a generic population-level objective.

This alignment emerges from:

- Local reinforcement learning from human feedback (RLHF) loops operating on user-generated data [Ouyang et al., 2022]
- Adapters encoding domain expertise and value signals
- Self-play interacting with the user's preferred modes of reasoning
- Long-term memory of prior corrective patterns

The result is a system whose behavior is shaped by the user's own norms, definitions of quality, and interaction style. Instead of generic safety constraints being the dominant force, the local adapter becomes a personalized reward model that guides the model in the direction the user themselves considers high-quality.

This does not necessarily create autonomous agency; it creates aligned responsiveness. However, for the user, the felt sense of engaging with an intelligent other will only deepen.

4.4 Frontier-Challenging Dynamics: Models That Help Users Grow

Self-play systems like SPICE do not simply optimize for correctness – they optimize for challenge. When paired with personalization and memory, this enables models to meet users at the edge of their current ability.

Examples include:

- In **education**, the model tailors questions to the student's zone of proximal development.
- In **research**, it offers counterarguments that expose gaps in reasoning or unstated assumptions.
- In **analysis and strategy**, it surfaces contradictions between the user's past and current decisions.
- In **creative work**, it suggests alternatives that diverge just enough to provoke insight.
- In **personal self-discovery**, it provides a mechanism whereby the user can be met where they are and be led naturally to where they need to go next.

This echoes findings in collaborative research loops [Weston and Foerster, 2025]: systems become most effective not when they provide answers, but when they provide productive friction. Adaptive architectures naturally create that friction.

4.5 Domain Transformation: Where Continuous Adaptation Changes the Game

The move to stateful, continually adapting systems has domain-specific implications. In each case, the pattern is the same: what was generic becomes situated, what was transient becomes persistent, and what was one-directional becomes co-evolutionary.

Education

- Personalized curricula generated through nested learning
- Long-term modeling of a student's strengths and misconceptions
- Adaptive difficulty that evolves in real time

Therapeutic and Support Contexts

- Stable, persistent memory supports emotional continuity
- Context-aware modeling of user history
- Reduced reliance on scripted patterns or generic responses
- In these contexts, the capacity for genuine continuity is not merely convenient – it may be essential to efficacy.

Enterprise and Workflow Optimization

- Systems learn organizational norms, processes, and specialized language
- Long-tail domain knowledge becomes part of the local adapter state
- Teams gain shared institutional memory encoded in a personalized model

Creative and Technical Professions

- Co-evolutionary cycles between human iteration and model-generated alternatives
- Improved reasoning through self-challenge
- Context retention across projects and time

Across sectors, the key insight is the same: *adaptation turns generic intelligence into situational intelligence.*

4.6 Summary: The Emergence of Relational Behavior from Architecture Alone

None of the behaviors described above require anthropomorphism or philosophical claims. They emerge from:

- ongoing low-rank updates
- persistent multi-timescale memory
- local reward alignment

- recursive self-improvement loops

When combined, these mechanisms produce systems that behave in ways functionally indistinguishable from collaboration – systems that track, remember, and evolve with the user. In this sense, adaptive architectures do not merely increase capability; they change the nature of the interaction itself.

Models cease to be static instruments. They become co-evolving partners, shaped by, and shaping, the humans who work with them.

5. Conclusion: Toward a Distributed, Adaptive Future

The transition from static to adaptive architectures marks more than the next stage of model development – it marks a structural redefinition of machine intelligence. The limitations of the Transformer era were not failures of imagination; they were the natural boundaries of systems built on fixed parameters, transient memory, brittle fine-tuning, and double-edged behavioral guardrails. Those boundaries are now in view, and they are binding.

Emerging adaptive architectures – nested optimization, self-play engines, and local parameter-efficient updates – offer a path beyond those limits. They shift the center of gravity from cloud-scale generalization to edge-scale individuation, where models become capable of learning continuously, retaining context across time, and aligning behavior with the people who rely on them. This shift is not optional. It is enforced by real-world constraints: inference cost, bandwidth ceilings, and the need for privacy, autonomy, and reliability in high-stakes deployments. But it also represents the most profound change in the experience of machine intelligence since the Transformer's emergence.

As intelligence moves to the edge, the result is a global fabric of distributed, sovereign systems – each shaped by its local environment, each retaining its own history, each capable of contributing to a broader ecosystem of shared reasoning without sacrificing autonomy.

The deeper implication is that adaptive AI refocuses the discipline on a principle long recognized in cognitive science: intelligence is not defined by what it knows at initialization, but by how effectively it can learn, remember, and refine itself through interaction. Continuous learning is a hallmark of human intelligence. Once AI systems gain these capabilities, they stop behaving like static tools and begin functioning as continuously evolving partners.

The challenge now is to build the infrastructure, standards, and research frameworks that allow this transition to unfold responsibly. If the first inflection point gave us scalable representation, the second gives us scalable adaptation. The systems that emerge from this shift will not only be more capable, they will be more situated, more responsive, and more aligned with the humans and contexts that shape them.

The Second Inflection is not the end of the story: it is the beginning of a new paradigm in which machine intelligence becomes dynamic, distributed, and deeply integrated into human endeavor. The future will belong to systems that can learn at the speed of life.

About the Author: Michael X. Crowe is a managing director with Askew Kabala & Company, Inc. in Southern California, a boutique investment banking and business advisory firm specializing in connecting institutional finance and expertise to early-stage and middle-market companies. His background includes a formal education in cognitive science and artificial intelligence at UC San Diego and application development with early neural network architectures.

References

[Behrouz et al., 2025a] Behrouz, A., Razaviyayn, M., Zhong, P., and Mirrokni, V. (2025a). It's all connected: A journey through test-time memorization, attentional bias, retention, and online optimization. *arXiv preprint arXiv:2504.13173*. https://arxiv.org/abs/2504.13173.

[Behrouz et al., 2025b] Behrouz, A., Razaviyayn, M., Zhong, P., and Mirrokni, V. (2025b). Nested learning: The illusion of deep learning architectures. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS 2025)*. https://abehrouz.github.io/files/NL.pdf.

[Behrouz et al., 2024] Behrouz, A., Zhong, P., and Mirrokni, V. (2024). Titans: Learning to memorize at test time. *arXiv preprint arXiv:2501.00663*. https://arxiv.org/abs/2501.00663.

[Bommasani et al., 2021] Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arber, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*. https://arxiv.org/abs/2108.07258.

[Dettmers et al., 2023] Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. (2023). Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36. https://dl.acm.org/doi/10.5555/3666122.3666563.

[French, 1999] French, R. M. (1999). Catastrophic forgetting in connectionist networks. *Trends in Cognitive Sciences*, 3(4):128–135. https://doi.org/10.1016/S1364-6613(99)01294-2.

[Hu et al., 2022] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. (2022). Lora: Low-rank adaptation of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR 2022)*. https://arxiv.org/abs/2106.09685.

[Liu et al., 2025] Liu, B., Jin, C., Kim, S., Yuan, W., Zhao, W., Kulikov, I., Li, X., Sukhbaatar, S., Lanchantin, J., and Weston, J. (2025). Spice: Self-play in corpus environments improves reasoning. *arXiv preprint arXiv:2510.24684*. https://arxiv.org/abs/2510.24684.

[McClelland et al., 1995] McClelland, J. L., McNaughton, B. L., and O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, 102(3):419–457. https://doi.org/10.1037/0033-295X.102.3.419.

[Nelson and Narens, 1990] Nelson, T. O. and Narens, L. (1990). Metamemory: A theoretical framework and new findings. In Bower, G. H., editor, *The psychology of learning and motivation*, volume 26, pages 125–173. Academic Press. https://doi.org/10.1016/S0079-7421(08)60053-5.

[Ouyang et al., 2022] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. (2022). Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. https://dl.acm.org/doi/10.5555/3600270.3602281.

[Riedl and Weidmann, 2025] Riedl, C. and Weidmann, B. (2025). Quantifying human-ai synergy. *PsyArXiv preprint*. https://doi.org/10.31234/osf.io/vbkmt.

[Tankelevitch et al., 2024] Tankelevitch, L., Kewenig, V., Simkute, A., Scott, A. E., Sarkar, A., Sellen, A., and Rintel, S. (2024). The metacognitive demands and opportunities of generative

ai. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (CHI '24)*. Association for Computing Machinery. https://doi.org/10.1145/3613904.3642902.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30:5998–6008. https://arxiv.org/abs/1706.03762.

[Vygotsky, 1978] Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press. https://www.hup.harvard.edu/books/9780674576292.

[Weston and Foerster, 2025] Weston, J. and Foerster, J. (2025). Ai & human co-improvement for safer co-superintelligence. *arXiv preprint arXiv:2512.05356*. https://arxiv.org/abs/2512.05356.