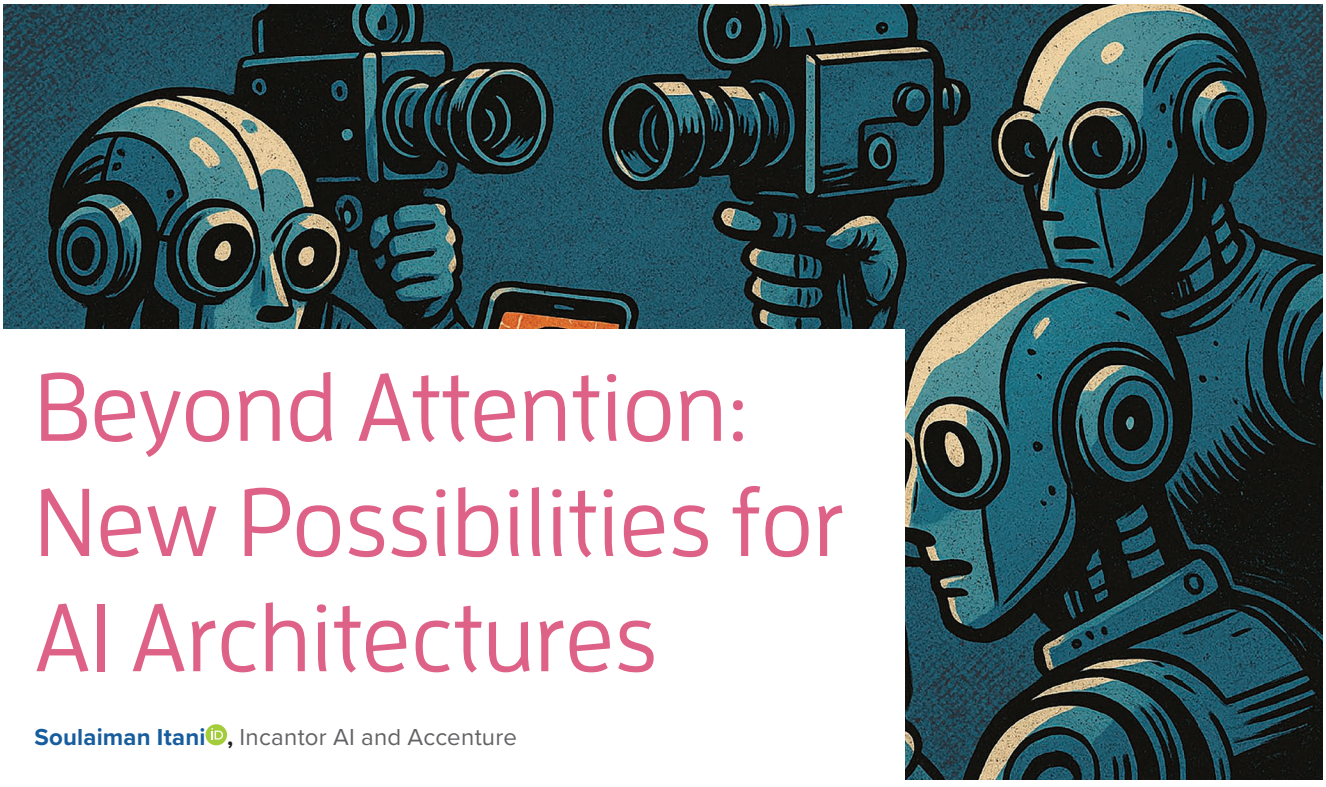

GRAPHICS GENERATED USING CHATGPT 5.0

Beyond Attention: New Possibilities for AI Architectures

Soulaiman Itani, Incantor AI and Accenture

As limitations and scaling challenges associated with large language models emerge, this article highlights several promising alternatives—including Cantor, a light fractal model currently being developed by the author.

By the beginning of 2026, global investment in artificial intelligence is estimated to have reached US$1.5 trillion.[1] Few developments in the history of computing have had as profound an impact as the 2017 paper "Attention Is All You Need,"[2] which introduced transformer architecture. The transformer's attention mechanism now forms the backbone of modern large language models (LLMs), such as ChatGPT, Claude, Gemini, and others—driving advances that have transformed research, industry, and the world's economy.

The transformer architecture represents one of the most consequential breakthroughs in the history of machine learning. The work of "Attention" authors Vaswani et al.[2] and the broader research community that expanded upon it has fundamentally shaped the field of artificial intelligence. Its conceptual clarity, mathematical elegance, and scalability established a new foundation for sequence modeling and generative systems—an achievement that continues to underpin state-of-the-art advancements across language, vision, and multimodal domains.

Yet, as the field matures, the scientific community has begun to explore architectures that move beyond transformers' limitations. Scaling attention mechanisms indefinitely may not yield proportionally greater reasoning, memory, or abstraction. As the cost of data and computation increases exponentially, new paradigms—more efficient, structured, and biologically inspired—are emerging.

This article surveys several of the most promising alternatives to transformers and introduces the author's Cantor, a light fractal model (LFM), as one example of how hierarchical and recursive computation can complement attention-based systems.

Digital Object Identifier 10.1109/MC.2025.3624444
Date of current version: 22 December 2025

EDITOR **MICHAEL ZYDA**
University of Southern California;
zyda@mikezyda.com

THE TRANSFORMER PARADIGM: STRENGTHS AND CONSTRAINTS

Transformers represent text input sequences as tokens, typically subword units, that are embedded into dense vector spaces. The architecture's core mechanism, self-attention, allows every token to consider every other token in the sequence through learned query (Q), key (K), and value (V) projections:

- Q vectors act as information seekers, identifying which other elements are relevant to the current representation.
- K vectors define the latent space where relevance is computed.
- V vectors carry the contextual information passed between tokens.

This architecture enables global context awareness and parallel computation, leading to major performance improvements over recurrent and convolutional predecessors. Transformers have achieved state-of-the-art results across natural language processing, computer vision, speech recognition, and multimodal reasoning.

However, several structural limitations have become increasingly evident:

1. *Quadratic complexity*: The self-attention mechanism scales as $O(n^2)$ with sequence length n, making long-context modeling expensive.
2. *Resource demands*: Expanding model size to achieve incremental gains requires enormous compute, energy, and data resources.
3. *Limited reasoning depth*: Transformers excel at pattern completion but struggle with hierarchical reasoning, abstraction, and symbolic composition.
4. *Context saturation*: Performance deteriorates once sequences exceed a model's effective attention window.
5. *Ephemeral memory*: Transformers lack a native separation between working and long-term memory; each input must be reprocessed in isolation.

These constraints have motivated researchers to seek architectures that can capture long-range dependencies, maintain persistent memory, and operate with greater efficiency.

- *Linear-time scalability*: SSMs achieve O(n) complexity in sequence length, eliminating the quadratic bottleneck of attention.
- *Content-aware transitions*: In models such as Mamba, input data modulates transition parameters, blending dynamical systems with contextual adaptivity.
- *Hardware-optimized parallelization*: Efficient scan algorithms restore the parallelism that earlier recurrent models lacked.

Empirical studies show[3] that Mamba-style architectures can outperform similarly sized transformers on language modeling tasks while operating with significantly lower computational cost [see Figure 2(a) and 2(b)]. This direction illustrates that sequence modeling can emerge from principles of signal propagation and control theory, not solely attention.

> Transformers have achieved state-of-the-art results across natural language processing, computer vision, speech recognition, and multimodal reasoning.

EMERGING DIRECTIONS BEYOND TRANSFORMERS

Several promising research directions are redefining what next-generation neural architectures could look like. While attention remains a powerful mechanism, it is no longer the sole foundation for modeling complex systems. Three representative paradigms illustrate the breadth of this exploration.

Selective and structured state space models

State space models (SSMs) treat sequence processing as a dynamic system evolving through continuous or discrete states rather than pairwise attention.

Recent work—most notably the Mamba architecture[3]—extends this idea through selective state transitions, allowing the model to dynamically adjust its internal state in response to input content (see Figure 1).

Key innovations include:

Implicit and gated convolutional operators

Another direction approximates long-range dependencies using implicit convolutions or gating mechanisms.

The Hyena architecture[4] exemplifies this approach: it replaces explicit attention with hierarchical, multiplicatively gated convolutions that achieve subquadratic scaling (O(n\log n)) while maintaining long-range sensitivity (see Figure 3).

These models view the attention matrix as a learned filtering process,

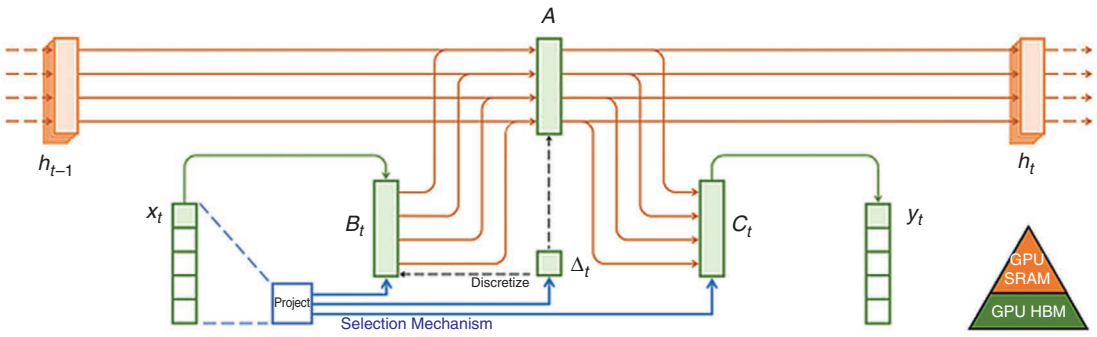

FIGURE 1. (Overview.) Structured selective state space models (SSMs) independently map each channel (for example, $D = 5$) of an input x to output y through a higher dimensional latent state h (for example $N = 4$). Prior SSMs avoid materializing this large effective state (DN, times batch size B and sequence length L) through clever alternate computation paths requiring time invariance: the (Δ, A, B, C) parameters are constant across time. Our selection mechanism adds back input-dependent dynamics, which also requires a careful hardware-aware algorithm to only materialize the expanded states in more efficient levels of the GPU hierarchy. (This is an image of the selective state space model with hardware-aware state expansion. Source: Gu and Dao[3]; used with permission.)

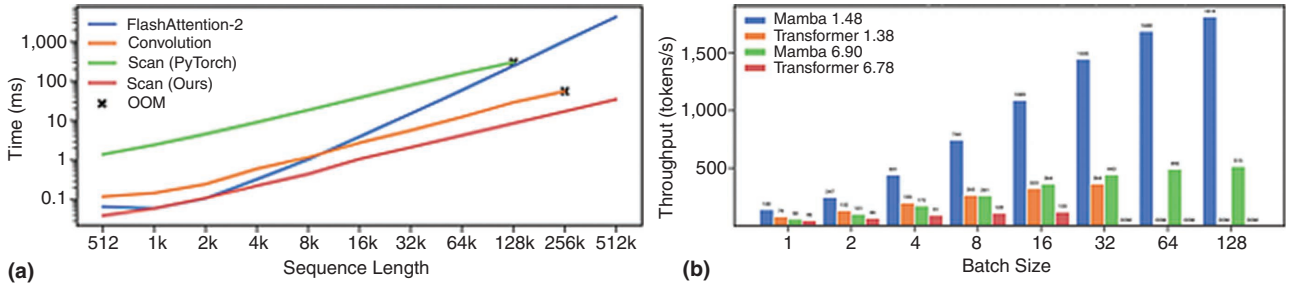

FIGURE 2. Efficiency benchmarks. (a) Scan versus convolution time versus attention (A100 80GB PCIe). Training: Our efficient scan is 40x faster than a standard implementation. (b) (These are images of the selective SSM. Source: Gu and Dao[3]; used with permission.)

enabling sparse or frequency-domain representations of context.

They are particularly efficient for modalities such as audio, video, and very long text, where locality and recurrence are natural.

Concept-level and hierarchical embedding models

Meta's large concept models (LCMs) propose an even more abstract perspective: operating directly in a concept space rather than at the token level.[5]

Instead of predicting individual words, LCMs represent and reason over higher-level semantic units—sentences, ideas, or multimodal concepts.

This paradigm introduces three potential advantages:

- *Cross-modal generality*: Concept embeddings unify textual, visual, and auditory information within a single latent structure.
- *Language independence*: Reasoning occurs at a semantic rather than syntactic level, enabling multilingual and cross-domain generalization.
- *Reduced sequence fragmentation*: Operating above the token level decreases the number of computational steps required for global reasoning.

Although LCMs are in early stages, they suggest that reasoning in concept space may overcome some of the contextual fragility inherent in token-based models.

FRACTAL AND HIERARCHICAL ARCHITECTURES: THE CANTOR LIGHT FRACTAL MODEL

Within this landscape of alternatives, fractal and hierarchical architectures present another line of inquiry. Developed after the author's tenure as AI lead at Google, the Cantor light fractal model (LFM) shown in Figure 4 extends the principles of multiscale representation by processing information concurrently across multiple abstraction levels.

Rather than operating solely at the token level, Cantor integrates understanding across words, phrases, sentences, and documents—analogous to how human cognition synthesizes meaning at several scales simultaneously.

Multilevel embedding structure

Cantor maintains concurrent embedding layers at distinct hierarchical levels (see Table 1).

Each higher level distills and contextualizes the information below it, while also influencing interpretation at finer scales through top-down feedback.

Fractal projection and feedback

The model employs recursive fractal projections:

- *Upward projection*: Lower-level representations are summarized into higher-order abstractions.
- *Downward conditioning*: Global context guides fine-grained interpretation, preserving coherence over long spans.
- *Residual pathways*: Detail not captured at higher levels is retained locally, ensuring precision.

This bidirectional recursion enables a balanced combination of global understanding and local fidelity.

Memory anticipation and sparse access

Cantor introduces a dedicated memory anticipation module designed to maintain long-term, associative memory independent of active processing.

Key properties include:

- separation of working and long-term memory for modular scalability

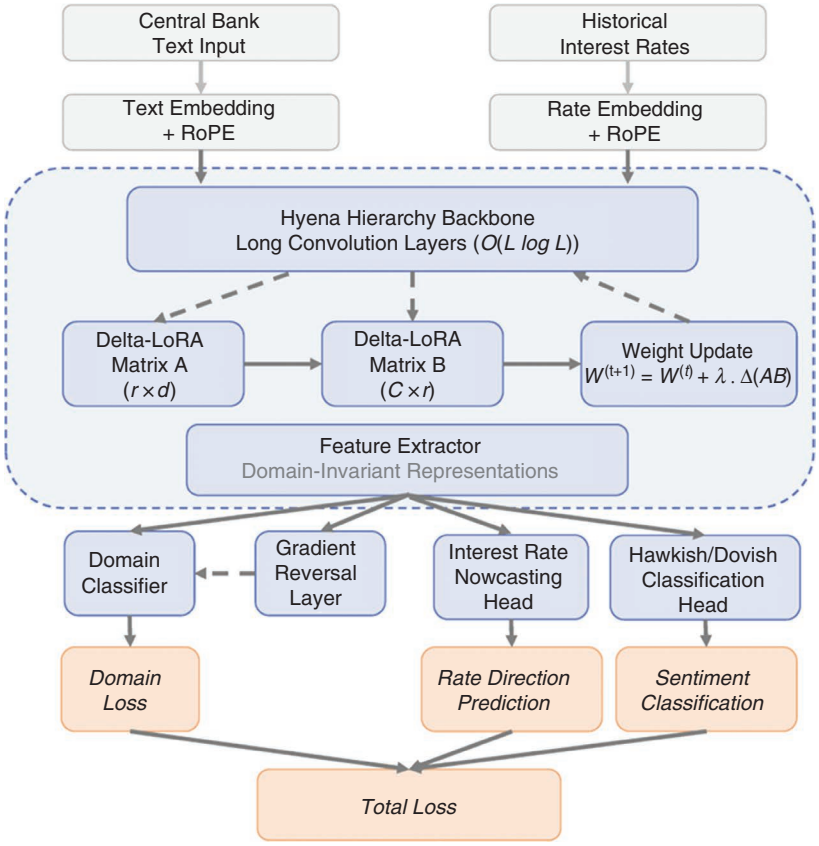

FIGURE 3. The Hyena model. (Source: Poli et al.[4]; used with permission.)

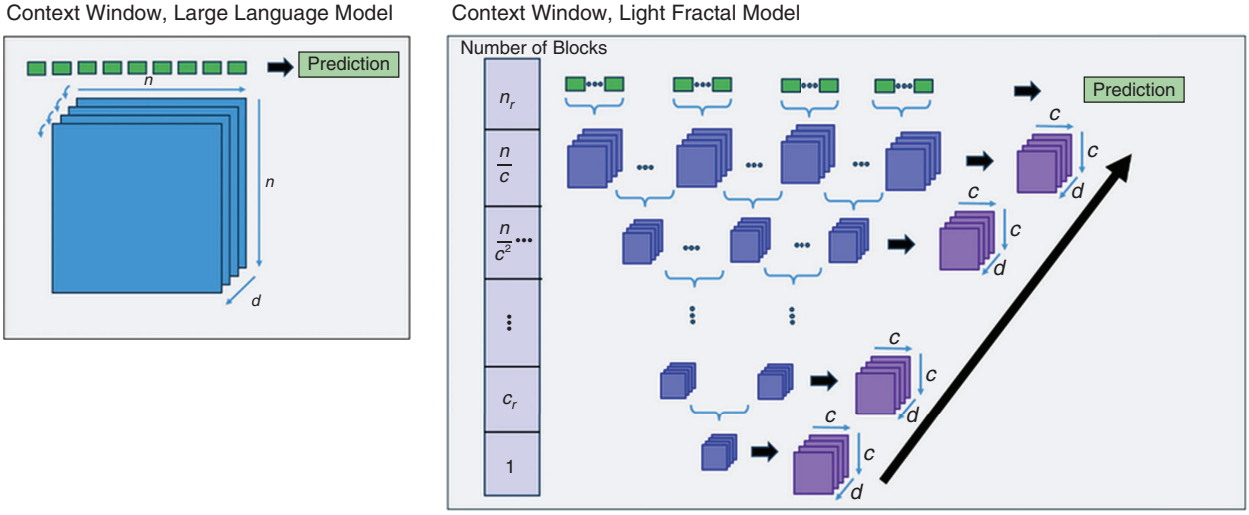

FIGURE 4. Fractal models break down the context into groups of constant size (c) and embed/summarize them into higher levels of abstractions. Inference is done with all levels of abstraction, allowing for global and local context. The compute scales as O(c*d*n)=O(n) (linearly in the context window) while traditional transformers scale as d*n2 =O(n2). We find c between 100 and 1,000 tokens usually gives great results. (Source: Soulaiman Itani and Incantor AI; used with permission.)

- sparse access patterns enabling terabyte-scale storage with constant retrieval cost
- dynamic reorganization and compression, allowing memory to evolve as knowledge accumulates.

As new data arrives, the memory system prestructures embeddings for efficient retrieval and updates its organization to reflect discovered patterns.

Training and convergence

Cantor's training framework integrates multiresolution gradient optimization: gradients at high abstraction levels influence global coherence, while lower levels refine local details.

Parameters across levels follow a geometric reduction—each successive layer containing roughly half the parameters of the one below—minimizing redundancy while preserving capacity.

This design enables faster convergence, lower memory footprint, and parallel optimization of submodules (such as memory and compression) across distributed hardware.

Inference efficiency

During inference, only a subset of parameters is activated, proportional to task complexity rather than total model size (for example, see Table 2).

A canonical LLM will require an estimated range of a few hundred billion active parameters; a rule of thumb is 10–20% of the total model parameters. By contrast, Cantor's selective activation allows large models to operate effectively on standard GPUs, improving accessibility and sustainability.

APPLICATIONS AND COMPARATIVE PERFORMANCE

Preliminary evaluations show that the Cantor LFM achieves:

- 40–60% faster convergence than transformer baselines of similar scale
- order-of-magnitude lower inference cost on long-context reasoning tasks
- enhanced coherence and reduced hallucination rates in extended text generation.

Its architecture also generalizes naturally to multimodal data, applying the same fractal hierarchy to image features, audio sequences, or structured data—where lower levels capture fine detail and higher levels encode compositional semantics.

Current practical applications of the LFM include audio localization, synthetic speech, and IP attribution. As these scale, the intention is to open it up to third party usage.

INTEGRATING FRACTAL, STATE-SPACE, AND CONCEPT MODELS

The future of neural architecture design may not be defined by a single model type. Instead, hybrid systems could combine the strengths of these paradigms:

- state-space dynamics for efficient temporal modeling
- implicit gating for scalable long-range context
- fractal hierarchy for multi-level abstraction
- concept embeddings for semantic reasoning.

Such systems would move toward architectures capable of reasoning hierarchically, storing knowledge persistently, and generalizing across modalities—all while operating within practical computational limits.

TABLE 1. Multilevel embedding structure.

Level	Function
Token	Fine-grained lexical semantics
Word	Complete word meaning
Phrase	Local compositional relationships
Sentence	Full thought representation
Paragraph	Discourse structure and transitions
Document	Global thematic understanding

TABLE 2. Inference efficiency.

Model size	Approximate active parameters
10 B	~100 K
100 B	~300 K
1 T	~1 M

The transformer and its attention mechanism remain milestones in artificial intelligence—an achievement of lasting scientific and practical value. Yet innovation in AI architecture did not end there. As the field confronts challenges of scalability, efficiency, and interpretability, new paradigms are beginning to emerge.

Selective SSMs, implicit convolutional operators, concept-level embeddings, and fractal hierarchical systems each offer distinct paths forward.

The Cantor LFM represents one contribution within this broader

exploration—an attempt to align computational design more closely with hierarchical cognition and efficient memory organization.

Rather than seeking to replace transformers, such approaches aim to expand the design space of intelligent systems. By integrating principles of recursion, sparsity, and multiscale abstraction, future architectures may achieve not only greater performance but also deeper alignment with the structure of reasoning itself.

REFERENCES

1. "Worldwide AI Spending Will Total $1.5 Trillion in 2025," *Gartner*, Sep. 17, 2025. [Online]. Available: https://www.gartner.com/en/newsroom/press-releases/2025-09-17-gartner-says-worldwide-ai-spending-will-total-1-point-5-trillion-in-2025
2. A. Vaswani et al., "Attention is all you need," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst. (NIPS)*, 2017, pp. 6000–6010.
3. A. Gu and T. Dao, "Mamba: Linear-time sequence modeling with selective state spaces," 2023, *arXiv 2312.00752*.
4. M. Poli et al., "Hyena hierarchy: Towards larger convolutional language models," in *Proc. 40th Int. Conf. Mach. Learn.*, 2023, pp. 28,043–28,078.
5. The LCM Team et al., "Large concept models: Language modeling in a sentence representation space," 2025, *arXiv:2412.08821*.

SOULAIMAN ITANI is the co-founder and chief technical officer of Incantor AI, San Francisco, CA 95055 USA and managing director for the Center of Advanced AI at Accenture, San Francisco, CA 94105 USA. Contact him at solomonitani@gmail.com.

Digital Object Identifier 10.1109/MC.2025.3638038