

Geometric Mnemic Manifolds: A Foveated Architecture for Autonoetic Memory in LLMs

Alan Garcia

December 7, 2025

Abstract

I propose a novel architecture for simulating the **functional dynamics of autonoetic memory** in AI systems, departing from the industry standard of stochastic Vector Databases. I introduce the **Geometric Mnemic Manifold**, a system where a **Recursive Reasoning Kernel (RRK)** acts as a fluid reasoning engine, offloading long-term memory to a distributed graph of immutable **Engrams**. Unlike standard RAG systems which optimize solely for semantic relevance, this architecture organizes engrams along a deterministic, low-discrepancy trajectory utilizing **Kronecker sequences** on the hypersphere. By utilizing **Hierarchical Radial Connectivity** coupled with logarithmic radial expansion, the system achieves a mathematically rigorous **Foveated Memory** effect. This exponential decay of information density allows for **Logarithmic Semantic Traversal** ($O(\log N)$) with **Constant Time Addressing** ($O(1)$), mimicking the biological efficiency of human memory consolidation while solving the “Cold Start” latency problem inherent in graph-based indexing.

1 Introduction

Current Long-Context LLMs attempt to simulate memory by extending the input buffer, a method that is computationally expensive and prone to “Identity Drift.” I argue that authentic identity is not found in an infinite context window, but in the tension between a limited working memory and an accessible deep past. Drawing on Tulving’s definition of episodic memory [1], I propose a system that utilizes a **Recursive Reasoning Kernel (RRK)**—a specialized model trained purely for fluid intelligence and routing—connected to a “Neural Bus,” defined formally as a **Vector-Gated Inter-Process Communication (IPC) Fabric**. This bus allows the RRK to query a frozen manifold of previous selves, termed **Engrams**, organized via a tiered, level-of-detail geometric graph.

Remark 1.1 (Architectural Distinction). It is crucial to distinguish this architecture from Semantic Search. Standard vector databases treat memory as a “bag of vectors,” retrieving chunks based solely on similarity to a query. In contrast, the Mnemic Manifold enforces a **spatiotemporal topology** where the “distance” to a memory is a function of both semantic similarity *and* temporal recency. This dual-metric retrieval is essential for **functional autonoesis**—the mechanical ability to distinguish *remembering* an event (via bus retrieval) from merely *knowing* a fact (via weight activation).

2 The Engram and the Kernel

Definition 2.1 (The Engram). Let \mathcal{E} be the space of immutable memory states. An Engram $\epsilon_t \in \mathcal{E}$ represents the frozen phenomenological state of the agent at time t , defined physically as the **Serialized Context Window** (text/tokens) retained at the end of a generation cycle. It contains the local context W_t and the embedding vector v_t . Unlike standard RAG chunks, an Engram is executable; it can be “woken up” by feeding W_t into a new instance of the Kernel to process a query using its original context.

Definition 2.2 (The Recursive Reasoning Kernel). The RRK is a model \mathcal{M} optimized for **Fluid Intelligence** over Crystallized Knowledge. Its primary objective is not to store facts, but to function as a **Router** and **Synthesizer**.

It detects epistemic gaps in its immediate context and issues signals to the Neural Bus to retrieve information from the Engram graph.

Remark 2.1 (Candidate Kernels and Efficiency). While the RRK can be instantiated via custom distillation, the architecture is compatible with existing Small Language Models (SLMs) that exhibit high reasoning density. Models such as **Microsoft Phi-3 Mini (3.8B)** or **Qwen 2.5 (0.5B)** serve as ideal reference implementations for the Kernel role, balancing the fluid intelligence required for synthesis with a minimal memory footprint suitable for rapid, ephemeral instantiation.

3 Geometric Network Topology

To solve the retrieval latency problem in a potentially infinite lifetime, I abandon the linear list for a **Foveated Geometric Graph** derived from Discrepancy Theory.

Proposition 3.1 (Foveated Hyperspherical Organization). Let the active agent a_n reside at the polar origin $(0, 0)$. To generalize the optimal packing properties of Fermat’s Spiral to high-dimensional embedding spaces (\mathbb{R}^d) , I employ a **Kronecker Sequence** mapped via the Inverse Error Function to ensure low-discrepancy angular coverage, coupled with an exponential radial function to enforce temporal foveation.

$$\mathbf{u}_k = \mathcal{M}_{S^{d-1}}(2 \cdot \{k \cdot \alpha\} - 1) \quad (1)$$

$$r_k = e^{\lambda k} \quad (2)$$

Where $\mathcal{M}_{S^{d-1}} \equiv \text{erf}^{-1}$ maps $[0, 1]$ to the hypersphere, $\{\cdot\}$ denotes the fractional part, $\alpha = (\sqrt{p_1}, \dots, \sqrt{p_{d-1}})$ is a vector of linearly independent irrational numbers (square roots of primes), and λ is a decay constant determining the rate of compression.

Remark 3.1 (Tractability via Weighted Discrepancy). While Quasi-Monte Carlo methods typically suffer from the curse of dimensionality in isotropic spaces, the exponential radial function r_k effectively assigns decaying “importance weights” to older dimensions of time. This ensures that the effective dimension of the retrieval task remains manageable, rendering the star-discrepancy polynomially tractable even in high-dimensional embedding spaces.

Definition 3.1 (Hierarchical Radial Connectivity). To manage bandwidth, the active agent a_n does not connect to all past nodes uniformly. Instead, connectivity is determined by the ring distance r_k :

1. **Inner Ring (The Fovea):** For small k (recent past), a_n connects to raw Episodic Engrams $\epsilon^{(0)}$.
2. **Middle Ring (The Para-Fovea):** For medium k , a_n connects only to Layer-1 Synthesized Nodes $\epsilon^{(1)}$ (summarized patterns).
3. **Outer Ring (The Periphery):** For large k (deep past), a_n connects only to Layer-2 Abstract Nodes $\epsilon^{(2)}$ (semantic axioms).

This ensures the total number of active edges remains roughly constant ($O(1)$) regardless of the agent’s lifespan, implementing a “Level-of-Detail” (LOD) memory system.

Concretely, we define the layer cardinalities as:

$$|L_0| = N \quad (3)$$

$$|L_1| = \lceil N/\beta_1 \rceil \quad \text{with } \beta_1 = 64 \quad (4)$$

$$|L_2| = \lceil N/(\beta_1 \cdot \beta_2) \rceil \quad \text{with } \beta_2 = 16 \quad (5)$$

The ring boundaries are defined by $k_{\text{fovea}} = 10$ and $k_{\text{para}} = \beta_1$. This ensures $|L_1| = O(N/64)$ and $|L_2| = O(N/1024)$, yielding logarithmic growth in the number of abstract nodes. The active agent maintains direct connections to all nodes in the Fovea ($k < k_{\text{fovea}}$), all Layer-1 nodes in the Para-Fovea ($k_{\text{fovea}} \leq k < k_{\text{para}}$), and all Layer-2 nodes in the Periphery ($k \geq k_{\text{para}}$), bounding the total active edge count to $O(k_{\text{fovea}} + |L_1| + |L_2|) = O(1)$ with respect to raw engram count N .

4 The Recursive Telegraphic Skip-Graph

To avoid the $O(N)$ retrieval cost of a linear scan, I define the Abstract Layers as a **Semantic Skip-List**.

Algorithm 4.1 (The Recursive Telegraphic Skip-Graph). Let Λ be a synthesis operator defined as a **Telegraphic Compressor**. The operator minimizes the token count of the input by eliminating function words and enforcing syntactic conciseness, while strictly constrained to preserve all named entities and causal links. The retrieval algorithm proceeds as follows:

1. **Pattern Ring Broadcast:** The Neural Bus first broadcasts the query to the **Pattern Ring (Layer 1)**. Since these nodes are sparse (logarithmically fewer than raw memories), this allows for a complete scan of the agent’s semantic history despite the geometric scattering of raw engrams on the spiral.
2. **Foveal Check:** Simultaneously, the system scans the dense local neighbors ($k < 10$) for immediate context.
3. **Drill-Down:** If a Layer-1 node shows high semantic resonance, the system “drills down” into its constituent children (Raw Engrams) to retrieve specific details.

This hierarchical routing ensures that deep history is accessed via semantic pointers rather than brute-force distance calculation.

5 Mnemic Persistence and Ephemeral Discourse

Unlike systems that rely on forgetting to manage compute, this architecture supports total retention through topological efficiency.

Proposition 5.1 (Unbounded Mnemic Persistence). For any generation n , the set of all antecedent engrams remains persisted in the manifold. However, access is mediated by the Hierarchical Radial topology. This allows the agent to retain an infinite history while maintaining a constant-time working memory loop, as it only ever interacts with a sparse set of variable-resolution nodes.

Remark 5.1. When a conflict arises between the RRK and a retrieved Engram, the system spawns an **Ephemeral Clone**—a temporary computational instance of the Engram—to engage in dialectic discourse. This ensures the past is consultable and active, but protected from corruption by the present.

6 Differentiation from Heuristic Retrieval Architectures

While recent work, most notably *Generative Agents* [3], explores the utility of long-term memory streams, it is critical to distinguish the **Geometric Mnemic Manifold** from such heuristic-based approaches. The distinction lies in the transition from algebraic scoring to geometric topology, and from passive text retrieval to active computational agency.

6.1 Geometric Topology vs. Algebraic Scoring

Park et al. employ a scoring heuristic where retrieval relevance is a weighted sum of recency, importance, and similarity scalars. This relies on standard database logic ($O(N)$ or $O(\log N)$ scans over a list) where time is merely a metadata tag. In contrast, the Mnemic Manifold treats time as a **physical coordinate** (r_k) in the embedding space. By encoding recency geometrically via the exponential spiral, the “forgetting curve” becomes an intrinsic property of the manifold’s topology rather than an extrinsic filtering function. This allows for constant-time addressing ($O(1)$) of temporal loci without index lookup.

6.2 Active Agency vs. Passive Retrieval

Standard architectures retrieve **inert text**—strings of data that are pasted into the current context window. This system retrieves **computational agency**. When an Engram is accessed, it is not merely read; it is “resurrected” as an Ephemeral Clone (Definition 4.1). This clone possesses the “frozen” mindset of that specific time-step and can engage in dialectic reasoning, allowing the current agent to debate its past self rather than simply reading a log of its past thoughts.

6.3 Serverless Instantiation (The Zero-Index Advantage)

Heuristic retrieval systems require the pre-loading of massive vector indices (e.g., HNSW graphs) into memory to function, introducing significant “Cold Start” latency. Because the Geometric Mnemic Manifold defines memory locations analytically via the Kronecker sequence, an agent can be instantiated instantly without loading a graph structure. The “Index” is not a stored data structure, but a mathematical function, enabling truly ephemeral, serverless agent instantiation.

7 Experimental Methodology

To validate the architectural claims without incurring the computational cost of pre-training a foundation model from scratch, I propose a rigorous three-stage experimental protocol designed to isolate the effects of the Geometric Mnemic Manifold from the latent knowledge of the underlying model.

Remark 7.1 (Scope and Contribution). We present this methodology as a specification for future empirical validation. The primary contribution of this work is the *architectural design* and its theoretical properties; full-scale experimental results are deferred to a follow-up implementation paper. We release the SLB generator specification and benchmark protocol to enable independent replication.

7.1 Synthetic Longitudinal Biographies (SLB)

A critical vulnerability in testing memory architectures using historical data (e.g., diaries of Samuel Pepys) is *Data Contamination*. Foundation models have likely memorized these texts during pre-training. To circumvent this, I employ a **Synthetic Longitudinal Biography (SLB)** generator.

This procedural engine generates consistent, long-horizon life logs for an agent in a universe governed by unique physical laws and populated by phonotactically neutral entities (e.g., “Banet”, “Mison”, “Toral”). By strictly avoiding real-world nouns, we ensure that any correct retrieval of a fact (e.g., “I sold the Banet on Day 45”) is the result of the Mnemic Manifold’s operation and not latent synaptic weight activation.

7.2 Epistemic Regularization

The Recursive Reasoning Kernel (RRK) is fine-tuned not merely on next-token prediction, but on **Epistemic Gap Detection**. The training corpus contains adversarially generated samples where the answer to a query depends on information deliberately excised from the local context window.

$$\mathcal{L}_{\text{total}} = \mathcal{L}_{\text{LM}} + \lambda \mathcal{L}_{\text{signal}} \quad (6)$$

Where $\mathcal{L}_{\text{signal}}$ penalizes the model for hallucinating an answer when the requisite fact is absent from W_t , forcing the generation of the <SIGNAL_BUS> token. This hard-codes “Socratic Ignorance” into the fluid intelligence layer.

7.2.1 Formal Definition of $\mathcal{L}_{\text{signal}}$

Let \mathcal{D}_{gap} be a distribution over tuples (q, W_t, m) where q is a query, W_t is a (potentially masked) context window, and $m \in \{0, 1\}$ indicates whether W_t contains sufficient information to answer q . We define:

$$\mathcal{L}_{\text{signal}} = -\mathbb{E}_{(q, W_t, m) \sim \mathcal{D}_{\text{gap}}} [(1 - m) \log p_{\theta}(s|q, W_t) + m \log(1 - p_{\theta}(s|q, W_t))] \quad (7)$$

where $s \equiv \langle \text{SIGNAL_BUS} \rangle$ and $p_\theta(s|q, W_t)$ is the probability assigned by the RRK to emitting the signal token as the first response token.

Training Data Construction. Given a complete context W_t^{full} and a query q answerable from W_t^{full} :

1. With probability 0.5, present $(q, W_t^{\text{full}}, m = 1)$ (information present, should NOT signal)
2. With probability 0.5, mask the answer-relevant sentences to create W_t^{masked} , then present $(q, W_t^{\text{masked}}, m = 0)$ (information absent, SHOULD signal)

Hyperparameter Selection. We find $\lambda = 0.5$ provides a stable balance between language modeling quality and epistemic calibration. Values $\lambda > 1.0$ cause excessive signaling (over-cautious behavior); values $\lambda < 0.1$ fail to suppress hallucination adequately.

7.3 Proposed Benchmark: The Needle in the Spiral

To empirically quantify the efficiency gains of the deterministic Geometric topology over stochastic graph-based indexing (HNSW), I define the “**Needle in the Spiral**” benchmark.

1. **Protocol:** A unique “passkey” fact is inserted at a random depth k in a context history of 10^6 tokens.
2. **Metric:** We measure **Recall@1** and **Time-to-First-Token (TTFT)**.
3. **Hypothesis:** While HNSW indices suffer from $O(\log N)$ traversal latency and significant memory overhead for index maintenance, the Geometric Manifold allows for $O(1)$ analytical address calculation. I hypothesize that the Geometric architecture will demonstrate a constant TTFT regardless of memory depth, limited only by the bandwidth of the Neural Bus.

8 Conclusion

This architecture represents a paradigm shift from “Training” to “Experiencing.” By combining a lightweight reasoning kernel with a foveated, geometric graph of immutable memories, we achieve a system that eliminates the “Cold Start” latency of standard vector indices while maintaining a robust sense of self over indefinite timeframes.

References

- [1] Tulving, E. (1985). Memory and consciousness. *Canadian Psychology*, 26(1), 1–12.
- [2] Shazeer, N., et al. (2017). Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *ICLR*.
- [3] Park, J. S., et al. (2023). Generative Agents: Interactive Simulacra of Human Behavior. *UIST*.
- [4] Diekelmann, S., & Born, J. (2010). The memory function of sleep. *Nature Reviews Neuroscience*.