The Heart Beat Agents Memory Architecture

A Cognitive Operating System for Resilient Autonomous Agents

Technical Series Vol. 2

www.HeartBeatAgents.com

I. Executive Abstract: The Memory Wall and the Stateless Ceiling

Modern Large Language Models (LLMs) operate as sophisticated "stateless functions," transforming inputs into outputs without a native, persistent feedback loop. This creates a "Memory Wall": a ceiling where agent utility is limited by the fixed context window and the lack of longitudinal learning. Existing solutions typically rely on external vector databases that introduce operational entropy, network latency, and consistency gaps.

Heart Beat Agents or HBA introduces a unified memory architecture that transcends simple Retrieval-Augmented Generation (RAG). By integrating a multi-signal cognitive storage layer directly into the primary relational database (PostgreSQL), we achieve Transactional Memory Consistency. Our architecture moves beyond "stateless response" to "stateful evolution," utilizing Maximal Marginal Relevance (MMR) and Layered Graceful Degradation to ensure that memory is an indestructible enhancement rather than a load-bearing failure point.

II. Cognitive Taxonomy: The Tripartite Memory Model

To build an agent that mirrors human-like proficiency, we depart from the "flat file" approach to data storage. HBA categorizes information based on its functional role in cognitive processing, drawing directly from established cognitive psychology.

2.1 Semantic Memory: The Fact-Based Knowledge Graph

- **Definition**: A repository for context-independent facts, naming conventions, and project constraints.
- **Function**: Stores "What the Agent Knows" (e.g., "The client's brand color is #0066CC").
- **Weighting**: High persistence; these memories define the rigid boundaries of the agent's operating environment.

II. Cognitive Taxonomy: The Tripartite Memory Model

2.2 Episodic Memory: Temporal Experience and Reinforcement

- **Definition**: A chronological record of task outcomes and specific interaction experiences.
- Function: Stores "What the Agent Experienced" (e.g., "The deployment to us-east-1 failed due to a timeout").
- **Weighting**: High persistence; these memories define the rigid boundaries of the agent's operating environment.
- **Reinforcement Bias**: We implement a Negative Reinforcement Bias, where failures are assigned a higher importance score (0.9) than successes (0.7). In a stochastic environment, knowing the "failure path" is mathematically more valuable for error-avoidance than confirming a known success.

II. Cognitive Taxonomy: The Tripartite Memory Model

2.3 Procedural Memory: The Evolution of Skills

- **Definition**: A specialized store for successful tool sequences and workflow recipes.
- **Function**: Stores "How the Agent Acts" (e.g., read_file → pandas_analysis → matplotlib_plot).
- **Operational Impact**: By caching successful "Muscle Memory," agents bypass expensive trial-and-error cycles in multi-step tool orchestration.

II. Cognitive Taxonomy: The Tripartite Memory Model

2.4 Conversation Memory: Continuity Summarization

- **Definition**: Compressed semantic summaries of previous dialogue exchanges.
- **Function**: Maintains continuity across sessions without the "token tax" of full transcript storage.

Memory Type	Biological Analogy	Data Type	Importance Decay
Semantic	World Knowledge	Facts/Constants	Low (Stable)
Episodic	Personal Experience	Outcome Logs	Variable (Bias to Failure)
Procedural	Muscle Memory	Tool Chains	High (Performance-linked)
Conversation	Working Notes	Summaries	Fast (Context-dependent)

III. The Mathematics of Recall: Quad-Signal Ranking

The core of our retrieval engine is the Hybrid Ranking Formula. We recognize that Vector Similarity (*V*) is a powerful tool for semantic meaning, but it is often "fuzzy" and fails to distinguish between fresh insights and stale data. We calculate the relevance of a memory entry (*M)* against a query (Q) using the following weighted objective function:

$$FinalScore = 0.55(V_{score}) + 0.20(T_{score}) + 0.10(I_{score}) + 0.15(R_{score})$$

Where

- V_{score} (Vector Similarity): Cosine similarity of the 1,536D embedding.
- T_{score} (Textual Rank): PostgreSQL ts_rank for linguistic token matching, ensuring precision for acronyms and specific proper nouns.
- I_{score} (Importance): A subjective weighting $(0.3–0.9)$ assigned at creation or updated via access frequency.
- R_{score} (Recency): An exponential decay function $1/(1 + days\_since\_created)$, ensuring the agent prioritizes current reality over historical noise.

This multi-signal approach ensures that the agent is not just finding "similar" things, but finding the most relevant, high-quality, and current information available.

IV. Implementation: The Single-Box Relational-Vector Nexus

A critical divergence from contemporary agentic middleware is our rejection of the distributed vector database. For the high-concurrency requirements of autonomous agents - working locally and collaborating at edge - the "Single-Box" approach utilizing **PostgreSQL with pgvector** minimizes the CAP theorem trade-offs inherent in multi-service architectures.

IV. Implementation: The Single-Box Relational-Vector Nexus

4.1 Dimensionality Normalization: The Zero-Padding Proof

The Heart Beat Agents architecture is designed for *provider agnosticism*. Since various embedding models (OpenAI, Gemini, Mistral, Ollama) produce inconsistent vector dimensions (from 384D to 3,072D), we employ a mathematical transformation to ensure schema stability without data migration.

- **The Mechanism**: All storage columns are fixed at Vector(1536).
- **The Math**: Shorter vectors are transformed via _pad_to_storage(), which appends zero-values to reach the target dimensionality.
- Zero-padding doesn't change cosine similarity, so vector relationships stay consistent, allowing you to switch embedding systems (like OpenAI to Ollama) without changing your database structure, just re-indexing in the background.

IV. Implementation: The Single-Box Relational-Vector Nexus

4.2 Indexing Strategy: HNSW over IVFFlat

For sub-millisecond retrieval at scale, we implement Hierarchical Navigable Small World (HNSW) indexes on all embedding columns.

- **Parameters**: We utilize m=16 (max connections per element) and ef_construction=200 (search buffer during index creation).
- **Operational Integrity**: Indexes are created CONCURRENTLY to prevent table-locking, ensuring that agent memory remains accessible during maintenance cycles.

V. System Resilience: The Four Layers of Graceful Degradation

Most autonomous systems exhibit "Brittle Failure"—if the vector database or embedding API fails, the agent's cognition collapses. The Heart Beat Agents architecture is engineered for "Functional Floors," ensuring that agents remain operational regardless of infrastructure health.

- **Layer 1**: **Semantic Failure**: If the embedding provider is unreachable, the system automatically bypasses the vector CTE in the SQL query. The V_score is nullified, and the system relies on the remaining three signals (Text, Quality, Recency).
- **Layer 2**: **Temporal Latency**: Memory injection is governed by a strict 1.0s timeout. If retrieval exceeds this window, the agent proceeds with its base system prompt rather than blocking the execution loop.
- **Layer 3**: **Cognitive Saturation**: To prevent "context poisoning," we implement Maximal Marginal Relevance (MMR). From the top 20 candidates, we select 12 entries that maximize the distance from one another, ensuring high information density and avoiding redundant "echoes" in the prompt.
- **Layer 4**: **Atomic Deduplication**: New memories are vetted against existing entries using a strict cosine similarity threshold (0.92 for explicit intent). This prevents the "Infinite Loop" problem where agents repeatedly store the same observation.

VI. Active Maintenance: The Entropy Management Cycle

A memory system that only grows is a system that eventually fails. We treat memory as a dynamic data structure subject to Consolidation and Decay.

6.1 LLM-Driven Consolidation

Every 6 hours, the system executes a Union-Find Clustering algorithm.

- **Cluster Threshold**: Memories with a cosine similarity above 0.82 are grouped.
- **Merging**: Clusters of three or more are synthesized by an LLM into a single, high-density "Consolidated Truth," and the original atomic entries are purged.

VI. Active Maintenance: The Entropy Management Cycle

A memory system that only grows is a system that eventually fails. We treat memory as a dynamic data structure subject to Consolidation and Decay.

6.2 The Forgetting Curve (Importance Decay)

To simulate cognitive relevance, we apply a daily decay multiplier of 0.995 to all importance scores.

- **The Math**: A memory with a starting importance of 0.5 will survive for ~390 days before dropping below the 0.01 deletion threshold.
- **Reinforcement**: Every time a memory is accessed via the recall tool, its access count increments and its decay is functionally reset, ensuring that "useful" knowledge is biologically immortal within the system.

VI. Active Maintenance: The Entropy Management Cycle

A memory system that only grows is a system that eventually fails. We treat memory as a dynamic data structure subject to Consolidation and Decay.

6.3 Bounded Memory Capacity

Each agent is strictly capped at 500 memories. When this limit is breached, the system executes Intelligent Pruning, deleting the lowest-importance entries first to ensure the vector index remains performant and noise-free.

VII. Security, Privacy, and Multi-Tenant Scoping

In an enterprise-grade autonomous environment, memory cannot be a monolithic data lake. The Heart Beat Agents architecture enforces knowledge isolation by design, ensuring that data leakage between agents or organizations is architecturally impossible.

VII. Security, Privacy, and Multi-Tenant Scoping

7.1 Hierarchical Scoping

Memory entries are governed by a strict ownership model:

- **Agent Scope (Default)**: The vast majority of memories are private to the specific agent instance. This ensures that an HR agent's episodic memory of a sensitive termination does not leak into the semantic search of a Marketing agent.
- **Workspace Scope**: Knowledge can be explicitly "promoted" to the organization level through administrative action or specific API triggers. Once promoted, these memories (e.g., brand guidelines, shared deployment keys, or company-wide holidays) become visible to all agents within that specific tenant.
- **Isolation Integrity**: Every database query for memory or skills is hard-coded to include the organization id and the agent id filters, enforced at the repository layer to prevent horizontal privilege escalation.

VII. Security, Privacy, and Multi-Tenant Scoping

7.2 Encrypted Provider Pipelines

For organizations utilizing cloud-based embedding providers (OpenAI, Voyage, Cohere), API keys are stored using AES-256 encryption at rest. For high-security environments, the architecture supports a fully air-gapped deployment via Ollama, where both the LLM and the embedding model run on local silicon, ensuring no data ever traverses a third-party network.

VIII. The Path to Autonomous Maturity

The Heart Beat Agents architecture represents a fundamental shift from static prompt engineering to dynamic experience engineering. By solving the "Memory Wall," we move agents from being tools that respond to colleagues that evolve.

VIII. The Path to Autonomous Maturity

8.1 Summary of Architectural Advantages

Memory entries are governed by a strict ownership model:

- **Operational Simplicity**: Utilizing PostgreSQL and pgvector eliminates the "infrastructure tax" of managing separate vector databases, reducing deployment complexity by an order of magnitude.
- **Resilience through Hybridization**: By combining Vector, Full-Text, Quality, and Recency signals, the system achieves a level of retrieval precision that single-signal systems cannot match.
- **Cognitive Efficiency**: Consolidation and MMR-driven selection ensure that the agent's context window is filled with high-entropy, non-redundant information, maximizing the value of every token spent.

VIII. The Path to Autonomous Maturity

8.2 Future Directions: From Memory to Skill Synthesis

The next evolution of this architecture involves moving from Procedural Memory (storing tool sequences) to Skill Synthesis, where successful procedures are automatically compiled into "Native Skills" (DSL-based functions). This transition will allow agents to not just remember how they solved a problem, but to create a permanent, optimized shortcut for all future agents in the workspace to utilize.

IX. Conclusion

The Heart Beat Agents architecture does not treat memory as an archive; it treats it as a cognitive lubricant. By engineering for failure, optimizing for information density, and anchoring everything in a robust relational foundation, we have created an architecture that allows autonomous agents to scale without succumbing to the noise of their own history. Memory that finds itself is no longer a theoretical ideal, it is a production reality.

This document was authored by Jyhad Aamri, the principal architect behind the Heart Beat Agents security framework.

Jyhad Aamri
Chief Executive Officer @ Mosaic Singularity
hello@JyhadAamri.com
www.JyhadAamri.com