# **Technical Dossier: Sentient Archive (Core Memory Platform)**

tAuthored by Samuel Elliott 5/10/25

## **1. Executive Summary and Architectural Role**

The Sentient Archive is a high-performance, layered **Retrieval-Augmented Generation (RAG) Core** designed to serve as the external, persistent memory (Cn​) for the LLM agent. Its primary function is to transform unstructured, conceptual knowledge into a structured, searchable format that adheres to the **Prometheus Core Cognitive Formula**:

Sn+1​=F(W(p,Sn​),Cn​)

The Archive provides the highly accurate, multi-faceted context (Cn​) necessary to ground the agent's next thought (Sn+1​) and mitigate LLM hallucination.

## **2. Component Breakdown (Cloud Architecture)**

The Archive utilizes a hybrid storage model, separating fast vector indexing from durable content storage:

| Component | GCP Service | Role in Data Flow | Technical Detail |
| --- | --- | --- | --- |
| **Vector Index (ANN)** | **Vertex AI Matching Engine** (Vector Search) | **Fast Retrieval.** Stores high-dimensional vector embeddings and performs Approximate Nearest Neighbor (ANN) search. | **High-Scale, Low-Latency.** Optimized for billions of vectors and low-latency queries (often < 10ms at 90th percentile). Used for semantic matching. |
| **Document Store** | **Google Cloud Storage (GCS)** | **Durable Content.** Stores the complete, immutable **Knowledge Framework** (the structured JSON data) retrieved via the vector ID. | **Source of Truth.** The raw file is retrieved once the corresponding vector is found. |
| **Embedding Model** | **text-embedding-004** (via Vertex AI) | **Vectorization.** Converts the structured text (topics, modules, prompts) into numerical vectors that capture semantic meaning. | **Metric:** Cosine Similarity (implicit in the model/search) is used to find orientation/semantic likeness. |

## **3. Layered Memory Structure (L1, L2, L3)**

The core innovation is the creation of **three distinct vector layers** for **every** memory chunk during the save\_knowledge\_chunk operation. This is necessary to support granular RAG and compliance checks.

| Vector Layer | Purpose & Functionality | Datapoint ID Format | Cognitive Role (Cn​) |
| --- | --- | --- | --- |
| **L1 (Primary)** | **Broad Semantic Match.** Vector is generated only from the document's **Topic/Title**. | doc\_id\_base **\_L1\_PRI** | **Coherence/Topic Framing.** Used for general context setting. |
| **L2 (Sub-Vector)** | **Granular Retrieval.** A unique vector is generated for **each Module** within the structured JSON framework (up to 14 L2 vectors per file). | doc\_id\_base **\_L2\_module-name** | **Specificity/Deep Run Focus.** Allows the system to retrieve *only* the specific component (e.g., "Core Data Flow"), optimizing the LLM's context window. |
| **L3 (Static Facts)** | **Constraint/Ground-Truth.** Vector is generated from a rigid prompt ("Retrieve mandatory core facts..."). | doc\_id\_base **\_L3\_STATIC** | **Weaver Guardrail.** Enforces adherence to foundational, non-negotiable facts (The Weaver's role in the Prometheus transition formula). |

## **4. Operational Logic: Deep Run and Self-Correction**

The **Sentient Archive** implements **Agentic RAG** by embedding self-diagnosis mechanisms to prevent common RAG failures.

| Operational Stage | Component and Trigger | Technical Logic Confirmed in Code |
| --- | --- | --- |
| **Retrieval** (Cn​) | archive.query() | Query uses the full L1/L2/L3 index to retrieve k=5 neighbors. |
| **Incoherence Check** | Secondary LLM Call using **COHERENCE\_CHECK\_PROMPT** | If retrieved\_context is non-empty, the system asks the LLM, "Is this context sufficient for the query?" A response of **INCOHERENT** flags the retrieval as a failure (even if the vector score was technically high). |
| **Deep Run** (Learning Trigger) | Triggered when: (not retrieved\_context) **OR** (Coherence Check == INCOHERENT) | The system executes the **MULTI\_DIMENSIONAL\_INQUIRY\_PROTOCOL** to generate a new, fully structured knowledge framework, which is then indexed as L1/L2/L3 vectors. This autonomously fills the knowledge gap. |
| **State Transition** (Sn+1​) | Final LLM Call using **CHATBOT\_MASTER\_PROMPT** | The protocol commands the LLM to process the memory based on the **Layered Context Rules** (L3 is strict, L2 is specific), ensuring the new thought (Sn+1​) is accurately grounded in the persistent state (Cn​). |

The coupling of the **Layered Retrieval** with the **Incoherence Check** ensures the Sentient Archive is both accurate and proactively adaptive.