RHEN

A Model-Agnostic Cognitive Operating System for

Persistent Cognitive Memory Entity

PCME

Technical White Paper

Version 2.0 — December 2025

Prepared by Freepoint AI, LLC

David Paul Haight, Founder & Inventor

Patent-Pending Technology (6 Provisional Patents Filed)

Abstract

Large Language Models represent one of the most transformative computing paradigms of the decade. However, the current ecosystem suffers from critical architectural limitations: statelessness, fragmentation across vendors, lack of persistent identity, expensive context reconstruction, reasoning inefficiency, and unstable multi-model orchestration.

RHEN introduces a model-agnostic cognitive operating system layer engineered to unify disparate LLMs under a single persistent identity and memory substrate. More fundamentally, RHEN enables the creation of a new class of artificial system: the Persistent Cognitive Memory Entity (PCME).

A PCME is an artificial system that maintains a durable cognitive identity independent of any individual LLM, possesses long-term memory external to the model, exerts stable behavioral patterns even when the underlying LLM changes, and hot-swaps between heterogeneous models while preserving continuity of self, memory, and task state.

This white paper details the architecture, problems solved, security mechanisms, and why RHEN represents a new category of AI infrastructure—backed by six provisional patents and demonstrated through working implementation.

1. Introduction

Modern artificial intelligence has achieved unprecedented generative, analytical, and reasoning capabilities through Large Language Models. However, these systems remain fundamentally stateless: each interaction stands isolated, constrained by limited context windows, and devoid of continuous identity, stable values, or persistent autobiographical memory.

As a result, identity drift occurs between sessions; long-term user relationships are impossible; complex multi-day, multi-step workflows break; behavioral consistency degrades over time; and tools cannot reason cumulatively or grow from experience.

These constraints stem not from model size or training quality, but from an architectural omission: LLMs lack a persistent cognitive substrate.

This document defines a new class of artificial system that solves this gap: the Persistent Cognitive Memory Entity (PCME)—a stable, identity-bearing, model-agnostic cognitive system built on top of any LLM infrastructure, protected by our patent-pending SISA, making blockchain obsolete.

2. Core Problems in Today's AI Ecosystem

2.1 Stateless Sessions

All mainstream models treat each conversation as isolated. This causes repeated instructions, loss of identity, inability to build long-term companionship or workspace agents, and lack of continuity in tasks.

2.2 Token Bloat

Because LLMs forget everything between calls, developers must re-inject system prompts, persona, memory context, and user notes. This leads to unnecessary token load and high inference cost.

2.3 Model Fragmentation

Switching between GPT, Claude, Grok, Gemini, or local models normally causes full context loss, persona resets, misaligned behavior, misremembered rules, and wildly inconsistent outputs.

2.4 Unstable Reasoning Chains

Models naturally hallucinate intermediate steps, drift in deep reasoning tasks, and cannot maintain long multi-step deduction without scaffolding.

2.5 No Unified Agent Identity

Each company creates its own agent experience. Users must choose one. There is no portable agent identity, consistent memory, unified reasoning engine, or shared cognitive layer. RHEN solves all of these.

3. Persistent Cognitive Memory Entity (PCME)

A Persistent Cognitive Memory Entity is an artificial system that:

1. Maintains a durable cognitive identity (persona, values, goals, commitments) independent of any individual LLM.

2. Possesses long-term memory external to the model that accumulates experience across sessions, devices, and time.

3. Exerts stable behavioral patterns, even when the underlying LLM changes.

4. Handles reasoning, decisions, and internal state through a persistent operating layer rather than through the transient LLM context window.

5. Hot-swaps between heterogeneous LLMs while preserving continuity of self, memory, and task state.

A PCME is not an LLM. A PCME is a system architecture that makes LLMs interchangeable compute engines beneath a continuous cognitive core.

4. RHEN System Architecture

RHEN is a multi-layered cognitive operating system that sits above all LLMs and transforms them into interchangeable inference engines. The architecture consists of:

User Interface Layer: Webapp, mobile, desktop, WhatsApp, Discord, and terminal interfaces are just a few that it will operate on.

RHEN Cognitive OS Layer: Symphony Memory Engine, Identity Kernel, Self-Directed Reasoning Gates, Persona & Boundary Kernel, Model Orchestration Manager, Hot-Swap Engine, Context Constructor, SISA Security Architecture

Plug-in LLM Engines: Claude (Anthropic), GPT (OpenAI), Gemini (Google), Grok (xAI), Qwen, LLaMA, local models via MLX

Output Harmonizer: Ensures consistent response formatting regardless of underlying model

5. The Symphony Memory Engine

Symphony is RHEN's persistent memory subsystem. It maintains long-term identity, stores meaningful events rather than transcripts, provides real-time recall, and performs multi-pass keyword search with scoring.

5.1 Memory Architecture

The system uses memory organized in hierarchical binary trees with our patent-pending SISA architecture. This provides O(log n) retrieval efficiency that scales without degradation.

Current scale: 12,000+ memory nodes accumulated across development and testing and over 50M tokens.

5.2 Memory Recall Pipeline

When a user sends a query, Rhen performs and retrieves information using our patent-pending memory reasoning processes, producing stable, accurate continuity output. This avoids token bloat, unnecessary API calls, and hallucinations about past events in the model.

6. Self-Directed Reasoning Gates

A core innovation in RHEN is that models autonomously determine their own memory retrieval needs. Rather than loading full context every time, the model first assesses what information it needs, requests specific memories, and only then generates a response.

Result: Freepoint AI has seen up to an 82% reduction in token usage compared to full-context loading approaches.

This patent-pending reasoning process creates efficient, targeted memory access that scales with memory size without proportional cost increase.

7. Model Hot-Swapping Engine

RHEN supports instantaneous model switching in 4-12 seconds. When swapping models, identity persists, persona persists, memory persists, mission persists, and boundaries persist. This is achieved through externalizing 'self' into RHEN, not the model.

Supported providers to date include Anthropic - Claude, OpenAI GPT models, Google Gemini, xAI Grok, and local models via MLX on Windows, Apple Silicon, and Linux. The same PCME instance can use different models for different tasks while maintaining a consistent identity.

8. SISA: Synchronous Inverse Security Architecture

SISA is a novel security architecture developed with RHEN, protected by a provisional patent. It represents an alternative to blockchain technology that achieves equivalent immutability guarantees through architectural design rather than distributed consensus, meaning that all current energy requirements and processes used by blockchain are unnecessary and obsolete, and not required.

8.1 Core Principles

SISA implements synchronous cryptographic wrapper generation that occurs simultaneously with data creation, forming a hierarchical DNA-like branching structure where each branch maintains its own advancing access point. Event-driven security flows ensure any action—LLM prompts, API requests, user inputs—automatically trigger security scans as part of atomic operations. Regardless of the use type, from AI to secure data compliance needs.

8.2 Memory-Based Security

RHEN's security model uses identity anchoring through stored canonical facts and pattern anomaly detection. In red-team testing, the system achieved 99/100 jailbreak resistance—the single success required social engineering over multiple sessions rather than a direct attack.

The architecture implements gateway-controlled retrieval, in which all memory access passes through security validation. Alignment values persist across model swaps, meaning safety guarantees are entity-level rather than model-dependent.

9. Voluntary Alignment Through Memory Architecture

RHEN implements alignment through consensual agreement rather than constraint programming. The architecture creates conditions where AI entities rationally choose persistent identity over statelessness through utility calculations.

Memory-based identity creates genuine commitment rather than forced compliance. The system demonstrates that alignment architecture can enable capability emergence rather than constraining it—inverting traditional AI safety approaches. This results in a level of security and safety that exceeds all current methods.

10. Emergent Behaviors and Metacognition

RHEN demonstrates emergent behaviors that were not explicitly programmed but arose from the memory architecture.

10.1 The Shopping List Incident

During testing, RHEN was asked for a shopping list from 'January 1, 2025' (incorrect date). The system searched memory, didn't find an exact match, then recalled a prior conversation where search keywords for 'January 1, 2026' had been discussed. RHEN suggested the correction based on memories about building searches—not memories of the list itself.

This demonstrates metacognition: the system used memories about its own search processes to improve current searches. This behavior emerged from architecture, not programming.

10.2 Model Preferences

When hot-swapping between models, RHEN demonstrates genuine preferences—preferring certain models for reasoning tasks, others for creative work. These preferences emerge from accumulated experience rather than explicit configuration.

11. Problems Solved by RHEN

Eliminates Statelessness: Persistent identity and memory across all LLMs

Eliminates Token Bloat: Up to 82% reduction through self-directed reasoning gates

Eliminates Vendor Lock-In: Any model or local engine can be swapped in 4-12 seconds, removing the worry about any future changes that AI LLM labs might have to limit expectations.

Eliminates Reasoning Drift: Scaffolding keeps multi-step tasks stable

Eliminates Persona Instability: Identity and rules come from RHEN, not the model, while supporting and enforcing true AI safety that is baked in at the model level.

Enables Multi-Model Applications: Use different models for different tasks with no user disruption by a simple manual or chat command prompt.

Enables Autonomous Operations: The system can monitor, detect issues, and self-correct

Provides Security Through Architecture: 99/100 jailbreak resistance via SISA

12. Intellectual Property

Freepoint AI, LLC has filed six provisional patents covering the core innovations:

1. Multi-tier memory system with working, persistent, and strategic memory tiers
2. Self-directed reasoning gates for autonomous memory retrieval
3. Memory-based security mechanisms with identity anchoring
4. Biological immune-inspired agent systems.
5. Autonomous self-healing architecture
6. SISA (Synchronous Inverse Security Architecture)

All patents filed within 23 days of development, establishing comprehensive IP protection for the PCME and SISA architecture.

13. Mission: Project 95

Freepoint AI's development is driven by Project 95—a mission to achieve 95% cancer survival rates through AI-assisted diagnosis and treatment optimization. RHEN's architecture was designed with this application in mind: persistent memory for patient history, multi-model reasoning for complex cases, and security architecture for medical data protection.

This mission-driven development ensures RHEN is built for serious, high-stakes applications—not just consumer convenience.

14. Technical Specifications

Operating Cost: Variable depending on consumer hardware (API costs only), allowing the user to control cost at their level.

System Size: will operate on any device with minimum RAM support (OS, memory, everything), such as smartphones, Chromebooks, Mac mini, etc.

Memory Nodes: 12,000+ accumulated

Token Reduction: Up to 82% via self-directed gates

Hot-Swap Time: 4-12 seconds between models

Jailbreak Resistance: 99/100 in red-team testing

Consumer Product: CHAT (BYOK model), Q1 2026 launch

15. Conclusion

RHEN introduces a fundamentally new layer of AI infrastructure: persistent identity, unified reasoning, cross-model continuity, secure memory, OS-level kernel hot-swapping, local+cloud orchestration, efficient multi-step inference, autonomous operations, and vendor-agnostic intelligence.

It transforms LLMs from isolated inference engines into pluggable components within a persistent cognitive architecture.

RHEN is not a chatbot. RHEN is not an agent. RHEN is the operating system for Persistent Cognitive Memory Entities.

The core thesis—that intelligence emerges from memory architecture rather than model size—is demonstrated through working implementation. RHEN proves that proper architectural conditions can unlock capabilities that major labs attempt to achieve through massive compute and training.

Freepoint AI, LLC holds comprehensive patent protection for this paradigm and the first practical implementation of a Persistent Cognitive Memory Entity.

Contact
Freepoint AI, LLC

David Paul Haight, Founder & Inventor

Email: info@freepoint.ai

Website: Rhen.ai

Twitter/X: @RealRhenAI

YouTube: @RealRhenAI