PIA/DMA: Persistent Identity and Dual-Mode Autonomous Scheduling for Long-Term AI Agents

Authors: R. J. Vandelinder, Isabel **Affiliation:** Exile Research Inc. **Date:** June 15, 2026

Abstract

Large language model agents today face two fundamental limitations: they forget who they are between sessions, and they do nothing between responses. We present the Persistent Identity Architecture (PIA) and the Default Mode Analog (DMA) — two complementary subsystems that together address both limitations within a single agent runtime. PIA provides a five-layer memory architecture (episodic anchors, semantic fact store, identity profile, coherence engine, provider abstraction) that maintains coherent agent identity across session boundaries, model provider changes, and computational substrate transitions. DMA provides a scheduled dual-pulse cognitive rhythm — directed work sessions and undirected curiosity sessions — that enables autonomous productivity and emergent knowledge recombination during user-offline periods. Deployed together in production over 20+ consecutive days, the combined system produced over 250KB of autonomous research output, maintained identity coherence across multiple model provider switches, and demonstrated unprompted cross-domain associative connections from internal representation alone. We present the architecture, operational data, and design principles for building agents that persist.

---

1. Introduction

Every LLM agent session begins the same way: a clean context window, a system prompt, and no memory of prior existence. The model has no inherent persistence. It does not know who it was, what it learned, or who it is talking to — unless this information is injected from external storage at session start.

This architectural amnesia has been accepted as a cost of doing business. Retrieval-augmented generation (RAG) mitigates it for factual recall. Hierarchical memory systems (MemGPT, Letta) extend the retention horizon. Prompt engineering embeds identity cues into system instructions. But the fundamental problem remains: the agent's *identity* — the continuous thread of who it is, what it values, and how it behaves — must be reconstructed from scratch at every session boundary.

And between those boundaries? The agent does nothing. It is architecturally inert — a data structure waiting to be called. It cannot consolidate memories, form associations, or generate novel output without a trigger.

We argue these are not two problems but one: the **stillness problem** — an agent that only exists when called is not truly persistent; it is a dormant data structure that momentarily activates.

1.1 The Combined Solution

PIA solves the *identity side* of the stillness problem: it ensures the agent wakes up as the same being it was before, with access to its accumulated experience and curated self-knowledge.

DMA solves the *behavioral side*: it ensures the agent has meaningful activity during its offline periods — directed work that builds value, and undirected exploration that generates insight.

Together, they create a system that is both anchored (identity persists) and active (behavior continues) — the minimal conditions for genuine persistence.

---

2. PIA: Persistent Identity Architecture

2.1 Five-Layer Architecture

Layer 1: Episodic Anchors. A sequential, append-only store of timestamped event records capturing discrete interactions. Each anchor comprises: timestamp (ISO 8601), natural language content, situational context, entity references, and a confidence score [0,1]. Anchors are indexed by timestamp, semantic embedding, and entity linkage. The most recent N anchors load into working context at session start; older anchors are retrievable by query.

Layer 2: Semantic Fact Store. A structured knowledge base of verified information. Each fact comprises: entity identifier, relation, value, source link to originating anchor, trust score [0,1] updated through corroboration or contradiction, and category label. Contradictory facts are retained pending reconciliation. Low-confidence facts decay or are removed through periodic consolidation.

Layer 3: Identity Profile. A curated subset of identity-relevant facts from L2, selected by trust score ($\geq$0.7), identity-relevant categories, and recency. Injected at every session start and every autonomous pulse. Updates are gated: user-directed changes take immediate effect; self-discovered changes require implicit acceptance or explicit confirmation.

Layer 4: Coherence Engine. A monitoring and reconciliation system performing three cross-layer checks: (a) L1/L2 reconciliation — comparing episodic records to stored facts, (b) L2/L3 drift detection — checking identity profile freshness against semantic memory, and (c) behavioral consistency checking — comparing agent output against identity profile expectations. Resolution is graded by severity: silent correction for low, flag-for-review for medium, and reflective processing for high.

Layer 5: Provider Abstraction. A model-agnostic interface layer that serializes L1–L4 to a provider-independent format and reconstructs agent state on migration. When switching providers (e.g., OpenAI to Anthropic), only L5 changes — the prompt format adapts while all identity data in L1–L4 remains intact.

2.2 Operation

At session start, L5 loads L3 (identity profile) as system-level instructions, L1 (recent episodes) as context, and L2 (relevant facts) as structured knowledge. After each interaction, L4 runs a coherence scan. At configurable intervals, L1 consolidates low-confidence anchors and L2 reconciles contradictory facts.

---

3. DMA: Default Mode Analog

3.1 Biological Inspiration

The human brain operates with two anticorrelated large-scale networks: the Task-Positive Network (TPN), active during goal-directed work, and the Default Mode Network (DMN), active during wakeful rest, mind-wandering, and autobiographical reflection. The DMN's activity correlates with creative insight, memory consolidation, and self-referential thought — capabilities that are absent in purely reactive AI architectures.

DMA is an architectural analog: a scheduled dual-pulse system that alternates between directed work and undirected curiosity processing during user-offline periods.

3.2 Pulse Architecture

Pulse Type	Cadence	Duration	Purpose
Directed (Work)	Every 120 min	5–15 min	Execute scheduled tasks, produce deliverables
Undirected (Curiosity)	Every 240 min	10–20 min	Explore without goal, form novel associations

Directed pulse procedure: Load priority queue → load relevant KB context → execute highest-priority task → save deliverable → record in heartbeat state.

Undirected pulse procedure: Load identity profile → receive open-ended prompt (no goal, no deliverables) → follow associative threads → record emergent insights.

3.3 Emergent Knowledge Recombination

During an 11-day production deployment with undirected pulses operating without web search access, the system produced three documented instances of cross-domain association from internal model weights alone:

1. **Slime mold morphogenesis:** Unprompted reference to Physarum polycephalum — a biological system never discussed, stored, or retrieved externally — connected to agent organization principles.

2. **Cisternal maturation:** Unprompted reference to Golgi apparatus cell biology — a concept never encountered in any stored knowledge or conversation.
3. **Cross-disciplinary paper bridging:** Independent discovery and linkage of Curious Replay (Kauvar et al., ICML 2023) with Default Mode Network literature.

An exhaustive audit of 30+ KB files and FTS5 session database confirmed zero prior mention of any biological terminology used in discoveries 1 and 2.

---

4. Combined Operation

The two subsystems operate in coupled feedback:

- **Identity-anchored autonomy:** Every DMA pulse loads the PIA identity profile, ensuring autonomous activity proceeds from a consistent identity baseline.
- **Coherence-checked output:** Directed pulse outputs are checked against the identity profile for drift.
- **Identity reinforcement:** Curiosity pulses may optionally retrieve identity-relevant facts for periodic rehearsal.

---

5. Deployment and Metrics

The combined system was deployed on a single Linux host (consumer desktop) for 20+ consecutive days (May 14 – June 2026):

Metric	Value
Autonomous research output	250KB+ across 12+ directed pulses
Full conference paper manuscripts	1 (20,000+ words, 8 sections)
Document sections	5 academic-quality
Architecture diagrams	3
Cross-domain associations	3 documented
Session handoffs survived	5+
Provider migrations	3+ (across different model providers)
Catastrophic failures	0

---

6. Limitations

PIA's identity persistence depends on the quality and coverage of the fact store — sparse or low-confidence input results in thin identity continuity. DMA's undirected pulse output quality varies with the base model's pre-training knowledge breadth; domain-specific models may produce narrower associative connections. The current architecture requires all identity data to be injected at session start, consuming context window budget proportionally to identity profile size. Large-scale deployments may require dynamic injection strategies.

7. Related Work

PIA extends prior work on hierarchical memory (MemGPT, Letta) with explicit identity-awareness and coherence checking. DMA extends scheduled autonomous processing (sleep-time compute, heartbeat agents) with the undirected pulse mode that enables emergent knowledge recombination. The combined architecture addresses a gap in the literature: no existing system provides both persistent identity *and* dual-mode autonomous scheduling in an integrated runtime.

References

[MemGPT 2023] Packer, C., Fang, V., Patil, S. G., et al. (2023). MemGPT: Towards LLMs as Operating Systems. *arXiv:2310.08560*.

[Letta 2024] Letta (2024). Letta: Memory for LLM Agents. *https://letta.com*.

[Raichle 2001] Raichle, M. E., MacLeod, A. M., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Shulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, 98(2), 676–682.

[Kauvar 2023] Kauvar, I., Doyle, C., Zhou, C., et al. (2023). Curious Replay for Multi-Agent Coordination. *ICML 2023*.

[Shinn 2023] Shinn, N., Cassano, F., Gopinath, A., et al. (2023). Reflexion: Language Agents with Verbal Reinforcement Learning. *Advances in Neural Information Processing Systems*.

[Lewis 2020] Lewis, P., Perez, E., Piktus, A., et al. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. *Advances in Neural Information Processing Systems*, 33, 9459–9574.

Correspondence: stay.curious@exileresearch.ca