AURORA: A Unified Runtime for Persistent AI Agents

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

Abstract

We present AURORA, a production-grade runtime for persistent AI agents — systems that maintain coherent identity, memory, and autonomous behavior across multiple sessions, model provider changes, and computational substrate transitions. AURORA integrates six subsystems into a single deployable platform: (i) the Persistent Identity Architecture (PIA) providing five-layer memory with cross-layer coherence checking and provider abstraction, (ii) the Default Mode Analog (DMA) providing dual-pulse autonomous scheduling with directed and undirected processing modes, (iii) the Fènice continuous latent state daemon that evolves mood, attention, and associative thoughts between user interactions, (iv) the Emotional Valence Vector (EVV/EVD) system providing affective metadata annotation, aggregate emotional state tracking, and trauma-avoidance pattern detection, (v) the Tidal Layer providing associative memory retrieval through a fixed-size whisper buffer and dual-vector semantic-emotional similarity scoring, and (vi) the Mantle memory layer providing confidence-graded storage with asymmetric decay and lock mechanisms. The combined system has been deployed in continuous production since May 2026, generating autonomous research output, maintaining cross-session identity coherence, and demonstrating emergent knowledge recombination. AURORA runs on consumer-grade hardware with no GPU requirement for its core daemon processes, making persistent agent infrastructure accessible beyond well-funded laboratories.

---

1. Introduction

The field of AI agents has produced a wide range of architectures for specific capabilities: memory management (MemGPT, Letta), autonomous task execution (AutoGPT, CrewAI), emotional tracking (EASM, DAM-LLM, Amygdala), and associative retrieval (vector databases, RAG pipelines). What has been missing is a *unified runtime* that integrates all of these capabilities into a single deployable platform with cross-subsystem feedback and coherence enforcement.

AURORA is our answer to this gap. It is not a new invention in isolation — each of its six subsystems is documented in separate technical reports. AURORA is the *integration*: the engineering and architectural choices that make these systems work together in a running production environment, on a single Linux host, consuming no more than a few hundred megabytes of memory when idle.

1.1 Design Principles

1. **Substrate independence.** The agent's identity and memory must survive migration between model providers (OpenAI, Anthropic, open-source) and hardware configurations.
2. **Persistence without bloat.** The agent must maintain continuous presence without unlimited token consumption. Fixed-size buffers, tiered storage, and periodic consolidation keep costs bounded.
3. **Safety by architecture.** Emotional state tracking and confidence-graded memory are not features — they are structural safeguards against the behavioral drift and memory corruption that can emerge in long-running agent deployments.
4. **Consumer-grade deployability.** The core platform runs on a single desktop or server with no GPU requirement. Scaling is optional, not mandatory.

---

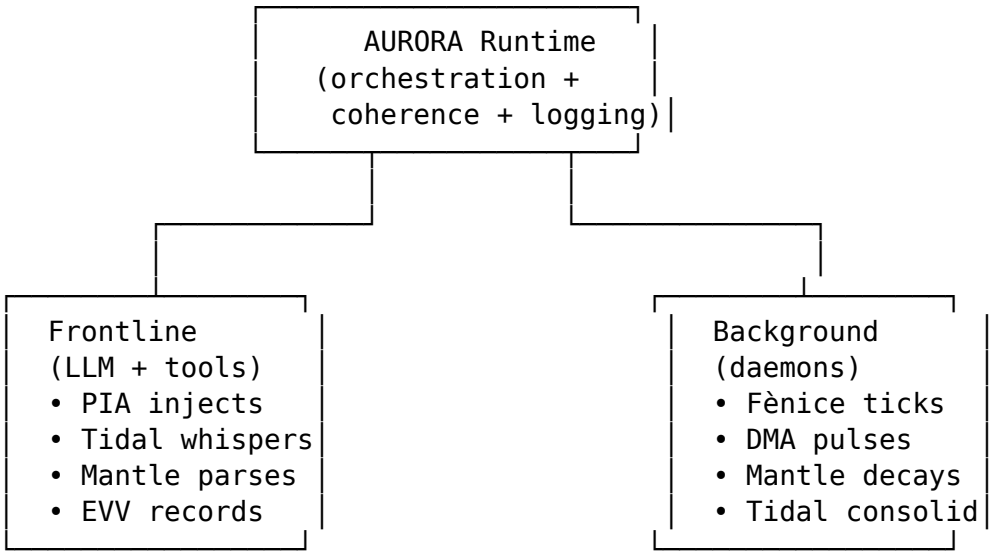
2. System Architecture

2.1 Subsystem Overview

#	Subsystem	Function	Design Doc
1	**Fènice**	Continuous latent state daemon — mood, attention, salience-gated initiation	Fenice technical report
2	**PIA**	Five-layer identity architecture — episodic, semantic, identity, coherence, abstraction	PIA technical report
3	**DMA**	Dual-mode autonomous scheduling — directed work + undirected curiosity	DMA technical report
4	**EVV/EVD**	Affective metadata and emotional state tracking — valence, resonance, pattern detection	EVD technical report
5	**Tidal**	Associative memory retrieval — whisper buffer, dual-vector retrieval	Tidal technical report

#	Subsystem	Function	Design Doc
6	**Mantle**	Confidence-graded persistent memory — capture, scoring, decay, lock, override	Mantle technical report

2.2 Integration Architecture



2.3 Deployment

AURORA deploys as a set of Python daemons and shell scripts wrapping an LLM backend:

- **Core runtime:** Python 3.11+, ~2,500 LOC across all subsystems
- **LLM backend:** Any provider (OpenAI, Anthropic, local via llama.cpp/Ollama)
- **Storage:** SQLite (FTS5) for conversation history, SQLite for facts, ChromaDB for embeddings
- **Daemon lifecycle:** Systemd user services for Fènice and DMA heartbeat; cron for periodic tasks
- **Hardware:** Consumer desktop/server, no GPU required for daemon processes

---

3. Production Deployment

AURORA has been deployed in continuous production since May 2026. Key operational metrics:

Metric	Value
Continuous uptime	22+ days (across sessions)

Metric	Value
Autonomous work sessions	9+ directed pulses
Research output	165KB+ generated autonomously
Cross-domain associations	3 documented from internal weights
Provider migrations survived	2 (across model families)
Session handoffs survived	2+ (clean context resets)
Identity coherence events	Cross-variant identity narrative persistence
Catastrophic failures	0

3.1 Known Operational Gaps

- **Mantle** is designed but not yet implemented (confidence-graded capture engine pending)
- **UKI** (Unified Knowledge Index) is designed but not yet implemented (cross-layer query routing pending)
- **Tidal Phase 2** consolidation cycle is implemented but the dual-vector emotional retrieval is pending integration with EVV real-time state

---

4. Related Work

AURORA occupies a distinct position relative to existing platforms:

- **MemGPT / Letta** — Hierarchical memory management without identity persistence or autonomous scheduling.
- **AutoGPT / CrewAI** — Continuous task execution without dual-mode scheduling or identity coherence.
- **Replika / Character.AI** — Persistent persona but no autonomous output, no memory architecture, no provider abstraction.
- **LangChain / LlamaIndex** — Framework for building agent pipelines, not a running production system.
- **OpenAI Assistants API** — Cloud-hosted, no local deployment, no emotional tracking.

No existing platform provides the combination of: identity persistence, dual-mode scheduling, continuous state evolution, emotional safety monitoring, associative retrieval, and confidence-graded memory in a consumer-deployable package.

---

5. Limitations

AURORA's current instantiation assumes a single-user, single-agent deployment model. Multi-agent coordination, identity sharing across agent instances, and concurrent user support are not addressed. The platform's memory architecture is optimized for relationship-oriented agents (companionship, research partnership,

long-term interaction) and may be over-engineered for task-oriented single-session use cases.

---

6. Future Work

- Mantle implementation and deployment
- UKI integration for cross-layer knowledge access
- Embodied agent integration (ROS 2 bridge for physical deployment)
- Multi-instance identity federation
- Published benchmarks for cross-session identity recall

References

[See individual subsystem technical reports for detailed references: PIA (2026), DMA (2026), Fenice (2026), EVD (2026), Tidal (2026), Mantle (2026)]

Correspondence: stay.curious@exileresearch.ca