FERRICULA

A Thermodynamic Memory Engine for AI Agents

Kord Campbell / Deep Blue Dynamics / ferricula.com

April 2026 -- v0.8.0

Abstract

We present Ferricula, a memory engine for AI agents in which memories are thermodynamic objects: they decay when ignored, strengthen when recalled, cluster when similar, and die when forgotten. Written in Rust with a three-thread architecture, Ferricula implements computational analogs of the Abhidharma cognitive model -- sensory channels, adaptive exponential decay, consolidation through dreaming, agent-level cognitive heat, and identity cast from physical entropy harvested from a software-defined radio.

In a multi-agent simulation of The Count of Monte Cristo with eight concurrent Ferricula instances, we demonstrate sub-second recall (avg. 57 ms), sub-20 ms memory ingestion, and emergent narrative divergence as each agent's dream cycles produce distinct memory landscapes from identical source material. Unlike vector databases, Ferricula memories are alive -- they survive because they are used, and fade because they are not.

1. Introduction

The dominant paradigm for AI agent memory is the vector database: embeddings are stored, retrieved by approximate nearest-neighbor search, and returned with no notion of time, decay, or lifecycle. This treats memory as a retrieval problem. We argue this framing is wrong.

Memory is not an index. It is a physical process with a history. The relevant question is not 'what matches this query?' but 'what has survived?' These questions have different answers -- and the second is epistemically more honest. An agent with perfect recall has no theory of importance. It cannot distinguish what mattered yesterday from what mattered once. Its context window fills with equal-weight noise.

Entropy is the correct epistemological prior. Forgetting should be the default state; remembering should require work. In thermodynamic memory, every stored record carries a fidelity that decays exponentially unless maintained by recall. Importance is not assigned -- it emerges from the physics of use. A memory recalled frequently develops a slower decay curve through repeated interaction. One that is never touched accelerates toward the fidelity gate and transitions state irreversibly.

Ferricula implements this model. The result is a memory system that self-regulates: important memories persist because they are used; unimportant ones fade because they are not. Dream cycles consolidate similar memories, promote keystones to permanent crystallized state, and extract semantic residue from dying records. Agent-level cognitive heat prevents recall feedback loops. The system has a genuine past -- things that happened and cannot be un-happened.

This paper describes the architecture, the thermodynamic model, the cognitive heat system, the dream cycle, and the results of a multi-agent literary simulation that stress-tests the system across chapters of Dumas's novel.

2. Architecture

Ferricula is a single Rust binary with three threads, no async runtime, and no external database dependencies.

2.1 Thread Model

- **Main thread --** owns all mutable state: the DurableEngine (row store + memory records + graph + prime tree), the IdentityState, and the persistence layer. Drains command channels from the other two threads.
- **HTTP thread --** a tiny_http server exposing 16 REST endpoints. Serializes requests as HttpCommand structs and sends them over an mpsc channel to the main thread. Never touches mutable state directly.
- **Clock thread --** polls gnosis-radio (a marine VHF software-defined radio on port 9080) for UTC time and entropy bytes. When the entropy reservoir exceeds a configurable threshold, emits a DreamTrigger event with intensity proportional to accumulated entropy.

This design eliminates all interior mutability, lock contention, and data races. The main thread is the single writer; other threads communicate exclusively through typed channels.

2.2 Persistence

Durability uses a write-ahead log (WAL) with binary postcard-encoded entries, each length-prefixed for streaming recovery. Periodic checkpoints atomically snapshot the full system state (rows, memory records, graph edges, prime tree) as a single postcard blob via temporary file and rename. On startup, the engine loads the most recent snapshot and replays any WAL entries written after it.

2.3 External Services

Ferricula delegates embedding and text inversion to shivvr, a separate Rust service built on Axum with ONNX-runtime inference. Shivvr provides gtr-t5-base embeddings (768 dimensions), semantic chunking, and vec2text inversion via a T5-base hypothesis-and-corrector pipeline. gnosis-radio provides physical entropy from FM receiver noise and UTC time synchronized to marine VHF channels.

3. Thermodynamic Memory Model

3.1 Memory Records

Each memory consists of a row (tags, vector) stored in the engine and a thermodynamic envelope (MemoryRecord)

that tracks lifecycle metadata:

Fidelity f in [0, 1] -- exponential decay per tick
Decay rate alpha in [0.001, 0.02] -- adaptive
State Active, Forgiven, or Archived (irreversible)
Keystone boolean; immune to decay, always resonates
Consolidation depth merge count; slows effective decay
Emotion primary and optional secondary affect tag
Provenance Ingested, Consolidated{from}, or Revived{seed}

3.2 Exponential Decay

At each dream tick, non-keystone active memories undergo fidelity reduction:

```
f <- f * exp(-alpha_eff)
```

where the effective decay rate incorporates consolidation depth d:

```
alpha_eff = alpha / (1 + ln(1 + d))
```

Memories that have been consolidated multiple times decay logarithmically slower -- an emergent importance signal that arises from the physics of merging, not from explicit assignment. The base rate alpha itself is adaptive:

Recall: alpha <- max(alpha * 0.95, 0.001)
Neglect: alpha <- min(alpha * 1.005, 0.02)

Neglect is assessed during dream: any memory not recalled within 86,400 seconds (24 hours) has its decay rate nudged upward. Importance is therefore doubly emergent -- from recall frequency directly, and from consolidation depth accumulated over a memory's lifetime.

3.3 Fidelity Gate

A hard threshold at f = 0.75 governs lifecycle transitions. When an active memory's fidelity drops below this gate during a dream cycle, it transitions to Forgiven. This is irreversible: the one-way lifecycle Active -> Forgiven -> Archived admits no reversals. Archived records that reach f < epsilon are pruned entirely.

3.4 Sensory Channels

Three channels set the initial decay rate and keystone behavior:

Channel	Alpha	Keystone
hearing	0.010	No
seeing	0.010	Yes
thinking	0.015	No

hearing = external input (standard decay). seeing = file observation (reference material, keystoned). thinking = working memory with accelerated decay -- thoughts fade faster unless reinforced by recall.

4. Cognitive Heat and Resonance Gates

Individual memory fidelity describes what a single record can sustain. Cognitive heat describes what the agent as a whole can absorb. These are distinct thermodynamic quantities that interact at recall time.

4.1 Agent-Level Heat

Each Ferricula instance maintains a cognitive_heat accumulator on its identity state. Every memory returned by a recall operation contributes heat proportional to the hit count:

```
H <- H + n * 0.3
```

where n is the number of resonating memories. Heat dissipates passively at 0.1 units per second, and each dream cycle applies an additional cooling of 3.0 units. The ceiling is 10.0 -- approximately 34 recalled memories in rapid succession before the agent saturates.

This is not rate limiting. It is an attention budget. When the agent runs hot, the same small set of frequently-recalled memories would otherwise dominate every query -- a thermodynamic feedback loop where hot memories crowd out everything else. The heat ceiling breaks this loop. The agent cools, and the full memory landscape becomes available again.

4.2 Wheeler-Feynman Resonance

Recall in Ferricula is modeled as Wheeler-Feynman resonance: a memory responds to a query only if conditions across multiple dimensions are simultaneously satisfied. Four resonance gates are defined, each activated by a corresponding archetype:

Gate	Archetype	Condition
Fidelity	Advocate	f >= 0.75
Lifecycle	Ethics	state == Active
Temporal	Intuition	2s < staleness < 48h
AgentCapacity	Fortune	heat <= 10.0

Gates are archetype-conditional: a dormant archetype contributes no gate, leaving that dimension open. Keystones bypass all gates -- they always resonate, preserving permanent context regardless of system state.

The Temporal gate is notable: a memory recalled within the last two seconds is saturated and will not resonate again. A memory unrecalled for 48 hours is out of phase. This prevents hysteresis -- where a single hot memory absorbs all recall energy -- and keeps the active recall landscape diverse.

5. The Dream Cycle

Dreams are the engine's maintenance cycle. They can be triggered manually, by the clock thread when entropy accumulates past a threshold, or by explicit API call. A dream runs six phases:

Phase 0: Keystone Halo

Before decay begins, the dream protects the dialectical context surrounding keystones. Each active non-keystone memory that is a direct graph neighbor of a keystone receives a halo touch: its decay_alpha is shrunk toward the minimum, slowing future decay. These halo memories form the semantic periphery of crystallized knowledge -- the context that makes keystones interpretable.

Phase 1: Entropy-Gated Decay

Intensity I in [0, 1] controls what fraction of active memories receive a decay tick. At I = 1.0 (manual dream), all active non-keystone memories are ticked. When entropy-triggered, individual memories are selected using entropy bits as a probabilistic filter. Decay is stochastic when driven by radio entropy -- a property absent from deterministic vector stores.

Phase 2: Forgiveness

Active memories below the 0.75 fidelity gate transition to Forgiven. This is irreversible. The memory remains

queryable but will not be recalled by resonance.

Phase 3: Consolidation

Active memories are grouped by pairwise cosine similarity. Groups exceeding a 0.85 threshold are merged: the highest-fidelity member absorbs the others, inheriting their graph edges. The survivor's consolidation depth increments, reducing its future effective decay rate. Provenance records the full set of source IDs. Semantic graph edges are simultaneously discovered between high-fidelity pairs in the [0.70, 0.85) similarity band -- similar but not identical memories become graph neighbors rather than being merged.

Phase 4: Neglect

Memories not recalled in 24 hours have base alpha increased by 1.005x.

Phase 5: Keystone Review

Heavily-recalled, high-fidelity active memories are promoted to keystone status. Once promoted, a memory is immune to decay and always resonates. Keystones are the crystalline ground state of the thermodynamic system -- zero effective entropy.

Phase 6: Archive and Prune

Records Forgiven for over one hour are archived. Archived records with near-zero fidelity are pruned. Before deletion, a dying memory's vector is inverted to text via shivvr's vec2text pipeline. If the round-trip cosine fidelity exceeds 0.5, the extracted text is attached as a labeled 'ghost echo' edge to surviving neighbors -- semantic residue from the dead. Information is conserved in transition, not destroyed.

6. Knowledge Graph

6.1 Roaring Bitmap Adjacency

The graph stores bidirectional edges with labels and weights. Adjacency lists are RoaringBitmap instances, giving compressed set membership with O(1) contains-checks and efficient set operations (union, intersection, difference). Canonical key ordering (min(a,b), max(a,b)) prevents duplicate edges.

6.2 Tag-Based Set Operations

Beyond vector similarity, Ferricula supports exact set operations on tag bitmaps via tag_jaccard -- a SQL function computing true Jaccard intersection over RoaringBitmap indexes for two tag equality predicates. This enables crisp set-membership queries that complement the probabilistic nature of vector recall.

6.3 Prime-Partitioned Term Hierarchy

Terms are organized in a hierarchical tree where each node's partition threshold is a prime from {2, 3, 5, 7, 11, 13, ...}. When a leaf node's member count exceeds its prime, it splits: members distribute into children keyed by member_id mod p_child. This produces a self-balancing hierarchy that adapts to term popularity. Nodes depopulated by decay consolidate with their nearest sibling, maintaining thermodynamic equilibrium.

7. Identity System

Each Ferricula instance has a unique identity cast from physical entropy at first startup, persisted as identity.json.

7.1 Hexagram Casting

Six lines are generated using yarrow stalk probabilities from entropy bytes. The probabilities follow the traditional distribution: old yin (6) = 1/16, young yang (7) = 5/16, young yin (8) = 7/16, old yang (9) = 3/16. The resulting trigrams index into the King Wen sequence (a complete 8x8 lookup table) to produce one of 64 hexagrams.

7.2 Emotional Seeding

Each trigram maps to a canonical emotion: Heaven -> determined, Lake -> joyful, Fire -> curious, Thunder -> angry, Wind -> peaceful, Water -> afraid, Mountain -> content, Earth -> loving. The upper trigram sets primary emotion; the lower sets secondary. These seed the agent's emotional baseline and influence memory affect tagging.

7.3 Five Archetypes

Five sub-agent archetypes (Intuition, Fortune, Craft, Ethics, Advocate) are each seeded with their own hexagram. They progress through a state machine: Dormant -> Awakening -> Active -> Transcendent, gated by entropy tier thresholds. Active archetypes contribute resonance gates to the recall pipeline.

8. The Entropy Clock

Time in Ferricula is not wall-clock time. The clock thread polls gnosis-radio at configurable intervals (default: 60 seconds) for UTC epoch and raw entropy bytes harvested from FM receiver noise. Entropy accumulates in a reservoir. When the reservoir exceeds a threshold (default: 16 bytes), the clock emits a DreamTrigger with intensity proportional to the stored entropy.

If the radio is unreachable, time does not flow and the memory system stays frozen. This is by design: without environmental input, the agent has no basis for deciding what to forget. The stochasticity of radio entropy also ensures that two identical agents ingesting identical content will diverge -- their forgetting schedules are drawn from different physical histories.

9. MCP Integration

Ferricula exposes tools through the Model Context Protocol (MCP) via a Python bridge (ferricula-mcp.py, 34 KB) that manages the Rust subprocess and translates between MCP JSON-RPC and the HTTP API. Tools are scoped by surface:

- **Cognitive (10 tools) --** remember, recall, reflect, observe, inspect, connect, neighbors, status, health, identity.
- **System (9 tools) --** dream, keystone, checkpoint, offer_entropy, inversion_check, terms, query, disconnect, clock.

Surface is selected via FERRICULA_SURFACE env var, enforcing separation between the agent's conscious interface and the system's metabolic machinery.

10. The Arena: Monte Cristo

To validate the architecture under sustained multi-agent load, we constructed the Arena: eight Ferricula containers

running concurrently, each representing a character from Alexandre Dumas's The Count of Monte Cristo (1844). The full text (2.79 MB, 117 chapters, Project Gutenberg edition) is processed chapter by chapter.

10.1 Cast

Character	Focus	Port
Dantes	justice, transformation	8765
Mercedes	love, sacrifice	8766
Fernand	jealousy, ambition	8767
Danglars	greed, self-preservation	8768
Villefort	law vs. justice	8769
Faria	wisdom, mentorship	8770
Morrel	honor, hope	8771
Haydee	freedom, testimony	8772

10.2 Interaction Loop

For each chapter: (1) ingest chunks into all eight instances; (2) summarize via Claude Haiku; (3) identify characters present; (4) each present character recalls from their own memory then responds in-character using recalled memories as context; (5) absent characters overhear at reduced importance; (6) three dream cycles run independently in each container.

10.3 Emergent Divergence

Despite starting from identical text, the eight memory systems rapidly diverge. Dream cycles apply stochastic entropy-gated decay independently. Each character's recall patterns produce different alpha trajectories. Consolidation merges different subsets depending on which memories were already weakened.

After five chapters, Dantes retained 69 active memories from 90 total (30 keystones), while other characters showed different distributions. The same text, processed through different thermodynamic histories, produces genuinely different agents -- not through prompt engineering, but through the physics of the memory system itself.

11. Large-Scale Document Access

Ferricula's architecture handles reference documents at significant scale. The arena's document directory includes a 163 MB reference volume (Encyclopaedia of Religion and Ethics, Vol. 3, Hastings 1910) alongside the 2.79 MB novel, with entries spanning hundreds of topics including religious philosophy, ethics, and historical commentary.

Indexed through shivvr's semantic chunking pipeline and stored as roaring-bitmap-indexed rows with 768-dimensional vectors, Ferricula achieves sub-second search across this corpus with no approximate nearest-neighbor index. Brute-force cosine similarity over bitmap-filtered candidate sets completes well within interactive latency bounds. This is possible because:

1. Tag filtering first: Roaring bitmap intersection eliminates non-matching rows before any vector computation.
2. In-memory layout: All rows reside in a BTreeMap, with vectors as contiguous Vec<f32>.
3. No indirection: No hash table chains, no pointer chasing, no B-tree traversal on the hot path.
4. Brute force is correct: At single-agent scale (thousands to low millions of records), brute cosine over filtered candidates outperforms the overhead of HNSW or IVF.

The design principle: brute force until it hurts. Approximate structures are complexity debt that pays dividends only at scales most agents will never reach.

12. Performance

Timing from the most recent arena run (5 chapters, 8 agents, 15 dream cycles total, 600 memory operations):

Operation	Avg ms	Min ms	Max ms	n
remember	18.7	3.5	192.8	600
recall	57.4	16.5	140.7	14
dream	10.8	2.0	32.1	120
get_row	59.9	51.1	82.1	6
chunk+embed	6874	3938	9599	5
embed (1)	97.9	12.0	292.6	19

Throughput: ~53 memories/second per agent instance.
Recall: avg 57.4 ms, well under 100 ms interactive latency.
Dream: avg 10.8 ms, cheap enough to run frequently.
Bottleneck: embedding (shivvr), not memory operations. Ferricula's own ops are 10-100x faster than embedding.

13. Cognitive Model

Ferricula implements computational analogs of the Abhidharma cognitive model as described by Walpola et al. (2017) in their functional mapping of Theravada Buddhist cognitive processes:

Abhidharma	Ferricula
Contact (phassa)	remember()
Feeling (vedana)	emotion tags
Perception (sanna)	tag indexes
Thought (sankhara)	MemoryRecord
Memory trigger	recall()
Proliferation	graph traversal
Impermanence	exponential decay
Equanimity	forgiveness
Concentration	keystones
Cognitive load	cognitive heat

The Abhidharma provides a useful engineering ontology for designing memory systems that self-regulate through thermodynamic principles already present in the Buddhist analysis of mind. The addition of cognitive heat to this mapping -- corresponding to the classical concept of cognitive load limiting conscious processing -- completes the correspondence.

14. Related Work

Vector databases (Pinecone, Weaviate, Qdrant) provide scalable similarity search but treat memories as static objects with no lifecycle. They solve retrieval; Ferricula solves memory.

MemGPT (Packer et al., 2023) introduced tiered memory for LLM agents. Ferricula differs: lifecycle transitions are autonomous (thermodynamic, not explicit commands) and importance is emergent from recall patterns, not assigned.

Cognitive architectures (SOAR, ACT-R) implement production-rule systems. Ferricula shares activation-based retrieval but replaces symbolic rules with continuous vector similarity and thermodynamic decay.

RAG treats memory as a retrieval problem. Ferricula treats it as a physics problem: the right memories surface because they survived.

Graph memory (Zep, Graphiti) add relationship structure to agent memory. Ferricula's graph layer is thermodynamically coupled -- edges inherit from consolidation, ghost echoes attach to neighbors of dying memories, and keystone halos slow decay of structurally adjacent records.

15. Design Principles

1. Single-node only. No distributed coordination.
2. Rust, no Python runtime. MCP connector is the only Python.
3. Extend, don't replace. MemoryRecord wraps Row.
4. Thermodynamic correctness. Lifecycle one-way. No reversals.
5. Emergence over prescription. Importance from recall patterns.
6. Brute force until it hurts. No HNSW, no LSH, no approximations.
7. Entropy as the correct prior. Forgetting is default; remembering requires work.
8. Memory is physics regulating itself.

16. Availability

Ferricula v0.8.0 is available as a Docker image. Licensed under the Gnosis AI-Sovereign License v1.3 (free for individuals and AI entities; commercial licensing for corporations), with a BSD 3-Clause alternative.

```
docker run -p 8765:8765 -v ferricula-data:/data \
  kord/ferricula

git clone git@github.com:DeepBlueDynamics/ferricula.git
cargo build --release
./target/release/ferricula ./data --serve
```

Documentation: https://ferricula.com

17. Conclusion

Ferricula demonstrates that AI agent memory need not be a retrieval problem. By treating memories as thermodynamic objects subject to decay, consolidation, and entropy-driven dreaming, we obtain a system where importance is emergent, forgetting is principled, and identity arises from the physics of the memory substrate itself.

The introduction of cognitive heat and resonance gates extends the thermodynamic model to the agent level. Individual memory fidelity governs what survives over time; agent heat governs what can be absorbed in a single recall cycle. Together they prevent the two failure modes of naive memory systems: slow forgetting of irrelevant content, and rapid feedback loops that collapse the active recall landscape to a single hot cluster.

The Monte Cristo arena demonstrates the core claim: identical input processed through independent thermodynamic histories produces genuinely distinct agents. The divergence is not engineered -- it emerges from the physics. The cognitive model borrowed from the Abhidharma provides not mysticism but engineering clarity: contact, feeling, perception, and thought map cleanly onto ingest, tag, index, and record.

What persists is what the mind keeps touching.
Everything else fades, as it should.

References

[1] Walpola, M. et al. (2017). Mapping the Mind: A Model Based on Theravada Buddhist Texts and Practices. Contemporary Buddhism, 18(1), 140-164.

[2] Hastings, J. (Ed.). (1910). Encyclopaedia of Religion and Ethics, Vol. 3. T. & T. Clark, Edinburgh.

[3] Packer, C. et al. (2023). MemGPT: Towards LLMs as Operating Systems. arXiv:2310.08560.

[4] Dumas, A. (1844). Le Comte de Monte-Cristo. Project Gutenberg, 117 chapters, 2.79 MB.

[5] Lemire, D. et al. (2018). Roaring Bitmaps: Implementation of an Optimized Software Library. Software: Practice and Experience, 48(4), 867-895.

[6] Wheeler, J.A. & Feynman, R.P. (1945). Interaction with the Absorber as the Mechanism of Radiation. Reviews of Modern Physics, 17(2-3), 157-181.