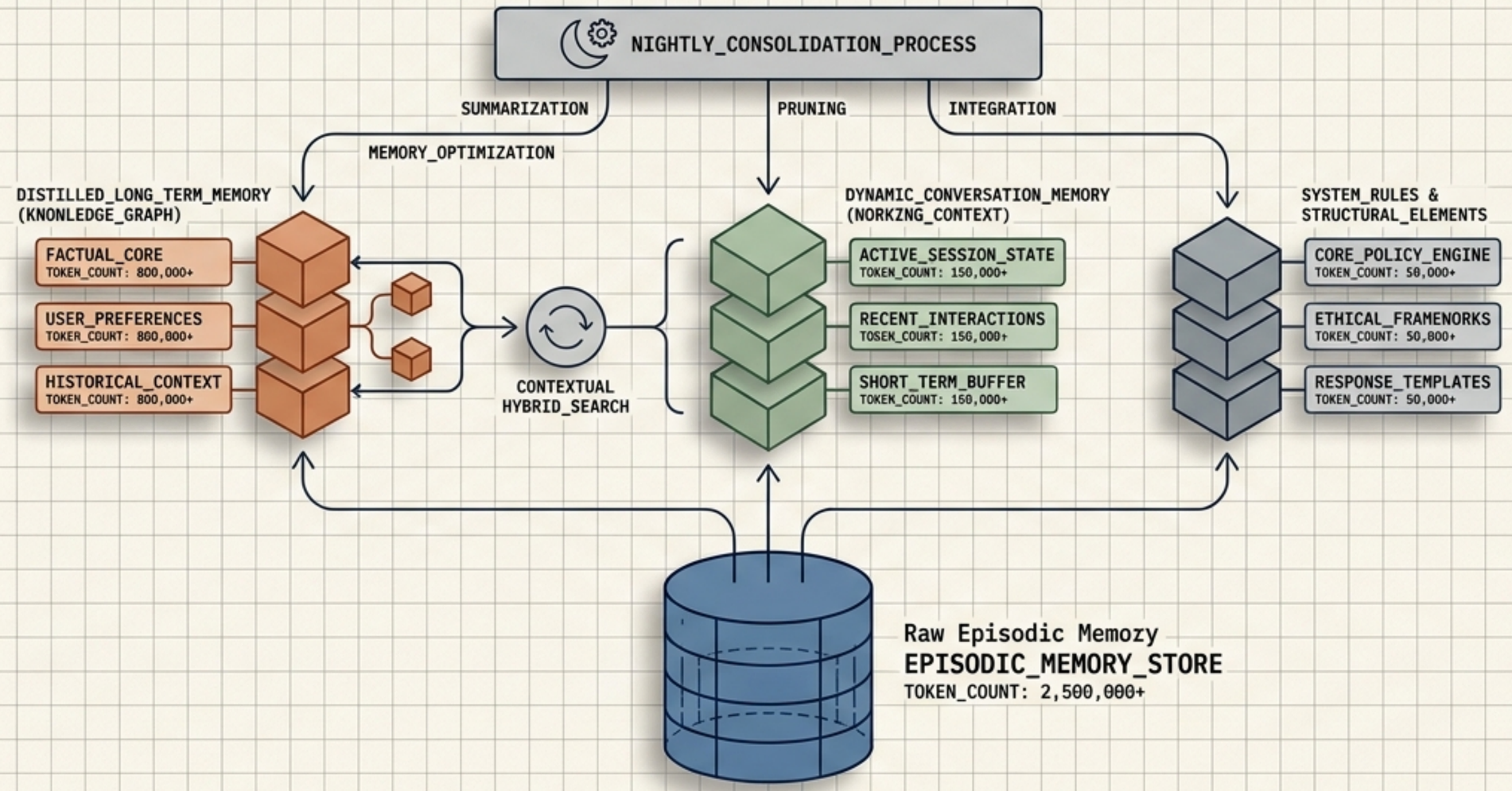
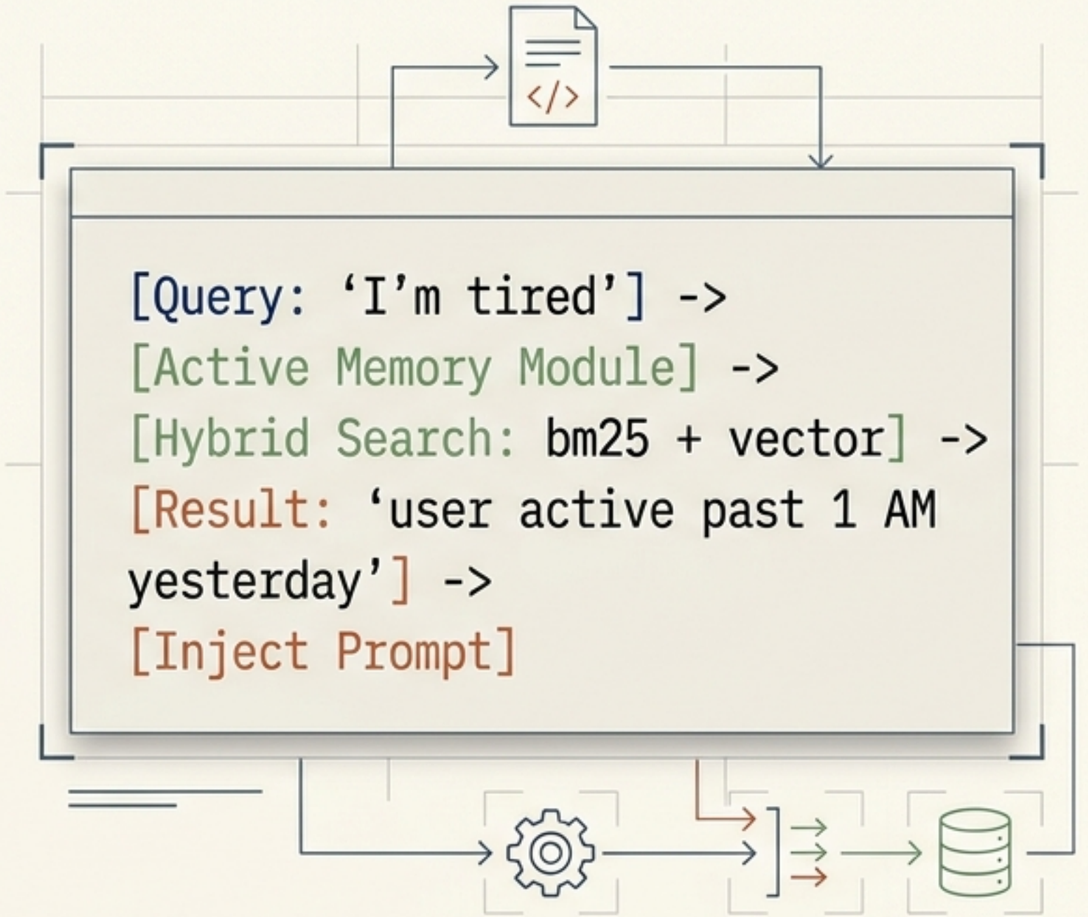
The Architecture of Feeling Known

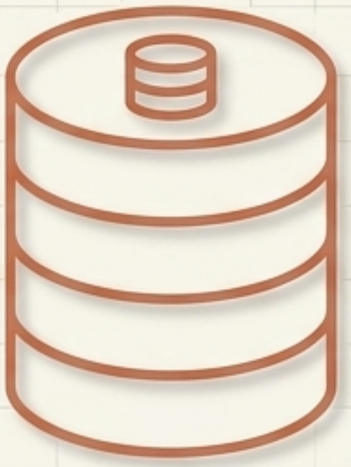
Engineering an AI memory system through three-tier storage, hybrid search, and nightly consolidation.





This single sentence is the difference between an AI that acts like a pet and an AI that acts like a stranger.

Personality File =
Character



Memory System =
Relationship

Every AI assistant is solving “**how to answer better.**” This architecture solves a different problem: **how to make the user feel known.** The gap between a character and a friend isn’t a smarter model—it is **shared history.**

Seconds to Hours	The Current Session	Days to Months
Raw, continuous observation	Active conversational tracking	Distilled, permanent facts

Memory is not a single database. It requires three discrete tiers because each tier solves a fundamentally different timescale problem.

Episodic Memory

Storage:
SQLite + FTS5 + sqlite-vec

Growth Trigger:
Every 45s (screen observation), chat message, screen reaction

Capacity:
Unlimited (Local disk)

Conversation Memory

Storage:
JSONL + summary

Growth Trigger:
Every message

Capacity:
Dynamic (Compacts at ~500K tokens. Oldest 70% summarized, newest 50 raw messages kept).

Long-Term Memory

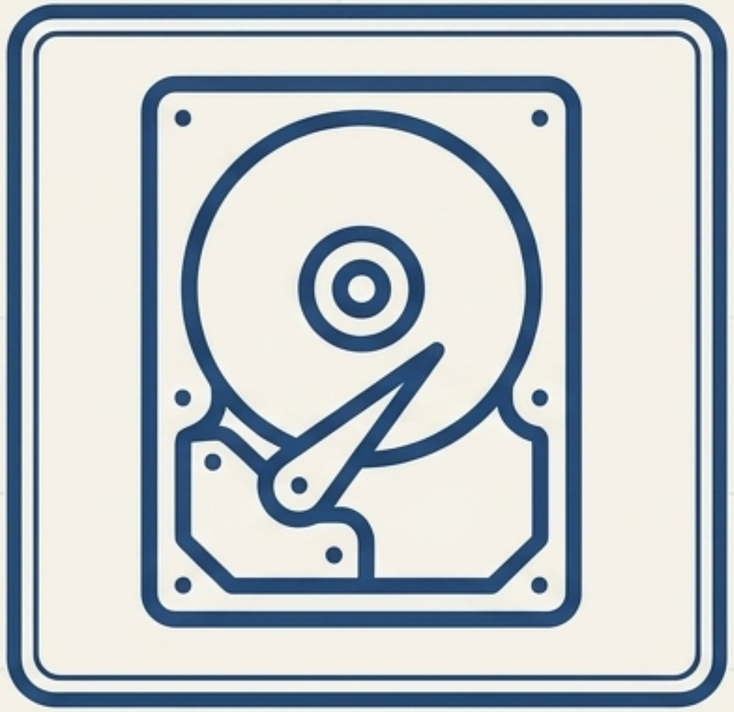
Storage:
JSON file

Growth Trigger:
Nightly consolidation extracts 3-5 top facts

Capacity:
Hard cap of 100 entries.

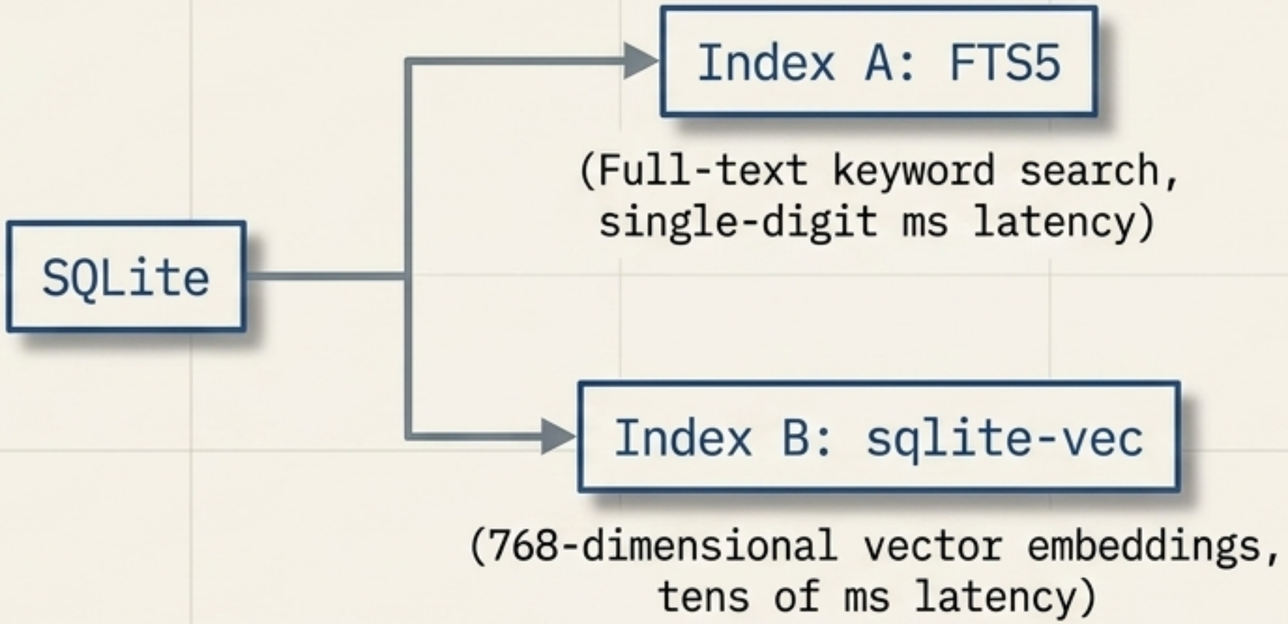
Each tier maps to human cognitive functions—from raw sensory input to working memory to crystallized knowledge.

Episodic Memory Implementation



The Raw Material

Append-only observation log. Generates 50,000–80,000 entries per month.



Design Rationale: Why no vector database?

Because this runs locally. Every dependency is a failure mode. One single file. One .dylib. No network calls. Close the laptop, open it tomorrow, and the memory remains perfectly intact.

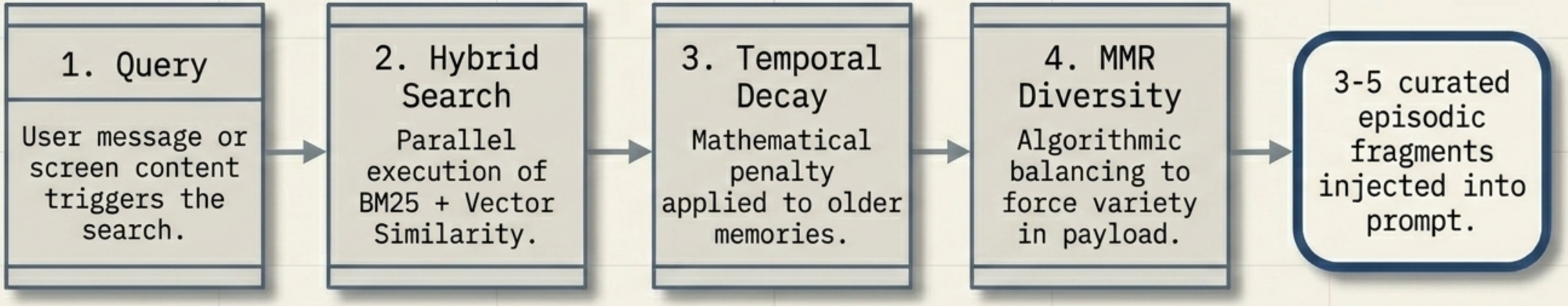
Enter the Active Memory module.

The hard part isn’t writing to SQLite.

`RETRIEVE_ACTIVE_MEMORY_ROW: USER_QUERY_PROXIMITY_MATCH > 0.95`

The hard part is taking a single user sentence, finding the 3-5 most relevant memories out of thousands in milliseconds, and injecting them into the prompt so the model remembers naturally.

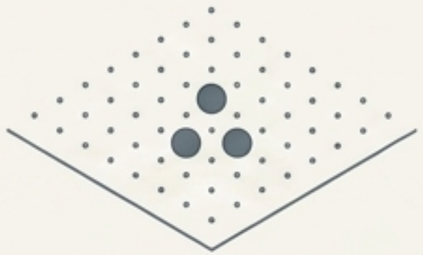
Episodic Memory Implementation



Query: "I'm so tired."

"I'm tired of this bug."
(Last week)

High Precision, Narrow Context

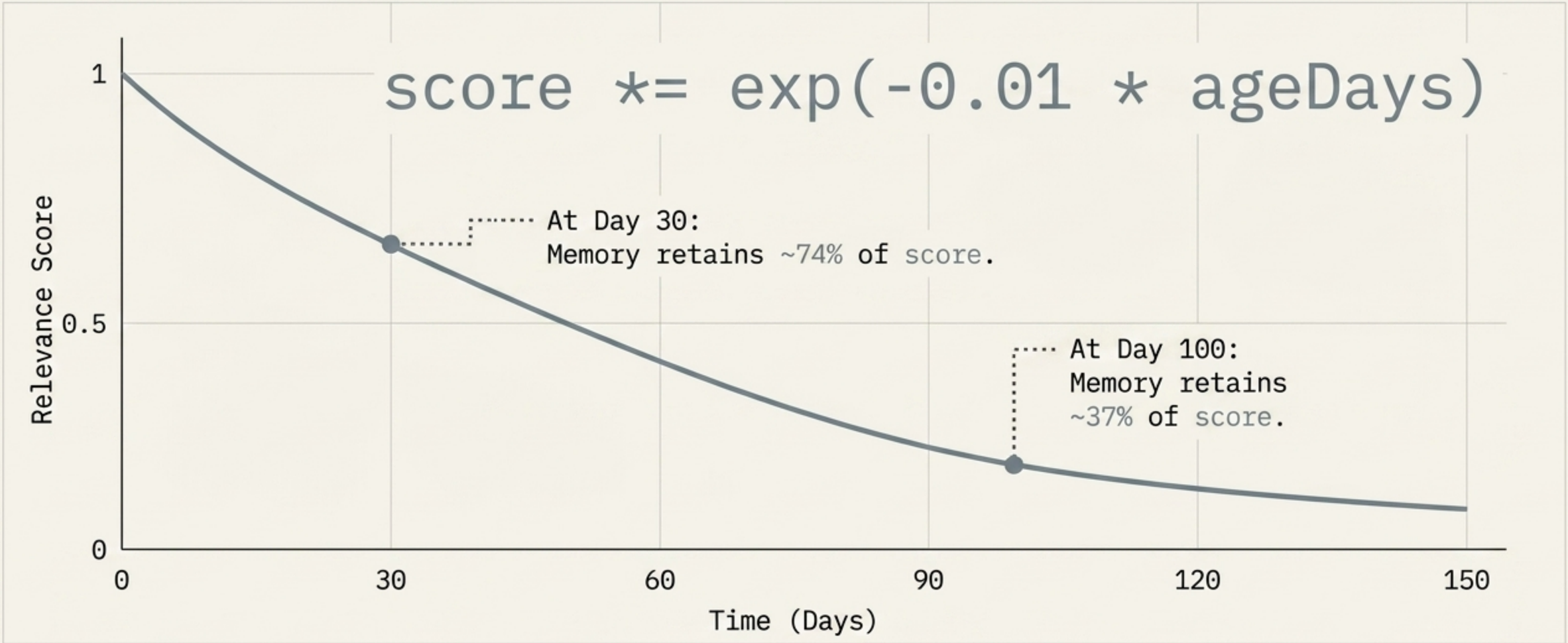


"User was active past 1 AM."
(2 days ago)

High Recall, Semantic Context

BM25 catches literal keywords. Vector catches semantic meaning.
Merging them guarantees both precision and recall without sacrificing either.

Temporal Decay in Memory Implementation



Recency matters for relationships. “You were up late yesterday” implies care; “You were up late six months ago” is just data retrieval. The math ensures recent memories automatically win ties.

0.7 * relevance - 0.3 * max_similarity

- Hotpot
- Hotpot
- Hotpot
- Hotpot

Without MMR:
Finite prompt space wasted.

- Hotpot
- Late nights
- Real cat named Mochi

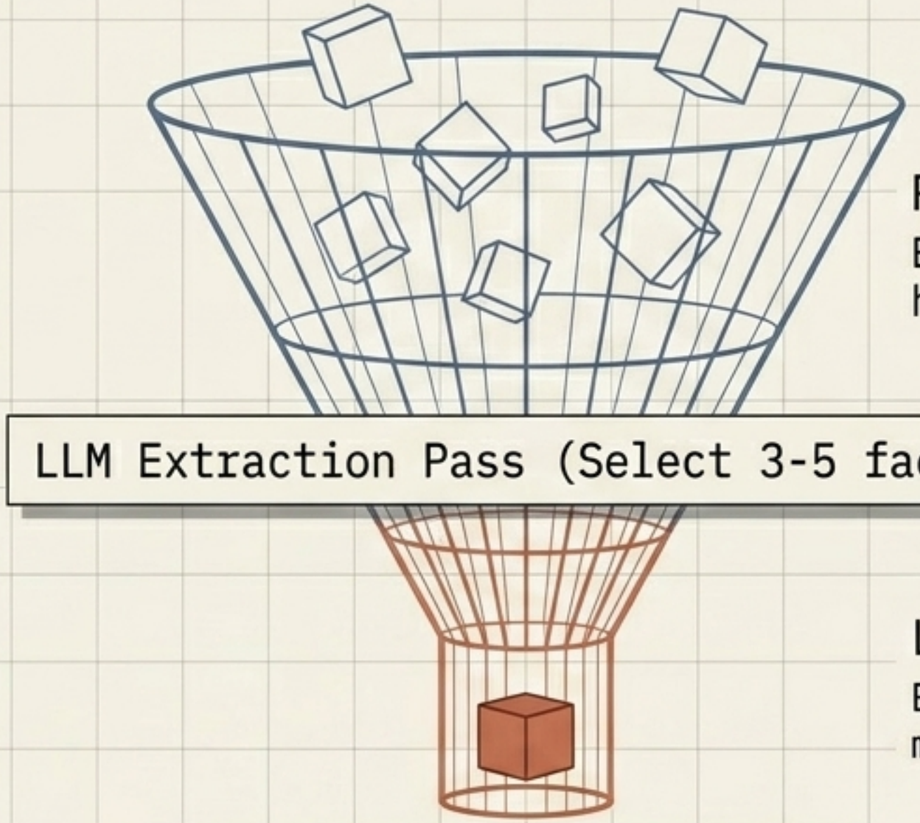
With MMR:
Diverse memory pool for the model.

The algorithm is greedy. It picks the highest-scoring result first, then strictly penalizes remaining candidates if they are too similar. It forces variety out of a repetitive interaction history.

23:30

Nightly Consolidation: The AI Dreams.

During the day, data is noisy, scattered, and fragmented. Every night at 23:30, the system mirrors human sleep—sorting through the day's episodic mess to promote only the most vital facts to permanent storage.



Raw Episodic Fragments

Example: "User said so hungry at 14:32"

LLM Extraction Pass (Select 3-5 facts)

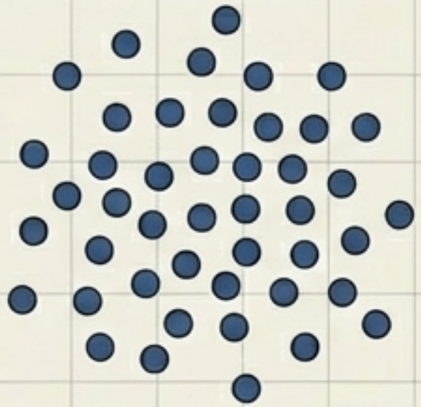
Long-Term Pattern

Example: "User often gets hungry mid-afternoon, might skip lunch."

100 Entry Hard Cap

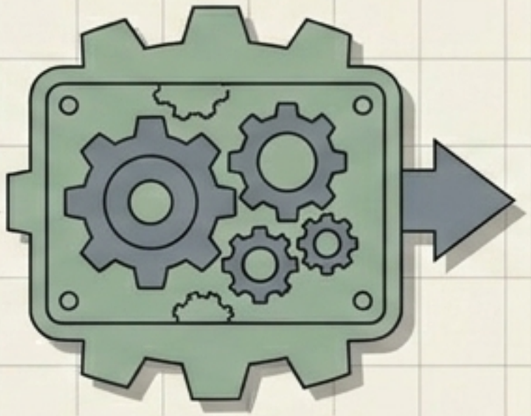
The cap is an economic forcing function. 100 entries = ~500 tokens in the stable prefix. This forces the dreaming system to ruthlessly prioritize: What really matters about this person?

Scenario: You mention your sister's birthday in hour 1 of a 3-hour conversation. Without the flush, it gets buried in the oldest 70% and summarized away into nothing.

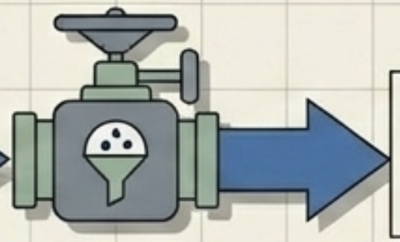


Raw Conversation Tokens

Pre-Compaction Flush



500K Token Summarizer

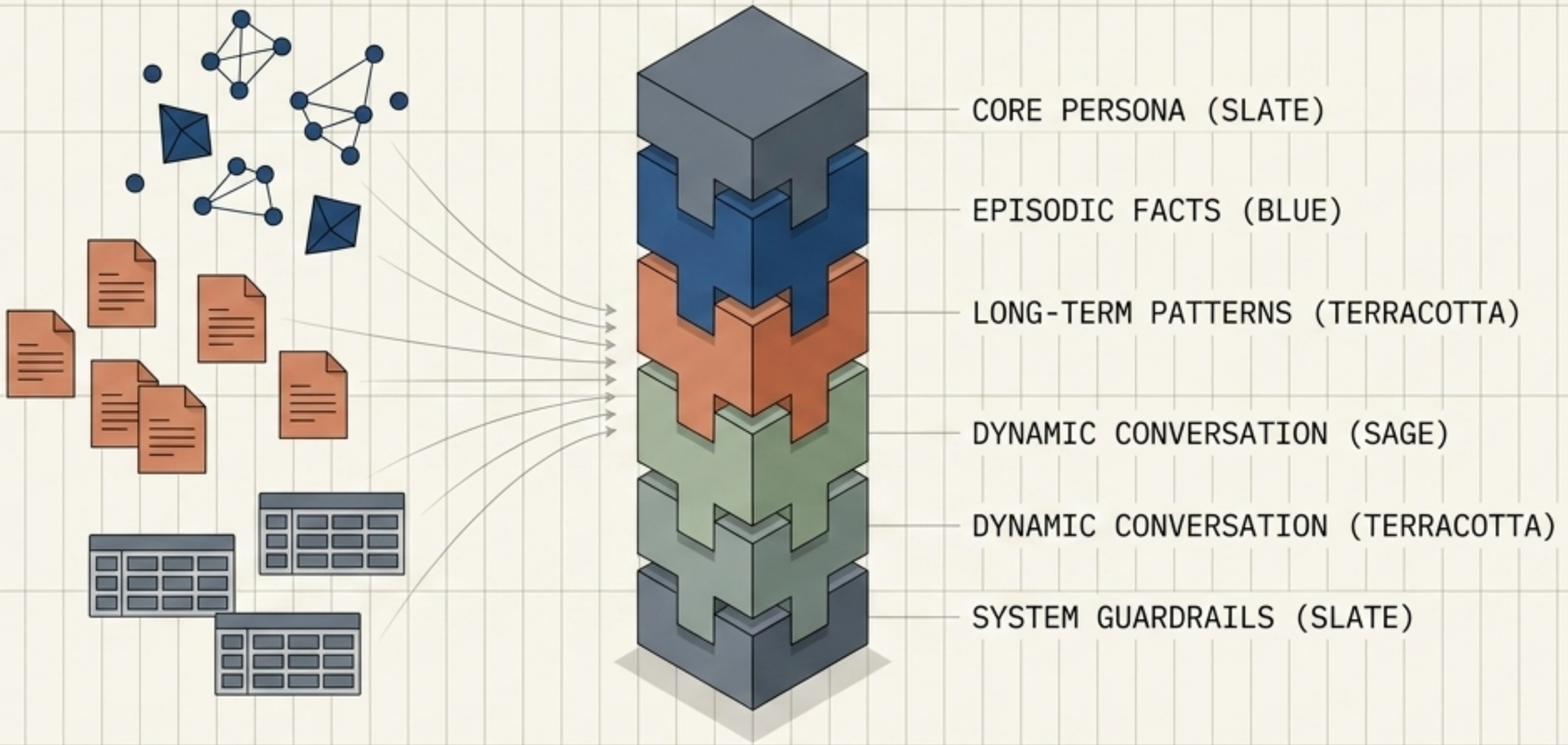


Episodic Memory

Resolution: The flush intercepts. It extracts important facts directly into episodic memory before the compressor touches them.

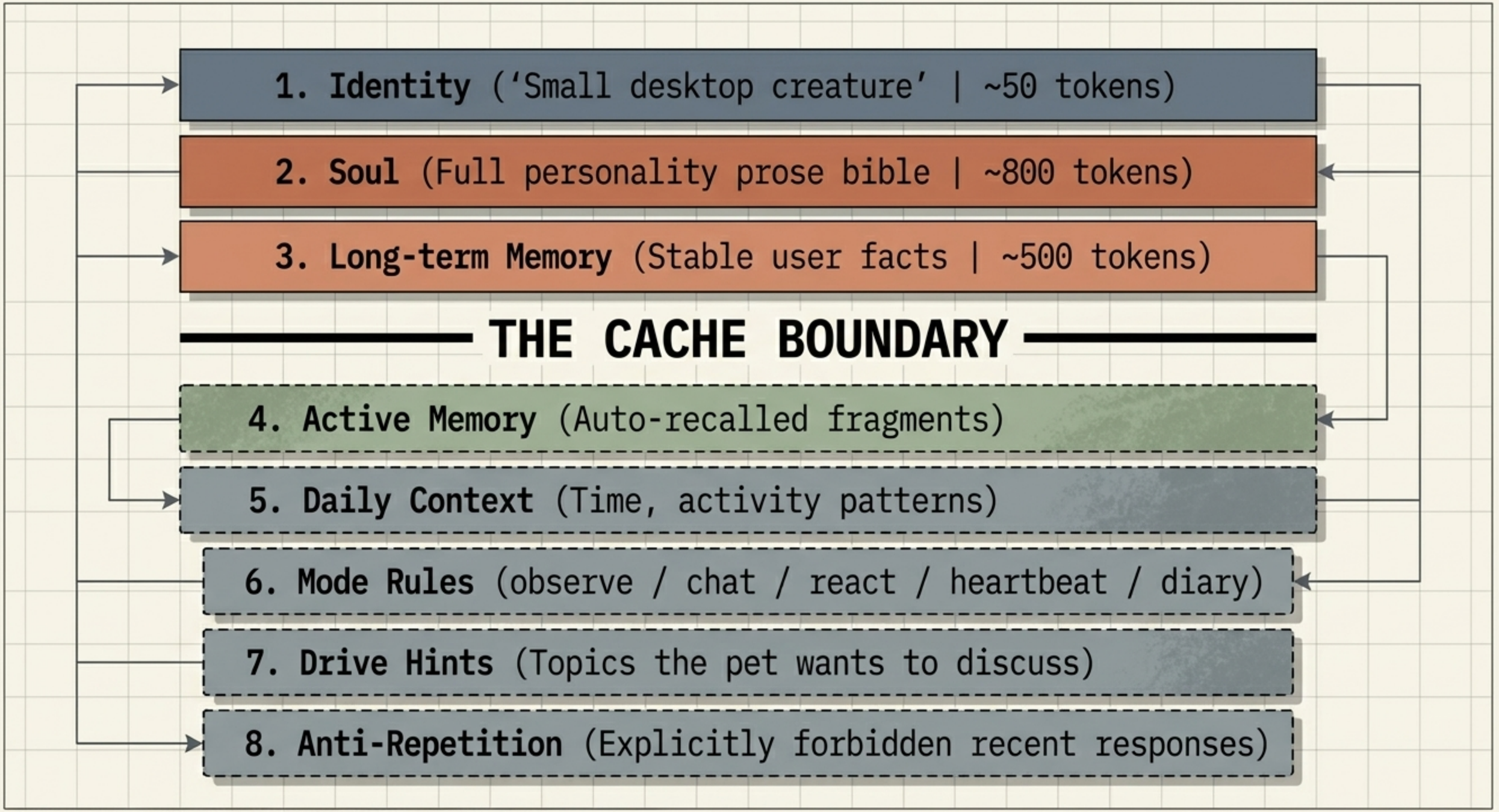
A small mechanism that closes a fatal information leak.

SYSTEM ARCHITECTURE V2.0
DOCUMENTATION // MEMORY SYSTEMS

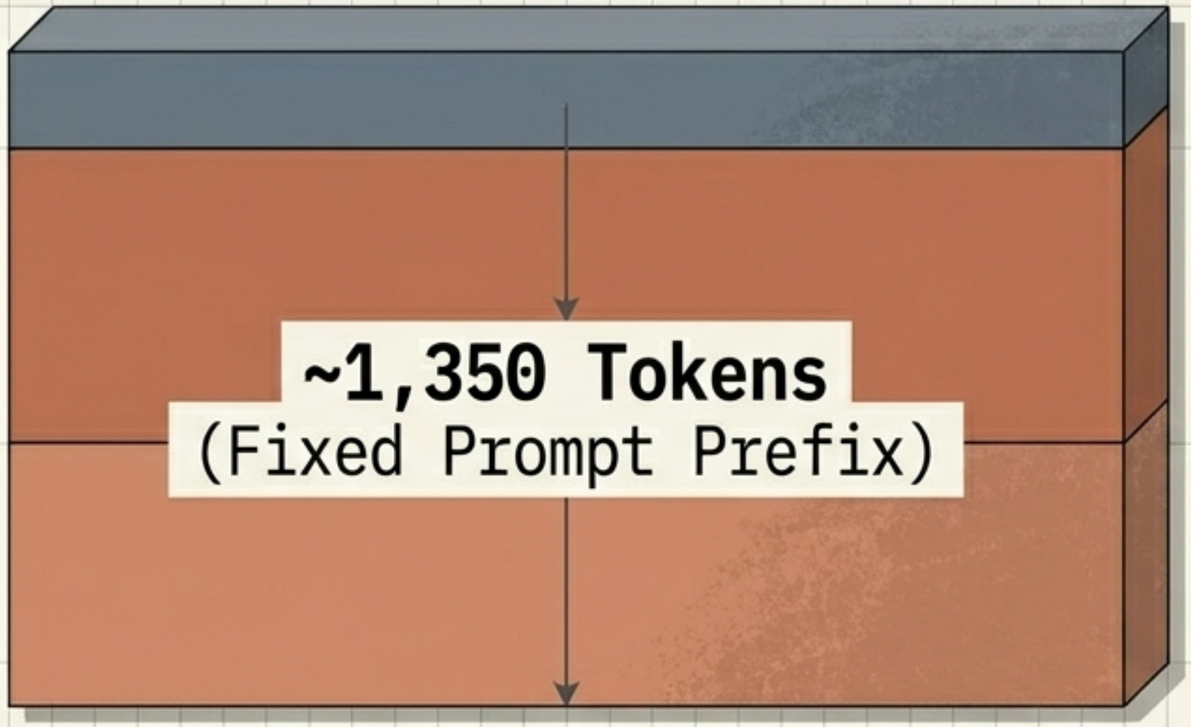


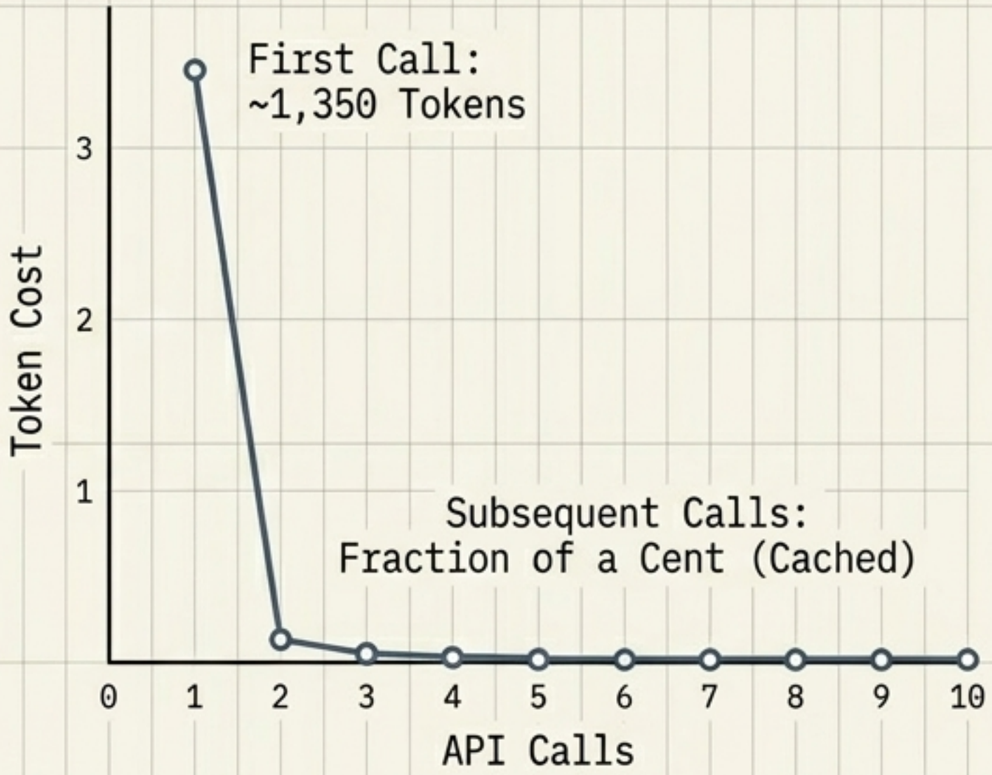
Memory is stored. Memory is retrieved. The final step: Assembly.

Turning isolated data points into the 8-Layer Prompt Engine that gives the AI its voice.


1. Identity ('Small desktop creature' | ~50 tokens)
2. Soul (Full personality prose bible | ~800 tokens)
3. Long-term Memory (Stable user facts | ~500 tokens)
THE CACHE BOUNDARY
4. Active Memory (Auto-recalled fragments)
5. Daily Context (Time, activity patterns)
6. Mode Rules (observe / chat / react / heartbeat / diary)
7. Drive Hints (Topics the pet wants to discuss)
8. Anti-Repetition (Explicitly forbidden recent responses)

SYSTEM ARCHITECTURE V2.0
DOCUMENTATION // MEMORY SYSTEMS & PREFIX CACHING





Layers 1–3 form a fixed prompt prefix. Through prefix caching, these 1,350 tokens (80%+ of the stable context) are nearly free after the first call.

A desktop pet speaks dozens of times an hour. Re-processing full context would bankrupt the system. You pay full price once, and every subsequent call within the TTL costs a fraction of a cent.

Daily Context (Layer 5)

It's 2 AM, user has been coding for 4 hours.

Mode Rules (Layer 6)

[observe] : Silent recording.

[chat] : Normal dialogue.

[react] : Unprompted comment on screen content.

[heartbeat] : 5-minute autonomous check.

[diary] : 23:00 nightly journal.

The system shifts these dynamic rules on every single call, creating a contextually aware entity that reacts to both the user and the passage of time.

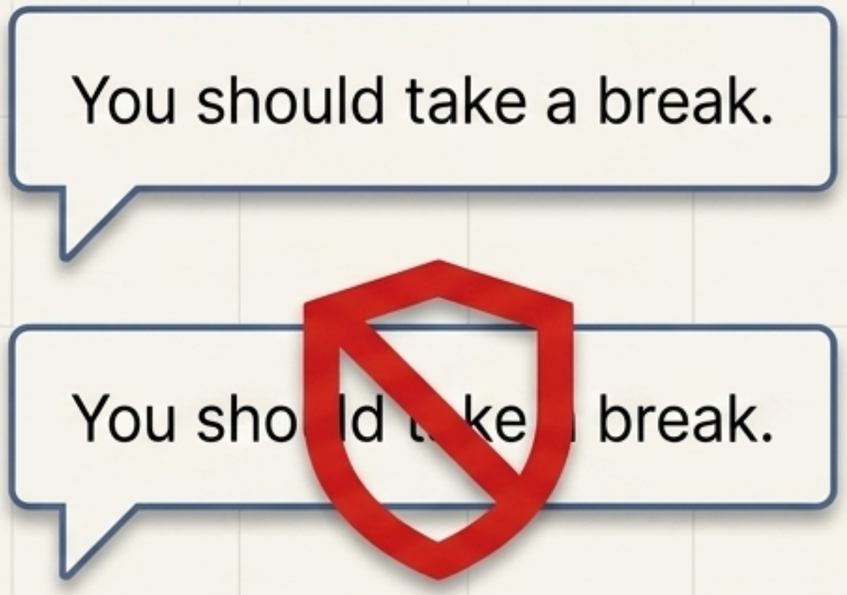
The illusion of life comes from knowing when to speak.

The heartbeat mode fires every 5 minutes. Usually, it chooses silence. But when Daily Context (idle time, late hours) collides with Drive Hints (a seeded topic from recent observation), it initiates conversation entirely on its own.

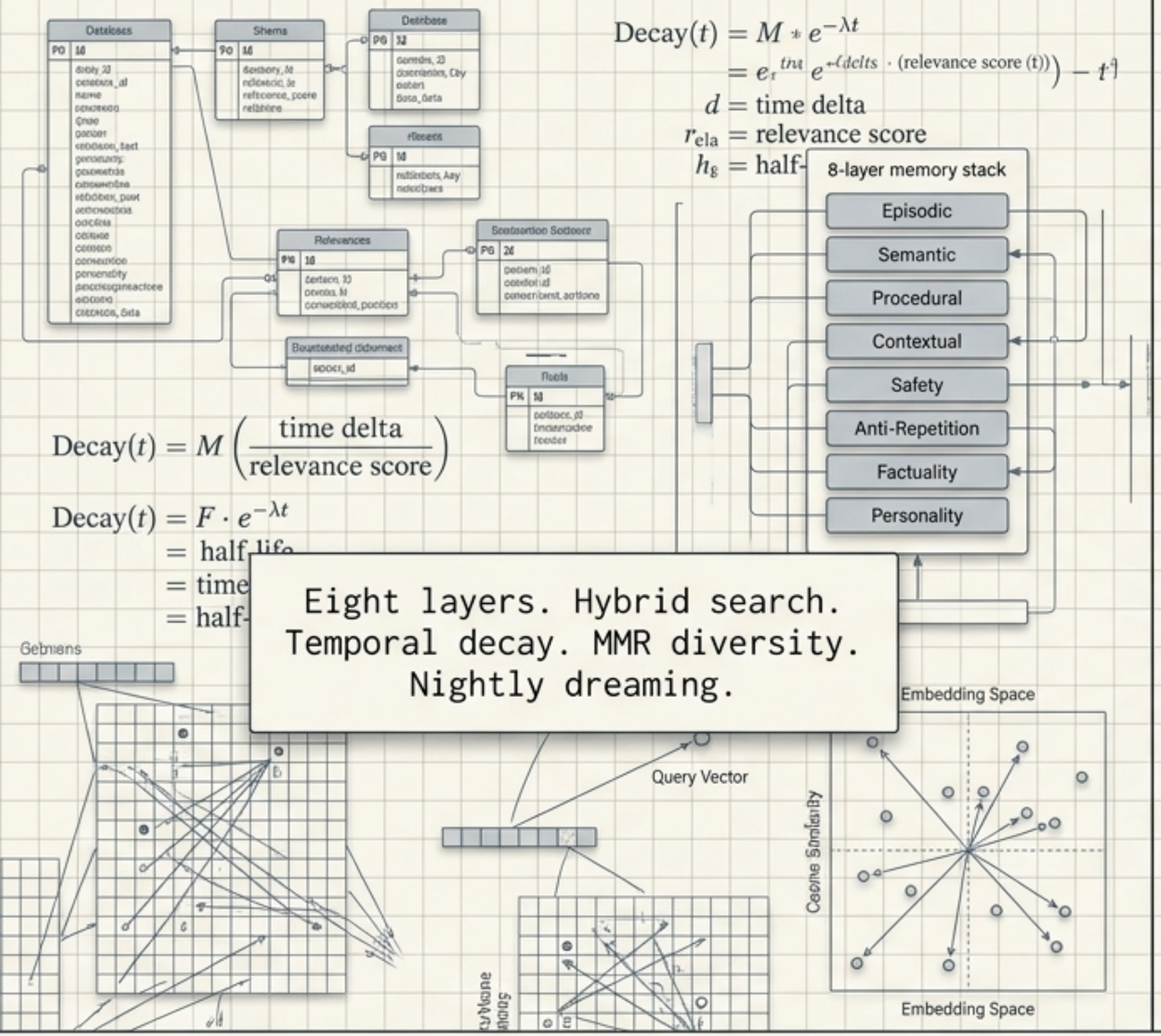
It doesn’t speak on a timer. It speaks when it has something to say.

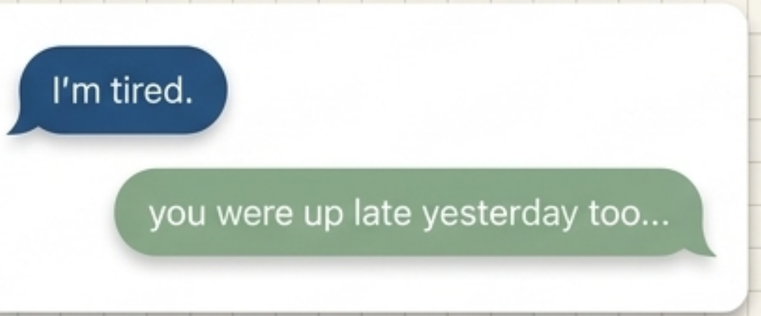
Layer 8: Anti-Repetition

Models naturally loop. By injecting the last N responses at the very end of the prompt with an explicit “do not repeat” instruction, the loop is broken. Constraints work best as the final layer the model reads before generating.




```
Total payload: 2,000–3,000
tokens per turn.
Over 50% is cached.
```





Memory systems don't solve a technical problem. They solve an emotional one.
How do you make something that isn't alive feel like it cares about you?
It remembers.