



A Stranger Without Memory

Why relationship-building requires shared history.

♥ Every AI assistant solves "how to answer better."
An AI pet solves "how to make you feel known."



Storage is Easy, Recall is Hard:

Auto-injecting context at conversation time is the true engineering challenge.



Sleep-Based Consolidation: Three memory tiers + nightly "dreaming" to mirror human cognition.



The 8-Layer Prompt Engine: Turning memory into conversation while strictly managing token caching and latency.

Three Tiers, Three Horizons



Why local SQLite?
Zero dependencies,
zero network calls.
You close the laptop,
the memory stays yours.

Storage: SQLite +
FTS5 + sqlite-vec

Capacity: Unlimited
(~50k-80k entries/month)

Input: Every 45s observation,
chat, or screen reaction.

Tier 1:
Episodic
(Raw Material)

Storage: JSONL +
Rolling Summary

Capacity: Dynamic
(70/30 split compression
at ~500k tokens)

Input: Real-time
message appending.

Tier 2:
Conversation
(Session Context)

Storage: JSON
(Stable Prompt Prefix)

Capacity: Max 100 entries
(~500 tokens)

Input: Nightly "dreaming"
extractions.

Tier 3:
Long-Term
(Curated Essence)

Nightly Dreaming & Active Recall

Query



Hybrid Search:
BM25 (Keyword) +
768-dim Vector (Semantic)



Temporal Decay:
 $\text{Score} *= \exp(-0.01 * \text{ageDays})$



MMR Diversity:
0.7 Relevance / 0.3 Variety



23:30 Daily:

The system reviews the day's episodic fragments.

Distillation: Compresses raw noise ('user said hungry at 14:32') into behavioral patterns ('user skips lunch').

Pre-Compaction Flush: Extracts facts before session summaries compress them, preventing slow information leaks.

The 8-Layer Prompt Engine

The Cache Boundary

1. **Identity:** Small creature on desktop (~50t)

2. **Soul:** Prose character bible (~800t)

3. **Long-Term Memory:** 100 stable facts (~500t)

~1350 Tokens /
80% Cached

4. **Active Memory:**
Auto-recalled episodic fragments.

5. **Daily Context:** Clock & idle time.

6. **Mode Rules:** Observe, Chat,
React, Heartbeat, Diary.

7. **Drive Hints:** Topic seeds from
observations.

8. **Anti-Repetition:** Forces novelty.

Variable /
Dynamic

Prefix caching makes a chatty pet economically viable.
The "Heartbeat" mode uses layers 5 & 7 to speak spontaneously, not on a timer.

Memory is an Emotional Bridge

Solving for connection, not just correctness.



"You were up late yesterday too."

This single sentence requires eight prompt layers, hybrid search, MMR diversity, and nightly sleep consolidation.

The user sees none of the engineering. They only experience the outcome: The pet remembers them.

Memory systems don't just solve the technical problem of data retrieval. They solve the emotional problem of making a digital entity feel alive and caring.