Anamnesis

The memory layer for AI

Persistent, provenance-aware memory that any model can share — and that knows when a fact has changed.

EXECUTIVE SUMMARY · ANAMNESIS.SMTRY.AI · HELLO@JIASHLEY.COM · 2026-08

The problem

Every AI session starts from zero. Teams pay a **search tax** on every task: without memory, an assistant greps and reads files for many turns before it can act — measured at ~5× more turns, ~4.6× more input tokens, ~5× slower. The naive fix, a vector store bolted onto a chat log, degrades as it grows: at 1,000 memories with realistic noise, answer fidelity collapsed from 0.94 to 0.42 — and when a decision changes, the stale record keeps being served as if it were current.

The product

Anamnesis captures working sessions automatically and consolidates them into long-term memory — distilled and organized, not a transcript dump. Three properties distinguish it:

- **Provenance and trust.** Every memory carries a provenance-derived trust tier; uncorroborated content is labeled and kept out of decision-weight.
- **Temporal belief revision.** Supersession is detected automatically; only the current belief is served as context, with the old one preserved for audit. A plain vector store structurally cannot do this.
- **Model-agnostic.** One pool serves every client; knowledge captured by one model is recalled and honored by another, verified end to end.

Measured results

From a pre-registered, reproducible, blind-graded benchmark suite (public at anamnesis.smtry.ai) — evidence, not claims.

- **Near-oracle quality, ~18× cheaper:** 94.2% fidelity vs 98.3% with the full source in context (431 vs 7,820 tokens). Blinded baseline: 7.5%.
- **Survives scale that breaks naive RAG:** trust-aware ranking restores 0.42 → 0.94 fidelity at 1,000 noisy memories.
- **Current decisions, not stale ones:** automatic supersession — every check passed.
- **One brain across models:** cross-model recall 1.00 vs 0.00 blinded.
- **Pays for itself:** ~5× fewer turns, ~2–8× cheaper per query in real sessions.

The moat

Data sovereignty, in tiers. Hosted with per-user encryption today; client-held keys on the roadmap; at the top, a fully on-premises tier where local models read shared memory through a **one-way mirror** — context flows out, nothing unvetted flows back, generated content never leaves the machine. For regulated and IP-sensitive organizations this tier is the product, and it is the hardest part to copy.

Memory compounds. Switching models is cheap; abandoning a year of institutional memory is not. A model-agnostic memory layer sits above the model churn and gains value with every session — the durable layer in a stack where everything else is being commoditized.

Status & contact

Live in production: automatic capture, consolidation, provenance trust tiers, belief revision. Daily multi-model use (Claude Code, Google Antigravity, Codex — one shared pool). Private beta underway; benchmark suite published and reproducible.

hello@jiashley.com — data room and pitch materials available to investors on request.