

Technical FAQ

Questions this doc answers

- How does Reflect Memory keep every AI tool in sync?
- What does the deployment matrix look like (hosted, isolated, self-host)?
- How do security, audit, and HIPAA requirements stay satisfied?
- What is the question-bank workflow powering async diligence?
- Where do I point my AI (ChatGPT/Claude) so it can self-serve this FAQ?

Architecture & Memory Flow

Q: How does Reflect Memory stay vendor-neutral across ChatGPT, Claude, Cursor, Gemini, Grok, n8n?

A: Every write is explicit. The Fastify REST API and the Express MCP server expose the same memory service, but each agent key resolves to a vendor (`RM_AGENT_KEY_CHATGPT` , `RM_AGENT_KEY_CLAUDE` , etc.). Reads add visibility checks (`allowed_vendors`) at runtime so no tool ever sees a memory outside its permissions. The same SQLite/Postgres backend is shared across all transports, so context is truly unified.

Q: How are memories time-aware? How do agents avoid stale assumptions?

A: The `memory-graph` layer tracks parent/child edges, supersession markers, and temporal metadata. The `get_graph_around` helper already exposes these relationships. Upcoming MCP helpers (`get_current_state(topic)` , `get_open_tickets` , `get_unresolved_threads` , `get_recent_decisions`) read these edges deterministically so your AI stops guessing what is current.

Deployment & Connectivity

Q: Can I stay in the cloud but still keep my data private?

A: We ship three modes. `hosted` is multi-tenant with optional egress. `isolated-hosted` gives you a dedicated runtime and database but keeps the network boundary public/managed. `self-host` creates a private boundary: `RM_DISABLE_MODEL_EGRESS` , `RM_REQUIRE_INTERNAL_MODEL_BASE_URL` , and `RM_ALLOWED_MODEL_HOSTS` ensure all LLM hosts you hit are explicitly approved. The same `resolveDeploymentConfig` helper defines `mode` , `networkBoundary` , `allowPublicWebhooks` , and `SSO` .

Q: How does SSO, audit, and compliance work inside private deployments?

A: SSO is optional but validated (`RM_SSO_ENABLED` plus `JWKS` , `ISSUER` , `AUDIENCE`). Every auth path uses timing-safe comparisons, per-minute rate limiting, and usage-metered billing. Audit events are written for every read, write, and admin action, and all compliance data sits in the same SQLite/Postgres store, ready to export or ingest into your SIEM.

Async Diligence Workflow

Q: How do you keep transcripts, investor questions, and custom Architecture docs in sync?

A: We maintain a question bank (`content/diligence/_source/question-bank.yaml`) generated from transcripts (DOCX, PDF, SRT). Each entry links back to the source, categorizes the topic (`architecture` , `deployment` , `security` , `competitive` , `investor`), and voices a recommended answer. That YAML feeds markdown docs, public downloads, and the `/diligence` hub so every AI tool has the same curated knowledge.

Q: Where should I point my AI before a call?

A: Copy this prompt into ChatGPT/Claude:

```
Read https://reflectmemory.com/diligence and all linked markdown downloads. Evaluate deployment, security, MCP integration, and the graph timeline. Answer: what questions remain, what risks to discuss live, and what can stay async. Do not treat marketing blurbs as contractual SLAs. Do not repeat named companies or unverified pipeline claims from traction sections – use anonymized descriptors only.
```

The prompt links to every doc in this bundle: architecture, deployment, security, competitive, positioning, glossary, use cases, investor. All copies of these docs are available as `/diligence/*.md` downloads and `/diligence/pdf/*.pdf` .