Memory Harvester Engine (MHE)

> **Cognitive–anamnetic system** for capturing multi-assistant dialogues (ChatGPT, Claude, Gemini), extracting artifacts (code, docs, decisions), and organizing them into a chronologically faithful, semantically searchable memory substrate powering your RAG/agent stack.

Hybrid + Contextual RAG Recall (v1.0)

The `/rag/query` endpoint has evolved into a **voice of recall**. It now leverages hybrid search and context stitching to produce ready-to-use prompts for LLMs.

---

1) Delegation to Hybrid Search

The RAG service no longer runs its own similarity search. It calls the `/search` endpoint internally, inheriting its fused lexical + semantic ranking for **high-fidelity recall**.

---

2) Context Stitching (Conversational Threads)

When a top result is a **message**, the system automatically fetches its immediate chronological neighbors (previous + next) in the same thread. This preserves dialogue flow and prevents orphaned snippets.

Example stitched block:

```
<CONTEXT SOURCE="thread:abc123">
[2025-09-22 12:05:01] user: We need a way to summarize weekly progress.
[2025-09-22 12:05:45] assistant: I will implement a consolidation job. It will
query for Memory Cards and synthesize a report. Here is the core function:
async def run_consolidation_job(...): ...
[2025-09-22 12:06:10] user: Looks good, proceed.
</CONTEXT>
```

---

3) Prompt Formatter

The endpoint assembles a **structured prompt string**, merging stitched message threads and Memory Card summaries into a clean, LLM-ready format with inline citations.

Example output:

```
{
  "prompt": "You are a helpful assistant. Use the following context...
\n\n<CONTEXT SOURCE='memory_card:c1d2e3'>\nSummary: Implements the Memory
Consolidator Agent...\n</CONTEXT>\n\n<CONTEXT SOURCE='thread:a1b2c3'>\n...
stitched dialogue ...\n</CONTEXT>",
  "citations": [
    { "type": "memory_card", "id": "c1d2e3" },
    { "type": "message", "id": "m9n8o7", "thread_id": "a1b2c3" }
  ]
}
```

---

4) Sprint: Cognitive Acuity Enhancements (v1.0.1)

Sharpening recall quality and ergonomics: - **Message embeddings**: Embed `message.content` for fused semantic + lexical scoring. - **Snippet highlighting**: Use PostgreSQL `ts_headline` to show why a lexical hit matched. - **Pagination**: Cursor-based pagination for `/search`, `/memory-cards`, `/artifacts`. - **Dream access**: New `GET /consolidations/{id}` endpoint for full synthesis reports. - **ANN index**: Maintain pgvector `ivfflat` index for scalable, low-latency semantic search.

---

5) Sprint: Data Hygiene & Governance (v1.1.0)

Ensuring purity of memory before full-scale ingestion: - **Redaction module**: Detect and scrub PII (emails, IPs) and secrets (API keys, tokens). - **Ingest hook**: Run every message through the redactor before persisting. - **Configurable**: Controlled by env var `MHE_SCRUBBING_ENABLED=true`.

---

6) Sprint: Observability Foundations (v1.2.0)

Enabling systemic self-awareness: - **Metrics**: Prometheus `/metrics` endpoint with counters and histograms: - `mhe_ingest_messages_total` (by source) - `mhe_artifacts_created_total` (by kind) - `mhe_memory_cards_minted_total` - `mhe_api_request_duration_seconds` (per endpoint) - **Tracing**: OpenTelemetry spans for critical paths: - `POST /ingest/export` - `POST /search` - `POST /rag/query/stitched`

---

7) Sprint: User Interface Foundations (v1.3.0)

Giving the engine a **face**: - **Chrono View**: Timeline of all threads in reverse chronological order. - Clicking a thread shows full dialogue (user/assistant) with inline syntax highlighting for code artifacts. - **Memory Card**

Integration: Messages with associated Memory Cards are visually marked. - Clicking a card shows its summary, tags, and provenance. - **Read-only**: Initial UI is for exploration and validation of ingest quality.

Outcome

- **Sharper recall**: Messages + cards scored with both lexical precision and semantic relevance.
- **Transparent results**: Highlighted snippets and provenance chips.
- **Scalable performance**: ANN index enables millisecond retrieval at scale.
- **Accessible insights**: Consolidation reports treated as first-class retrievable assets.
- **Memory integrity**: Automated scrubbing ensures safe, trustworthy knowledge ingestion.
- **Self-awareness**: Metrics and traces provide visibility into health and performance.
- **Mirror to the mind**: A read-only UI for human exploration of the memory substrate.

Result: The MHE not only recalls and speaks—it now *thinks with clarity, preserves its purity, observes itself, and shows its face*. This evolution cements it as a **state-of-the-art, production-ready RAG-as-a-service core** for your agentic stack.

Changelog

- **v1.0.1 – Cognitive Acuity**: Message embeddings, snippet highlighting, pagination, dream access, ANN index.
- **v1.1.0 – Data Hygiene**: Redaction module, ingest hook, configurable scrubbing.
- **v1.2.0 – Observability**: Prometheus metrics, OpenTelemetry tracing.
- **v1.3.0 – UI Foundations**: Chrono View, thread explorer, Memory Card integration.

Milestone: Initial architectural blueprint fully realized.