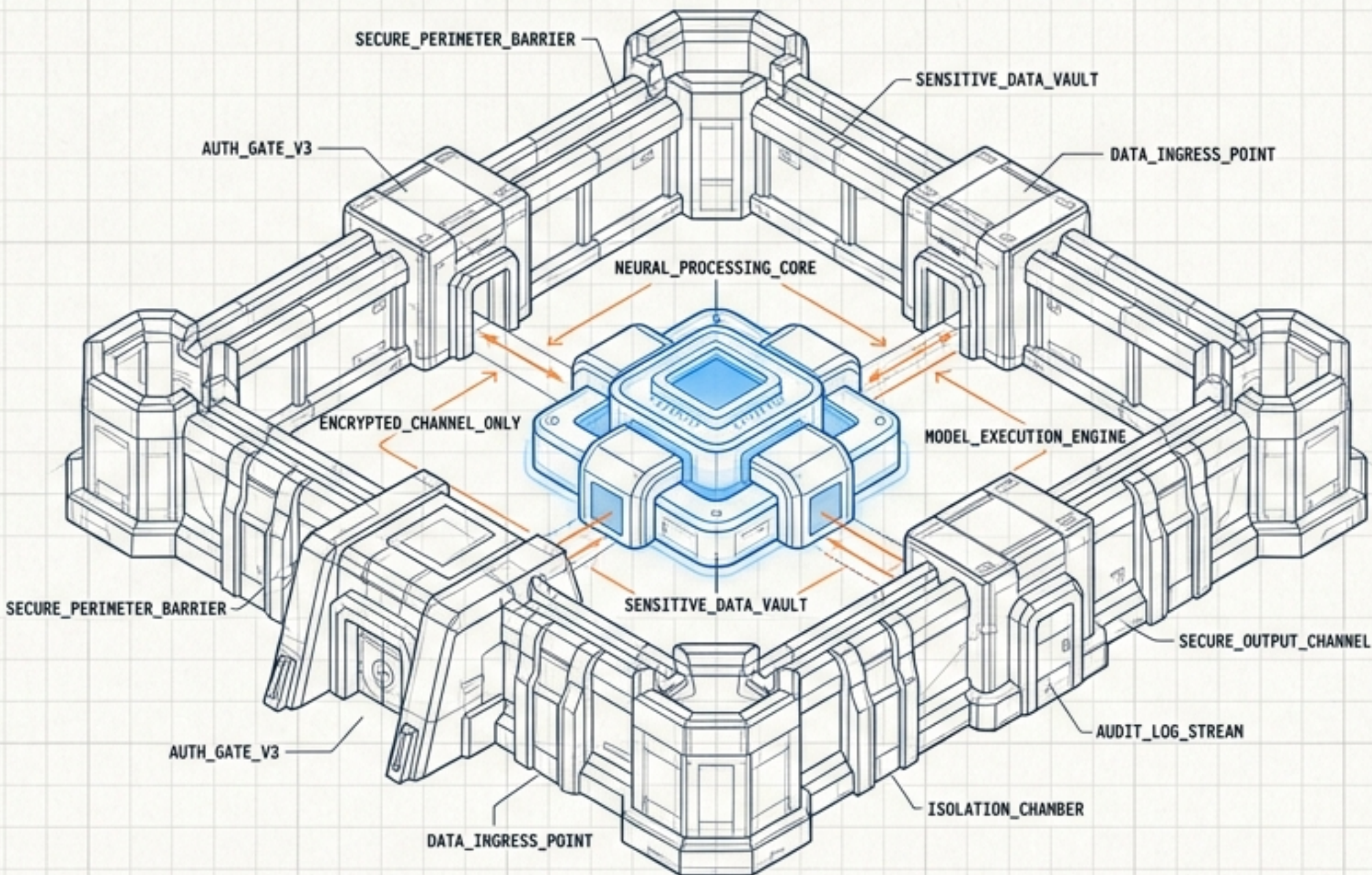




bulletproof-monet

The Anatomy of a Sovereign AI Architecture

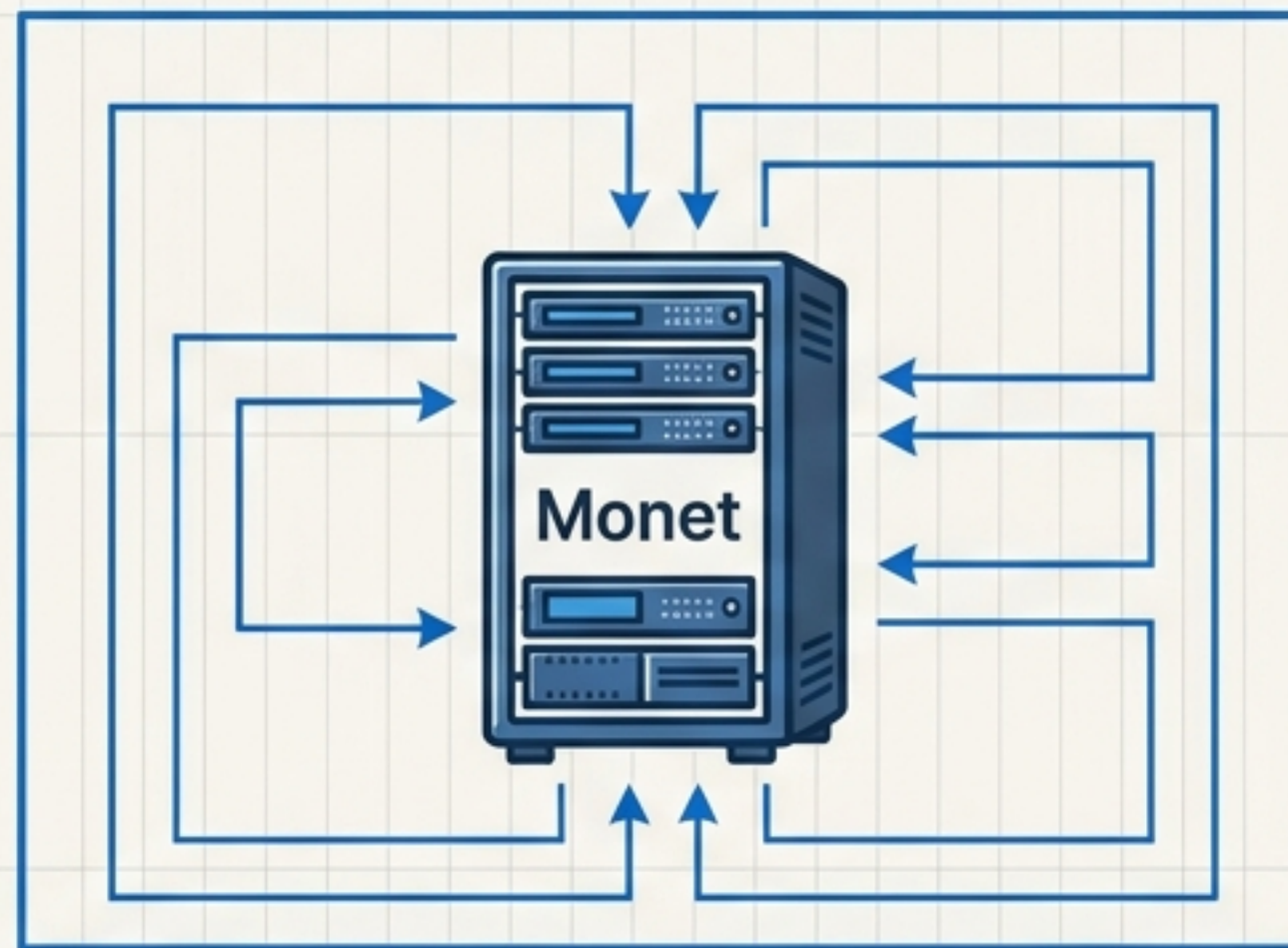
MECHANICAL METADATA

SYS_DEPLOY: SINGLE_HOST | **ENGINE:** CLAUDE_CLI | **AUTH:** DENY_BY_DEFAULT

A personal intelligence residing on a single hardened VPS.

Monet is not a wrapper or a library; it is an operational source tree deployed to `/opt/monet`. It drives the Claude Code CLI using your persistent memory, orchestrated by local tools, and reachable securely via Telegram.

Amnesia, Vendor Lock-in, Privacy Leaks

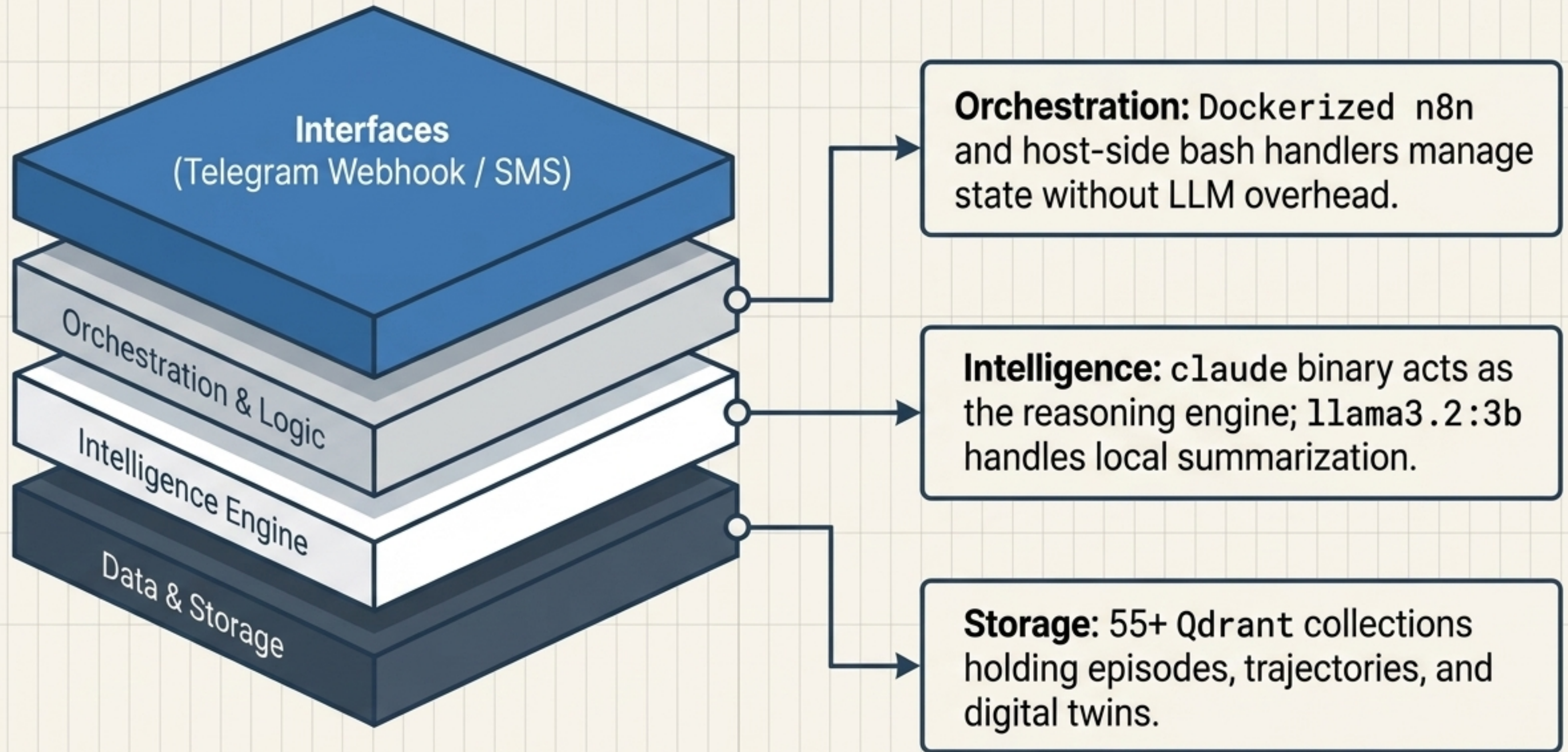


SOVEREIGN: Zero external APIs beyond your existing Claude subscription.

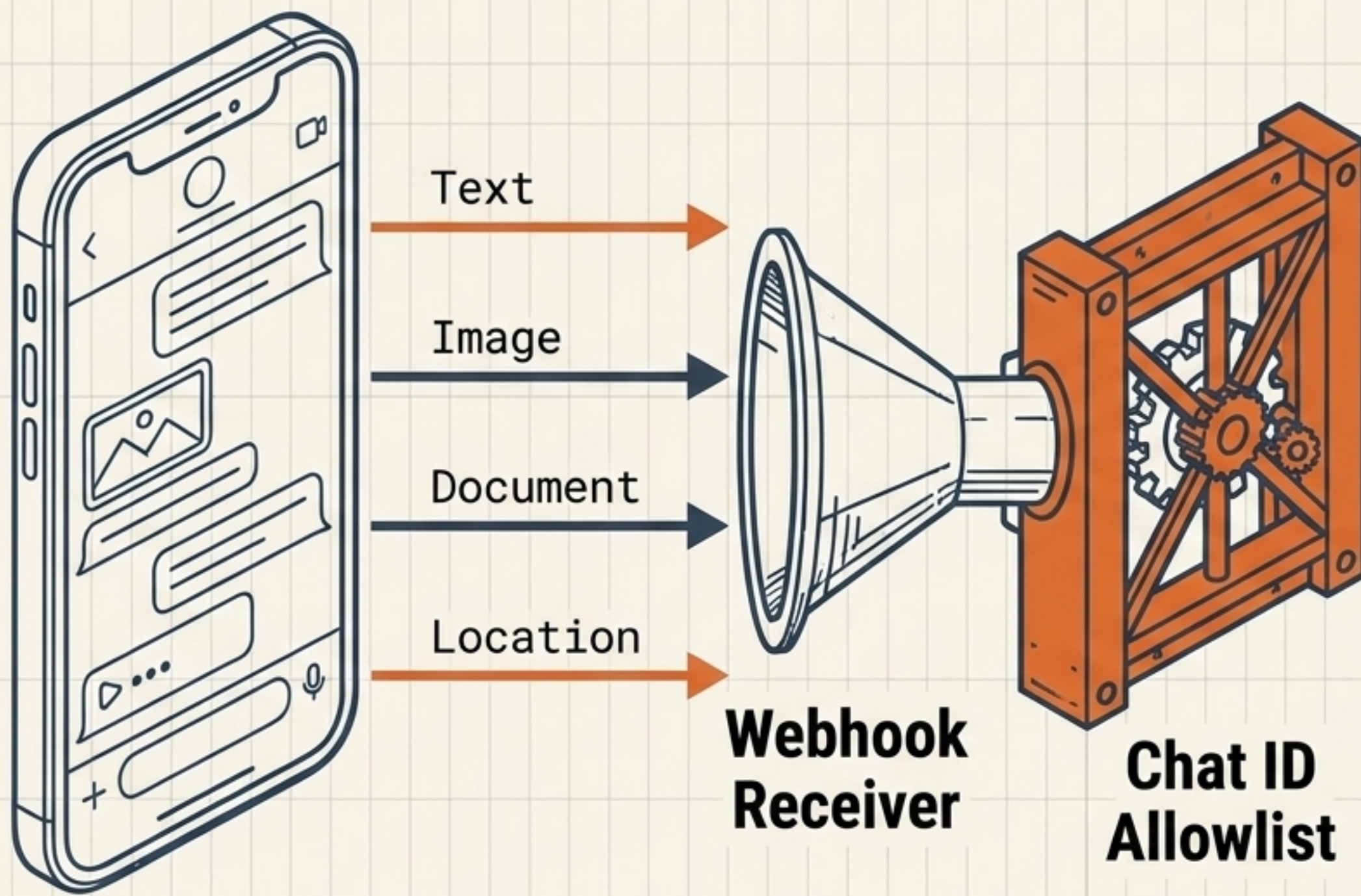
PERSISTENT: Deep vector-memory backend (Qdrant + Ollama) that remembers across sessions.

SECURE: Single-user system locked to one Telegram Chat ID with an append-only audit trail.

The architecture isolates orchestration, intelligence, and storage.



Multi-modal sensing locked to a single authorized user.



- **Universal Ingestion:** Accepts text, images (auto-tagged by type: receipt, whiteboard, etc.), documents, and locations.
- **Voice Support:** Voice notes are transcribed locally (or answered from captions).
- **Absolute Access Control:** The webhook receiver compares every inbound update against a single `TG_AUTHORIZED_USER_ID`. All other traffic is silently dropped.

A strictly bounded command surface area.

Conversation

/new (fresh session)

/export (session summary)

Memory & Search

/search (unified local data)

/memory-stats (Qdrant point counts)

/memory-export

DENY_BY_DEFAULT

Clips & Bookmarks

/clip, /clips, /find

/mark, /recall (semantic search
across saved snippets)

Automation & Utilities

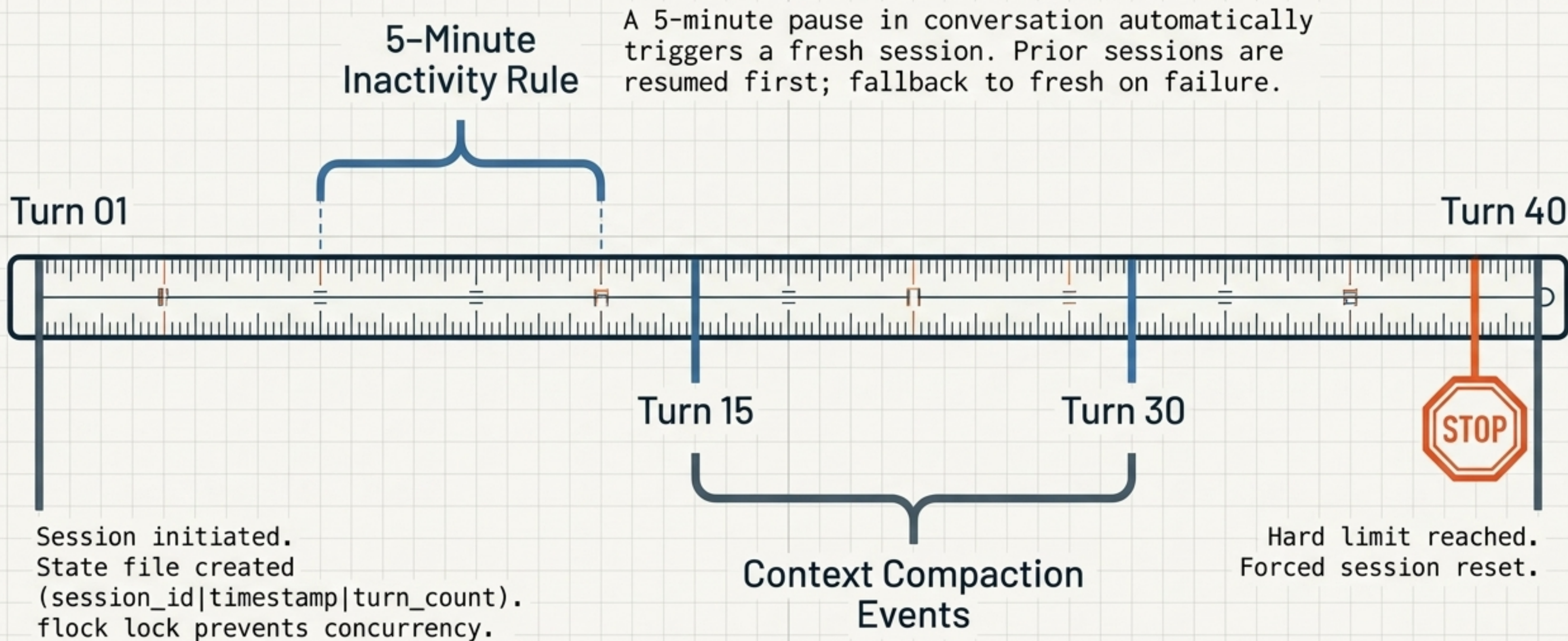
/remind, /reminders

/ss (screenshot)

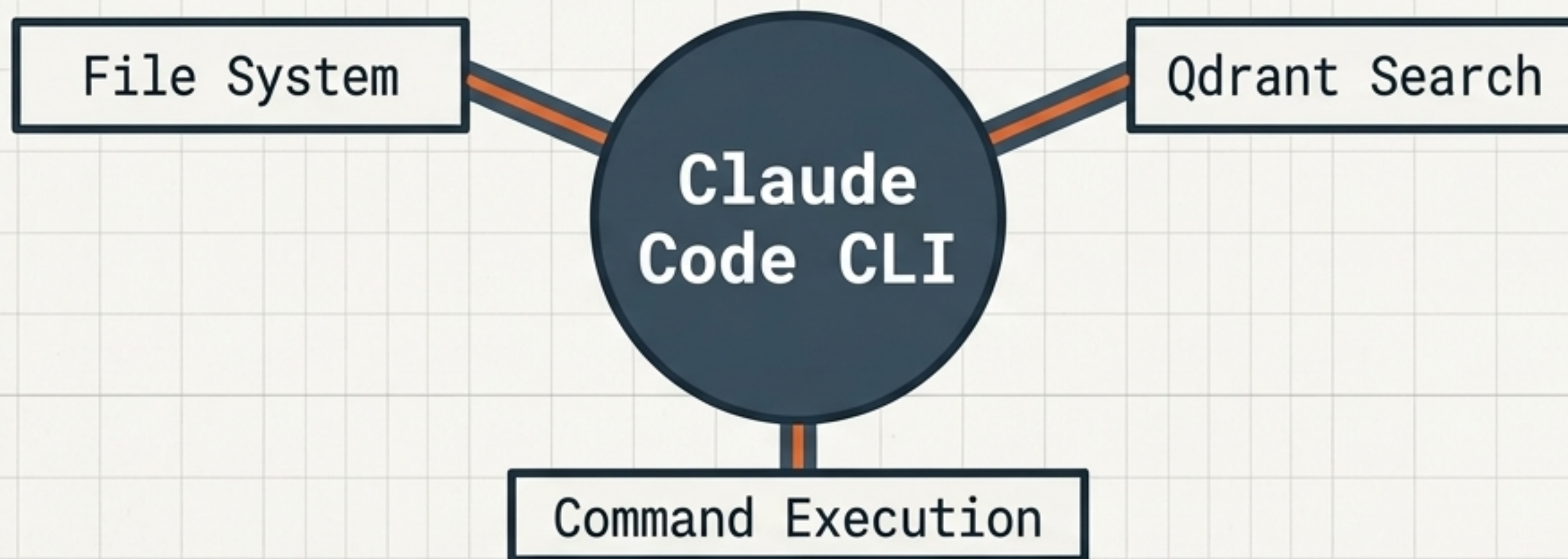
/photo-log, /usage

Anything not explicitly on this 17-command list is treated as ordinary conversation and routed to Claude.

Mechanical session lifecycles ensure stability and prevent context collapse.



The execution engine operates with unrestricted local authority.



Monet delegates reasoning to the Claude Code CLI, running as a subprocess with **--dangerously-skip-permissions**. It executes commands and reads/writes files on the host without human prompting.

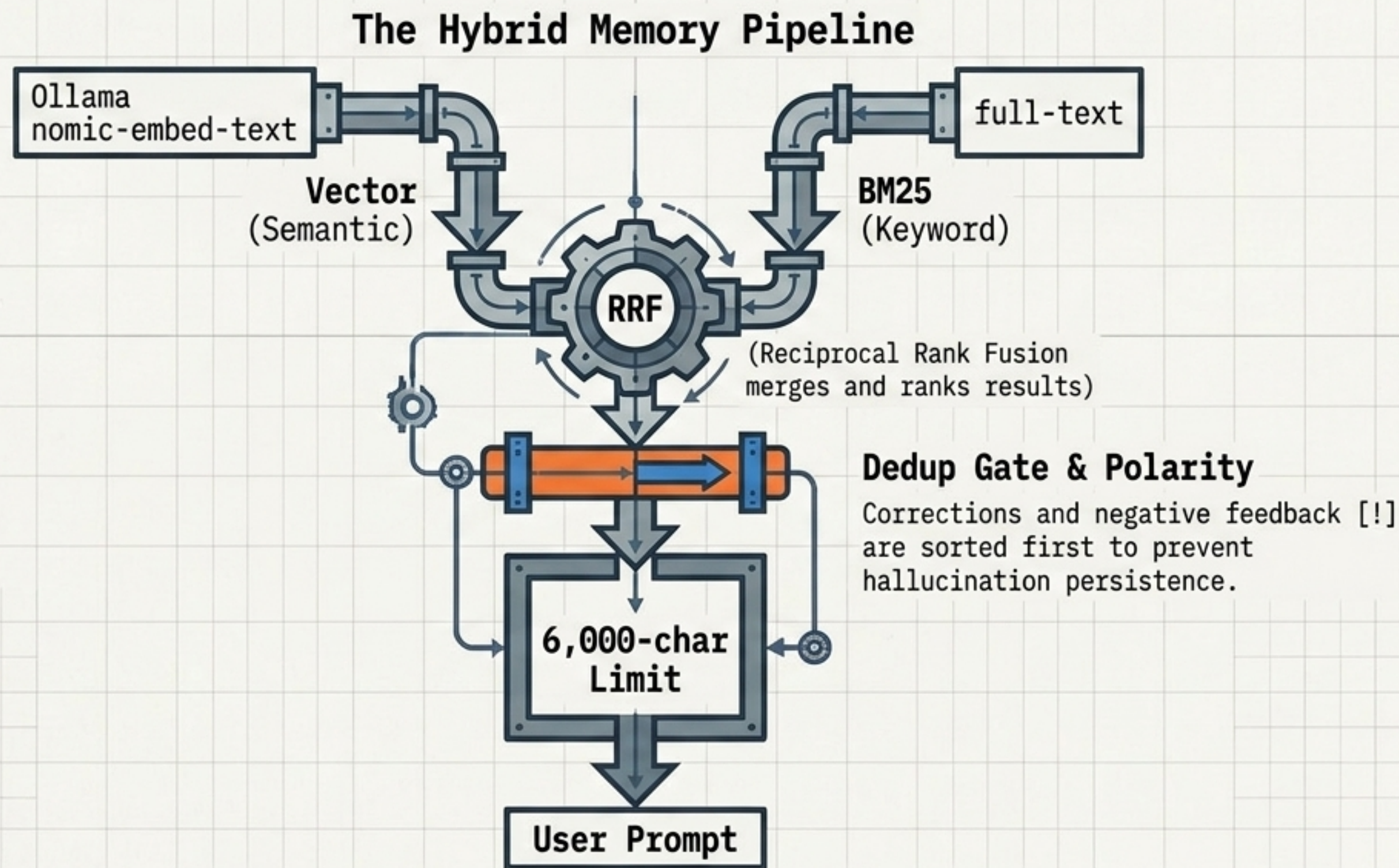
Extensibility: Powered by a custom TypeScript MCP server (**claude-memory-mcp**) exposing over 60 specific memory tools (e.g., **memory_boost**, **context_budget**) directly to the LLM.

Requires a dedicated VPS. The single-chat-id allow-list is the only thing standing between a stranger and a shell.

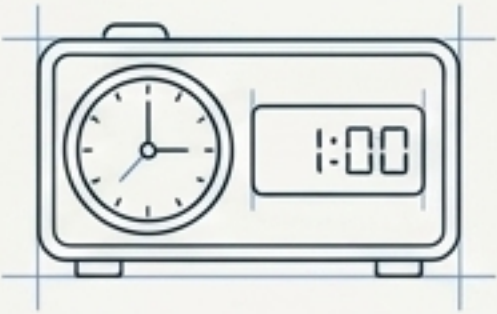
A two-layer memory model separates immediate constraints from deep history.

	Layer 1 (Hot Context)	Layer 2 (Deep Reference)
Technology	Local File (MEMORY.md)	Vector DB (Qdrant) + Ollama
Scope	Server identity, connectivity rules, persistent system facts	55+ collections (episodes, trajectories, learnings, digital twin)
Injection Method	Auto-loaded every session by Claude	Retrieved per-prompt via Hybrid Search
Volatility	Highly curated, rarely changes	Continuously appended via memory hooks and skill-harvesting

The recall pipeline fuses keyword and semantic search under strict token budgets.



Automation distinguishes between programmatic echoes and generative reasoning.



Programmatic Echoes

Reminders

- Time-based triggers.
- Evaluated every minute via `monet-remind.sh`.
- Echoes a fixed, pre-written user message (e.g., “Take out the trash”).
- Zero LLM cost.



Generative Reasoning

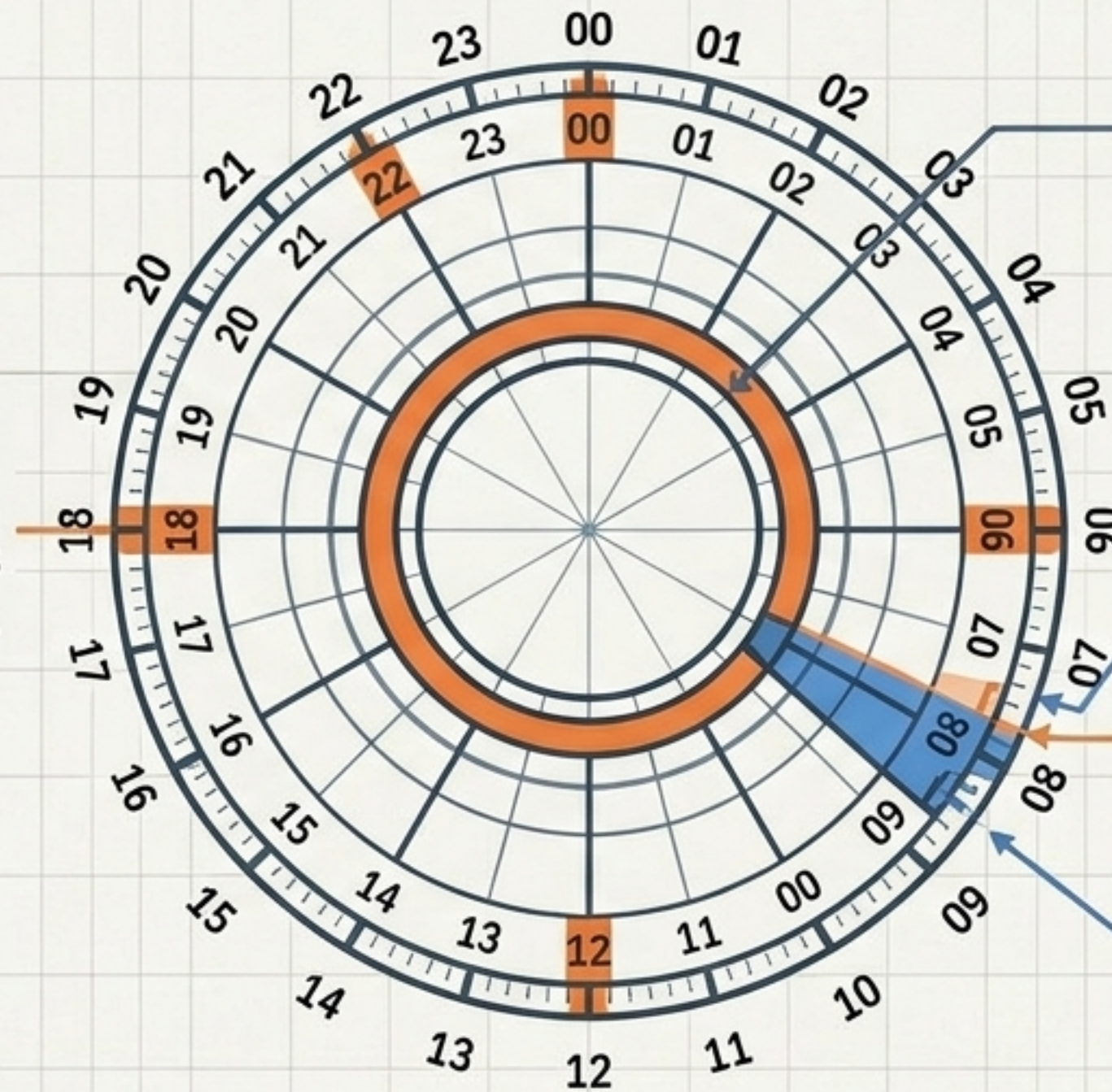
Scheduled Jobs

- Recursive AI tasks.
- Driven by `monet-cron.sh`.
- Executes a fresh prompt through the Claude engine (e.g., “Generate a daily threat brief”).
- Delivers newly synthesized, generative results.

The background heartbeat continuously refines system intelligence.

The 24-Hour Heartbeat

Every 6h (e.g., 17:00 UTC):
Learning & Skill-Harvesting
loop (extracts reusable patterns
from Qdrant "pheromone trails").



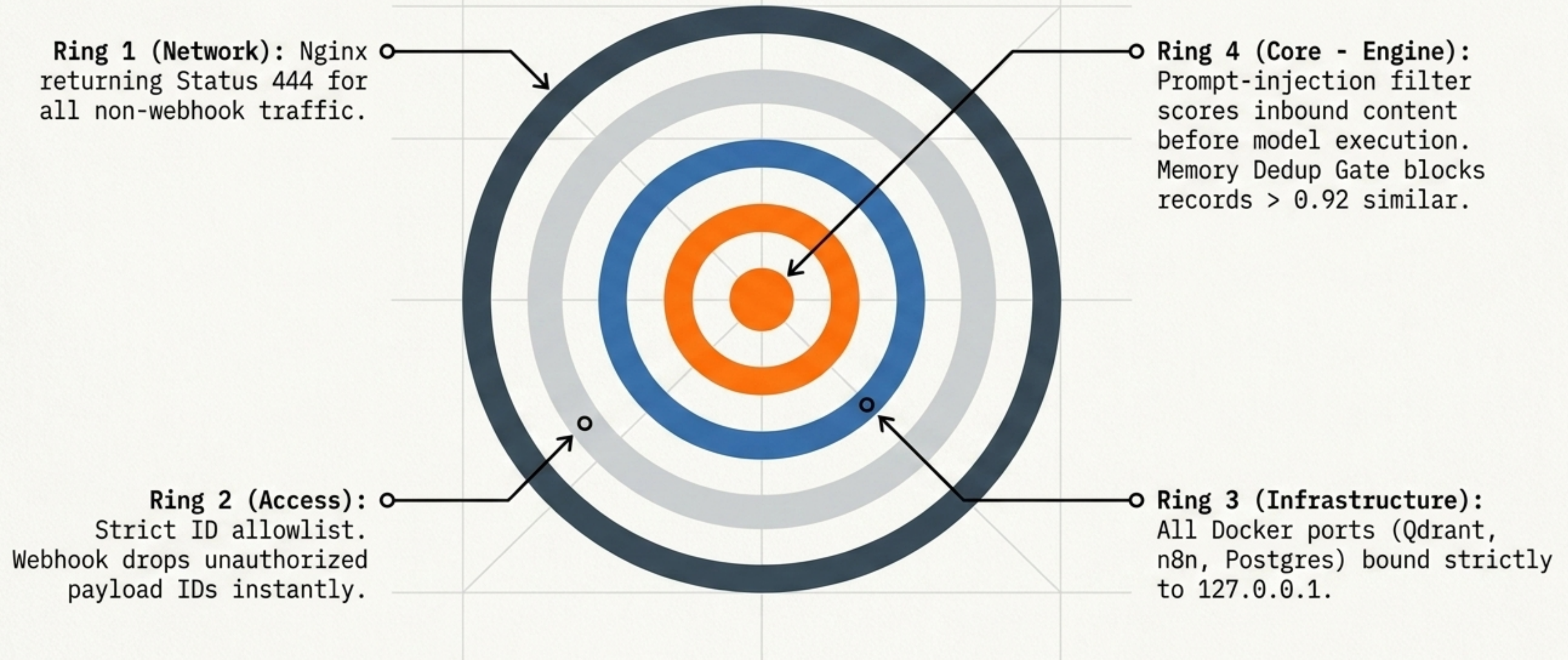
Every Minute: Watchdog
(kills stuck sessions > 6 mins)
& Agent-job execution.

Every 6h (e.g., 17:00 UTC):
Learning & Skill-Harvesting
loop (extracts reusable patterns
from Qdrant "pheromone trails").

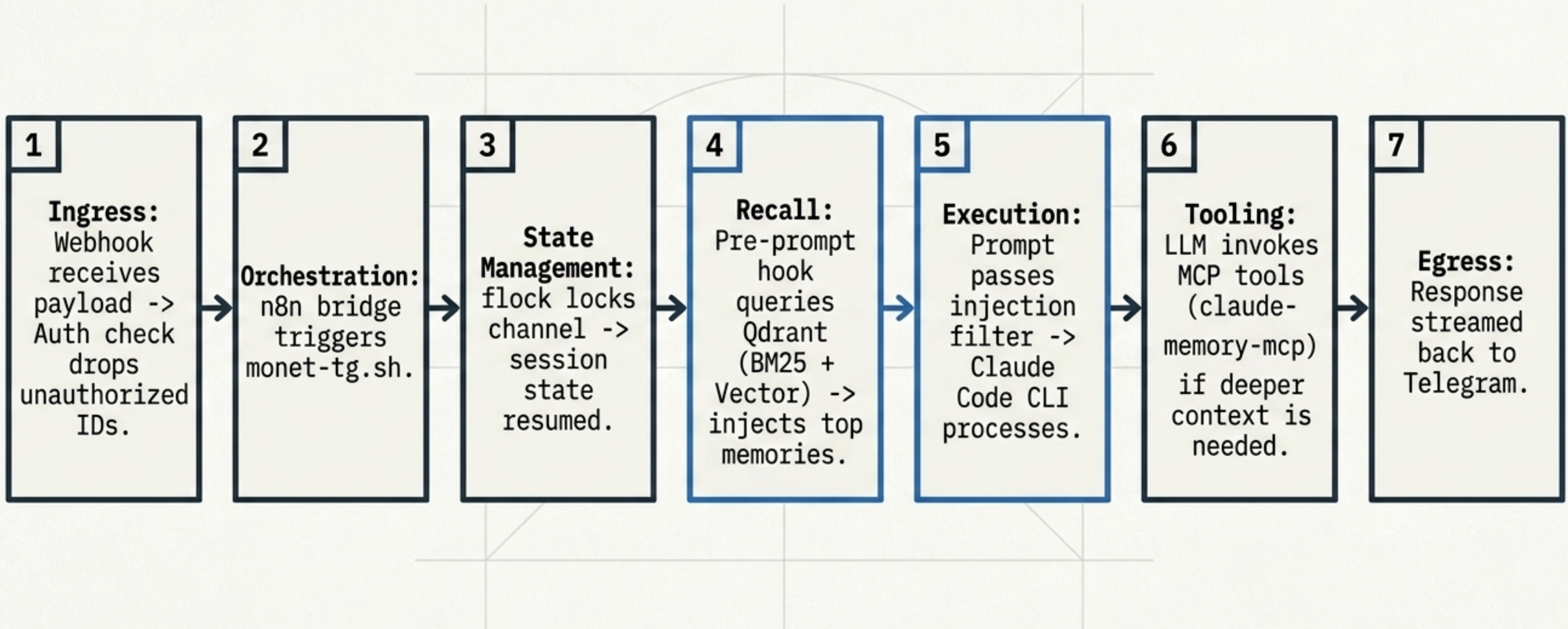
07:30 UTC: Daily Briefing
([monet-pulse.sh](#)) delivered
to Telegram.

08:30 UTC: Nightly skill-usage
aggregation.

Paranoid, defense-in-depth isolation for a highly privileged engine.

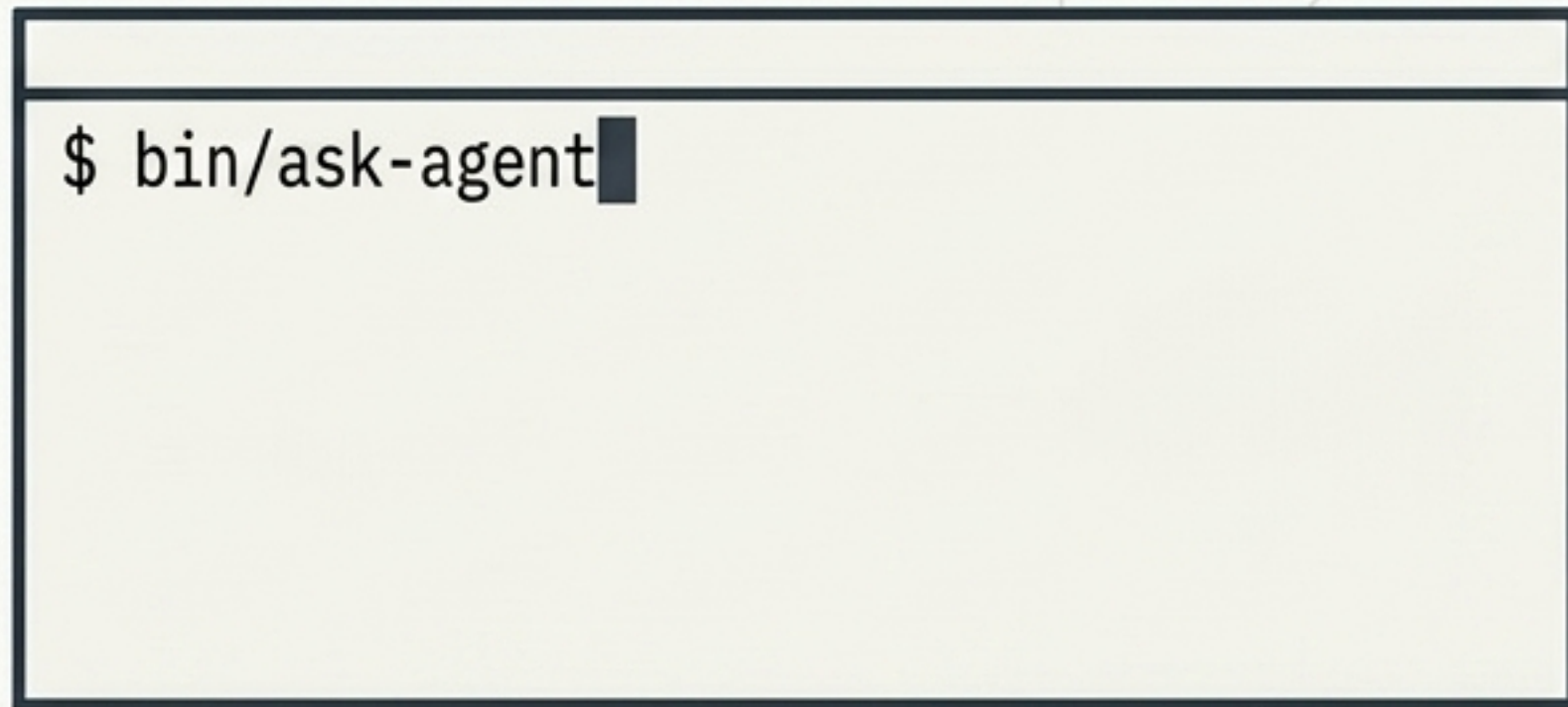


The Anatomy of a Single Turn.



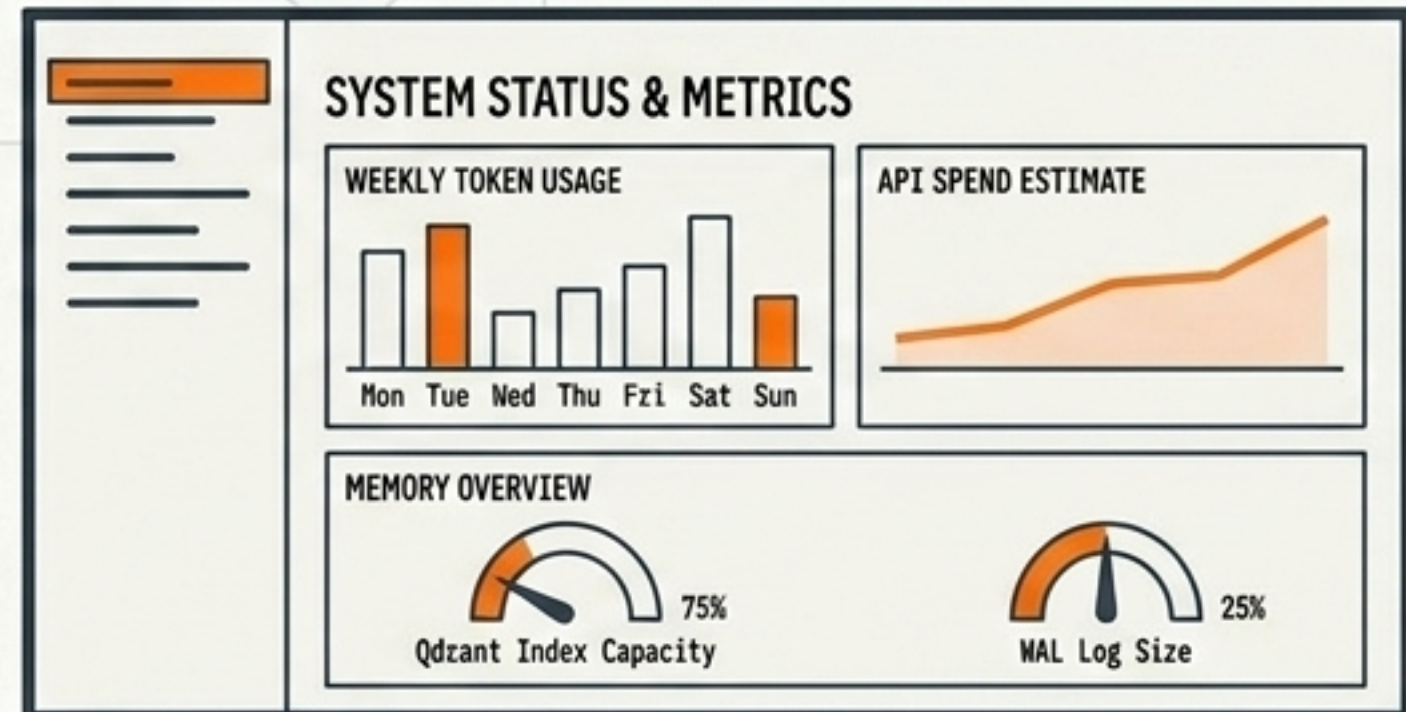
Host-side administration and optional agent delegation.

Optional Agent Personas:



bin/ask-agent CLI tool dispatches queries to markdown persona definitions in **\$MONET_AGENTS_DIR**. Operates entirely outside the Telegram chat state.

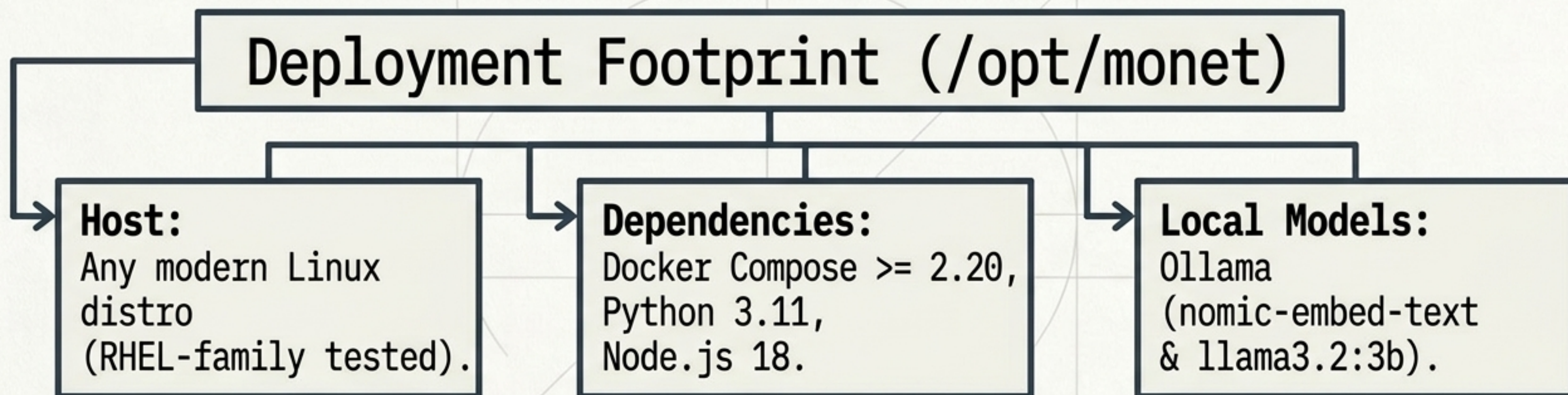
Management API (Port 8091):



Read-only, bearer-token authenticated endpoints bound to localhost.

- **/audit**: Queries the append-only SQLite log (WAL mode) tracking every turn.
- **/usage**: 7-day token consumption and API spend thresholds.
- **/memory/stats**: Qdrant collection capacities.

A complete sovereign intelligence deployed via one unified stack.



ZERO_THIRD_PARTY_DATA_BROKERS TOTAL_DATA_SOVEREIGNTY
0 Critical / 0 High vulnerabilities in automated scans.