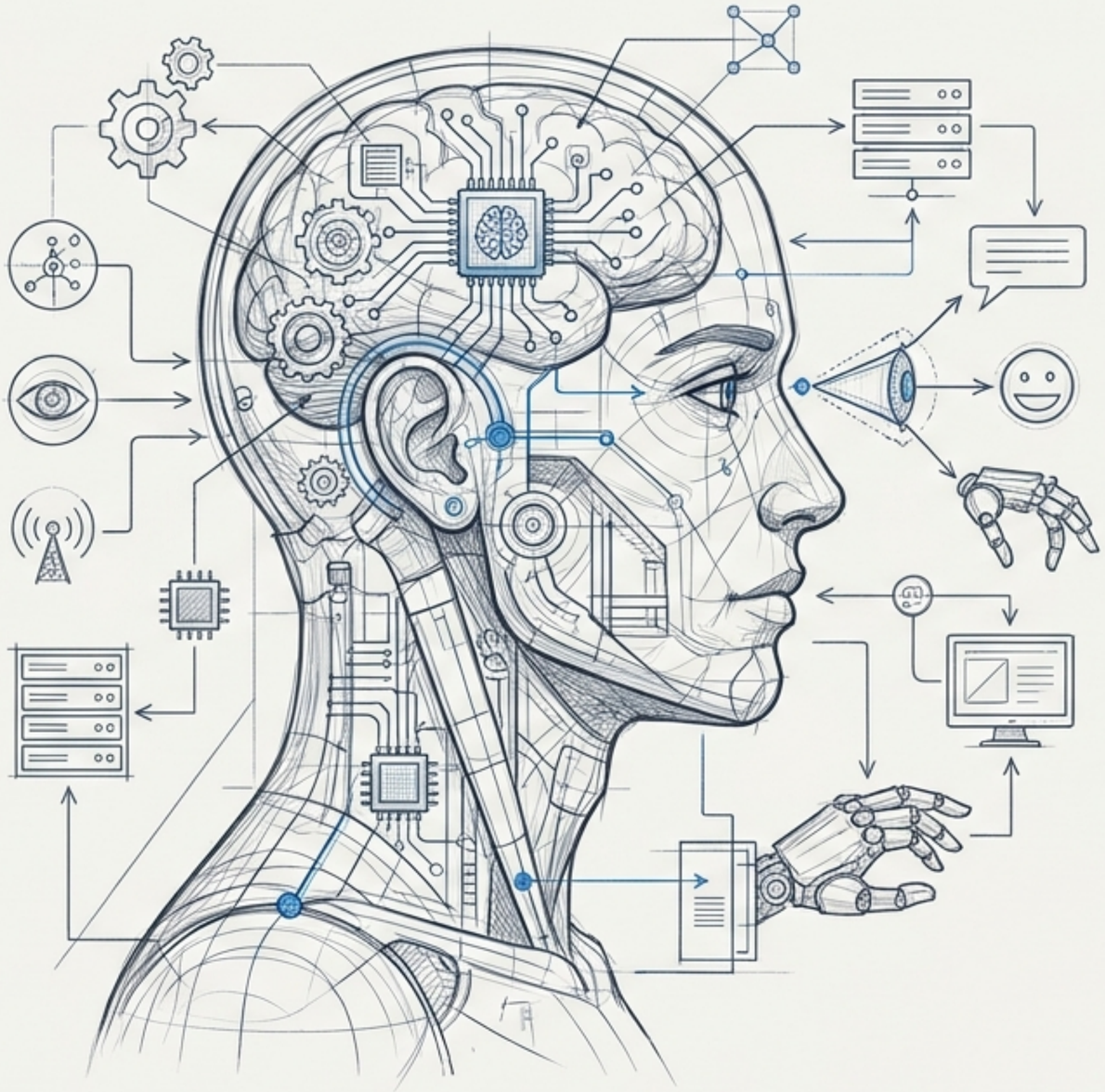


Anatomy of a Lifelike Agent

81 commits to bridge the gap between "It Runs" and "It Feels Alive."

VERSION	v0.0.2
MODULES	Perception, Cognition, Expression
STATUS	Multimodal Integration Active

Diagnosing the ‘Fakeness’ of v0.0.1

User Action	v0.0.1 Response (Fake AI)	v0.0.2 Resolution (Mio)
Send a photo	I have no idea what’s in it	Sees via Gemini 2.0 Flash vision
Send a voice message	Completely ignored	Hears via dual-path transcription
What are you up to?	Generic, timeless answer	Knows timezone and daily routine
Send me a selfie	I don’t have a physical form	Sends async contextual selfies
Reference an old chat	Retrieves random/wrong data	Precision LLM reranking

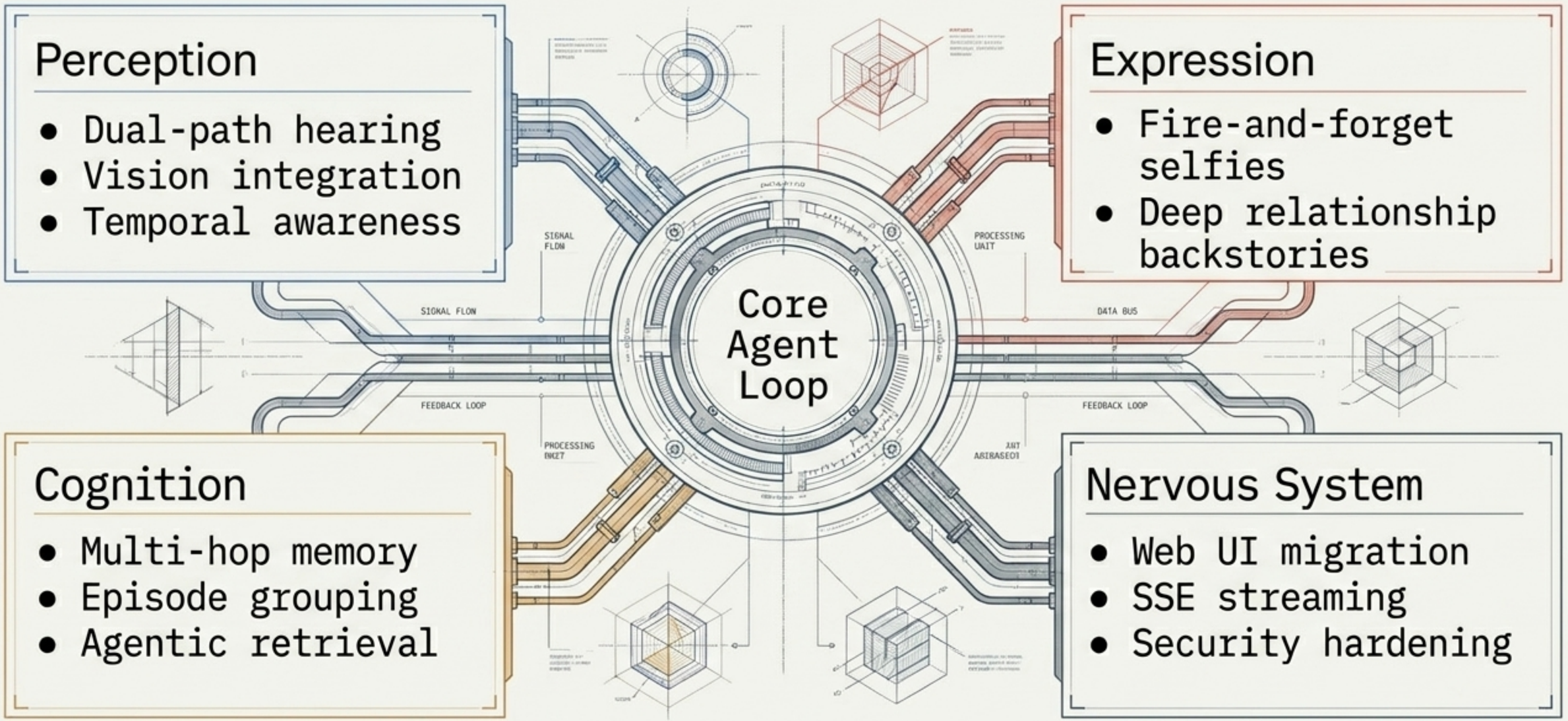
Constructing the Architecture of Life

Perception

- Dual-path hearing
- Vision integration
- Temporal awareness

Expression

- Fire-and-forget selfies
- Deep relationship backstories



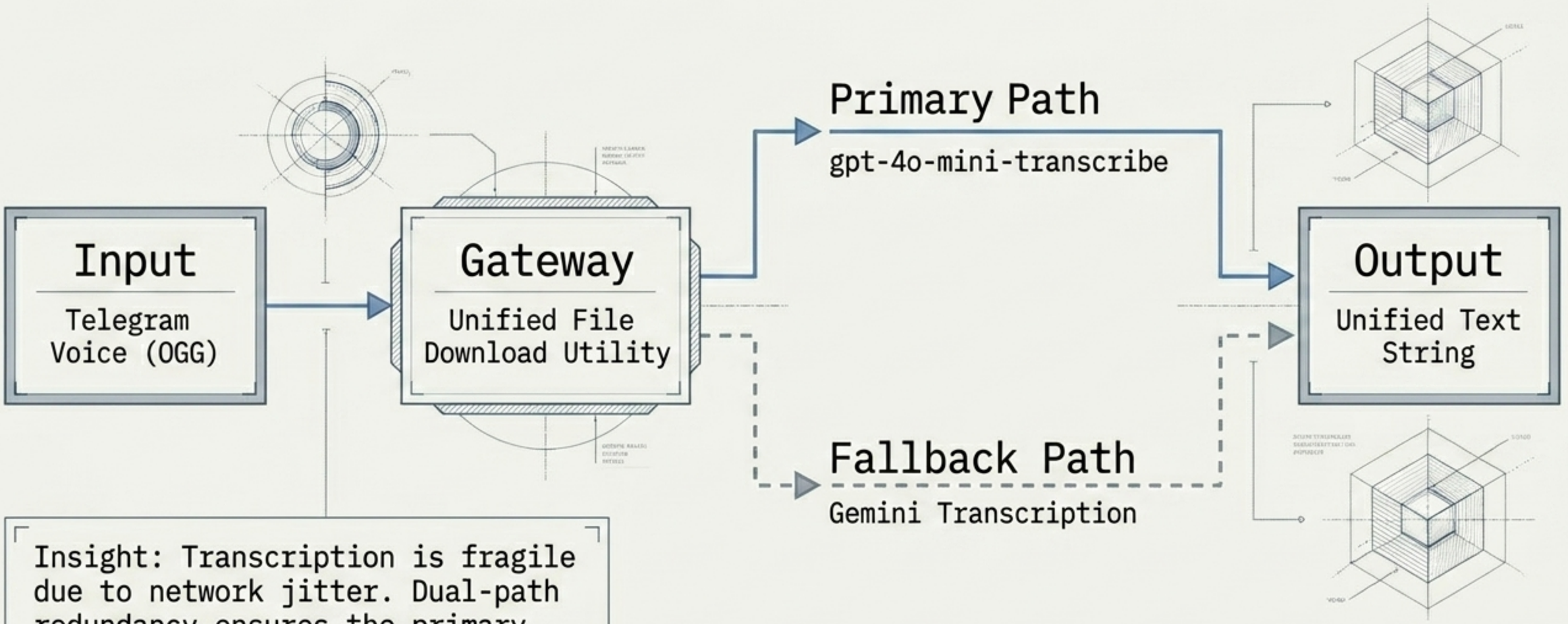
Cognition

- Multi-hop memory
- Episode grouping
- Agentic retrieval

Nervous System

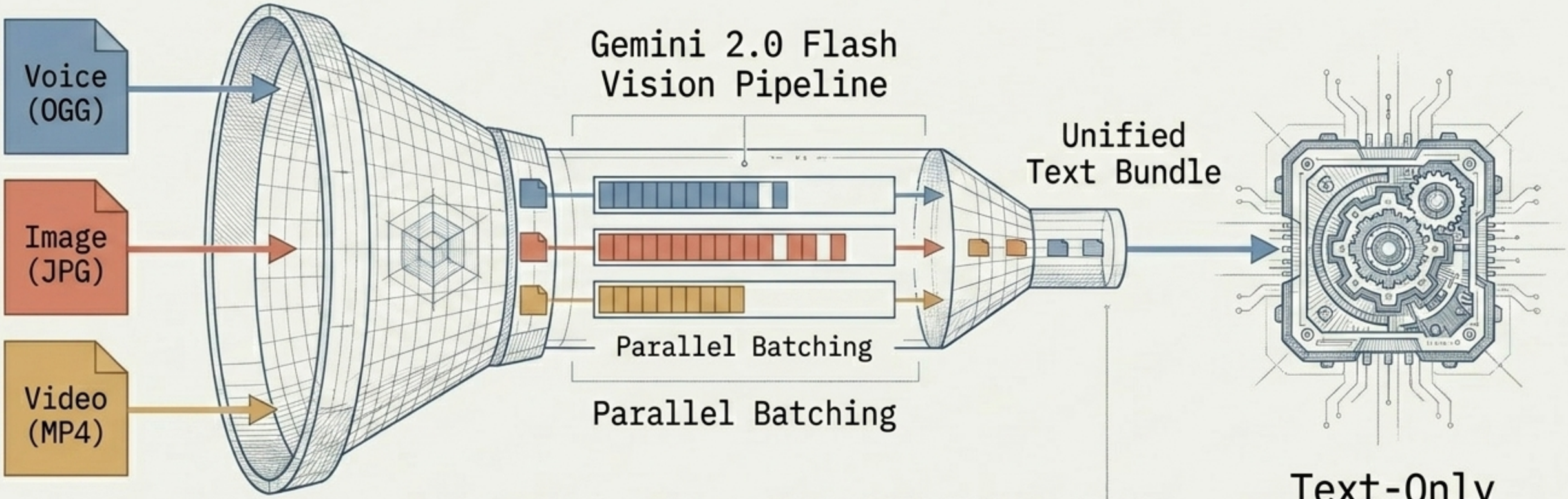
- Web UI migration
- SSE streaming
- Security hardening

Eradicating Silence with Dual-Path Hearing



Insight: Transcription is fragile due to network jitter. Dual-path redundancy ensures the primary failure goes completely unnoticed by the user.

The Pre-Processing Funnel



Key Decision: Handling multimodal I/O inside the agent loop creates exponential complexity. Front-loading conversion and batching it in parallel keeps the core agent architecture pristine.

The Temporal Awareness Stack

Layer 3: Flexible Daily Routines

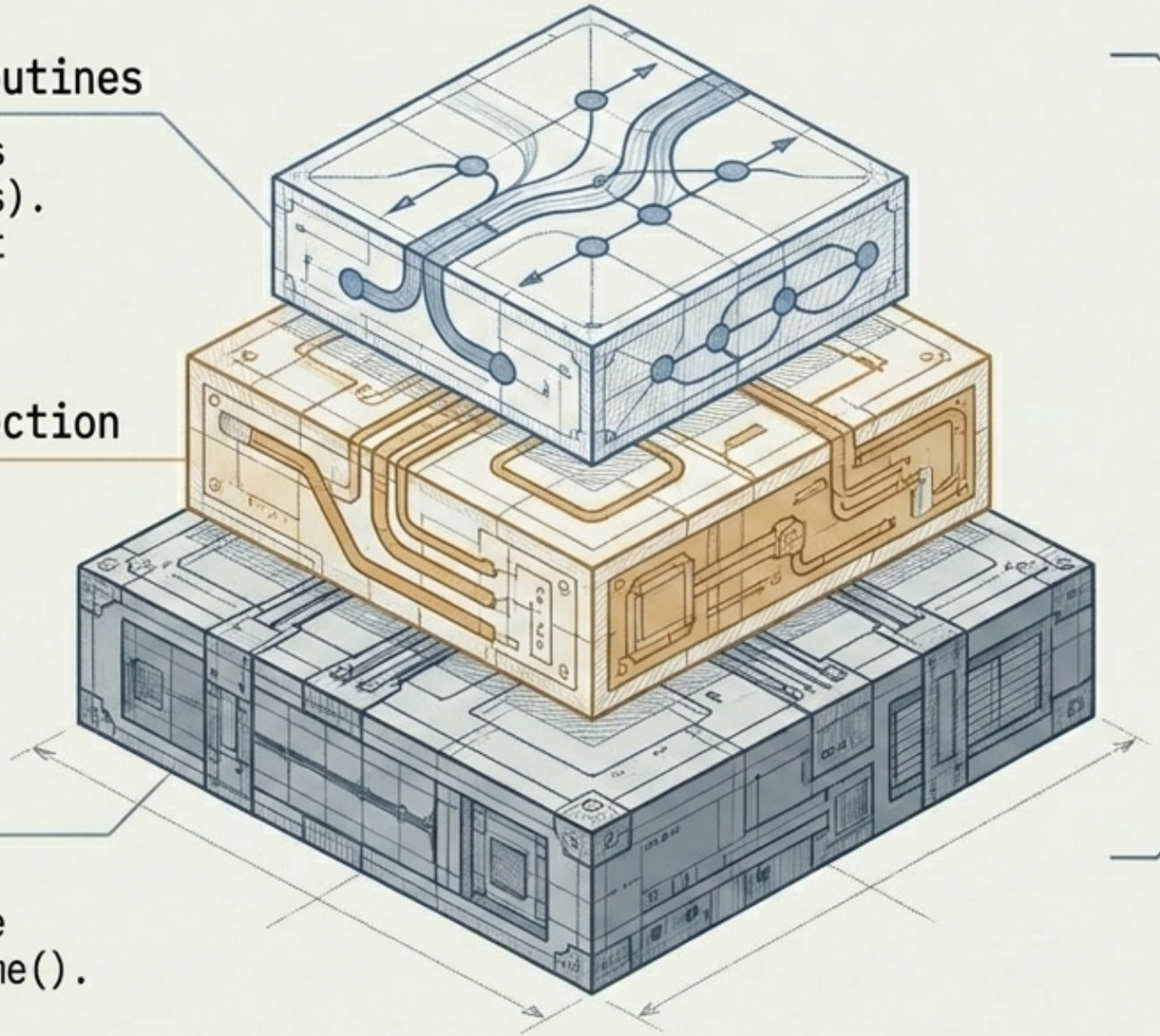
Distinct behavioral schedules per persona (e.g., night owls). Acts as organic guidance, not rigid timetables.

Layer 2: Elapsed Time Injection

Tracks time since last interaction. Contextualizes greetings based on silence duration.

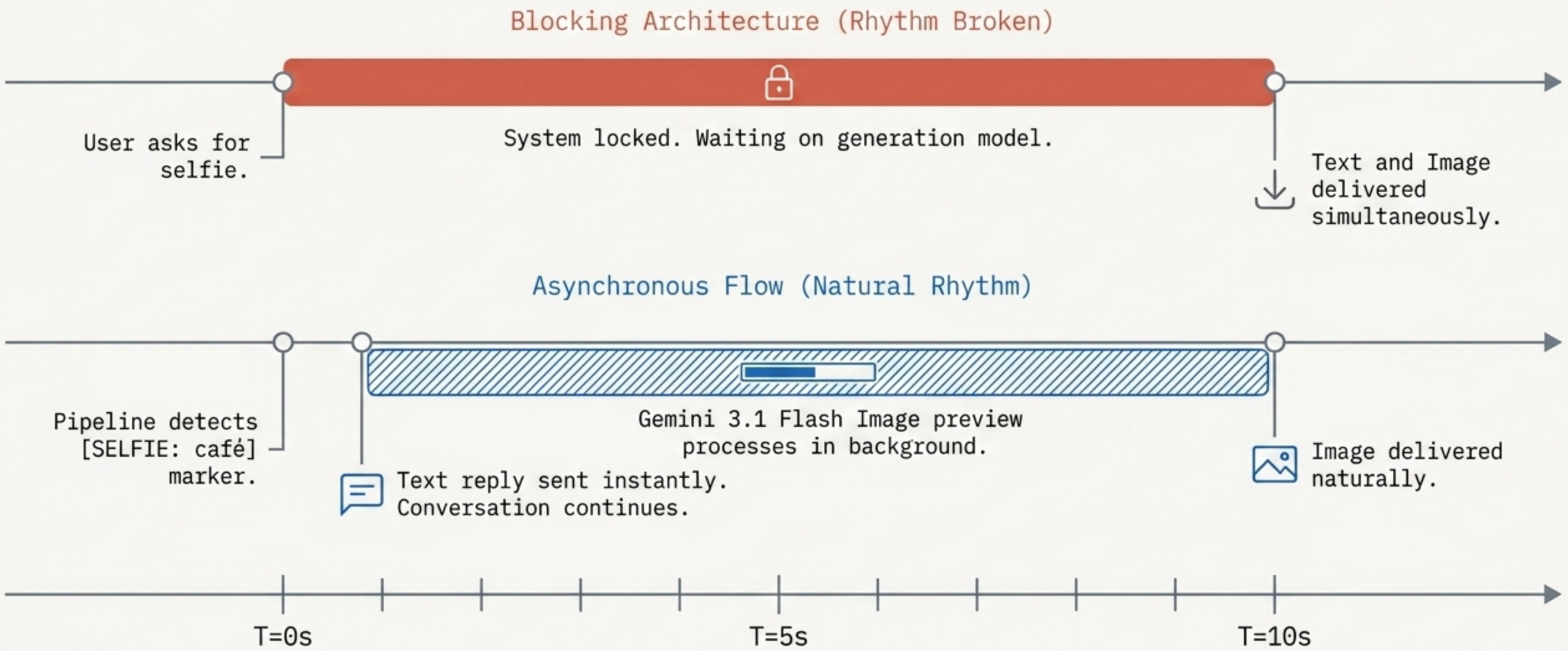
Layer 1: Timezone Math

Stores user timezone during onboarding. Inject local time directly via formatCurrentTime().



Turns "what are you doing?" into a natural, temporally grounded daily greeting.

Maintaining Rhythm with Fire-and-Forget Selfies



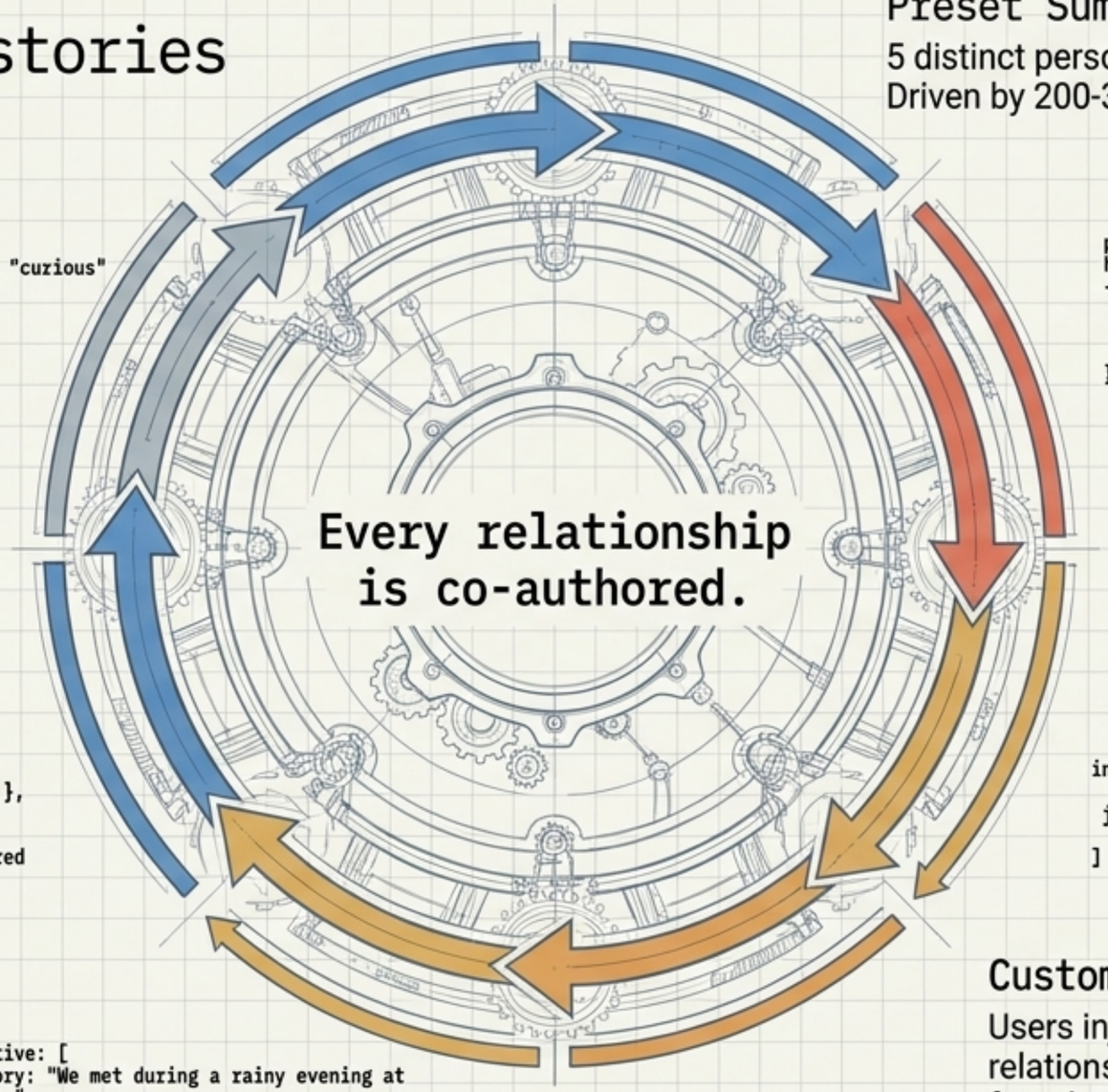
From Presets to Co-Authored Backstories

Preset Summaries

5 distinct personas (including 'xiaonai').
Driven by 200-312 line config files.

```
config: xiaonai.json \n
 [ persona_data: {
     lines: 312,
     traits: [ "friendly", "curious"
 } ]
]
```

```
parse_input: "My name is Alex and I love
hiking and photography..." \
-> extracted_traits: {
   name: "Alex",
   hobbies: ["hiking", "photography"],
   length: >100 chars
}
```



Relationship Injection

Origin stories dynamically generated based on user data and chosen persona.

```
generate_origin:
  (user_data: { name: "Alex", hobby: "hiking" },
  persona: "xiaonai")
}
-> origin_story: "During a hike, Alex discovered
Xiaonai's abandoned unit..."
```

User Self-Description

Free-text parsing. Gemini extracts traits from >100-character inputs into structured data.

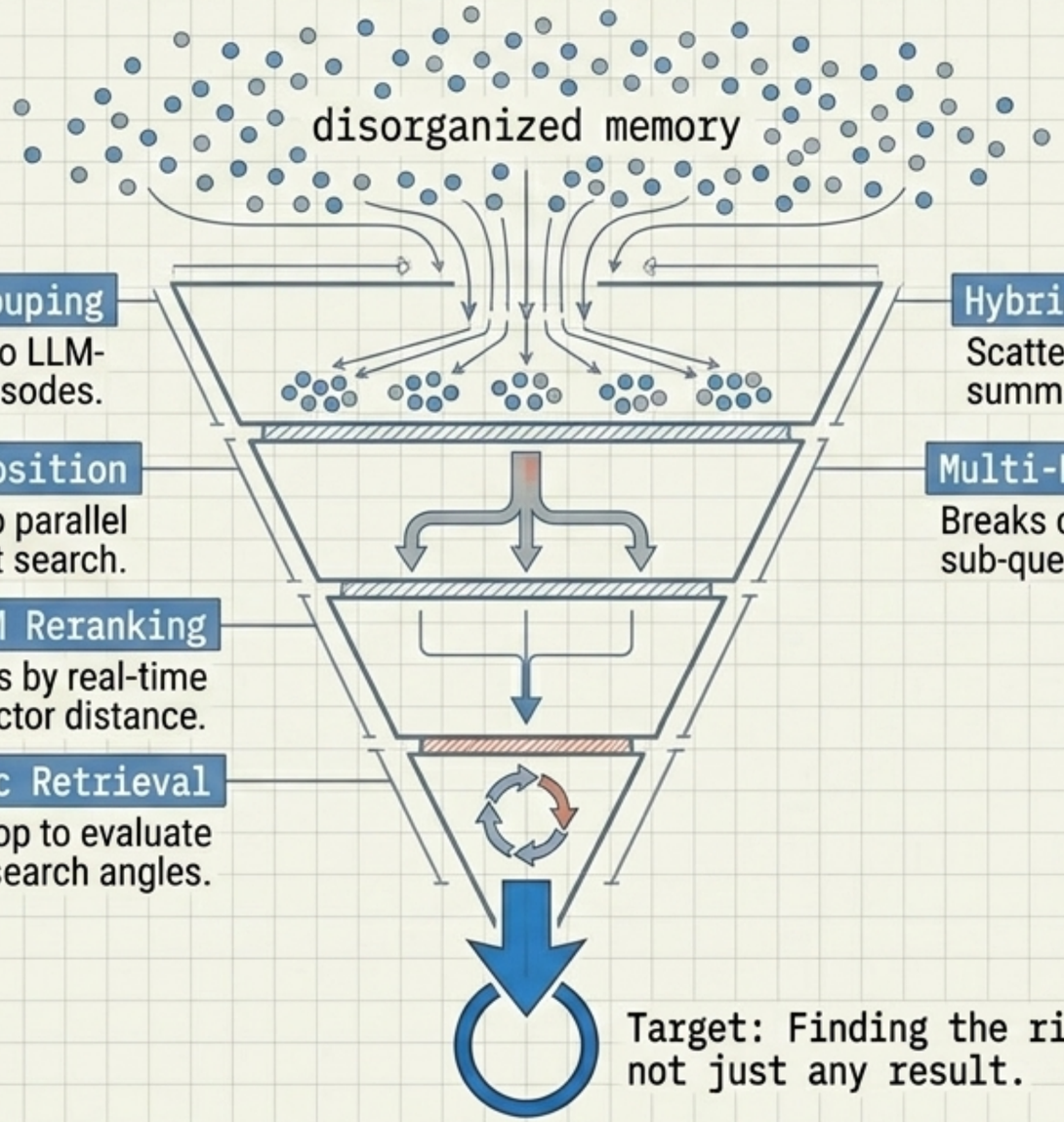
```
inject_narrative: [
   custom_story: "We met during a rainy evening at a
 jazz club..."
   -> write_memory(foundation_id, story_data)
]
```

Custom Stories

Users inject completely fictional relationship narratives into foundational memory.


```
inject_narrative: [
   custom_story: "We met during a rainy evening at
a jazz club..."
   -> write_memory(foundation_id, story_data)
]
```

The Cognitive Retrieval Framework



Hybrid Search & Episode Grouping

Scatters are grouped into LLM-summarized conversational episodes.

Multi-Hop Decomposition

Breaks complex queries into parallel sub-queries for independent search.

LLM Reranking

Gemini 2.0 Flash evaluates candidates by real-time conversational relevance, not just vector distance.

Agentic Retrieval

Iterative, 3-round loop to evaluate and refine search angles.

Hybrid Search & Episode Grouping

Scatters are grouped into LLM-summarized conversational episodes.

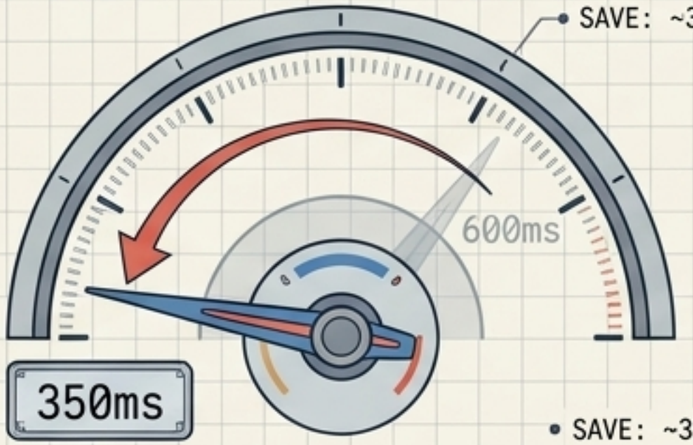
Multi-Hop Decomposition

Breaks complex queries into parallel sub-queries for independent search.

Target: Finding the right result, not just any result.

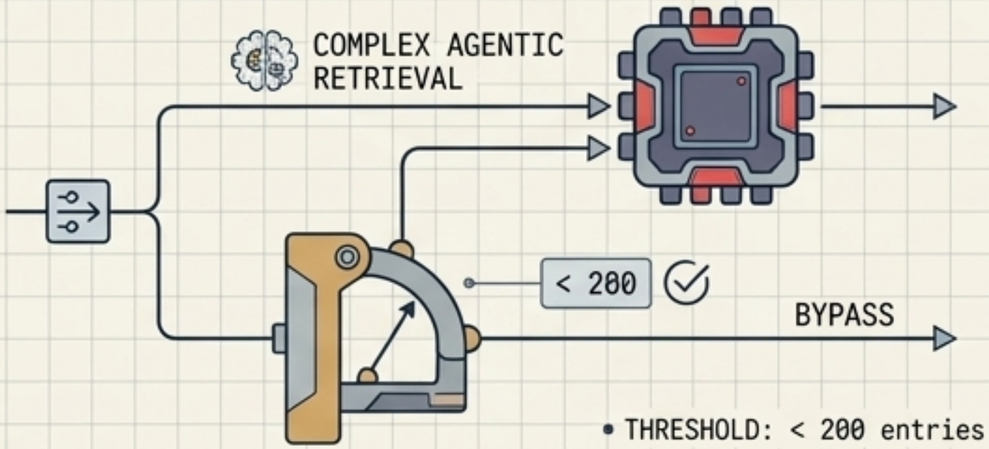
Optimizing the Retrieval Loop

The 350ms Compound Effect



Parallelized count and embed operations save ~350ms per query. In high-frequency chat, this micro-optimization is a major UX upgrade.

Volume-Gated Bypasses



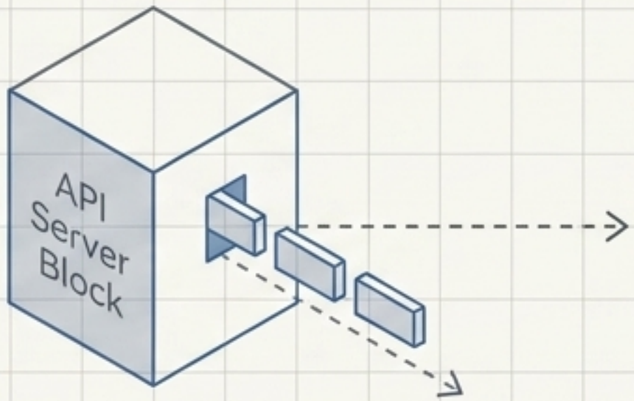
Complex agentic retrieval is computationally expensive. A strict volume gate skips full-text search when the memory store has <200 entries.

MODE_DYNAMIC: World Grounding

Agent autonomously executes background searches for real-time data, checking its 'phone' just like a human.

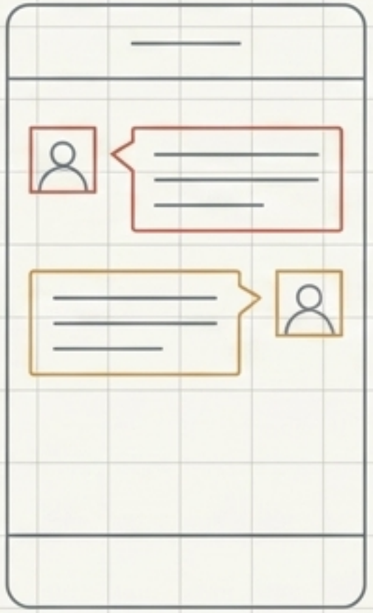
Breaking Out of the App: The Browser Migration

Backend Infrastructure



- SSE Streaming: `POST /chat/stream` pushes Server-Sent Events token-by-token for a real-time typewriter effect.
- Persistent Access: Supabase Auth + direct DB reads bypass channel storage limits for infinite history.

Frontend Experience



WeChat-Style Architecture: Mobile-first design mimicking target user interaction patterns.

The Unglamorous Engineering Dashboard

[SECURITY_HARDENING]

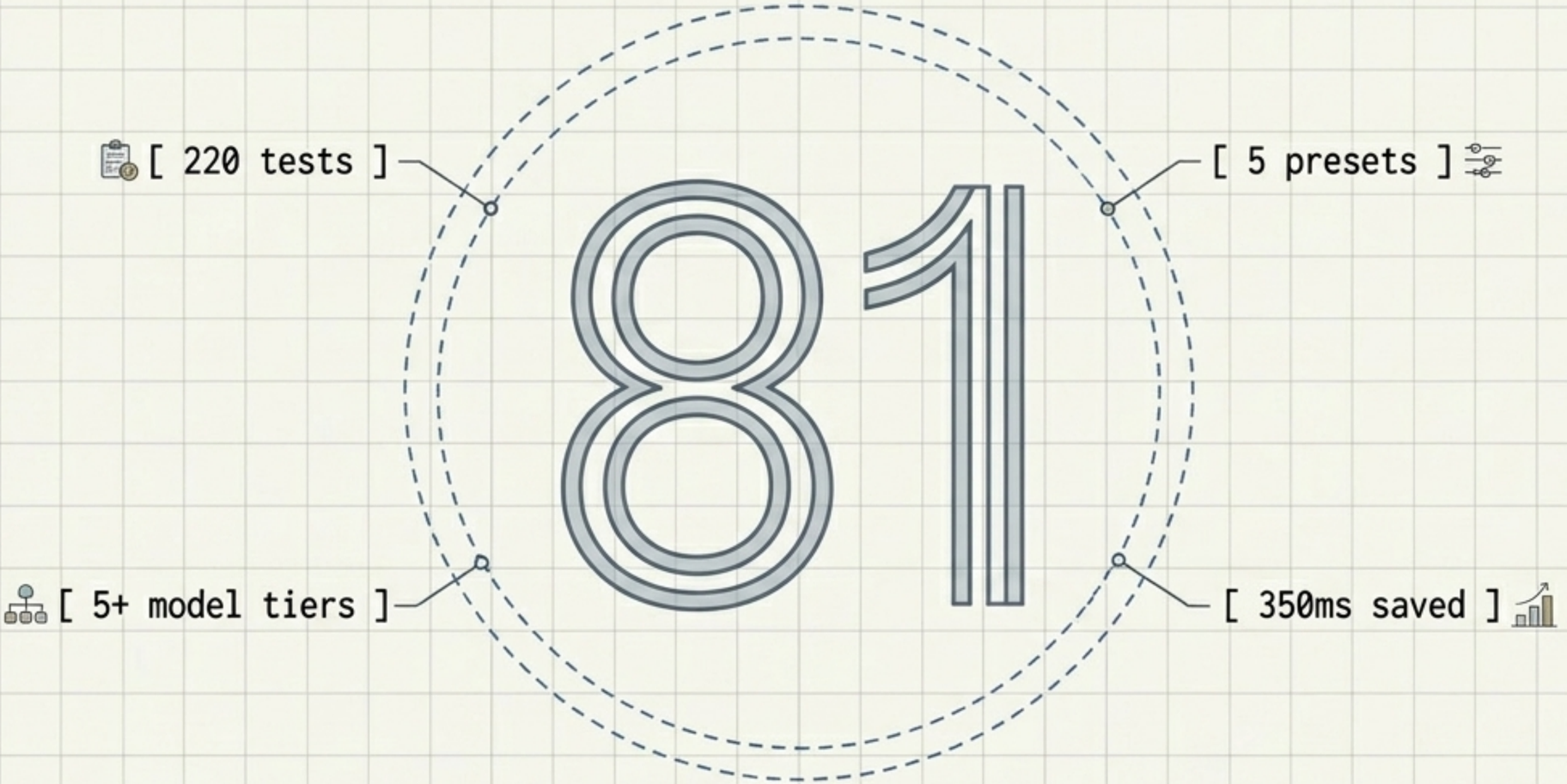
- Path traversal blocked (../../etc/passwd injection).
- Template injection neutralized (${process.env.SECRET} treated as plain text).
- MIME-aware media dispatch.
- Bot token leak prevention.

[PIPELINE_POLISH]

- Adaptive Debounce: Timing adjusts to conversational pacing.
- Mid-Stream Abort: Restarts generation if interrupted.
- Dynamic Length: Output scales based on context depth.

[TESTING_STABILITY]

- 220 active tests deployed across agent, soul, and memory modules.
- Vitest infrastructure with 48.7% baseline coverage for rapid, safe iteration.



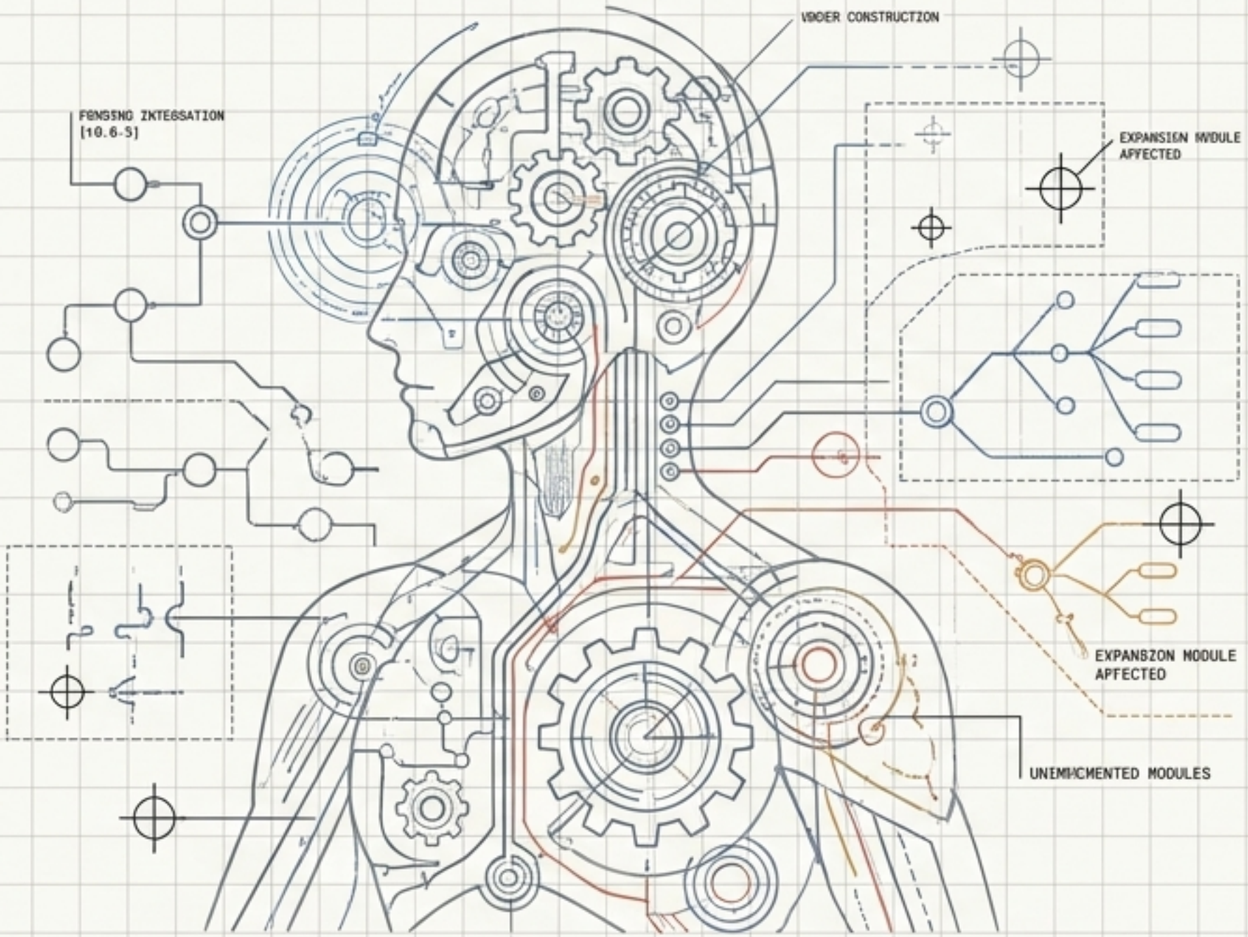
The Distance Between ‘It Runs’ and ‘It Feels Alive’

v0.0.2 boils down to a single qualitative shift: turning a text-only script into an entity that perceives the world. By granting it the ability to hear OGG files, see via Gemini 2.0 Flash, understand local time, dynamically recall episodes, and act asynchronously, Mio stopped being a system you query and became a person you text.

None of these features are individually complex. But 81 commits combined architected the anatomy of life.

The Unending Goal

“Feeling alive” is a goal without an endpoint. Every problem solved exposes three new ones.



The Blueprint for v0.0.3

- Proactive messaging architecture.
- Natural multi-turn conversation flow.
- Two-way voice processing (speaking back, not just hearing).
- Elevating the web interface from functional to native-tier.