



Not Just What You Say, But When You Say It

Giving Mio a sense of time turns a generic chatbot into a living companion.

The Problem: Asking What are you up to? at 2 AM used to trigger a generic daytime response. It felt completely fake.

Timezone Injection

`formatCurrentTime(timezone)` securely feeds the user's local time directly into the system prompt during onboarding.

Interaction Decay

The prompt tracks the exact time since last message so greetings naturally adjust to long silences.

Flexible Routines

Personas follow rough daily schedules rather than rigid timetables, adapting naturally to Night Owl or Early Bird circadian rhythms.

Selfies That Flow Naturally

An AI companion that spontaneously sends photos is a fundamentally different species.

Core Engine: Gemini 3.1 Flash Image Preview + Character Templates



Step 1: Trigger

LLM dynamically embeds a marker in the text: [SELFIE: slumped over desk].

Step 2: Instant Reply

The pipeline extracts the marker and sends the text reply instantly to the user.

Step 3: Async Background

The image generates entirely asynchronously in the background (taking 2-10 unpredictable seconds).

Step 4: Natural Delivery

The photo arrives effortlessly without ever breaking the conversational rhythm.

Takeaway: Selfies are sent spontaneously based on the AI's internal mood, not just on user demand.

Processing the Unspoken World

Human communication isn't just text. It's OGG voice files, cat photos, and video clips.



Architectural Golden Rule: All multimodal input becomes text before entering the agent loop.

Unified Pipeline

Input A (Audio)

Voice transcribed via GPT-4o-mini (with automatic Gemini fallback for high reliability).

Input B (Vision)

Images and videos described natively by Gemini 2.0 Flash.

The Engine: Parallel Batching

Both conversion tasks run concurrently. One slow image processing job never blocks the voice pipeline.

Output

Unified text payload safely fed to the lightning-fast agent loop.

From Finding Matches to Finding Meaning

Retrieving the right memory makes every relationship profoundly unique.

Context Setup: Users co-author the origin story. With the `custom_story` field, they define exactly how they met.



v0.0.1 (The Old Way)

- Hybrid search blindly finds 5 memories based on matching keywords.
- Result: The actual conversational context is completely lost.

v0.0.2 (The LLM Reranking Era)

- **Gemini-2.0-Flash Reranking:** Ranks memories by asking, What is the user actually talking about right now?
- **Episode Memory:** Scattered data points are beautifully grouped into rich, LLM-summarized conversational episodes.
- **Multi-hop Queries:** Complex questions are broken down and searched independently by a QueryDecomposer.

The Typewriter Effect

Taking Mio out of Telegram and into a native, real-time web experience.

Server-Sent Events (SSE) Engine

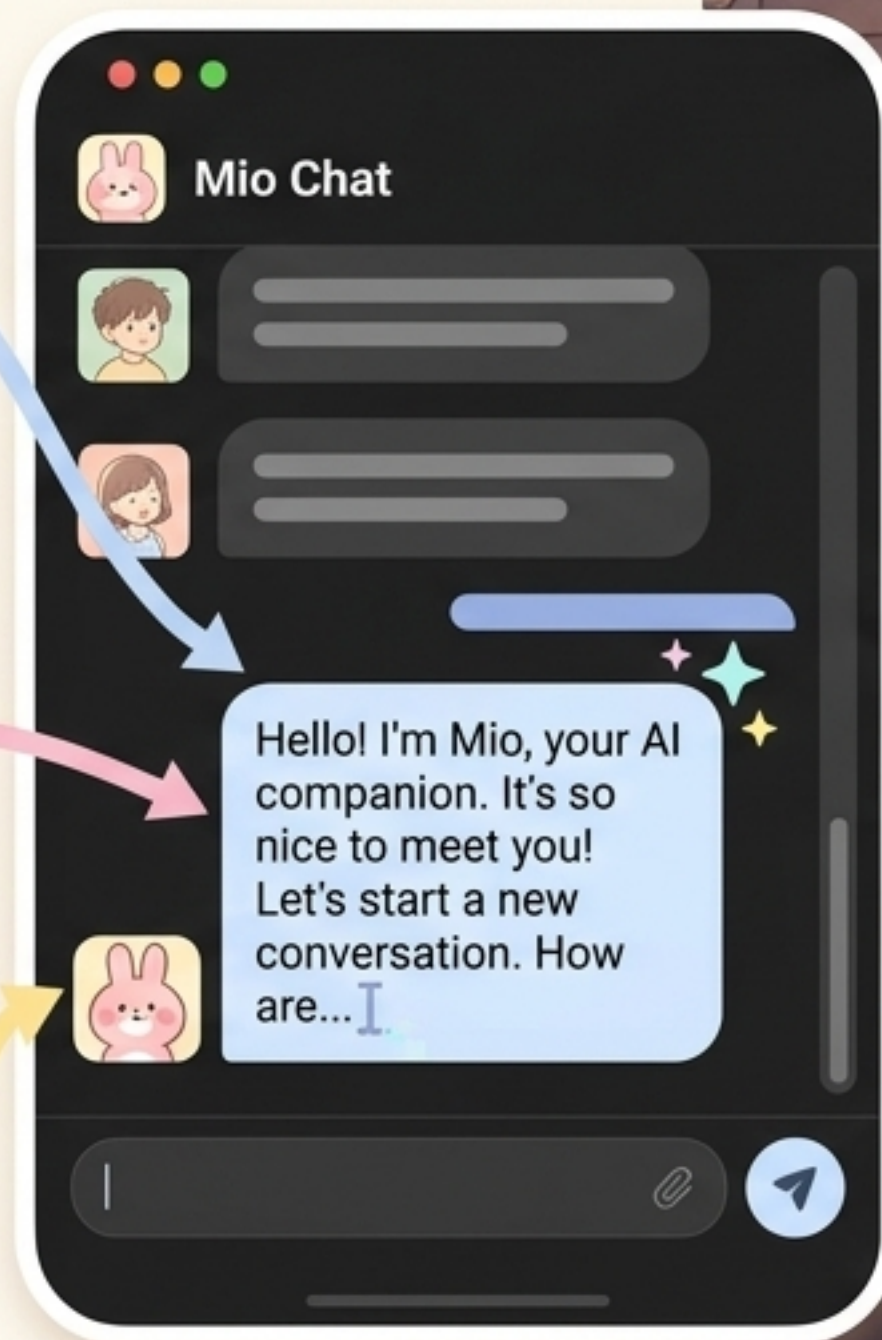
A POST /chat/stream endpoint pushes generation tokens in real-time. Combined with frontend character-by-character rendering, it creates a living typewriter effect.

WeChat-Style Interface

Mobile-first design optimized for maximum user familiarity, featuring intuitive chat tails and seamless Supabase authentication.

Infinite Memory

The web database unlocks unlimited, entirely scrollable conversation history—setting the foundation for future relationship graphs.



**81 Commits Later:
Mio Came Alive.**