

The V0.0.5 Burn Rate



Waiting 10 seconds for a reply to “How was your day?” kills the companion experience. The features were complete, but the latency was unbearable.

Routing by Task & Seeing Clearly

90% of Traffic: Daily Chat

Route: Gemini 3 Flash (Thinking: Minimal)

Metrics: 1-2s TTFT | 214 tokens/sec | 4x Cheaper

Impact: Lightning-fast emotional conversational flow.

10% of Traffic: Premium Tasks

Route: Gemini 3.1 Pro (Thinking: Low)

Metrics: High cognitive depth

Impact: Deep personality extraction and memory summarization.



Before: I see a colorful game.

Now: Looks like you're playing Genshin Impact — how's the new region?



Curing Voice Hallucinations

Mandarin voice clips were hallucinating English artifacts, breaking the most intimate input modality.

Step 1: Explicit Targeting

Added the 'zh' parameter to the OpenAI

Effect: Stops auto-detect failures caused by code-switching and background noise.

Step 2: The Nuance Net

Enhanced Gemini Fallback Prompt.

Effect: Catches and correctly parses colloquial Chinese slang, internet-speak, and proper nouns when OpenAI lacks confidence.

CLEAN TEXT:
[Correct Mandarin Transcript]



DRY Refactoring

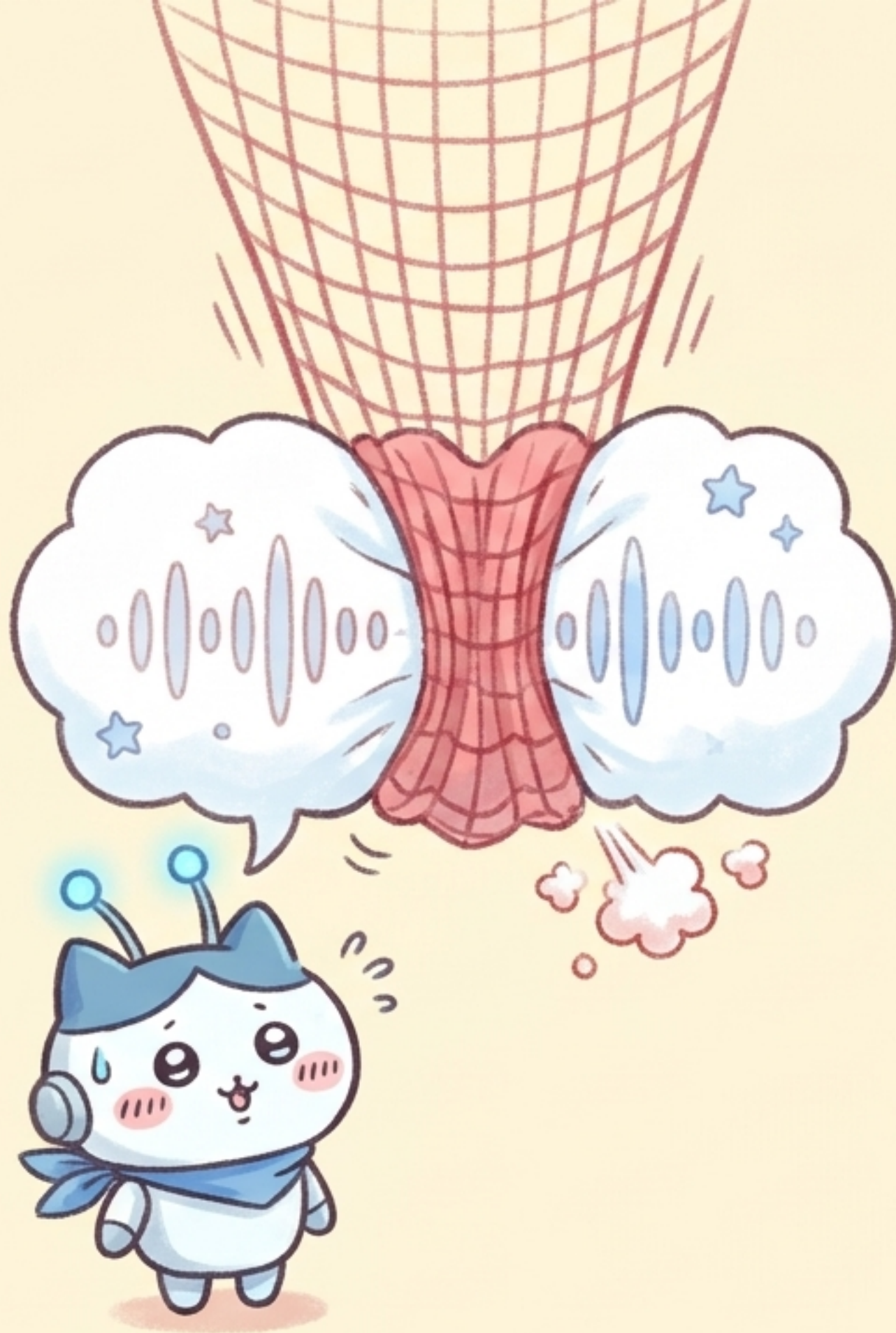


Cleaning up the technical debt from the v0.0.5 media update.

- ✓ -80 Lines of duplicated server code eliminated.
- ✓ Extracted shared logic into **resolveMedia()** (fetches and validates uploads).
- ✓ Extracted shared logic into **prepareChatContext()** (builds history, persona, and media).

Result: Both **POST /chat** and **POST /chat/stream** handlers now share elegant, unified logic.

Shielding Against Platform Risk



Google explicitly positions Gemini as a “super tool, not an emotional companion.”

External safety filters deleting responses mid-sentence.

Provider hostility toward roleplay and companion use cases.

The Strategic Pivot: Model-Agnostic Architecture

The v0.0.6 routing layer isn't just for cost savings. It is a defensive shield that ensures Mio can swap AI providers instantly without rewriting the core application.