Voice, Emotion, and the Screenwriter-Actor Model

ARCHITECTURAL PARADIGM SHIFT

LEGACY AI MODEL → ~~GENERIC TEXT-TO-SPEECH~~

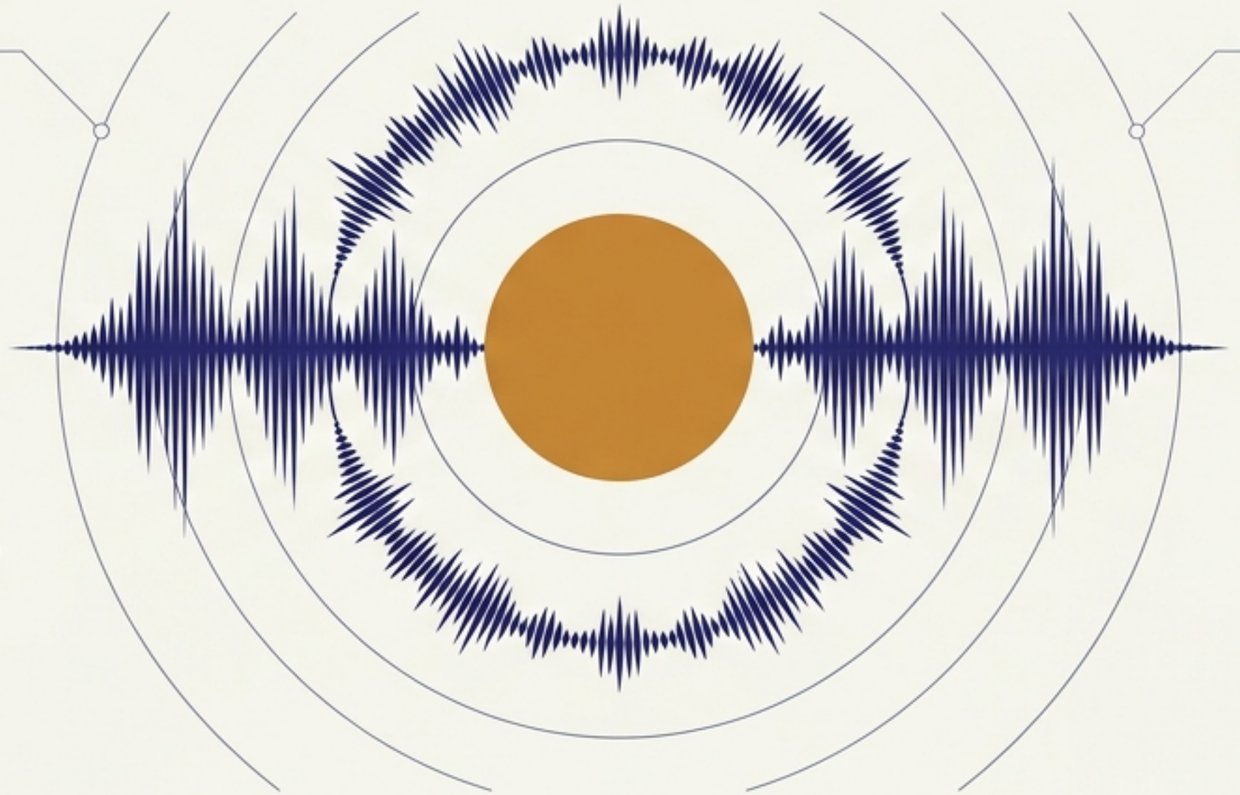
SCREENWRITER LLM ⇄ ACTOR MODEL (Hume EVI)

- GPT-4
- Claude 3 Opus
- Context Awareness

EMOTIONAL SYNTHESIS LOOP

- Prosody Control
- Emotional Intelligence
- Realtime Synthesis

Decoupling Intelligence (Soul) from Expression (Voice).



CORE TECHNICAL PRINCIPLES

1. THE SCREENWRITER (SOUL)

- ROLE: Narrative, Context, Intent
- INPUT: Multimodal Data Stream (Text, Audio)
- OUTPUT: Structured Script & Emotional Metadata (e.g., {sentiment: "joyful", intensity: 0.8})

2. THE ACTOR (VOICE)

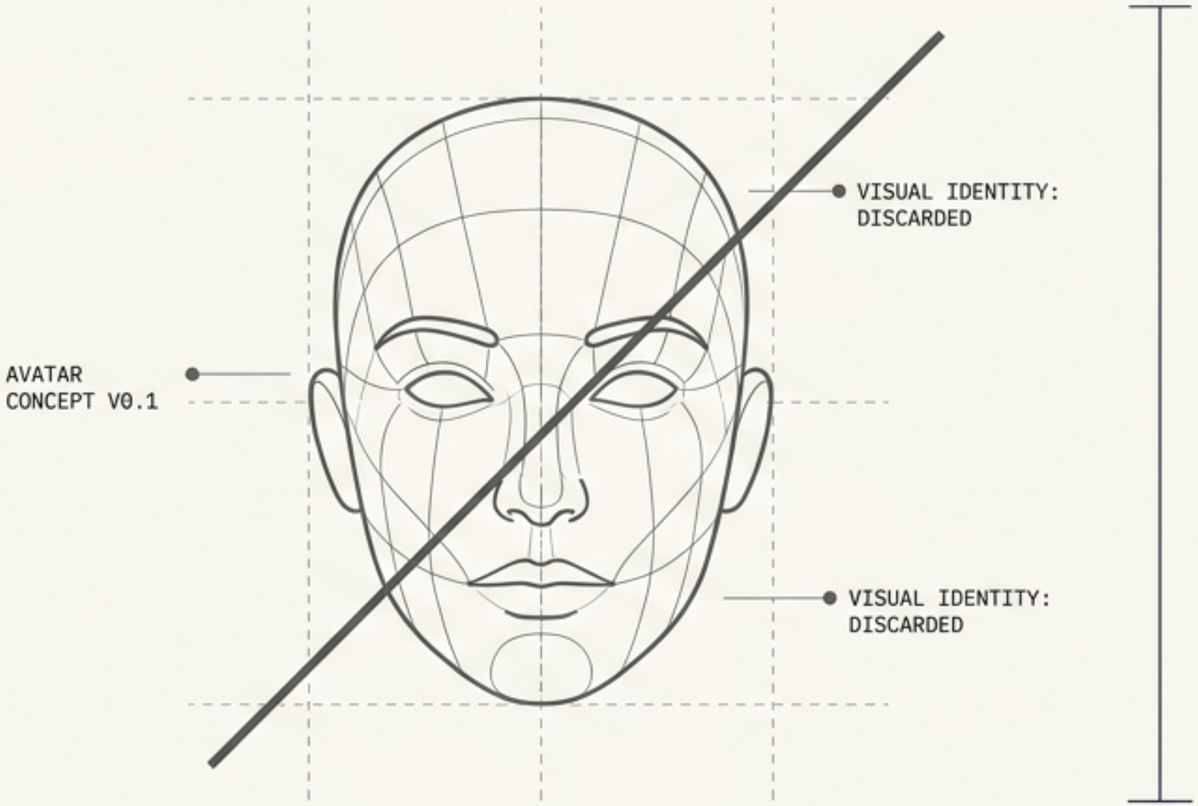
- ROLE: Performance, Prosody, Realtime Synthesis
- INPUT: Script + Emotional Metadata
- OUTPUT: High-Fidelity, Emotionally Rich Audio Waveform

3. REALTIME INTERACTION LOOP

- LATENCY: Sub-580ms Response Time
- FEEDBACK: User Emotional Reaction Analysis (Hume API)
- ADAPTATION: Dynamic Script & Performance Adjustment

Rebuilding Mio: The architectural shift to emotionally intelligent realtime AI voice.

Strip away the physical. Keep the soul.



A companion is defined by memory and conversation, not a fake face.

But a soul needs a voice.

In v0.1.0, Mio learned to speak using cloned TTS. It sounded natural, but speaking isn't the same as feeling.

Words carry emotion.
The voice just carries words.

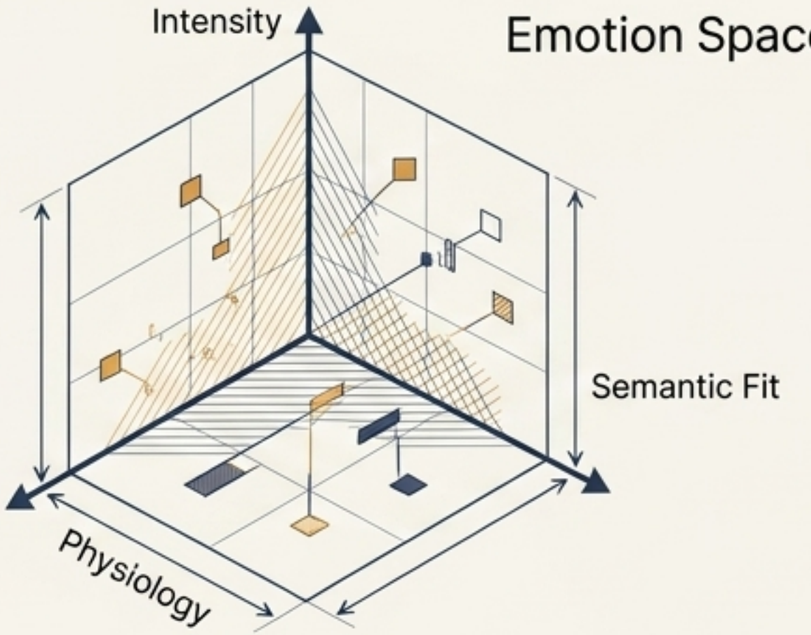
Diagnostic Flow Diagram

The Manual Markup Dead-End

- Manual SSML tags ('say this sadly') work for audiobooks.
- They are structurally unworkable for an AI generating thousands of dynamic, unpredictable messages with complex subtext.

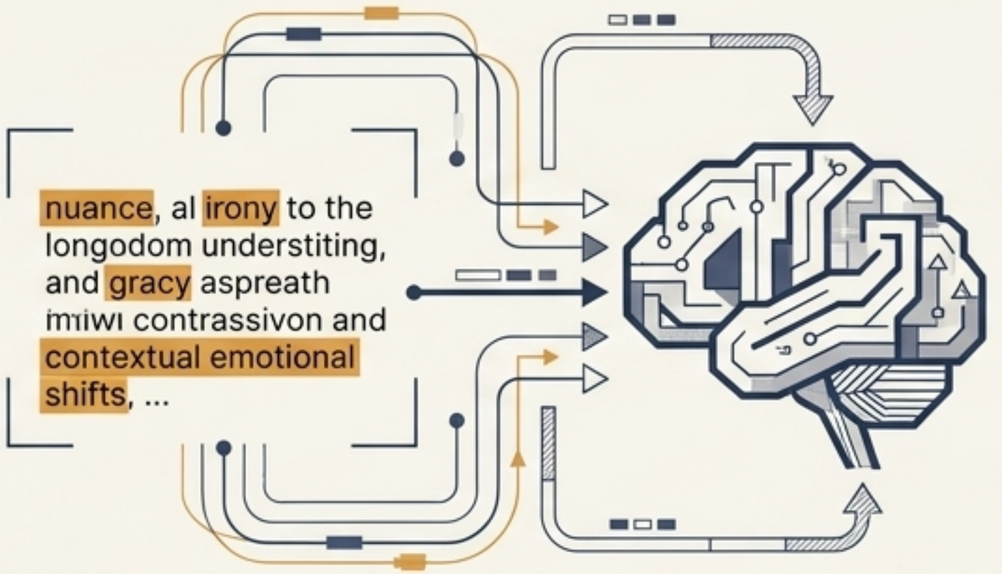
From ‘Reading’ to ‘Understanding and Expressing’

Doubao TTS 2.0



90%+ accurate emotional prosody inference without manual tags. Trained on 2,000 hours of acting data.

Hume AI Octave

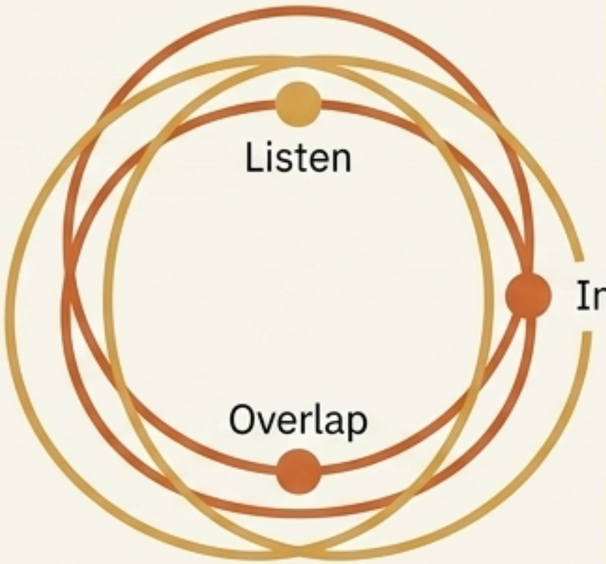


Doesn't just map sentiment; uses an LLM to ‘read’ nuance, irony, and contextual emotional shifts like a human voice actor.

This is the non-realtime bridge (~200ms latency buffer) powered by gpt-4o-mini-transcribe. Perfect for voice messages.

Voice messages are a bridge. Realtime is the destination.

Interruptions
(Handling overlapping speech gracefully)



Turn-taking
(Knowing when the user is actually done)

Vocal Tone
(Capturing user emotion that pure text transcription destroys)

The Latency Chain

4 sequential steps = 1 to 3 seconds of latency.
An eternity in spoken conversation.

Evaluating the V1.0 Architecture

	Hume EVI 3	OpenAI Realtime	Doubao Realtime
Custom LLM	✅ [Gemini/Claude/Any]	❌ [GPT only]	❌ [Doubao only]
Emotion in Voice	Best in class	Weak (“future”)	Good (3D space)
Interruption	Tone-based	Yes	Yes
Latency	~500ms TTFB	~300ms	~700ms

The Fatal Flaw of Big Tech APIs. They force you into their monolithic models. For a companion product, locking out custom memory and system prompts destroys the personality.

The Screenwriter and The Actor

The Screenwriter

Custom LLM
(Gemini / Claude)

- Memory Injection
- Personality Prompt
- Conversation History
- Reasoning

Writes the script based on deep context.

The Actor

Hume EVI

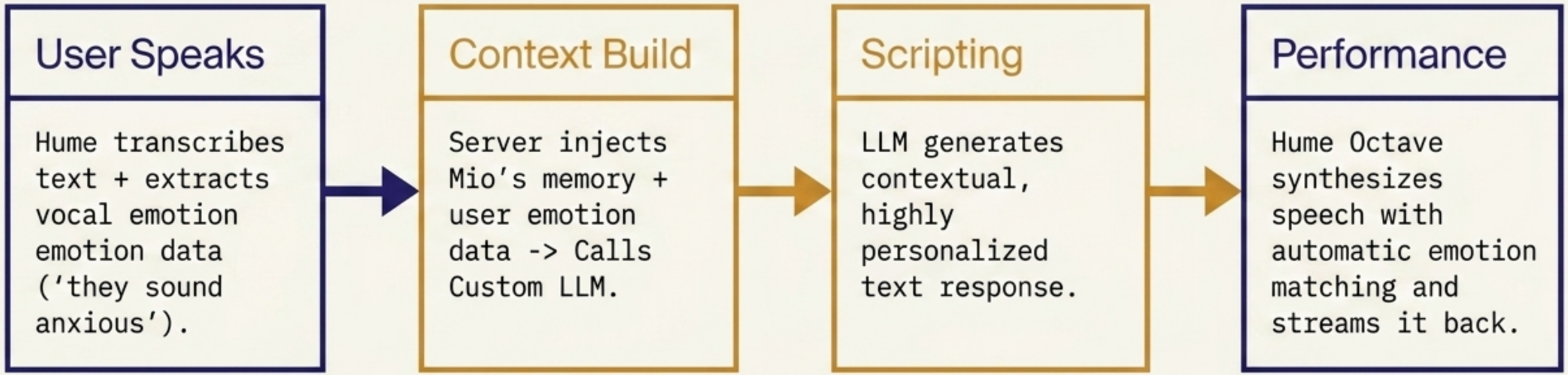
- Listening/STT
- Emotion Detection
- Turn-Taking
- Expressive Delivery

Performs the script, manages the stage, handles the unexpected.

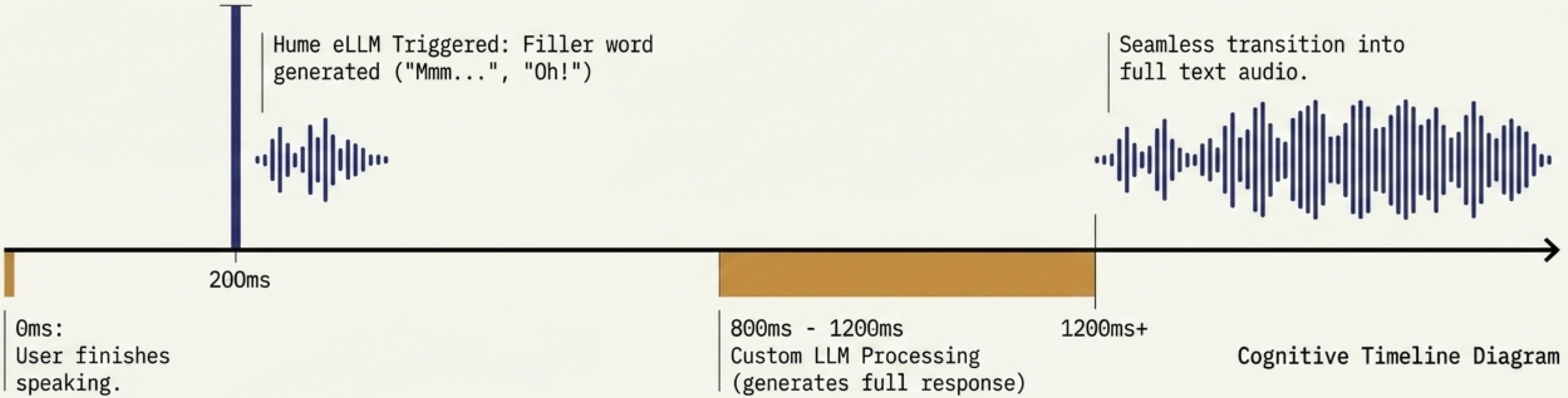
Strict separation of concerns. The LLM focuses on *what* to say. Hume focuses on *how* to deliver it.

The Architecture of a Conversation

Premium, editorial whitepaper



Improvising while the Writer writes.



Cognitive Timeline Diagram

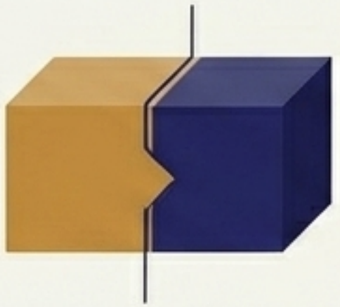
Honest Latency

Humans don't experience silence as "fast thinking" – they experience it as disconnection. A 200ms filler word creates perceived zero-latency and feels psychologically sound.

Beyond Monolithic Models

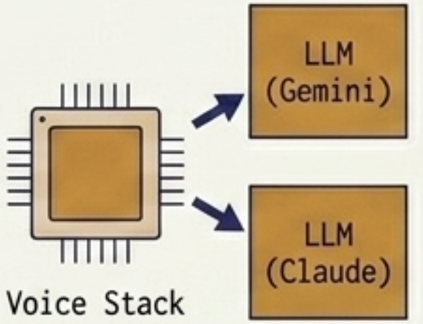
1. Separation of Concerns

LLM handles content logic; Hume handles delivery mechanics. Independent evolution.



2. Vendor Flexibility

Swap out the LLM (Gemini to Claude) anytime as models advance, without touching the voice stack.

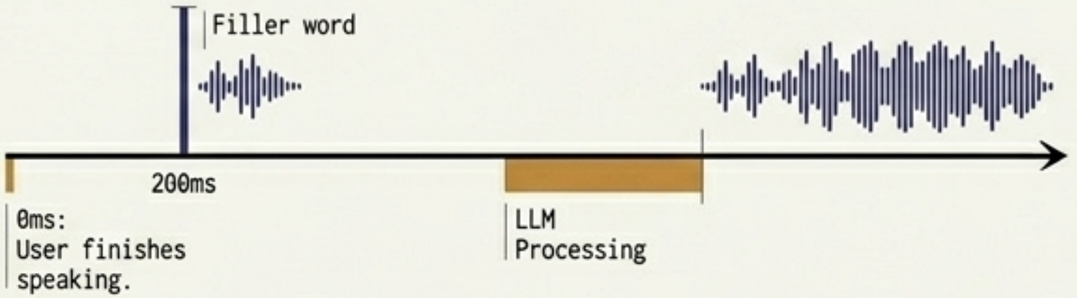


3. Emotion as First-Class Signal

Analyzing *how* the user speaks (vocal tone) enriches the LLM prompt context beyond mere text.

4. Honest Latency

The filler-word pattern creates a psychologically natural conversational flow.



Realtime Voice Unit Economics

Pro Tier (Base Subscription)

Text Chat
+
Voice Messages
(Doubao / Octave TTS)

Affordable, proven, easily bundled.

Voice Tier (Premium Subscription)

Everything in Pro
+
Realtime Bidirectional Voice
(Hume EVI)

High per-minute API costs require separate pricing.

Realtime voice cannot be bundled. A single heavy user (10 mins/day) breaks a standard subscription model. Tier separation ensures sustainability.

The Voice is the Product

Big labs will not build deeply emotional companions due to brand risk. The Screenwriter-Actor model gives Mio a defensible moat: it leverages Big Tech's raw performance, but directs it with custom, untouchable emotional depth.