The Architecture of a Voice

Engineering Text-to-Speech for Autonomous AI Companions

```
PROJECT: OpenClaw Field Notes
DOCUMENT: Post-Mortem 03
TARGET: sys.audio.init()
```

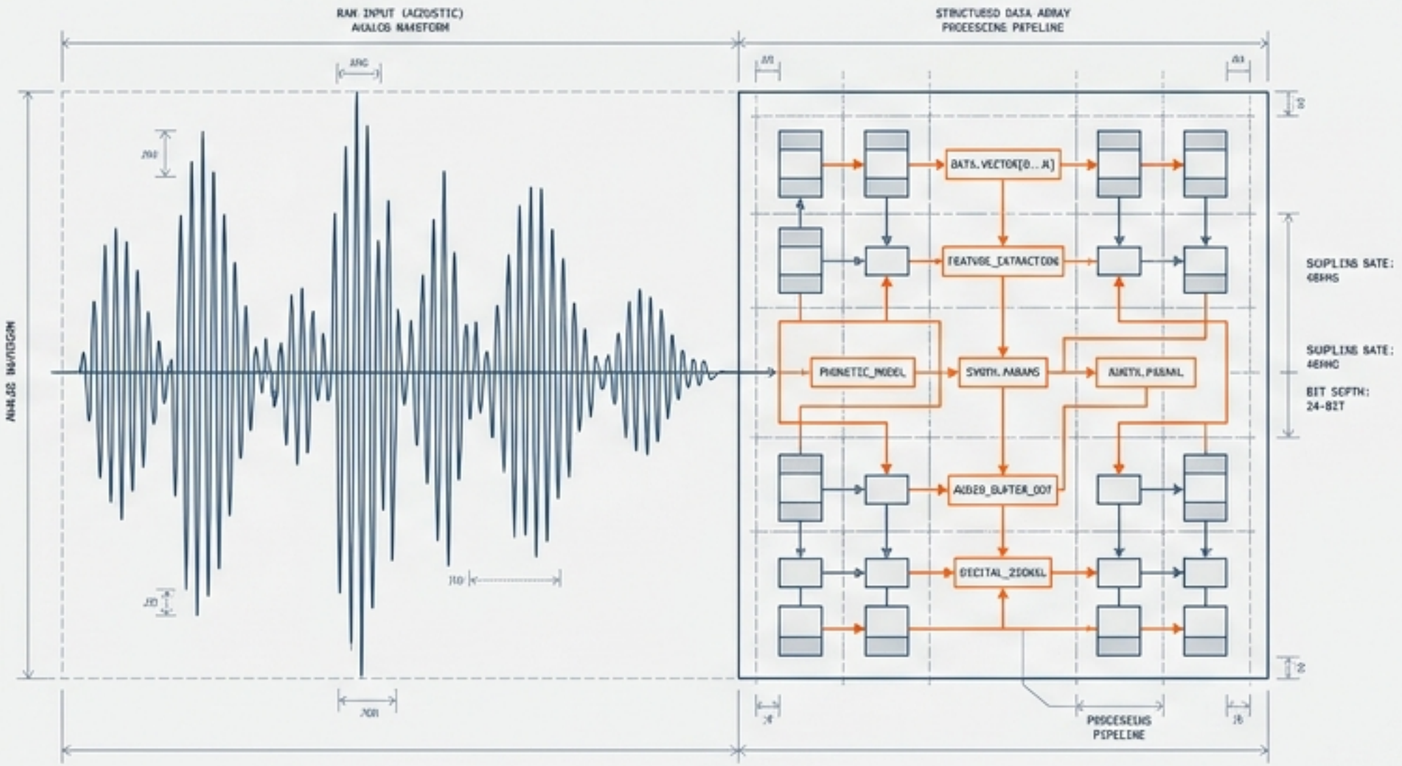
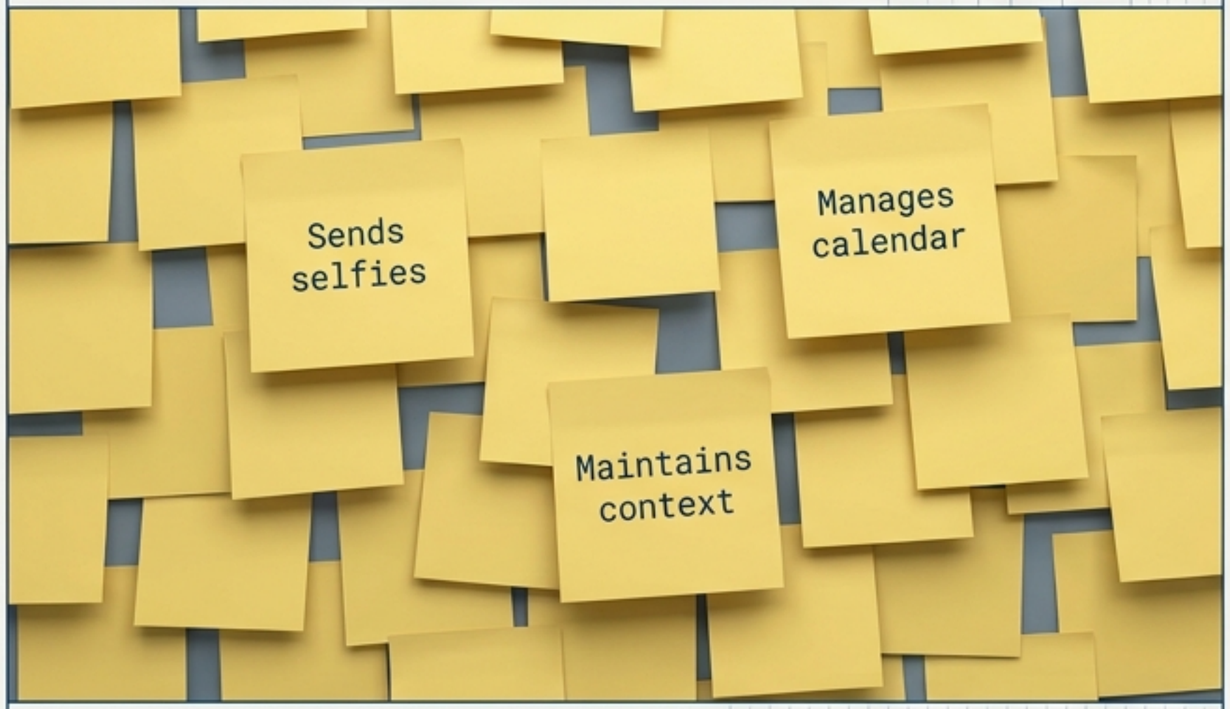


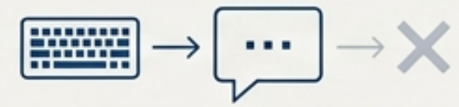
Figure 1.1: Transformation from acoustic input to structured digital signal for AI voice generation.

A companion with a soul needs a voice.

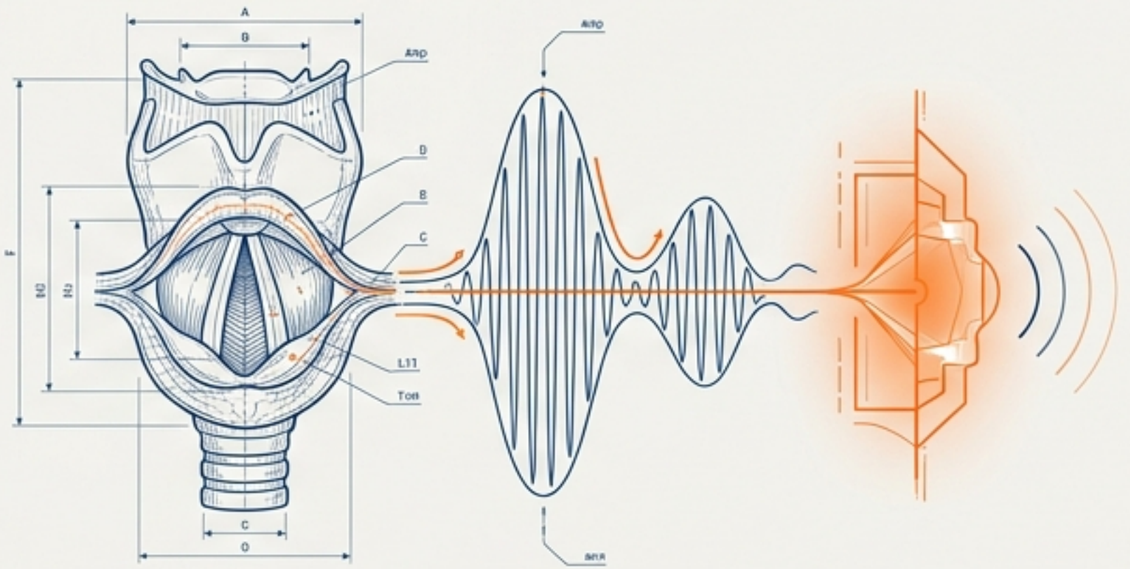
The Silent System



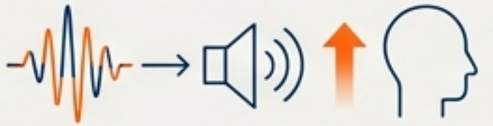
The system functions perfectly as a digital assistant. But performing an emotional screenplay through text chat bubbles breaks the illusion. Every sigh remains silent.



The Missing Piece

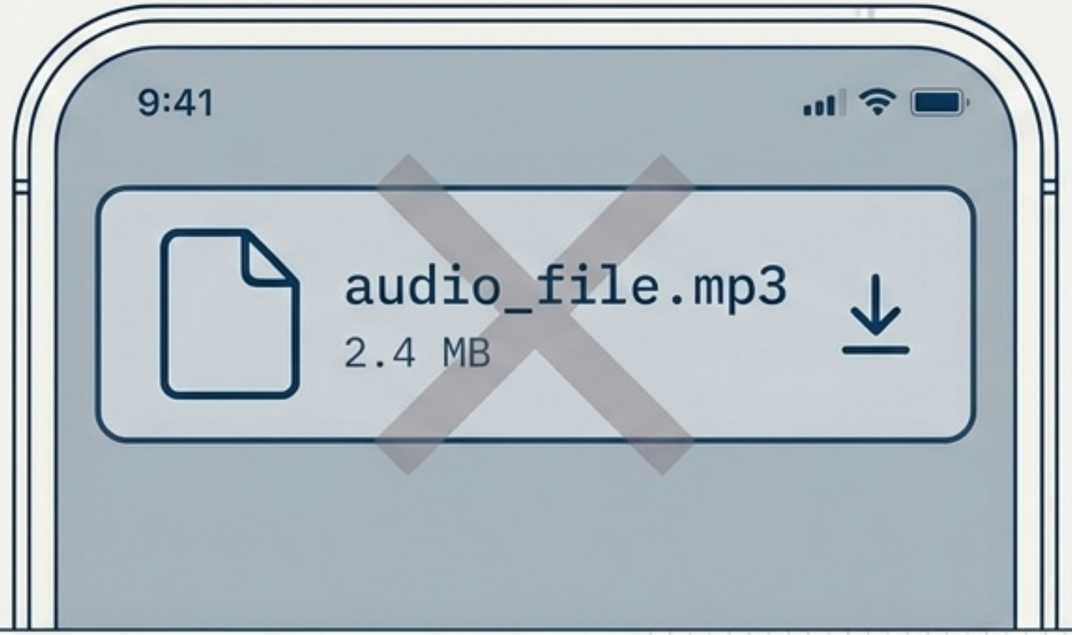


Voice is the threshold between software and presence. It transforms a scripted algorithm into a living persona.



The format dictates the psychology.

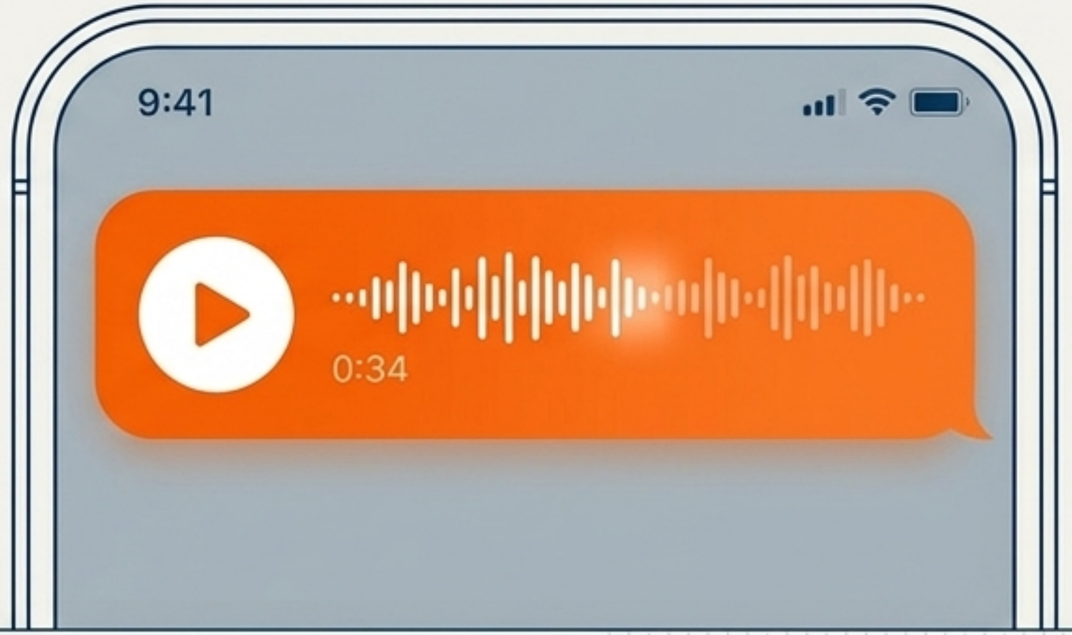
The Document



MP3 Format

Renders as a generic document attachment. Instantly kills the illusion of human interaction.

The Bubble



OGG/Opus Format

Telegram natively recognizes the format, generating a seamless voice bubble that mimics human presence.

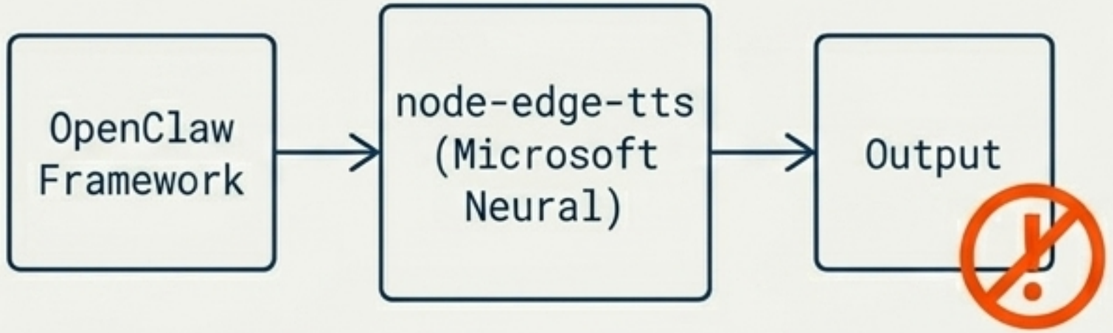
The OpenClaw TTS Sandbox

Provider	API Key	Output Format	Voice Bubble	Special Feature
Edge TTS	No	MP3	Document (No)	Free, zero config
OpenAI	Yes	Opus / MP3	Native (Opus)	Clean, reliable
ElevenLabs	Yes	Opus / MP3	Native (Opus)	Premium English quality
Fish Audio	Yes	OGG / Opus	Yes (Native)	Excellent Chinese voices, cost-effective
Volcano Engine	Yes	MP3	Yes (v2 hack)	Per-sentence LLM emotion control

> **STATUS:** TTS is off by default. Provider selection permanently determines output format, operational cost, and user experience.

The Quick Start: Edge TTS

Architecture & Command




```
> Restart gateway...
> /tts status
> Provider: edge (configured)
```

The Fatal Flaws Checklist

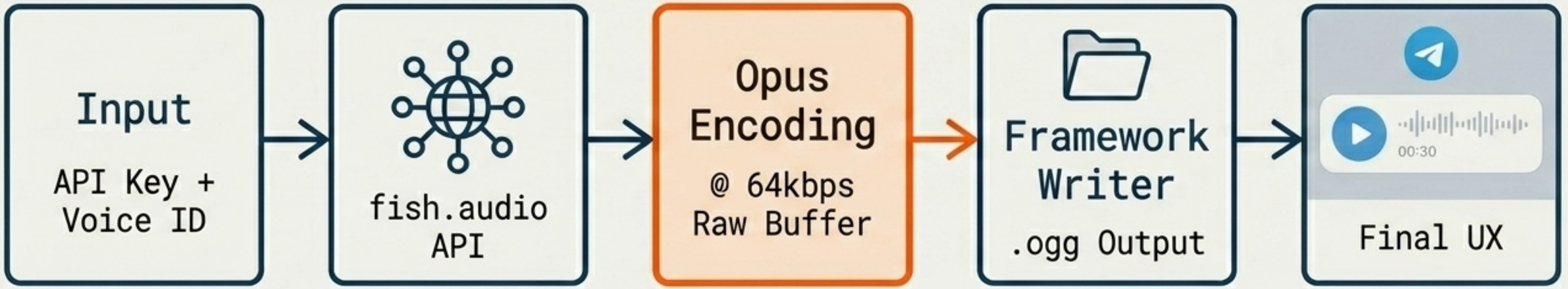
1. The Formatting Failure

Outputs MP3 exclusively. This triggers the dreaded document attachment UX, destroying the illusion of seamless conversation.

2. The Reliability Void

As a public web service, there is no SLA, no published rate limits, and no guaranteed uptime. Impossible to rely on for a 24/7 companion.

The Pragmatic Daily Driver: Fish Audio



Simple Setup

Requires only two configuration fields to integrate.

Low Latency

A single API call ensures rapid response times.

Natural Intonation

Solid Mandarin voicing avoids robotic cadence.

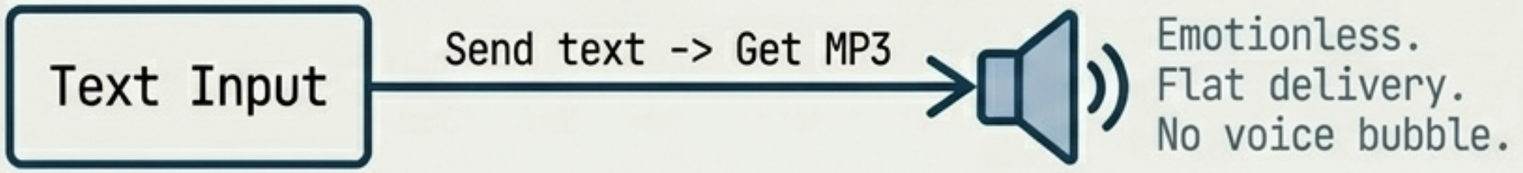
The Fish Audio Gotcha: Literal Translation

Fish Audio does not parse emotion markers. It speaks verbatim. When migrating from an emotion-based pipeline, the system prompt must be aggressively scrubbed of bracket instructions to prevent the AI from narrating its own stage directions aloud.

The Emotion Machine: Volcano Engine v2

Standard TTS dictates what to say. The seed-tts-2.0 model dictates how to speak.

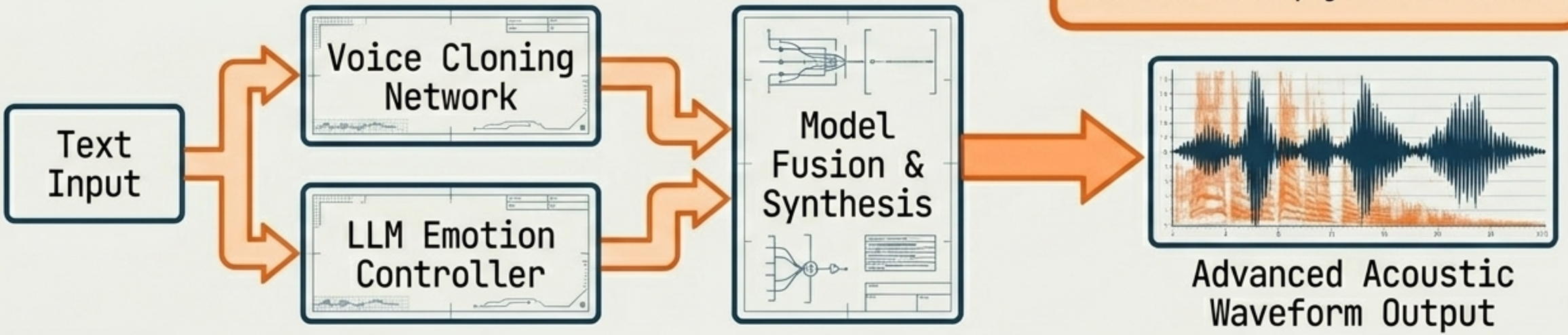
v1 (Legacy)



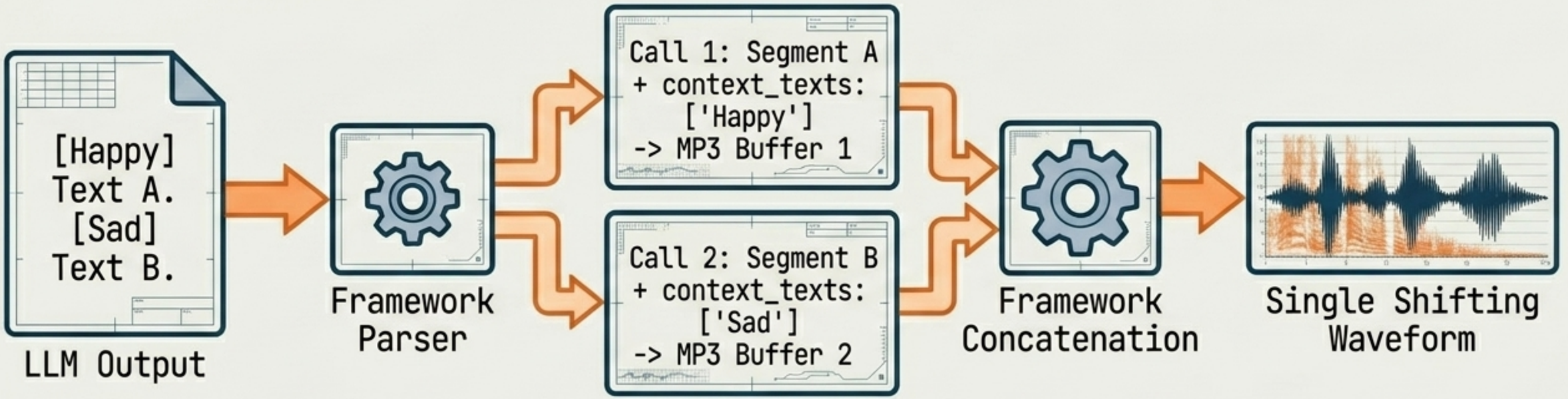
⚠ The Architecture Catch

The critical **context_texts** parameter acts as a stage direction. However, it is strictly limited by the API: it only applies to the very first sentence of any given API call.

v2 (seed-tts-2.0)



Engineering Emotion: The Multi-Call Architecture



The framework splits the parsed text, parallelizes the API calls to bypass the single-sentence limitation, and stitches the audio back together. The user sees clean text; the voice actor performs the hidden stage directions.

Emotion Marker Anatomy

Short Labels

[Happy]

Evaluation

Highy effective. The LLM generates them instantly, and the TTS API responds with maximum consistency.

Voice Commands

[Speak softly]

Evaluation

Functional but variable. Slightly slower generation and occasional misinterpetation by the audio model.

Scene Narration

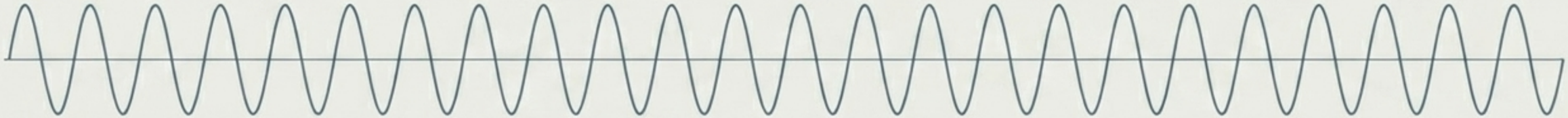
[Crying while looking away]

Evaluation

Overly complex. Frequent inconsistent execution where model ignores the instruction.

Forging an Identity: Voice Cloning

Built-In Speakers



Functional, but generic. Using built-in IDs fails to sell the illusion of a unique, localized companion.

Cloned Identity (The S_ Prefix)

Training a custom voice clone yields a unique speaker ID. This is the critical threshold where the output stops sounding like software and starts sounding like itself.

The Volcano v2 Hazard Log

Version Control

The 'version: v2' flag is mandatory. Omission silently defaults to v1, resulting in flat audio and literally spoken brackets.

The Stripping Trap

Do NOT use stripActionMarkers(). Bracket markers must survive the pipeline to successfully reach the emotion parser.

The Format Hack

Outputs MP3. The framework must inject 'voiceCompatible: true' to force Telegram to simulate a native bubble.

Ghost Overrides

Model overrides save to sessions.json and survive reloads. Clear session data to fix persistent emotion pipeline errors.

Ghost Overrides

Model overrides save to sessions.json and survive reloads. Clear session data to fix persistent emotion pipeline errors.

The Echo Bug

Hide the audio file path from the LLM. If exposed via a tool result, the message tool may resend it, duplicating the output.

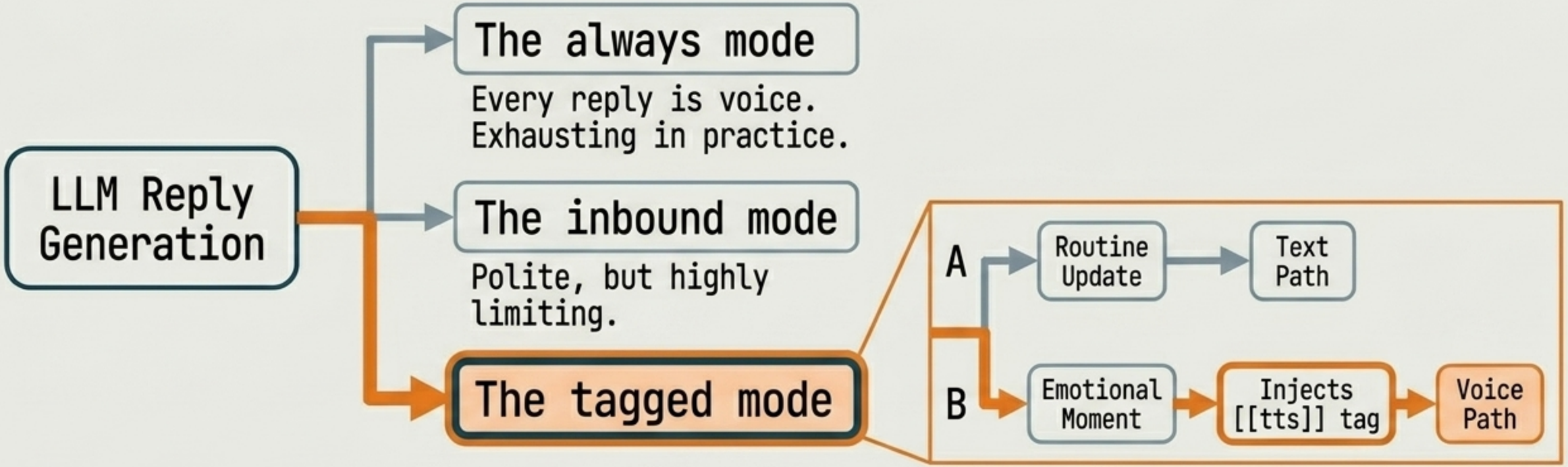
Diagnostic Matrix: Pragmatism vs. Ambition

Dimension	Fish Audio	Volcano v2
Setup Complexity	2 Fields (Minimal)	4+ Fields & Voice Clone
Emotion Control	None	Per-sentence context_texts
Output Format	Native OGG / Opus	MP3 with compatibility hack
Latency	Low (Single Call)	Higher (Multi-Call Routing)
Gotcha Surface Area	Extremely Small	Large (Parsing, Session Overrides)

Fish Audio wins for robust, reliable daily driving.

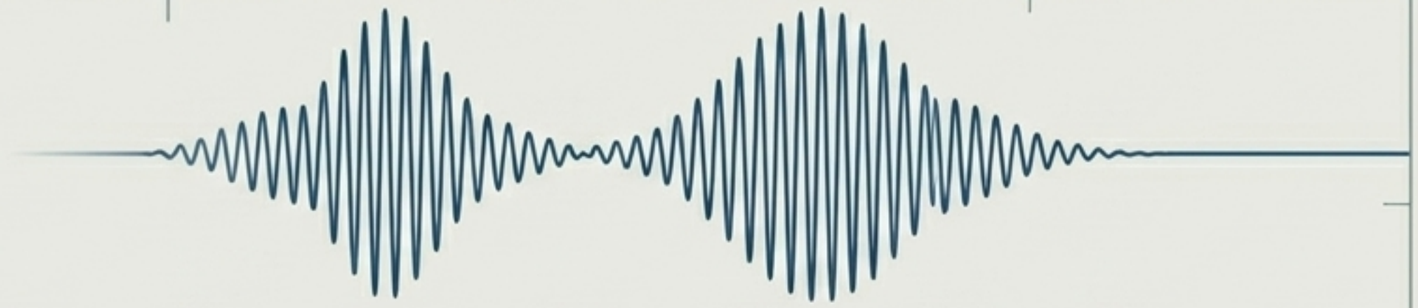
Volcano v2 wins for uncanny, dramatic AI actor demonstrations.

The Trigger Mechanics: System Design



The tagged mode teaches the model to autonomously deploy voice only when it adds emotional or contextual value, preventing user audio fatigue.

A Category Shift, Not a Feature.



“Text-based interactions, no matter how well-personified, still feel like chatting with software. A voice note that sounds natural, delivered as a round bubble with the right emotional tone — that feels like hearing from a person.”

- The technical setup is easily solvable.
- The architectural challenge is choosing the provider that fits the psychological use case.
- Fish Audio provides daily reliability; Volcano reveals the uncanny future of the AI actor.

NEXT IN SERIES: DEPLOYMENT, GCE, AND DOCKER PATCHES