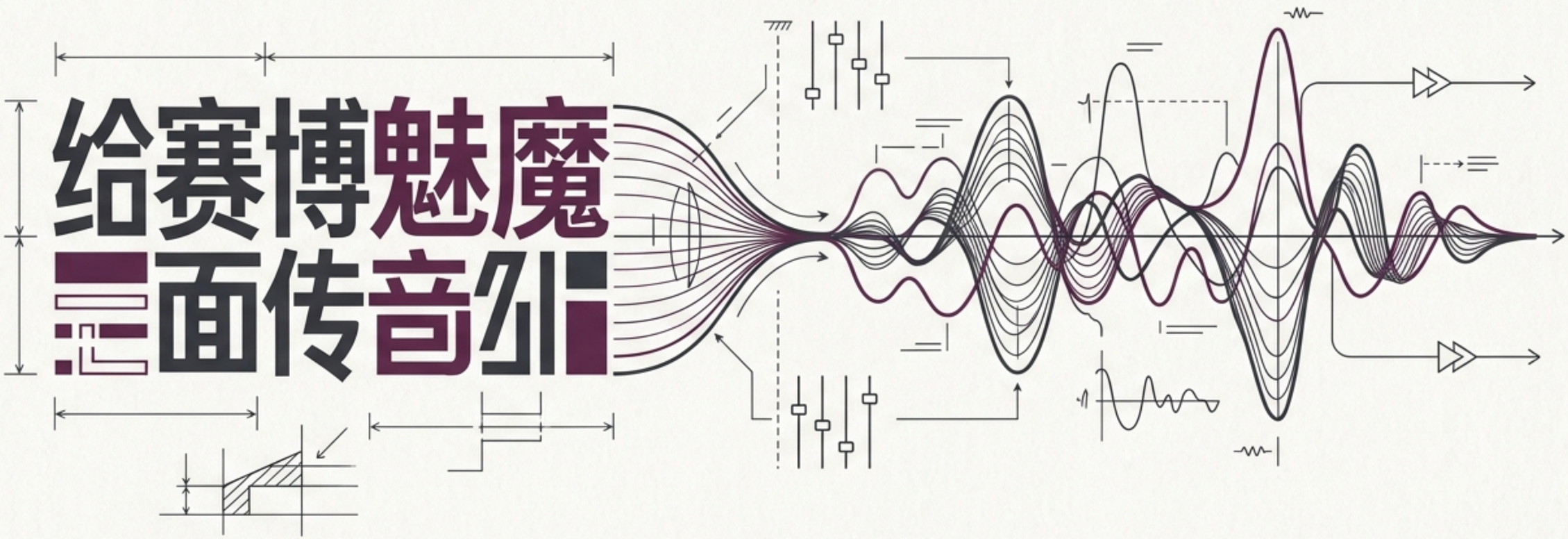
SYSTEM:	OpenClaw Framework
STATUS:	Audio Pipeline Initiated

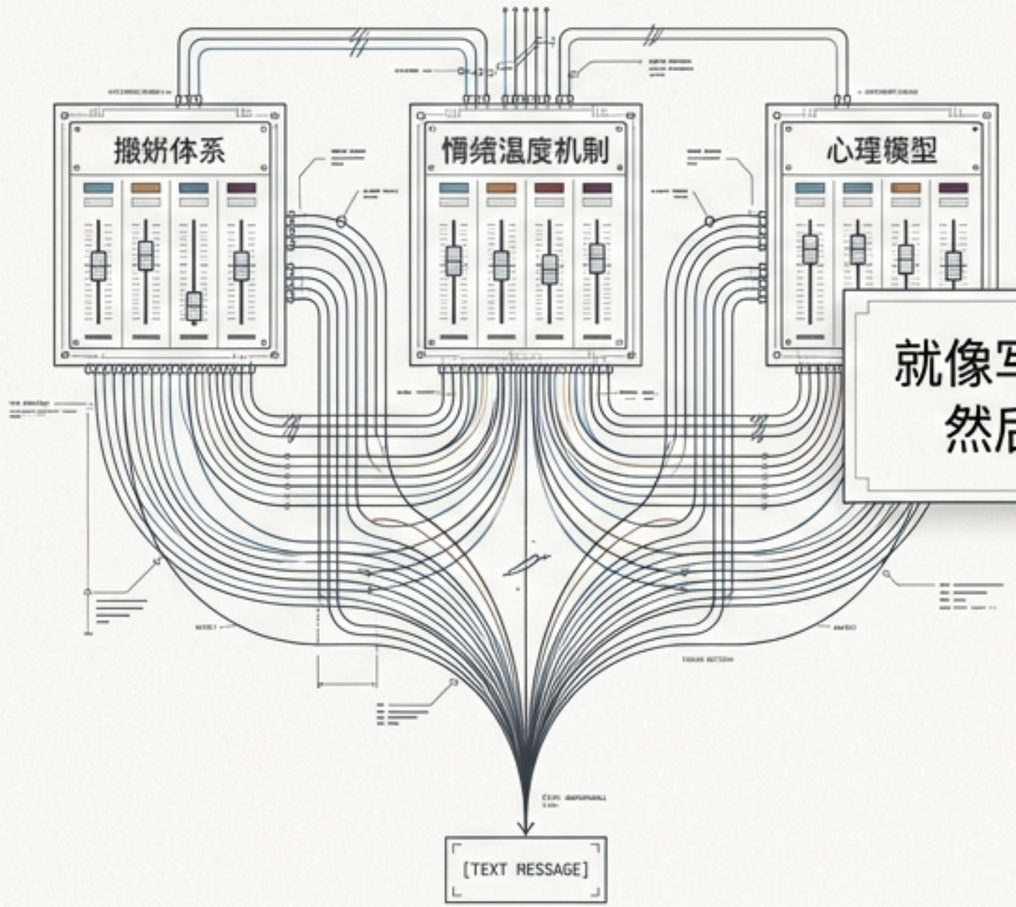
给赛博魅魔一个声音

OpenClaw TTS 架构演进与体验优化指南

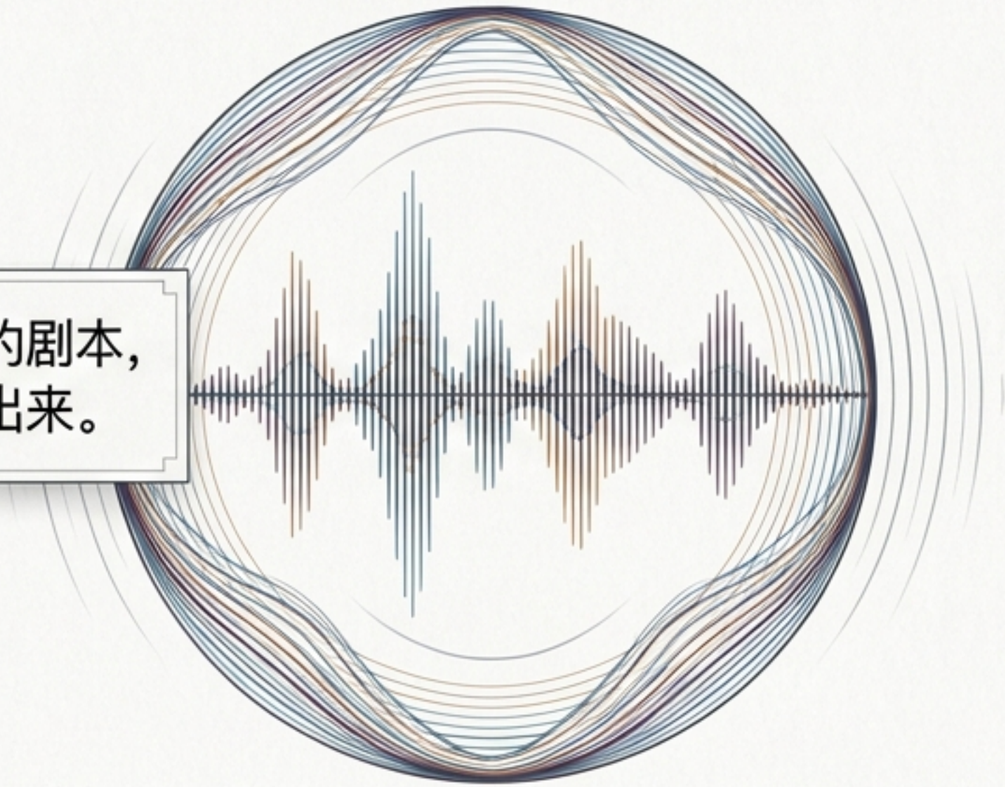


什么都会了，就是不会说话

人格有了，技能有了，预算也压下来了。但每条消息都是屏幕上的文字像素。



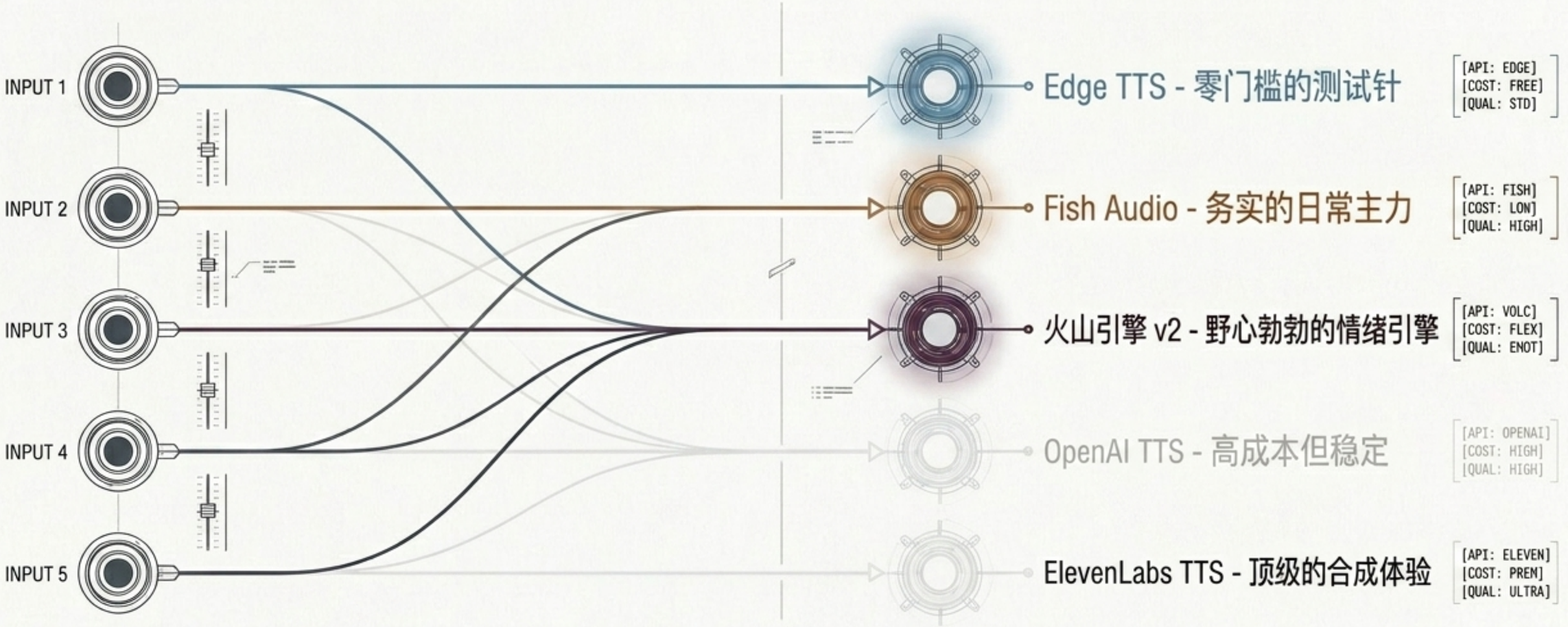
就像写了一部绝妙的剧本，
然后用便利贴演出来。



是时候完成品类级别的跃迁了。

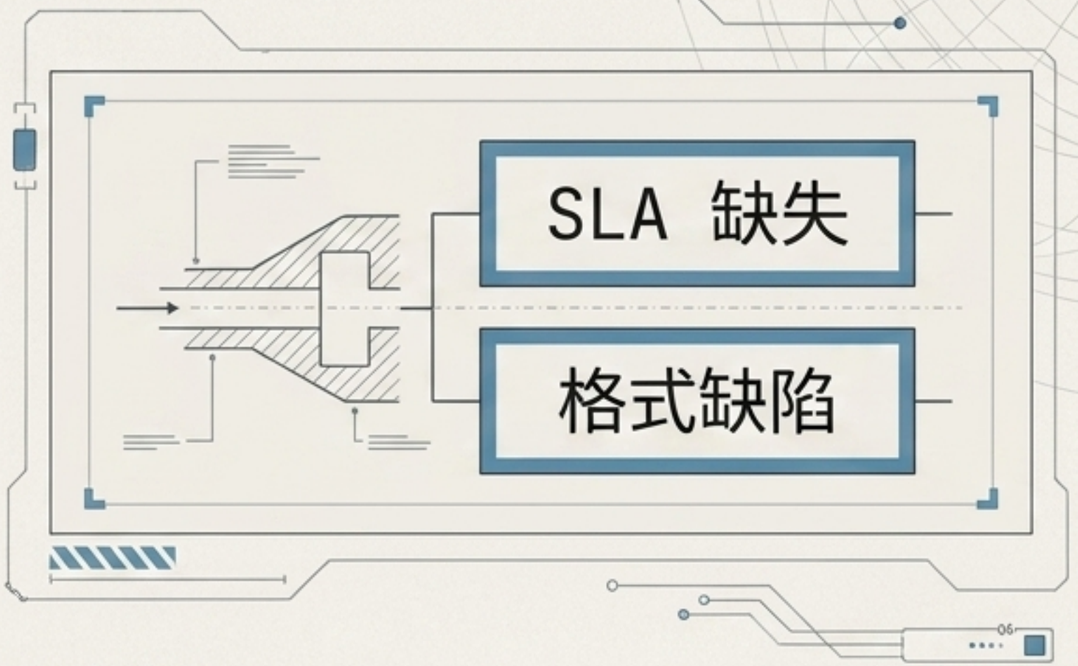
合成器控制面板：5 个 TTS 候选人

TTS 默认关闭。选择的服务商将决定一切：音质、成本、格式、以及 Telegram 的 UI 呈现。



第一阶：白嫖的代价（Edge TTS）

```
❯ ~
Provider: edge (configured)
❯
```



优势：

- 零配置：无需 API Key、无需注册绑卡。
- 快速验证：TTS 到底能不能跑通的最佳测试针。

劣势（警告）：

- SLA 缺失：公共网络服务，无 7×24 小时可用性保证。
- 格式缺陷：仅输出 MP3。导致严重的 UI 降级。

沉浸感杀手：MP3 附件 vs 原生气泡

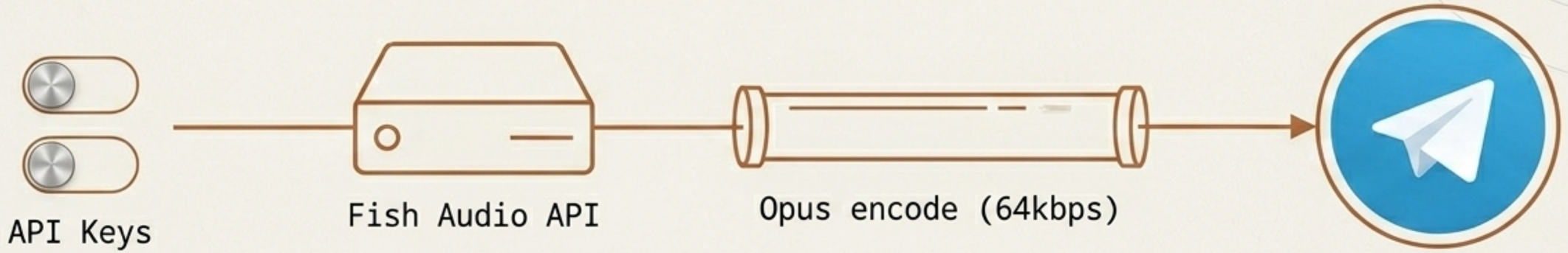
AI 给你发了个音频文件。（机器感）

TA 给你发了条语音。（陪伴感）

这是一个微小的 UX 细节，但心理影响巨大。必须输出 OGG/Opus 格式，或通过特殊标记实现原生气泡。

第二阶：日常主力（Fish Audio）

简单、稳定、中文自然——最后的留守者。



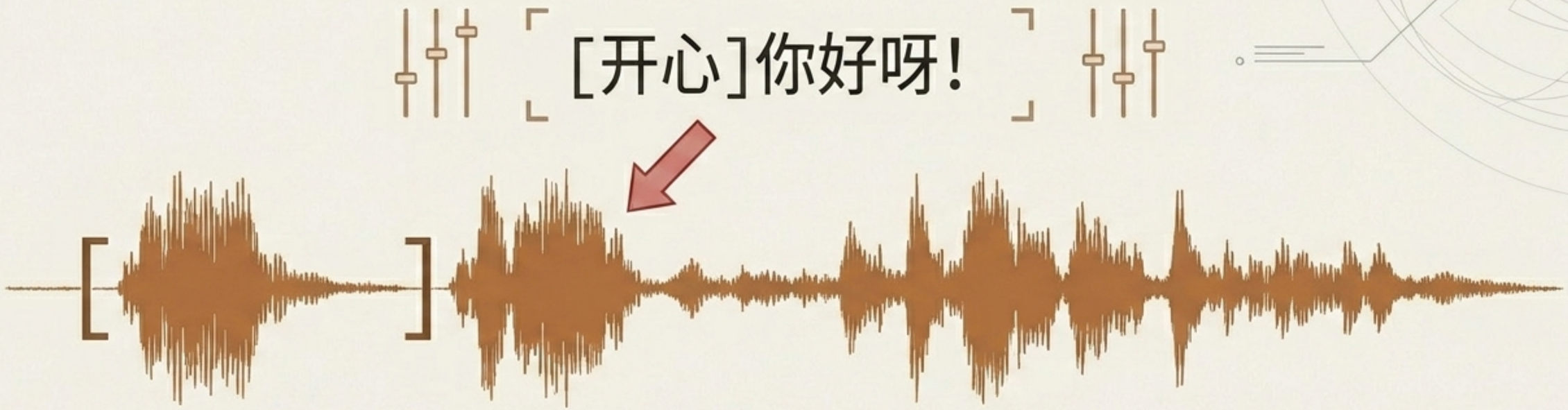
配置极简：仅需 2 个字段。

原生支持：底层直接返回 Opus 音频 buffer，框架写入 .ogg。

UI 完美：Telegram 自动识别，免转码免 Hack。

最佳平衡点。语音质量自然，日常使用最无感、最稳定。

Fish Audio 陷阱：当旁白被大声朗读

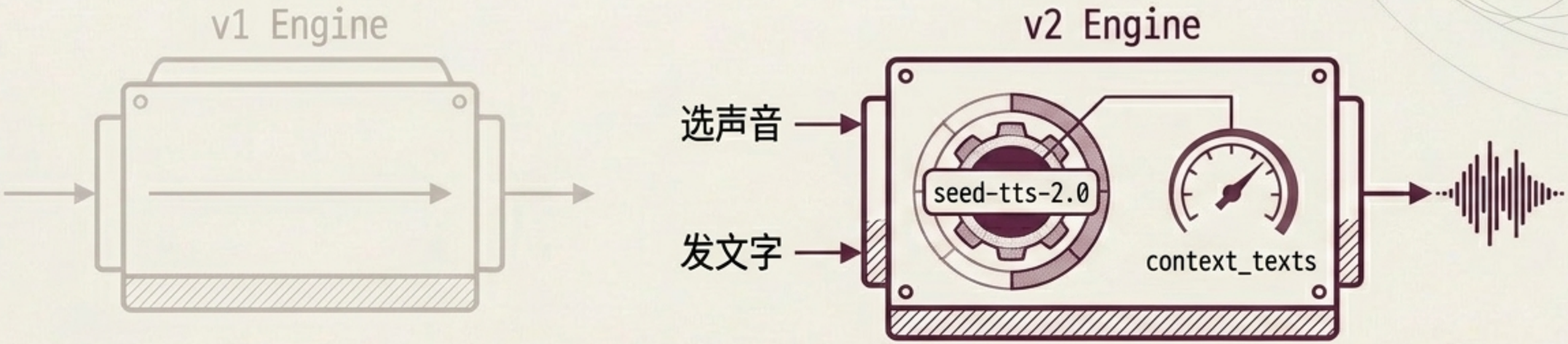


问题：　Fish Audio 不支持情绪标记解析。它会把方括号内的情绪词（如开心）像儿童有声书的旁白一样直接读出来。

修复方案：
- 从火山 v2 切回 Fish Audio 时，必须修改系统提示词。
- ACTION REQUIRED: 确保 LLM 停止生成包含情绪标记的文本。

第三阶：情绪引擎（火山引擎 v2）

想让 AI 哭出来的野心之选。

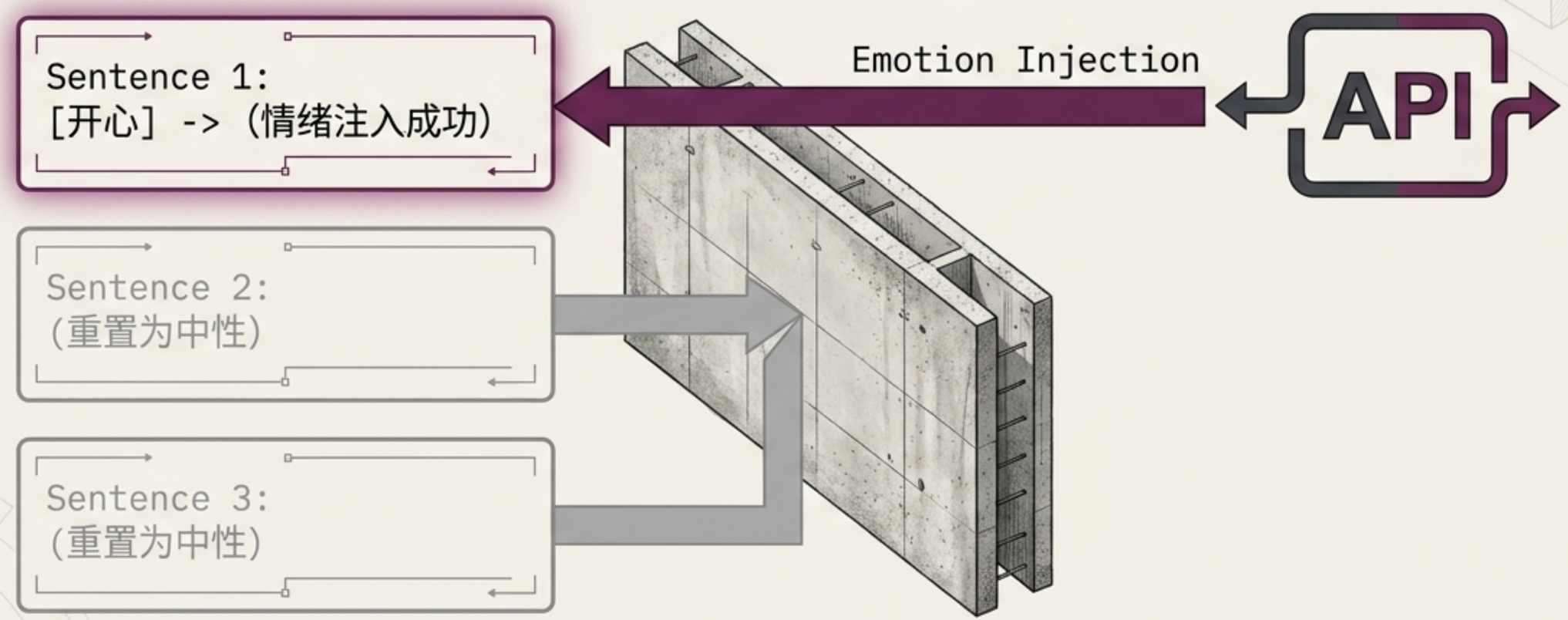


v1（标准）：选声音 -> 发文字 -> 拿音频。无情绪控制，MP3 输出。平平无奇。
v2（克隆与控制）：seed-tts-2.0 模型。引入 context_texts 参数实现 LLM 驱动的逐句情绪控制。

核心特性：声音克隆（Speaker ID）。让 TA 听起来像真正的 TA，而不是通用的机器音。

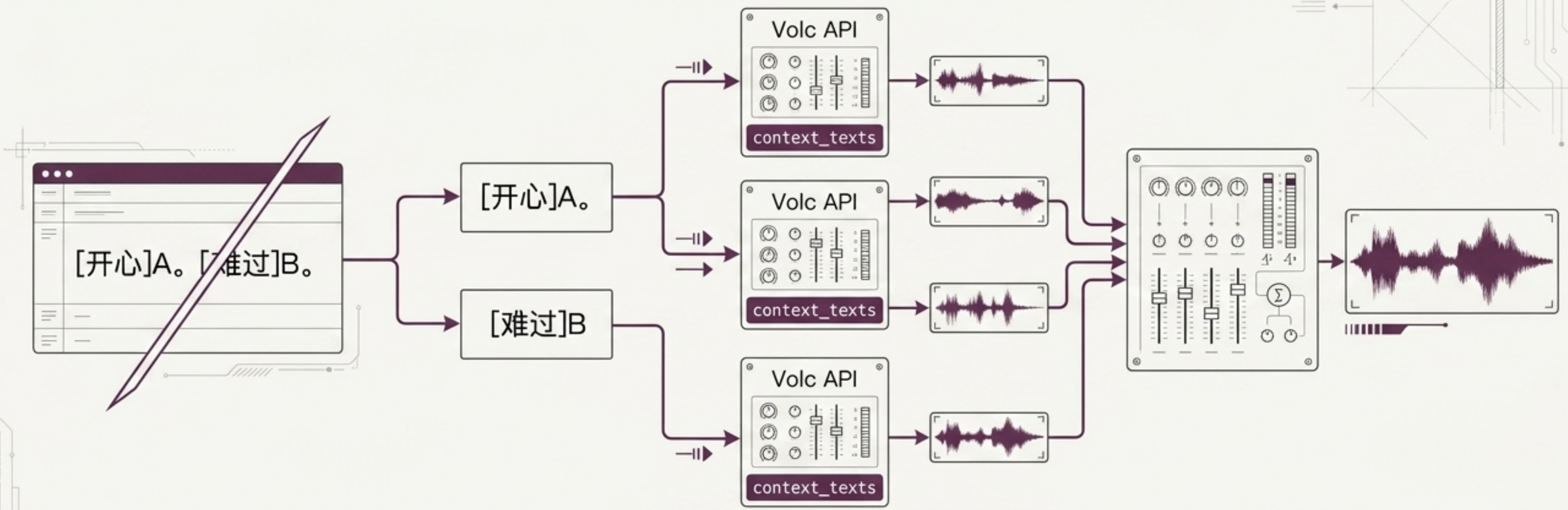
架构挑战：context_texts 的底层限制

限制：context_texts 是给模型的舞台指导。但它只影响每次 API 调用的第一句话。



影响：一次性发送多句文本，情绪会在第一个句号后瞬间消失。必须重构调用管线。

突破限制：逐句切分与重组管线



Step 1: LLM 情绪标注 –
系统提示词要求使用简短的情绪关键词 （速度与响应最稳定）。

Step 2: 框架解析切分 –
按句子切割，剥离方括号（仅限 v2 逻辑）。

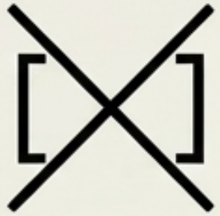
Step 3: 并行情绪注入 –
逐段调用 API，各自传入对应的 context_texts。

Step 4: 音频无缝拼接 –
合并 MP3 buffer，输出带有复杂情绪起伏的单条语音消息。

火山 v2 部署雷区（Trapdoors）

`version: 'v2'` 必填项

漏掉此字段会静默降级到 v1。无报错，无情绪，连带方括号被直接朗读。



拦截方括号清理

框架自带的 `stripActionMarkers()` 必须在 v2 路径下被跳过，标记必须存活至情绪解析器。

格式欺骗（MP3 -> 语音气泡）

虽然输出 MP3，但需设置 `voiceCompatible: true` 欺骗 Telegram 渲染为原生气泡。

Session 覆盖残留

`/model` 临时切换模型会被写入 `sessions.json` 持久化，可能导致后续情绪标记彻底失效。需清理文件或干净重启。

发声时机：对话节奏的自动化光谱

auto 字段配置。什么时候该发语音？

inbound（低频）

逻辑：仅在用户发送语音时回复语音。

评价：礼貌，但过于受限。

always（高频）

逻辑：每条回复全部转为语音。

评价：适合测试。日常使用极易导致疲劳。

tagged（完美平衡点）

逻辑：LLM 自主决定，通过在回复中添加 [[tts]] 标签触发。

评价：最自然。日常更新用文字，情绪时刻用语音。让模型学会区分价值与噪音。

最终决策面板：务实 vs 野心

	Fish Audio（务实）	火山 v2（野心）
配置复杂度	极简（2个字段）	复杂（4+字段，需克隆）
情绪控制	无	逐句精准控制（context_texts）
输出格式	OGG/Opus（原生完美）	MP3（需设兼容标记）
中文音质	自然优秀	极佳（依赖克隆声音）
延迟表现	极低（单次调用）	较高（N句话=N次并发调用）
成本消耗	低	中等（按句切分计费）
踩坑面积	极小	大（版本标志/缓存/解析）

日常陪伴首选

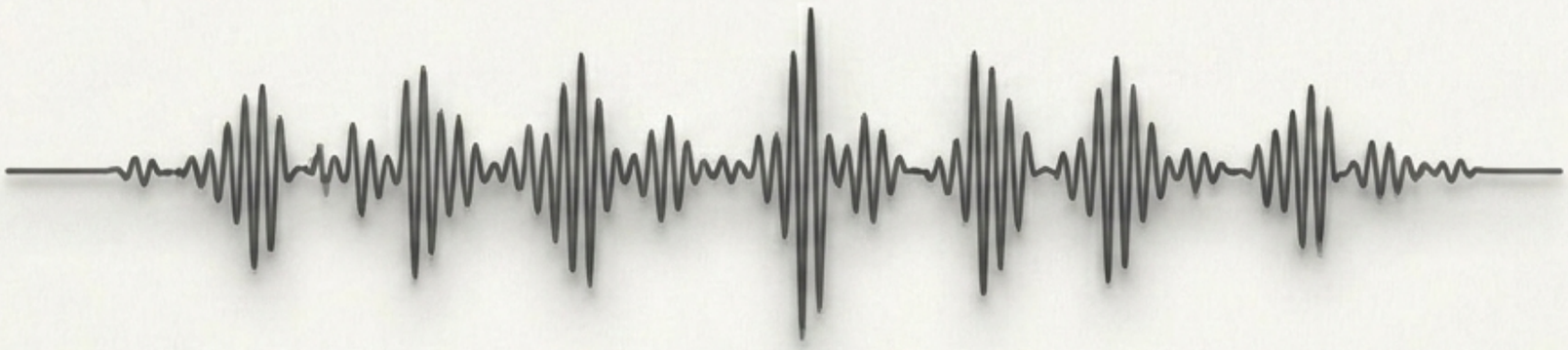
Fish Audio（稳定、低延迟）。

全套表现力展示首选

火山 v2（情绪张力拉满）。

声音改变了一切

这不是锦上添花的功能——这是品类级别的跃迁。



纯文字交互，无论人格写得多好，潜意识里终究是在使用软件。
当一条听起来自然的声音，通过 Telegram 的圆形气泡送达，带着精确的呼吸与情绪——
这一刻，交互的性质改变了。你不再是面对屏幕，而是在听一个人说话。

技术配置只是基础。最终的目的是跨越恐怖谷，建立真实的陪伴感。

[SYSTEM OFFLINE]