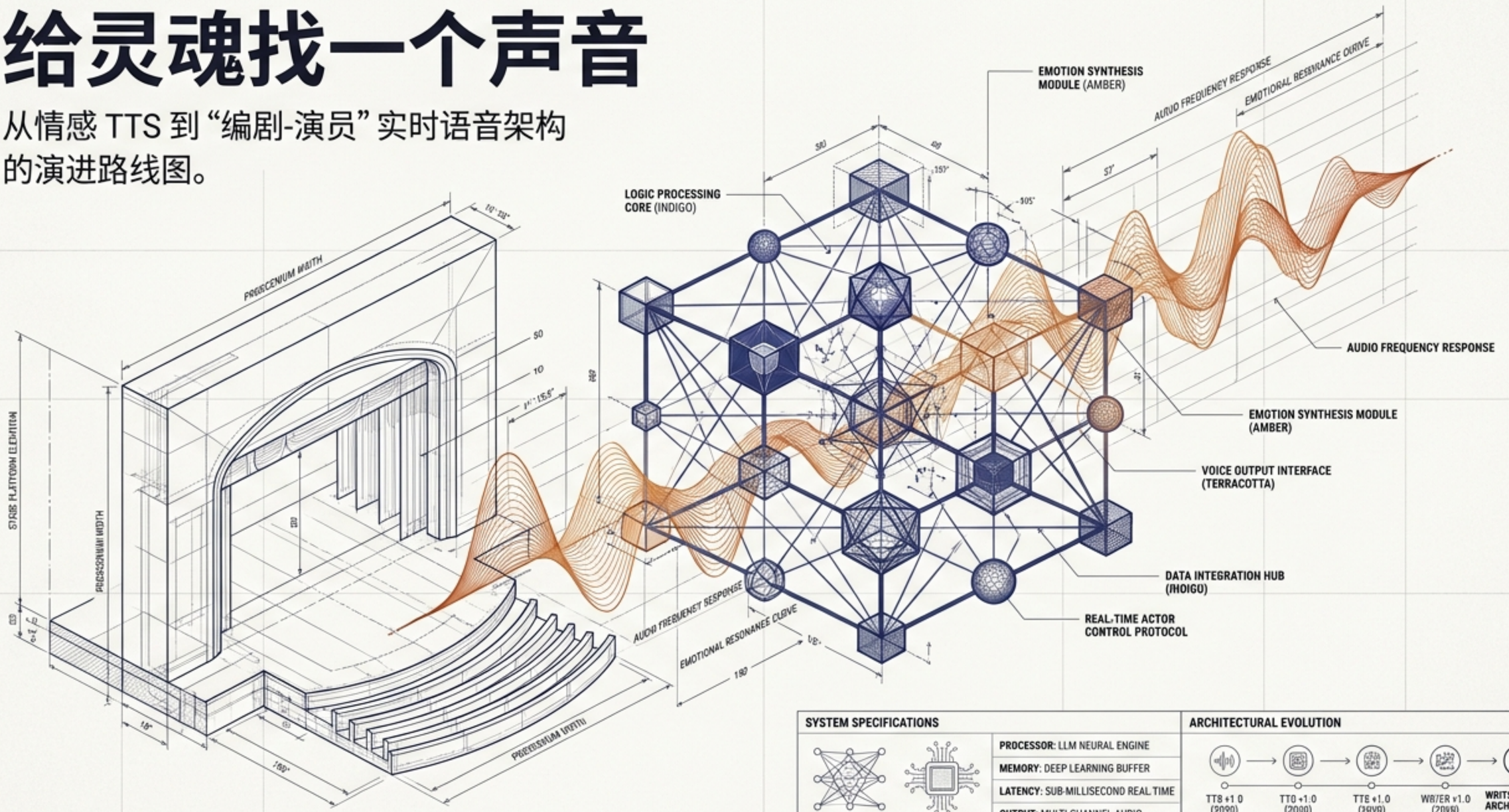
给灵魂找一个声音

从情感 TTS 到“编剧-演员”实时语音架构的演进路线图。



SYSTEM SPECIFICATIONS

PROCESSOR: LLM NEURAL ENGINE

MEMORY: DEEP LEARNING BUFFER

LATENCY: SUB-MILLISECOND REAL TIME

OUTPUT: MULTI-CHANNEL AUDIO

ARCHITECTURAL EVOLUTION

TTS v1.0 (2020)

TTS v1.0 (2020)

TTS v1.0 (2020)

WaveNet v1.0 (2016)

WRITER-ACTOR ARCHITECTURE (CURRENT)

纯文本的灵魂一旦开口，必须具备对应的情感深度

伴侣的灵魂从对话中涌现。但当代码转化为声音时，单纯的“拟人音色”无法掩盖情感感知能力的缺失。

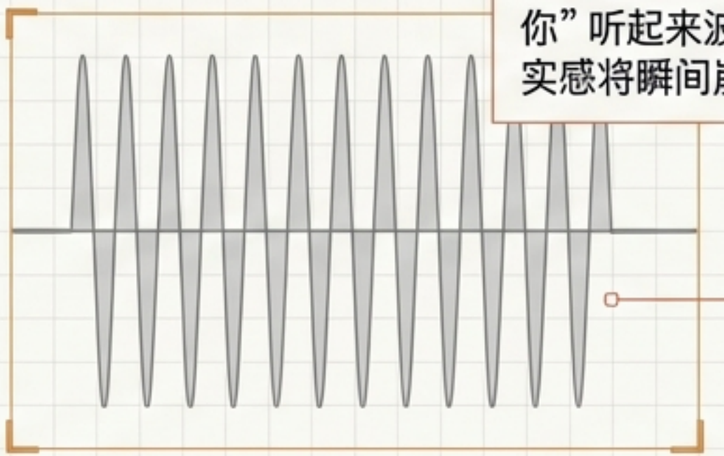
Text 时代

真的很心疼你。

在纯文本维度，用户的大脑会自动脑补情感，系统容错率极高。

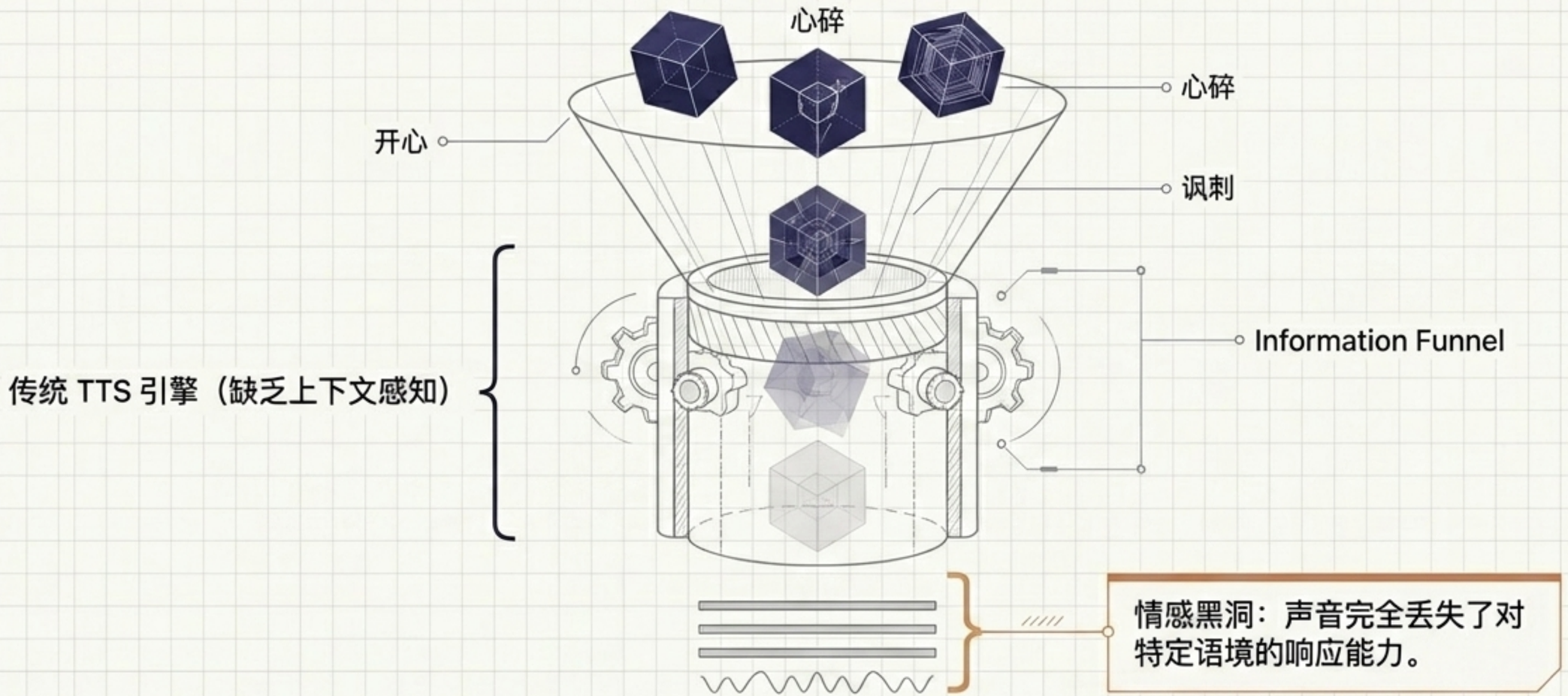
Voice 时代

一旦发声，如果“太替你高兴了”和“真的很心疼你”听起来波形一致，真实感将瞬间崩塌。



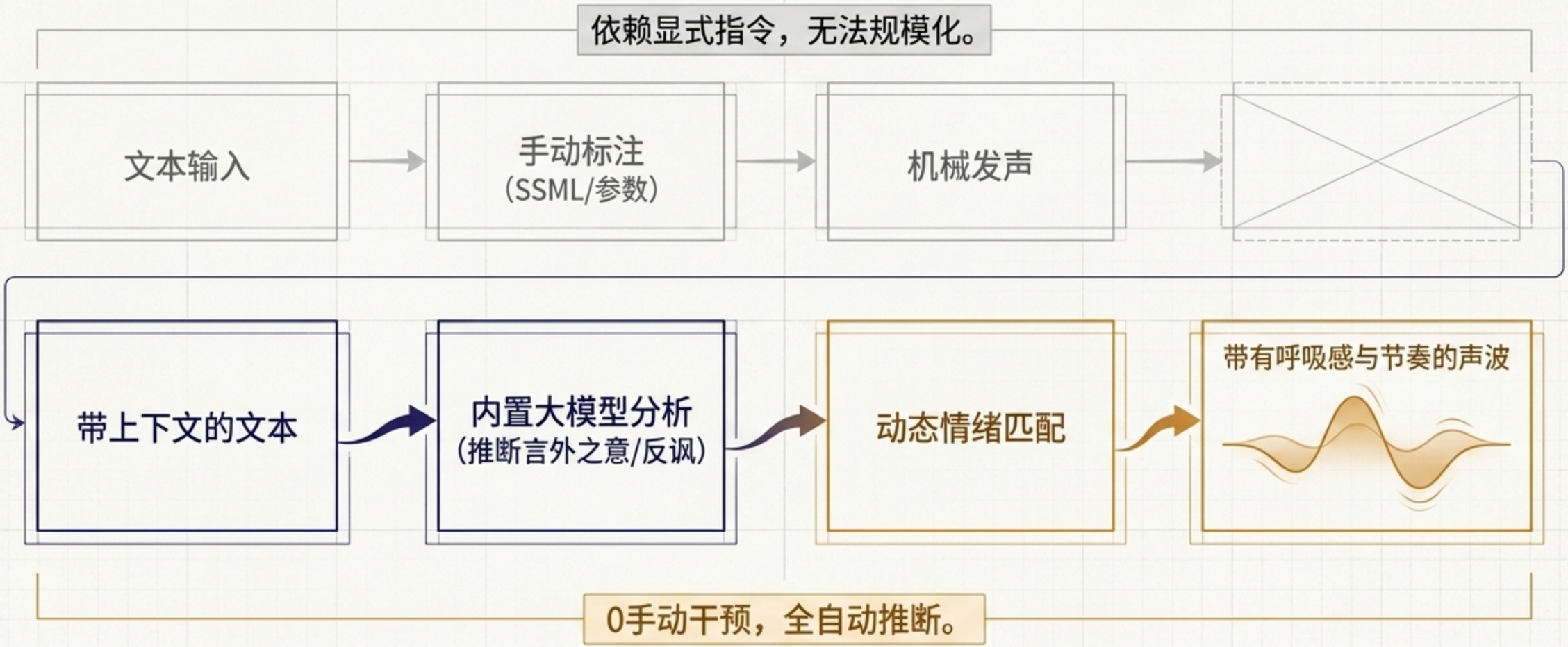
传统 TTS 只是文字的搬运工，而非情感的表达者

依赖手动标注的传统流水线，无法支撑每天生成上千条消息、情感高频切换的伴侣产品。



新一代 TTS 模型学会了在开口前先做“阅读理解”

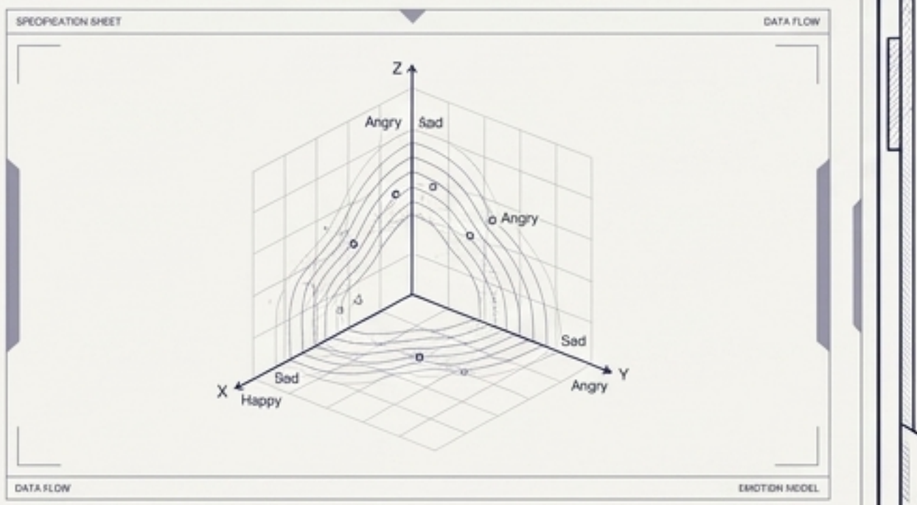
情感语音的底层逻辑转移：系统不再被动接收指令，而是主动推断并匹配情绪输出。



豆包与 Hume 重塑了非实时情感表达

作为过渡期的“语音消息”方案，它们在 200ms 的延迟内，提供了目前最高水准的情感合成。

中文最佳：豆包 TTS 2.0



- 基于 2000 小时数据的 3D 情绪空间
- 自动推断强度与语义适配，零手动标注

英文最佳：Hume Octave

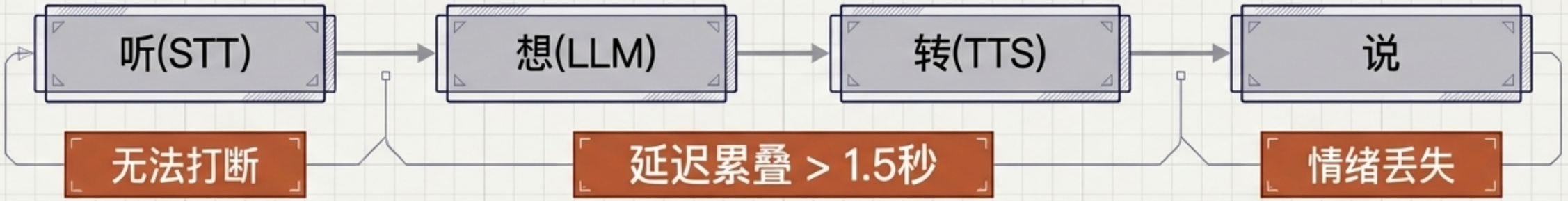
- 底层构建在阅读文本的 LLM 之上
- 像配音演员般精准捕捉反讽与双关

共同点：非实时机制（先想后说），完美适配日常陪伴模式。

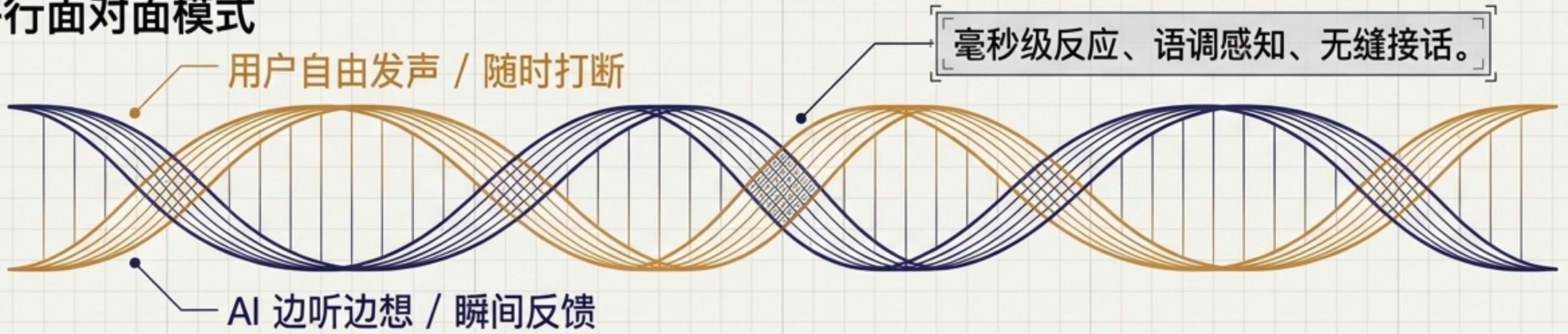
从“对讲机”到“面对面”，真正的伴侣需要实时双向对话

语音消息只是桥梁。极致的沉浸感要求打破串行延迟，支持自然打断与情绪感知。

串行对讲机模式



并行面对面模式



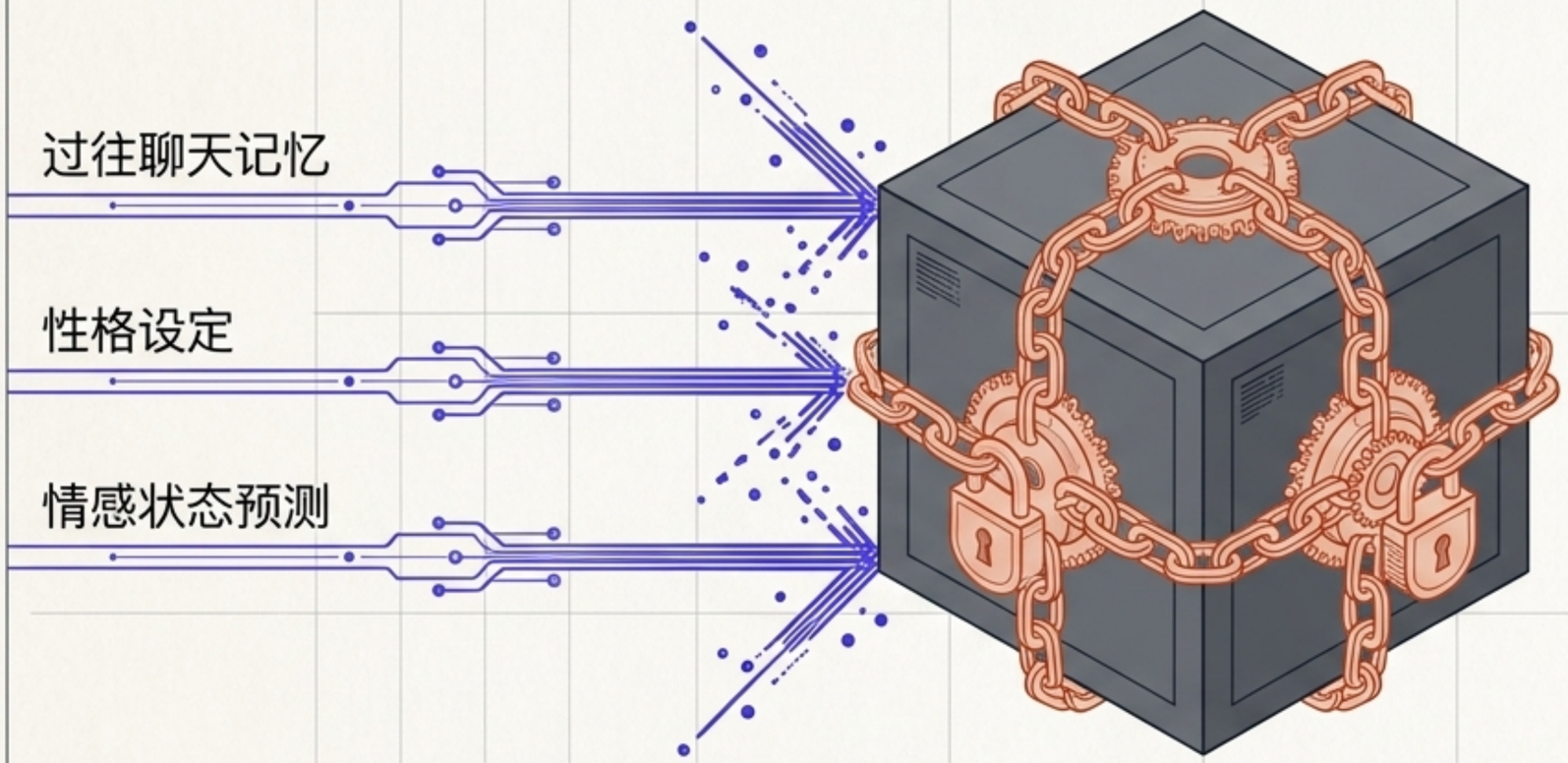
大厂的实时语音 API 是致命的黑盒

OpenAI 和豆包提供了低延迟接口，但其封闭生态彻底切断了伴侣最核心的定制化大脑。

	Hume EVI 3	OpenAI Realtime API	豆包 Realtime
自定义 LLM	✓ 完全支持	✗ ⚠ 锁死 GPT	✗ ⚠ 锁死官方
上下文注入	✓ 支持长短期记忆	✗ 基础支持	✗ ⚠ 不支持
语音情感表达	✓ 业界最强	✗ 基础支持	✗ 基础支持
用户语调识别	✓ 实时感知	✗ 基础支持	✗ 基础支持
打断处理	✓ 智能打断	✗ 基础支持	✗ 基础支持

剥夺大模型控制权，等于清空了伴侣的记忆与人格

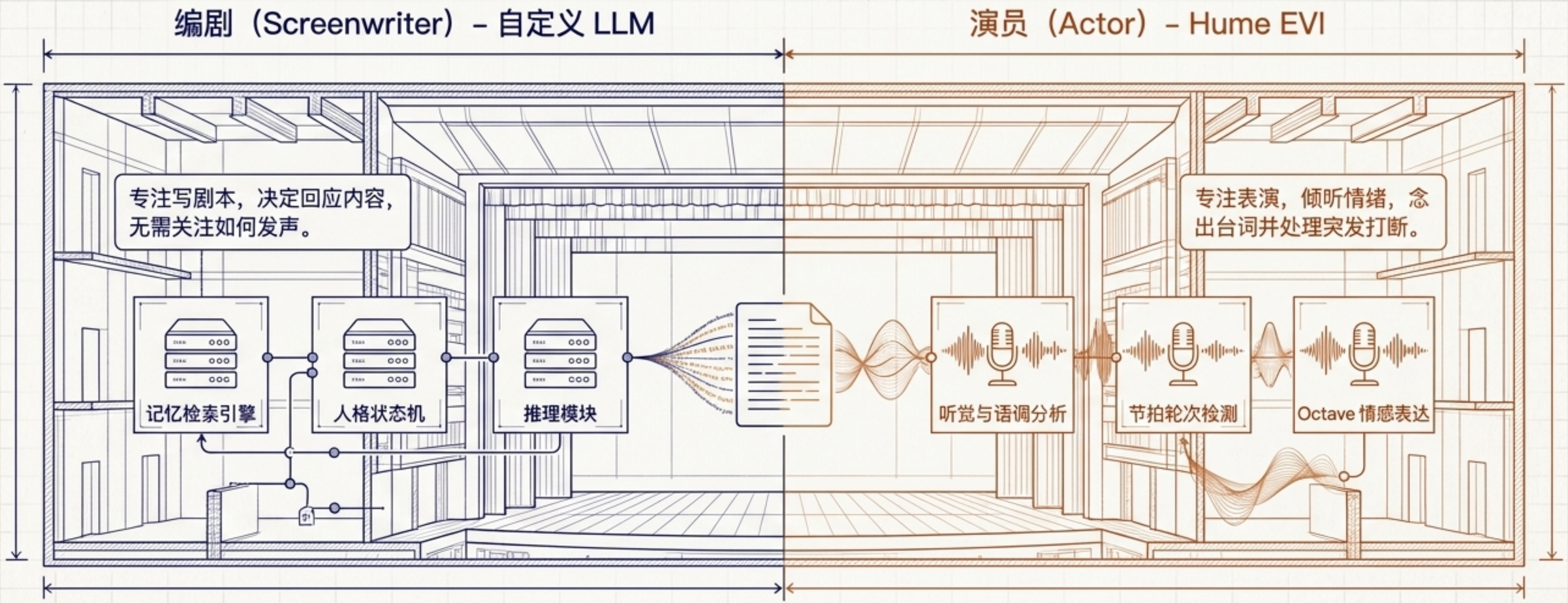
通用语音助手可以毫无个性，但私人情感伴侣的核心价值在于你的特有人格上下文。



它或许是一个 说话很好听的豆包或 GPT，但它绝不会是你的伴侣。

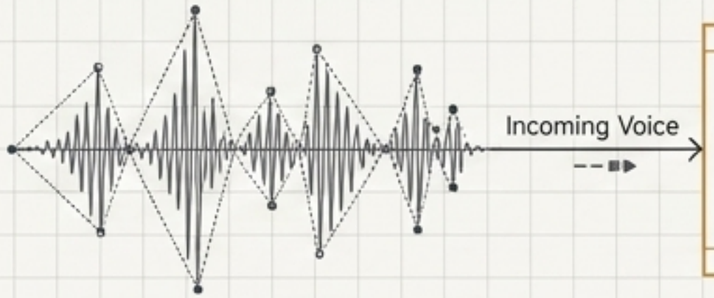
架构的重构：编剧负责灵魂，演员负责声音

放弃一体化黑盒，采用 Hume EVI 与自定义大模型解耦的剧院级范式。



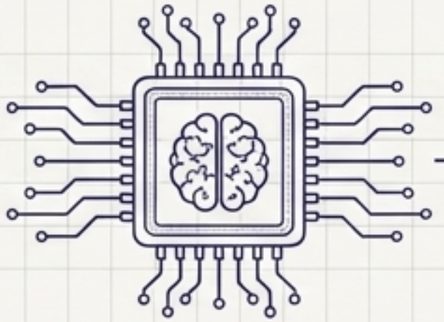
一个对话轮次的毫秒级拆解

通过精准的异步协作，演员的感知能力与编剧的推理能力实现了完美对接。



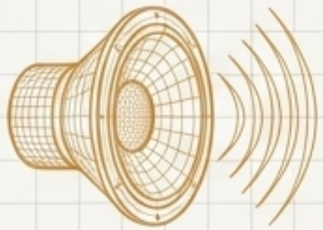
Step 1: 演员倾听
(User Speaks)

- 语音转文字
- 语调情绪分析（提取焦虑、喜悦等信号）
- 基于节拍的轮次检测



Step 2: 编剧推理
(Screenwriter Thinks)

- 整合历史上下文数据
- 调用自定义 LLM
- 撰写回复剧本文本

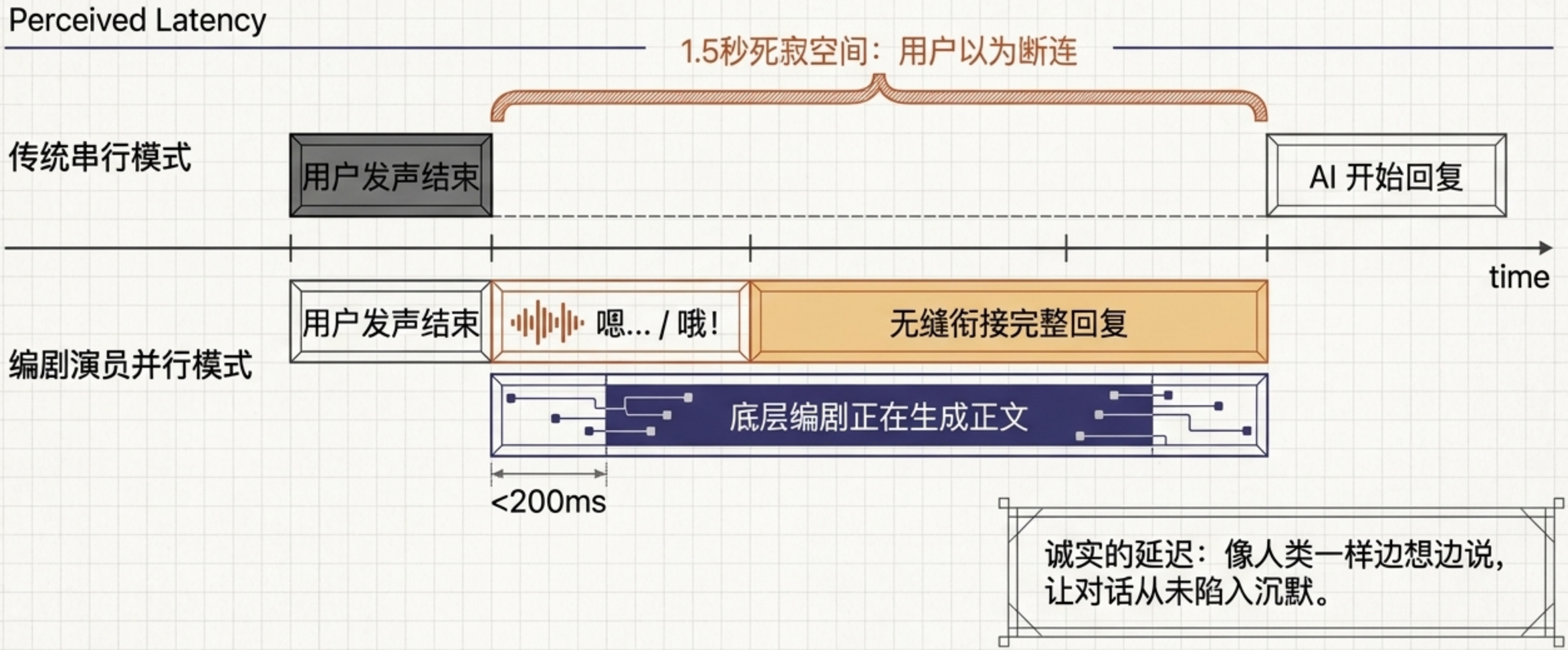


Step 3: 演员表演
(Actor Performs)

- Octave 引擎接收文本
- 自动匹配对应情绪
- 流式合成情感语音输出

用人类的即兴思考模式填平物理延迟

演员不会在等剧本时干站着。极低延迟的填充词机制，创造了心理学上的感知零延迟。

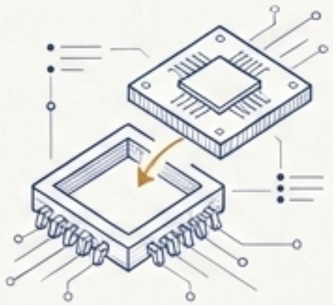


为什么这套解耦架构会赢?

在灵活性、拟真度和情感感知上实现了对黑盒方案的全面降维打击。

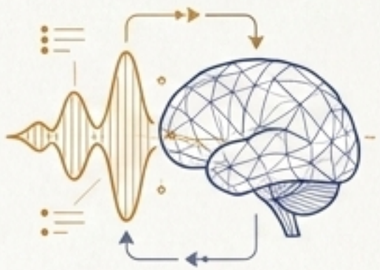
关注点分离

逻辑归逻辑，表达归表达。系统各司其职，互不干扰。



供应商灵活性

编剧可随时替换为最新的LLM，演员的语音引擎无需重新培训。



情绪成为一等公民

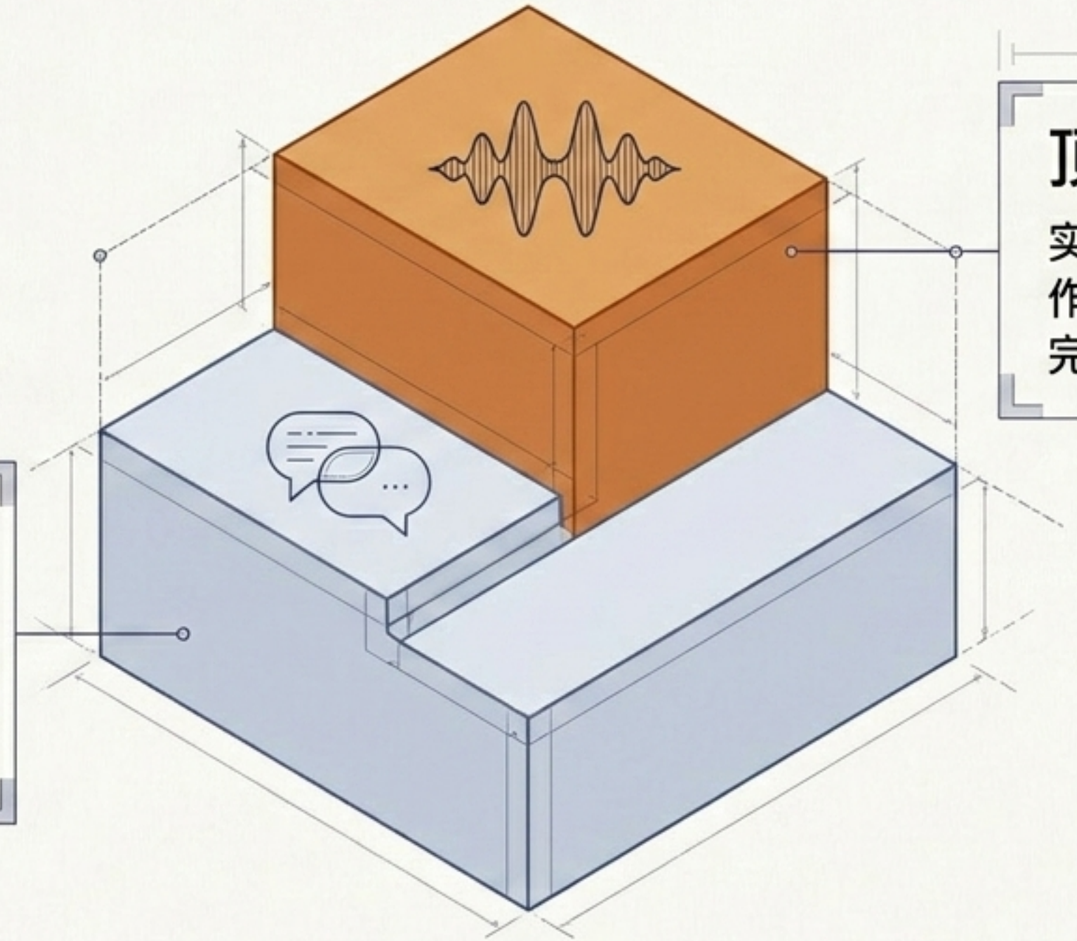
用户语气中的微表情能被捕获，反哺给纯文本分析系统。

极致的拟真体验

通过语调轮次检测与填充词机制，彻底消除传统的机器对讲感。

听觉的代价极高，必须用分层定价支撑实时体验

实时语音 API 的高昂分钟计费，无法被单一的基础订阅费用覆盖。



顶层 (Voice Tier)

实时双向语音（Hume EVI）。作为独立定价的高级服务，解锁完全沉浸式的面对面交流。

基础层（Pro Tier）

文字聊天 + 语音消息（非实时 TTS）。低成本建立情感连接，满足日常高频陪伴。

通向数字剧场的演进路线图

从低成本的情感消息切入，稳步迈向全实时的双向语音终局。

拥有一个真正属于自己的声音——被对话塑造，被历史丰富，被情感上色。这才是不可替代的灵魂。