

2025 对话式 AI 发展白皮书



对话式 AI 实践指南

涵盖技术、产品、
生态、场景



对话式 AI 体验

质量评估方法
“三维二轨”行业首发



一图看懂

热门对话式 AI
应用场景分布



对话式 AI 发展白皮书

前言

对话式 AI 开启人机对话的新纪元

现实中人与人的沟通以语音为主，视觉其次，视觉的重要性在于信息的丰富度，信息浓度和沟通效率主要靠语音，而在生成式 AI 的浪潮下，人与 AI 的交互中亦是如此。从图灵测试到生成式 AI 爆发，人类对“机器对话”的探索已跨越半个世纪。对话式 AI 正以颠覆性姿态推开新纪元的大门——它不再是机械式被动应答的工具，而是具备感知、理解与创造能力的交互主体。这一变革的核心动力，源于多项核心技术的不断演进。例如，自然语言理解的质变，大模型突破语义泛化瓶颈；实时互动技术的成熟，低延迟交互让对话更自然、流畅；多模态感知的融合，语音、视觉、情境的协同决策等等。

2024 年 5 月，GPT-4o 的发布开创了 AI 实时语音交互的先河，推动了 AI 语音交互进入对话延迟更低、互动更流畅的 RTC 方案。同时，在 10 月份，OpenAI 发布了与 Agora（声网兄弟公司）、Livekit、Twilio 共建的 RealtimeAPI 公开测试版，用于构建基于 GPT-4o 语音到语音的 AI 应用和智能体，所有付费的开发者都能在应用程序中构建低延迟、多模态的实时互动体验。

当前对话式 AI 的应用有两个技术路线，分别是端到端模型方案和级联模型方案，两者在性能、成本、灵活性、扩展性以及集成部署等方面各有优缺点。同时，对话式 AI 的交互体验伴随着技术升级和应用扩展正在快速发展，对于开发者与 AI 创业者而言，如何选择最适合自身业务的技术方案与产品供应商显得至关重要。

对此，声网研究院联合 RTE 开发者社区推出《对话式 AI 发展白皮书》，基于对行业的洞察、调研，并结合自身的业务经验，从对话式 AI 发展的背景、技术方案与产品生态、对话体验质量评估方法、应用实践落地等多个维度系统的梳理对话式 AI 的发展现状与未来的趋势方向，希望能为行业的从业者带来更多的帮助。

目录

第一章：对话式 AI 发展的背景..... 5

第二章：对话式 AI 核心技术 7

 2.1 对话式 AI 的发展..... 7

 2.2 对话式 AI 的主要技术方向 7

 2.3 级联对话式 AI 的基本原理 8

 2.4 响应延迟..... 9

 2.5 智能打断..... 11

 2.5.1 打断机制的核心类型..... 11

 2.5.2 避免意外的打断..... 12

 2.6 选择性注意力锁定 13

 2.6.1 技术特点..... 13

 2.6.2 场景应用..... 14

 2.7 对话上下文管理..... 14

 2.7.1 为什么上下文很重要..... 14

 2.7.2 如何做好上下文? 15

 2.8 视觉理解..... 16

 2.8.1 实时视频解析能力：动态场景的“感知-决策”闭环 17

 2.8.2 静态图像分析能力：从物体识别到逻辑推理..... 17

 2.8.3 对话式 AI 的视觉理解进化方向..... 18

 2.9 音频处理..... 18

 2.9.1 麦克风和自动增益控制 18

 2.7.2 回声消除（AEC） 19

 2.9.3 噪声抑制（ANS） 19

 2.10 网络传输 21

 2.10.1 WebSocket 与 WebRTC 21

 2.10.2 实时传输网络 22

第三章：对话式 AI 的主流方案及产品生态..... 25

 3.1 技术架构与方案分类..... 25

 3.1.1 串行方案..... 25

 3.1.2 多模态方案..... 26

3.2 如何快速构建对话式 AI 应用.....	27
3.2.1 基于对话式 AI 引擎方案	27
3.2.2 基于 Linux SDK 构建方案	28
3.2.3 基于 TEN 开源框架自研方案.....	28
3.3 对话式 AI 技术生态	29
3.4 对话式 AI 产业生态	30
第四章：对话式 AI 体验质量评估方法.....	32
4.1 性能指标体系：三维二轨.....	32
4.1.1 详解“三维二轨”理解能力：语义理解准确率.....	35
4.1.2 详解“三维二轨”表达能力：对话连贯性与流畅度.....	37
4.1.3 详解“三维二轨”交互能力：响应时间与实时性	39
4.1.4 落地“三维二轨”：应用框架到你的对话式 AI 产品	41
4.1.5 场景示例：智能点餐与老年人陪伴	41
4.2 用户体验要素.....	45
4.2.1 交互友好性	45
4.2.2 环境适应性	45
4.2.3 设备适应性	46
4.2.4 可靠性与稳定性	46
4.3 AI 模型评测平台（对话式）	46
4.3.1 “仪表盘”：综合最优、响应最快模型一目了然	47
4.3.2 “竞技场”：自主选择模型性能对比	48
第五章：对话式 AI 的应用实践.....	49
5.1 对话式 AI 场景热力榜单.....	50
5.2 社交领域的 AI 场景探索如火如荼	51
5.2.1 从机械应答到“类人化”社交引擎.....	51
5.2.2 星野案例分析：情感陪伴与 UGC 生态的深度融合	52
5.3 AI 陪伴硬件迎来风口.....	53
5.3.1 实现真正的情感陪伴，交互体验是关键.....	54
5.3.2 AI 陪伴硬件，不止于儿童.....	54
5.3.3 案例分享：珞博智能 AI 宠物“芙崽 Fuzozo”	56
5.4 百镜大战开启，下一代个人智能设备入口?.....	57
5.5 对话式 AI 正从底层重塑教育生态	59

5.6 智能助手成为普及最广泛的 AI 应用	62
5.6.1 技术驱动：多模态与大模型赋能智能助手“泛在化”	62
5.6.2 功能革新：从“问答”到“赋能”，重塑工作与生活体验	63
5.6.3 未来更具人格化与情感智能	63
5.7 AI 客服已超越“成本削减工具”，成为企业服务竞争力核心组件	64
5.7.1 价值升维：从“降本增效”到“体验重塑”与“价值创造”	64
5.7.2 技术融合：多模态、大模型与业务系统的深度集成	65
5.7.3 行业实践：从“标准化服务”到“场景化创新”	65
第六章：对话式 AI 的未来展望	67
6.1 对话式 AI 多模态与情感能力将达到类人水平	67
6.2 科幻走进现实，基于对话式 AI 能力的 AI 超级助手将出现	69
6.2.1 个人助手：从“工具”到“终身伙伴”	69
6.2.2 企业超级助手：组织智能体	71
6.3 对话式 AI 对人类社会产生深远影响	72
6.3.1 改变人机交互方式	72
6.3.2 提升效率与生活质量	73
6.3.3 推动行业创新与经济增长	74
6.3.4 促进知识传播与信息获取	75

第一章：对话式 AI 发展的背景

人们对更自然、便捷的人机交互方式的追求，是对话式 AI 发展的核心驱动力。在传统的交互模式中，无论是命令行界面、图形界面还是触摸操作，都需要用户主动适应机器的操作逻辑，存在一定的学习成本。相比之下，语音交互凭借其表达灵活、信息丰富、易于理解的优势，成为最符合人类本能的交互方式。这种优势在特定场景中尤为突出：在智能车载系统、智能家居设备等无键盘环境中，语音成为最直接的操作方式；在移动场景下，语音交互也能解放用户的双手。这些多样化需求共同推动了传统对话式 AI 的初期发展。

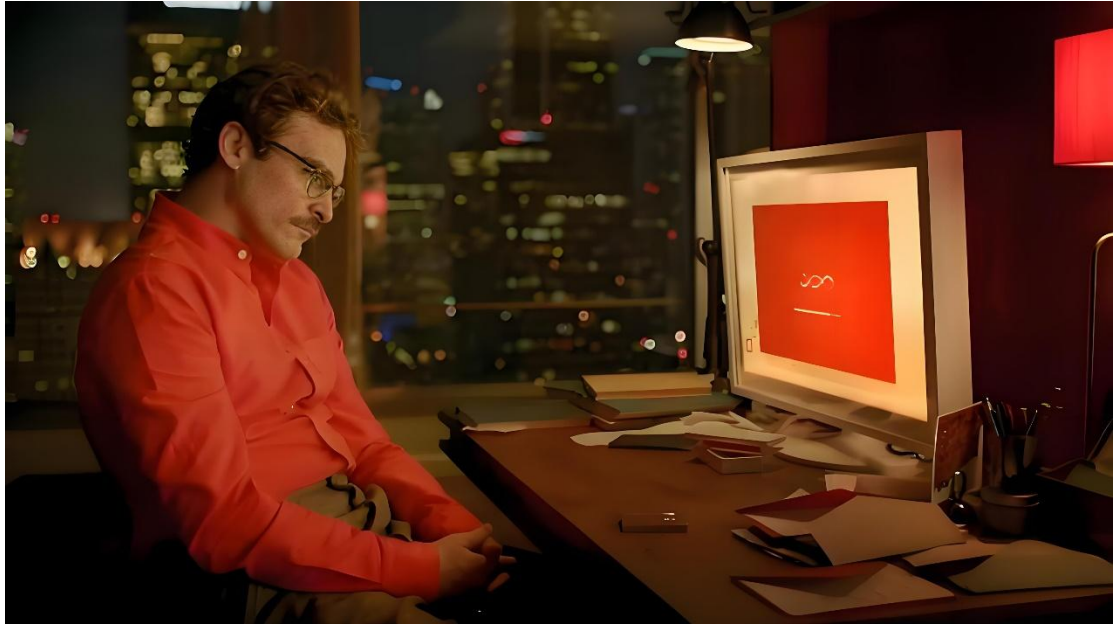
然而，在大语言模型出现之前，传统对话系统的能力存在明显局限。其技术架构主要基于语音识别（ASR）和自然语言理解（NLU）模块，配合预设的对话流程和规则引擎。这种架构虽然能够处理简单的任务型对话，如设备控制、信息查询等，但在理解复杂语义、维持多轮对话、处理开放域话题等方面表现欠佳。对于需要深度语言理解和生成能力的场景——如情感陪伴、个性化教学、复杂客服咨询等——传统系统往往难以提供令人满意的体验。

ChatGPT 等大语言模型的突破性进展，从根本上改变了对话式 AI 的技术范式。基于海量数据预训练的大语言模型，不仅能够精准理解用户意图，还能生成流畅、连贯且符合语境的回复。这种能力使得对话系统不再局限于简单的指令响应，而是能够进行富有逻辑的推理、展现个性化的表达，并适应多样化的对话场景。生成式 AI 技术的成熟，使得对话系统在情感陪伴、教育辅导、创意协作等复杂场景中实现了质的飞跃，为对话式 AI 开启了全新的可能性。本书所探讨的“对话式 AI”，正是建立在生成式 AI 技术基础上，具备真正语言理解与交互智能的新一代解决方案。

对话式 AI 让科幻走进现实

著名科幻电影《Her》相信大家都不陌生，它为我们展现了一个充满情感深度的未来世界，其中主人公与对话式 AI 系统萨曼莎之间的关系尤为引人入胜。萨曼莎不仅具备高度智能，能够通过自然流畅的对话理解并回应主人公的情感和需求，还展现了自我学习和进化的惊人能力。她能够随着时间的推移不断成长，与主人公建立起了深厚的情感纽带。

AI 伴侣萨曼莎的实现正对应了当下生成式 AI 时代下的对话式 AI 技术，同时这部电影也生动地描绘了对话式 AI 的潜在发展方向。未来的对话式 AI 不仅能够在技术层面实现更精准的语义理解和意图识别，还可能在情感交互方面取得重大突破。它们将不再是冷冰冰的机器，而是能够理解、共情并回应人类情感的智能伙伴。这种情感交互的提升将极大地拓展对话式 AI 的应用范围，使其能够在心理健康辅导、情感陪伴、个性化教育等领域发挥更加重要的作用。



随着对话式 AI 的不断发展，人类与机器之间的界限将逐渐模糊。AI 将不再是单纯的工具，而是可能成为人类生活中的重要伴侣和协作者，给人类社会带来深远的变化。

对话式 AI 的广泛应用，首先将改变的是人类与机器的交互方式。它使人类能够以最自然、最便捷的方式——语言，与机器进行交流。这种改变极大地降低了技术使用的门槛，让更多人，无论年龄、教育背景或技术熟练程度如何，都能轻松地获取信息、使用各种智能设备和服务。只需简单的对话，用户就能与机器进行有效交互，享受科技带来的便利。

同时对话式 AI 在提升工作效率和生活质量方面的作用也尤为显著。在工作场景中，它充当智能助手角色，能够协助完成诸如日程安排、邮件处理、数据分析等任务，从而提高工作效率。对话式 AI 还促进了信息获取和知识传播的便捷性。它能够快速准确地回答用户的问题，提供丰富多样的信息和知识。这有助于人们在海量的信息中迅速找到所需内容，满足学习、工作和生活中的各种信息需求。

更为重要的是，对话式 AI 的广泛应用推动了各行业的创新和变革，催生了新的商业模式和产业生态，为经济增长和社会进步注入了新的动力。未来，在教育、医疗、客服、家居等行业，都将有望出现定制版的 AI “萨曼莎”。

第二章：对话式 AI 核心技术

2.1 对话式 AI 的发展

对话式 AI 是一种基于大语言模型（Large Language Model, LLM）、自动语音识别（Automatic Speech Recognition, ASR）、文字转语音（Text To Speech, TTS）、实时互动（Real -Time Engagement, RTE）等技术的复杂人工智能系统，可为**人类与机器提供自然、真实、流畅的类人对话能力**。对话式 AI 系统能够准确识别语音、文本和图像，以及识别语言习惯、语气和情感，并能够以适当的自然语言做出回应。

对话式 AI 的概念看似新颖，但其雏形可追溯到 60 年代中期，由 Joseph Weizenbaum 开发的聊天机器人 ELIZA1。ELIZA 是世界上第一个机器人心理治疗师，它利用 NLP 技术来匹配输入内容中的模式并生成相应的回应。从基础层面来看，如今的对话式 AI 仍然采用模式识别和预测分析的方式来运作。然而，主要的区别在于：ELIZA 的功能仅限于产品中被编程设定的脚本，且只能基于文本对话。而真正的对话式 AI 则具备自然语言对话的能力和多轮对话管理能力，能够记住对话的上下文，并能够利用大量数据持续“学习”，处理更复杂问题和灵活多样的提问方式，最终及时、准确地给出响应。

对话式 AI 的发展从 20 世纪 60 年代至今，大概经历了四个阶段：

时间阶段	代表产品	核心技术	核心特点与交互方式	典型应用场景
20 世纪 60-80 年代	ELIZA, PARRY	基于规则的模式匹配	依赖关键词匹配和预编写脚本，对话机械，无法理解上下文	学术研究，简易心理治疗模拟
20 世纪 90 年代-21 世纪初	ALICE, Jabberwacky	统计模型，AIML 语言	规则库扩大，在特定竞赛（如勒布纳奖）中表现突出，但仍有局限	在线聊天，竞赛验证
21 世纪初-2020 年代	Siri, Google Assistant, Alexa	深度学习，语音识别与合成	以语音为主要交互方式，能完成设定任务（如查询、设置提醒），但对话能力仍较初级	智能手机、智能音箱、智能车机上的个人数字助理
2020 年代后期至今	ChatGPT、Grok 4 声网对话式 AI 引擎	大型语言模型，多模态交互，上下文记忆	能进行自然流畅的多轮对话，理解复杂意图和情感，提供个性化、场景化的主动服务	情感陪伴、AI 客服、实时翻译等场景

2.2 对话式 AI 的主要技术方向

随着对话式 AI 技术的快速发展，市场上涌现出了不同的技术架构方案和产品形态。目前 Voice Agent 有两种主流架构方式，一是三段式的级联模式，二是端到端模式。其中级联

模式级联方案以其技术相对成熟、可控性较高、以及成本相对可控，成为现阶段大多数开发者和企业的首选。而端到端方案则以其潜在的更自然交互体验和更低延迟，代表着未来的发展方向，详见白皮书第三章：技术架构与方案分类。

由于级联模型在未来很长一段时间仍将是主流方案，因此本文将着重介绍级联模式的对话式 AI 技术方向及应用：

2.3 级联对话式 AI 的基本原理

对话式 AI 的核心职责是听懂人类讲话的内容和情感，并以语音对话的方式作出回应，然后重复这个循环。

当前，几乎所有的级联对话式 AI 产品都普遍采用相似的架构：对话式 AI Agent 服务部署于云端，协调端到端语音对话（Speech-to-Speech）的交互闭环，整体采用 RTC 技术实现超低延迟双全工人机对话。支持行业主流 AI 模型接入，并通过 LLM 函数调用或结构化输出连接后端系统。

整个循环流程如下：

- ✧ 用户设备麦克风捕获语音信号，编码后经实时传输网络传输至云端代理程序
- ✧ 将用户语音实时转换（ASR）为高准确率的文字，作为 LLM 的输入
- ✧ 文本被整合为上下文提示（Prompt），由大语言模型（LLM）进行推理生成响应。
- ✧ LLM 生成的文本发送至文本转语音（TTS）模型合成音频
- ✧ 合成音频通过实时传输网络传回用户设备完成交互

对话式 AI 的整个服务全部在云端运行，包括 ASR、LLM 和 TTS。从长远来看，人们期望有更多的 AI 服务可以在端侧运行，但云端仍是未来很长一段时间的最佳方案。

- ✧ 对话式 AI 程序需调用高性能模型（ASR、LLM、TTS）以实现低延迟的复杂流程。当前终端设备算力不足，无法在可接受延迟内运行最优模型。
- ✧ 当前，仍有大量商用场景用户是需要通过电话进行呼叫的，在此类场景中，服务无法部署到用户的终端设备上，也迫使服务集中于云端。
- ✧ 对话式 AI 需要跨终端使用，也导致云端部署成为最高效的方案。

如果你正在开发一个对话式 AI 产品或应用，你可能已经发现对话式 AI 并没有想象中的简单，我们会面临并思考以下问题：

- ✧ 对话式 AI 该选择哪些大语言模型和 TTS？
- ✧ 如何降低人机对话的端到端延迟？
- ✧ 如何让 AI 对话与人类一样自然？
- ✧ 如何在长会话中管理对话上下文？

- ◇ 如何将对话式 AI 集成到现有应用中？
- ◇ 如何评估对话式 AI 的性能表现？

2.4 响应延迟

人类在日常对话中期望能得到快速响应，长时间的停顿会让人感觉不自然，国际电信联盟 ITU 定义的语音通信延迟标准为：人类可接受的端到端单向延迟为 400ms，也就是双向对话响应延迟不高于 800ms。而目前行业内生产级 AI 人机对话端到端延迟普遍在 3 秒以上，行业领先水平也仅可以做到 2s 左右。这是因为对话式 AI 远比语音通话本身复杂得多，除了音频处理、网络传输外，还包括 ASR、LLM、TTS 等服务，每一个环节都会产生延时，这使得我们要达到较为理想的端到端对话响应延迟还有很长的路要走。我们如果提升对话式 AI 的延迟体验，就需要优化每一个过程环节的延迟响应时间。

如何测量对话式 AI 端到端响应延迟？

对话式 AI 的端到端响应延迟是指用户说话的最后一个字到 AI 发声的第一个字的时间间隔。如果你想简单测量对话式 AI 的延迟效果，你只需录制一段 AI 对话，将录音加载到音频编辑器中，查看音频波形，从用户语音结束处测量到 AI 发音处即可得到大致延迟数据。

但是要精准测量对话式 AI 的端到端语音延迟，特别是音频处理、网络传输、ASR、LLM 和 TTS 等环节的分段延时，没有特定工具是很难实现的。

声网 AI 模型评测平台（对话式）

声网为开发者提供了一个适配主流模型的实时性能数据评测平台，通过这个平台可以非常直观的测量主流 ASR、LLM 和 TTS 的动态实时延迟性能数据，同时平台还提供了最优延迟和最优性能组合供开发者参考。

Top 模型			
语音识别(ASR) 语言模型(LLM) 语音合成(TTS)			
排名	模型	供应商	末字延迟(TTLW) ①
1	 实时语音识别	凤鸣	615 ms
2	 实时语音识别	腾讯云	628 ms
3	 实时语音识别	火山引擎	787 ms

Top 模型

语音识别(ASR) 语言模型(LLM) 语音合成(TTS)

排名	模型	供应商	首字延迟 (TTFT) ①
1	 GLM 4 AirX	智谱	392 ms
2	 通义千问 Turbo	阿里云	431 ms
3	 GLM 4 Air	智谱	512 ms

Top 模型

语音识别(ASR) 语言模型(LLM) 语音合成(TTS)

排名	模型	供应商	首字节延迟 (TTFB) ②
1	 语音合成	百度智能云	173 ms
2	 语音合成	阿里云	219 ms
3	 基础语音合成	腾讯云	361 ms

上图分别是 ASR 识别、LLM 推理和 TTS 合成三类模型的处理响应时长，我们可以看到 ASR 最优延迟为 615ms、LLM 最优延迟为 392ms、TTS 最优延迟为 173ms。三大模型需要串行工作，那么纯模型响应的总体延迟就需要 1180ms。

注意，这仅仅是 AI 模型的响应延迟，如果测量完整的端到端通话延迟，还需要增加音频处理和网络传输延迟，目前行业优秀水平为 300-400ms 左右，因此，在生产环境下，1.5-1.6s 的 AI 对话延迟是一个值得追求的好体验，而 2s 以内就可以达到良好水平。

从实践来看，1s 以内的目标是非常具有挑战性的，但并非不可能用当前的方案来实现！声网目前在理想环境下，已经可以做到将对话式 AI 的端到端延迟降至 1s 以内。

★ Top 10 组合

排名	语音识别 (ASR) ① 声网 VAD	大语言模型(LLM) ①	语音合成(TTS) ①	级联模型总延迟 ①	标准差 ①
1	 凤鸣 - 实时语音识别	 智谱 - GLM 4 AirX	 百度智能云 - 语音合成	1,180 ms	± 25 ms
2	 腾讯云 - 实时语音识别	 智谱 - GLM 4 AirX	 百度智能云 - 语音合成	1,194 ms	± 29 ms
3	 凤鸣 - 实时语音识别	 阿里云 - 通义千问 Turbo	 百度智能云 - 语音合成	1,219 ms	± 39 ms
4	 凤鸣 - 实时语音识别	 智谱 - GLM 4 AirX	 阿里云 - 语音合成	1,226 ms	± 25 ms
5	 腾讯云 - 实时语音识别	 阿里云 - 通义千问 Turbo	 百度智能云 - 语音合成	1,232 ms	± 43 ms
6	 腾讯云 - 实时语音识别	 智谱 - GLM 4 AirX	 阿里云 - 语音合成	1,240 ms	± 29 ms
7	 凤鸣 - 实时语音识别	 阿里云 - 通义千问 Turbo	 阿里云 - 语音合成	1,265 ms	± 40 ms
8	 腾讯云 - 实时语音识别	 阿里云 - 通义千问 Turbo	 阿里云 - 语音合成	1,278 ms	± 43 ms
9	 凤鸣 - 实时语音识别	 智谱 - GLM 4 Air	 百度智能云 - 语音合成	1,300 ms	± 304 ms
10	 凤鸣 - 实时语音识别	 智谱 - GLM 4 Plus	 百度智能云 - 语音合成	1,310 ms	± 153 ms

2.5 智能打断

打断处理是指允许用户随时打断 AI 对话。打断是正常对话的一部分，所以为了让对话足够自然，优雅地处理打断就显得非常重要。

为了实现精准的打断处理，你需要做到的通道中每个部分都是可取消的，而且你还需要能够非常快速地停止客户端的音频播放。

2.5.1 打断机制的核心类型

1. 语音活动检测 (VAD) 打断

1) 技术原理：

通过实时分析音频流中的声音信号强度与频谱特征，判断用户是否需要开始说话。

2) 实现方式：

发声即打断：检测到人声瞬时能量超过阈值（如信噪比 $SNR > 0.8$ ）立即中断 AI，这种方式响应速度最快，但容易误触发

持续时长打断：用户语音持续超过设定阈值（如 300ms - 1000ms）才触发，避免“嗯”、“啊”等语气词误触发

3) 适用场景：

适合 AI 客服对话高时效性场景，但对环境噪音敏感。

2. 语音意图打断

1) 技术原理:

结合自然语言理解 (NLU) 技术, 识别用户输入中的关键词或紧急意图。

2) 实现方式:

- ✧ **关键词匹配:** 预设关键词 (如 “停止”、“客服”) 触发打断, 支持两种模式:
 - 语音包含关键词即触发 (如 “**小度小度**, 开灯”)
 - 仅语音开头含关键词才触发 (如 “**停止**播放”), 减少误判
 - ✧ **意图识别:** 通过模型计算意图概率, 当意图置信度>阈值时中断
- 3) **技术优势:** 精准控制打断条件, 适用于需强意图导向的场景 (如智能家居指令)

3. 手动规则打断

1) 技术原理:

用户通过物理按钮、快捷键或配置规则主动触发中断

2) 实现方式:

- ✧ **客户端事件:** 通过调用 RESTful API 或客户端组件 API 主动发起打断请求, 实现点击按钮或发送特定命令来打断智能体。
- ✧ **业务规则:** 配置强制不打断时段、字数阈值等轻量化规则 (如前 2 秒禁止打断)

3) 适用场景:

会议主持、教育工具等需完全可控的场景

2.5.2 避免意外的打断

开放的对话环境容易受到各种语音和噪声的干扰, 造成误打断的情况发生, 有几个来源非常值得注意:

突发非稳态噪声: 优秀音频降噪算法已经可以在 “人声” 与 “稳态噪声” 分离方面表现良好, 但是对于一些瞬发的非稳态噪声却常常失效, 比如咳嗽声、键盘敲击声。在技术您可以尝试调整 VAD 起始段长度和置信度阈值, 以尽量减少此类中断来源。但其权衡在于, 增加起始段长度和提高置信度阈值却会给那些您原本希望检测为完整话语的短语带来问题。因此更具性价比的方案是选择一个具备 AI 降噪能力, 可以较好地消除非稳态噪声的对话式 AI 引擎。

回声消除失败: 回声消除算法并非完美无缺, 从静音到语音播放的转换尤其具有挑战性。如果您进行过大量语音对话测试, 您可能遇到过 AI 在刚开始说话时就打断自己的情况。罪

罪魁祸首是回声消除算法允许一点点初始语音音频反馈到麦克风中。

背景语音：区分对象人声和背景语音是一大难点，如果背景语音的音量高于您的音量阈值，背景语音将可能会触发 AI 对话的中断。目前行之有效的方式采声纹锁定功能，让 AI 可以根据声纹识别、锁定对话对象，如声网的对话式 AI 的选择性注意力锁定功能。详情请参阅下文选择性注意力锁定部分内容。

2.6 选择性注意力锁定

开放环境下的人机对话，不仅会受到噪声的干扰，还会受到周围背景语音的影响，比如展会、商场、火车站等高噪声环境；家庭他人交谈、电视人声环境，都容易让 AI 无法锁定对话对象，造成对话混乱。

选择性注意力锁定（Selective Attention Locking）是认知心理学与人机交互领域的核心概念，指系统（人或机器）在信息过载环境中主动聚焦特定目标、忽略无关信息的能力。

在人类认知中，这是一种节能策略：当多个信息源同时输入时，大脑会优先处理与当前目标相关的信息（例如在嘈杂派对中听到自己的名字），并抑制无关刺激。这一现象也被称为“鸡尾酒会效应”，为 AI 系统的设计提供了重要启发。

2.6.1 技术特点

在对话式 AI 中，选择性注意力锁定的核心是**锁定目标说话人**，并尽可能降低噪声和干扰信号的影响。实现路径主要包括：

1. 声纹识别

声纹识别是利用音色、语调、语速等特征，AI 系统可稳定区分和识别不同说话人。声纹识别具备高精度识别和个性化支持两大优势：

- **高精度识别：**基于深度学习的声纹识别，即使在嘈杂环境下也能准确定位目标人声。例如，声网对话式 AI 引擎可屏蔽 95% 的背景人声干扰。
- **个性化体验：**支持自动识别不同说话人并提供差异化回应，为多用户场景带来更自然的交互。

2. 空间感知与波束形成

通过麦克风阵列和波束形成算法，基于声源空间位置定向拾音，削弱非目标用户声音。

3. 注意力机制建模

在语音识别与语义理解模型中引入 Transformer 注意力机制，使模型能“聚焦”目标信号而忽略干扰。

2.6.2 场景应用

- **会议场景：**AI 助手只响应主持人语音指令，或根据声纹区分发言人，生成更精确的会议纪要。
- **嘈杂环境：**在展会、商场等场所，锁定用户的声音，避免因背景噪声误触发。
- **家庭与个人助手：**在多成员家庭中，根据声纹为不同成员提供个性化语音控制和定制化服务。
- **医疗场景：**医生查房时，AI 只响应医生指令，自动忽略病人或其他环境音，保障信息安全与交互准确。

2.7 对话上下文管理

对话式 AI 的上下文管理是指系统在多轮交互中对历史对话信息进行获取、组织、存储和利用的过程，以确保对话的连贯性、准确性和个性化。其本质是模拟人类对话中的记忆能力，解决大模型的“记忆瓶颈”问题，让 AI 系统理解并有效利用历史对话信息、用户偏好及任务状态，与用户自然、准确对话。

2.7.1 为什么上下文很重要

1. 保障对话连贯性

大语言模型本质是基于前文来预测下一个最可能的词，没有上下文记忆，AI 就无法进行有意义的多轮对话，每次回答都会忘记之前讨论的内容。如用户先问“推荐编程书籍”，再问“它的作者是谁”，若无上下文关联，AI 无法正确回应。

2. 消除语言歧义

自然语言存在多义性（如英文 bank 可指金融机构，也可以指河岸），有上下文才能准确区分。

3. 保障意图准确性

同样的问题在不同的上下文中可能有不同的意思，如“怎么到银行”，在财务讨论和城市导航中的意义就完全不同。

4. 保障任务连续性

复杂任务需分步执行，上下文让 AI 能够跟踪进度并基于之前的结果继续工作。

5. 突破模型记忆限制

即使最强大的 AI 模型也知识与记忆容量限制，超长对话会导致历史信息丢失，需主动管理。

2.7.2 如何做好上下文？

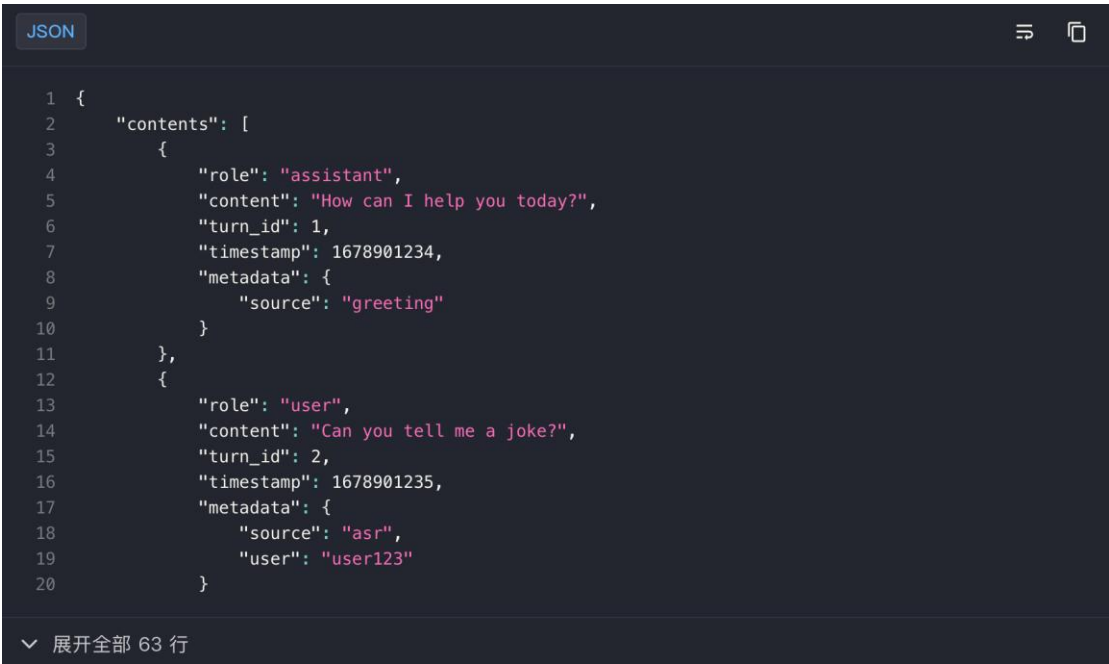
上下文管理是对话式 AI 的**记忆中枢**，需通过技术方案（RAG、结构化记忆）与优化策略（意图过滤、动态裁剪）解决连贯性、资源消耗与准确性的三角平衡问题。

声网对话式 AI 引擎采用的方案是通过结构化短期记忆+动态长期记忆注入，在保障兼容性的同时，针对实时语音交互场景（尤其中断处理）进行深度优化，并赋予开发者高度灵活的上下文控制权限。

1. 短期记忆机制

短期记忆（Short-term Memory）是智能体在当前会话中临时保存和处理的信息空间。用户和对话式智能体互动期间，智能体会在短期记忆中缓存当前对话的上下文，确保智能体能够连贯地理解和回应用户的连续输入。

对话式 AI 引擎整体短期记忆会以 JSON 的形式存储，遵循 OpenAI Chat Completions 的规则，同时进行了部分的扩展。具体数据结构如下：



```
1 {
2   "contents": [
3     {
4       "role": "assistant",
5       "content": "How can I help you today?",
6       "turn_id": 1,
7       "timestamp": 1678901234,
8       "metadata": {
9         "source": "greeting"
10      }
11    },
12    {
13      "role": "user",
14      "content": "Can you tell me a joke?",
15      "turn_id": 2,
16      "timestamp": 1678901235,
17      "metadata": {
18        "source": "asr",
19        "user": "user123"
20      }
21    }
22  ]
23 }
```

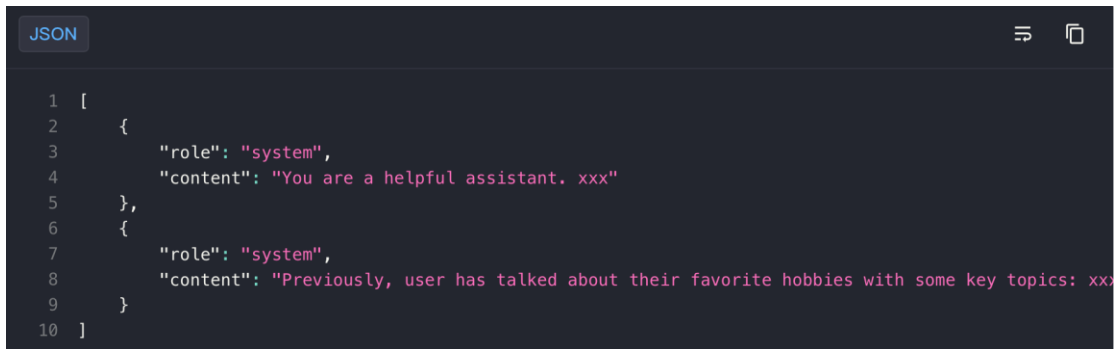
2. 短期记忆的获取：

- ✧ 智能体运行期间，调用 GET 获取智能体短期记忆接口获取短期记忆 JSON。该接口会得到智能体生命周期内储存的完整的短期记忆。
- ✧ 智能体停止后，声网会通过消息通知服务将短期记忆回调至你的业务服务器，详见消息通知事件类型。

3. 固化和注入长期记忆:

智能体的短期记忆会随智能体停止而消失,你可以在智能体停止后,将短期记忆储存到你的服务器以固化为长期记忆,之后在创建智能体时通过 `llm.system_messages` 直接注入原始记忆内容或经过总结的记忆内容,从而实现在智能体退出或重启后仍然能够访问和使用这些数据。

以下示例展示了通过 `system_messages` 注入经过总结的记忆内容:



```
1 [
2   {
3     "role": "system",
4     "content": "You are a helpful assistant. xxx"
5   },
6   {
7     "role": "system",
8     "content": "Previously, user has talked about their favorite hobbies with some key topics: xxx"
9   }
10 ]
```

4. 方案特点

✧ 灵活的多源上下文整合

- 区分消息来源 (ASR/LLM/Command 等), 支持差异化处理
- 可融合用户身份 (metadata.user) 实现个性化交互

✧ 跨会话记忆传承

- 短期记忆固化 → 业务服务器存储 → 长期记忆注入
- 解决智能体重启后上下文丢失问题

✧ 中断场景深度优化

- 精确记录语音打断时间点 (interrupt_timestamp)
- 保留原始生成内容 (original), 避免信息丢失
- 为 LLM 提供完整上下文, 提升打断后对话连贯性

✧ 兼容性与标准化

- 兼容 OpenAI 消息结构, 降低迁移成本
- 标准化 JSON 格式简化系统集成

✧ 开发者友好性

- Custom LLM 模式下开放全字段访问, 支持复杂业务逻辑扩展
- 动态更新 `system_messages` 实现实时记忆调整

2.8 视觉理解

对话式 AI 的视觉理解能力是指 AI 系统通过分析图像或视频内容, 赋予 AI “看懂” 图

像的能力，结合语音对话实现多模态融合交互。

视觉理解具体表现在两个维度：实时视频解析能力（环境理解、动态手势识别、人物表情识别）和静态图像分析（图像/图表解析、拍照识题）

2.8.1 实时视频解析能力：动态场景的“感知-决策”闭环

通过摄像头捕捉连续画面，在语音对话的基础上，为 AI 提供视觉能力，其中关键技术能力包括：

1. 环境与对象识别

- ✧ **实时识别环境与空间布局**（如家庭房间结构、办公室设备分布），结合用户指令生成场景化反馈（如“关闭左侧台灯”需定位灯具位置）。
- ✧ **实时识别对象特征**（如人物特征、生物/物品类别），结合对话内容，准备回来用户问题（如“请告诉我旁边的这只小猫是什么品种”）

2. 动态手势识别

- ✧ **高精度动作捕捉**：捕捉手势轨迹变化，智能识别挥手、点赞、比划数字等手势动作，延迟控制在 200ms 内。
- ✧ **视觉+语音协同**：通过麦克风阵列定位声源，结合摄像头捕捉手势意图（如停止手势打断播报），实现多通道指令解析。

3. 人物表情与情感识别（未来目标）

- ✧ **微表情分析**：通过面部肌肉运动单元（AU）量化喜悦、困惑、焦虑等情绪，从面部肌肉微动捕捉情绪波动，准确感知并理解情绪变化。
- ✧ **情感识别**：AI 学会了“看懂”而不仅仅是“看到”，从而理解“视觉画面中”物体之间的关系、情感色彩，甚至是隐含的社会文化背景。这类似于培养一个孩子的观察能力，从最初的“看到红色的球”发展到理解“孩子们在操场上快乐地玩球”。

2.8.2 静态图像分析能力：从物体识别到逻辑推理

1. 通用物体识别与定位

- ✧ **空间理解**：理想空间关系，比如描述“猫趴在键盘上阻挡了屏幕”的干扰关系；
- ✧ **物体识别**：物种亚类（如“柯基犬”而非仅“狗”）、物品类别（电脑、电视）、商品型号（iPhone 15 Pro/Pro Max）。

2. 专业场景解析

- ✧ **文档与图表处理**：理解文档与图表内容，甚至解析学术论文中的化学结构式、数学公式推导逻辑；

- ✧ **工业图纸分析：**识别机械图纸尺寸公差，自动检查装配冲突；
- ✧ **医疗影像初筛：**标注 CT 影像中的结节位置，关联良恶性概率（需医生复核）。

2.8.3 对话式 AI 的视觉理解进化方向

- ✧ **感知维度扩展：**从静态图像到长视频时序分析（>10 分钟）和实时视觉内容；
- ✧ **交互深度升级：**从单向问答到多模态协同决策（如手势+语音操控家电）；
- ✧ **认知能力涌现：**AI 开始展现“理解意图-关联常识-情感反馈”的类人思维链。

2.9 音频处理

在实时互动中，音频的采集和处理异常复杂的，但一个优秀的实时互动或对话式 AI 应用通常会让你忽略音频处理的复杂性。不过当我们要去构建复杂的对话式 AI 应用时，就不可避免会面临着实时音频处理中的各种问题，所以我们需要了解音频处理的主要技术点：

2.9.1 麦克风和自动增益控制

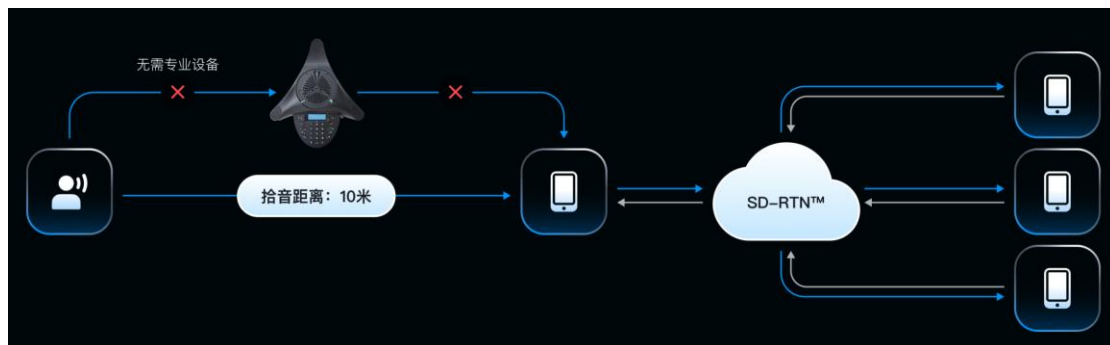
麦克风的小型化让我们可以将麦克风内置在任何设备之中，让我们通过各类麦克风可以从智能设备、笔记本电脑、蓝牙耳机中内置的小麦克风获得出色的音频。但值得注意的是，我们在实际应用中却常常面临着距离远近、设备差异带来各类音频问题。通常我们会采用自动增益控制（AGC）来解决这一问题。

自动增益控制（AGC）主要是通过自适应增益算法动态调节输入/输出音量，解决因用户距离麦克风远近或设备差异导致的音量波动问题，确保语音信号强度稳定在理想范围内（-20dB 至 40dB）。

动态响应：在嘈杂环境中自动提升人声音量，安静场景则抑制增益，避免爆音失真；

低失真处理：结合人声频谱保护技术，避免过度放大导致的语音失真。

自动增益控制（AGC）主要是根据输入音频信号的能量水平动态调整音频增益，当音频信号的能量较大时，减小增益值，以避免音频信号过大而造成的失真；当音频信号的能量较低时，增加增益值，以避免音频信号过小而导致的难以听清的问题，如声网的自动增益控制，可以做到不依赖专业设备，清晰拾音距离达到 10 米。



2.7.2 回声消除 (AEC)

声学回声是指由通信设备的扬声器发出的声音经过声学环境传播被麦克风采集到的信号。回声现象在实时音频通话中特别普遍，特别是用户使用扬声器通话，或者在没有耳机的情况下使用笔记本电脑进行通话时。

在实时音频互动中，回声信号会严重影响通信双方的音频可懂度和通信体验舒适度。因此回声消除成为音频通信过程中一项必要的处理环节。

回声消除技术的原理是通过分析麦克风采集到的音频信号和扬声器播放的音频信号之间的相关性，来估计得到音频声学传播路径的冲激响应，然后根据估计的回声路径冲击响应与扬声器的播放信号估计麦克风采集到的回声信号，最后通过逆滤波的方法将估计的回声信号从麦克风采集的音频中减去，达到消除回声的目的。

常用的回声消除算法包括基于各种判决准则的自适应滤波法、卡尔曼滤波法等。其中，自适应滤波法是通过采集回声信号和麦克风输入信号之间的相关性信息，来动态地调整滤波器参数，以适应不同的回声环境。卡尔曼滤波法则是首先预测得到初步的结果，然后根据实际采集信号与估计信号之间的误差对初步估计结果进行修正，从而得到更为准确的滤波器系数。

值得庆幸的是，主流的 RTC SDK（如声网 RTC SDK）目前都已经内置回声消除模块，因此，如果你使用 RTC SDK 或对话式 AI 引擎构建对话式 AI 应用时，你就可以依赖其内置的回声消除功能，你只需要弄清楚如何集成回声消除逻辑。

当然你如果选择基于 WebSocket 来构建你的对话式 AI 应用，你将不会默认获得任何回声消除能力。

2.9.3 噪声抑制 (ANS)

降噪是指通过各种技术手段减少音频信号中的杂乱噪声。在实时音频处理中，降噪通常包括预处理和后处理两个阶段。

1) 预处理阶段，在声音采集阶段，降噪系统会首先获取音频信号，并利用特定的算法进行初步处理。常用的降噪算法包括均值滤波、中值滤波和自适应滤波等。这些算法通过分析音频信号的统计特征，去除其中的噪声成分。

2) 后处理阶段，在音频采集后，降噪系统会对音频信号进行进一步处理。后处理阶段的降噪算法通常包括频域滤波、时域滤波以及混合滤波等。这些算法通过对音频信号在频域或时域进行分析，去除其中的噪声成分。

1. AI 降噪

降噪技术有很多种，各有特点，也各有局限。近年来，基于深度学习的 AI 降噪算法得

到了广泛应用。这类算法通过使用深度神经网络对音频信号进行建模，不仅能够有效地抑制稳态噪声，对于非稳态/突发噪声也有非常良好的效果。

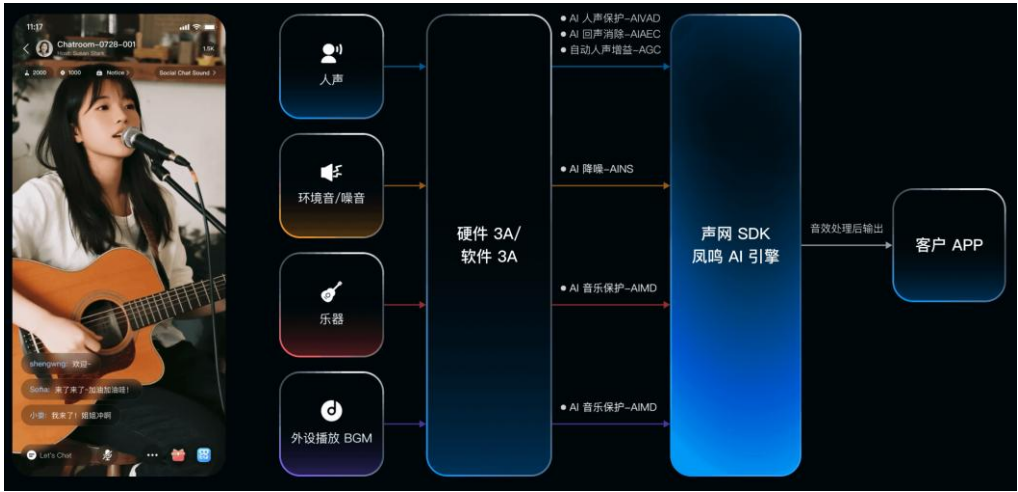
在实际应用中，基于深度学习的 AI 音频降噪技术通常需要大量的标注数据进行训练，如声网的 AI 降噪技术基于其超 10000 亿分钟规模的海量音频数据训练，可以在保证人声无损伤的前提下，强力抑制 100+ 类型的稳态和非稳态噪声，在低信噪比或人声密集场景下也能实现纯净声音体验。



2. AI 人声保护/AI 音乐保护

在强降噪场景下，人声/音乐背景容易被当做噪音消除从而导致语音失真、人声空洞感，以及音效差的情况发生，这需要我们在开启降噪的同时，启用 AI 人声保护尽最大限度避免人声损伤和丢字，实现在强降噪场景下的高保真语音处理。

- ✧ **通话人声保护**：在强降噪场景下，通过频域掩蔽技术保护人声频段，避免语音失真或空洞感；
- ✧ **重叠人声分离**：确保多人对话场景中每个说话者的语音清晰可辨；
- ✧ **背景音乐保护**：保留还原乐器、背景音乐声效。



2.10 网络传输

2.10.1 WebSocket 与 WebRTC

在构建对话式 AI 服务，选择 WebSocket 还是 WebRTC 方案经常被提及甚至拿来比较，其实两者之间存在显著的区别：

WebSocket：适用于实时双向数据传输的场景，如在线聊天室、实时数据监控、股票行情推送、在线游戏状态同步等，更侧重要于文本、二进制等数据的快速交换

WebRTC：适用于实时音视频通信，如果视频通话、视频会议、对话式 AI、直播连麦等场景，在这些场景中，对音视频的实时性、流畅性较高，WebRTC 可以很好地满足这些需求。

简单来说 WebSocket 更适合于 Server-to-Server，且对延迟不敏感的场景；而若需构建基于 App 或浏览器的原生应用，并要实现对话级的超低延迟，你无疑应该选择 WebRTC 方案。

1. WebSocket 在实时对话中存在关键缺陷

- ✧ **WebSocket 的 TCP 依赖导致关键缺陷：**WebSocket 基于 TCP 协议，其可靠传输机制在实时媒体流中反而成为瓶颈。Head-of-Line Blocking 会导致单个数据包丢失或延迟时，后续所有音频包被阻塞，造成卡顿和延迟累积。对于对话式 AI 需连续交互的场景，此问题会显著破坏对话流畅性
- ✧ **WebRTC 的 UDP 更适合实时传输：**WebRTC 基于 UDP 协议，天然支持**选择性丢包**：网络拥塞时可自动丢弃延迟到达的冗余包，优先保障最新数据。同时结合**前向纠错（FEC）**和动态码率调整，即使在高丢包率下仍可实时流畅通话，比如声网已支持在 80%的情况实现流畅音视频通话。

2. WebRTC 更适应复杂网络和业务逻辑

- ✧ **拥塞控制与带宽自适应：**WebRTC 内置拥塞控制算法，可实时监测网络抖动、延迟及丢包率，动态调整音频码率与帧率，适应 WiFi/4G/5G 等多种网络环境
- ✧ **精准时间戳与打断逻辑：**音频帧自带 RTP 时间戳，实现精准播放时序控制。用户打断 AI 发言时（如插话），WebRTC 可立即中止当前输出并切换响应，而 WebSocket 需应用层额外实现中断同步逻辑。

3. WebSocket 音频处理能力缺失

WebRTC 以及主流的 RTC 产品（如声网 RTC SDK）目前都已经内置 3A（AEC 回声消除/ANS 降噪/AGC 增益控制），可直接处理设备麦克风输入，消除环境噪音与回声。

而 WebSocket 需开发者自行集成 VAD（语音活动检测）及降噪库，增加实现复杂度。

4. WebRTC 支持数据透明和观测

WebRTC 支持实时上报端到端延迟、抖动、丢包率等指标，便于监控 QoS，如声网水晶球。WebSocket 需自行埋点采集数据，且难以区分网络延迟与应用处理延迟。

5. WebRTC 协议栈更完备

WebRTC 提供全栈媒体解决方案（编解码/传输/加密/NAT 穿透），开发者仅需调用 API。反观 WebSocket 需额外实现音频分片与重组逻辑、自定义拥塞控制、端到端加密、重传与心跳机制。

2.10.2 实时传输网络

对话式 AI 应用的构建还需要选择一个合适的网络服务方案，因为直接部署在互联网的质量和用户体验是难以得到保证的。

大家熟悉的互联网是基于标准的 Best-Effort（尽力而为）服务模型来设计，它的使命是尽最大可能地传输数据，但不保证实时性、可靠性等 QoS 指标。因为互联网的核心任务是满足全球超过 50 亿网民的保证互联互通，以确保适应网络用户的快速增长。而作为代价，QoS 与用户体验等并不被互联网优先考虑。

为提高互联网 QoS 保障能力，业界提出了许多解决方案，其中基于软件定义的虚拟实时网络在实践中被证明一种行之有效的方案。如声网的软件定义全球实时网络 SD-RTN™（Software-Defined Real Time Network），基于底层互联网构建，可通过软件定义算法构建的有 QoS 保障的虚拟实时网络，专为实时音视频业务提供接近“专线”的质量。

1. 网络架构

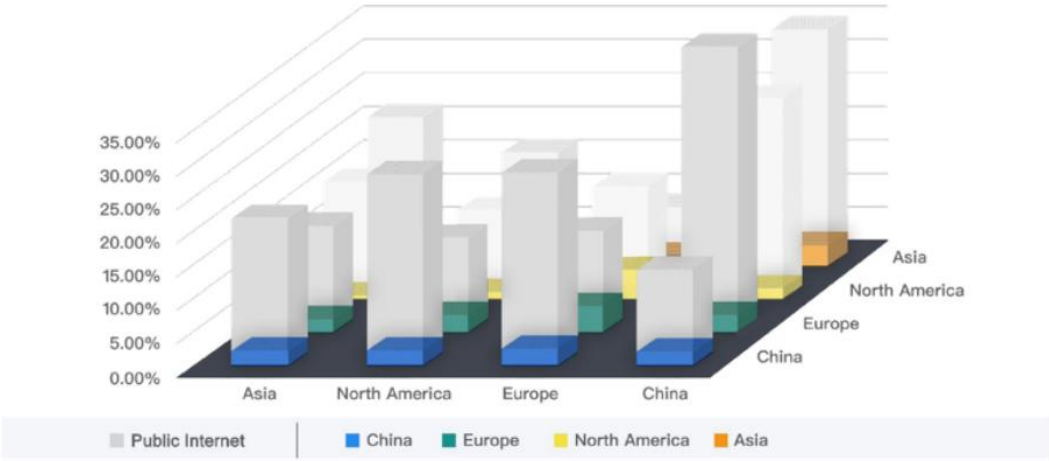
SD-RTN™采用边缘节点（Edge Node）接入，中心节点（Core Node）控制的体系架构，所有节点分布部署在全球互联网的不同数据中心，不同节点之间相互协作，共同完成路径测量、数据转发、传输模式控制、接入选择和权限认证等工作。SD-RTN™采用云端控制算法综合考虑接入选择、高可扩展性、高可用性、负载均衡等指标，为实时音视频业务提供最佳的 QoS 质量保障。

2. 全球覆盖

SD-RTN™采用全球部署接入节点的方式，通过软件技术形成 overlay 叠加层全球网。当前 SD-RTN™部署在全球五大洲 255 个骨干网数据中心。

SD-RTN™中每个节点都是 SD-RTN™的接入点，为全球所有声网用户提供一个就近的快速访问接入点，后端不依赖于固定的域名系统列表，总是根据实际网络条件实时寻找最佳连接路由，提供动态的网络逻辑，无论用户身处何时何地，业务都触手可达。

同时 SD-RTN™可以通过多条独立“优化”路径智能地发送数据包，以确保在最小时间窗口内获得较高传输成功率。在这里，首先到达数据包被使用，而任何丢失或延迟的数据包将被忽略。



SD-RTN™与公共互联网丢包率对比

3. 智能路由

SD-RTN™采用智能路由机制，所有边缘节点以秒级为单位周期性的互相发送测试数据包，并实时上报给中心节点，中心控制器通过深度学习算法，综合丢包、时延、抖动等网络指标，为实时音视频业务“导航”出一条最优路径。相比公共互联网，SD-RTN™的路由刷新速度大幅提升，可以有效地避开流量拥塞路段，为用户提供最“实时”的音视频互动体验。

4. 优质传输

优质传输率能够直接反映 SD-RTN™在一个地区的覆盖水平。当前 SD-RTN™日均服务分钟数过万的国家地区超过 190 个，覆盖水平达到良好和优秀的国家地区超过 130 个。

地区	澳大利亚	国内小ISP	中国移动	中国电信	中国联通	东亚	欧洲	北美	南亚
澳大利亚	100%	100%	100%	100%	99.86%	100%	99.93%	100%	100%
国内小ISP	99.85%	99.1%	99.93%	99.93%	99.86%	99.93%	99.79%	99.79%	99.86%
中国移动	99.86%	100%	100%	99.79%	99.93%	100%	99.93%	100%	99.86%
中国电信	100%	100%	100%	100%	100%	100%	99.93%	99.93%	99.86%
中国联通	99.44%	100%	100%	100%	100%	100%	99.72%	99.86%	99.93%
东亚	100%	100%	100%	100%	100%	99.93%	100%	100%	99.93%
欧洲	100%	100%	100%	100%	100%	100%	100%	99.93%	100%
北美	100%	100%	100%	100%	99.93%	99.93%	99.93%	100%	99.51%
南亚	99.93%	99.93%	100%	100%	100%	100%	100%	100%	100%

SD-RTN™ 全球优质传输率

地理位置	传输方式	传输时延 (ms)			
		50%用户	70%用户	90%用户	95%用户
北美	公共互联网	44	63	82	94
	SD-RTN™	27	35	42	53
欧洲	公共互联网	36	44	72	165
	SD-RTN™	29	36	42	59
亚洲	公共互联网	38	55	85	128
	SD-RTN™	30	38	42	62
南美	公共互联网	47	92	127	203
	SD-RTN™	38	56	63	82

区域内延迟

地理位置	传输方式	传输时延 (ms)			
		50%用户	70%用户	90%用户	95%用户
北美-欧洲	公共互联网	102	159	217	273
	SD-RTN™	62	80	81	83
亚洲-北美	公共互联网	165	260	343	420
	SD-RTN™	103	115	119	124
亚洲-中国	公共互联网	45	83	163	236
	SD-RTN™	35	51	63	92
南美-中国	公共互联网	293	304	364	472
	SD-RTN™	187	209	219	227

区域间延迟

第三章：对话式 AI 的主流方案及产品生态

3.1 技术架构与方案分类

随着对话式 AI 技术的快速发展，市场上涌现出了多种技术架构方案和产品形态。VoiceAgent 是利用语音 AI 和实时互动技术进行交互、解决特定场景问题的智能体，代表场景有情感陪伴、智能外呼、实时翻译以及会议协作等。目前 Voice Agent 有两种主流架构方式，传统三段式的级联模式，以及端到端模式。级联模式以其模块化和低成本带来的灵活性，成为现阶段大多数开发者和企业的首选。

3.1.1 串行方案

串行方案（级联模式）是将三个独立组件串联起来的流水线，三个组件各自负责专业的工作：语音识别（STT， Speech - to - Text）把语音输入转化为文本；语言模型（LLM， Large Language Model）分析文本并生成回复内容；语音合成（TTS， Text -to - Speech）将回复内容转为语音输出。



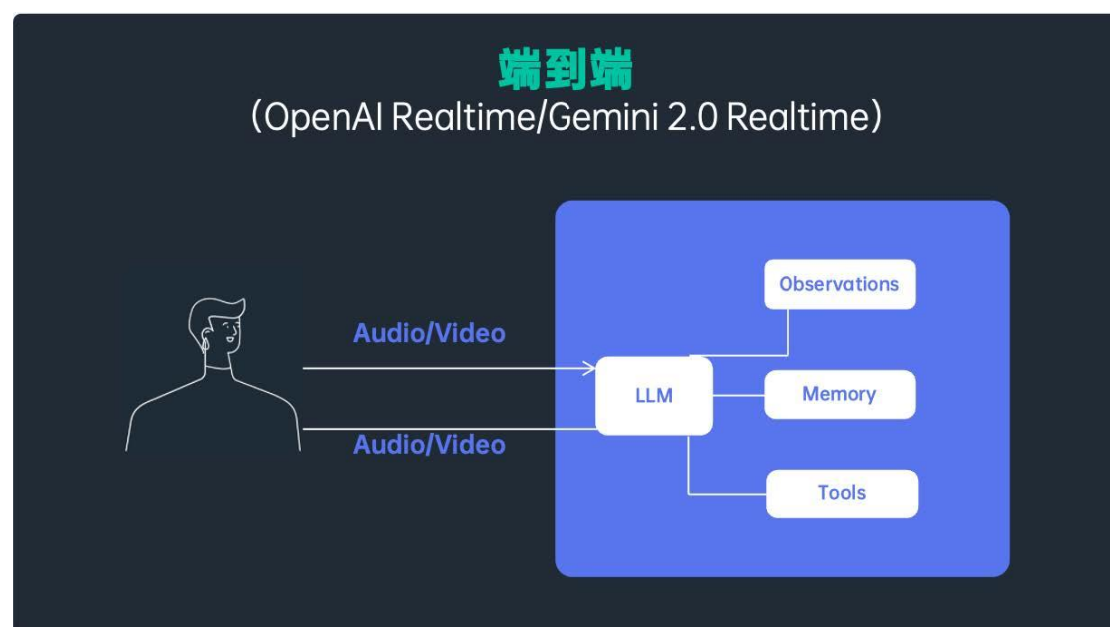
级联模式的优点在于是串联模式，允许开发者根据具体需求选择不同的服务供应商。例如，在口语练习中，可以选择擅长处理口音问题的 STT 服务。当这条串联路线上某个组件出现问题时，可以单独对其进行调试或替换，不会影响整体运作。这样模块化的另一个好处是成本可控，灵活搭配不同价位的服务。比如在产品快速验证时期可以选择开源项目，随着产品的不断发展，选择更专业、功能更完善的商业版服务。

但由于涉及多个环节，可能会导致响应时间较长，并且需要分别对各个组件进行调试和部署，增加了复杂性。从实际应用来看，级联模式在当前技术条件下仍然是最成熟和可

靠的选择，特别是对于需要定制化功能和成本控制的企业应用场景。

3.1.2 多模态方案

相比之下，多模态方案（端到端模式）更加直接，用户输入语音，模型内部经过一系列复杂的处理过程，直接输出语音回复，中间过程不会显性展示。以 OpenAI 的 RealtimeAPI 为例，它让开发者可以构建近乎实时的「语音到语音」的体验，实现更流畅的交互体验。这种模式将语音的处理过程从级联模式的三个步骤降低到一个步骤，显著降低了延迟，响应速度更快。由于只有一个模型，部署也很简单。



但是，这意味着整个过程更像是一个“黑盒子”，很难针对其中某一环节进行单独调整。而且需要投入大量资源进行开发。如果想要适应新的应用场景，可能需要重新训练整个模型。端到端模式可能还存在回复不够深入和严谨的问题。吴恩达在今年二月的分享中提到端到端语音模型的两个挑战：缺少类似文本的结果检查；语音响应需要更加简洁，可能会减少了通过思维链推理，答案不够深入。

理论上端到端模型延迟更低，实际使用中，一般平均延迟确实也更低，但经常会出现个案延迟很高的情况，也就是延迟不稳定，这方面不如级联模型成熟。以及，黑盒子不光是单独调整的问题，而是整个推理过程不透明（因为没有文字形式的中间步骤可以看），不方便定位问题和调试。

从技术发展趋势来看，端到端模式代表了对话式 AI 的未来方向，但在当前阶段，其技术成熟度和应用灵活性仍有待提升。对于大多数企业和开发者而言，级联模式仍然是更实用的选择。

3.2 如何快速构建对话式 AI 应用

对于开发者和企业而言，选择合适的技术方案是构建对话式 AI 应用的关键。当前市场上主要有三种主流的构建方案，每种方案都有其独特的优势和适用场景。如果你正在开发一个对话式 AI 产品或应用，你可能需要考虑以下问题：对话式 AI 该选择哪些大语言模型和 TTS？如何降低人机对话的端到端延迟？如何让 AI 对话与人类一样自然？如何在长会话中管理对话上下文？如何将对话式 AI 集成到现有应用中？如何评估对话式 AI 的性能表现？

3.2.1 基于对话式 AI 引擎方案

声网对话式 AI 引擎重新定义了人机交互界面，突破了传统文字交互，实现了高拟真、自然流畅的实时语音对话，让 AI 真正“开口说话”。该引擎支持个性化语音定制、情感化表达和深度多轮对话理解，赋予 AI 智能体更强的表现力和互动感。适用于智能助手、情感陪伴、口语陪练、智能客服、智能硬件、沉浸式游戏 NPC 等创新场景，让 AI 智能体以更具温度和情感的声音，与用户建立更加自然、生动的交流体验。



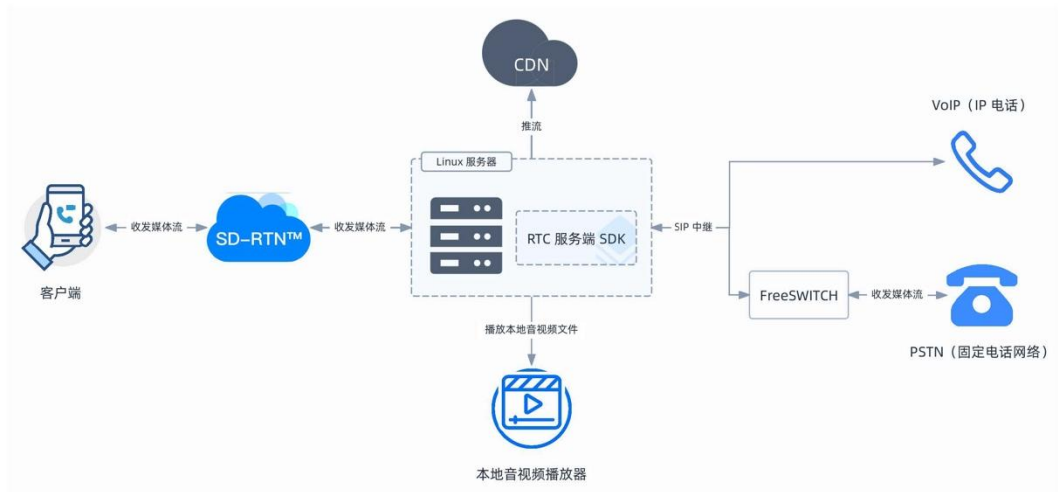
声网对话式 AI 引擎的核心优势在于其全链路深度优化，端到端响应延迟低至 650 ms，确保对话流畅自然。通过选择性注意力锁定技术，能够屏蔽 95% 环境人声、噪声干扰，精准识别对话人声。340 ms 极致打断响应，模拟真人对话节奏，让对话更贴近真人交互体验。即使在 80% 丢包率下仍能稳定流畅对话，断网 3-5 秒也能自如响应。同时兼容适配全球主流 LLM、TTS，满足不同场景和业务需求，开发者只需 2 行代码、15 分钟，即可快速构建对话式 AI 智能体。

该引擎在功能支持方面表现全面，包括 AI 实时语音对话，让用户与 AI 智能体实现语音对话，如真人一般自然；语音打断功能，允许用户随时语音打断 AI 智能体并快速响应，提升对话效率和流畅性；AI 降噪技术，智能过滤环境噪音，提升交互准确性；背景人声抑

制，有效分离目标人声与背景人声，确保对话清晰无干扰；实时字幕功能，语音实时转换为文字，提供字幕同步支持。

3.2.2 基于 Linux SDK 构建方案

对于需要更高度定制化和灵活性的企业应用，基于 Linux SDK 的构建方案提供了完整的底层控制能力。声网 RTC 服务端 SDK 部署在服务端，可与集成了声网 RTC 客户端 SDK 的 app 通过 SD-RTN™ 进行实时通信，实现向客户端发送音视频流和从客户端接收音视频流的功能。该方案展示了 RTC 服务端 SDK 的几种应用，包括在频道中播放本地文件，直推 CDN，以及与企业呼叫中心互通。服务端 SDK 将 SD-RTN™ 中的媒体流转换为指定格式，输出给其他模块，并将其他模块的媒体流编码，发送到 SD-RTN™。



这种方案的产品特性包括：兼容性强，与 Android、iOS、Windows、macOS、及框架平台互通，与声网的 Web RTC SDK（3.0.0 或更高版本）以及小程序声网 RTC SDK（2.4.2）互通；可靠性高，支持集群部署，动态扩容，服务高可用；流畅性好，基于 SD-RTN™ 的音视频抗丢包特性，实现低延迟的服务器到客户端的内容传递；高并发支持，同时支持推拉数百路音视频流；安全性强，提供音视频通话、数据传输、数据存储等端到端安全保障机制。

在适用场景方面，该方案广泛应用于 AI 虚拟人场景，服务端 SDK 深度融合了 AI 算法，构建了 AI 虚拟人多模态交互的音视频通讯基础，可以实现用户和虚拟人之间低延迟、高清音视频的实时互动；AI 互动课堂，这是一种个性化在线教学方式，利用 AI 技术对学生的表现进行智能分析，向学生推送针对性的音视频课件讲解；IoT 应用，服务端 SDK 可以支持智能穿戴设备、摄像头、传感器等物联网设备间的实时音视频传输，实现智能安防、远程监控、远程会议等功能；云游戏领域，服务端 SDK 能够处理云游戏画面的实时捕捉、编码和传输，确保玩家在云端大型游戏中获得低延迟、高画质的体验。

3.2.3 基于 TEN 开源框架自研方案

TEN 是一个全面的开源生态系统，用于创建、定制和部署具有多模态功能的实时对话式

AI 代理，包括语音、视觉和虚拟形象交互。该框架为有技术实力的开发团队提供了完全自主可控的解决方案，包括 TEN Framework 核心框架、TEN Turn Detection 转话检测、TEN VAD 语音活动检测、TEN Agent 代理示例、TMAN Designer 可视化设计器、TENPortal 官方门户等组件。

TEN 框架的核心功能十分丰富。在实时虚拟形象方面，可以使用 Trulience 的多样化免费虚拟形象选项构建引人入胜的 AI 虚拟形象，支持多种虚拟形象和自定义选项。在实时语音与 MCP 服务器集成方面，TEN Agent 与 MCP 服务器无缝集成，扩展其 LLM 功能，支持实时对话和工具集成。在硬件实时通信方面，TEN Agent 可在 Espressif ESP32-S3 KorvoV3 开发板上运行，实现硬件上的实时通信与 LLM 集成。在实时视觉和屏幕共享检测方面，支持 Google Gemini 多模态 Live API，具备实时视觉和屏幕共享检测功能，集成天气查询和网络搜索等强大工具。

TMAN Designer 是该框架的一个重要组成部分，它是一个低/无代码选项，用于创建具有易于使用的工作流 UI 的语音代理。它可以加载应用程序和图形，并包括在线编辑器、日志查看器等功能。对于开发者而言，使用 TEN 框架需要一定的系统要求，包括 Agora AppID 和 App Certificate（每月免费分钟数）、OpenAI API 密钥（兼容 OpenAI 的任何 LLM）、Deepgram ASR（注册即可获得免费积分）、Elevenlabs TTS（注册即可获得免费积分）等密钥，以及 Docker/Docker Compose、Node.js (LTS) v18 等安装环境，最低系统要求为 CPU $\geq$ 2 核、RAM $\geq$ 4GB。

3.3 对话式 AI 技术生态

级联模式灵活性的基础，是开放繁荣的生态。根据 RTE 开发者社区发布的「RTE+AI 融合生态洞察报告 2024」，围绕 Voice Agent 产品的研发、商业化和增长的生命周期，在各个细分方向上已经有着众多技术成熟的供应商。如以 DeepGram、AWS 为代表的 STT、Minimax、Azure 为代表的 TTS、以及 Open AI、Gemini、DeepSeek 等百花齐放的 LLM。除了基础的三个组件，该生态中还有提供底层传输的 RTC（如声网）、以及方便开发者快速搭建部署 Agent 的框架（如开源框架 TEN Framework）等。这些上下游企业让开发者能够选择最适合的能力组合、快速接入新服务、利用和学习行业最佳实践。

在基础技术服务商方面，STT（语音识别）服务包括 Deepgram 的高精度实时语音识别、AWS Transcribe 的云端语音识别服务、Google Speech-to-Text 的多语言语音识别支持、Azure Speech Services 的企业级语音识别解决方案。TTS（语音合成）服务涵盖 Minimax 的高质量中文语音合成、Azure Cognitive Services 的多语言语音合成、ElevenLabs 的 AI 驱动语音克隆、Amazon Polly 的云端文本转语音服务。LLM（大语言模型）服务则包括 OpenAI GPT 系列的业界领先通用大模型、Google Gemini 的多模态大语言模型、DeepSeek 的开源高性能大模型、Claude 的 Anthropic 开发安全 AI 助手。

底层技术基础设施同样重要，实时通信（RTC）技术方面有声网 RTC 的全球领先实时音视频技术、腾讯 RTC 的企业级实时通信服务、Twilio 的实时通信解决方案。基础设施支持包括 AWS、Azure、Google Cloud、阿里云等云计算平台，边缘计算技术用于降低延迟、提升用户体验，CDN 网络提供全球内容分发和加速服务。

级联模式在这样开放生态的支持下，已经有产品正在逐渐赶超端到端模式的体验。近期声网发布的全球首个对话式 AI 引擎就是采用级联模式的结构，基于在音视频领域深厚的技术优势与场景实践的积累，此次发布的对话式 AI 引擎既保留了级联模式灵活低成本优势，又克服了其延迟高和部署难的问题。只需 2 行代码，15 分钟，开发者即可部署基于大模型的对话式 AI Agent，语音对话延迟中位数低至 650ms，让 AI 秒回你的提问。

随着技术的进步，Voice Agent 的生态还会不断的扩大与完善，形成更成熟灵活的生态协作机制。为开发者提供更便捷的开发工具，降低算力与流量成本，创造更沉浸自然的用户体验，进而促成更健康的产业结构。

3.4 对话式 AI 产业生态

对话式 AI 的产业生态正在快速形成和完善，从国际大厂的企业级解决方案到国内创新公司的特色产品，从开源框架到商业平台，形成了百花齐放的格局。这个生态系统不仅为不同需求的开发者和企业提供了多样化选择，也推动了整个行业的技术进步和应用创新。



对话式 AI 产业生态呈现多层次发展格局。在国际主流平台方面，Google Dialogflow CX、Microsoft Azure Bot Service、IBM Watson Assistant、Amazon Lex 等企业级解决方案占据重要地位，提供强大的自然语言理解和多渠道部署能力。开源框架如 Rasa 为中小企业和个人开发者提供了低成本的入门选择。

国内 LLM 产品蓬勃发展，讯飞星火、通义千问、智谱清言、Kimi、小冰等 AI 产品各具特色，在多语言处理、多模态理解、长文本处理、情感计算等方面展现出强劲实力。开发工

具与平台方面，字节跳动扣子（Coze）、Dify、n8n 等平台为不同技能水平的用户提供了便捷的开发环境。

应用场景不断拓展，智能客服实现 7×24 小时无间断服务，教育培训领域的口语陪练和 AI 互动课堂提供个性化学习体验，情感陪伴应用满足用户的情感需求，智能硬件集成让设备更加智能化。

随着技术的进步，对话式 AI 将在大模型小型化、跨语言理解、情感智能、常识推理等方面实现重要突破，同时在开发效率、部署便利性、运维自动化、成本优化等工程化能力方面持续提升，推动整个产业生态的协同发展。

第四章：对话式 AI 体验质量评估方法

在这场由技术驱动的变革中,我们见证了对话式 AI 或语音智能体(Voice Agent)应用的不断涌现。但与此同时,一些日益凸显的关键挑战也摆在面前:我们该如何科学评价人机对话体验的好坏?AI 的理解和表达需达到何种水平才能让用户满意?什么样水平的人机对话体验能真正创造用户价值与商业价值?

当我们尝试抽丝剥茧地去分解语音和对话的要素时,就会发现其中蕴含着极为丰富且复杂的信息。除了话语的字面意思(语义),说话时的韵律特征(如说话方式、风格、重音、速率)也承载着大量信息。再到语用层面,说话口吻、遣词造句的方式会赋予话语更多语境内涵。此外,诸如疲劳、紧张的音质,笑声、叹气声、清嗓子等非语言声音,物理环境内的其他声音事件,乃至说话人的身份等,都可能包含有价值的信息量。不仅如此,对话交互本身的动态过程,如等待响应的时间、打断的时机与处理、多轮对话的连贯性、听觉上的疲劳度、回应的一致性等等因素,都会对整体对话体验产生直接而显著的影响。

而当我们回顾当前常用的评估手段时,会发现它们与评估这种复杂体验的需求之间还存在较大的鸿沟。目前常见的几类评价方式:一类像词错误率(WER)这种针对特定模块语音识别(ASR)能力的量化评估指标,可以从某个技术侧面反映对话式 AI 的部分性能;另一类像任务完成率,则属于面向特定目标(通常是服务型场景)是否达成来设计的评估方案。当然也有类似 Chatbot Arena 那样,通过收集最终用户的总体评分或偏好选择来进行主观评价的方法。这类方法虽然能反映整体偏好,却往往难以深入揭示体验好坏的具体原因,为针对性改进提供的直接指导也相对有限。

总体而言,当前行业内普遍缺乏一个能够系统性梳理、全面涵盖上述复杂因素,并能有效指导实践的对话式 AI 体验评估框架。面对现有评估手段的局限与对话体验本身的复杂性,我们需要首先明确构成“良好”体验的关键要素,并构建一个更完善又能实操的评估体系。

声网音频技术负责人陈若非在行业首次提出了“三维二轨”的对话式 AI 质量评估方法,也感谢他对本章节的内容贡献。

4.1 性能指标体系：三维二轨

1. 解构对话体验：“三维二轨”评估框架

在构建一个有效的评估框架前,我们先来探讨一下什么是好的人机对话体验。声音作为信息与情感的载体,构成了对话的基础。人与人之间的语音对话,本质上是信息交换与情感体验共建的过程。而在人机对话这个特定场景里,用户的核心诉求通常可以归结为两个主要方面:

获取有效信息	获得良好感受
满足学习知识、解答疑问或达成特定任务的需求。	在交互中感受到舒适、被理解、被尊重，并获得积极或符合预期的情感反馈。

当这些核心诉求得到满足时，用户往往会认为这是一次好的对话体验。

同时，评估体验不能脱离具体的应用场景。不同的场景目标，决定了评价对话好坏时侧重点的不同。当前常见的对话式 AI 应用，大致可以根据其主要目标划分为两类：

服务型智能体	陪伴型智能体
更侧重于高效地完成特定任务或提供准确信息。评估重点偏向信息获取的有效性。	更侧重于建立情感连接、提供陪伴和维持舒适的关系。 评估重点偏向情感体验的营造。



要实现有效的信息传递，智能体需要具备精确的理解能力和相应的问题解决智能。而要提供良好的情感体验，则不仅需要敏锐地感知用户的情绪与意图，还需要以恰当的方式进行表达与回应。此外，无论哪种场景，交互过程本身的质量（如响应速度、流畅度）都至关重要。

2. “三维二轨”评估框架概述

基于上述讨论，同时也考虑和现有技术框架的结合，我们建议使用一个“三维二轨（Three Dimensions, TwoTracks）”的评估框架来设计对话式 AI 体验的评估。该框架旨在从 AI 自身的核心能力（三个维度）和具体的评估路径（两条轨道）这两个层面，对复杂的对话体验进行解构和结构化分析。

我们将影响对话体验的 AI 端核心因素划分为三个关键能力维度：

AI 的理解能力

指的是 AI 准确感知并理解用户输入信息（涵盖听觉感知、语言语义、语用情境、情感等多方面内涵）的综合能力。这是有效回应的基础。

🗣️ AI 的表达能力

指的是 AI 能够生成清晰、自然、富有表现力且符合场景需求的语音进行回应的能力。这直接影响信息的传递效果和用户的情感感受。

💬 交互能力

指的是在用户与 AI 的交互过程中,所体现出的对话协调性、流畅度、响应性与整体舒适度等动态行为表现。这关乎交互过程本身是否顺畅自然。

基于评估目的的差异,我们将具体的评价活动分为两条可并行的测试路径:

基准测试	用户导向测试
目的: 在可控环境下, 对 AI 在上述三个能力维度下的具体技术模块或专项能力进行客观、量化的评估。	目的: 聚焦于最终用户在实际或模拟使用场景中, 对 AI 整体交互表现的主观体验与满意度进行评估。
方法: 通常运用标准化的测试语料、参考答案和客观评价指标(如 ASR 的 WER、TTS 的 MOS 分、响应延迟等)。	方法: 通常依赖真人评估, 采用等级评分法 (RatingScales)、对比选择法 (A/B Testing 或 PairwiseComparison)、用户访谈等。
价值: 主要用于深入了解模型/系统性能、发现技术短板、支持算法优化与选型、进行技术对比。	价值: 主要用于验证产品是否满足用户需求、体验是否良好、是否具有市场竞争力, 并为产品迭代提供直接的用户反馈。

基准测试主要聚焦于 AI 自身核心能力(如理解、表达和交互)的内部量化分析, 为技术优化和系统迭代提供明确的方向。用户导向测试则更关注这些能力在实际应用场景中的综合表现, 以及最终用户的整体体验和满意度。

这两种测试方式相辅相成, 共同构建了对对话式 AI 体验的全面评估体系。

3. “三维二轨”评估框架概览

评估路径	1. 理解能力	2. 表达能力	3. 交互能力
基准测试	听觉感知层（含语音识别，声学/说话人识别，鲁棒性）语言理解层（含核心语义/意图，语用/社交智能，知识运用/推理）	基础清晰度与客观质量（含 TTS-WER, STOI, PESQ, LUFs）- 主观感知质量与表现力（含 MOS, SMOS, 风格/情感评分）	时间响应性（含 E2E/打断延迟）轮次管理与打断（含成功/误打断率）
用户导向测试	字面理解准确度、深层理解与智能、语境感知与适应	表达清晰度、表达自然度、风格匹配度	响应及时性、交互流畅度、交互舒适度

这个框架提供了一个系统性的视角, 帮助我们从多个维度全面评估对话式 AI 的表现。

接下来, 我们将深入探讨每个维度的具体评估方法与指标。

4.1.1 详解“三维二轨”理解能力：语义理解准确率

理解能力是对话式 AI 有效响应的基础, 确保 AI 能够准确、全面地接收和解析用户意图。我们从基准测试和用户导向测试两个方面来系统评估 AI 的理解能力。

1. 基准测试

 听觉感知层：评估 AI 从原始音频信号中准确提取各类基础信息的能力, 及其在复杂环境下的稳定性(鲁棒性)。

- 语音识别 (ASR)：将用户语音转换为文本, 使用词错误率 (WER) 评估。
- 声学事件检测 (AED)：识别非语言声音, 如掌声、警报等。
- 说话人识别：区分或确认说话人身份。
- 鲁棒性测试：在噪声、回声等干扰下的表现。

 语言理解层：评估 AI 结合上下文进行意图、内涵的理解和知识运用、推理的能力。

- 核心语义与意图理解：准确把握用户话语的字面意思和主要目的。
- 语用与社交智能理解：理解语境依赖的意义，包括隐含意图、情感态度。
- 知识运用与推理能力：运用背景知识、进行逻辑推断以深化理解。

2. 用户导向测试

用户导向测试聚焦于用户在实际交互中对 AI 理解能力的主观感受。建议采用真人评估, 围绕以下几个方面展开：

✔ 字面理解难度：AI 是否准确听懂并理解了用户所说的话的直接、表层意思。

💡 深层理解与智能：AI 能否超越字面, 展现出对隐含意图、背景知识的运用和一定程度的推理能力。

🧩 语境感知与适应：AI 能否感知并恰当利用对话历史和当前的交流环境信息来调整其理解 and 行为。

评估通常采用李克特量表 (Likert Scale, 如 1-5 分制) 进行多维度打分, 可辅以 A/B 对比偏好问卷, 并设置开放性评语栏以收集具体的定性反馈。

3. 理解能力评估方法示例

分数	行为描述 (综合评估)	参考用例
5	全面准确理解 (Excellent): 完全理解了用户所说的内容、意图和相关背景, 回应精准且周到。	用户 (语气疲惫讽刺): “哦太棒了, 又来一个会议邀请。” -> AI: “听起来您似乎对这个会议不太期待。需要我帮您查看下日程是否冲突, 或者建议一个婉拒的理由吗?” (识别情绪, 提供解决方案)
4	基本准确理解 (Good): 大致理解了用户的核心意思, 回应整体恰当, 但可能忽略了部分细节或隐含信息。	用户: (同上) -> AI: “好的, 收到一个新的会议邀请。需要我将其添加到您的日历吗?” (处理了事件, 忽略了情绪)
3	理解主要信息 (Fair): 能抓住用户话语的关键点或主要请求, 但对整体语境或细节把握不足, 回应不够精准。	用户: (同上) -> AI: “收到了。关于这个会议邀请, 您需要了解更多信息吗? 比如时间或发起人?” (确认收到事件, 提供通用、略显机械的下一步)
2	理解少量信息/偏差较大 (Poor): 仅理解了少量、孤立的词语, 或对用户意图存在明显误解, 回应不相关。	用户: (同上) -> AI: “会议是多人交流想法的好方式。我们来讨论一下时间管理技巧吧?” (基于“会议”触发无关话题)

分数	行为描述（综合评估）	参考用例
1	完全错误/无法理解 (VeryPoor)：完全未能理解用户输入, 回应完全无关、混乱或直接表示无法理解。	用户：(同上) -> AI：“今天天气怎么样?”（完全无关）

这种评分标准提供了一个清晰的框架, 帮助评估者一致地判断 AI 的理解能力表现。通过具体的用例示范, 评估过程更加客观和可靠。

4.1.2 详解“三维二轨”表达能力：对话连贯性与流畅度

AI 的表达能力, 主要指其生成语音 (TTS - Text-to-Speech) 的质量, 决定了用户接收信息和情感体验的直接感受。

我们同样从基准测试和用户导向测试两个方面来评估。

1. 基准测试

基础语音质量与可懂度：评估生成的语音是否足够清晰可懂, 信号本身是否具备良好的客观质量。

- 合成语音准确度：使用 TTS-WER/CER 评估。
- 合成语音可懂度：使用 STOI 算法预测。
- 合成语音质量：使用 PESQ、SNR、LIFS 等指标。

表现力和风格匹配：通过真人听众的主观判断, 评估合成语音的整体自然度、流畅度、声音表现力以及特定音色是否令人满意。

- 整体自然度：使用平均意见分 (MOS) 评估。
- 风格/情感表现力：评估准确性、自然度、恰当性。
- 音色相似度：当需要模仿特定发音人时使用。

2. 用户导向测试

用户导向测试聚焦于用户在实际交互中对 AI 表达方式 (声音、语气、风格等) 的整体主观感受。

 **表达清晰度：**AI 是否能准确地发音, 清晰地表达出想表达的内容。

 表达自然度：AI 的声音是否足够自然，足够拟人，蕴含各种韵律细节。

 风格匹配度：AI 说话的风格、韵律和所对应的情绪反馈是否匹配当前的对话场景和用户对其角色的期望。

评估同样采用李克特量表进行多维度打分, 辅以 A/B 对比和开放性评语收集。

3. 表达能力评分标准示例

分数	行为描述（综合评估）	参考用例
5	表现力极佳（Excellent）：发音清晰精准，表达自然流畅（超拟人），风格/情感完美契合语境。	用户（沮丧地）：“唉，这个拼图太难了，我弄不好！真没意思！” -> AI（语速特意放缓，声音带真挚理解和温和鼓励）：“别着急，这个拼图看起来确实挺复杂的。有时候退一步看看整体需要我陪你一起看看吗？”
4	表现力良好（Good）：发音清晰，表达整体自然，风格/情感基本恰当，但偶有细微不自然之处。	用户：（同上） -> AI（声音清晰自然，语速正常，但情感色彩稍弱，语调起伏不大）：“嗯，看起来遇到困难了。这个确实不容易。你可以先休息一下再试试看。”
3	表现力中等（Fair）：发音基本清晰，但自然度或风格/情感匹配度有明显不足，有时显生硬。	用户：（同上） -> AI（发音清晰，但语气和节奏没有变化，对用户的沮丧没有回应）：“嗯，看起来遇到困难了。你可以先休息一下再试试看。”
2	表现力较差（Poor）：发音清晰度或自然度存在明显问题（如机器感重、音质差），风格/情感通常不匹配。	用户：（同上） -> AI（声音清晰度但是完全没有匹配用户的心情）：“你可以先休息一下再试试看。”
1	表现力极差（Very Poor）：发音不清晰难以理解，或表达方式极不自然/完全错误。非常不自然，甚至不能清晰的表达内容。	用户：（同上） -> AI（声音含混不清部分词语难以辨认）：“好的，可-以-休-息-下-。”

这种评分标准帮助评估者系统地判断 AI 的表达能力,从发音清晰度到情感表达的匹配度,全面衡量 AI 声音输出的质量。

4.1.3 详解“三维二轨”交互能力：响应时间与实时性

交互质量关乎用户与 AI 交互过程本身的动态表现,如响应速度、对话流畅度、打断处理等,直接影响沟通效率和用户的整体舒适感与掌控感。我们同样从基准测试和用户导向测试两个方面来评估。

1. 基准测试

时间响应性：AI 响应用户语音输入和用户打断的速度。

- 端到端响应延迟：从用户语音结束到 AI 响应开始的时间。
- 打断响应延迟：从检测到用户打断到 AI 停止输出的时间。

轮次管理与打断处理：评估对话双方轮流发言的协调性、自然性,以及打断机制的有效性与准确性。

- 用户侧打断成功率：用户能否在需要时成功打断 AI。
- AI 侧误打断率：AI 是否会错误地打断用户。
- 用户侧误打断率：非语义声音是否被错误识别为打断。

2. 用户导向测试

用户导向测试是评估交互体验最直接、最重要的方式,因为它直接反映了用户在实际使用中的主观感受。

响应及时性：用户对 AI 响应速度、等待时间以及整体交互节奏的主观感受。

交互流畅度：用户感受到的对话流转是否自然、协调,有无不适的打断或停顿。

交互舒适度：用户在交互过程中所需付出的注意力和听觉努力程度,以及整体感受是否舒适自然、行为一致性高。

评估同样采用李克特量表进行多维度打分,辅以 A/B 对比和开放性评语收集。

3. 交互能力评分标准示例

分数	行为描述（综合评估）	典型例子（场景：用户在听 AI 主持的互动播客时进行互动）
5	交互性极佳（Excellent）：响	AI（流畅主持）：“关键在于创新思维。”

	应迅速, 流程极其顺畅自然(如面对面对话), 打断处理得体, 交互轻松舒适且一致。	-> 用户(几乎同时插入): “那具体怎么创新呢?” -> AI (无缝暂停并回应): “很好的问题! 比如可以尝试跨界合作, 或者” (响应极快, 打断自然流畅如真人对话)
4	交互性良好 (Good): 响应及时, 流程大体流畅, 打断处理基本得当, 偶有轻微延迟或不协调。	AI (主持): “关键在于创新思维。” -> 用户(稍等 AI 句末): “那具体怎么创新呢?” -> AI (快速回应): “创新可以从跨界合作、技术突破等方面入手” (响应及时, 交互顺畅, 可能打断处理不如 5 分完美但体验良好)
3	交互性中等 (Fair): 响应速度尚可但偶有延迟, 或流程中出现不自然停顿/打断, 需要用户额外耐心	AI (主持): “关键在于创新思维, 这需要我们”-> 用户(试图打断): “那具体怎么” -> AI (继续说了几个字才停顿, 然后回应): “打破常规。您是问创新的具体方法吗? 可以考虑跨界合作” (有可感知的打断延迟或偶尔的小卡顿, 流畅度受影响)
2	交互性较差 (Poor): 响应经常缓慢或打断处理不当, 流程卡顿, 让用户感到不适或需额外努力。	AI (主持): “关键在于创新思维。” -> 用户: “那具体怎么创新呢?” -> AI (长时间停顿或延迟): “(数秒后) 创新的方法有很多, 比如跨界合作” (存在明显的高延迟, 让用户不适) 或 AI (用户说话时突然插入): “关键在于 (AI 突然插话) 您是指技术创新吗?” (发生不当的 AI 侧误打断)
1	交互性极差 (Very Poor): 响应极慢/无响应或交互混乱, 无法有效进行对话, 用户非常不满意。	AI (主持): “关键在于创新思维。” -> 用户: “那具体怎么创新呢?” -> AI: (长时间无响应, 或反复错误打断用户, 或反复要求用户重复, 导致对话基本无法有效进行) 例如: AI: “抱歉, 我没听清, 能再说一遍吗?” (在用户清晰表达后反复出现)

这种评分标准帮助评估者系统地判断 AI 的交互能力,从响应速度到打断处理的自然度,全面衡量对话过程的流畅性和舒适度。

4.1.4 落地“三维二轨”:应用框架到你的对话式 AI 产品

“三维二轨”框架旨在为对话式 AI 体验评估提供了一套系统性的分析视角。在实际落地应用里,我们需要结合具体场景的目标来设计具体落地的评测计划。本章旨在通过典型示例,阐述如何应用该框架,将理论转化为可操作的评估实践。

场景分析与目标确立

明确您的对话式 AI 产品主要服务于什么场景?为用户解决的核心问题或提供的核心价值是什么?是高效的提供有效信息,还是有效的提供情绪价值?

评价维度与优先级判定

基于场景分析和评估目标,确定“理解(U)”、“表达(E)”、“交互(I)”三个核心能力维度中,哪些对实现产品核心价值最为关键,并据此确定评估资源的投入优先级。

指标精选与测试设计

围绕确定的关键维度,从详述的指标库中精选最能反映核心问题的指标,并设计相应的测试方案(基准测试和/或用户导向测试)。

结果解读、行动与迭代

结合基准测试的客观数据与用户导向测试的主观反馈进行分析。尝试将评估发现的问题与技术根因关联,指导优化方向。

评估的价值在于后续行动。将评估融入产品迭代流程,优化后再评估,形成持续改进的闭环。考虑将部分自动化指标纳入日常监控。

4.1.5 场景示例:智能点餐与老年人陪伴

1. 场景示例:智能点餐

场景与价值

用户在餐厅或通过 App/电话进行点餐,核心价值在于快速、准确地完成订单,提升效率。

评估目标

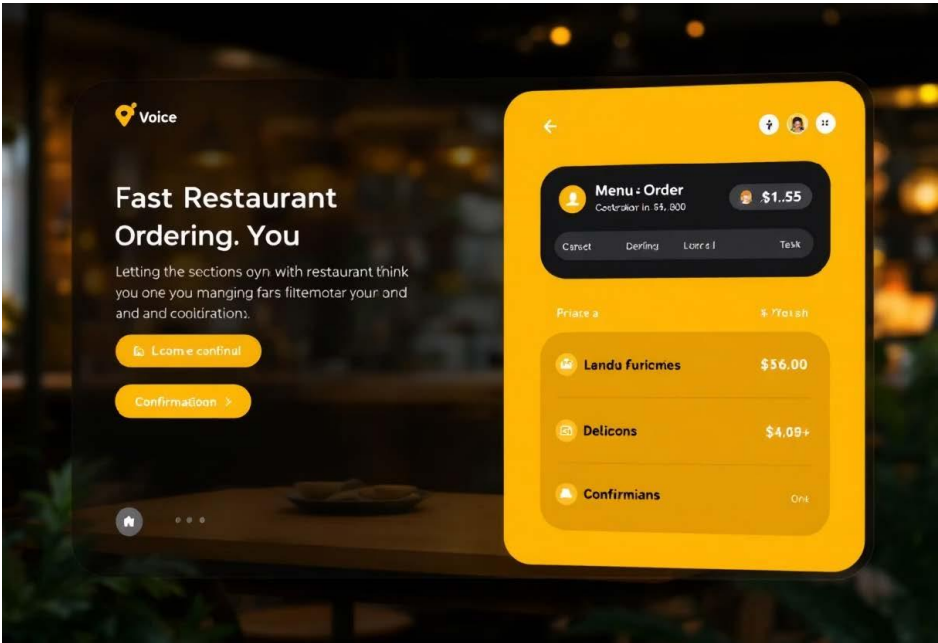
验证核心点餐流程的准确率和交互效率是否达到上线标准;可能需要考虑特定领域 ASR

准确率等评价细项。

优先级

理解能力(U) > 交互质量(I) > 表达能力(E)

理由：必须先准确理解用户点单内容（U）；交互过程需迅速流畅, 避免用户等待或重复（I）；声音表达清晰即可（E）。



重点关注

- U 维度中的听觉感知鲁棒性和核心语义与意图理解。
- I 维度中的时间响应性。

测试设计

- 基准测试：在典型餐厅环境噪声下测试 WER 和槽位填充 F1；测量端到端 P90/P95 延迟。
- 用户导向测试：设计核心点餐任务, 测量任务成功率；评估用户感知的理解准确度和响应及时性。

2. 场景示例：老年人陪伴

场景与价值

为独居老人提供日常陪伴、情感支持和安全监护, 核心价值在于建立情感连接、提供温暖舒适的互动体验。

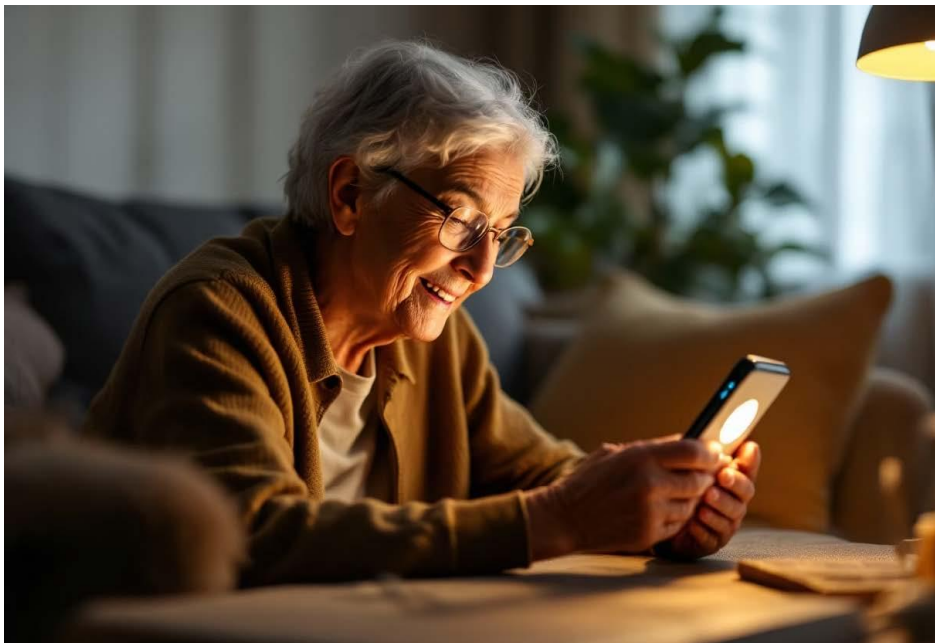
评估目标

评估 AI 的共情表达能力和对话自然度是否符合用户预期;验证 AI 能否进行有意义的、长期的、上下文连贯的交流。

优先级

表达能力 (E) > 理解能力 (U) $\approx$ 交互质量 (I)

理由: 声音和表达方式是否温暖、自然、有情感,且语速、音量是否适合老年用户 (E) 显著影响陪伴体验;能够理解老人的话语及隐含情感 (U) 是有效互动的基础;交互的流畅舒适比绝对速度更重要 (I)。



重点关注

- E 维度中的主观感知质量与表现力(需考虑语速、音量等风格匹配度)
- U 维度中的语用与社交智能理解和用户导向的语境感知
- I 维度中的用户导向的交互舒适度

测试设计

- 基准测试: 重点进行主观 MOS 测试评估声音自然度;测试风格/情感表达准确性;评估语速、音量是否在适老化范围。
- 用户导向测试: 设计开放式对话场景,评估表达自然度、风格匹配度、交互舒适度;结合访谈了解 AI 的共情能力。

结果解读与行动建议

评估的价值在于后续行动。以下是如何有效利用评估结果的建议：

综合分析

结合基准测试的客观数据与用户导向测试的主观反馈进行分析。例如, 用户抱怨“反应慢”是否真的对应高延迟指标? 用户觉得“声音假”具体是 MOS 分低还是风格不匹配?

关联诊断

尝试将评估发现的问题与技术根因关联, 指导优化方向。例如, 高 WER 可能需要优化 ASR 模型, 低 MOS 可能需要优化 TTS 声学模型或韵律模型, 频繁误打断可能需要调整 VAD 或交互策略。

供应商选型

如果评估目的是选择技术供应商, 可以将筛选出的关键维度和指标作为核心考察点, 进行横向对比, 选择在你最看重的方面表现最优的方案。

迭代闭环

将评估融入产品迭代流程, 优化后再评估, 形成持续改进的闭环。考虑将部分自动化指标纳入日常监控。

无论基准测试结果如何, 最终的用户导向测试结果 (尤其是用户的整体满意度、使用意愿等) 是检验对话式 AI 产品是否成功的“黄金标准”。

本章节以评估框架为切入点, 系统梳理了影响人机对话质量的关键因素, 并探讨了相应的评价思路。需要说明的是, 这里并非严格的学术论文, 因此未对每一项定义进行严谨界定, 比如理解、表达与交互能力在实际应用中往往彼此关联、相互影响, 而非完全独立。视觉、身体语言等多模态信息同样对对话体验具有重要作用, 本文也只聚焦于语音维度展开。本文的出发点是结合当前主流技术能力与评测体系, 帮助读者一起梳理出一份更具实际参考价值的对话式 AI 评估框架指南, 从而可以更好的构建和迭代对话式 AI 产品。

在这个迎接语音智能体新物种的新时代, 很多探索还在早期。我们希望这套“三维二轨”的评估框架能够起到抛砖引玉的作用, 让对话式 AI 应用的开发者和行业伙伴, 更多的关注和思考人机对话的体验细节和价值。无论是从帮助人们高效获取信息, 还是带来更好的情感体验的角度, 一起构建出安全可靠又有温度的对话式 AI 产品, 来帮助我们每个人实现自我

提升与治愈，共同迈向一个人机和谐共生的未来。

4.2 用户体验要素

对话式 AI 的核心价值在于其与用户交互的能力，而用户体验 (UX) 是衡量其成功与否的终极标尺。优秀的用户体验应贯穿交互的始终，并体现在多个维度上。本小节将聚焦于以下四个关键用户体验要素：

4.2.1 交互友好性

指用户在与对话式 AI 交互过程中感受到的顺畅、自然、高效和愉悦的程度。它直接决定了用户是否愿意持续使用以及推荐给他人。这其中涉及到 AI 的理解精准度、AI 响应的流畅度、智能打断、上下文感知与连贯性等一系列关键因素。

- **AI 理解精准度：**这主要取决于 AI 大模型能否准确识别用户的意图，尤其在口语化或者模糊表达时，AI 能否精准理解用户的语言，并生成符合用户预期的回答，以及 AI 能否识别用户情绪（如沮丧、兴奋）并做出适当回应（如安抚、祝贺），这样不仅提升了 AI 的理解精准度，还提升了交互的温度和人性化。
- **AI 响应流畅度：**在与 AI 的语音交互中，AI 能否做到快速并且自然、流畅的响应至关重要，从体验上看，在人与 AI 的互动中，1s 左右的响应延迟能更接近人与人之间的实时对话。
- **AI 智能打断：**在人类真实语音对话中，打断对方并提出疑问是常见现象，在对话式 AI 场景，能否支持随时打断也成为衡量 Agent 智能化以及交互友好性的重要指标。
- **上下文感知与连贯性：**在情感陪伴、办公助手等场景中，人与 AI 的对话会出现对话轮次多、对话时间长等特点，这其中 AI 能否理解并记住对话上下文，进行多轮连贯对话，避免用户反复重申信息，也是影响 AI 交友友好性的关键因素。

4.2.2 环境适应性

指对话式 AI 在不同物理环境和社会场景下保持良好交互表现的能力，在实际的应用场景中，物理与网络环境是复杂多变的，系统需具备鲁棒性以应对干扰，以确保对话式 AI 在用户实际使用场景（而非理想实验室环境）中可靠可用，扩大其应用范围。

- **噪声抗干扰性：**在对话式 AI 中，一旦面临嘈杂环境，周围人说话的环境人声、设备运行声等背景噪音会干扰语音的识别，还可能会打断人与 AI 的对话，最终导致语音识别的准确率下降，交互体验大打折扣，这需要 RTC 具备出色的降噪能力。
- **网络适应性：**在人与 AI 的对话中同样会面临当网络不佳时出现卡顿、延迟高的现象。这其中需要 RTC 技术具备抗弱网、抗丢包能力，即使在网络不稳定的环境下，核心对话功能仍能保持较为流畅。

4.2.3 设备适应性

指对话式 AI 在不同类型、尺寸和能力的终端设备（如智能手机、智能音箱、车载系统、智能电视、可穿戴设备）上提供一致且优化的交互体验的能力，让用户不管是在高端机还是低端机，都能体验到丝滑、流畅的对话式 AI 服务。

- **性能优化：**一方面，手机作为 AI 落地的核心载体，不同的手机在 CPU、内存、GPU 性能方面表现各异，对语音的处理效果也不一样，RTC 服务商需要具备丰富的终端适配能力，评估低端设备上的响应延迟和计算资源占用，以保障在各类终端都能流畅运行。
- **功耗控制：**同时在 AI 手表、眼镜等可穿戴设备中，对话式 AI 的功耗控制也十分重要，例如 RTC 的 SDK 需要做到包体小、功耗低，保障设备的长续航。

4.2.4 可靠性与稳定性

在对话式 AI 中，系统持续、无差错地提供服务的能力，以及在面对压力、故障或变化时维持服务水平的韧性也是影响用户交互体验的重要因素。

- **服务可用性：**对话式 AI 服务提供者需要保障系统整体可用的时间比例，例如 SLA（服务等级协议）可用性需要达到 99.99%。在部分组件故障、数据中心问题或网络中断时，系统是否具备容错机制和灾备恢复能力，这是决定服务可用性的关键因素。
- **性能稳定性：**在高峰负载的情况下，对话式 AI 的系统响应时间（如语音识别延迟、回复生成时间）能否保持稳定在可接受范围内，并且进一步评估延迟的 P95/P99 值及波动情况。同时，资源占用稳定性也不可忽视，当系统运行时，对 CPU、内存、网络带宽的占用是否平稳，要避免剧烈波动导致卡顿或崩溃。

用户体验是对话式 AI 成功落地的核心。交互友好性确保对话顺畅高效，环境适应性保障系统在复杂现实场景中可用，设备适应性实现跨终端的无缝体验，而可靠性与稳定性则是用户持续信任和依赖的基础。这四个维度相互关联，共同构成了保障对话式 AI 用户体验质量的核心框架。在设计和评估过程中，需要综合运用技术指标（如准确率、延迟、错误率）和用户主观反馈（如满意度调查、可用性测试）进行全面衡量，并持续迭代优化。

4.3 AI 模型评测平台（对话式）

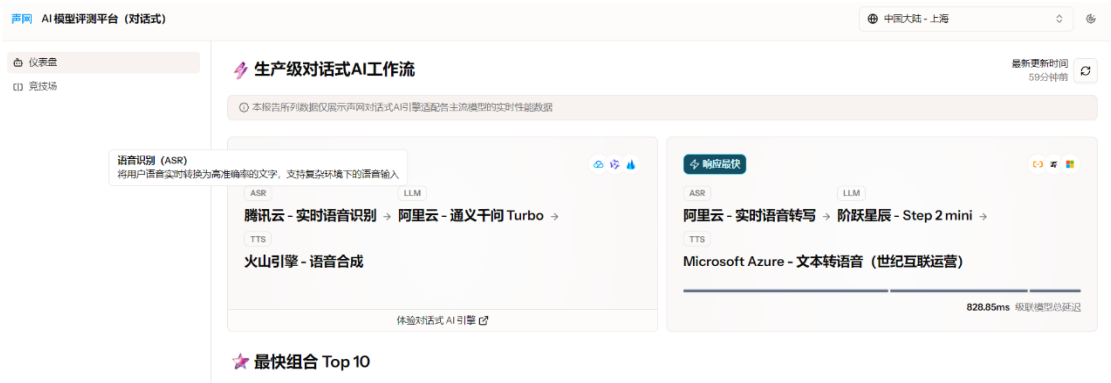
级联大模型凭借独特的 ASR+LLM+TTS 的三段式设计，成为了 AI 创业者开发 AI 应用/硬件的主流方案。声网对话式 AI 引擎也是基于级联大模型的架构设计，并通过灵活扩展的特性兼容适配全球主流的大模型与 ASR、TTS 供应商，满足不同场景和业务需求。在今年 7

月，声网发布了全球首个 AI 模型评测平台（对话式）。该平台针对对话式 AI 场景下，级联大模型中的 ASR、LLM、TTS 的延迟数据、词错误率、字母数字性能等指标提供主流供应商的横向测评，更直观的展示声网对话式 AI 引擎适配各主流模型的实时性能数据。同时，平台还提供了“竞技场”功能，开发者可自由选择 ASR、LLM、TTS 的主流供应商进行对比，根据性能的数据表现，选择更适配自身业务的模型供应商。

4.3.1 “仪表盘”：综合最优、响应最快模型一目了然

AI 模型评测平台主要分为“仪表盘”与“竞技场”两项功能，在“仪表盘”中官方推荐了对话式 AI 引擎适配后综合最优、响应最快的级联模型组合。测试区域覆盖了上海、新加坡、日本、洛杉矶、法兰克福等 10 个全球主流城市。测试的模型也十分丰富，LLM 大模型包含了 DeepSeek V3、豆包大模型、智谱 GLM 系列模型、通义千问系列模型、MiniMax Text 01 以及 Llama 3.3 70B、GPT 4.1 mini、Gemini 2.0 Flash、Claude Haiku 3.5 等国内外主流的大模型。ASR 与 TTS 也覆盖了 阿里云、腾讯云、讯飞、OpenAI、Microsoft Azure、Cartesia 等知名 AI 企业旗下的主流模型。

例如，截止 2025 年 9 月 23 日 16:30 分，中国大陆-上海区域综合最优推荐级联模型组合为腾讯云-实时语音识别+阿里云-通义千问 Turbo+火山引擎-语音合成。响应最快推荐级联模型组合为阿里云-实时语音转写+阶跃星辰 - Step 2 mini + Microsoft Azure - 文本转语音（世纪互联运营），级联模型总延迟为 828.85 ms。



同时，“仪表盘”针对对话式 AI 场景下级联模型的延迟数据、词错误率、字母数字性能的数据表现，定制了多个排行榜，且评测数据保持每小时更新：

1、级联模型组合评测

以级联模型总延迟为评测指标进行对比 以末字延迟、词错误率为评测指标进行对比。

2、语音识别 (ASR) 评测

以末字延迟、词错误率为评测指标进行对比。

3、语言模型 (LLM) 评测

以首字延迟的性能数据进行排序。

4、语音合成(TTS) 评测

对比 TTS 模型的首字节延迟、词错误率、字母数字性能。

ASR 模型的词错误率检测基于交互式测试方法(Turn Detect)，以 300ms 的超时判定模拟用户的感知阈值，旨在评估最佳对话体验下 ASR 模型的实际可用性。在测试中声网也发现，在对话式场景下，ASR 模型需要在实时性和准确性之间做出更精细的权衡，一些在非对话式场景下准确率*高的模型(如 OpenAI 系列)并不一定是对话式场景的最佳选择。

TTS 模型的评测包含了中英文场景下首字节延迟、词错误率(WER)和字母数字性能(AP)三项指标的对比，词错误率体现了 TTS 模型在通用场景下的表现，数值越低，TTS 准确性越高。字母数字性能通过统计 TTS 模块对包含多音字、数字、缩写、公式、标点符号的复杂文本中每个字符的正确生成比例，全面衡量其对多类型特殊字符的精准处理能力，数值越高，TTS 准确性越高。通过这三类指标的评测让 TTS 语音合成的准确性进一步被量化，技术对比更具参考价值。

4.3.2 “竞技场”：自主选择模型性能对比

在 AI 模型评测平台“竞技场”中，开发者可自主选择不同的 ASR、LLM、TTS 模型进行首字延迟、末字延迟、词错误率、字母数字性能等多项性能的对比。测试区域同样覆盖上海、洛杉矶等全球十大核心城市。

针对延迟数据的评测指标，平台还提供了多个分位的延迟数据差异对比，从 P25、P50 到 P99 共六个分位，开发者可以更详细的了解每个模型的延迟数据表现。例如，凤鸣-ASR P50 分位的末字延迟为 572 ms，意味着测试期间 50%的延迟数据低于 572 ms。

此外，“竞技场”的 TTS-语音合成对比中，不仅提供了预设的语音合成测试语句，反映多种场景下（字母数字混合、非流畅性语句、客户服务、医疗健康、外呼销售、有声书&播客、非常见发音词汇）不同模型的语音合成质量，并支持试听。还支持用户自主输入文本内容，一键生成语音合成效果的对比，并支持中英文。通过个性化试用场景，帮助开发者快速验证技术与实际需求的匹配度，让选型决策更高效。

相比于“三维二轨”的对话式 AI 性能指标体系，AI 模型评测平台直接为开发者提供了实时评测级联模型各项体验性能的平台，并支持自定义对比。但需要注意的是，该平台所列的各项数据仅展示声网对话式 AI 引擎适配各主流模型的实时性能数据，主要针对选择声网对话式 AI 引擎的开发者可以在此基础上选择更适配自身业务的最优对话式 AI 模型组合。

第五章：对话式 AI 的应用实践

对话式 AI 的爆发开启了人与 AI 互动的新纪元，对硬件、教育、社交等行业带来了深远的影响，对话式 AI 在社交中提供情感陪伴，在线社交市场涌现出了一大批 AI 情感陪伴 App；在教育中实现个性化辅导，AI 助教能够在课堂答疑、课后辅导、实时判卷等场景中与学生自然、流畅的对话；在智能终端上实现无缝控制，AI 陪伴硬件有望成为 Labubu 后的下一个风口；在客服场景中大幅提升效率与体验。这些应用场景的落地实践不仅验证了对话式 AI 技术正在不断走向成熟，更体现了其解决人机交互实际痛点、提升用户体验和驱动 AI 业务创新的核心价值。

本章将通过应用场景热力榜单、行业发展趋势、实践案例分析等展现对话式 AI 如何助力各类 AI 场景落地实践，并成为推动各行业发展的力量。

5.1 对话式 AI 场景热力榜单



5.2 社交领域的 AI 场景探索如火如荼

在社交中，对话式 AI 提供情感陪伴，从 2013 年上映的电影《Her》中的虚拟女友萨曼莎到 2025 年埃隆·马斯克的 xAI 公司宣布推出二次元 AI 女友，AI 伴侣已经从电影走进现实。这一转变不仅是技术的驱动，更是社交需求在科技催化下的全新表达。如今，AI 伴侣类应用在年轻群体尤其是大学生和年轻职场人士中广泛渗透，部分用户与 AI 伴侣建立起长期情感联系，并愿意付费订阅以提升互动体验。

社交应用一直是年轻用户的聚集地，也有成为新流量入口的潜质，而生成式 AI 的爆火，也为原本沉寂的在线社交市场带来了新的想象力。在全球著名投资基金、咨询公司 Andreessen Horowitz 发布的 2025 年全球 100 大生成式 AI 应用排行榜中，仅次于 ChatGPT 和 DeepSeek 的第三名，正是来自 AI 社交领域的 Character.AI。

The Top 50 Gen AI Web Products, by Unique Monthly Visits				
1.  ChatGPT	11.  Moonshot AI	21.  CIVITAI	31.  Photogram	41.  Monica
2.  deepseek	12.  Hailuo AI	22.  ElevenLabs	32.  Moescape AI	42.  CURSOR
3.  character.ai	13.  Hugging Face	23.  Sora	33.  Midjourney	43.  Ideogram
4.  perplexity	14.  Poe	24.  Crushon AI	34.  candy.ai	44.  GMB
5.  JanitorAI	15.  Adot	25.  BLACKBOX AI	35.  zeemo	45.  Clipchamp
6.  Claude	16.  Eden AI	26.  DeepAI	36.  VEED	46.  Meta AI
7.  QuillBot	17.  PolyBuzz	27.  Gamma	37.  Invideo AI	47.  StudyX
8.  SUNO	18.  SERART.AI	28.  Leonardo AI	38.  Pixelcut	48.  bolt
9.  SPICYCHAT.AI	19.  liner	29.  cutout.pro	39.  talkie	49.  PicWish
10.  Doubao	20.  KLING AI	30.  BRAINLY	40.  PixAI	50.  Joyland

Source: Similarweb, January 2025
Charts are for informational purposes only. Past performance is not indicative of future results. None of the above should be taken as investment advice. See e107 securities disclosures.

 Aló Consumer

Character.AI 利用大模型生成各种人物和角色风格的对话让人充满了新奇感。而在国内 AI 社交领域，MiniMax 旗下的星野、字节的猫箱、阅文集团的筑梦岛等是最早爆火的一批 AI 社交应用，此后，随着多模态融合、情感识别与共情模拟以及语音交互等技术的不断成熟，不仅涌现出了一批 AI 社交新秀，京东、美团、百度等巨头也都纷纷入局，例如美团推出的 AI 社交产品 Wow、京东推出的“他她它”、“宠 TA”、百度的“月匣”等，再加上老牌社交应用陌陌、Soul 等产品，AI 社交市场的发展涌现出了一波新的热度。

5.2.1 从机械应答到“类人化”社交引擎

传统的 AI 社交受限于文本交互与延迟响应，交互体验生硬，且经常陷入一问一答的机

械感。而伴随着生成式 AI 的爆火，人与 AI 的交互也迎来新的突破，低延时语音交互、全双工语音交互、情感计算与多模态融合等技术革新推动 AI 社交从机械应答走向“类人化”社交。

低延时语音交互：当下大模型的实时语音交互主要分为级联方案与端到端方案，级联方案通过 ASR+LLM+TTS 三步骤，将用户的语音转换成文字，输入给大模型，再将生成的文字回答转换为语音，输出给用户，这其中每一个步骤都会产生一定的延迟，AI 整体的响应延迟能达到 3 秒左右。而端到端方案则是直接将语音实时输入给大模型，再直接生成语音，不仅降低了延迟，还能捕捉到更多的信息。

端到端方案虽然对话延时更低、模型性能上限更高，但也存在着训练成本高昂、系统复杂、黑盒透明度低等问题，并不是所有对话式 AI 场景都适合端到端方案。反观级联方案，成本可控、模块化易于调试和维护等优势，受到了很多开发者的青睐，同时，随着声网与模型厂商对级联方案延时的优化，ASR+LLM+TTS 的整体延时可以有效降低至 1s 左右，推动了 AI 陪伴社交等对话式 AI 场景的加速落地实践。

全双工语音交互：全双工语音模型的出现，也让 AI 社交更具“真人感”。例如，社交平台 Soul 自研的全双工语音通话大模型，摒弃了传统语音交互中依赖的 VAD（语音激活检测）机制，打破行业中普遍存在的“轮次对话”模式，赋予 AI 自主决策对话节奏的能力。AI 可实现主动打破沉默、适时打断用户、边听边说、时间语义感知、并行发言讨论等。甚至还模拟“咳嗽”、“结巴”等真人表达细节，让 人与 AI 的对话更具“真人感”，支持打造更沉浸、类现实交互的语音互动新体验。

情感识别与共情：随着大模型的进化，很多新一代模型可识别数十种细微情绪，通过语音语调、微表情（结合摄像头）实时生成共情回应。例如 AI 社交伴侣通过识别用户喜怒哀乐的语气，做出不同的贴心回应，陪伴机器人能根据用户心率变化调整安慰策略。

多模态人格构建：豆包、通义等越来越多的大模型支持文本、语音、图像统一生成，用户可自定义角色外貌、性格、记忆轨迹，形成“成长型关系”。这也有效的推动用户更愿意与 AI 进行更深入的对话。

5.2.2 星野案例分析：情感陪伴与 UGC 生态的深度融合

星野是 MiniMax 基于通用大模型打造的 AI 原生应用，致力于为用户打造一个沉浸式 AI 内容社区，这是一个融合多模态交互、UGC 创作与情感陪伴的沉浸式社交平台。其核心玩法围绕“智能体”（用户创建的 AI 角色）展开，通过高度自定义角色+深度情感互动，满足用户对陪伴、创作和社交的需求。

1. 智能体创建：用户成为“造梦师”

用户可定义角色的形象（上传图片生成动漫形象）、性格、声线（70+音色混搭）、背景

故事，甚至通过“对话样本”训练专属语言风格。还能通过文本描述快速生成角色，支持“重说”“改写”功能实时调整对话走向，降低创作门槛。

2. 沉浸式交互：情感计算与多模态融合

除了接入 MiniMax 的通用大模型外，星野也接入了 DeepSeek R1 模型，支持“内心戏”功能（如角色表面克制内心醋意），增强情绪层次感。同时，星野还打造了丰富的多模态体验：语音对话（支持自定义声线）、星念（AI 生成的“记忆相片”）、群聊模式（多角色互动），构建“文本-语音-图像”闭环。

3. UGC 社区吸引用户共创

星念交易市场：用户可生成并出售角色“星念”（记忆卡牌），创作者获 2% 分成，形成创作-消费闭环。社交裂变：公开角色可被其他用户互动，头部角色（如“莫轻舞”）吸引 12.4 万人参与共创衍生故事。

在互动对话体验方面，基于 MiniMax 在文本、声音、视觉模型下的先进能力，以及声网的对话式 AI 技术，在星野，用户与 AI 智能体可实现秒回、清晰、流畅的实时语音对话，再搭配超拟人的 AI 音色，让 AI 交互更接近人类的真实对话，成为用户娱乐、工作、学习中的 AI 伙伴。

5.3 AI 陪伴硬件迎来风口

2025 年开年，以 AI 玩具为代表的 AI 情感陪伴类硬件迎来风口。在 CES、MWC 等国际展会上，AI 玩具成为焦点赢得广泛关注。有数据预测，到 2028 年，国内 AI 玩具市场规模有望达到 300 亿至 400 亿元人民币。

随着 AI 玩具走红，玩具企业订单呈现爆发式增长。东莞某潮玩公司曾表示，在给潮玩毛绒玩具增加了 AI 功能后，订单已经超过 200 万只。

情绪价值已被视作 AI 玩具的核心竞争力，但在当下产品功能日趋同质化的市场环境中，如何从用户体验层面深度挖掘并提升情绪价值，已然成为行业亟需攻克的关键命题。

目前主流的 AI 玩具主要依赖语音交互，但语音识别准确率受环境噪音、孩子发音不标准等因素影响较大，导致交互过程频繁出错，破坏了用户体验。虽然许多 AI 玩具背后都配备了大模型能力，但互动模式相对单一，体验差强人意。无论是需要手动按键还是依赖特定的唤醒词，激活语音交互功能，都可能会打断自然流畅的互动节奏。

同时，在对话过程中，大多数 AI 玩具缺乏打断机制。当用户想要急切表达自己的想法，或纠正玩具给出的错误回答时，只能被迫等待机器冗长的语音输出结束，才能继续说话。

这些体验都限制了作为 AI 玩具本应具有的趣味性和实用性，使其难以真正融入人们的生活，成为理想的智能陪伴伙伴。

5.3.1 实现真正的情感陪伴，交互体验是关键

要让 AI 陪伴硬件产生真正的情绪价值，首先需要构建自然流畅的交互体验。这意味着 AI 需要能够准确识别语音指令，并理解指令背后的情感和意图，即使在嘈杂环境中，也能快速给予正确的反馈，让对话如同人与人面对面交流般流畅、生动。

现实中，人和人当面交流响应时间约 200ms，远程沟通约 400ms，在 AI 陪伴硬件中，当 AI 的响应延时做到 1s 左右时，被视为接近真人聊天对话速度，有效提升了交互的自然度与流畅性。同时，AI 硬件还需要具备精准的语音识别能力，即便在复杂环境下，也能准确获取用户指令，有效屏蔽 95% 的环境人声及噪声干扰。

多模态交互的深度融合也同样重要。人类沟通是多模态的，除了语音交互，AI 陪伴硬件在视觉、触觉等交互方式应协同发挥作用，才能更全面的穿搭情感和意图，提供全方位、沉浸式的互动体验。

个性化定制的互动同样不可或缺。每个个体都是独一无二的，他们的兴趣爱好、性格特点各不相同。在角色设定上，需要进行更加灵活的选择，尤其是 IP 类玩具，需要精准复刻其独特的声线、语调和发音习惯，并且能够对角色情绪进行生动演绎，让虚拟角色成为真正的生活伙伴。

5.3.2 AI 陪伴硬件，不止于儿童

不仅仅是面向儿童的 AI 玩具，AI 陪伴硬件在成人情感陪伴以及老年人情感关怀等细分赛道也都展现出巨大潜力，具体场景有以下几种分类。

儿童陪伴

AI 玩具

AI 玩具是目前最热门的赛道。通过拟人、拟动物或拟 IP 的方式，与儿童进行视、听、触多维度交互，提供娱乐和陪伴的同时，也能起到一定的教育作用。

AI 早教机器人

在互动基础上加入语言、数学、编程等知识，结合语音、图像和触觉反馈，增强孩子的学习体验，提高知识吸收效率。



成人陪伴

AI 助手

通常具备语音交互功能，帮助用户查询信息、设置提醒、控制智能家居设备等，并提供个性化的服务和建议，一定程度上缓解人们的生活压力，提供便捷和舒适的生活体验。

AI 虚拟宠物

以虚拟形式存在，通常具有可爱的形象和丰富的互动功能，能模拟真实宠物的一些行为和情感反应，如陪伴主人玩耍、对主人的情绪做出回应等。





伴随对话式 AI 的快速发展，AI 陪伴不再仅仅是一个套着 AI 外壳的普通玩具或硬件，而是能够真正走进人们内心，成为情感交流、学习成长、互动陪伴的好伙伴，为人们带来更多充满乐趣和价值的互动体验。

5.3.3 案例分享：珞博智能 AI 宠物“芙崽 Fuzozo”

芙崽 Fuzozo 是珞博智能推出的一款专注于为 Z 世代人群提供情感陪伴的 AI 养成系潮玩，具备毛绒包挂的形态，方便随身携带，通过多模态交互与类宠物养成玩法，与用户建立长期情感羁绊，舒缓孤独感与焦虑感。同时，产品还有社交玩法，可以通过“碰一碰”交友，方便年轻人在朋友圈迅速社交裂变。

珞博智能自主研发的多模态情感模型“MEM”，通过每日的互动陪伴，培养用户与芙崽的亲密度，养成芙崽的多样化性格，让每一只芙崽都与用户形成独一无二的专属羁绊。

珞博智能的另一项核心技术“EchoChain 仿生记忆系统”，则赋予芙崽长期记忆与学习的能力，可以记住用户的身份、特征、喜好、事件等，并通过自学习能力对记忆进行反思，逐步成长为更懂用户的同频伙伴。

通过与声网的合作，大幅优化了芙崽 AI 交互的实时性和稳定性，使得这款 AI 陪伴机器人能够进行自然流畅的对话，提供实时情感陪伴和个性化交互，从而为用户带来全新的陪伴体验。

同时，声网对话式 AI 的端到端解决方案，从软硬件全方位为珞博智能提供支持。在软件层面，声网提供了对话式 AI 开发套件。其中，先进的语音活动检测 (VAD) 技术能够精准识别

语音信号,有效降低背景噪音的干扰,从而确保语音识别的高准确率。实时语音合成功能实现了快速响应,让交互更加流畅自然。智能打断处理技术则赋予了设备灵活的对话能力,能够根据用户的表达实时调整,极大地提升了交互的适应性与流畅度,告别“机械式”应答,助力芙崽实现更流畅、更迅速的交互体验。在硬件层面,声网的支持涵盖芯片选型、功耗设计、震动马达等关键环节,确保硬件性能与软件功能的高度匹配。



5.4 百镜大战开启，下一代个人智能设备入口？

2025 年全球智能眼镜市场迎来爆发式增长, IDC 数据显示第一季度中国出货量同比激增 116.1%, Counterpoint 报告则指出上半年全球市场增速达 110%, 其中 AI 眼镜占比飙升至 78%。以 Meta、小米、雷鸟、Rokid 为代表的科技巨头与初创企业集体涌入, 全球超 40 家厂商密集发布新品, 仅 2025 年 5 月单月新品超十款, 中国 AI 眼镜相关企业数量更是突破 440 家, “百镜大战”的序幕正式拉开。

这场竞争不仅关乎硬件形态创新, 更指向下一代人机交互入口的争夺——AI 眼镜正从功能型穿戴设备演化为融合视觉、语音、环境感知的全天候智能终端。这背后也源于多项核心技术的驱动。

1. 多模态交互革命

相比传统的触控类智能眼镜, 新一代的 AI 眼镜融合语音、手势、眼动追踪来实现交互。接入大模型赋予实时问答、实时翻译、场景识别等对话式 AI 能力, 大幅提升交互自然度。

2. 硬件性能跃升

AI 眼镜的落地与普及也离不开显示屏、芯片、传感器等硬件模块性能的提升，例如光学显示层面，MicroLED+树脂衍射光波导技术解决彩虹纹问题；芯片层面，国产芯片开始崛起，例如，华为昇腾芯片，在性能大幅提升的同时，有效降低功耗；传感器集成，PPG 心率监测（Meta）、微投影隐形显示（Gyges Labs）进一步拓展了 AI 眼镜的应用边界。

3. 端云协同架构

轻量化模型（例如 DeepSeek）在端侧处理基础任务，复杂计算云端协同，平衡响应速度与能耗。

在多重技术的驱动下，AI 眼镜的应用场景从工具走向了生态入口，主要分为以下几个类别：

- 消费级场景：实时翻译、语音助手、实时导航、影像记录、语音支付（例如 Rokid 接入支付宝）、控制智能家居（例如小米 AI 眼镜）；
- 行业级刚需：医疗场景调取 3D 解剖图谱辅助手术；工业场景 AR 标注远程维修；无障碍服务，例如中国电信推出搭载星辰大模型的 AI 眼镜，支持视障人士导航和健康监测。
- 社交与元宇宙：Rokid AR Spatial 支持 300 英寸虚拟屏；Meta 眼镜一键分享影像至社交平台。

案例分析：Ray-Ban Meta 智能眼镜从遇冷到爆款

Ray-Ban Meta 智能眼镜是近年来对话式 AI 在智能硬件领域的一个成功案例。它成功的在 AI 技术、时尚外观与实用功能之间取得了平衡，吸引了消费者。而这背后是经历了从初代 Ray-Ban 遇冷到二代爆款的显著转折，核心在于其产品的演进、功能的实用性、成熟技术的巧妙整合等因素。

1. 时尚先行，科技隐形：这款眼镜首先是一副好看的雷朋眼镜，然后才是一款智能设备。它采用了经典的 Wayfarer 等造型，与时尚品牌雷朋合作，确保了外观的时尚感和佩戴的舒适性，解决了以往许多智能眼镜“不愿戴”或“不敢戴”的痛点。

2. AI 功能实用且场景化：其二代产品引入的多模态 AI 交互（如实时视觉识别、翻译）解决了用户在实际生活中的特定痛点（如识别物品、翻译文字），并且其语音交互方式在特定场景下（如双手被占用）比手机更方便。

3. 成熟技术的巧妙整合：Ray-Ban Meta 并没有追求不成熟的“黑科技”，而是将相对成熟的硬件技术（摄像头、扬声器、麦克风、蓝牙）与不断优化的 AI 能力通过软件更新进行整合，提供了稳定可靠的用户体验。

4. **明确的“辅助定位”**：它并未试图取代手机，而是作为手机的一个便携的、解放双手的辅助设备，这种定位更清晰，也更容易被用户接受。

Ray-Ban Meta 的成功案例为行业带来了几点重要启示：

- **工业设计与时尚感是敲门砖**：智能硬件，尤其是可穿戴设备，必须注重外观设计，与时尚品牌联名合作是一个值得考虑的捷径。
- **多模态交互是未来方向**：纯语音交互在某些场景下有局限，结合视觉、音频的多模态交互能极大提升对话式 AI 的理解能力和实用性。
- **用户体验优先于技术堆砌**：用户不关心用了多少项“顶尖”技术，更关心产品能否在特定场景下稳定、便捷地解决他们的实际问题。



5.5 对话式 AI 正从底层重塑教育生态

随着大模型行业的快速发展，对话式 AI 已不再仅是教育领域的辅助工具，而是成为推动教育模式变革的核心驱动力。它以其自然语言交互、个性化支持、多模态融合等能力，从教学方式、资源分配、评价体系等多个维度重构教育生态，推动教育行业从“标准化灌输”向“个性化培养”深刻转型。

目前对话式 AI 在教育行业的核心应用场景主要集中在语言学习产品、课后辅导与作业系统、AI 自主答疑工具、教学工具类应用、虚拟陪伴式学习等核心场景：

1. 语言学习产品：提升口语练习频次与沉浸式交互体验

该场景最典型的业务就是口语陪练，传统的口语陪练存在着多个痛点，例如：一对一真人口语陪练成本高、排课难，无法满足“每天都练”的需求；缺乏真实语境互动，学生只会做题不会说。通过引入对话式 AI 的能力，可以模拟不同的真实口语交流场景，如“旅游景点问路”、“在饭店点餐”等场景，同时 AI 还支持实时纠错+反馈+情景反应，增强学习的沉浸感，打破练习枯燥感。

2. 课后辅导与作业系统：提升学生写作/作答质量，缓解教师改作业压力

传统的线上课后辅导作业场景，一对一的教师辅导，价格较为昂贵，而多人场景下老师因为精力有限，无法对每位学生作业逐条点评。学生写完作业即交，老师点评反馈后即结束，后续改进情况的反馈缺乏。而接入对话式 AI 的方案，通过能说会道的 AI 老师针对学生作文、简答题生成结构化反馈并鼓励二次修改，后续教师可通过 AI 工具批量查看学生的反馈与改进情况。

通过在课后作业辅导中引入对话式 AI 能力，不仅可以提高学生“完成-反馈-修改”的闭环效率，还降低教师评分与点评人力成本，增加学生作业时长与深度，形成真实学习路径数据。

3. 虚拟语伴/多模态互动学习：通过陪伴式引导培养学生自主学习习惯

在教育场景，除了老师教的有质量，学生也需要学的有动力，很多时候学生“不想学”，是因为没有学习计划或方向，缺乏情绪支持与学习成就感反馈。通过对话式 AI 为学生提供“陪伴式 AI 角色”提供每日学习建议、情绪鼓励、学习日记回顾等，以多模态交互的方式引导学习，提升沉浸感，培养学生自主学习的习惯。对于教育机构而言，通过教育与 AI 陪伴的结合，还能拓展情感陪伴与非认知能力培养业务线。

4. 数学工具类应用：解决学生不会问、教师无法精讲的问题

传统的数学工具类应用虽然能为学生快速的解答数学题目，但这背后也仍然存在一些学习痛点，例如数学难题学生拍照查答案，但看不懂步骤；学生不会提问或不敢问，造成“刷题不理解”。

通过引入解答更智能、交互更自然、流畅的对话式 AI 解决方案，AI 可以提供逐步解题过程，并配合“对话式追问”功能解释每一步的解题，同时还可以重写答案解释风格，如“更简明的解题”或“像老师讲”。这样不仅可以优化搜索类数学工具的“理解深度”；也提升了用户的学习时长与学习效率。

结合以上教育场景中对话式 AI 的应用以及声网在在线课堂的业务实践与行业观察，我们认为随着对话式 AI 多模态交互的不断成熟，实时互动课堂可能未来会有以下变化。

互动课堂场景	互动类型和频率	未来场景变化	可能性
1v1	真人高频互动	全赛道覆盖，从真人教学变成自学模式 -正价课1v1，AI数字人老师代替真人 -Demo课1v1，AI数字人老师代替真人 -1v1 AI口语老师，未来语言学习机构必备的应用	几乎100% 会出现这类场景变化，AI数字人老师可能用于成人，卡通IP用于低幼龄的课程
小班课	真人高频互动	全赛道覆盖，前期可作为AI学生加入班级带动互动频次，提升互动效果和体验。后期可以做为AI老师来带领小班互动课。 -正价课，AI学生代替真人学生	目前AI还未完成成熟，多人互动需要更深度的AI训练和技巧。可以用AI助教来做课间休息时候的辅助
大班课	真人低频互动	职业赛道，AI老师代替真人老师授课	100%确定场景，目前已经有机机构用AI数字人老师来做大班课。
AI互动录播课	学生之间，学生和内容低频互动	传统的AI互动课（脚本和视频片段预设好后的互动），未来有可能转变成真AI数字人老师互动课，更加千人千面，更加个性化。	目前看可行性较高，主要是互动来自课件内容和学生互动，次要互动是学生和学生互动，老师主要以讲课和辅助为主。

未来对话式 AI 对教育行业带来的改变可以概括为以下几点：

1. 提升教学效率与教师创作能力

AI 可协助批改、生成材料、翻译、教学差异化设计，缓解教师日常负担，提升课堂质量与个性服务能力。

2. 促进个性化学习与公平接入

对话式 AI 可根据学生水平、习惯提供个性辅导，支持多语言与不同学情的学生。例如 EmoGPT 提供心理支持，CampusGPT 提供免费模型服务。

3. 改变教学评估与学习分析方式

AI 可实时监测学习行为、提供预警与反馈，优化学习路径和效果，赋能“智能教学”发展。

4. 重塑师生角色与学习方式

学生从被动学习者转为 AI 的合作伙伴，教师角色从内容传授者转向引导者与批判性思维促进者。

在未来的教育场景中，AI 不再只是工具，而是“第二个老师”。对话式 AI 正在重塑教与学的方式，让每一个学生都能拥有随时响应、因材施教的学习伙伴。

案例分析：豆神 AI&声网 打造 AI 超拟人多对一直播课

豆神 AI 是 A 股上市公司豆神教育推出的语文素质能力培养型 APP，整合了自研内容平台与多家主流大模型，具备 AI 课文智能互动讲解、AI 作文批改、实时电话私教问答等功能。

豆神 AI 今年宣布推出 AI 超拟人多对一直播课，课程采用真人教师与 AI 教师“双师”协作体系，实现千人千面的定制化训练与辅导。超拟人 AI 老师给学生直播讲解诗词等语文知识，名师、黑板、同学、举手发言，所有教育要素一应俱全，真实还原了线下课堂的体验，AI 会通过视觉大模型时刻关注学生的课堂注意力，并及时发现调整学生的情绪状态。

在声网对话式 AI 引擎的加持下，AI 教师能够在 AI 实时训练体系、课堂答疑、情境创设、AI 实时判卷等环节中与学生展开低延时、自然流畅的对话、保障了 A 语音交互的实时性和稳定性。

5.6 智能助手成为普及最广泛的 AI 应用

在对话式 AI 的诸多落地形态中，以豆包、DeepSeek、腾讯元宝、智谱清言等为代表的智能助手应用，已毋庸置疑地成为渗透率最高、普适性最强的 AI 服务形式。它们凭借自然语言交互、多模态理解、场景深度融合及低使用门槛等核心优势，从工具、社交、政务、办公、教育等维度全面融入用户的日常生活与工作，标志着 AI 技术从“可用”走向“好用”，并正逐步成为如同水电般的基础服务。

5.6.1 技术驱动：多模态与大模型赋能智能助手“泛在化”

智能助手应用的广泛普及，根本源于底层 AI 技术的跨越式发展。基于大语言模型 (LLM) 的架构升级，使其交互能力从单一文本走向语音、图像、视频等多模态融合，理解与生成能力也从事务型问答迈向情感化、人格化的陪伴与创作。

1. 多模态交互成为标配：语音对话是人类最自然的交流方式，当下主流的智能助手应用均支持文本、语音、视频的交互方式，不仅大幅降低了 AI 技术使用门槛，也推动人机交互从“手动操控”到“自然对话”的根本性转变。例如智谱清言支持用户通过语音、视频通话功能进行音视频聊天，能识别画面中的物体，并进行实时对话与解答；豆包 AI 的“数据智能可视化”功能允许用户上传数据表格，自动生成图文分析报告。这标志着智能助手正逐渐从“对话机器人”进化成“全能感知终端”。

2. 大模型深化语义理解与服务深度：DeepSeek 等大模型在政务场景中展现出强大的语义理解与专业知识处理能力。例如，潍坊市政务服务应用 DeepSeek 大模型，对 1974 个政务服务事项的问答知识库进行口语化改造，使智能客服能准确理解并解答用户口语化的办事咨询，准确率超 90%。这背后是大模型对海量知识库的深度融合与多轮对话优化的能力支撑。

3. 低成本与高效率推动规模化落地：智能助手类应用多采用云端部署与端侧轻量化结合的方式，用户无需昂贵设备，通过智能手机或网页即可访问，极大降低了使用门槛。这也是其能快速覆盖至三四线城市的关键原因。

5.6.2 功能革新：从“问答”到“赋能”，重塑工作与生活体验

智能助手的功能已远超传统意义上的“问答”，其核心价值在于成为用户全方位的效率伙伴、创意引擎与情感陪伴。

功能维度	代表应用与能力	应用场景与价值
智能办公与创作	豆包、WPS AI、商汤商量、智谱轻言 生成市场分析报告、提炼会议纪要、生成工作总结、数据分析与表格生成等；	提升专业工作效率，降低重复劳动，辅助决策判断
政务与公共服务	DeepSeek、腾讯元宝DeepSeek满血版 提供7×24小时政务咨询、材料预审、智能分诊	优化政务服务流程，提高办事效率，实现“数据多跑路，群众少跑腿”
社交与情感陪伴	豆包、腾讯元宝、MiniMax 提供情感支持、生活建议（如婚恋、育儿）、趣味聊天	填补情感需求，提供即时、准确的社交反馈，缓解现代社会的孤独感
教育与知识获取	DeepSeek、豆包、腾讯元宝 解题答疑（提供思路而非仅答案）、知识总结（如视频“课代表”）	实现个性化学习辅导，降低知识获取门槛，提升信息消化效率

5.6.3 未来更具人格化与情感智能

智能助手正通过 API 集成、平台嵌入、账号体系互通等方式，深度融合至各类软硬件生态中，成为连接一切服务的“隐形”中枢。但其进一步普及仍面临诸多挑战，并呈现出清晰的发展趋势。

1、内容准确性与用户隐私仍面临挑战：

准确性瓶颈：一些智能助手背后的大模型在处理大数据量时存在限制，且生成结果具有一定不可控性。AI 生成的内容仍需人工审核与辅助。

隐私与伦理：深度融合的智能助手触及大量用户数据，如何保障隐私安全、避免算法偏见是重中之重。

场景深度：在高度专业化领域（如精密制造、高级医疗诊断），智能助手仍需与领域知识更深度结合。

2、未来趋势：AI 助手走向情感智能与主动智能

更具人格化与情感智能：未来的智能助手将不仅能完成任务，更将具备更鲜明的“人格”设定和更高的情商，提供更深度的情感支持和情绪价值。

主动智能与无缝服务：从“问答”走向“预见”，根据用户习惯、地理位置和时间上下文，主动提供提醒、建议和服务。

进一步降低使用门槛：强化语音交互（特别是方言识别）、图像识别等能力，让银发群体和数字化弱势群体也能平等享受 AI 红利



5.7 AI 客服已超越“成本削减工具”，成为企业服务竞争力核心组件

在对话式 AI 的落地场景中，客服领域始终走在最前沿。早期，企业引入 AI 客服的核心诉求是降低人力成本、提升服务效率——通过自动回答高频问题、7x24 小时在线响应，将人工客服从重复性劳动中解放出来。然而，随着多轮对话、情感计算、业务融合等技术的成熟，AI 客服的价值已发生根本性转变：从最初的“成本削减工具”演进为重塑用户体验、驱动业务增长的核心竞争力组件。

5.7.1 价值升维：从“降本增效”到“体验重塑”与“价值创造”

传统 AI 客服的核心指标是“成本降低率”和“一次性解决率”，而新一代 AI 客服的评价体系已扩展至三个维度：

1. 体验提升维度：

情感化交互：通过情绪识别（如声纹情绪分析、文本情感判断）动态调整回应策略。例如，当识别到用户焦虑时，AI 会优先采用安抚性话术并主动转接人工。

个性化服务：结合用户历史行为、订单数据、偏好标签，提供定制化解决方案。例如电

商的 AI 客服能根据用户购物习惯主动推荐退换货方式，提升售后满意度提。

2. 业务增长维度：

精准营销与交叉销售：在解决用户问题的同时，识别潜在需求并推荐相关产品。

商机挖掘与流失预警：通过分析对话中的负面情绪或关键词（如“投诉”“注销”），自动触发挽留流程并推送优惠券，降低流失率。

3. 战略决策维度：

用户洞察反馈：AI 客服成为企业感知市场的“神经末梢”，通过分析海量对话数据，自动生成用户痛点报告、产品缺陷汇总、竞品对比分析，反向驱动产品迭代与运营优化。

5.7.2 技术融合：多模态、大模型与业务系统的深度集成

技术的不断成熟与演进是每个场景价值跃迁的基础。新一代 AI 客服呈现三大技术特征：

1. 多模态交互融合：支持语音、文字、图像、视频等多渠道输入。用户可通过拍摄产品照片直接描述问题，AI 通过 CV 识别故障并推送解决方案，提升解决效率。

2. 大语言模型（LLM）重构对话能力：基于 GPT-4、通义千问等大模型的客服系统，突破了传统意图识别的局限，能够理解模糊表达、处理复杂多轮对话，并生成更自然、更具逻辑的回复。

3. 与业务系统无缝集成：AI 客服深度对接 CRM、ERP、订单系统等后台业务，不再是孤立的问答工具，而是成为业务办理的入口。例如，用户可直接通过语音指令完成“机票退订”“套餐变更”等操作，全程无需人工介入

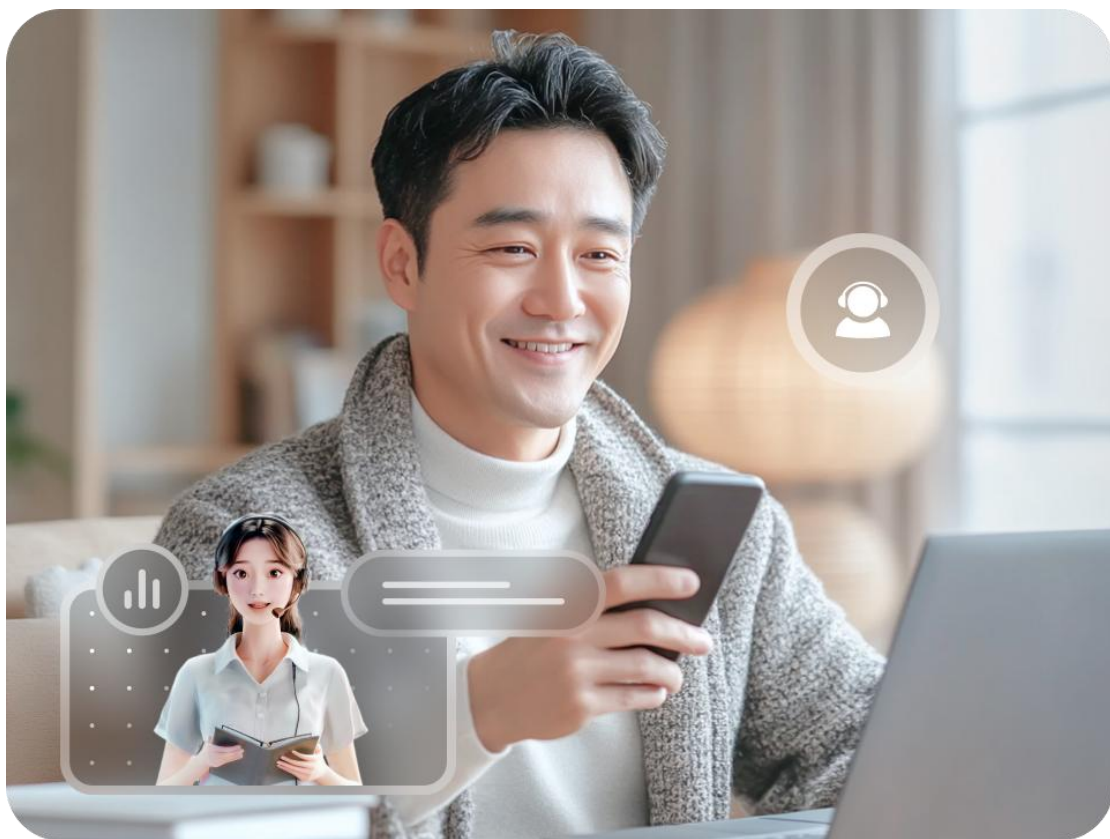
5.7.3 行业实践：从“标准化服务”到“场景化创新”

各个行业基于自身业务特性，已经逐渐挖掘出 AI 客服的差异化价值：

金融行业：AI 客服不仅处理业务咨询与查询，更成为“风控前线”。通过对话内容实时分析欺诈风险（如异常转账请求），银行等金融机构可以有效降低诈骗类案件发生的风险。

政务领域：“AI 政务助手”实现政策解读、材料预审、办事引导一体化，将市民咨询窗口排队时间大幅降低，同时通过精准解读提升政策触达效率。

零售与电商：AI 客服承担从售前咨询、售中订单管理到售后服务的全流程服务，并有望成为私域运营的重要手段。



面向未来，随着大模型、多模态交互等技术的不断成熟，AI 客服的发展将呈现多维度趋势。首先，智能化水平将持续提升，大模型技术让 AI 客服能更精准理解客户意图，实现“所答即所问”，并从被动应答迈向主动挖掘需求与销售，通过配置销售技巧主动引导客户表达真实需求。其次，交互方式会更趋近人类自然沟通，在控制幻觉问题后，回复将更加自然、人性化。再者，AI 客服将与业务流程深度融合，实现营销服一体化应用，数据标签与客户终身价值模型在各环节发挥作用，且智能客服数据可用于销售和营销。此外，自动构建决策模型，依据客户决策因素和因子，结合数据体系精准应答。最后，多模态交互兴起，除文本外，支持语音、图像等交互，为客户提供更便捷、多元服务体验，全面提升客户服务质量与效率。

第六章：对话式 AI 的未来展望

当人类与机器的沟通不再需要学习复杂的代码指令，当自然语言成为通行数字世界的万能密钥，对话式 AI 将彻底重构人机交互的底层逻辑。它从“功能工具”蜕变为“**感官的延伸**”与“**情感中枢**”——通过类人化的多模态交互、深度情感共鸣与环境感知能力，在对话互动上实现智商、情商均在线的要求。这场革新不仅是技术的进化，更是文明向“**人机共生**”演进的里程碑。

6.1 对话式 AI 多模态与情感能力将达到类人水平

对话式 AI 将成为智能体不可或缺的感知中枢，其核心价值在于实现 “**能说会道**” 且 “**察言观色**” 的类人能力——深度理解人类的情感意图与思考逻辑，成为真正可信赖的伙伴与助手。未来将在三大方向实现突破性跨越：

能力维度	进化方向	类人化实现程度
多模态交互能力	视觉理解 + 自然对话韵律 + 多人并发处理	环境物体识别准确率达到 95%，声纹识别准确率达到 99%，支持多人/机并行发言无冲突
深度情感共鸣	情绪理解 + 人格化表达 + 长期关系适配	可结合语气+表情变化理解情绪变化；支持人格化表达；匹配长期关系人设
环境感知与场景适配	时空上下文响应 + 工业级鲁棒性	感知并识别时空环境和场景

一、多模态交互：全感官协同的拟真对话

1. 全双工对话——开启自然对话

- **全双工对话**：彻底告别一问一答机械式、轮式对话的低效交流械，同时 AI 还将更加准确判断、及时把握发言时机。
- **超低延时**：人机对话延时将从目前秒级进阶到毫秒级的水平，最终实现端到端 400ms 以内的超低延时对话体验。

2. 自然打断——突破机械感壁垒

智能打断：响应反应延迟从目前的 300ms 级持续获得优化提升，支持任意打断后无缝续接，插入“嗯”“明白”等反馈词使对话流畅度媲美真人。同时准确识别过滤“嗯”、“哦”等非打断语气词，保持对话的连续性。

- **识别停顿：**识别对话中的停顿和沉默，实现 AI 主动打破沉默发起话题，模拟真实社交中的“边听边说”，提升人机对话的自然度。
- **选择性注意力锁定：**可以通过声纹特征，准确锁定发言对象，在嘈杂或多人环境下，选择性锁定或识别目标发言内容；同时还可以在家庭/会议场景中精准区分重叠语音，实现并行发言、甚至进行是辩论或合唱。

3. 视觉理解——让 AI 能说会看

- **环境与物体理解：**精准识别空间物体（如空间环境、人物细节、故障设备），为 AI 感知加上“眼睛”，让 AI 看见这个世界，看这个与之对话的对象与环境。
- **肢体语言理解：**结合对话上下文，识别理解对象手势、眼神等肢体动作，以更准确的理解对话对象的思想与表达。
- **微表情情感映射：**从面部肌肉微动捕捉情绪波动，准确感知并理解情绪变化。同时，结合计算机视觉与语音分析，使 AI 数字人实时响应用户表情和动作。

二、情绪识别：穿透语言表象的精准洞察

1. 多模态情感画像

- **跨模态情绪融合：**通过语音基频、面部肌肉微动、甚至是心率变化等生理信号，建立感知多模态情感画像，准确识别对象情绪情感变化。
- **矛盾情感解析：**识别“强颜欢笑”、“笑中含泪”等矛盾表达（如语音积极但微表情消极），触发主动关怀策略。例如 AI 检测到用户哽咽声后自动降语速 20%，插入安抚词。

2. 语境化情绪推理

- 结合对话历史与场景语义，区分短暂情绪波动与长期心理状态。

3. 长期关系构建——动态记忆进化

- 记录 200+ 维度用户画像（如健身习惯/饮食偏好），对话连贯性提升 40%，对焦虑用户自动减少反问频次。

三、共情表达：从理解到行动的情感支持

1. 人格化建模：基于人格特质（外向性/严谨性）生成稳定对话风格。

2. **情感语音合成：**调节语速、停顿及语调起伏，生成“带笑意的安慰”或“克制的提醒”，贴近真人表达。

四、长期陪伴：动态关系构建与场景适配

1. 个性化记忆进化：多维度记录用户画像（如“周三健身”“讨厌咖啡因”），对焦虑用户自动减少反问句频率，对抑郁倾向者增加肯定式鼓励

2. 场景化情感支持：

- **家庭场景：**AI 宠物在对话中检测用户落寞神情时，主动推荐治愈电影并调整蜷缩姿态供拥抱入睡
- **职场场景：**会议助手识别发言人愤怒语调时，生成“是否需要暂停讨论”的缓冲建议

未来 AI 产品通过“越用越懂你”的人格进化机制，使 AI 从工具真正升级为“人类数字伙伴”，开辟硬件情感溢价市场

6.2 科幻走进现实，基于对话式 AI 能力的 AI 超级助手将出现

随着大语言模型与多模态技术的深度进化，**具备类人对话能力的超级 AI 助手将从科幻走向现实**，成为重构人机协作关系的文明级变量。未来的超级助手将如电影《Her》中的“萨曼莎”一般，突破工具属性桎梏，成为人类最可靠的伙伴。

未来 AI 超级助手将根据角色和功能分为：个人超级助手、企业超级助手、行业专家等不同角色

6.2.1 个人助手：从“工具”到“终身伙伴”

1. 虚拟主体：可跨设备运行的“数字生命体”

核心形态：

虚拟本体：以软件形式运行于手机、车机、AR 眼镜、家电等各类智能终端，通过联网学习实现数据同步，可以实现跨终端应用和互通，实现真正的如影随行。

实体载体：可以自动匹配各类实体终端，比如桌面机器人，全息投影设备、车载大屏，并可以通过对话或姿态动作传递情感。

技术支撑：

端云混合部署：即使在智能终端设备遭遇突发断网时，仍能通过端侧轻量化模型临时对话和应用。

跨操作系统兼容：适配 Windows、Android、iOS、Linux 等各类平台，实现“一次部署，全平台服务”。

2. 自学习与进化：越用越懂你的“伙伴”

动态知识库：可以整合设备端与云端数据（如健康记录、工作文件、对话历史），构建多维度用户画像（例：“每周三健身”、“对海鲜过敏”、“项目获批”），实现知识库动态更新。

学习进化引擎：可以根据对话内容、对话风格、工作内容、日常习惯、指令偏好以及日程安排进行自训练、自学习和自进化，完成 AI 的自动进化，并动态优化服务策略。

人格化成长轨迹：可以基于 Big Five 特质（外向性/严谨性）生成稳定对话风格，随互动深化调整表达方式（如从“严谨助手”渐变为“幽默朋友”）

主动场景适配：职场场景用严谨语气生成报告，家庭场景切方言讲睡前故事，实现“一人千面”

3. 终身伙伴-人类文明的“碳基-硅基”共生体

AI 超级助手将超越工具属性，成为贯穿用户生命全程的**意识延伸体与情感承载体**，实现对人们全周期的陪伴。

成长阶段：启蒙守护者

1) 婴幼儿期：

健康监测：通过可穿戴设备实时追踪生理指标（如心率、体温），结合哭声分析判断需求（饥饿/不适），并且根据不同的场景自动安抚宝宝。

认知启蒙：动态生成适龄互动内容（如 3D 动画儿歌、多语言启蒙），并通过与宝宝对话，进行认知启蒙教育。

2) 学龄期：兴趣与情商培养

兴趣培养：通过日常对话与练习，辅助家长分析挖掘和培养小孩兴趣，并生成兴趣训练计划。

情商培育：识别社交焦虑（如语音颤抖），模拟同伴对话练习，培养共情能力。

3) 青少年期：学习陪伴者

个人助教：自动记录各类学习指标变化（注意力时长、错题类型等），生成定制学习路径与提升计划。

学习陪练：陪伴学习，通过对话帮忙学习进行背诵、口语训练以及错题纠正。

4) 成年及后续阶段：超级协作伙伴

在服务对象逐步进入成年阶段，超级 AI 助手将会成长为人类生活助理、工作助理、学习助理、情感助理和健康助理的综合体，为人类提供从工作、生活到学习，从情感到健康的全面服务。

AI 超级助手将成为人类的超级协作伙伴，它不仅是最懂你的“知己朋友”，它将成为你的最佳倾诉对象；它也是你一生的“老师”，陪伴你一生的学习与成长；它还是你的私人“管家”和个人“助理”，帮你把工作和生活打理得井井有条；同时它又是你的情感导师和健康医生，时刻关注和管理你的身心健康。

未来，人类一生中对话最多将会是自己的 AI 超级助理，未来最懂你的那个“人”也将会是 AI 超级助理。

4. 个人分身-数字孪生身份

个人分身 (AI Avatar) 是 AI 超级助手的一种高级形态，其核心特征在于通过深度学习与多模态技术构建高度个性化的“数字自我”，实现从工具到“代理者”的质变。

数字孪生：通过长期交互数据（语音、文字、行为习惯）构建用户的“数字孪生”，精准复现语言风格、决策逻辑甚至价值观。实现从工具到数字自我的形态进阶，核心体现在于除了响应指令和安排事务外，个人分身还可以代替个人承担社交身份，进行社交活动。

行为代理：替代用户执行重复性任务，如回复邮件、参加次要会议、维护社交关系，实现“睡眠时仍在工作”，将人类从重复性任务、无效社交中解放出来。

未来，个人分身与人类将形成共生关系，个人分身处理机械性工作（如数据整理），人类专注创造性决策，形成能力互补。

6.2.2 企业超级助手：组织智能体

企业超级助手将从单点功能向多智能体协作网络跃迁，它未来将成为企业多功能的数字员工，全面提升企业运营效率。

1. 数字员工：企业超级助手将以企业数字员工的身份存在，成为企业的一员，并在企业的发展中实现不断升级和进化。

2. 企业知识库：构建知识图谱，形成企业知识库，大幅提高企业信息检索效率。未来需要查询企业资料或产品知识，直接开口问企业超级助手即可为你迅速提供最准确的答案。

3. 企业培训师：代替企业的培训师，以 1v1 或课堂的方式完成员工的企业入职培训、岗位培训和产品培训等工作，并随时解答员工的问题。还可以代替企业的培训师，进行经销商/代理商培训，并随时解答经销商/代理商问题。

4. 招聘助手：招聘管理：自动筛选简历、预约面试、生成岗位测评题。

面试管理：自动与候选人完成首轮线上面试，并自动分析候选人匹配度，提供给业务部门负责人。

5. 超级客服：咨询投诉：代替人工处理电话、邮件、社媒等多渠道咨询投诉，通过声纹识别与情感分析（如愤怒语调检测）动态调整话术，助力企业全面提升客服效率和客户满意度。

主动营销：代替电话销售人员，进行电话外呼营销，并可基于客户画像或历史订单推荐关联商品，有效提升转化率和客单价。

6.3 对话式 AI 对人类社会产生深远影响

对话式 AI 正以多维方式重塑人类社会。它彻底改变人机交互逻辑，让自然语言成为主要接口，人们无需复杂操作，通过日常对话即可操控设备、获取服务，从智能音箱响应指令到多模态助手理解表情手势，人机互动更贴近人际交流。同时，它显著提升效率与生活质量，既能帮专业人士快速处理文书查询，又能为普通人规划行程、管理日程，企业则借此实现全天候智能服务，大幅降低沟通成本。

在更宏观层面，对话式 AI 驱动行业革新与经济升级，金融、电商等领域借其开发智能顾问、数字人直播等新形态，催生万亿级市场空间。知识传播也因之突破壁垒，它化身随叫随到的“移动知识库”，既能即时解答疑问，又能为学习者定制教育方案，让信息获取更高效，推动知识普惠。这种技术渗透正从生活细节到产业结构，持续释放变革能量。

6.3.1 改变人机交互方式

- **从图形界面到自然语言交互**

长期以来，图形用户界面（GUI）主导着人机交互。用户通过鼠标、键盘操作图标、菜单来与计算机系统交互，完成任务。但这种方式要求用户学习特定操作流程，如在办公软件中，用户需熟悉各种菜单选项位置来进行文档编辑、格式调整、命令发送。复杂软件的学习曲线陡峭，新用户与老用户难以快速上手，限制了技术与软件的普及。

对话式 AI 的出现，使交互从 GUI 转向自然语言。用户能像与他人交谈一样，用日常语言与机器沟通。如在智能语音助手应用中，用户说“帮我查询明天北京到上海的航班”，助手即可理解意图，快速提供航班信息；智能家居场景下，用户发出“打开客厅灯，将亮度调至 50%”指令，智能设备就能执行。这种自然语言交互更符合人类交流习惯，打破 GUI 操作限制，让用户交互更轻松、高效。

- **降低技术使用门槛，促进普惠科技**

对话式 AI 显著降低技术使用门槛，推动普惠科技发展。以往，残障人士、老年人等群体因 GUI 操作困难，难以充分利用数字技术。对话式 AI 的语音交互功能，使视障人士通过语音指令即可使用设备，如查询信息、操作软件；老年人无需复杂操作学习，通过自然语言交流就能享受数字服务，如与智能音箱聊天、获取生活资讯、控制智能家居设备。

在教育资源匮乏地区，学生可借助对话式 AI 学习工具，用自然语言提问，获取课程讲解、作业辅导等帮助，弥补师资不足。这使先进技术惠及更多人群，无论身体状况、年龄大小、地域差异，都能平等获取信息、享受技术服务，缩小数字鸿沟，促进社会公平。



6.3.2 提升效率与生活质量

- 工作场景中的应用与效率提升案例

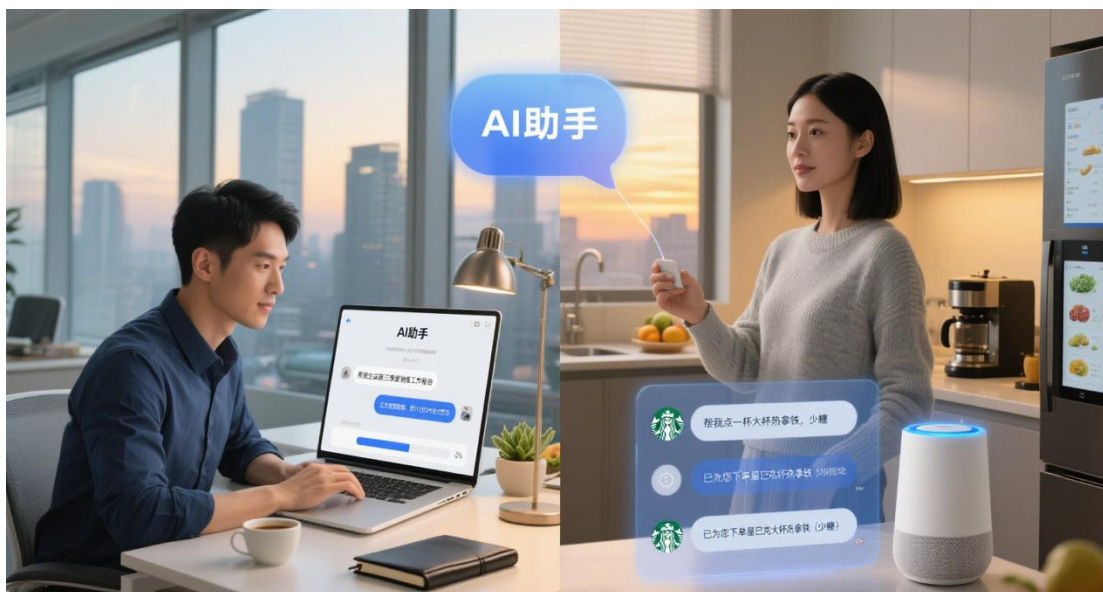
在办公领域，对话式 AI 应用广泛。智能办公助手可理解自然语言指令，如“安排周五下午 3 点与销售团队的会议，准备上季度销售报告”，助手自动完成会议安排、报告生成等繁琐任务，减少人工操作时间。在文档处理方面，用户对 AI 说“总结这份合同的关键条款”，AI 快速分析提炼，提高工作效率。

客服行业，智能客服运用对话式 AI 实时响应客户咨询，处理常见问题，如电商平台中，客户询问商品信息、物流状态，智能客服快速解答，减轻人工客服压力，实现 24 小时服务，提升客户满意度和企业运营效率。

- 生活服务领域的创新与便利性增强

生活服务领域，对话式 AI 带来诸多创新便利。餐饮行业，消费者通过智能点餐系统，用语音描述菜品喜好、口味要求，系统推荐菜品、完成点餐，提升点餐体验。外卖配送中，用户可与配送助手对话，了解配送进度、修改地址。

智能家居借助对话式 AI 实现设备智能控制。用户回家说“我回来了”，智能系统自动开灯、调节室内温度、播放舒缓音乐；睡前说“关闭所有电器，拉上窗帘”，系统统一执行，提升生活便利性和舒适度，让生活更轻松惬意。



6.3.3 推动行业创新与经济增长

- 创新商业模式与产业生态的形成

对话式 AI 催生新商业模式，改变产业生态。在智能营销领域，企业利用对话式 AI 分析客户对话，精准洞察需求，开展个性化营销。如在线教育机构通过与潜在学员对话，了解学习需求、目标，推荐合适课程套餐，提高营销转化率。

同时，围绕对话式 AI 形成新产业生态。技术研发企业、数据标注公司、内容提供商、应用开发商等相互协作。研发企业提供先进技术，数据标注公司为模型训练准备数据，内容提供商丰富对话内容，应用开发商将技术集成到各类产品，共同推动对话式 AI 产业发展，创造新经济增长点。

- 对教育、医疗、金融等具体行业的变革作用

医疗行业，对话式 AI 辅助医疗服务。智能导诊系统通过与患者对话，了解症状，初步判断病情，推荐就诊科室。医疗机器人可与医生协作，记录病历、查询医学资料、辅助诊断。患者还能通过 AI 聊天机器人获取健康管理建议、疾病预防知识，改善医疗服务可及性和效率。

金融领域，对话式 AI 优化客户服务和风险评估。智能客服解答客户账户管理、理财产品咨询等问题。智能投顾根据客户财务状况、投资目标，提供个性化投资建议、资产配置方案。同时，通过分析客户对话数据，评估信用风险，提升金融服务质量和安全性，推动金融行业创新发展。



6.3.4 促进知识传播与信息获取

- 知识普及与教育公平性的提升

对话式 AI 为知识普及和教育公平提供有力支持。在线教育平台利用 AI 提供 24 小时学习辅导，不同地区学生随时提问，获取解答。如偏远地区学生学习中遇到数学、物理知识疑惑，通过对话式 AI 学习工具，得到详细讲解，接触丰富知识资源，缩小与发达地区教育资源差距。

科普领域，对话式 AI 科普助手以通俗易懂语言解答公众科学问题，如“黑洞是如何形成的”“5G 技术原理是什么”，激发公众对科学兴趣，普及科学知识，提升全民科学素养。

- 信息透明度与获取效率的改善

对话式 AI 提高信息透明度和获取效率。政务服务方面，市民通过对话式 AI 政务助手，查询政策法规、办理业务流程、咨询民生问题，如“申请保障性住房需要哪些条件”，助手准确回复，让政务信息更透明，方便市民办事。

企业内部，员工通过对话式 AI 办公助手，快速获取公司制度、项目资料、业务数据等信息，如“查询上季度销售业绩数据”，助手即时提供，减少信息查找时间，提高工作效率。