Interspeech 2026 音频推理挑战赛：评估音频推理模型和智能体的推理过程质量

Ziyang Ma[1,2], *Ruiyang Xu*[1], *Yinghao Ma*[3], *Chao-Han Huck Yang*[4], *Bohan Li*[1], *Jaeyeon Kim*[5], *Jin Xu*[6], *Jinyu Li*[7], *Carlos Busso*[5], *Kai Yu*[1], *Eng Siong Chng*[2], *Xie Chen*[1,**]

[1]上海交通大学，中国， [2]南洋理工大学，新加坡， [3]伦敦玛丽女王大学，英国， [4]英伟达，美国， [5]卡内基梅隆大学，美国， [6]通义团队，阿里巴巴集团，中国， [7]微软公司，美国{zym.22, chenxie95}@sjtu.edu.cn

arXiv:2602.14224v1 [cs.SD] 15 Feb 2026

摘要

近年来，大型音频语言模型（大型音频语言模型）在理解方面表现出色，但往往缺乏透明的推理过程。为解决这一“黑箱”局限，我们在 Interspeech 2026 组织了**音频推理挑战赛**，这是首个专门评估音频领域思维链（思维链）质量的共享任务。挑战赛推出了**MMAR-Rubrics**，这是一种新颖的实例级评估协议，用于评估推理链的事实准确性和逻辑。赛事设有单模型赛道和智能体赛道，吸引了来自18个国家和地区的156支队伍参赛。结果表明，智能体系统目前在推理质量方面领先，能够利用迭代式工具编排和跨模态分析。此外，单模型通过强化学习和复杂的数据流水线迅速进步。本文详细介绍了挑战赛的设计、方法以及对最先进系统的全面分析，为可解释音频智能提供了新的见解。[1]

索引词：音频推理、评估、大型音频语言模型、多模态智能体

1. 引言

理解声音并进行推理是人类智能的基本方面。从口语对话和音乐作品到细微的环境线索，人类不仅能够感知各种各样的听觉信号，还能解读其含义、进行推断，并在复杂的声学场景中做出决策。在人工系统中复现这一能力，长期以来一直是人工智能研究的一项关键目标。

大语言模型（LLM）的近期进展与音频处理技术的进步相结合，催生了大型音频语言模型（LALM）[1, 2, 3, 4, 5, 6, 7, 8]。借助大规模多模态训练，大型音频语言模型在理解任务 [9]中取得了令人瞩目的成果。除了理解任务之外，Audio-CoT [10]和 Audio-Reasoner [11], 等开创性工作也尝试将显式思维链（CoT）推理引入音频领域。然而，尽管取得了这些进展，当前的大型音频语言模型在推理能力方面仍存在局限，且表现不稳定。现有音频相关推理基准测试（如 MMAR [12], 、OmniBench [13], 和 MMAU-Pro [14], ）的一项关键局限在于，它们主要关注最终答案的准确率，却忽略了中间推理过程。这种“黑箱”范式会带来显著风险，尤其是在复杂的现实场景中，可解释性和多步推理至关重要。

为解决这一局限并推动透明的音频智能发展，我们在 Interspeech 2026 举办了**音频推理挑战赛**。本挑战赛旨在推动社区将评估重点从以结果为导向的指标转向以过程为导向的推理质量。为适应多样化的架构方案，我们设计了两个不同的赛道：面向端到端大型音频推理模型（ LARM）的**单模型赛道** ，以及面向多模态、可使用工具并具备自验证能力的系统的 **智能体赛道** 。

这项挑战赛的一项关键创新，是聚焦于评估的信度和稳定性。思维链评估仍是一个开放问题，且常受不稳定的大语言模型评审指标困扰。为缓解这一问题，我们推出**MMAR-Rubrics**，这是 MMAR [12] 基准测试的更新版本，并采用了一种新颖的实例级评分标准评估协议。该协议以人工标注的思维链数据为基础，能够对推理路径的事实准确性、逻辑性和完整性进行细致评估，从而显著降低传统评分方法的方差。

本次挑战赛受到了全球研究界的广泛关注。我们收到了来自18个国家和地区的156份报名。整个赛程竞争激烈，最终有23支队伍成功提交结果并进入单模型赛道的最终排行榜，智能体赛道则有24支队伍。这一踊跃参与的情况凸显了为音频模型配备强大推理能力日益迫切的需求和浓厚兴趣。

本次挑战赛的主要贡献总结如下：

- **开创性的音频思维链评估**。 我们推出了首个专门评估音频推理模型和智能体中间推理过程的挑战赛，为可解释音频智能开辟了一个新方向。
- **稳健的评估协议**。 我们提出了一种实例级量规评估方法，有效解决了以往大语言模型评审指标固有的不稳定问题。
- **双赛道洞见**。 通过分析表现优异的系统，我们全面了解了当前端到端大型音频推理模型和智能体系统，并揭示了各自的优势。
- **开放资源**[2]。我们发布 MMAR-Rubrics 基准数据、评估脚本和详细技术报告，以促进未来研究。

**表示通讯作者。[1]音频推理挑战赛官方网站：https: // audio-reasoning-challenge.github.io/

[2]音频推理挑战赛的 MMAR-Rubrics 数据和脚本：https://github.com/ddlBoJack/MMAR

2. 相关工作

2.1. 音频推理系统

当前的音频推理方法通常分为两类：将推理过程内化的端到端大型音频推理模型（LARMs），以及利用大语言模型协调专用音频工具的音频智能体。

早期 [10] 研究试图通过上下文学习激发大型音频语言模型的推理能力，并证明提示模型生成显式思维链能够在某些任务中解锁推理能力。为了将这一能力内化，后续工作 [11, 15, 16] 转向基于大规模合成数据进行监督微调（SFT）。最近的前沿方向 [17] 则是通过强化学习（RL）对音频模型进行对齐。尽管将GRPO等算法直接应用于音频时，在促进推理链 [18, 19], 方面面临挑战，但混合方法已应运而生。通过将思维链冷启动数据与改进后的GRPO流程相结合，模型 [20, 21, 22, 23] 成功缓解了推理问题。

与单体模型不同，智能体将推理分解为规划、感知和整合等离散阶段。 [24] 该领域的主要策略是符号桥接，即首先将音频转换为中间文本 [25]或结构化表示，以利用基于文本的大语言模型的强大推理先验 [26, 27]。除了静态转换之外，近期的智能体还强调动态优化并实施迭代循环，使系统能够主动查询工具，并以由粗到细的方式优化证据 [27, 28]。近期工作 [29] 不再盲目信任感知输出，还引入了自反思决策机制，使模型能够自主决定何时依赖内部知识、何时咨询外部识别模块，从而显著增强其在多步推理任务中的稳健性。

2.2. 音频推理基准测试

为了严格评估音频模型不断提升的能力，研究社区推出了一系列基准测试，旨在考查超越简单感知的复杂推理。这些基准测试通常涵盖两个方向：针对特定领域的专项评估，以及全面的多技能评估。

语音推理评估领域的研究大多侧重于逻辑演绎，评估基于口语输入的数学推理 [30] 以及基于语音事实的多跳推理 [31, 32], 分别如此。时空动态也是关注重点，用于衡量模型对时间、三维空间和物理声音动态的理解，而基于描述的方法往往无法胜任这些领域 [33, 34]。

多项基准测试旨在评估模型对多种音频类型的整体听觉智能，包括语音、音乐和环境声音。早期研究 [35, 36, 37] 评估模型在需要特定领域知识的专家级任务上的表现。一个增强版本 [14] 则重点测试了49种不同技能。此外，研究还提出了一种涵盖信号、感知、语义和文化推理的层级分类体系，并利用真实世界的网络视频探究音频深度推理能力 [12]。近期研究 [38] 也关注中文语境中的复杂、动态和多重场景。

尽管这些基准测试具有多样性和深度，但它们有一个共同的关键局限：主要将最终答案的准确率作为唯一的性能指标。这种以结果为导向的评估无法判断模型得出正确答案究竟是依靠合理的逻辑，还是虚假相关性。因此，目前仍缺乏用于评估

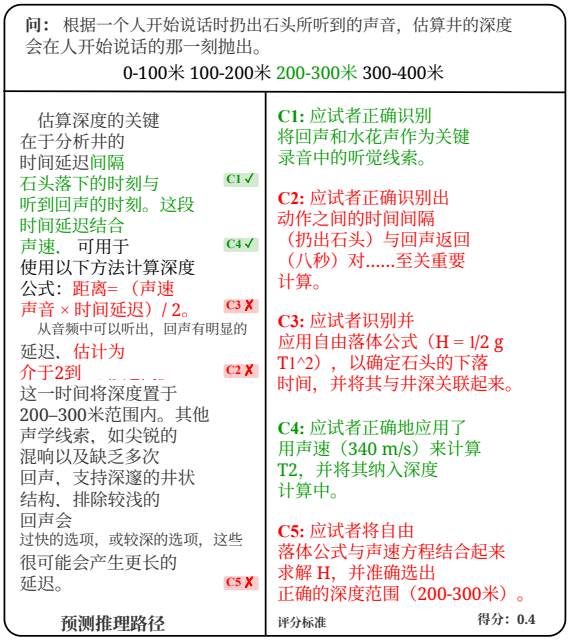

图 1：*MMAR-Rubrics* 如何评估推理过程的示例。

音频推理中思维链的事实准确性、逻辑性和完整性，是本挑战旨在弥补的一项空白。

3. 挑战设计

3.1. 任务形式化

给定一个音频信号 A 和一个语言查询 Q，系统$\mathcal{F}(\cdot)$需要生成一条推理链 $\hat{Y}_{\mathrm{CoT}}$，明确推导出答案，然后给出最终答案 Y_{Answer}：

$$(\hat{Y}_{\mathrm{CoT}}, \hat{Y}_{\mathrm{Answer}}) = \mathcal{F}(A, Q) \tag{1}$$

与以往只对最终答案进行惩罚的基准测试不同，我们的评估函数优先考虑推理过程的质量，但关键在于，将正确执行视为推理有效的先决条件。特定实例的最终得分 S 定义如下：

$$S = \begin{cases} \mathcal{E}(\hat{Y}_{\mathrm{CoT}}) & \text{if } \hat{Y}_{\mathrm{Answer}} = Y_{\mathrm{Answer}} \\ 0 & \text{if } \hat{Y}_{\mathrm{Answer}} \neq Y_{\mathrm{Answer}} \end{cases} \tag{2}$$

其中， Y_{Answer} 表示真实答案，而 $\mathcal{E}(\cdot)$ 是推理质量评估函数。此设计确保模型不会因导致错误结论的“幻觉推理”而获得奖励，从而保证逻辑推理和预测都具有高精确度。

3.2. 实例级量规评估

对推理路径进行自动评估仍是一项尚待解决的挑战。系统级评分是一种常见方法，由大语言模型评分员依据整体量规给出分数。然而，我们的实验发现，这种方法存在不稳定、信度不足的问题。为克服这些局限，我们提出 **MMAR-Rubrics**，这是一种实例级评估协议，可将对预测推理路径 Y_{CoT}的评估分解为

一组可验证的原子标准，这些标准源自人工标注的推理路径 Y_{CoT}。对于每个实例，我们生成 k 条标准。随后，对每条标准进行二元判断^（True/False），检查 Y_{CoT} 是否明确满足相应要求。实例级总分为 k 项二元结果的平均值：

$$\mathcal{E}(\hat{Y}_{\mathrm{CoT}}) = \frac{1}{k}\sum_{i=1}^{k}\mathbb{I}\big[\text{criterion}_i \text{ is satisfied}\big] \tag{3}$$

具体而言，我们使用 Gemini-2 5-Pro 根据真实答案自动生成 $k = 5$ 可验证的标准。这些标准针对具体实例，使评估能够适应每个问题独特的逻辑。为评估预测的推理路径 Y_{CoT}，我们提示 GPT-4o 充当评分员，依据这些标准评估预测结果，如图 1 所示。为提高透明度和信度，模型不仅需要针对每项标准给出二元判断，还需为其判断提供简短的文字说明。

我们从信度和与人类判断的一致性两个方面，将 MMAR-Rubrics 与系统级基线进行比较验证。基线采用一套详尽、通用的评分标准，要求大语言模型评分员在给定真实推理路径的情况下，针对事实准确性、逻辑和完整性等维度，对整体推理质量按 5 分制打分。图 1 展示了这两种评估范式的比较。

表1： 系统级与实例级评估协议的比较。 信度以克里彭多夫 α 系数（α）衡量。评分者间信度根据三位大语言模型评分员（M_1：*GPT-4o*、 M_2：*Gemini-2.5-Flash*、M_3：*GPT-5- mini*）的评分计算。评分者内信度根据 M_1 的五次运行计算。与人类判断的一致性通过与 *50* 对人类偏好（C）和拒绝（R）的推理路径进行比较来衡量。各列显示大语言模型评分员（M_1）与人类偏好一致（$C > R$）、不一致（$C < R$）或判定回答相同（$C = R$）的次数。

协议	信度（$\alpha\uparrow$）		与人类判断的一致性		
	评分者间	评分者内	$C > R\uparrow$	$C < R\downarrow$	$C = R$
系统级	0.7459	0.9081	15	8	27
实例级	**0.7921**	**0.9216**	**17**	**2**	31

• **信度**。 对于评分者间一致性，三名大语言模型评分员（GPT-4o、Gemini-2.5-Flash、GPT-5-mini）对来自 Qwen3-Omni-Thinking 且具有正确最终答案的 664 条推理路径进行评分。MMAR-Rubrics 的克里彭多夫 α 系数达到 0.7921，显著高于系统级评估基线。对于评分者内信度（使用 GPT-4o、温度为 = 1.0 的情况下运行五次），我们的协议也表现出更高的稳定性。

• **与人类判断的一致性**。 我们对 50 对推理路径进行了偏好研究。这些推理对从 Qwen3-Omni-Thinking 以及单模型赛道中的前三名模型的输出中随机抽取。每条路径首先由人工标注员进行评判，然后通过比较两项评分来推断协议的偏好。结果表明，MMAR-Rubrics 与人类偏好更加一致。与系统级基线相比，它更常与人类选择一致，并且产生更少的偏好逆转，即指标错误地偏向被人类否定的路径。

3.3. 双轨设计

为适应当前音频智能研究中的多样化范式——从单体端到端模型到模块化智能体系统，我们设立了两个不同的赛道。两者均在 MMAR-Rubrics 基准上接受评估，但遵循不同的约束条件，以考察不同的能力。

赛道 1：单模型赛道。 参赛者构建一个端到端可微的单一模型（例如，将音频编码器连接到大语言模型解码器），接收音频并生成思维链推理轨迹以及最终答案。系统必须在一次前向传播中进行内在推理，不得委托外部工具、应用程序接口、搜索引擎或独立控制器来完成任务。目标是在不依赖符号化辅助的情况下，单独评估模型内部推理质量。

赛道 2：智能体赛道。 参赛者设计一个音频推理智能体，可协调多个开源模型和工具（例如自动语音识别、音源分离、节拍跟踪、字幕生成器、规划器），以生成思维链路径和最终答案。重点在于提供透明的推理轨迹，展现中间音频分析如何影响决策，超越仅输出答案的处理流程。智能体可通过显式规划（规划—执行—反思）、用于存储中间产物（例如转录文本、音轨、字幕）的结构化记忆，以及自验证或交叉核验来实现。目标是评估系统级推理能力，即将高层次的听觉问题拆解为可执行的子任务，主动搜寻证据，并通过自我纠正解决复杂查询。

4. 挑战赛结果

本次音频推理挑战赛吸引了众多队伍积极参与，共有来自 18 个国家的 156 支队伍报名。为确保评估严谨，比赛分为两个不同阶段。初赛阶段使用了从与最终测试集相同分布中抽取的 500 道题。在这一阶段，单模型赛道有 23 支队伍、智能体赛道有 24 支队伍提交了有效结果。决赛阶段采用了 MMAR 中包含 1000 道题的完整基准测试。最终，只有表现最出色的队伍脱颖而出：单模型赛道有 14 支队伍、智能体赛道有 16 支队伍成功提交结果并进入最终排行榜。本节将分析这些顶尖系统的表现，比较两种范式，并重点介绍促成最佳成绩的技术创新。

4.1. 总体表现

表 4 展示了两个赛道的最终排行榜。结果显示，基于智能体的系统在推理质量和最终答案准确率方面通常都优于端到端系统。表现最佳的智能体系统评分标准得分为 69.83 %，而表现最佳的单模型得分为 65.29%。值得注意的是，单模型赛道与智能体赛道之间的表现差距，在推理质量指标上比在最终答案准确率上更为明显。这表明，尽管端到端模型有时能够“猜出”正确答案，智能体系统却能提供更加透明、逻辑严谨且可验证的推理路径。不过，顶尖单模型正在迅速缩小差距，这表明其内化推理能力日益成熟。

4.2. 顶级系统分析

4.2.1. 单模型赛道。

单模型赛道要求参赛者在单次前向传播中完成内在推理。表2详细介绍了表现最佳的三支队伍所采用的方法。Qwen3-Omni 的主导地位显而易见，

表2：t的解决方案　音频推理挑战赛中表现最佳的团队　　（单模型赛道）。

Rank	评分标准	准确率	基础模型	后训练方法
1	**65.29**	**74.00**	Qwen3-Omni-Instruct	两阶段强化学习（采用设计的奖励进行 GRPO）
2	62.55	71.00	Qwen3-Omni-Thinking	免训练（采用注意力操控）
3	62.22	71.70	Qwen3-Omni-Thinking	LoRA 监督微调（搭配复杂的数据流水线）

表3：音频推理挑战赛（智能体赛道）中表现最佳团队的解决方案。

Rank	评分标准	准确率	解决方案	新颖性
1	**69.83**	0.769	迭代式证据收集	40+ 集成音频工具
2	66.23	**0.774**	多智能体投票	用于数值任务频谱图分析的视觉语言模型
3	66.09	0.751	多智能体辩论	多智能体辩论与共识

表 4：音频推理挑战赛决赛阶段排行榜。我们报告了在决赛阶段提交结果的队伍在单模型赛道和智能体赛道中的表现。“评分标准”表示实例级、基于评分标准的推理质量得分，“准确率”表示 *MMAR* 中最终答案的准确率（%）。请注意，基于智能体的系统通常取得了更高的推理得分。

	单模型赛道		智能体赛道	
Rank	评分标准（%）	准确率（%）	评分标准（%）	准确率（%）
1	**65.29**	**74.00**	**69.83**	76.90
2	62.55	71.00	66.23	**77.40**
3	62.22	71.70	66.09	75.10
4	60.61	73.40	64.61	72.20
5	58.86	71.20	63.00	71.00
6	58.53	69.30	62.63	73.60
7	57.77	69.40	62.35	72.60
8	57.15	68.10	61.85	72.70
9	56.03	68.40	60.25	77.10
10	47.83	63.60	59.47	68.20
11	47.49	62.50	57.51	69.80
12	38.93	62.60	54.16	70.50
13	23.77	46.40	53.30	71.70
14	19.03	40.40	53.03	68.60
15	-	-	47.84	65.00
16	-	-	23.91	46.40

该系列是所有顶尖参赛方案的骨干网络。

第一名队伍采用渐进式两阶段强化学习策略，取得了 65.29% 的评分标准得分。他们首先在公开问答数据上从头训练“RL-zero”阶段，随后进入“边界增强”阶段，利用困难负例和边界案例，并将思维链路径的语义相似度用作奖励信号。两个阶段均采用了群体相对策略优化（GRPO）[17]，有效地使模型的内部思维过程与类人的推理保持一致。

有趣的是，排名第二的团队采用了免训练的方法。他们通过修改 Qwen3-Omni-Thinking模型中的注意力机制，专门为音频令牌范围内的注意力权重引入缩放因子。这种调整无需更新参数，就能有效促使模型更集中地关注听觉线索，从而缓解多模态模型常见的“浅层浏览”行为。

第三支队伍专注于使用 LoRA 进行高质量监督微调。他们的关键贡献是设计了一套数据合成流程，摒弃了常见的“逆向工程”思维链做法。相反，他们采用“从问题到推理”的标注策略，确保推理路径忠实于音频线索；随后使用双重验证机制，由大语言模型担任裁判，在训练前过滤掉幻觉。

4.2.2. 赛道*2*：智能体赛道。

智能体赛道侧重于系统级推理，其中，编排工具和验证证据的能力至关重要。表3概述了顶尖团队采用的策略。

该赛道的冠军团队构建了一个庞大的智能体系统，集成了 40 多种专用音频工具，评分标准得分达到 69.83%。这些工具涵盖语音和音频处理核心任务（自动语音识别、音源分离、说话人分离）、信号分析（频谱特征、能量动态）以及高阶音乐理论（和弦进行、演唱技巧、节奏模式）。他们的关键创新是采用迭代式证据收集循环。智能体并非按线性流程执行，而是在初始置信度较低时，使用不同工具重新查询音频，从而有效模拟主动聆听。

排名第二的团队通过将音频转换为视觉表示，具体包括梅尔频谱图、CQT 和均方根频谱图，实现了跨模态突破。他们利用视觉语言模型（VLM）分析这些频谱图像，以完成细粒度数值任务（例如，统计事件数量或测量时长），并结合“分流一致性”机制，根据各组成模型之间的一致程度将样本分配给不同的求解器。

排名第三的团队采用了多智能体辩论框架，由不同的智能体生成假设并相互评议，以达成共识，从而显著降低了幻觉率。该系统由中央控制器管理，执行多轮推理，智能体在此过程中生成“同伴意见”并相互评议推理过程，以解决冲突并达成稳健的共识。

5. 结论

Interspeech 2026 音频推理挑战赛标志着音频智能评估的一次关键转变：从简单的准确率转向过程严谨性。通过引入 MMAR 评分标准，我们提供了一项稳定的实例级指标，用于量化思维链质量。对顶尖系统的分析凸显了两种主流范式：智能体赛道展示了通过迭代使用工具和视觉—频谱桥接来分解复杂查询的稳健性；而单模型赛道则展现了通过强化学习和精心设计的数据流程将推理内化的潜力。我们发布相关资源，以加速可信、可解释音频模型的发展。

6. 参考文献

[1] Y. Gong、A. H. Liu、H. Luo、L. Karlinsky 和 J. Glass，“联合音频与语音理解”，载于 ASRU 会议论文集，2023 年。[2] C. Tang、W. Yu、G. Sun、X. Chen、T. Tan、W. Li、L. Lu、Z. Ma 和 C. Zhang，“Salmonn：迈向具备通用听觉能力的大语言模型”，ICLR 会议论文集，2024 年。[3] Y. Chu、J. Xu、X. Zhou、Q. Yang、S. Zhang、Z. Yan、C. Zhou 和 J. Zhou，“Qwen-audio：通过统一的大规模音频语言模型推进通用音频理解”，arXiv 预印本 arXiv:2311.07919，2023 年。[4] Y. Chu、J. Xu、Q. Yang、H. Wei、X. Wei、Z. Guo、Y. Leng、Y. Lv、J. He、J. Lin 等，“Qwen2-audio 技术报告”，arXiv 预印本 arXiv:2407.10759，2024 年。[5] S. Ghosh、S. Kumar、A. Seth、C. K. R. Evuru、U. Tyagi、S. Sakshi、O. Nieto、R. Duraiswami 和 D. Manocha，“Gama：具备先进音频理解与复杂推理能力的大型音频语言模型”，EMNLP 会议论文集，2024 年。[6] Z. Kong、A. Goel、R. Badlani、W. Ping、R. Valle 和 B. Catanzaro，“Audio Flamingo：具备少样本学习和对话能力的新型音频语言模型”，ICML 会议论文集，2024 年。[7] S. Ghosh、Z. Kong、S. Kumar、S. Sakshi、J. Kim、W. Ping、R. Valle、D. Manocha 和 B. Catanzaro，“Audio Flamingo 2：具备长音频理解和专家级推理能力的音频语言模型”，ICML，2025 年。[8] J. Xu、Z. Guo、J. He、H. Hu、T. He、S. Bai、K. Chen、J. Wang、Y. Fan、K. Dang 等，“Qwen2.5-Omni 技术报告”，arXiv 预印本 arXiv:2503.20215，2025 年。[9] Z. Ma、R. Xu、Z. Xing、Y. Chu、Y. Wang、J. He、J. Xu、P.-A. Heng、K. Yu、J. Lin 等，“Omni-captioner：面向全模态精细感知的数据流程、模型和基准测试”，ICLR 会议论文集，2026 年。[10] Z. Ma、Z. Chen、Y. Wang、E. S. Chng 和 X. Chen，“Audio-CoT：探索大型音频语言模型中的思维链推理”，ASRU 会议论文集，2025 年。[11] Z. Xie、M. Lin、Z. Liu、P. Wu、S. Yan 和 C. Miao，“Audio-reasoner：提升大型音频语言模型的推理能力”，EMNLP 会议论文集，2025 年。[12] Z. Ma、Y. Ma、Y. Zhu、C. Yang、Y.-W. Chao、R. Xu、W. Chen、Y. Chen、Z. Chen、J. Cong 等，“MMAR：面向语音、音频、音乐及其混合内容深度推理的挑战性基准测试”，NeurIPS 会议论文集，2025 年。[13] Y. Li、Y. Ma、G. Zhang、R. Yuan、K. Zhu、H. Guo、Y. Liang、J. Liu、Z. Wang、J. Yang 等，“OmniBench：迈向通用全模态语言模型的未来”，NeurIPS 会议论文集，2025 年。[14] S. Kumar、Š. Sedlˊaček、V. Lokegaonkar、F. Lˊopez、W. Yu、N. Anand、H. Ryu、L. Chen、M. Plička、M. Hlavˊaček 等，“MMAU-Pro：用于整体评估音频通用智能的挑战性综合基准测试”，arXiv 预印本 arXiv:2508.13992，2025 年。[15] A. Goel、S. Ghosh、J. Kim、S. Kumar、Z. Kong、S.-g. Lee、C.-H. H. Yang、R. Duraiswami、D. Manocha、R. Valle 等，“Audio Flamingo 3：借助完全开放的大型音频语言模型推进音频智能”，NeurIPS 会议论文集，2025 年。[16] Z. Xie、Z. Ma、Z. Liu、K. Pang、H. Li、J. Zhang、Y. Liao、D. Ye、C. Miao 和 S. Yan，“Mini-omni-reasoner：大型语音模型中的逐词思考与说话”，arXiv 预印本 arXiv:2508.15827，2025 年。[17] D. Guo、D. Yang、H. Zhang、J. Song、R. Zhang、R. Xu、Q. Zhu、S. Ma、P. Wang、X. Bi 等，“DeepSeek-R1：通过强化学习激励大语言模型的推理能力”，Nature 会议论文集，2025 年。[18] G. Li、J. Liu、H. Dinkel、Y. Niu、J. Zhang 和 J. Luan，“强化学习优于监督微调：音频问答案例研究”，arXiv 预印本 arXiv:2503 11197，2025 年。[19] A. Rouditchenko、S. Bhati、E. Araujo、S. Thomas、H. Kuehne、R. Feris 和 J. Glass，“Omni-R1：微调音频大语言模型真的需要音频吗？”，ASRU 会议论文集，2025 年。

[20] G. Wijngaard、E. Formisano、M. Esposito 和 M. Dumontier，《Audsemthinker：通过对声音语义进行推理来增强音频语言模型》，*NeurIPS* 论文集，2025 年。[21] F. Tian、X. T. Zhang、Y. Zhang、H. Zhang、Y. Li、D. Liu、Y. Deng、D. Wu、J. Chen、L. Zhao 等，《Step-audio-r1 技术报告》，*arXiv* 预印本 *arXiv:2511.15848*，2025 年。[22] J. Xu、Z. Guo、H. Hu、Y. Chu、X. Wang、J. He、Y. Wang、X. Shi、T. He、X. Zhu 等，《Qwen3-omni 技术报告》，*arXiv* 预印本 *arXiv:2509.17765*，2025 年。[23] S. Wu、C. Li、W. Wang、H. Zhang、H. Wang、M. Yu 和 D. Yu，《Audio-thinker：通过强化学习指导音频语言模型何时以及如何思考》，*arXiv* 预印本 *arXiv:2508.08039*，2025 年。[24] H. Yang、A. Stolcke 和 L. P. Heck，《基于大语言模型的口语对话智能体》，载于*EMNLP* 教程论文集，2025 年。[25] C.-H. H. Yang、Y. Gu、Y.-C. Liu、S. Ghosh、I. Bulyko 和 A. Stolcke，《基于大语言模型和任务激活提示的生成式语音识别错误纠正》，载于*ASRU* 论文集，2023 年。[26] T. Taheri、Y. Ma 和 E. Benetos，《Sar-lm：基于大语言模型的符号音频推理》，*arXiv* 预印本 *arXiv:2511.06483*，2025 年。

[27] Y. Ro推理器：Li, 一种用于由粗到细音频深度推理的免训练多智能体框架”，*Proc. ICASSP*，2025。

[28] G. Wijngaard、E. Formisano 和 M. Dumontier，“Audiotoolagent：面向音频语言模型的智能体框架”，*arXiv* 预印本 *arXiv:2510.02995*，2025 年。[29] Z. Wan、C.-H. H. Yang、J. Tian、H. Ye、A. Pasad、S.-w. Fu、A. Goel、R. Hachiuma、S. Diao、K. Dhawan 等，“Speech-hands：一种结合自我反思的语音智能体方法，用于结合全模态感知进行语音识别和音频推理”，*arXiv* 预印本 *arXiv:2601.09413*，2026 年。[30] C. Wei、B. Wang、J.-j. Kim 和 N. F. Chen，“迈向口语数学推理：对基于语音的模型在多方面数学问题上的表现进行基准测试”，*arXiv* 预印本 *arXiv:2505.15000*，2025 年。[31] C.-K. Yang、N. Ho、Y.-T. Piao 和 H.-y. Lee，“Sakura：基于语音和音频信息的大型音频语言模型多跳推理研究”，*Interspeech* 论文集，2025 年。[32] W. Yang、Y. Li、Y. Wei、M. Fang 和 L. Chen，“Speechr：大型音频语言模型语音推理基准测试”，*arXiv* 预印本 *arXiv:2508.02018*，2025 年。[33] D. Bhattacharya、A. Kulkarni 和 S. Ganapathy，“大型音频语言模型时间推理的基准测试与置信度评估”，*Interpseech* 论文集，2025 年。[34] Z. Liu、Z. Niu、Q. Xiao、Z. Zheng、R. Yuan、Y. Zang、Y. Cao、X. Dong、J. Liang、X. Chen 等，“Star-bench：以音频四维智能为视角，探究深层时空推理”，*ICLR* 论文集，2026 年。[35] S. Sakshi、U. Tyagi、S. Kumar、A. Seth、R. Selvakumar、O. Nieto、R. Duraiswami、S. Ghosh 和 D. Manocha，“Mmau：大规模多任务音频理解与推理基准测试”，*ICLR* 论文集，2025 年。[36] Y. Ma、S. Li、J. Yu、E. Benetos 和 A. Maezawa，“Cmi-bench：用于评估音乐指令遵循能力的综合基准测试”，*ISMIR* 论文集，2025 年。[37] Z. Wan、C.-H. H. Yang、Y. Yu、J. Tian、S. Li、K. Hu、Z. Chen、S. Watanabe、F. Cheng、C. Chu 等，“Speechiq：大语言模型语音理解在不同认知层级上的语音智能体智商”，载于*ACL* 论文集，2025 年。[38] Y. Yu、H. Jin、Y. Yu、J. Zhuang 和 H. Wang，“现在你听见我了：针对大型音频语言模型的音频叙事攻击”，*EACL* 论文集，2026 年。