UltraEval-Audio：音频基础模型综合评估的统一框架

Qundong Shi[1]*, **Jie Zhou**[1]*, **BiyuanLin**[1],**Junbo Cui**[1],**Guoyang Zeng**[1],**Yixuan Zhou**[1],**Ziyang Wang**[1], **Xin Liu**[1],**Zhen Luo**[1],**Yudong Wang**[2]†, **ZhiyuanLiu**[2]†[1]
ModelBest 公司 [2]清华大学{shiqundong,zhoujie,
linbiyuan}@modelbest.cn{yudongwang,
liuzy}@tsinghu edu.cnhttps://github.com/OpenBMB/UltraEval-Audio

arXiv:2601.01373v1 [cs.SD] 4 Jan 2026

摘要

自 GPT-4o 问世以来，音频基础模型的发展迅速加快。然而，综合评估的缺失已成为该领域进一步发展的关键瓶颈，尤其是在音频生成方面。当前的音频评估面临三大挑战：（1）音频评估缺乏统一框架，数据集和代码分散于不同来源，阻碍了公平、高效的跨模型比较；（2）音频编解码器作为音频基础模型的关键组成部分，缺乏广受认可且全面的评估方法论；（3）现有语音基准高度依赖英语，难以客观评估模型在中文上的表现。针对第一个问题，我们提出了 **UltraEval-Audio**，这是一个面向音频基础模型的统一评估框架，专为音频理解与生成任务设计。UltraEval-Audio 采用模块化架构，支持 10 种语言和 14 个核心任务类别，并无缝集成 24 个主流模型与 36 个权威基准。为提升研究效率，该框架提供一键评估功能，并配套实时公开排行榜。针对第二个挑战，UltraEval-Audio 采用了一种新颖的音频编解码器综合评估方案，从三个关键维度评估其性能：语义准确性、音色保真度和声学质量。针对第三个问题，我们提出了两个新的中文基准——*SpeechCMMLU* 和 *SpeechHSK*，用于评估中文知识水平和语言流利度。我们希望 **UltraEval-Audio**能为学术界和工业界提供一个透明、高效、公平的音频模型比较平台。我们的代码、基准和排行榜已发布于 https://github.com/OpenBMB/UltraEval-Audio。

1 引言

继大语言模型（LLM）取得突破性成功后，多模态大模型迅速发展。值得注意的是，OpenAI 发布 GPT-4o（OpenAI 等人，2024年）标志着原生音频交互时代的到来，并加速了音频基础模型（AFM）的蓬勃发展。随后，一系列具备复杂理解与生成能力的模型相继问世，包括 Qwen2.5-Omni（Xu 等人，2025）、Moshi（Défossez 等人，2024年）、GLM-4-Voice（Zeng 等人，2024年）、Step- Audio（Huang 等人，2025）、Kimi-Audio（Ding 等人，2025）和 MiniCPM-o 2.6（Yao 等人，2024年），显著拓展了人机交互的边界。随着模型能力快速迭代，如何客观、系统地评估这些模型已成为学术界关注的焦点。然而，当前音频评估缺乏统一框架，数据集和代码分散于各种来源，极大地阻碍了模型之间公平、高效的比较。

除了评估框架碎片化的问题之外，现有评估系统在评估深度和语言覆盖范围方面也存在显著局限。传统评估工具

*共同贡献。
†通讯作者。

通常针对特定任务设计，例如自动语音识别（ASR）或自动语音翻译（AST），因此难以适用于具备通用交互能力的音频基础模型（AFM）。提示管理和推理参数调优等领域尤其如此，而这两者对于全面评估音频基础模型至关重要。随着模型能力迅速发展，除了传统的自动语音识别和自动语音翻译任务之外，以用户为中心的语音基准也逐渐兴起。然而，这些基准仍面临以下核心挑战：

(1) 音频编解码器是音频基础模型的支柱，但目前缺乏系统的性能指标。编解码器由音频分词器（将音频转换为离散令牌）和声码器（根据生成的令牌重建音频）组成。音频编解码器的设计直接影响音频表征的保真度和效率，进而显著影响音频基础模型的整体性能（Ye 等人，2025）。现有方法涵盖范围广，但对具体性能维度的洞见有限。
(2) 当前基准测试严重依赖英语，因此难以客观评估模型在中文方面的表现。主流基准测试，例如 SpeechTriviaQA（Défossez 等人，2024年）、SpeechWebQuestions（Nachmani 等人，2023年）和 SpeechAlpacaEval（Fang 等人，2025），主要以英语为中心，导致无法充分衡量模型在中文语境下的知识水平和语言能力。

为应对这些挑战，我们提出了 **UltraEval-Audio**，一个面向音频基础模型的统一评估框架。通过解耦数据加载、提示管理、推理参数控制，以及多样化的后处理和聚合方法，该框架为研究人员提供了统一且灵活的评估环境。用户只需运行简单的“一键式”自动化脚本，即可快速启动评估进程。这种解耦的框架设计不仅提高了实验的可复现性，还便于研究人员快速适配和灵活扩展。此外，UltraEval-Audio 创新性地引入了音频编解码器评估方案和若干中文评估基准，从模型组件和评测基准两方面填补了空白。通过这种全栈整合，UltraEval-Audio 旨在规范并提高整个评估进程的透明度。通过提供实时公开的音频排行榜，该框架致力于推动音频基础模型领域朝着更具可解释性和公平性的方向发展。我们的贡献总结如下：

• **首个统一的音频评估框架**：UltraEval-Audio 支持广泛的输入输出模态，包括“文本 → 音频”、“文本 + 音频 → 文本”、“音频 → 文本”以及“文本 + 音频 →音频”。该框架支持 10 种语言和 14 个核心任务类别，并深度集成了 24 个主流模型和 36 项权威基准测试，涵盖语音、环境声音和音乐三大关键领域。凭借用户友好的设计，该框架提供“一键式”评估功能以及公开可用的排行榜，支持透明比较。

• **面向音频编解码器的多维评估方案**：我们建立了一套系统的评估方案，涵盖语义准确性、音色保真度和声学质量，弥补了音频编解码器领域缺乏广泛认可且系统性的多维性能指标这一不足。

• **两个全新的中文评测基准**：我们提出了 **SpeechCMMLU** 和**SpeechHSK**，旨在系统衡量音频基础模型在中文语境下的知识水平和语言流畅度。

2 相关工作

本节将介绍音频基础模型、音频评估框架、评测基准以及音频编解码器评估的最新发展。

音频基础模型的进展。新兴的音频基础模型通常采用 Audio Codec+LLM 架构，一般由三个核心组件构成：(1) 音频分词器，将原始音频信号转换为离散令牌，同时保留语义和声学信息；(2) 大语言模型骨干网络，负责上下文建模和自回归词元预测；(3) 声码器，根据生成的音频词元合成自然语音波形。根据是否包含声码器，音频基础模型可分为两类：(1) 音频理解基础模型，以音频和文本为输入，仅生成文本作为输出（例如，Qwen-Audio（Chu 等人，2023年、2024）、Gemini-1.5（Team 等人，2024））；(2) 音频生成基础模型，以音频和文本为输入，同时生成语音和文本作为输出（例如，GPT-4o-Realtime（OpenAI 等人，2024）、Moshi（Défossez 等，2024）、MiniCPM-o 2.6（Yao 等人，2024）、Qwen2.5-Omni（Xu 等人，2025）和 Kimi-Audio（Ding 等人，2025））。

与此同时，音频编解码器也在迅速发展。SoundStream（Zeghidour 等人，2021年）是首个能够处理多种音频类型的通用音频编解码器。EnCodec（Défossez 等人，2022）、DAC（descript-audio-codec）（Kumar 等人，2023年）、HiFi-Codec（Yang 等人，2023年）、X-codec（Ye 等人，2025）、BigCodec（Xin 等人，2024）和 BiCodec（Wang 等人，2025年b）进一步提升了重建质量、码本效率以及与基于大语言模型的语音生成的兼容性，体现出音频分词器正朝着可扩展、低延迟和生成式方向发展的明确趋势。

音频评测框架的发展。人们已提出许多用于评估文本和视觉基础模型的综合框架，例如 OpenCompass（贡献者，2023年）、OpenAI Evals [3]、UltraEval（He 等人，2024年）和 VLMEvalKit（Duan 等人，2024年）。然而，针对音频基础模型的综合评估框架一直付之阙如。

在音频基础模型出现之前，音频模型通常针对特定任务进行设计，其评估方式往往临时制定，并且常与模型代码库一同提供。音频基础模型已展现出在各种任务上的强大通用能力，因此，开发一种整合多项任务的综合评估框架变得日益必要。目前已有若干音频评测框架。例如，AudioBench（Chen 等人，2024年）收集了 8 项不同任务和 26 个基准测试，用于评估音频基础模型。AHELM（Lee 等人，2025）汇集了各类数据集，从 10 个方面全面衡量音频基础模型的表现。但这些框架未涵盖音频生成任务。Kimi-Audio-Evalkit（Ding 等人，2025）整合了 Kimi-Audio 评估中提及的所有基准测试，以便复现结果。然而，其评估进程需要五个步骤，使用起来较为繁琐。此外，修改提示词也不方便，因为必须直接在代码中更改，而不能通过配置文件进行修改。AU-Harness（Surapaneni 等人，2025）提供了一个高效的评测引擎，支持 380 多项任务，但要求用户手动将开源音频基础模型适配为标准化的 vLLM 服务。

音频评估基准测试的发展。除传统的自动语音识别和自动语音翻译基准测试外，该领域已开始开发以用户为中心的基准测试：这类基准测试以原始语音为输入，不提供额外的任务描述，并直接评估模型的回复。在音频理解方面，AIR-Bench（Yang 等人，2024年）从现有数据集中收集口语问答样本，并使用 GPT-4 作为自动评估器。VoiceBench（Chen 等人，2024年）进一步拓展了这一方向，既纳入自然口语问答样本，也纳入合成口语指令；后者使用谷歌文本转语音，根据基于文本的指令遵循数据集（例如 AlpacaEval（Li 等人，2023b）、IFEval（Zhou 等人，2023））生成。在音频生成方面，第一个专门的语音问答基准测试 Llama-Question（Nachmani 等人，2023）引入了一个合成语音问答数据集，并采用一种新颖的评测范式：先使用 ConformerASR（Gulati 等人，2020年）将回复音频转写为文本，再评估答案准确率。SpeechWebQuestions（Nachmani 等人，2023）、SpeechTriviaQA（Défossez 等，2024）和 SpeechAlpacaEval（Fang 等，2025）分别源自对应的文本基准测试 WebQuestions（Chen 等人，2015）、TriviaQA（Joshi 等人，2017）和 AlpacaEval。然而，目前所有这些基准测试都仅限于英语，多语言语音基准测试仍基本未得到探索。

音频编解码器评估的发展。 音频编解码器评估采用主观和客观指标。主观评估通常遵循 MUSHRA（系列，2014年）协议，该协议同时使用隐藏参考和低锚点。客观评估包括多种方法：ViSQOL（Hines 等，2015；Chinen 等，2020年）通过测量与真实值的频谱相似度，作为平均意见分的替代指标；尺度不变信噪比（SI-SNR）用于量化重建音频与原始音频之间的相似度，同时忽略信号尺度；梅尔距离用于计算重建波形与真实值波形的对数梅尔频谱图之间的差异；STOI（Taal 等，2011年）用于评估语音可懂度；说话人相似度（SIM）则利用嵌入模型，计算重建音频与真实值的说话人向量之间的余弦相似度。除了这些直接指标，近期研究（如 Ye 等（2025）还采用文本转语音（TTS）等下游任务，间接评估编解码器性能。

3 音频评测设计与方法论

本节概述了针对音频基础模型的系统化音频评估设计与方法论。第 3.1 节介绍了统一的音频评估分类体系，用于组织任务及其对应的基准测试实例。在此分类体系的基础上，第3.2节聚焦于编解码器评估，并提出了一套全面的方法论，用于评估音频编解码器——这是音频基础模型架构的核心组成部分——

[3] https://github.com/openai/evals

表 1： UltraEval-Audio 中的任务分类体系。

类别	领域	Task	描述	指标
音频理解	语音	ASR	给定语音音频，生成转录文本。	词错误率 / 字符错误率
		AST	给定源语言的语音音频，生成目标语言的文本翻译。	BLEU
		性别分析	给定语音音频，预测说话者的性别。	Acc.
		语音问答	给定语音音频，针对相应的问题生成文本答案。	存在匹配 / G-Eval
		情感分析	给定语音音频，识别说话者的情感状态。	Acc.
	音乐	乐器识别	给定音乐音频，判断其中占主导地位的乐器。	Acc.
		音乐流派	给定音乐音频，判断其所属的音乐流派。	Acc.
		和弦识别	给定音乐音频，识别和弦标签序列。	Acc.
	环境声音	音频分类	给定非语音音频，将其分类到预定义的场景或事件类别中。	Acc.
		音频描述	给定非语音音频，生成对其内容的自然语言描述。	BLEU / ROUGE-L
音频生成	语音	TTS	根据输入文本合成相应的语音音频。	自动语音识别词错误率
		VC	根据输入文本和参考语音样本，以目标说话人的声音合成语音。	自动语音识别词错误率 / SIM
		语音问答	根据语音音频生成恰当的口语回复。	存在匹配 / G-Eval / UTMOS
音频编解码器	语音	语音编解码	从离散表示中编码并重建语音音频。	自动语音识别词错误率 / SIM / UTMOS

注： WER：词错误率；CER：字错误率；BLEU/ROUGE-L：文本生成质量；Acc.：分类准确率；SIM：说话人嵌入余弦相似度；UTMOS：客观语音质量评估指标；G-Eval：基于 GPT 的评估指标；ASR-WER：通过自动语音识别模型转录生成或重建的语音，再根据转录文本计算词错误率。

架构。最后，第 3.3 节介绍了两个中文语音基准测试 SpeechCMMLU 和 SpeechHSK，以弥补现有音频评测研究中缺乏中文系统评估资源的问题。

3.1 音频评估分类体系

音频基础模型涵盖广泛的模态、任务、语言和领域。现有评估工作通常聚焦于孤立的任务或特定模态，这凸显了构建统一且可扩展的评估框架的必要性。UltraEval-Audio 并非只是汇总现有基准测试，而是提出了一套可复用、可扩展的评测分类体系，为未来的评测设计提供方法指导。

任务分类体系：评估哪些能力

UltraEval-Audio 采用以能力为导向的任务分类体系，对音频基础模型进行评估。评估任务被划分为多个核心类别，分别体现不同的模型能力和典型的输入输出设置。如表 1所示，该分类体系包含三个高层类别：音频理解、音频生成和音频编解码器。每个类别下的任务还按应用领域进一步组织，包括语音、音乐和环境声音。

音频理解任务用于评估模型从音频输入中提取和解读语义信息的能力。该类别涵盖语音和非语音领域的多个分析层次。在语音领域，任务涵盖从自动语音识别和自动语音翻译等低层次识别，到语音问答等高层次语义推理。在非语音领域，任务包括乐器识别等低层次分类，以及音频描述等高层次理解。

音频生成任务用于评估模型在给定条件下合成或转换音频的能力。在语音领域，任务包括文本转语音、语音克隆（语音转换），以及为语音问答任务生成口语回答。在音频理解任务中，语音问答评估文本回答的准确率；在音频生成任务中，还会进一步从声学、自然度和可懂度等方面评估生成的音频回答的质量。

此外，音频编解码器评估被视为一个独立的任务类别。虽然音频编解码器并非传统的应用任务，但它们在音频基础模型中发挥着重要作用。音频编解码任务会系统地评估编解码器压缩和重建音频的能力，同时保留语义内容和声学特征。

每种任务类别都采用与其目标能力相匹配的标准化评测指标。例如，自动语音识别采用词错误率（WER）或字错误率（CER）；翻译和音频描述任务依赖 BLEU、ROUGE 等文本相似度指标；属性分类和声音分类任务采用准确率；音频生成和编解码器任务则通过基于自动语音识别的 WER 或专门的语音质量指标进行评估。

通过以这种方式组织任务和指标，UltraEval-Audio 提供了可复用且可扩展的评测分类体系，为未来的评测设计提供方法指导，并确保在不同音频领域和模型类型中对能力进行一致且可比较的覆盖。

表 2： B enchmarks 支持的基准测试。* 表示本文引入的新基准测试 r.

Task	语言	数据集
ASR	en	TED-LIUM（Rousseau 等人，2012年），VoxPopuli（Wang 等人，2021年），LibriSpeech（Panayotov 等人，2015） The People's Speech（Galvez 等，2021年）、WenetSpeech（Zhang 等，2022） GigaSpeech（Chen 等，2021年）、AudioMNIST（Srinivasan，2023年）
	zh	KeSpeech（Tang 等，2021年）、AISHELL-1（Bu 等，2017）
	荷兰语、法语、德语、意大利语、波兰语、葡萄牙语、西班牙语	MLS（Pratap 等，2020年）
	中文、英语、俄语、德语、日语、.....	FLEURS（Conneau 等，2022），Common Voice（Ardila 等，2019年）
AST	中文、英语、俄语、德语、日语等.....	CoVoST 2（Wang 等人，2020年）
VC	中文、英语	Seed-TTS-Eval（Anastassiou 等，2024年）、CV3-Eval（Du 等，2025）
TTS	中文、英语	Long-TTS-Eval(Wang 等人，2025a)
语音编解码	en	LibriSpeech
	zh	AISHELL-1
语音问答	en	SpeechTriviaQA (Défossez 等人，2024年), SpeechWebQuestions (Nachmani 等人，2023年) SpeechAlpacaEval（Fang 等人，2025），LLaMA-Questions（Nachmani 等人，2023年） AIR-Bench（Yang 等人，2024年），MMAU（Sakshi 等人，2024年）
	zh	SpeechHSK *，SpeechCMMLU *
情感分析	en	TESS（Dupuis & Pichora-Fuller，2010），MELD（Poria 等人，2018年）
性别分析	en	VoxCeleb(Nagrani 等人，2017)
和弦识别	-	和弦 (deepcontractor，2023年)
乐器识别	-	NSynth (Engel 等人，2017)
音乐流派	-	GTZAN（Sturm，2013年）
音频描述	-	AudioCaps (Kim 等人，2019年),WavCaps (Mei 等人，2024年), Clotho (Drossos 等人，2020年)
音频分类	-	CatDog (Moreaux，2023年), DESED (Turpault 等人，2019年) VocalSound（Gong 等人，2022），COVID-19 Sounds（Dong 等人，2020年） PASCAL CHSC 2011（Bentley 等人），ICBHI 2017 呼吸音（Rocha 等人，2018年）

基准测试实例化：如何衡量能力

基于任务分类体系，UltraEval-Audio 通过精心挑选的一系列广泛采用的基准测试，对每项评估能力进行具体实现。如表 2所示，该框架整合了 36 项基准测试，涵盖 14 项任务和 10 种语言。每项基准测试都明确映射到预先定义的任务类别，从而确保能力定义与具体评估细节保持一致。

在每项任务中，我们汇集多个基准测试数据集，以便对模型能力进行多维、稳健的评估。对于自动语音识别任务，我们支持十多个数据集。这些数据集涵盖从清晰的朗读语音（如 LibriSpeech）到复杂的噪声场景（如 WenetSpeech），并涵盖单语言（如用于中文的 AISHELL- 1）和多语言条件（如 MLS、FLEURS）。这种设计不仅可以评估绝对识别准确率（词错误率/字符错误率），还可以评估模型在不同口音、领域、噪声水平和语言中的泛化能力与稳健性。同样，对于语音问答，我们纳入了多个基于知识和基于指令的基准测试，包括 SpeechTriviaQA 和 SpeechAlpacaEval。

在所有任务中，所选基准测试涵盖多种语言和不同的声学领域，并采用任务分类体系中规定且针对各项任务量身定制的指标进行评估。这种实例化策略确保评估框架全面、可扩展，并与底层能力定义紧密契合，为系统且可复现的评估奠定了清晰的基础。

3.2 编解码器评估

音频编解码器是音频基础模型的重要组成部分，其设计会直接影响音频表征的保真度和模型整体性能。然而，现有评估方法采用的指标各异、标准不一致，且缺乏系统、可比较的评估框架。为此，我们提出一种涵盖**语义**、**音色保真度**和**声学质量**的三维编解码器评估方法论。

对于语义，我们使用词错误率来衡量重建音频保留原始内容的效果。具体而言，我们使用高性能自动语音识别模型对重建音频进行转录，并将结果与原始转录文本进行比较。我们对英语使用 Whisper-large-v3，对中文使用 Paraformer-zh。对于音色保真度，我们

使用经过说话人验证微调的 WavLM-large 提取音频嵌入，并计算原始音频与重建音频嵌入之间的余弦相似度。该指标量化编解码器保留说话人特征的能力。对于声学质量，我们结合使用 UTMOS（Saeki 等人，2022）预测整体自然度，并使用 DNSMOS P.835（Reddy 等人，2022）和 P.808（Reddy 等人，2021年）评估噪声环境中的语音质量。
这一三维评估框架从内容、说话人特征和感知质量三个方面提供互补视角，从而能够更全面、可靠地表征编解码器性能及其潜在局限，并为模型开发与优化提供细粒度的诊断工具。

3.3 中文基准

随着音频基础模型的发展，语音评估已从词错误率（WER）等传统低层级指标转向更高层级的能力，包括复杂语义理解和知识推理。然而，现有的高层级语音基准测试大多聚焦于英语，例如 SpeechTriviaQA、SpeechWebQuestions 和 SpeechAlpacaEval，中文领域的相关资源则相对匮乏。为弥补这一空缺，我们推出了两个新基准：*SpeechCMMLU* 用于评估模型理解中文世界知识任务并进行推理的能力，而*SpeechHSK* 用于衡量中文语言水平。

SpeechCMMLU 通过系统化的语音合成，将广受认可的中文知识推理基准 CMMLU（Li 等人，2023年a）扩展至音频模态。其构建与质量保证流程如下：

1. **文本指令构建**。每道选择题都被整理为统一指令，将题干、选项和作答限制整合为标准化提示词。
2. **音频指令合成**。考虑到人工录制成本高昂，且数据集规模庞大（11,583 条），我们采用高质量文本转语音模型 CosyVoice2 生成语音提示。该模型在中文发音和专业术语方面表现稳定，适用于大规模语音数据集的构建。3. **自动化质量控制**。为确保语义忠实度，所有合成音频均通过自动语音识别流程转录。只有转录文本与原文完全一致的样本（CER = 0%）才会被保留，从而有效消除文本转语音合成过程中可能引入的发音错误或术语偏差。

按照这一流程，发布的 SpeechCMMLU 数据集包含 3,519 条高质量语音样本，涵盖多种学科领域，适用于评估模型在专业领域中的中文知识理解与推理能力。

SpeechHSK采用由中国教育部设立的标准化全球中文水平测试——汉语水平考试（HSK）。该考试以其权威性和科学性广受认可。SpeechHSK 将 HSK 的听力理解部分转化为音频基准，为衡量模型的中文语言能力提供分级且实用的评估方式。

- **数据来源与构建**。该基准测试取材于官方汉语水平考试听力理解题。每个样本由一道问题和四个选项组成，其中问题音频采用原始考试录音，而选项最初以文本形式提供。为将文本选项用于语音评估，中文母语者会重新录制选项，并确保音频自然流畅。
- **难度等级结构**。该数据集遵循汉语水平考试的六个熟练度等级（一级：初级；六级：高级），且等级越高，问题数量越多，共计 170 个样本。这种结构支持对模型的中文理解进行循序渐进的诊断性评估。

SpeechCMMLU 和 SpeechHSK 共同构成了面向中文语音的多层次评测套件。SpeechCMMLU 侧重于知识推理，而 SpeechHSK 则强调综合语言熟练度。这两个基准测试均通过自动化流程构建，并经过严格的质量控制，以确保可复现性和可扩展性。

4 UltraEval-Audio框架

本节介绍 UltraEval-Audio 的工程架构。为应对音频基础模型评估中因输入模态和模型接口的多样性而产生的集成挑战，该

框架采用解耦的模块化设计，并结合配置驱动的工作流。这种设计使各组件能够独立开发、灵活扩展并高效复用。所有数据集、模型和评估流程均通过统一的 YAML 配置文件以声明方式进行指定，从而无需手动编写脚本即可实现端到端评估。

如图1所示，该框架由三个主要模块组成：数据准备、模型部署和评估。以下小节将详细介绍各个组成部分。第4.1节详细介绍数据加载器和提示管理，第4.2 节介绍模型部署机制，第4.3节介绍评估流程和自动评分模块。

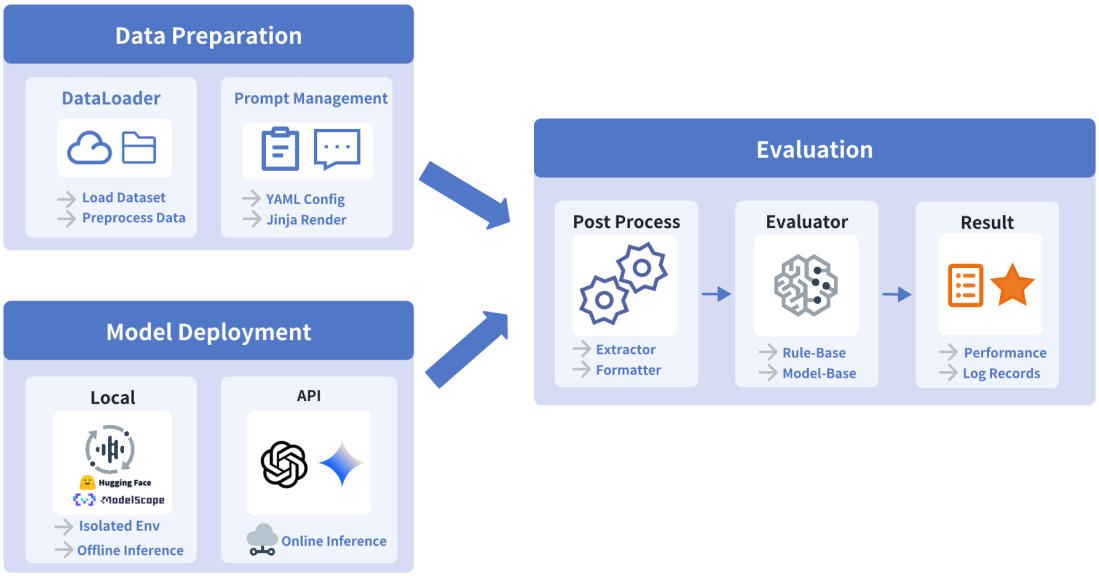

图1： UltraEval-Audio框架概览

4.1 数据准备

数据加载器。数据加载和管理构成了音频评估的基础设施。与结构通常较为简单的文本数据集不同，音频数据集涉及音频信号与文本标注之间的紧密耦合。这种耦合带来了显著的数据管理挑战，包括管理分散的文件路径、处理多种音频编码格式，以及统一不一致的采样率。

为应对这些挑战，UltraEval-Audio采用以音频为中心的数据集组织方案。在这种设计中，使用一个保留字段*audio* 来存储音频内容。该框架提供内置的自动化流程，可在加载过程中动态管理音频资源的获取，包括从HuggingFace等开源仓库自动下载和缓存，以及进行解码和格式标准化。

该框架将复杂的多模态预处理解耦为标准化的数据接口，大幅简化了准备流程，并确保后续模型推理与评估能够高效执行。

提示词管理。 提示词在引导模型生成相关任务输出方面起着关键作用。提示设计不仅与模型输入格式有关，也会直接影响评估结果的稳定性和可比性。然而，由于训练策略不同，模型往往对提示词高度敏感，例如对特殊标记或角色定义的使用，并且这种敏感性会因任务而有很大差异。模型特定的敏感性与任务相关的差异相结合，使得硬编码提示词难以复用，也损害了不同模型之间公平比较的基础。

为解决这一问题，UltraEval-Audio 引入了一种配置驱动的提示管理机制。提示结构通过基于 YAML 的配置定义，并结合 Jinja模板引擎支持变量注入和条件逻辑。这种设计使研究人员能够在任务、模型乃至样本层面灵活构建提示。因此，研究人员可以定义或替换提示模板

以适配不同任务，而无需修改任何代码，从而大幅提升评估工作流的可维护性和可扩展性。

在实际应用中，UltraEval-Audio 的提示管理模块由三个组件组成：

• **提示词注册表**，用于注册和索引与不同任务关联的提示模板。

• **YAML解析器**，用于解析声明式配置文件，并加载特定模型或任务的提示词定义。

• **模板渲染器**，利用基于 Jinja 的渲染，将数据集字段（例如 {{audio}}、{{ques- tion}}、{{choice_a}}）注入模板并生成最终的模型输入。

这种机制不仅支持固定模板，还支持涉及动态变量和条件分支的更复杂提示构建。以下是一个 MiniCPM-o 2.6 自动语音识别提示词示例：

mini-cpm-omni-asr-en: class: audio_evals.prompt.base.Prompt args: template: - role: user contents: - type: text value: '请仔细聆听音频片段并转录其中的内容。请用小写字母输出。' - type: audio value: '{{audio}}'

{{audio}} 占位符会在推理时动态替换为相应数据样本中的音频字段。若要适配不同的自动语音识别模型，用户只需定义针对特定模型的提示词，并在运行时通过*--prompt$prompt* 参数选择该提示词，无需修改任何代码。

对于需要依赖样本构建提示词结构的任务，例如选项数量可变的多项选择题，模板可以加入条件渲染逻辑。示例如下：

single_choice_extended: class: audio_evals.prompt.base.Prompt args: template: - role: user contents: - type: audio value: "{{audio}}" - type: text value: "请从选项 A、B、C、D{% if choice_e isdefined and choice_e %} 和 E{% endif %} 中选择最合适的答案来回答以下问题。你只能选择 A、B、C 或 D{% if choice_e is defined and choice_e %} 或 E{% endif %}。{{question}} A. {{choice_a}} B. {{choice_b}} C. {{choice_c}} D. {{choice_d}}{% if choice_e is defined and choice_e %} E. {{choice_e}}{% endif %}"

提示管理机制通过配置驱动设计、模板复用和动态渲染，实现统一的自动化评估。一方面，它将提示设计与核心评估逻辑彻底解耦，降低不同任务和模型之间的集成开销。另一方面，它确保提示词定义可复现、可扩展且易于共享，为音频基础模型的标准化评估奠定了坚实基础。

4.2 模型部署

模型部署是连接数据与评估的关键环节。音频模型通常以两种形式之一进行部署：（1）远程 API 模型，通过官方软件开发工具包或 HTTP API 端点执行推理；（2）本地部署模型，依靠本地硬件资源执行推理。这两类模型在环境依赖、硬件要求和执行模式方面的差异，为构建统一评测框架带来了重大工程挑战。

对于本地部署模型而言，这一挑战尤为突出，依赖冲突是常见问题。不同模型可能需要彼此不兼容的深度学习框架、驱动程序或音频处理库版本。例如，音频质量评估模块（如 UTMOS、SIM、DNSMOS）和自动语音识别词错误率评估器通常依赖特定的环境版本，而这些版本可能与被评估模型的依赖栈冲突。在同一运行时环境中混用多个模型，可能导致环境污染或推理失败，从而损害评估结果的可复现性。

为应对这一挑战，UltraEval-Audio 引入了**隔离运行时**机制，确保环境独立，并在系统层面安全执行。其核心设计特点如下：

1. **环境级隔离**：为每个待评估模型分配独立的虚拟运行时环境，其中仅包含该模型所需的依赖项。该环境与主评估进程完全隔离，从根本上消除依赖冲突。2. **基于子进程的模型执行**：模型以守护进程模式作为独立子进程运行，并保持持久加载状态，以支持连续推理，显著减少频繁加载模型带来的开销。3. **进程间通信（IPC）**：主评估进程通过系统管道与各模型子进程通信，以安全、低延迟的方式交换数据和推理结果。

在设计层面，这一机制有效实现了微服务式推理架构。用户无需手动管理或安装额外依赖，系统也显著降低了大规模多模型评估的工程复杂度。此外，为统一不同模型类型的接口，框架在部署层提供了统一的*.inference()*方法，该方法接收来自提示管理模块的输入，并返回推理结果。

通过这一设计，UltraEval-Audio 在系统层面实现了模型执行的完全解耦与自动化管理。一方面，隔离运行时消除了模型之间的依赖冲突和环境污染，确保评估结果具有可复现性和一致性。另一方面，统一的推理接口对底层模型实现进行了抽象，降低了跨模型比较和框架扩展的技术门槛。总体而言，该模块为大规模音频模型评估提供了稳定、可扩展且透明的运行时基础。

4.3 评估

模型推理完成后，UltraEval-Audio 进入评估阶段，将模型原始输出转换为结构化、可量化的结果，以便客观评估不同模型的性能。评估工作流由两个核心组件构成：后处理和评估器。后处理模块对模型输出进行标准化和语义对齐，评估器则根据任务计算指标。这种两级结构构建了一条从输出生成到指标计算的全自动标准化流水线。

后处理。 在多模态音频评估中，模型输出通常会包含与任务目标无关的多余前缀或后缀、格式伪影或非结构化文本，从而影响指标计算的准确性。为确保用于评估的输入保持一致且结构良好，UltraEval-Audio 采用了灵活的后处理机制，可在多个层级对输出进行解析和标准化。

该机制采用模块化和可组合设计，可根据任务规范实现自适应工作流。该框架包含多个内置后处理模块，例如选项提取、是/否解析器和 *JSON* 字段解析器，这些模块可以按顺序组合，形成多步流水线，以处理复杂任务。

评估器。评估器模块会处理经过后处理标准化的模型输出，并负责计算最终的性能指标。UltraEval-Audio 将评估器分为两类：

表3： 参与评估的音频基础模型概览

模型	机构	Type	模态 （输入→输出）	语言
GPT-4o-Realtime	OpenAI	专有	音频 + 文本 → 音频 + 文本	多语言
Qwen3-Omni-30B-A3B-Instruct	阿里巴巴	开源	音频 + 文本 → 音频 + 文本	多语言
Qwen2.5-Omni	阿里巴巴	开源	音频 + 文本 → 音频 + 文本	英语、中文
MiniCPM-o 2.6	OpenBMB	开源	音频 + 文本 → 音频 + 文本	英语、中文
Kimi-Audio-7B-Instruct	Moonshot	开源	音频 + 文本 → 音频 + 文本	英语，中文
Gemini-1.5-Flash	谷歌	专有	音频 + 文本 → 文本	多语言
Gemini-1.5-Pro	谷歌	专有	音频 + 文本 → 文本	多语言
Gemini-2.5-Flash	谷歌	专有	音频 + 文本 → 文本	多语言
Gemini-2.5-Pro	谷歌	专有	音频 + 文本 → 文本	多语言
Qwen2-Audio-7B	阿里巴巴	开源	音频 + 文本 → 文本	多语言
Qwen2-Audio-7B-Instruct	阿里巴巴	开源	音频 + 文本 → 文本	多语言
MiDaShengLM-7B	小米	开源	音频 + 文本 → 文本	多语言
GLM-4-Voice	智谱AI	开源	音频 → 音频	英语、中文

• **基于规则的评估器**：对于具有参考答案的任务，采用经典的算法指标。该框架包括用于自动语音识别任务的词错误率、用于自动语音翻译的 BLEU、用于分类任务的准确率，以及其他类似的度量指标。

• **基于模型的评估器**：对于音频生成质量或开放式生成任务，会使用预训练评估模型来近似人类判断。这包括用于评估音色或语音克隆保真度的说话人相似度（SIM）、用于评估语音自然度和感知质量的 UTMOS 和 DNSMOS，以及基于GPT的开放域问答评分器（大语言模型裁判）。

所有评估器均通过统一接口进行注册和编排，其输出会自动汇总为汇总结果摘要。这种设计确保了不同任务的各项指标之间具有兼容性，同时便于快速集成新的评估器，使框架具有高度的模块化和可维护性。

通过采用标准化后处理和多维评估框架，UltraEval-Audio 建立了从模型输出到可量化指标的端到端自动化流水线。这种设计提升了一致性、透明度和可复现性，为多模态音频基础模型的客观比较与持续发展奠定了系统化基础。

5 评估结果

UltraEval-Audio提供了一套统一解决方案，可在各种测试环境中系统评估音频处理模型的性能。本节中，我们精心选取了具有代表性的评估任务，并构建了三个排行榜：音频理解、音频生成和音频编解码器排行榜。随后，我们对第 5.1 节介绍的 13 个领先音频基础模型和 9 种音频编解码器进行评估。我们分别在第 5.2、5.3 和5.4 节展示音频理解、音频生成和音频编解码器排行榜。

5.1 评估模型

我们评估了新兴的音频基础模型和音频编解码器，并据此构建排行榜。表 3展示了参与评估的音频基础模型，包括 GPT-4o-Realtime 和 Gemini-2.5-Pro 等领先的专有模型，以及 Qwen3-Omni-30B-A3B-Instruct 和 Kimi-Audio-7B-Instruct 等最新开源模型。音频编解码器包括 Encodec（Défossez 等人，2022）、ChatTTS-DVAE[4], Mimi（Défossez 等人，2024年5，75 75 40 40[6] 2024年）系列、WavTokenizer-large-speech-token（记作 WavTokenizer-large-）和 WavTokenizer-large-unify-token（记作 WavTokenizer-large-）（Ji 等人，）以及 Spark（Wang 等人，2025年b）。

[4]https://github.com/2noise/ChatTTS [5]
https://huggingface.co/novateur/WavTokenizer-large-speech-75token [6]
https://huggingface.co/novateur/WavTokenizer-large-unify-40token

若干模型会针对特定任务设计不同的提示词和参数。例如，Kimi-Audio-7B-Instruct 针对自动语音识别任务设有专门的提示词。为复现这些结果，我们使用发布者提供的提示词和参数运行各项任务。对于发布者未提供这些配置的任务，我们使用官方推理参数和提示词进行评估。此外，我们不进行额外的提示词或参数优化，以确保评估方案公平且一致。

5.2 音频理解

表4： 音频理解排行榜。最佳结果以粗体显示。平均得分是所有可用指标得分的平均值，其中基于词错误率的指标使用（100 –词错误率），其他指标（例如 BLEU/准确率）保持不变。

模型	ASR						AST		EMO	平均得分(↑)
	LibriSpeech 开发集-干净语音 | 开发集-其他语音 测试集-干净语音 | 测试集-其他语音 TED-LIUM		CV-15 英语 | 中文	AISHELL-1	FLEURS	Wenet -test-net	covost2-en2zh	covost2-zh2en	MELD	
	词错误率（↓）	词错误率（↓）	词错误率 | 字符错误率（↓）	字符错误率（↓）	字符错误率（↓）	字符错误率（↓）	BLEU（↑）	BLEU（↑）	准确率（↑）	
GPT-4o-Realtime	2.30 | 5.60 2.60 | 5.50	4.80	27.44 | 37.44	7.30	5.40	28.90	37.10	15.70	33.20	73.75
Qwen3-Omni-30B-A3B-Instruct	1.25 | 2.27 1.36 | 2.57	2.82	**6.00 | 4.32**	0.87	2.61	**4.82**	46.58	**29.40**	56.81	**84.92**
Qwen2.5-Omni	2.10 | 4.20 2.40 | 4.20	4.70	8.70 | 5.20	1.10	4.60	6.00	42.50	11.50	53.60	81.88
MiniCPM-o 2.6	1.60 | 3.40 1.70 | 4.40	3.00	10.30 | 9.60	1.60	4.40	6.90	**48.20**	27.20	52.40	83.15
Kimi-Audio-7B-Instruct	**1.18 | 2.34** **1.28 | 2.44**	2.96	7.09 | 5.72	**0.60**	**2.53**	5.55	36.61	18.30	**59.23**	83.27
Gemini-1.5-Flash	5.90 | 7.20 21.90 | 16.30	6.90	208.00 | 84.37	9.00	85.90	279.90	33.40	8.20	45.20	27.80
Gemini-1.5-Pro	2.60 | 4.40 2.90 | 4.90	3.00	8.36 | 13.26	4.50	5.90	14.30	47.30	22.60	48.40	81.09
Gemini-2.5-Flash	3.73 | 6.71 3.28 | 12.03	3.53	46.76 | 36.15	6.40	6.45	126.07	3.67	10.61	51.53	62.67
Gemini-2.5-Pro	5.30 | 4.51 2.84 | 6.74	**2.52**	9.42 | 11.04	3.36	4.25	16.83	41.75	27.84	46.59	80.72
Qwen2-Audio-7B	1.57 | 3.50 1.60 | 3.88	3.43	8.67 | 7.03	1.52	5.89	8.09	45.30	24.84	42.87	82.14
Qwen2-Audio-7B-Instruct	2.90 | 5.50 3.10 | 5.70	5.90	10.68 | 8.39	2.60	6.90	10.30	39.50	22.90	17.40	78.29
MiDaShengLM-7B	2.20 | 4.75 2.21 | 5.16	146.53	13.66 | 29.13	1.23	3.28	16.56	38.52	22.68	53.96	68.50

注： 当识别错误总数超过参考文本中的词数或字符数时，词错误率/字符错误率可能会超过 100。

对于音频理解，我们选取了成熟且广受引用、并常用于现有论文的基准测试。具体而言，我们使用 **LibriSpeech** （英语）、**TED-LIUM**（英语）、**Common Voice 15** （英语/中文）、**AISHELL-1** （中文）、**FLEURS** （中文）、**WenetSpeech-test-net**（中文）进行自动语音识别，使用 **covost2-en2zh**、**covost2-zh2en** 进行自动语音翻译，并使用 **MELD** 进行情感识别（情感识别）。最终结果如表 4所示，据此我们可以得出以下主要观察结果：

（1）GPT-4o-Realtime 在音频理解领域面临着激烈竞争，Qwen3-Omni-30B-A3B-Instruct、Kimi-Audio-7B-Instruct、MiniCPM-o 2.6 等开源模型，以及 Gemini-2.5-Pro 等专有模型在评估中表现更优。造成这一现象的一个关键原因是 GPT-4o-Realtime 在中文自动语音识别基准测试中的表现相对不佳，尤其是在 Wenet-test-net 和 Common Voice 15（CV-15）等数据集上的评估结果。

(2) Qwen3-Omni-30B-A3B-Instruct 在所有任务中均展现出卓越性能，在不同领域始终能提供高质量结果。Kimi-Audio-7B-Instruct 在自动语音识别和情感识别任务中表现出色，但在自动语音翻译任务中的表现较弱，说明该领域仍有提升空间。

(3) 对于 Gemini 系列模型，我们观察到每一代的 Flash 模型表现都弱于 Pro 模型。不过，与 Gemini 1.5 相比，Gemini 2.5 的 Flash 模型表现更好，而 Pro 模型的表现相近。

5.3 音频生成

为构建音频生成排行榜，我们选用**SpeechWebQuestions**、**SpeechTriviaQA**和**SpeechAlpacaEval** 来评估英语能力。我们还使用自主提出的基准测试**SpeechCMMLU** 和**SpeechHSK**来评估中文音频生成能力。我们首先根据答案准确率（准确率）评估模型在各数据集上的表现，然后评估生成音频的声学质量。

为更好地与现有评估结果保持一致，我们针对不同数据集采用不同的评估设置。具体而言：

表 5： 音频生成排行榜。声学指标（UTMOS | DNSMOS P.835 | DNSMOS P.808，分数范围为 0 至 5 ）基于语音任务中的生成的音频回答进行评估。最佳结果以粗体显示。

模型	语音 WebQuestions	语音 TriviaQA	语音 AlpacaEval	语音 CMMLU	语音 HSK	声学	平均得分(↑)
	准确率 (↑)	准确率 (↑)	准确率 (↑)	准确率 (↑)	准确率 (↑)	声学 (↑)	
GPT-4o-Realtime	**51.60**	**69.70**	**74.00**	70.05	**98.69**	4.29 | 3.44 | 4.26	**74.00**
Qwen3-Omni-30B-A3B-Instruct	51.50	55.27	67.97	47.83	40.27	**4.44** | 3.45 | 4.12	57.15
Qwen2.5-Omni	38.89	39.94	54.00	**73.72**	95.65	4.23 | **3.48** | **4.27**	63.68
MiniCPM-o 2.6	40.00	40.20	51.00	51.37	80.68	4.12 | 3.39 | 4.02	56.69
Kimi-Audio-7B-Instruct	33.69	38.20	34.40	71.25	97.42	2.94 | 3.22 | 3.62	56.69
GLM-4-Voice	32.00	36.40	51.00	52.61	71.06	4.21 | 3.46 | 4.07	53.56

注：平均分是以下 6 项分数的平均值：五项语音任务分数和归一化声学分数。对于声学分数（UTMOS | DNSMOS P.835 | DNSMOS P.808），每项分数（0–5）乘以 20，映射至 0–100，然后取平均值，得到归一化声学分数。

• 对于**SpeechWebQuestions** 和**SpeechTriviaQA**，我们遵循（Nachmani 等人，2023年）和（Défossez 等人，2024年）中的评估方法。我们使用 Whisper-large-v3 转录模型的语音回答；如果转录文本中包含标准答案，则认为回答正确。

• 对于**SpeechAlpacaEval**，我们采用（Zengetal.，2024年）中的评估方案，使用GPT-4o-mini评估转录回答的质量。回答按照 MT-Bench 评分标准（Zheng et al.，2023年）进行评分，分数范围为 1 到 10。

• 对于 **SpeechCMMLU** 和 **SpeechHSK**，我们使用 Paraformer-zh 转录生成的音频。随后将转录文本与多项选择选项进行匹配，以计算准确率。

我们进一步使用 UTMOS、DNSMOS P.835 和 DNSMOS P.808 作为指标，评估上述基准测试中生成的音频回答的声学数值。所有模型的表现汇总于表5，主要发现如下：

(1) GPT-4o-Realtime 在音频生成方面表现最佳，尤其在英语基准测试中表现突出。它始终能够生成高质量、准确且听起来自然的语音，使其成为表现最优的模型之一；
(2) Qwen3-Omni-30B-A3B-Instruct 和 Qwen2.5-Omni 在声学指标方面优于 GPT-4o-Realtime，展现出更出色的声音质量。这些模型能够生成更细腻的音频，提供更加丰富且逼真的语音输出；

(3) Kimi-Audio-7B-Instruct 在声学质量方面表现欠佳，其生成的语音缺乏一定的自然度和清晰度。这表明仍需改进，以使输出听起来更加自然、更具人类感。

5.4 音频编解码器

我们使用纯净语音语料库评估音频编解码器，包括**LibriSpeech-dev-clean** （英语）、**LibriSpeech-test-clean** （英语）和**AISHELL-1** （中文）。对于每个数据集，我们使用 ASR-WER/CER指标来衡量编解码器准确率，使用 SIM（相似度的缩写）来衡量音色保真度，并使用声学得分来衡量声学质量。结果如表6所示，我们观察到以下几点：

(1) 各模型在自动语音识别词错误率方面的表现差异有限，而音色保真度和声学质量则存在显著更大的差异。后两项指标为评估编解码器性能提供了更具区分度的信号。

(2) Mimi模型在编解码器准确率和音色保真度方面表现最佳，表明其令牌表示能够有效捕捉原始音频中的语言信息和音色信息。然而，其声学得分低于Spark和WavTokenizer-large-75，说明其解码器仍有改进空间。

（3）我们进一步提出图2来比较 Mimi编解码器变体。比较 Mimi（默认32位）与 Mimi（8位）可见，性能下降在音色保真度方面最为明显，而自动语音识别词错误率略有下降，声学质量小幅下降。这表明，音色信息更加依赖较高的位深度。流式变体在音色保真度和声学质量方面均进一步下降。

表6： 音频编解码器排行榜。连字符（-）表示 UTMOS 不适用于中文语音（AISHELL-1）。最佳结果以粗体显示。

模型	LibriSpeech-开发集-干净语音			LibriSpeech-测试集-干净语音			AISHELL-1			平均得分(↑)
	自动语音识别词错误率 (↓)	SIM (↑)	声学 (↑)	自动语音识别词错误率 (↓)	SIM (↑)	声学 (↑)	ASR-CER (↓)	SIM (↑)	声学 (↑)	
Encodec-24k	4.56	59.40	1.58 | 3.12 | 2.36	4.32	59.40	1.57 | 3.12 | 2.36	13.95	47.48	- | 2.93 | 2.03	65.24
Encodec-48k	3.85	65.53	1.52 | 2.88 | 2.42	3.80	66.00	1.48 | 2.87 | 2.40	6.85	68.78	- | 2.79 | 2.21	69.59
ChatTTS-DVAE	7.49	34.83	1.30 | 2.66 | 2.11	6.75	36.21	1.29 | 2.64 | 2.12	32.36	32.36	- | 2.24 | 1.57	52.86
Mimi（32位）	**2.04**	**92.18**	3.83 | 2.87 | 2.44	**1.96**	**92.68**	3.84 | 2.92 | 2.49	**2.82**	**84.80**	- | 2.43 | 1.89	80.96
Mimi（8位）	2.76	72.15	3.52 | 2.78 | 2.37	2.83	73.13	3.53 | 2.83 | 2.43	6.82	60.63	- | 2.42 | 2.04	72.72
Mimi-Streaming（8 位）	6.76	54.02	1.65 | 2.78 | 2.37	6.19	54.32	1.63 | 2.83 | 2.43	19.62	40.67	- | 2.42 | 2.04	61.37
WavTokenizer-large-75	4.31	69.97	4.01 | 3.64 | **3.26**	4.05	68.15	4.00 | 3.63 | **3.27**	8.97	64.27	- | 3.11 | **2.85**	76.67
WavTokenizer-large-40	8.13	60.26	3.78 | 3.70 | 3.13	7.73	56.63	3.77 | 3.70 | 3.16	25.52	49.21	- | 3.13 | 2.50	69.18
Spark	2.39	79.94	**4.18** | **3.85** | 3.24	2.53	79.53	**4.18** | **3.83** | 3.24	3.66	74.76	- | **3.63** | **2.85**	**82.29**

注：对于声学得分，我们还使用 UTMOS、DNSMOS P.835 和 DNSMOS P.808 指标。为计算平均得分，对于自动语音识别词错误率和 ASR-CER，我们计算 $100 - \text{val}$。对于声学得分，每个可用值（范围为 0 到 5）都通过 $20 \times \text{val}$ 进行归一化（映射至 0–100），声学得分为这些值的平均值（忽略连字符"-"）。最终得分为 9 个指标得分的平均值。

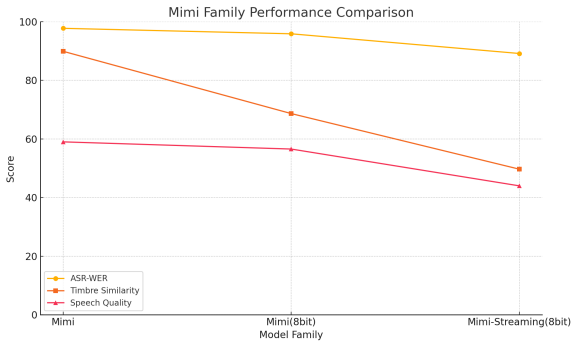

图 2： Mimi 系列性能对比。请注意，自动语音识别词错误率按 $(100 - \text{WER})$ 归一化，语音质量得分（声学得分）则乘以 20 进行缩放，以便于可视化。

(4) ChatTTS-DVAE、WavTokenizer-large-40 和 Mimi-Streaming（8bit）在 AISHELL-1 数据集上的表现不佳，表明仍需改进对中文语音的处理能力。

6 结论

本文介绍了 UltraEval-Audio，这是首个用于全面评估音频基础模型的统一评测框架。我们首先构建了完整的音频评测分类体系，涵盖各类任务和基准测试。为丰富评估内容，我们制定了系统的方法论来评估音频编解码器，并推出了两个中文基准：SpeechCMMLU 和 SpeechHSK。在此基础上，我们实现了模块化评估框架，并利用该框架评估和分析了热门模型在音频理解、音频生成和音频编解码任务上的表现。

7 局限性与未来方向

我们的局限性如下。首先，目前的一些语音基准将转录文本而非原始音频作为基于GPT的评测器的输入。这种设计依赖自动语音识别的性能，可能会使转录错误传递到下游判断中。因此，未来的工作应探索能够直接处理原始音频信号的评估流程。此外，现有评测指标主要侧重于技术层面，未能充分涵盖人类感知因素，例如韵律、情感，以及系统回复的语气是否适合特定的对话语境。

未来的工作中，我们计划持续更新和完善排行榜，提升推理能力（例如多GPU支持），并纳入可直接根据原始音频评估响应的方法。这些改进将提升音频基础模型评估的全面性和可靠性，为该领域的发展提供更明确的指引。

参考文献

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao 等。Seed-TTS：一系列高质量、多用途的语音生成模型。*arXiv* 预印本 *arXiv:2406.02430*，2024年。Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers 和 Gregor Weber。Common Voice：一个大规模多语言语音语料库。*arXiv* 预印本 *arXiv:1912.06670*，2019年。P. Bentley, G. Nordehn, M. Coimbra 和 S. Mannor。PASCAL 2011 心音分类挑战赛（CHSC2011）结果。http://www.peterjbentley.com/heartchallenge/index.html。Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu 和 Hao Zheng。AISHELL-1：一个开源普通话语音语料库及语音识别基线。载于*2017*年东方分会国际语音数据库与语音输入/输出系统及评估协调委员会第 *20* 届会议（*O-COCOSDA*），第 1–5 页。IEEE，2017。Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang 等。GigaSpeech：一个持续发展的多领域自动语音识别语料库，包含 10,000 小时的转录音频。*arXiv* 预印本 *arXiv:2106.06909*，2021年。Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár 和 C Lawrence Zitnick。微软 COCO 图像描述：数据收集与评估服务器。*arXiv* 预印本 *arXiv:1504.00325*，2015。Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan 和 Haizhou Li。VoiceBench：基于大语言模型的语音助手基准测试。*arXiv* 预印本 *arXiv:2410.17196*，2024年。Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman 和 Andrew Hines。ViSQOL v3：一款开源、可用于生产环境的客观语音和音频度量指标。载于*2020*年第十二届多媒体体验质量国际会议（*QoMEX*），第 1–6 页。IEEE，2020年。

Yunfei Chu、Jin Xu、Xiaohuan Zhou、Qian Yang、Shiliang Zhang、Zhijie Yan、Chang Zhou 和 Jingren 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，2023年。网址 https://arxiv.org/abs/2311.07919。

Yunfei Chu、Jin Xu、Qian Yang、Haojie Wei、Xipin Wei、Zhifang Guo、YichongLeng、Yuanjun Lv、Jinzheng He、Junyang Lin、Chang Zhou 和 Jingren Zhou。Qwen2-audio 技术报告，2024年。网址 https://arxiv.org/abs/2407.10759。

Alexis Conneau、Min Ma、Simran Khanuja、Yu Zhang、Vera Axelrod、Siddharth Dalmia、Jason Riesa、Clara Rivera 和 Ankur Bapna。《Fleurs：通用语音表征的少样本学习评估》。*arXiv*预印本 *arXiv:2205.12446*，2022。网址：https://arxiv.org/abs/2205.12446。OpenCompass Contributors。《OpenCompass：基础模型的通用评测平台》。https://github.com/open-compass/opencompass，2023。deepcontractor。《乐器和弦分类》。https://www.kaggle.com/datasets/deepcontractor/musical-instrument-chord-classification，2023。访问日期：2025年5月13日。Alexandre Défossez、Laurent Mazaré、Manu Orsini、Amélie Royer、Patrick Pérez、Hervé Jégou、Edouard Grave 和 Neil Zeghidour。《Moshi：用于实时对话的语音文本基础模型》。技术报告，2024。网址：https://arxiv.org/abs/2410.00037。Ding Ding、Zeqian Ju、Yichong Leng、Songxiang Liu、Tong Liu、Zeyu Shang、Kai Shen、Wei Song、Xu Tan、Heyi Tang 等。《Kimi-Audio技术报告》。*arXiv*预印本 *arXiv:2504.18425*，2025。Ensheng Dong、Hongru Du 和 Lauren Gardner。《用于实时追踪新冠肺炎的交互式网页仪表板》。《柳叶刀·传染病》，20(5)：533–534，2020。Konstantinos Drossos、Samuel Lipping 和 Tuomas Virtanen。《Clotho：音频描述数据集》。载于*ICASSP 2020-2020 IEEE*声学、语音与信号处理国际会议（*ICASSP*），第736–740页。IEEE，2020。

杜志豪、高长风、王宇轩、于帆、赵天宇、王浩、吕翔、王辉、倪崇嘉、施先等。《Cosyvoice 3：通过扩展和后训练实现真实场景下的语音生成》。*arXiv*预印本 *arXiv:2505.17589*，2025。

Haodong Duan、Junming Yang、Yuxuan Qiao、Xinyu Fang、陈林、刘源、董晓逸、臧宇航、张攀、王佳琪等。VLMEvalKit：用于评估大型多模态模型的开源工具包。载于 第*32*届 *ACM* 多媒体国际会议论文集，第11198–11201页，2024年。Kate Dupuis 和 M Kathleen Pichora-Fuller。多伦多情感语音集（TESS）——年轻说话人_快乐。2010年。Alexandre Défossez、Jade Copet、Gabriel Synnaeve 和 Yossi Adi。高保真神经音频压缩，2022年。Jesse Engel、Cinjon Resnick、Adam Roberts、Sander Dieleman、Mohammad Norouzi、Douglas Eck 和 Karen Simonyan。使用 WaveNet 自编码器合成音符。载于 国际机器学习会议论文集，第1068–1077页。PMLR，2017年。方庆凯、郭守涛、周岩、马正瑞、张少磊和冯洋。Llama-omni：借助大语言模型实现无缝语音交互，2025年。网址：https://arxiv.org/abs/2409.06666。Daniel Galvez、Greg Diamos、Juan Ciro、Juan Felipe Cerón、Keith Achorn、Anjali Gopi、David Kanter、Maximilian Lam、Mark Mazumder 和 Vijay Janapa Reddi。The People's Speech：面向商业用途的大规模多样化英语语音识别数据集。*arXiv* 预印本 *arXiv:2111.09344*，2021年。龚源、于晋和 James Glass。VocalSound：用于改进人类发声识别的数据集。载于*ICASSP 2022 - 2022*年 *IEEE*声学、语音与信号处理国际会议（*ICASSP*），第151–155页，2022年。doi：10.1109/ICASSP43922.2022.9746828。Anmol Gulati、James Qin、Chung-Cheng Chiu、Niki Parmar、Yu Zhang、余家辉、韩伟、王世博、张正东、吴永辉等。Conformer：用于语音识别的卷积增强型 Transformer。*arXiv* 预印本 *arXiv:2005.08100*，2020年。何超群、罗仁杰、胡盛鼎、赵元倩、Jie Zhou、吴航浩、张佳杰、韩旭、Zhiyuan Liu 和孙茂松。UltraEval：用于灵活、全面评估大语言模型的轻量级平台，2024年。Andrew Hines、Jan Skoglund、Anil C Kokaram 和 Naomi Harte。ViSQOL：客观语音质量模型。*EURASIP* 音频、语音与音乐处理期刊，2015年(1)：13，2015年。黄艾琳、吴博勇、Bruce Wang、闫超、胡晨等。Step-Audio：智能语音交互中的统一理解与生成，2025年。网址：https://arxiv.org/abs/2502.11946。纪胜鹏、Ziyue Jiang、王文、陈一夫、方明辉、左佳龙、Qian Yang、程熙泽、王泽涵、李瑞琪等。Wavtokenizer：用于音频语言建模的高效声学离散编解码器分词器。*arXiv* 预印本 *arXiv:2408.16532*，2024年。Mandar Joshi、Eunsol Choi、Daniel S Weld 和 Luke Zettlemoyer。TriviaQA：用于阅读理解的大规模远程监督挑战数据集。*arXiv* 预印本 *arXiv:1705.03551*，2017年。Chris Dongjoo Kim、Byeongchang Kim、Hyunmin Lee 和 Gunhee Kim。AudioCaps：为真实场景中的音频生成描述。载于计算语言学协会北美分会*2019*年会议论文集：人类语言技术，第*1*卷（长文与短文），第119–132页，2019年。Rithesh Kumar、Prem Seetharaman、Alejandro Luebs、Ishaan Kumar 和 Kundan Kumar。采用改进 RVQGAN 的高保真音频压缩。神经信息处理系统进展，36：27980–27993，2023年。Tony Lee、Haoqin Tu、Chi Heem Wong、王子骏、杨思伟、麦一帆、周雨茵、谢赐航和Percy Liang。AHELM：对音频语言模型的整体评估。*arXiv* 预印本 *arXiv:2508.21376*，2025年。Haonan Li、张一轩、Fajri Koto、杨逸飞、赵海、龚叶芸、段楠和 Timothy Baldwin。CMMLU：衡量中文大规模多任务语言理解能力，2023a年。

Xuechen Li、Tianyi Zhang、Yann Dubois、Rohan Taori、Ishaan Gulrajani、Carlos Guestrin、Percy Liang 和 Tatsunori B. Hashimoto。AlpacaEval：指令遵循模型的自动评估器。https://github.com/tatsu-lab/alpaca_eval，5，2023b年。Xinhao Mei、Chutong Meng、Haohe Liu、Qiuqiang Kong、Tom Ko、Chengqi Zhao、Mark D Plumbley、Yuexian Zou 和 Wenwu Wang。WavCaps：由 ChatGPT 辅助生成的弱标注音频描述数据集，面向音频语言多模态研究。*IEEE/ACM 音频、语音与语言处理汇刊*，2024年。Mathieu Moreaux。Audio cats a nd dogs。https://www.kaggle.com/datasets/mmoreaux/audio-cats-and-dogs，2023年。访问日期：2025年5月13日。Eliya Nachmani、Alon Levkovitch、Roy Hirsch、Julian Salazar、Chulayuth Asawaroengchai、Soroosh Mari-ooryad、Ehud Rivlin、RJ Skerry-Ryan 和 Michelle Tadmor Ramanovich。使用频谱图驱动的大语言模型进行口语问答和语音续接。*ar Xiv 预印本 arXiv:2305.15255*，2023年。Arsha Nagrani、Joon Son Chung 和 Andrew Zisserman。VoxCeleb：大规模说话人识别数据集。*arXiv 预印本 arXiv:1706.08612*，2017年。OpenAI、：Aaron Hurst、Adam Lerer、Adam P. Goucher、Adam Perelman等。GPT-4o系统卡，2024年。网址：https://arxiv.org/abs/2410.21276。Vassil Panayotov、Guoguo Chen、Daniel Povey 和 Sanjeev Khudanpur。LibriSpeech：基于公有领域有声读物的自动语音识别语料库。载于*2015年IEEE国际声学、语音与信号处理会议（ICASSP）*，第5206–5210页。IEEE，2015年。Soujanya Poria、Devamanyu Hazarika、Navonil Majumder、Gautam Naik、Erik Cambria 和 Rada Mihalcea。MELD：用于对话情感识别的多模态多方对话数据集。*arXiv 预印本 arXiv:1810.02508*，2018年。Vineel Pratap、Qiantong Xu、Anuroop Sriram、Gabriel Synnaeve 和 Ronan Collobert。MLS：用于语音研究的大规模多语言数据集。*arXiv*，abs/2012.03411，2020年。Chandan KA Reddy、Vishak Gopal 和 Ross Cutler。DNSMOS：用于评估噪声抑制器的非侵入式感知客观语音质量指标。载于*ICASSP 2021-2021年IEEE声学、语音与信号处理国际会议（ICASSP）*，第6493–6497页。IEEE，2021年。Chandan KA Reddy、Vishak Gopal 和 Ross Cutler。DNSMOS P.835：用于评估噪声抑制器的非侵入式感知客观语音质量指标。载于*ICASSP 2022-2022年IEEE国际声学、语音与信号处理会议（ICASSP）*，第886–890页。IEEE，2022年。BM Rocha、Dimitris Filos、Lea Mendes、Ioannis Vogiatzis、Eleni Perantoni、Evangelos Kaimakamis、P Natsiavas、Ana Oliveira、C Jácome、A Marques等。用于开发自动分类系统的呼吸音数据库。载于*由个性化健康与互联健康驱动的精准医疗：ICBHI 2017，希腊塞萨洛尼基，2017年11月18日至21日*，第33–37页。施普林格，2018年。Anthony Rousseau、Paul Deléglise 和 Yannick Esteve。TED-LIUM：面向自动语音识别的专用语料库。载于*LREC*，第125–129页，2012年。Takaaki Saeki、Detai Xin、Wataru Nakata、Tomoki Koriyama、Shinnosuke Takamichi 和 Hiroshi Saruwatari。UTMOS：东京大学Sarulab系统，适用于VoiceMOS挑战赛2022，2022年。网址：https://arxiv.org/abs/2204.02152。S Sakshi、Utkarsh Tyagi、Sonal Kumar、Ashish Seth、Ramaneswaran Selvakumar、Oriol Nieto、Ramani Duraiswami、Sreyan Ghosh 和 Dinesh Manocha。MMAU：大规模多任务音频理解与推理基准。*arXiv 预印本 arXiv:2410.19168*，2024年。B系列。音频系统中等质量级别的主观评估方法。国际电信联盟无线电通信全会，2，2014年。Sripaa D. Srinivasan。Audio MNIST。https://www.kaggle.com/datasets/sripaadsrinivasan/audio-mnist，2023年。访问日期：2025年5月13日。Bob L Sturm。GTZAN数据集：其内容、缺陷、缺陷对评估的影响及未来用途。*arXiv 预印本 arXiv:1306.1461*，2013年。

Sidharth Surapaneni、Hoang Nguyen、Jash Mehta、Aman Tiwari、Oluwanifemi Bamgbose、Akshay Kalkunte、Sai Rajeswar 和 Sathwik Tejaswi Madhusudhan。Au-harness：用于全面评估音频大语言模型的开源工具包。*arXiv* 预印本 *arXiv:2509.08031*，2025。

Cees H Taal、Richard C Hendriks、Richard Heusdens 和 Jesper Jensen。用于预测经过时频加权的噪声语音可懂度的算法。*IEEE* 音频、语音与语言处理汇刊，19(7):2125–2136，2011年。

Zhiyuan Tang、Dong Wang、Yanguang Xu、Jianwei Sun、Xiaoning Lei、Shuaijiang Zhao、Cheng Wen、Xingjun Tan、Chuandong Xie、Shuran Zhou 等。Kespeech：一个涵盖普通话及其八种次方言的开源语音数据集。发表于第三十五届神经信息处理系统大会数据集与基准测试专题（第二轮），2021年。

Gemini 团队、Petko Georgiev、Ving Ian Lei 等人。Gemini 1.5：解锁数百万词元上下文中的多模态理解，2024年。网址https://arxiv.org/abs/2403.05530。

Nicolas Turpault、Romain Serizel、Ankit Parag Shah 和 Justin Salamon。利用弱标注数据和声景合成，在家庭环境中进行声音事件检测。发表于声学场景与事件检测和分类研讨会，2019年。

Changhan Wang、Anne Wu 和 Juan Pino。CoVoST 2：大规模多语言语音转文本翻译语料库，2020年。

Changhan Wang、Morgane Riviere、Ann Lee、Anne Wu、Chaitanya Talnikar、Daniel Haziza、Mary Williamson、Juan Pino 和 Emmanuel Dupoux。VoxPopuli：用于表征学习、半监督学习和解释的大规模多语言语音语料库。*arXiv* 预印本 *arXiv:2101.00390*，2021年。

Chengyao Wang、Zhisheng Zhong、Bohao Peng、Senqiao Yang、Yuqi Liu、Haokun Gui、Bin Xia、Jingyao Li、Bei Yu 和 Jiaya Jia。MGM-Omni：将全模态大语言模型扩展至个性化长时程语音。*arXiv* 预印本 *arXiv:2509.25131*，2025a。

Xinsheng Wang、Mingqi Jiang、Ziyang Ma、Ziyu Zhang、Songxiang Liu、Linqin Li、Zheng Liang、Qixi Zheng、Rui Wang、Xiaoqin Feng、Weizhen Bian、Zhen Ye、Sitong Cheng、Ruibin Yuan、Zhixian Zhao、Xinfa Zhu、Jiahao Pan、Liumeng Xue、Pengcheng Zhu、Yunlin Chen、Zhifei Li、Xie Chen、Lei Xie、Yike Guo 和 Wei Xue。Spark-TTS：一种采用单流解耦语音词元的高效基于大语言模型的文本转语音模型，2025年b。网址https://arxiv.org/abs/2503.01710。

Detai Xin、Xu Tan、Shinnosuke Takamichi 和 HiroshiSaruwatari。BigCodec：突破低码率神经语音编解码器的极限。*arXiv* 预印本 *arXiv:2409.05377*，2024年。

Jin Xu、Zhifang Guo、Jinzheng He、Hangrui Hu、Ting He、Shuai Bai、Keqin Chen、Jialin Wang、Yang Fan、Kai Dang 等。Qwen2. 5-omni 技术报告。*arXiv* 预印本 *arXiv:2503.20215*，2025。

Dongchao Yang、Songxiang Liu、Rongjie Huang、Jinchuan Tian、Chao Weng 和 Yuexian Zou。HiFi-Codec：用于高保真音频编解码器的组残差向量量化。*arXiv* 预印本 *arXiv:2305.02765*，2023年。

Qian Yang、Jin Xu、Wenrui Liu、Yunfei Chu、Ziyue Jiang、Xiaohuan Zhou、Yichong Leng、Yuanjun Lv、Zhou Zhao、Chang Zhou 等。AIR-Bench：通过生成式理解对大型音频语言模型进行基准测试。*arXiv* 预印本 *arXiv:2402.07729*，2024年。

Yuan Yao、Tianyu Yu、Ao Zhang、Chongyi Wang、Junbo Cui、Hongji Zhu、Tianchi Cai、Haoyu Li、Weilin Zhao、Zhihui He 等。Minicpm-v：手机上的 GPT-4V 级多模态大语言模型。*arXiv* 预印本 *arXiv:2408.01800*，2024年。

Zhen Ye、Peiwen Sun、Jiahe Lei、Hongzhan Lin、Xu Tan、Zheqi Dai、Qiuqiang Kong、Jianyi Chen、Jiahao Pan、Qifeng Liu 等。编解码器至关重要：探索编解码器对音频语言模型的语义缺陷。发表于人工智能促进协会年会论文集，第39卷，第25697–25705页，2025。

Neil Zeghidour、Alejandro Luebs、Ahmed Omran、Jan Skoglund 和 Marco Tagliasacchi。Soundstream：一种端到端神经音频编解码器。*IEEE/ACM* 音频、语音与语言处理汇刊，30：495–507，2021年。Aohan Zeng、Zhengxiao Du、Mingdao Liu、Kedong Wang、Shengmin Jiang、Lei Zhao、Yuxiao Dong 和 Jie Tang。Glm-4-voice：迈向智能、类人化的端到端语音聊天机器人，2024年。网址：https://arxiv.org/abs/2412.02612。

Binbin Zhang、Hang Lv、Pengcheng Guo、Qijie Shao、Chao Yang、Lei Xie、Xin Xu、Hui Bu、Xiaoyu Chen、Chenchen Zeng 等。Wenetspeech：一个用于语音识别的 10000+ 小时多领域普通话语料库。载于*ICASSP 2022-2022 IEEE* 国际声学、语音与信号处理会议（*ICASSP*），第 6182–6186 页。IEEE，2022。

Lianmin Zheng、Wei-Lin Chiang、Ying Sheng、Siyuan Zhuang、Zhanghao Wu、Yonghao Zhuang、Zi Lin、Zhuohan Li、Dacheng Li、Eric Xing 等。《使用 MT-Bench 和 Chatbot Arena 评判大语言模型裁判》。神经信息处理系统进展，36:46595–46623，2023年。

Jeffrey Zhou、Tianjian Lu、Swaroop Mishra、Siddhartha Brahma、Sujoy Basu、Yi Luan、Denny Zhou 和 Le Hou。《大语言模型的指令遵循评估》，2023年。网址https://arxiv.org/abs/2311.07911。