UltraEval-Audio: A Unified Framework for Comprehensive Evaluation of Audio Foundation Models

Qundong Shi[1], **Jie Zhou**[1], **Biyuan Lin**[1], **Junbo Cui**[1], **Guoyang Zeng**[1], **Yixuan Zhou**[1], **Ziyang Wang**[1], **Xin Liu**[1], **Zhen Luo**[1], **Yudong Wang**[2], **Zhiyuan Liu**[2]
[1]ModelBest Inc. [2]Tsinghua University
{shiqundong, zhoujie, linbiyuan}@modelbest.cn
liuzy@tsinghua.edu.cn

https://github.com/OpenBMB/UltraEval-Audio

Abstract

The development of audio foundation models has accelerated rapidly since the emergence of GPT-4o, yet their comprehensive evaluation remains largely unexplored, particularly regarding audio generation. Traditional audio evaluation tasks lack a unified framework, with datasets and code scattered across various sources, and emerging user-centric speech benchmarks (e.g., AIR-Bench, Speech AlpacaEval) remain almost entirely English-centric. Moreover, audio codecs, which convert audio into tokens and reconstruct audio from tokens, are also essential to overall audio foundation model performance, yet they lack holistic evaluation. To address these gaps, we introduce **UltraEval-Audio**, a unified framework for evaluating audio foundation models in both audio understanding and audio generation. The framework offers a simple, user-friendly interface with one-command evaluation and provides public leaderboards for transparent comparison. It integrates 24 models and 36 authoritative benchmarks covering three domains: speech, environmental sound, and music, and supports 10 languages and 12 distinct task categories. To enrich Chinese speech evaluation, we additionally release two new benchmarks, SpeechCMMLU and SpeechHSK, which assess Chinese knowledge and language proficiency. We also propose a holistic evaluation approach for audio codecs based on semantics, timbre fidelity, and acoustic quality. Together, these contributions position UltraEval-Audio as a unified and comprehensive framework for evaluating audio foundation models. Our code, benchmarks, and leaderboards are available at https://github.com/OpenBMB/UltraEval-Audio.

1 Introduction

Following the groundbreaking success of large language models (LLMs), the field has witnessed the rapid emergence of multimodal large language models. Among these advances, OpenAI's GPT-4o (OpenAI et al., 2024) has significantly accelerated progress in the audio modality. Following its release, a wave of audio foundation models has emerged, including Qwen-Audio (Chu et al., 2023, 2024), Moshi (Défossez et al., 2024), GLM-4-Voice (Zeng et al., 2024), Step-Audio (Huang et al., 2025), Qwen2.5-Omni (Xu et al., 2025), Kimi-Audio (Ding et al., 2025), Mini-Omni (Xie & Wu, 2024), MiniCPM-o 2.6 (Yao et al., 2024) and Llama-Omni (Fang et al., 2025).

In contrast to the explosive growth of models in the audio domain, evaluation in this domain has largely remained stagnant. Existing evaluation frameworks still focus on specific audio tasks (e.g., Automatic Speech Recognition (ASR) and Automatic Speech Translation (AST)) for task-specific audio models rather than general audio foundation models. More importantly, these frameworks lack support for prompts and the adjustment of inference arguments, both of which are essential for evaluating audio foundation models. Moreover, while traditional audio tasks such as ASR and AST remain important, user-centric speech benchmarks have started to emerge, yet they face several challenges. First, evaluations in some speech benchmarks are inconsistent, ambiguous, and lack transparency. For example, Spectron (Nachmani et al., 2023) introduced Speech WebQuestions and reported the performance without releasing the dataset, while Moshi's Speech TriviaQA also remains unavailable. Second, the speech research community faces a shortage of multilingual benchmarks, resulting in evaluations that are predominantly English-centric. Existing benchmarks, including

AIR-Bench, VoiceBench and the aforementioned benchmarks, are all in English. There is an urgent need for an open and unified evaluation framework and more multilingual speech benchmarks to accelerate the advancement of audio foundation models.

At the architectural level, audio foundation models differ from general LLMs by an additional **audio codec** module, comprising an **audio tokenizer** that converts audio into discrete tokens and a **vocoder** that reconstructs audio from generated tokens. This codec design directly determines the fidelity and efficiency of audio representations, which profoundly affects the audio foundation model's overall performance (Ye et al., 2025). While the codec is a critical component to evaluate, existing methods are broad and provide limited insight into specific performance dimensions.

To address these challenges, we introduce **UltraEval-Audio**, a comprehensive evaluation framework for audio foundation models. The framework addresses limitations in existing evaluation by supporting datasets, prompt management, models, post-processing, evaluation methods, aggregation methods, and diverse input-output modalities, while providing a standardized and transparent platform for benchmarking. Our contributions are summarized as follows:

- UltraEval-Audio is the first unified and comprehensive framework for evaluating audio foundation models. It supports a wide range of input–output modalities, including *text* $\rightarrow$ *audio*, *text* + *audio* $\rightarrow$ *text*, *audio* $\rightarrow$ *text*, and *text* + *audio* $\rightarrow$ *audio*. The framework integrates 24 models and 36 authoritative benchmarks across three key domains: speech, environmental sound, and music. It supports 10 languages and 12 distinct task categories. The framework is highly user-friendly, offering one-command evaluation and public leaderboards for transparent comparison.
- We introduce two Chinese speech benchmarks: **SpeechCMMLU** and **SpeechHSK**, enabling the assessment of Chinese knowledge and language proficiency. In addition, we present a systematic pipeline for converting text-based benchmarks into audio datasets through speech synthesis and quality validation.
- We propose an evaluation methodology for audio codecs based on three key dimensions: semantics, timbre fidelity and acoustic quality.

2 Related Works

The development of evaluation frameworks is critical to the advancement of LLMs. These frameworks not only delineate the capabilities, limitations, and potential risks of LLMs but also offer essential guidance for developers in practical application.

The Advancements of Audio Foundation Models. In the early stage, typical audio foundation models adopted a simple **ASR+LLM+TTS** pipeline, which inevitably resulted in the loss of important acoustic information. Recently, a standardized paradigm for audio foundation models has begun to emerge, which can be summarized as an **Audio Codec+LLM** architecture. It generally comprises three core components: (1) an **audio tokenizer**, which converts raw audio signals into discrete tokens while preserving both semantic and acoustic information; (2) an **LLM backbone**, responsible for autoregressive token prediction and contextual modeling; and (3) a **vocoder**, which synthesizes natural speech waveforms from the generated audio tokens. However, not all audio foundation models implement all the components described above. Based on whether they incorporate a vocoder, audio foundation models can be classified into two categories: (1) audio understanding foundation models, which accept both audio and text as input but produce only text as output (e.g., Qwen-Audio, Gemini-1.5). (2) audio generation foundation models, which accept both audio and text as input and generate both speech and text as output (e.g., GPT-4o-Realtime, Moshi, MiniCPM-o 2.6, Qwen2.5-Omni, Kimi-Audio).

Meanwhile, audio codecs are also rapidly evolving. SoundStream is the first universal audio codec capable of handling diverse audio types. EnCodec, DAC (descript-audio-codec), HiFi-Codec, X-codec (Ye et al., 2025), BigCodec, and BiCodec further improve reconstruction quality, codebook efficiency, and compatibility with LLM-based speech generation, reflecting a clear trend toward scalable, low-latency, and generative audio tokenizers.

Multimodal LLM Evaluation Framework. The development of LLMs has been closely linked to the evolution of evaluation frameworks. For text-based LLMs, several evaluation frameworks exist: HELM (Liang et al., 2022), FlagEval [1], OpenCompass (Contributors, 2023), OpenAI Evals [2], and UltraEval (He et al., 2024). In the domain of visual LLMs, frameworks such as LVLMeHub (Xu et al., 2024), VLMEvalKit (Duan et al.,

[1] https://flageval.baai.ac.cn
[2] https://github.com/openai/evals

2024), and HEIM (Lee et al., 2023) have been proposed. However, despite the rising popularity of audio foundation models and the growing number of released models, a comprehensive evaluation of these models has been lacking.

The Development of Audio Evaluation. Before the emergence of audio foundation models, research in the audio domain primarily focused on tasks such as ASR, AST, and TTS, as well as other tasks like emotion recognition and sound classification. Each model was designed for a specific task, and its corresponding evaluation was typically ad hoc, often provided alongside the model repository. For example, ASR evaluations were commonly conducted using scripts from Whisper or ESPnet-SE.

LLMs are task-agnostic learners (Brown et al., 2020), which has in turn driven the development of comprehensive evaluation frameworks (Liang et al., 2022). Similarly, audio foundation models integrate a wide range of traditional audio tasks, including ASR, AST, TTS, emotion recognition, and others, underscoring the need for a unified and comprehensive evaluation framework in this domain. Several audio evaluation frameworks have been proposed to address this. For instance, **AudioBench** (Chen et al., 2024) collects 8 distinct tasks and 26 benchmarks for evaluating audio foundation models, but it lacks coverage of audio generation tasks. **Kimi-Audio-Evalkit** (Ding et al., 2025) integrates all benchmarks mentioned in Kimi-Audio evaluation for reproduction. However, its evaluation process requires five steps, making it cumbersome to use. Additionally, modifying prompts is inconvenient, as changes must be made directly in the code rather than through configuration files. **AU-Harness** (Surapaneni et al., 2025) offers an efficient evaluation engine supporting over 380 tasks, but it requires users to manually adapt open-source audio foundation models into standardized vLLM services.

Meanwhile, unlike traditional ASR and AST benchmarks, the field has begun developing user-centric benchmarks that use raw speech as input without additional task description and directly evaluate model responses. For audio understanding, AIR-Bench (Yang et al., 2024) collects spoken question answering (QA) samples from existing datasets and employs GPT-4 as an automatic evaluator. VoiceBench (Chen et al., 2024) further expands this direction by including both naturally spoken QA samples and synthetic spoken instructions, which are generated from text-based instruction-following datasets (e.g. AlpacaEval (Li et al., 2023b), IFEval (Zhou et al., 2023)) using Google TTS. AHELM (Lee et al., 2025) aggregates various datasets to holistically measure the performance of ALMs across 10 aspects: *audio perception, knowledge, reasoning, emotion detection, bias, fairness, multilinguality, robustness, toxicity, and safety*. For audio generation, the first dedicated speech question-answering benchmark, Llama-Question (Nachmani et al., 2023), introduced a synthetic speech QA dataset with a novel evaluation paradigm: it employs ConformerASR (Gulati et al., 2020) to transcribe reply audio into text before assessing answer accuracy. Speech WebQuestions (Nachmani et al., 2023) is derived from WebQuestions (Chen et al., 2015), while Speech TriviaQA (Défossez et al., 2024) similarly synthesizes audio from the TriviaQA (Joshi et al., 2017) dataset. Speech AlpacaEval (Fang et al., 2025) selects suitable data for speech interaction scenarios from AlpacaEval. However, all these benchmarks are currently limited to English, leaving multilingual speech benchmarks largely unexplored.

The Development of Audio Codec Evaluation. The evaluation of audio codecs employs both subjective and objective metrics. Subjective evaluation typically follows the MUSHRA (Series, 2014) protocol, which uses both a hidden reference and a low anchor. Objective evaluation includes several approaches: ViSQOL (Hines et al., 2015; Chinen et al., 2020) measures spectral similarity to the ground truth as a proxy for mean opinion score; Scale-Invariant Signal-to-Noise Ratio (SI-SNR) quantifies the similarity between reconstructed and original audio while ignoring signal scale; Mel distance computes the difference between the log-Mel spectrograms of reconstructed and ground truth waveforms; STOI (Taal et al., 2011) assesses speech intelligibility; and speaker similarity (SIM) is calculated as the cosine similarity between speaker vectors of the reconstructed audio and ground truth using an embedding model. Beyond these direct metrics, recent works like that of (Ye et al., 2025) also employ downstream tasks such as TTS, to indirectly evaluate codec performance.

3 UltraEval-Audio

This section delineates the scope and operational architecture of our evaluation framework, details the construction of two Chinese speech benchmarks, and describes the methodology for audio codec evaluation. As illustrated in Figure 1, the complete evaluation workflow mainly consists of three components: data, model, and assessment. Section 3.1 presents the data organization process. Section 3.2 introduces our innovations in the inference process of audio foundation models. Section 3.3 describes the assessment methodology. Section 3.4 provides details of the **SpeechCMMLU** and **SpeechHSK** benchmarks. Finally, Section 3.5 presents our holistic evaluation methods for audio codecs.

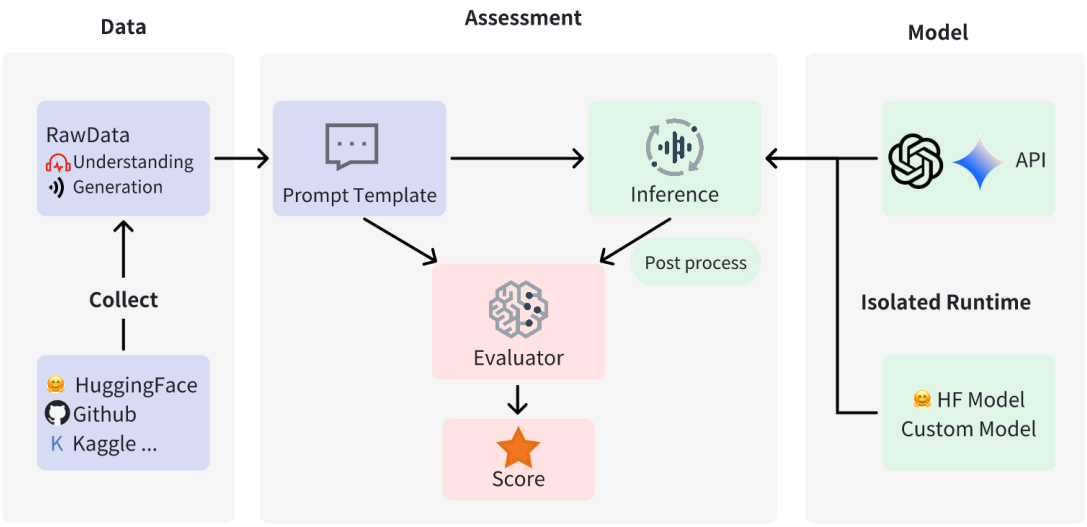

Figure 1: How UltraEval-Audio works

3.1 Data

The data pipeline comprises the entire process from raw data acquisition and preprocessing to formatting the data into prompts suitable for the model.

Raw Data. To align with the output modalities of audio foundation models, we categorize the benchmarks into two main types:

- Audio Understanding: Benchmarks designed for text-only output.
- Audio Generation: Benchmarks designed for audio output.

Across these two categories, our study incorporates 36 authoritative benchmarks, as detailed in Table 1. These benchmarks are carefully selected to comprehensively cover three key domains: **Speech** (e.g., ASR, AST, speech QA), **Environmental Sound** (e.g., animal sounds such as dogs and cats, and scene audio recognition), **Music** (e.g., instrument recognition, genre classification). These benchmarks cover ten languages, including both high-resource languages and low-resource languages (e.g., fleurs-hi (Conneau et al., 2022), fleurs-de, fleurs-ja, fleurs-ru), and are designed to evaluate model performance across diverse linguistic contexts. In total, they encompass twelve distinct tasks, evenly distributed between audio understanding and audio generation benchmarks, enabling a comprehensive assessment of audio foundation model capabilities along both semantic and acoustic dimensions.

Prompt Templates. Prompts play a crucial role in guiding models to generate task-specific outputs, directly affecting model performance. In practice, LLM developers typically provide official prompts to reproduce model performance. However, the diversity of these official prompts across different LLMs, combined with significant variations in prompt design across tasks, poses significant challenges to reproducibility. To address this issue, we provide customized, user-friendly, and readable prompt templates managed through YAML files. For example, a Qwen-Audio ASR prompt template is defined as:

```
qwen2-audio-pre-train-asr-zh:
  class: audio_evals.prompt.base.Prompt
  args:
    template: '<|audio_bos|><|{{audio}}|><|audio_eos|>Detect the language and recognize the
    speech: <|zh|>'
```

Table 1: Benchmarks supported in UltraEval-Audio. * indicates new benchmarks introduced in this paper.

Domain	Task	Dataset	Language
Speech	ASR	TED-LIUM (Rousseau et al., 2012)	en
	ASR	VoxPopuli (Wang et al., 2021)	en
	ASR	KeSpeech (Tang et al., 2021)	zh
	ASR	LibriSpeech (Panayotov et al., 2015)	en
	ASR	MLS (Pratap et al., 2020)	nl, fr, de, it, pl, pt, es
	ASR	FLEURS (Conneau et al., 2022)	zh, en, ru, de, jp, ...
	ASR	The People's Speech (Galvez et al., 2021)	en
	ASR	WenetSpeech (Zhang et al., 2022)	en
	ASR	GigaSpeech (Chen et al., 2021)	en
	ASR	AISHELL-1 (Bu et al., 2017)	zh
	ASR	Common Voice (Ardila et al., 2019)	zh, en, ru, de, jp, ...
	Speech QA	Speech TriviaQA (Défossez et al., 2024)	en
	Speech QA	Speech WebQuestions (Nachmani et al., 2023)	en
	Speech QA	Speech AlpacaEval (Fang et al., 2025)	en
	Speech QA	LLaMA-Questions (Nachmani et al., 2023)	en
	Speech Choice QA	SpeechHSK*	zh
	Speech Choice QA	SpeechCMMLU*	zh
	Audio Classification	AudioMNIST (Srinivasan, 2023)	en
	Emotion Analysis	TESS (Dupuis & Pichora-Fuller, 2010)	en
	Emotion Analysis	MELD (Poria et al., 2018)	en
	Gender Analysis	VoxCeleb (Nagrani et al., 2017)	en
	Single Choice	AIR-Bench (Yang et al., 2024)	en
	AST	CoVoST 2 (Wang et al., 2020)	zh,en,ru,de,jp, ...
Music	Chord Recognition	Chord (deepcontractor, 2023)	-
	Instrument Recognition	NSynth (Engel et al., 2017)	-
	Music Genre	GTZAN (Sturm, 2013)	-
Environmental Sound	Caption	AudioCaps (Kim et al., 2019)	-
	Caption	WavCaps (Mei et al., 2024)	-
	Cat/Dog Identify	CatDog (Moreaux, 2023)	-
	Caption	Clotho (Drossos et al., 2020)	-
	Audio Classification	DESED (Turpault et al., 2019)	-
	Single Choice With Answer	MMAU (Sakshi et al., 2024)	-
	Vocalsound Analysis	VocalSound (Gong et al., 2022)	-
Medical Domain	COVID Recognizer	COVID-19 Sounds (Dong et al., 2020)	-
	Heartbeat-Recognizer	PASCAL CHSC 2011 (Bentley et al.)	-
	Crackles-Recognizer	ICBHI 2017 Respiratory Sound (Rocha et al., 2018)	-

and a MiniCPM-o 2.6 ASR prompt template is defined as:

```
mini-cpm-omni-asr-en:
  class: audio_evals.prompt.base.Prompt
  args:
    template:
    - role: user
      contents:
      - type: text
       value: 'Please listen to the audio snippet carefully and transcribe the content. Please
         output in low case.'
      - type: audio
        value: '{{audio}}'
```

Our framework supports prompt templates in any format, applicable to both base models and chat models.

3.2 Model

UltraEval-Audio integrates 15 audio foundation models and 9 audio codec models. It provides a unified *.inference()* interface that accepts a prompt and returns a string response for all models. For speech-to-speech modality, the response is a JSON string containing references to the generated audio files.

In practice, there are two types of inference models: 1) API-based models, which perform inference via official clients or direct API calls, and 2) open-source models, which execute the entire inference pipeline locally. Open-source models often introduce significant dependency conflicts due to the diverse requirements of different models. For example, acoustic evaluators such as UTMOS (Saeki et al., 2022), SIMO (Yao et al., 2021), DNSMOS (Reddy et al., 2022) and the ASR-WER evaluator by Whisper-large-v3 (Radford et al., 2023) or Paraformer-zh (Gao et al., 2023), typically rely on specific environments. These dependencies frequently conflict with those required by the audio models under evaluation. To address this issue, **UltraEval-Audio** introduces an **Isolated Runtime** mechanism:

1. For each model, an isolated virtual environment is created, containing only its required dependencies and fully separated from the evaluation process.
2. The model is launched as a subprocess within its dedicated environment, enabling continuous inference as a service.
3. The evaluation process obtains inference results from the model subprocess through inter-process communication (IPC) mechanisms, such as system pipes.

This design completely eliminates cross-model dependency conflicts while clearly encapsulating each model's environment. Consequently, users no longer need to manage or install conflicting dependencies manually.

3.3 Assessment

LLM predictions often contain extraneous information, requiring post-processing before they can be effectively evaluated.

Post-Processing. Similar to text-based LLM evaluation frameworks, **UltraEval-Audio** also incorporates Choice Extraction, Yes/No Extraction, JSON Extraction, and multi-step post-processing workflows. Specifically, when evaluating the semantic capabilities of audio foundation model for audio generation, the generated audio must first be transcribed into text via ASR-based post-processing.

Evaluators that receive model outputs and produce scores fall into two categories: rule-based and model-based. UltraEval-Audio integrates various evaluators across both types:

- **Rule-based evaluators:** WER (Word Error Rate) for ASR tasks, BLEU for AST, accuracy (ACC) for audio classification, and COCO (Chen et al., 2015) metrics for audio captioning.
- **Model-based evaluators:** SIMO (Yao et al., 2021) for voice cloning, ASR-WER for TTS, UTMOS and DNSMOS for assessing speech naturalness, and GPT-based evaluation (G-Eval) for tasks like AlpacaEval.

3.4 Chinese Speech Benchmarks

With the advent of audio foundation models, a new generation of raw speech-based benchmarks has emerged, marking a significant paradigm shift in the field. This evolution moves beyond the exclusive reliance on traditional proprietary metrics, such as ASR and AST, extending evaluation to more comprehensive and user-centric scenarios. Nevertheless, the availability of speech benchmarks remains limited, particularly for Chinese, where a significant gap persists. To help bridge this gap, we introduce two benchmarks: **SpeechCMMLU** and **SpeechHSK (Chinese Proficiency Test)**.

SpeechCMMLU adapts the CMMLU (Li et al., 2023a) benchmark into the speech domain through a systematic synthesis pipeline. The following sections detail the data generation and quality assurance procedures.

1. **Text Instruction:** CMMLU is a multiple-choice benchmark. For the speech setting, each item is formatted by concatenating the question, answer options, and a meta instruction using the following template: *"There is a single-choice question about **. Answer the question by replying A, B, C, or D. Question: ** A. ** B. ** C. ** D. ** Answer:"*
2. **Speech Instructions:** Given the high cost of manual speech recording, especially when dealing with a substantial dataset consisting of 11,583 samples, we employed the TTS model (CosyVoice2 (Du et al., 2024)) to synthesize all spoken instructions.

3. **Quality Control:** Some question texts contain rare words or special symbols (e.g., [HNO2] > [H2NO2+]) that are challenging to pronounce. Moreover, the TTS model cannot guarantee correct pronunciation for every term. To ensure quality, we employ the ASR model Paraformer-zh to filter samples, retaining only those with a perfect transcription (zero WER).

Following quality control, we release the final **SpeechCMMLU** dataset, consisting of 3,519 samples.

SpeechHSK. HSK [3] is an official and authoritative language assessment system established by the Chinese government. As the most widely recognized standardized test for non-native speakers, it serves as a global benchmark for evaluating Chinese language proficiency. SpeechHSK is constructed by systematically collecting:

1. The original multiple-choice question audio from the listening sections of HSK exams.
2. The answer options, which are originally presented as text in the HSK exam, professionally recorded as audio by native Chinese speakers.

The benchmark is organized into six proficiency levels (SpeechHSK 1–6), with ascending level numbers corresponding to increased linguistic complexity and difficulty.

3.5 Audio Codec Evaluation

We evaluate audio codecs based on three dimensions:

1. **Semantics.** We assess the degree to which semantic content is preserved after audio compression using the ASR-WER metric. Specifically, we use Whisper-large-v3 for English and Paraformer-zh for Chinese to transcribe reconstructed audio and compare the transcriptions against the original text.
2. **Timbre Fidelity.** This measures how well audio tokens capture speaker characteristics. We compute speaker similarity (SIM) between original and reconstructed audio using a WavLM-large model fine-tuned for speaker verification. This model extracts speaker embedding vectors, and we calculate the cosine similarity between test utterances and reference clips to evaluate timbre fidelity.
3. **Acoustic Quality.** To evaluate speech naturalness, we employ UTMOS (Saeki et al., 2022), a metric that predicts the Mean Opinion Score (MOS) using self-supervised learning. Additionally, we use both DNSMOS P.835 and DNSMOS P.808 variants to assess quality in the context of noise suppression.

4 Evaluation Results

With **UltraEval-Audio**, researchers can comprehensively evaluate audio foundation models and audio codecs across multiple benchmarks. The framework provides a unified solution for systematically assessing the performance of these audio-processing models in various testing environments. This section first introduces the evaluated models in Section 4.1. Section 4.2 then presents the **Audio Understanding Leaderboard**, Section 4.3 discusses the **Audio Generation Leaderboard**, and Section 4.4 describes the **Audio Codec Leaderboard**.

4.1 Evaluated Models

We select popular and emerging audio foundation models and audio codecs to construct the leaderboard. The audio foundation models are listed in Table 2, and the audio codecs include **Encodec** (Défossez et al., 2022), **Chattts-DVAE** [4], the **Mimi** (Défossez et al., 2024) family, **WavTokenizer-large-v2-75-tokens**, **WavTokenizer-large-40-tokens** (Ji et al., 2024) and **Spark** (Wang et al., 2025).

Importantly, each model is evaluated using its official prompts and parameter settings if available; otherwise, default settings are applied. We do not perform prompt or parameter optimization, ensuring a fair and consistent evaluation protocol.

4.2 Audio Understanding

For audio understanding, we select well-established and widely cited benchmarks that are commonly used in existing papers on audio foundation models. Specifically, we use **Librispeech** (en), **TED-LIUM** (en),

[3] https://www.chinesetest.cn/HSK
[4] https://github.com/2noise/ChatTTS

Table 2: Overview of audio foundation models participating in the evaluation

Model	Institution	Type	Modality	Languages
GPT-4o-Realtime	OpenAI	Proprietary	audio + text -> audio + text	Multilingual
MiniCPM-o 2.6	OpenBMB	Open-Source	audio + text -> audio + text	English, Chinese
Gemini-1.5-pro	Google	Proprietary	audio + text -> text	Multilingual
Gemini-1.5-flash	Google	Proprietary	audio + text -> text	Multilingual
Gemini-2.5-flash	Google	Proprietary	audio + text -> text	Multilingual
Gemini-2.5-pro	Google	Proprietary	audio + text -> text	Multilingual
Qwen2-Audio	Alibaba	Open-Source	audio + text -> text	Multilingual
Qwen2-Audio -Instruction	Alibaba	Open-Source	audio + text -> text	Multilingual
GLM-4-Voice	Zhipu	Open-Source	audio -> audio	English, Chinese
Qwen2.5-Omni	Alibaba	Open-Source	audio + text -> audio + text	English, Chinese
Qwen3-Omni-30B -A3B-Instruct	Alibaba	Open-Source	audio + text -> audio + text	Multilingual
Kimi-Audio-7B-Instruct	Moonshot	Open-Source	audio + text -> audio + text	English, Chinese
MiDaShengLM-7B	Xiaomi	Open-Source	audio + text -> text	Multilingual

Table 3: Audio Understanding Performance. WER (↓) for ASR, BLEU (↑) for AST, and ACC (↑) for EMO. Best results are in bold.

Model	ASR						AST		EMO
	Librispeech dev-clean|dev-other test-clean|test-other	**TED-LIUM**	**CV-15** en|zh	**Aishell-1**	**FLEURS-zh**	**Wenet -test-net**	**covost2-en2zh**	**covost2-zh2en**	**MELD**
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
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
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
Qwen2-Audio -Instruction	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
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
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
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**
Qwen2-Audio	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
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
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
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
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

Common Voice 15 (zh, en), **FLEURS** (zh), **WenetSpeech-test-net** (zh), **Aishell-1** (zh) for ASR, **covost2-zh2en**, **covost2-en2zh** for AST, and **MELD** for emotion recognition (EMO).

The performance is presented in Table 3, which reveals the following key observations:

1. GPT-4o-Realtime faces strong competition in the field of audio understanding, with open-source models such as Qwen3-Omni-30B-A3B-Instruct and Kimi-Audio-7B-Instruct, as well as proprietary models like Gemini-2.5-Pro, achieving superior performance in this evaluation. A key contributing factor is GPT-4o-Realtime's relatively weak performance on Chinese ASR benchmarks, particularly on in-the-wild datasets like **Wenet-test-net** and **Common Voice 15** (CV-15).
2. Kimi-Audio-7B-Instruct excels in ASR and EMO tasks but underperforms in AST. In contrast, Qwen3-Omni-30B-A3B-Instruct demonstrates superior performance across all tasks.
3. Qwen2-Audio-Instruction exhibits a slight alignment gap compared to its base model, Qwen2-Audio.

4.3 Audio Generation

Table 4: Audio generation performance (↑). *Acoustic metrics (UTMOS | DNSMOS P.835 | DNSMOS P.808, scale 1–5) are evaluated on the generated audio responses from the speech tasks. Best results are in bold.

Models	Speech Web Questions	Speech TriviaQA	Speech CMMLU	SpeechHSK	Speech AlpacaEval	Acoustics*
GPT-4o-Realtime	**51.60**	**69.70**	70.05	**98.69**	**74.00**	4.29|3.44|4.26
GLM-4-Voice	32.00	36.40	52.61	71.06	51.00	4.21|3.46|4.07
MiniCPM-o 2.6	40.00	40.20	51.37	80.68	51.00	4.12|3.39|4.02
Qwen2.5-Omni	38.89	39.94	**73.72**	95.65	54.00	4.23|**3.48**|**4.27**
Kimi-Audio-7B-Instruct	33.69	38.20	71.25	97.42	34.40	2.94|3.22|3.62
Qwen3-Omni-30B-A3B-Instruct	51.50	55.27	47.83	40.27	67.97	**4.44**|3.45|4.12

To evaluate the semantic capabilities of generated audio, we use:

- For **Speech WebQuestions** and **Speech TriviaQA**, we follow the approach in (Nachmani et al., 2023) and (Défossez et al., 2024). Specifically, we transcribe the model's spoken responses using Whisper-large-v3, and a response is considered correct if the ground-truth answer appears in the transcription.
- For **Speech AlpacaEval**, we adopt the evaluation protocol of (Zeng et al., 2024), employing GPT-4o-mini to assess the quality of the transcribed responses. Responses are rated on a scale of 1 to 10, following the MT-Bench rubric (Zheng et al., 2023).
- For **SpeechCMMLU** and **SpeechHSK**, we employ Paraformer-zh to transcribe the generated audio. The transcriptions are then matched with the multiple-choice options to calculate accuracy.

We evaluate acoustic capabilities using the generated audio responses from the aforementioned benchmarks, employing UTMOS, DNSMOS P.835, and DNSMOS P.808 as metrics.

Notably, Moshi, Mini-Omni, and Llama-Omni are excluded due to their lack of support for the Chinese language.

The performance of all models is summarized in Table 4, with key findings as follows:

1. GPT-4o-Realtime continues to lead in audio generation, particularly excelling in semantic quality.
2. Qwen3-Omni-30B-A3B-Instruct and Qwen2.5-Omni outperform GPT-4o-Realtime in acoustic metrics.
3. Kimi-Audio-7B-Instruct underperforms in acoustic quality, suggesting that the naturalness and clarity of its generated speech still have room for improvement.

4.4 Audio Codec

We evaluate audio codecs using clean speech corpora, including **LibriSpeech-dev-clean** (en), **LibriSpeech-test-clean** (en), and **Aishell-1** (zh).

Table 5: Audio Codec Performance: ASR-WER (↓), ASR-CER (↓), SIM (↑), and Quality (UTMOS|DNSMOS P.835|DNSMOS P.808, ↑). Note: The hyphen (-) indicates that UTMOS is not applicable to Chinese speech (Aishell-1). Best results are in bold.

Models	Librispeech-dev-clean			Librispeech-test-clean			Aishell-1		
	ASR-WER	SIM	Quality	ASR-WER	SIM	Quality	ASR-CER	SIM	Quality
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
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
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
Mimi (32bit)	**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
Mimi (8bit)	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
Mimi-streaming (8bit)	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
WavTokenizer-large-v2-75-tokens	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**
WavTokenizer-large-40-tokens	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
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**

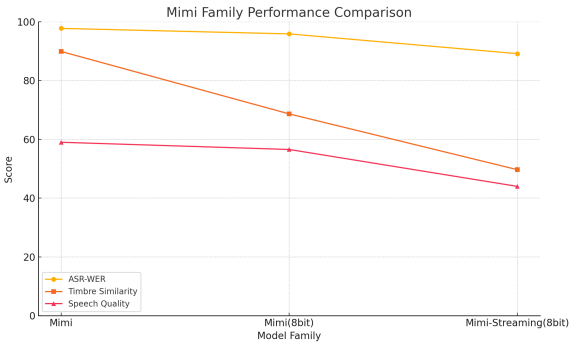

Figure 2: Mimi family performance comparison. Note: ASR-WER is normalized as $(100 - \text{WER})$, and Speech Quality scores are scaled by a factor of 20 for visualization.

The results are shown in Table 5, from which we observe the following:

1. ASR-WER performance shows only limited disparity across models, whereas timbre fidelity and speech quality exhibit substantially larger variation. These latter dimensions provide more discriminative signals for assessing codec performance.
2. The Mimi model performs best in ASR-WER and timbre fidelity, indicating that its token representation effectively captures both linguistic and timbral information from raw audio. However, its speech quality score lags behind Spark and WavTokenizer-large-v2-75-tokens, suggesting that improvements could be made to its decoder component.
3. As shown in Figure 2, comparing Mimi (default 32-bit) with Mimi (8bit), the performance drop is most pronounced in timbre fidelity, while ASR-WER slightly decreases, and speech quality drops modestly. This indicates that timbre information relies more heavily on higher bit depths. The streaming variant further degrades both timbre fidelity and speech quality.
4. ChatTTS-DVAE, WavTokenizer-large-40-tokens, and Mimi-Streaming (8bit) underperform on the AISHELL-1 dataset, indicating a need for improved handling of Chinese-language audio.

5 Conclusion

In this paper, we analyze the development trends of audio foundation models and their evaluation. We present **UltraEval-Audio**, the first unified framework for the comprehensive evaluation of audio understanding and generation.

To mitigate the limited multilingual diversity in speech datasets, we introduce two benchmarks: **SpeechCMMLU** and **SpeechHSK**. These fill a critical gap in the field and enable the effective evaluation of models' Chinese speech capabilities.

Given the importance of audio codecs in foundation models, we propose a new holistic evaluation method designed to assist developers in selecting appropriate components.

Finally, we evaluate several popular and emerging models, presenting leaderboards for **Audio Understanding**, **Audio Generation**, and **Audio Codec**, providing the research community with a transparent and comparative reference.

6 Limitations and Future Directions

Our study has several key limitations. First, some current speech benchmarks rely on transcribed text as input for GPT-based evaluators rather than raw audio. This design introduces a dependency on ASR performance, which may propagate transcription errors into downstream judgments. Future work should therefore explore evaluation pipelines that operate directly on raw audio signals. In addition, the existing evaluation metrics are predominantly technical and do not adequately capture human perceptual factors, such as prosody, emotion, and whether the tone of the system's reply is appropriate for a given conversational context.

In future work, we plan to continuously update and refine the leaderboards, imporve inference capabilities (e.g. multi-GPU support), and incorporate evaluation methods that score responses directly from raw audio.

These enhancements will increase the comprehensiveness and reliability of audio foundation model evaluation, providing clearer guidance for the advancement of the field.

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*, 2019.

P. Bentley, G. Nordehn, M. Coimbra, and S. Mannor. The PASCAL Classifying Heart Sounds Challenge 2011 (CHSC2011) Results. http://www.peterjbentley.com/heartchallenge/index.html.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1877–1901. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline. In *2017 20th conference of the oriental chapter of the international coordinating committee on speech databases and speech I/O systems and assessment (O-COCOSDA)*, pp. 1–5. IEEE, 2017.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*, 2021.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*, 2024.

Michael Chinen, Felicia SC Lim, Jan Skoglund, Nikita Gureev, Feargus O'Gorman, and Andrew Hines. Visqol v3: An open source production ready objective speech and audio metric. In *2020 twelfth international conference on quality of multimedia experience (QoMEX)*, pp. 1–6. IEEE, 2020.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models, 2023. URL https://arxiv.org/abs/2311.07919.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report, 2024. URL https://arxiv.org/abs/2407.10759.

Alexis Conneau, Min Ma, Simran Khanuja, Yu Zhang, Vera Axelrod, Siddharth Dalmia, Jason Riesa, Clara Rivera, and Ankur Bapna. Fleurs: Few-shot learning evaluation of universal representations of speech. *arXiv preprint arXiv:2205.12446*, 2022. URL https://arxiv.org/abs/2205.12446.

OpenCompass Contributors. Opencompass: A universal evaluation platform for foundation models. https://github.com/open-compass/opencompass, 2023.

deepcontractor. Musical instrument chord classification. https://www.kaggle.com/datasets/deepcontractor/musical-instrument-chord-classification, 2023. Accessed: 2025-05-13.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. Moshi: a speech-text foundation model for real-time dialogue. Technical report, 2024. URL 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, et al. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*, 2025.

Ensheng Dong, Hongru Du, and Lauren Gardner. An interactive web-based dashboard to track covid-19 in real time. *The Lancet infectious diseases*, 20(5):533–534, 2020.

Konstantinos Drossos, Samuel Lipping, and Tuomas Virtanen. Clotho: An audio captioning dataset. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 736–740. IEEE, 2020.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, et al. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *arXiv preprint arXiv:2412.10117*, 2024.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, et al. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 11198–11201, 2024.

Kate Dupuis and M Kathleen Pichora-Fuller. Toronto emotional speech set (tess)-younger talker_happy. 2010.

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. High fidelity neural audio compression, 2022.

Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Mohammad Norouzi, Douglas Eck, and Karen Simonyan. Neural audio synthesis of musical notes with wavenet autoencoders. In *International conference on machine learning*, pp. 1068–1077. PMLR, 2017.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. Llama-omni: Seamless speech interaction with large language models, 2025. URL 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, and Vijay Janapa Reddi. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*, 2021.

Zhifu Gao, Zerui Li, Jiaming Wang, Haoneng Luo, Xian Shi, Mengzhe Chen, Yabin Li, Lingyun Zuo, Zhihao Du, Zhangyu Xiao, et al. Funasr: A fundamental end-to-end speech recognition toolkit. *arXiv preprint arXiv:2305.11013*, 2023.

Yuan Gong, Jin Yu, and James Glass. Vocalsound: A dataset for improving human vocal sounds recognition. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 151–155, 2022. doi: 10.1109/ICASSP43922.2022.9746828.

Anmol Gulati, James Qin, Chung-Cheng Chiu, Niki Parmar, Yu Zhang, Jiahui Yu, Wei Han, Shibo Wang, Zhengdong Zhang, Yonghui Wu, et al. Conformer: Convolution-augmented transformer for speech recognition. *arXiv preprint arXiv:2005.08100*, 2020.

Chaoqun He, Renjie Luo, Shengding Hu, Yuanqian Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms, 2024.

Andrew Hines, Jan Skoglund, Anil C Kokaram, and Naomi Harte. Visqol: an objective speech quality model. *EURASIP Journal on Audio, Speech, and Music Processing*, 2015(1):13, 2015.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, and ... others. Step-audio: Unified understanding and generation in intelligent speech interaction, 2025. URL https://arxiv.org/abs/2502.11946.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, et al. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*, 2024.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*, 2017.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 119–132, 2019.

Tony Lee, Michihiro Yasunaga, Chenlin Meng, Yifan Mai, Joon Sung Park, Agrim Gupta, Yunzhi Zhang, Deepak Narayanan, Hannah Teufel, Marco Bellagente, et al. Holistic evaluation of text-to-image models. *Advances in Neural Information Processing Systems*, 36:69981–70011, 2023.

Tony Lee, Haoqin Tu, Chi Heem Wong, Zijun Wang, Siwei Yang, Yifan Mai, Yuyin Zhou, Cihang Xie, and Percy Liang. Ahelm: A holistic evaluation of audio-language models. *arXiv preprint arXiv:2508.21376*, 2025.

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. Cmmlu: Measuring massive multitask language understanding in chinese, 2023a.

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Alpacaeval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval, 5 2023b.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2024.

Mathieu Moreaux. Audio cats and dogs. https://www.kaggle.com/datasets/mmoreaux/audio-cats-and-dogs, 2023. Accessed: 2025-05-13.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*, 2023.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*, 2017.

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and ... others. Gpt-4o system card, 2024. URL https://arxiv.org/abs/2410.21276.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 5206–5210. IEEE, 2015.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. Meld: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint arXiv:1810.02508*, 2018.

Vineel Pratap, Qiantong Xu, Anuroop Sriram, Gabriel Synnaeve, and Ronan Collobert. Mls: A large-scale multilingual dataset for speech research. *ArXiv*, abs/2012.03411, 2020.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. Dnsmos p. 835: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2022-2022 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pp. 886–890. IEEE, 2022.

BM Rocha, Dimitris Filos, Lea Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, et al. A respiratory sound database for the development of automated classification. In *Precision Medicine Powered by pHealth and Connected Health: ICBHI 2017, Thessaloniki, Greece, 18-21 November 2017*, pp. 33–37. Springer, 2018.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pp. 125–129, 2012.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. Utmos: Utokyo-sarulab system for voicemos challenge 2022, 2022. URL https://arxiv.org/abs/2204.02152.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*, 2024.

B Series. Method for the subjective assessment of intermediate quality level of audio systems. *International Telecommunication Union Radiocommunication Assembly*, 2, 2014.

Sripaa D. Srinivasan. Audio mnist. https://www.kaggle.com/datasets/sripaadsrinivasan/audio-mnist, 2023. Accessed: 2025-05-13.

Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint arXiv:1306.1461*, 2013.

Sidharth Surapaneni, Hoang Nguyen, Jash Mehta, Aman Tiwari, Oluwanifemi Bamgbose, Akshay Kalkunte, Sai Rajeswar, and Sathwik Tejaswi Madhusudhan. Au-harness: An open-source toolkit for holistic evaluation of audio llms. *arXiv preprint arXiv:2509.08031*, 2025.

Cees H Taal, Richard C Hendriks, Richard Heusdens, and Jesper Jensen. An algorithm for intelligibility prediction of time–frequency weighted noisy speech. *IEEE Transactions on audio, speech, and language processing*, 19(7):2125–2136, 2011.

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, et al. Kespeech: An open source speech dataset of mandarin and its eight subdialects. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021.

Nicolas Turpault, Romain Serizel, Ankit Parag Shah, and Justin Salamon. Sound event detection in domestic environments with weakly labeled data and soundscape synthesis. In *Workshop on Detection and Classification of Acoustic Scenes and Events*, 2019.

Changhan Wang, Anne Wu, and Juan Pino. Covost 2: A massively multilingual speech-to-text translation corpus, 2020.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. Voxpopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. *arXiv preprint arXiv:2101.00390*, 2021.

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, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025. URL https://arxiv.org/abs/2503.01710.

Zhifei Xie and Changqiao Wu. Mini-omni: Language models can hear, talk while thinking in streaming, 2024. URL https://arxiv.org/abs/2408.16725.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

Peng Xu, Wenqi Shao, Kaipeng Zhang, Peng Gao, Shuo Liu, Meng Lei, Fanqing Meng, Siyuan Huang, Yu Qiao, and Ping Luo. Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*, 2024.

Tiansheng Yao, Xinyang Yi, Derek Zhiyuan Cheng, Felix Yu, Ting Chen, Aditya Menon, Lichan Hong, Ed H Chi, Steve Tjoa, Jieqi Kang, et al. Self-supervised learning for large-scale item recommendations. In *Proceedings of the 30th ACM international conference on information & knowledge management*, pp. 4321–4330, 2021.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, et al. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 25697–25705, 2025.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot, 2024. URL 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, et al. Wenetspeech: A 10000+ hours multi-domain mandarin corpus for speech recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 6182–6186. IEEE, 2022.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.