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
{yudongwang, liuzy}@tsinghua.edu.cn

Abstract

The development of audio foundation models has accelerated rapidly since the emergence of GPT-4o. However, the lack of comprehensive evaluation has become a critical bottleneck for further progress in the field, particularly in audio generation. Current audio evaluation faces three major challenges: (1) audio evaluation lacks a unified framework, with datasets and code scattered across various sources; (2) audio codec, as a key component of audio foundation models, lacks a widely accepted and holistic evaluation methodology; (3) existing speech benchmarks are heavily reliant on English, making it challenging to objectively assess models' performance on Chinese. We introduce **UltraEval-Audio**, a unified framework addressing these challenges through a modular architecture supporting 10 languages, 14 task categories, 24 models, and 36 benchmarks with one-command evaluation and real-time leaderboards. For audio codec, we propose a three-dimensional evaluation scheme covering semantic accuracy, timbre fidelity, and acoustic quality. For Chinese evaluation, we introduce two new benchmarks: *SpeechCMMLU* and *SpeechHSK*. Our code, benchmarks, and leaderboards are available at https://github.com/OpenBMB/UltraEval-Audio.

1 Introduction

The rapid advancement of audio foundation models (AFMs), such as GPT-4o (OpenAI et al., 2024), Qwen2.5-Omni (Xu et al., 2025), Moshi (Défossez et al., 2024), GLM-4-Voice (Zeng et al., 2024), Step-Audio (Huang et al., 2025), Kimi-Audio (Ding et al., 2025), and MiniCPM-o 2.6 (Yao et al., 2024) has significantly expanded the boundaries of human-computer interaction. However, the lack of comprehensive evaluation has become a critical bottleneck for further progress. Current audio evaluation faces three major challenges: (1) the absence of a unified framework, with datasets and code scattered across sources, hindering fair cross-model comparison; (2) the lack of systematic evaluation for audio codec—a core component of AFMs; and (3) the heavy reliance on English benchmarks, which inadequately assess model performance on Chinese.

Existing evaluation tools are often designed for specific tasks (e.g., ASR, AST) and fail to accommodate general-purpose AFMs. While recent efforts like AudioBench (Chen et al., 2024) and AHELM (Lee et al., 2025) have made progress, they lack support for audio generation tasks or require cumbersome procedures. Moreover, mainstream benchmarks such as SpeechTriviaQA (Défossez et al., 2024) and SpeechAlpacaEval (Fang et al., 2025) are English-centric, leaving a gap for Chinese evaluation.

To address these issues, we introduce **UltraEval-Audio**, a unified evaluation framework for AFMs. It features a modular architecture with "one-command" evaluation, supporting 10 languages, 14 task categories, 24 mainstream models, and 36 benchmarks. Our key contributions are (1) **The first unified audio evaluation framework**. UltraEval-Audio supports a wide range of input-output modalities, including "Text → Audio", "Text + Audio → Text", "Audio → Text", and "Text + Audio → Audio". The framework supports 10 languages, 14 core task categories and integrates 24 mainstream models and 36 authoritative benchmarks, covering three key audio types: speech, environmental sound, and music. With its user-friendly design, the framework offers a "one-command" evaluation feature and publicly available leaderboards for transparent comparisons. (2) **A multi-dimensional evaluation scheme for audio codec**. We have established a systematic evaluation scheme that covers semantic accuracy, timbre fidelity, and acoustic quality, ad-

[*]Equal Contributions.
[†]Corresponding Authors.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 566–577

dressing the lack of widely accepted and systematic multi-dimensional performance metrics for audio codec. (3) **Two new Chinese evaluation benchmarks**. We propose **SpeechCMMLU** and **SpeechHSK**, which are designed to systematically measure the knowledge proficiency and language fluency of AFMs in the Chinese context. The framework, code, and leaderboards are publicly available at https://github.com/OpenBMB/UltraEval-Audio under the Apache-2.0 license.

2 Related Work

Audio Foundation Models. Emerging AFMs typically adopt an **Audio Codec+LLM** architecture with three components: an audio tokenizer, an LLM backbone, and a vocoder. Based on vocoder inclusion, they fall into two categories: (1) **audio understanding models** (e.g., Qwen-Audio (Chu et al., 2023), Gemini-1.5 (Team et al., 2024)), which output text only; and (2) **audio generation models** (e.g., GPT-4o-Realtime (OpenAI et al., 2024), Moshi (Défossez et al., 2024), MiniCPM-o 2.6 (Yao et al., 2024)), which output both speech and text. Meanwhile, audio codec like EnCodec (Défossez et al., 2022), DAC (Kumar et al., 2023), and X-codec (Ye et al., 2025) has evolved toward scalable, low-latency tokenization for LLM-based speech generation.

Audio Evaluation Frameworks. While comprehensive frameworks exist for text (e.g., OpenCompass (Contributors, 2023), UltraEval (He et al., 2024)) and vision (e.g., VLMEvalKit (Duan et al., 2024)), a unified audio evaluation framework has been lacking. Recent efforts include: AudioBench (Chen et al., 2024) (8 tasks, 26 benchmarks) and AHELM (Lee et al., 2025) (10 evaluation aspects), but both lack coverage of audio generation tasks. Kimi-Audio-Evalkit (Ding et al., 2025) reproduces Kimi-Audio's benchmarks but requires five-step procedures and code-level prompt modifications. AU-Harness (Surapaneni et al., 2025) supports 380+ tasks but demands manual model adaptation to vLLM services.

Audio Evaluation Benchmarks. Beyond traditional ASR/AST, user-centric benchmarks have emerged. AIR-Bench (Yang et al., 2024) collects spoken QA with GPT-4 evaluation, and VoiceBench (Chen et al., 2024) adds synthetic instructions, Llama-Question (Nachmani et al., 2023) introduced speech QA with ASR-based evaluation, followed by SpeechWebQuestions (Nachmani et al., 2023), SpeechTriviaQA (Défossez et al., 2024), and SpeechAlpacaEval (Fang et al., 2025). However, all these benchmarks are English-only, leaving Chinese evaluation unexplored.

Audio Codec Evaluation. Recent audio codec evaluation combines subjective (MUSHRA (Series, 2014)) and objective metrics: ViSQOL (Hines et al., 2015) for perceptual quality, SI-SNR for reconstruction fidelity, STOI (Taal et al., 2011) for intelligibility, and speaker similarity (SIM) via embedding models. (Ye et al., 2025) also employs downstream TTS tasks to indirectly evaluate codec performance, but a systematic framework remains absent.

3 Audio Evaluation Design and Methodology

We outline a systematic audio evaluation design for AFMs by establishing a unified evaluation taxonomy. Building on this taxonomy, we develop a comprehensive audio codec evaluation methodology and introduce two Chinese speech benchmarks, SpeechCMMLU and SpeechHSK.

3.1 Audio Evaluation Taxonomy

UltraEval-Audio designs a capability-driven task taxonomy to organize evaluation. As shown in Table 1, tasks are grouped into three high-level categories: **audio understanding**, **audio generation**, and **audio codec**, further divided by domain (speech, music, environmental sounds). Each task is associated with standardized metrics (e.g., WER for ASR, BLEU for translation, accuracy for classification, and ASR-WER/UTMOS for generation/codec tasks).

Based on this taxonomy, we instantiate evaluation with **36 benchmarks** covering 14 tasks and 10 languages (Table 2 summarizes key datasets). This design ensures comprehensive, multi-dimensional assessment, e.g., ASR spans clean, studio-quality recordings (LibriSpeech) to in-the-wild speech (WenetSpeech), and single-language (AISHELL-1) to multilingual (FLEURS) conditions.

3.2 Codec Evaluation

Audio codec is a fundamental component of audio foundation models, yet existing evaluation approaches rely on diverse metrics with inconsistent standards. To address this, we propose a three-dimensional codec evaluation methodology encompassing **semantics**, **timbre fidelity**, and **acoustic quality**.

Category	Domain	Task	Description	Metrics
Audio Understanding	Speech	ASR	Given speech audio, produce a transcription.	WER / CER
		AST	Given speech audio in the source language, generate a text translation in the target language.	BLEU
		Gender Analysis	Given speech audio, predict the speaker's gender.	Acc.
		Speech QA	Given speech audio, generate a textual answer to the corresponding question.	Exist Match / G-Eval
		Emotion Analysis	Given speech audio, identify the speaker's emotional state.	Acc.
	Music	Instrument Recognition	Given music audio, classify the predominant instrument.	Acc.
		Music Genre	Given music audio, classify the corresponding music genre.	Acc.
		Chord Recognition	Given music audio, identify the sequence of chord labels.	Acc.
	Environment Sounds	Audio Classification	Given non-speech audio, classify it into a predefined scene or event category.	Acc.
		Audio Captioning	Given non-speech audio, generate a natural language description of its content.	BLEU / ROUGE-L
Audio Generation	Speech	TTS	Given input text, synthesize the corresponding speech audio.	ASR-WER
		VC	Given input text and a reference speech sample, synthesize speech in the target speaker's voice.	ASR-WER / SIM
		Speech QA	Given speech audio, generate an appropriate spoken response.	Exist Match / G-Eval / UTMOS
Audio Codec	Speech	Speech Codec	Encode and reconstruct speech audio from discrete representations.	ASR-WER / SIM / UTMOS

Table 1: Task taxonomy in UltraEval-Audio. WER: Word Error Rate; CER: Character Error Rate; BLEU/ROUGE-L: metrics for evaluating the quality of generated text; Acc.: classification accuracy; SIM: speaker embedding cosine similarity; UTMOS: an objective speech quality evaluation metric; G-Eval: GPT-based evaluation metric; ASR-WER: computed by transcribing the generated or reconstructed speech with an ASR model and then calculating WER on the transcriptions.

Task	Language	Dataset
ASR	en	TED-LIUM (Rousseau et al., 2012), VoxPopuli (Wang et al., 2021), LibriSpeech (Panayotov et al., 2015) The People's Speech (Galvez et al., 2021), WenetSpeech (Zhang et al., 2022) GigaSpeech (Chen et al., 2021), AudioMNIST (Srinivasan, 2023)
	zh	KeSpeech (Tang et al., 2021), AISHELL-1 (Bu et al., 2017)
	nl, fr, de, it, pl, pt, es	MLS (Pratap et al., 2020)
	zh, en, ru, de, jp, ...	FLEURS (Conneau et al., 2022), Common Voice (Ardila et al., 2019)
AST	zh, en, ru, de, jp, ...	CoVoST 2 (Wang et al., 2020)
VC	zh, en	Seed-TTS-Eval (Anastassiou et al., 2024), CV3-Eval (Du et al., 2025)
TTS	zh, en	Long-TTS-Eval (Wang et al., 2025a)
Speech Codec	en	LibriSpeech
	zh	AISHELL-1
Speech QA	en	SpeechTriviaQA (Défossez et al., 2024), SpeechWebQuestions (Nachmani et al., 2023) SpeechAlpacaEval (Fang et al., 2025), LLaMA-Questions (Nachmani et al., 2023) AIR-Bench (Yang et al., 2024), MMAU (Sakshi et al., 2024)
	zh	SpeechHSK*, SpeechCMMLU*
Emotion Analysis	en	TESS (Dupuis and Pichora-Fuller, 2010), MELD (Poria et al., 2018)
Gender Analysis	en	VoxCeleb (Nagrani et al., 2017)
Chord Recognition	-	Chord (deepcontractor, 2023)
Instrument Recognition	-	NSynth (Engel et al., 2017)
Music Genre	-	GTZAN (Sturm, 2013)
Caption	-	AudioCaps (Kim et al., 2019), WavCaps (Mei et al., 2024), Clotho (Drossos et al., 2020)
Audio Classification	-	CatDog (Moreaux, 2023), DESED (Turpault et al., 2019) VocalSound (Gong et al., 2022), COVID-19 Sounds (Dong et al., 2020) PASCAL CHSC 2011 (Bentley et al.), ICBHI 2017 Respiratory Sound (Rocha et al., 2018)

Table 2: Benchmarks supported in UltraEval-Audio. * indicates new benchmarks introduced in this paper.

Semantic preservation is measured via word error rate (WER) by transcribing reconstructed audio with Whisper-large-v3 (English) and Paraformer-zh (Chinese) and comparing to original transcripts. Timbre fidelity quantifies speaker characteristic retention through cosine similarity of WavLM-large speaker embeddings extracted from original and reconstructed audio. Acoustic quality combines UTMOS (Saeki et al., 2022) for overall naturalness with DNSMOS P.835 (Reddy et al., 2022) and P.808 (Reddy et al., 2021) for perceptual quality assessment in noisy environments.

3.3 Chinese Benchmarks

Existing speech QA benchmarks are English-centric. We introduce two Chinese benchmarks: **SpeechCMMLU** extends CMMLU (Li et al., 2023) to the speech modality by converting each multiple-choice item into a standardized instruction, synthesizing speech with a high-quality Chinese TTS system (CosyVoice2), and filtering samples via ASR back-transcription to ensure exact text fidelity (CER = 0), resulting in 3,519 high-quality samples spanning diverse subjects for evaluating Chinese knowledge reasoning. **SpeechHSK** is built from the listening section of official Hanyu Shuiping Kaoshi (HSK[1]) exams: question stems use the original recordings, while answer options are re-recorded by native speakers. It is organized into six proficiency levels with 170 samples for diagnostic assessment of Chinese listening comprehension. Together, they form a multi-level evaluation suite for Chinese speech. SpeechCMMLU targets

[1]https://www.chinesetest.cn

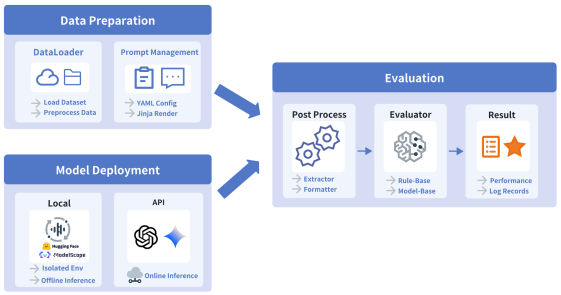

Figure 1: Overview of the UltraEval-Audio framework

knowledge reasoning, while SpeechHSK focuses on general language proficiency. Both benchmarks are constructed with rigorous quality control.

4 UltraEval-Audio Framework

We present the engineering architecture of UltraEval-Audio, designed to address the integration challenges arising from diverse input modalities and heterogeneous model interfaces in AFM evaluation.

As illustrated in Figure 1, the system consists of three main modules: data preparation, model deployment, and evaluation. The data preparation module standardizes diverse audio-text formats into a unified intermediate representation. The model deployment module integrates diverse inference backends through a unified interface, allowing different AFMs to be seamlessly integrated. The evaluation module manages post-processing, evaluator, and result generation in an automated manner.

Beyond modality coverage, we place equal emphasis on engineering extensibility and practical usability. In addition to supporting audio understanding, generation and codec tasks, the framework is designed to be fully configuration-driven and easy to deploy through one-command evaluation with automatic data and model management. The framework adopts a decoupled modular design with a configuration-driven workflow.

To position UltraEval-Audio within the landscape of existing evaluation toolkits, we compare it with representative frameworks as shown in Table 3.

4.1 Data Preparation

Data loader. UltraEval-Audio adopts an audio-centric dataset organization scheme with a reserved *audio* field. The framework provides built-in pipelines for automatic downloading, caching, decoding, and format standardization from sources like HuggingFace, significantly simplifying multimodal preprocessing.

Framework	Understanding	Generation	Configurable	One-command Eval
AudioBench	✓	✗	✗	✗
Kimi-Audio-Evalkit	✓	✗	✗	✗
AU-Harness	✓	✗	✓	✗
Ours	✓	✓	✓	✓

Table 3: Comparison with existing frameworks. **Understanding**: support for audio understanding tasks. **Generation**: support for audio generation tasks. **Configurable**: supports configuration-driven runtime without requiring any code modification. **One-command Eval**: support for one-command evaluation with automatic data/model downloading.

Prompt management. Given AFMs can be sensitive to prompts, we introduce a configuration-driven mechanism using YAML-based configurations and Jinja templating. This enables flexible prompt construction at task, model, or sample level without code modification. The module consists of three core components: a Prompt Register that indexes templates for different tasks, a YAML Parser that loads model- or task-specific prompt definitions from configuration files, and a renderer that injects dataset fields (e.g., audio) into templates. Figure 2 shows an example ASR prompt for MiniCPM-o 2.6, where {{audio}} is dynamically replaced at inference time.

```
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 lower case."
      - type: audio
        value: '{{audio}}'
```

Figure 2: An example ASR prompt for MiniCPM-o 2.6

4.2 Model Deployment

Model deployment poses two practical challenges: supporting both remote APIs and local models, and managing dependency conflicts for local models. To address these issues, we propose an **Isolated Runtime** mechanism. Concretely, each local model is executed within its own dedicated virtual environment that contains only its required dependencies, and is launched as a daemonized subprocess to preserve persistent state for continuous inference. The main process communicates with these model subprocesses through lightweight inter-process communication (IPC) over system pipes, enabling secure and low-latency data ex-

change.

This microservice-style architecture eliminates dependency conflicts while providing a unified *.inference()* interface across all model types.

4.3 Evaluation

The evaluation phase converts raw model outputs into quantifiable metrics through two components: **Post-processing** standardizes outputs using modular, composable modules (e.g., *Option Extraction*, *Yes/No Parser*, *JSON Field Parser*) that can be chained for complex tasks. **Evaluators** then compute final scores using either rule-based metrics (e.g., WER for ASR, BLEU for AST, and accuracy for classification) or model-based metrics (e.g., Speaker Similarity (SIM) for timbre, UTMOS/DNSMOS for speech quality, and LLM-as-a-Judge for open-ended QA). All evaluators are exposed through a unified registration interface, and their results are automatically aggregated to ensure consistent and reproducible reporting.

5 Evaluation Results

UltraEval-Audio provides a unified solution for systematically assessing audio models. We construct three leaderboards, covering audio understanding, generation, and codec evaluation, over 13 leading audio foundation models (Table 4) and 9 audio codecs (with detailed descriptions provided in Appendix A). These leaderboards span diverse multilingual benchmarks and assessment dimensions, enabling systematic comparison across different model architectures and interaction paradigms. Beyond reporting aggregate results, we further analyze benchmark relationships and the robustness of leaderboard rankings. All results are publicly available on our leaderboard (https://github.com/OpenBMB/UltraEval-Audio/blob/main/README.md#leaderboard).

5.1 Audio Understanding

We evaluate ASR (LibriSpeech, TED-LIUM, AISHELL-1, WenetSpeech, CommonVoice), AST (CoVoST 2), and emotion recognition (MELD). As shown in Table 5, key observations include: (1) GPT-4o-Realtime faces strong competition in the field of audio understanding, with open-source models such as Qwen3-Omni-30B-A3B-Instruct, Kimi-Audio-7B-Instruct, MiniCPM-o 2.6 as well as proprietary models like Gemini-2.5-Pro, achieving superior performance in the evaluation; (2) Qwen3-Omni-30B-A3B-Instruct demonstrates superior performance across all tasks, consistently delivering high-quality results in various domains. Kimi-Audio-7B-Instruct excels in ASR and EMO tasks but underperforms in AST, indicating room for improvement in the latter area.

5.2 Audio Generation

We assess spoken QA on English benchmarks (SpeechWebQuestions, SpeechTriviaQA, SpeechAlpacaEval) and our new **Chinese benchmarks** (SpeechCMMLU, SpeechHSK). Acoustic quality is measured via UTMOS and DNSMOS. Table 6 summarizes the results, with key findings: (1) GPT-4o-Realtime performs best in audio generation, particularly excelling in English benchmarks. It consistently produces high-quality, accurate, and natural-sounding speech, making it one of the top performers; (2) Qwen3-Omni-30B-A3B-Instruct and Qwen2.5-Omni outperform GPT-4o-Realtime in acoustic metrics, demonstrating superior sound quality. These models offer more nuanced audio generation, providing richer and more realistic speech outputs; (3) Kimi-Audio-7B-Instruct underperforms in acoustic quality, with its generated speech lacking some naturalness and clarity. This suggests that improvements are needed to make the output sound more natural and human-like.

5.3 Audio Codec

We evaluate 9 codecs using our three-dimensional scheme (semantics: ASR-WER; timbre: speaker similarity; acoustics: UTMOS/DNSMOS) on LibriSpeech and AISHELL-1. Table 7 and Figure 3 highlight key findings: (1) ASR-WER performance shows only limited disparity across models, whereas timbre fidelity and acoustic quality exhibit substantially larger variation. These latter dimensions provide more discriminative signals for assessing codec performance. (2) The Mimi model performs best in semantics accuracy and timbre fidelity, indicating that its token representation effectively captures both linguistic and timbral information from raw audio. (3) We further propose Figure 3 to compare Mimi codec variants. Comparing Mimi (default 32-bit) with Mimi (8bit), the performance drop is most pronounced in timbre fidelity, while semantics performance slightly decreases, and acoustic quality drops modestly. This indicates that timbre information relies more heavily on higher bit depths. The streaming variant further degrades on both timbre fidelity and acous-

Model	Institution	Type	Modality (Input→Output)	Languages
GPT-4o-Realtime	OpenAI	Proprietary	Audio + Text → Audio + Text	Multilingual
Qwen3-Omni-30B-A3B-Instruct	Alibaba	Open-Source	Audio + Text → Audio + Text	Multilingual
Qwen2.5-Omni	Alibaba	Open-Source	Audio + Text → Audio + Text	English, Chinese
MiniCPM-o 2.6	OpenBMB	Open-Source	Audio + Text → Audio + Text	English, Chinese
Kimi-Audio-7B-Instruct	Moonshot	Open-Source	Audio + Text → Audio + Text	English, Chinese
Gemini-1.5-Flash	Google	Proprietary	Audio + Text → Text	Multilingual
Gemini-1.5-Pro	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-7B	Alibaba	Open-Source	Audio + Text → Text	Multilingual
Qwen2-Audio-7B-Instruct	Alibaba	Open-Source	Audio + Text → Text	Multilingual
MiDaShengLM-7B	Xiaomi	Open-Source	Audio + Text → Text	Multilingual
GLM-4-Voice	Zhipu AI	Open-Source	Audio → Audio	English, Chinese

Table 4: Overview of audio foundation models participating in the evaluation.

Model	ASR (WER/CER↓)	AST (BLEU↑)	EMO (Acc.↑)	Avg. Score↑
GPT-4o-Realtime	8.04 / 19.76	37.1 / 15.7	33.2	73.8
Qwen3-Omni-30B-A3B-Instruct	**2.71** / **3.16**	46.6 / **29.4**	56.8	**84.9**
Qwen2.5-Omni	4.38 / 4.23	42.5 / 11.5	53.6	81.9
MiniCPM-o 2.6	4.07 / 5.63	**48.2** / 27.2	52.4	83.2
Kimi-Audio-7B-Instruct	2.88 / 3.60	36.6 / 18.3	**59.2**	83.3
Gemini-1.5-Flash	44.37 / 114.79	33.4 / 8.2	45.2	27.8
Gemini-1.5-Pro	4.36 / 9.49	47.3 / 22.6	48.4	81.1
Gemini-2.5-Flash	12.67 / 43.77	3.7 / 10.6	51.5	62.7
Gemini-2.5-Pro	5.22 / 8.87	41.8 / 27.8	46.6	80.7
Qwen2-Audio-7B	3.78 / 5.63	45.3 / 24.8	42.9	82.1
Qwen2-Audio-7B-Instruct	5.63 / 7.05	39.5 / 22.9	17.4	78.3
MiDaShengLM-7B	29.09 / 12.55	38.5 / 22.7	54.0	68.5

Table 5: Overall audio understanding results (Full results and detailed annotations in Appendix Table 9). WER is the mean over LibriSpeech (dev/test clean/other), TED-LIUM, and CV-15 (en); CER is the mean over CV-15 (zh), AISHELL-1, FLEURS, and Wenet-test-net; EMO is accuracy on MELD.

Model	English QA (Acc.)	Chinese QA (Acc.)	Acoustic (0-5)	Avg. Score↑
GPT-4o-Realtime	**51.6/69.7/74.0**	70.1 / **98.7**	4.29 / 3.44 / 4.26	**74.0**
Qwen3-Omni-30B-A3B-Instruct	51.5/55.3/68.0	47.8 / 40.3	**4.44** / 3.45 / 4.12	57.2
Qwen2.5-Omni	38.9/39.9/54.0	**73.7** / 95.7	4.23 / **3.48** / **4.27**	63.7
MiniCPM-o 2.6	40.0/40.2/51.0	51.4 / 80.7	4.12 / 3.39 / 4.02	56.7
Kimi-Audio-7B-Instruct	33.7/38.2/34.4	71.3 / 97.4	2.94 / 3.22 / 3.62	56.7
GLM-4-Voice	32.0/36.4/51.0	52.6 / 71.1	4.21 / 3.46 / 4.07	53.6

Table 6: Overall audio generation results (Full results and detailed annotations in Appendix Table 10). English QA scores are reported on SpeechWebQuestions, SpeechTriviaQA, SpeechAlpacaEval; Chinese QA scores are reported on SpeechCMMLU and SpeechHSK; acoustic quality is assessed using UTMOS, DNSMOS P.835 and P.808.

Model	Semantics (ASR-WER↓)	Timbre (SIM↑)	Acoustics (0-5)	Avg. Score↑
Encodec-24k	4.3 (en) / 14.0 (zh)	59.4 (en) / 47.5 (zh)	1.58/3.12/2.36	65.2
Encodec-48k	3.8 (en) / 6.9 (zh)	66.0 (en) / 68.8 (zh)	1.52/2.88/2.42	69.6
ChatTTS-DVAE	6.8 (en) / 32.4 (zh)	36.2 (en) / 32.4 (zh)	1.30/2.66/2.11	52.9
Mimi (32bit)	**2.0** (en) / **2.8** (zh)	**92.7** (en) / **84.8** (zh)	3.83/2.87/2.44	81.0
Mimi (8bit)	2.8 (en) / 6.8 (zh)	73.1 (en) / 60.6 (zh)	3.52/2.78/2.37	72.7
Mimi-streaming (8bit)	6.2 (en) / 19.6 (zh)	54.3 (en) / 40.7 (zh)	1.65/2.78/2.37	61.4
WavTokenizer-large-75	4.1 (en) / 9.0 (zh)	68.2 (en) / 64.3 (zh)	**4.01**/3.64/**3.26**	76.7
WavTokenizer-large-40	7.7 (en) / 25.5 (zh)	56.6 (en) / 49.2 (zh)	3.78/3.70/3.13	69.2
Spark	2.5 (en) / 3.7 (zh)	79.5 (en) / 74.8 (zh)	**4.18**/**3.85**/3.24	**82.3**

Table 7: Overall audio codec results (Full results and detailed annotations in Appendix Table 11). Semantics: ASR-WER on LibriSpeech test-clean (en) and AISHELL-1 (zh); Timbre: speaker similarity (SIM); Acoustics: UTMOS / DNSMOS P.835 / P.808.

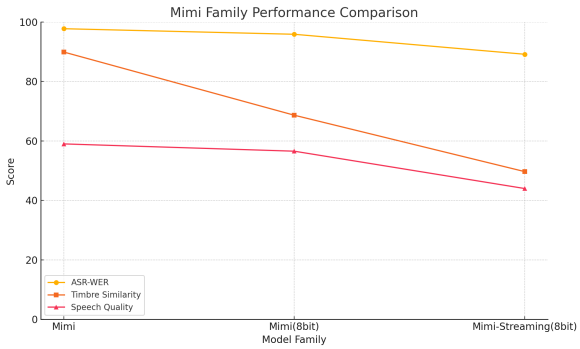

Figure 3: Mimi family performance comparison. Note that ASR-WER is normalized as (100 - WER), and speech quality scores (acoustic scores) are scaled by a factor of 20 for visualization.

tic quality.

5.4 Benchmark Relationship Analysis

To better understand the relationships among benchmarks in the evaluation suite, we analyze model-level performance correlations across benchmarks. Specifically, we compute Spearman rank correlation between benchmark pairs across evaluated models.

As shown in Figure 4, benchmarks within the same task family tend to exhibit stronger correlations. For example, the LibriSpeech splits are highly correlated with each other, indicating that they provide consistent ASR ranking signals under different splits. Similarly, several Chinese ASR benchmarks, such as CV-15 zh and Wenet-test-net, also show strong correlations, suggesting partial overlap in the model capabilities they measure. These high correlations should not be interpreted as making the benchmarks unnecessary, since different splits and datasets still cover different acoustic conditions and data distributions. However, they only indicate partially overlapping ranking signals on the currently evaluated models, rather than intrinsic redundancy between benchmarks.

In contrast, cross-task correlations are generally

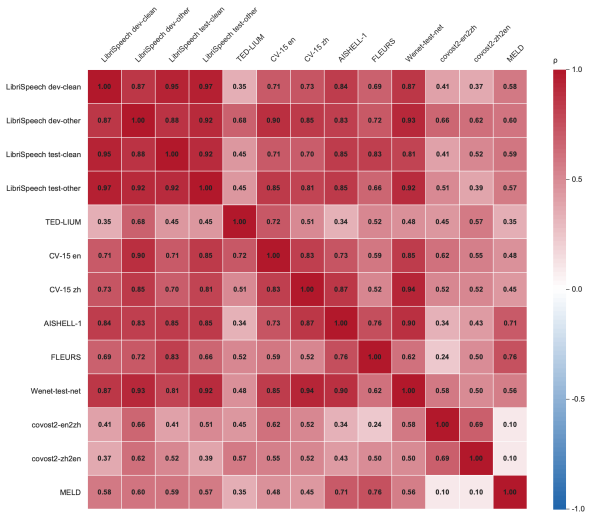

Figure 4: Spearman correlation heatmap among audio understanding benchmarks. WER/CER metrics are converted to $100 - x$, while other metrics are kept unchanged. Correlations are computed across evaluated models.

Model	Original	LOBO		LOTGO		Alt-norm	
		Best	Worst	Best	Worst	Best	Worst
Qwen3-Omni-30B-A3B-Instruct	**1**	**1**	**1**	**1**	**1**	**1**	**1**
Kimi-Audio-7B-Instruct	2	2	4	2	6	2	2
MiniCPM-o 2.6	3	2	3	2	4	3	3
Qwen2-Audio-7B	4	3	5	3	7	4	4
Qwen2.5-Omni	5	4	5	3	8	5	5
Gemini-1.5-Pro	6	6	7	3	6	6	6
Gemini-2.5-Pro	7	6	7	4	7	7	7
Qwen2-Audio-7B-Instruct	8	8	9	8	11	9	9
GPT-4o-Realtime	9	9	10	9	10	10	10
MiDaShengLM-7B	10	8	11	5	10	8	8
Gemini-2.5-Flash	11	10	11	11	12	11	11
Gemini-1.5-Flash	12	12	12	9	12	12	12

Table 8: Ranking sensitivity analysis on the Audio Understanding leaderboard. **Original**: rank under the published rule. **LOBO** (Leave-One-Benchmark-Out): the best and worst rank a model attains when one benchmark is removed at a time; a wide gap exposes over-reliance on a specific dataset. **LOTGO** (Leave-One-Task-Group-Out): the best and worst rank when an entire task group (ASR / AST / EMO) is removed in turn. **Alt-norm**: the best and worst rank under two alternative normalization schemes: z-score normalization and min-max normalization.

weaker. The AST benchmarks and the MELD emotion recognition benchmark show relatively low correlation with many ASR benchmarks, indicating that they capture capability dimensions that are not fully explained by speech recognition performance. Due to space constraints, we provide the corresponding correlation analysis for the other leaderboards in Appendix B and Appendix C.

5.5 Ranking Robustness

The average score reported in our leaderboards is designed as a simple and transparent summary for quick comparison. For heterogeneous evaluation suites, the choice of aggregation rule is partly a convention. Different normalization or benchmark weightings may emphasize different aspects of model behavior. We therefore test leave-one-benchmark-out (LOBO), leave-one-task-group-out (LOTGO), and two alternative normalizations: z-score and min-max normalization. As shown in Table 8, the original normalization and aggregation protocol exhibits reasonable robustness overall, with the main exception being MiDaShengLM-7B, whose ranking is sensitive due to its TED-LIUM result in Table 9.

6 Conclusion

In this paper, we introduce UltraEval-Audio, the first unified evaluation framework for comprehensive assessment of audio foundation models. We construct a complete audio evaluation taxonomy encompassing tasks and benchmarks. Building upon this, we implement a modular evaluation framework, which we then use to evaluate and analyze the performance of popular models across audio understanding, audio generation, and audio codec tasks.

7 Limitations and Future Directions

Our limitations are as follows. 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.

For future work, we plan to continuously update and refine the leaderboards, improve inference capabilities (e.g. multi-GPU support), and incorporate evaluation methods that evaluate 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

Philip Anastassiou, Jiawei Chen, Jitong Chen, Yuanzhe Chen, Zhuo Chen, Ziyi Chen, Jian Cong, Lelai Deng, Chuang Ding, Lu Gao, and 1 others. 2024. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*.

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

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.

Hui Bu, Jiayu Du, Xingyu Na, Bengu Wu, and Hao Zheng. 2017. 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)*, pages 1–5. IEEE.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, and 1 others. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

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

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

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

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

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. Technical report.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, and 1 others. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

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

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

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, and 1 others. 2025. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *arXiv preprint arXiv:2505.17589*.

Haodong Duan, Junming Yang, Yuxuan Qiao, Xinyu Fang, Lin Chen, Yuan Liu, Xiaoyi Dong, Yuhang Zang, Pan Zhang, Jiaqi Wang, and 1 others. 2024. Vlmevalkit: An open-source toolkit for evaluating large multi-modality models. In *Proceedings of the 32nd ACM international conference on multimedia*, pages 11198–11201.

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

Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi. 2022. High fidelity neural audio compression. *Preprint*, arXiv:2210.13438.

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

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

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

Chaoqun He, Renjie Luo, Shengding Hu, Yuanqian Zhao, Jie Zhou, Hanghao Wu, Jiajie Zhang, Xu Han, Zhiyuan Liu, and Maosong Sun. 2024. Ultraeval: A lightweight platform for flexible and comprehensive evaluation for llms. *Preprint*, arXiv:2404.07584.

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

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, and 1 others. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *Preprint*, arXiv:2502.11946.

Shengpeng Ji, Ziyue Jiang, Wen Wang, Yifu Chen, Minghui Fang, Jialong Zuo, Qian Yang, Xize Cheng, Zehan Wang, Ruiqi Li, and 1 others. 2024. Wavtokenizer: an efficient acoustic discrete codec tokenizer for audio language modeling. *arXiv preprint arXiv:2408.16532*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. 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)*, pages 119–132.

Rithesh Kumar, Prem Seetharaman, Alejandro Luebs, Ishaan Kumar, and Kundan Kumar. 2023. High-fidelity audio compression with improved rvqgan. *Advances in Neural Information Processing Systems*, 36:27980–27993.

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

Haonan Li, Yixuan Zhang, Fajri Koto, Yifei Yang, Hai Zhao, Yeyun Gong, Nan Duan, and Timothy Baldwin. 2023. Cmmlu: Measuring massive multitask language understanding in chinese. *Preprint*, arXiv:2306.09212.

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

Mathieu Moreaux. 2023. Audio cats and dogs. https://www.kaggle.com/datasets/mmoreaux/audio-cats-and-dogs. 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. 2023. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*.

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

OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, and 1 others. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

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

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

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

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2021. Dnsmos: A non-intrusive perceptual objective speech quality metric to evaluate noise suppressors. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6493–6497. IEEE.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 2022. 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)*, pages 886–890. IEEE.

BM Rocha, Dimitris Filos, Lea Mendes, Ioannis Vogiatzis, Eleni Perantoni, Evangelos Kaimakamis, P Natsiavas, Ana Oliveira, C Jácome, A Marques, and 1 others. 2018. 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*, pages 33–37. Springer.

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

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. 2022. Utmos: Utokyo-sarulab system for voicemos challenge 2022. *Preprint*, arXiv:2204.02152.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha.

2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

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

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

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

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

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

Zhiyuan Tang, Dong Wang, Yanguang Xu, Jianwei Sun, Xiaoning Lei, Shuaijiang Zhao, Cheng Wen, Xingjun Tan, Chuandong Xie, Shuran Zhou, and 1 others. 2021. 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)*.

Gemini Team, Petko Georgiev, Ving Ian Lei, and 1 others. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

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

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

Changhan Wang, Anne Wu, and Juan Pino. 2020. Covost 2: A massively multilingual speech-to-text translation corpus. *Preprint*, arXiv:2007.10310.

Chengyao Wang, Zhisheng Zhong, Bohao Peng, Senqiao Yang, Yuqi Liu, Haokun Gui, Bin Xia, Jingyao Li, Bei Yu, and Jiaya Jia. 2025a. Mgm-omni: Scaling omni llms to personalized long-horizon speech. *arXiv preprint arXiv:2509.25131*.

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, and 6 others. 2025b. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens. *Preprint*, arXiv:2503.01710.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, and 1 others. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and 1 others. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, and 1 others. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Zhen Ye, Peiwen Sun, Jiahe Lei, Hongzhan Lin, Xu Tan, Zheqi Dai, Qiuqiang Kong, Jianyi Chen, Jiahao Pan, Qifeng Liu, and 1 others. 2025. Codec does matter: Exploring the semantic shortcoming of codec for audio language model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25697–25705.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *Preprint*, arXiv:2412.02612.

Binbin Zhang, Hang Lv, Pengcheng Guo, Qijie Shao, Chao Yang, Lei Xie, Xin Xu, Hui Bu, Xiaoyu Chen, Chenchen Zeng, and 1 others. 2022. 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)*, pages 6182–6186. IEEE.

A Audio Codec Models

Audio codec models participating in the evaluation include Encodec (Défossez et al., 2022), ChatTTS-DVAE[2], the Mimi (Défossez et al., 2024) family, WavTokenizer-large-speech-75token (denoted as WavTokenizer-large-75)[3] , WavTokenizer-large-unify-40token (denoted as WavTokenizer-large-40)[4] (Ji et al., 2024) and Spark (Wang et al., 2025b).

B Audio Generation Spearman Correlation

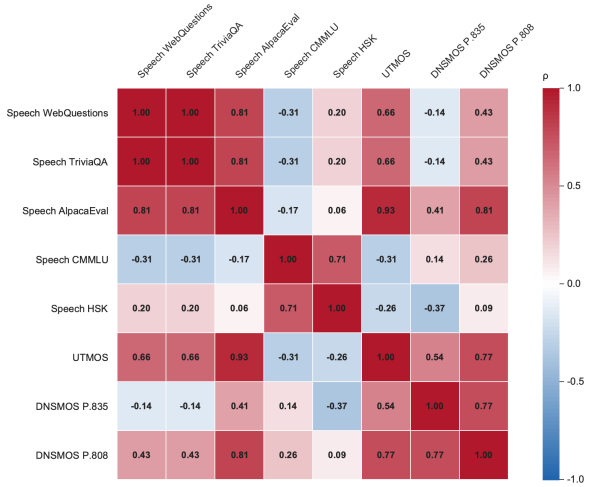

Figure 5: Spearman correlation heatmap among audio generation benchmarks.

C Audio Codec Spearman Correlation

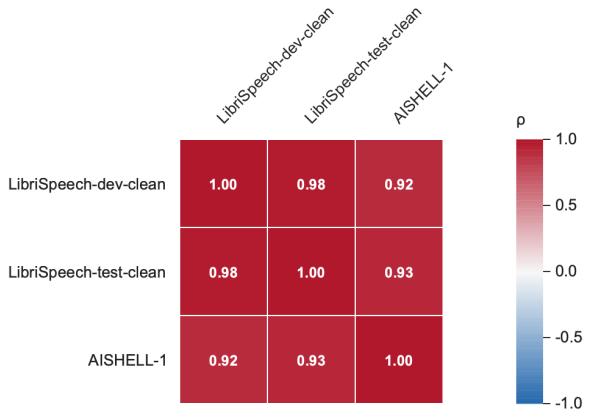

Figure 6: Spearman correlation heatmap among audio codec benchmarks.

[2]https://github.com/2noise/ChatTTS

[3]https://huggingface.co/novateur/WavTokenizer-large-speech-75token

[4]https://huggingface.co/novateur/WavTokenizer-large-unify-40token

Model	ASR						AST		EMO	Avg. Score (↑)
	LibriSpeech dev-clean | dev-other test-clean | test-other	TED-LIUM	CV-15 en | zh	AISHELL-1	FLEURS	Wenet -test-net	covost2-en2zh	covost2-zh2en	MELD	
	WER (↓)	WER (↓)	WER | CER (↓)	CER (↓)	CER (↓)	CER (↓)	BLEU (↑)	BLEU (↑)	Acc. (↑)	
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

Table 9: The audio understanding leaderboard. Best results are in bold. The average score is computed as the mean of all available metric scores, where WER-based metrics use (100-WER) and other metrics (e.g., BLEU/Acc.) are unchanged. WER/CER values can be greater than 100 when the total number of recognition errors exceeds the number of reference words/characters.

Models	Speech WebQuestions	Speech TriviaQA	Speech AlpacaEval	Speech CMMLU	Speech HSK	Acoustics	Avg. Score (↑)
	Acc. (↑)	Acc. (↑)	Acc. (↑)	Acc. (↑)	Acc. (↑)	Acoustics (↑)	
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

Table 10: The audio generation leaderboard. Acoustic metrics (UTMOS | DNSMOS P.835 | DNSMOS P.808, scores range from 0 to 5) are evaluated on the generated audio responses from the speech tasks. Best results are in bold. Note: The average score is computed as the average of 6 scores: five speech-task scores and normalized acoustic scores. For acoustic scores (UTMOS | DNSMOS P.835 | DNSMOS P.808), each value (0–5) is multiplied by 20 to map to 0–100, then averaged to obtain the normalized acoustic score.

Models	LibriSpeech-dev-clean			LibriSpeech-test-clean			AISHELL-1			Avg. Score (↑)
	ASR-WER (↓)	SIM (↑)	Acoustics (↑)	ASR-WER (↓)	SIM (↑)	Acoustics (↑)	ASR-CER (↓)	SIM (↑)	Acoustics (↑)	
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 (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	80.96
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	72.72
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	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**

Table 11: The audio codec leaderboard. The hyphen (-) indicates that UTMOS is not applicable to Chinese speech (AISHELL-1). Best results are in bold. Note: For acoustic scores we also use UTMOS, DNSMOS P.835, and DNSMOS P.808 metrics. To calculate the average score, for ASR-WER and ASR-CER, we calculate $100 - val$. For acoustic scores, each available value (ranges from 0 to 5) is normalized by $20 \times \text{val}$ (mapping to 0–100), and the acoustic score is their average (the hyphen '-' is ignored). The final score is the average of 9 metric scores.