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

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

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, hindering fair and efficient cross-model comparison; (2) audio codecs, as a key component of audio foundation models, lack 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. To address the first issue, we introduce **UltraEval-Audio**, a unified evaluation framework for audio foundation models, specifically designed for both audio understanding and generation tasks. UltraEval-Audio features a modular architecture, supporting 10 languages and 14 core task categories, while seamlessly integrating 24 mainstream models and 36 authoritative benchmarks. To enhance research efficiency, the framework provides a one-command evaluation feature, accompanied by real-time public leaderboards. For the second challenge, UltraEval-Audio adopts a novel comprehensive evaluation scheme for audio codecs, evaluating performance across three key dimensions: semantic accuracy, timbre fidelity, and acoustic quality. To address the third issue, we propose two new Chinese benchmarks, *SpeechCMMLU* and *SpeechHSK*, designed to assess Chinese knowledge proficiency and language fluency. We wish that **UltraEval-Audio** will provide both academia and industry with a transparent, efficient, and fair platform for comparison of audio 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 rapid development of multimodal large models has ensued. Notably, the release of OpenAI's GPT-4o (OpenAI et al., 2024) marked the advent of the native audio interaction era, accelerating the explosive development of audio foundation models (AFMs). Subsequently, a series of models with complex understanding and generation capabilities emerged, including 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), significantly expanding the boundaries of human-computer interaction. With the rapid iteration of model capabilities, the challenge of objectively and systematically evaluating these models has become a focal point of academic interest. However, current audio evaluation lacks a unified framework, with datasets and code scattered across various sources, greatly hindering fair and efficient comparisons between models.

In addition to the issue of fragmented evaluation frameworks, the existing evaluation systems also exhibit significant limitations in terms of evaluation depth and language coverage. Traditional evaluation tools are

*Equal Contributions.
†Corresponding Authors.

often designed for specific tasks, such as automatic speech recognition (ASR) or automatic speech translation (AST), making it difficult to adapt them to audio foundation models (AFMs) with general-purpose interactive capabilities. This is particularly true in areas like prompt management and inference parameter tuning, both of which are critical for the comprehensive assessment of AFMs. As model capabilities rapidly evolve, user-centered speech benchmarks are emerging in addition to traditional ASR and AST tasks. However, these benchmarks still face the following core challenges:

(1) Audio codecs, which serve as the backbone of AFMs, lack systematic performance metrics. A codec consists of an audio tokenizer (which converts audio into discrete tokens) and a vocoder (which reconstructs audio from generated tokens). The design of audio codecs directly affects the fidelity and efficiency of the audio representation, which significantly impacts the overall performance of AFMs (Ye et al., 2025). Existing methods are broad and provide limited insight into specific performance dimensions.

(2) Current benchmarks are heavily reliant on English, making it challenging to objectively assess models' performance on Chinese. Mainstream benchmarks, such as SpeechTriviaQA (Défossez et al., 2024), SpeechWebQuestions (Nachmani et al., 2023), and SpeechAlpacaEval (Fang et al., 2025), are primarily English-centric, leading to inadequate measurement of models' knowledge and language proficiency in the Chinese context.

To address these challenges, we propose **UltraEval-Audio**, a unified evaluation framework for audio foundation models. By decoupling data loading, prompt management, inference parameter control, and diverse post-processing and aggregation methods, this framework provides researchers with a unified and flexible evaluation environment. Users can quickly initiate the evaluation process using a simple "one-click" automated script. This decoupled framework design not only enhances the reproducibility of experiments but also facilitates rapid adaptation and agile extension for researchers. Furthermore, UltraEval-Audio innovatively introduces an audio codec evaluation scheme and several Chinese-language evaluation benchmarks, filling the gaps from both model components and evaluation benchmarks. Through this full-stack integration, UltraEval-Audio aims to standardize and enhance the transparency the entire evaluation process. By providing real-time public audio leaderboards, it strives to advance the field of audio foundation models towards greater interpretability and fairness. Our contributions are summarized as follows:

- **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 deeply integrates 24 mainstream models and 36 authoritative benchmarks, covering three key areas: speech, environmental sound, and music. With its user-friendly design, the framework offers a "one-click" evaluation feature and publicly available leaderboards for transparent comparisons.
- **A multi-dimensional evaluation scheme for audio codecs**: We have established a systematic evaluation scheme that covers semantic accuracy, timbre fidelity, and acoustic quality, addressing the lack of widely accepted and systematic multi-dimensional performance metrics for audio codecs.
- **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.

2 Related Work

In this section, we introduce the latest developments in audio foundation models, audio evaluation frameworks, evaluation benchmarks, and evaluation of audio codecs.

The Advancements of Audio Foundation Models. Emerging audio foundation models typically adopt an **Audio Codec+LLM** architecture, which 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 contextual modeling and autoregressive token prediction; and (3) a **vocoder**, which synthesizes natural speech waveforms from the generated audio tokens. 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 and produce only text as output (e.g., Qwen-Audio (Chu et al., 2023, 2024), Gemini-1.5 (Team et al., 2024)). (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 (OpenAI et al., 2024), Moshi (Défossez et al., 2024), MiniCPM-o 2.6 (Yao et al., 2024), Qwen2.5-Omni (Xu et al., 2025), and Kimi-Audio (Ding et al., 2025)).

Meanwhile, audio codecs are also rapidly evolving. SoundStream (Zeghidour et al., 2021) is the first universal audio codec capable of handling diverse audio types. EnCodec (Défossez et al., 2022), DAC (descript-audio-codec) (Kumar et al., 2023), HiFi-Codec (Yang et al., 2023), X-codec (Ye et al., 2025), BigCodec (Xin et al., 2024), and BiCodec (Wang et al., 2025b) 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.

The Development of Audio Evaluation Frameworks. Many comprehensive frameworks have been proposed for evaluating textual and visual foundation models such as OpenCompass (Contributors, 2023), OpenAI Evals [3], UltraEval (He et al., 2024), and VLMEvalKit (Duan et al., 2024). However, a comprehensive evaluation framework for audio foundation models has been lacking.

Before the emergence of audio foundation models, audio models were usually designed for specific tasks, with their evaluation typically being ad hoc and often included alongside the model repository. Audio foundation models have demonstrated strong general capabilities across various tasks, making it increasingly necessary to develop a comprehensive evaluation framework that integrates multiple tasks. Several audio evaluation frameworks have been proposed. For instance, AudioBench (Chen et al., 2024) collects 8 distinct tasks and 26 benchmarks for evaluating audio foundation models. AHELM (Lee et al., 2025) aggregates various datasets to holistically measure the performance of AFMs across 10 aspects. But they lack 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.

The Development of Audio Evaluation Benchmarks. Beyond 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. 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. SpeechWebQuestions (Nachmani et al., 2023), SpeechTriviaQA (Défossez et al., 2024), and SpeechAlpacaEval (Fang et al., 2025) are derived from corresponding textual benchmarks WebQuestions (Chen et al., 2015), TriviaQA (Joshi et al., 2017), and 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 Audio Evaluation Design and Methodology

This section outlines a systematic audio evaluation design and methodology for audio foundation models. Section 3.1 introduces a unified audio evaluation taxonomy that organizes tasks and their corresponding benchmark instantiations. Building upon this taxonomy, Section 3.2 focuses on codec evaluation and develops a comprehensive methodology for assessing audio codecs, a core component of audio foundation model archi-

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

Table 1: Task taxonomy in UltraEval-Audio.

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

Notes: WER: Word Error Rate; CER: Character Error Rate; BLEU/ROUGE-L: text generation quality; 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.

tectures. Finally, Section 3.3 introduces two Chinese speech benchmarks, SpeechCMMLU and SpeechHSK, to address the lack of systematic evaluation resources for Chinese in existing audio evaluation research.

3.1 Audio Evaluation Taxonomy

Audio foundation models cover a wide range of modalities, tasks, languages and domains. Existing evaluation efforts typically focus on isolated tasks or specific modalities, underscoring the need for a unified and extensible evaluation framework. Rather than simply aggregating existing benchmarks, UltraEval-Audio introduces a reusable and extensible evaluation taxonomy that offers methodological guidance for future evaluation design.

Task taxonomy: What Capabilities Are Evaluated

UltraEval-Audio adopts a capability-driven task taxonomy to organize the evaluation of audio foundation models. Evaluation tasks are grouped into core categories that reflect distinct model capabilities and typical input–output settings. As summarized in Table 1, the taxonomy consists of three high-level categories: audio understanding, audio generation, and audio codec. Within each category, tasks are further organized by application domains, including speech, music, and environment sounds.

Audio understanding tasks evaluate a model's ability to extract and interpret semantic information from audio inputs. This category spans multiple levels of analysis across speech and non-speech domains. In the speech domain, tasks range from low-level recognition such as ASR and AST to higher-level semantic reasoning such as speech QA. In the non-speech domain, tasks include low-level classification like instrument recognition and higher-level comprehension such as audio captioning.

Audio generation tasks assess a model's ability to synthesize or transform audio under given conditions. In the speech domain, tasks include TTS, voice clone (VC), and spoken answer generation for speech QA tasks. In audio understanding tasks, Speech QA evaluates the accuracy of textual responses. In audio generation tasks, it additionally assesses the quality of generated audio responses in terms of acoustics, naturalness, and intelligibility.

Additionally, audio codec evaluation is treated as a separate task category. Although audio codecs are not traditional application tasks, they play an important role in audio foundation models. Audio codec tasks systematically assess a codec's ability to compress and reconstruct audio while preserving both semantic content and acoustic characteristics.

Each task category is associated with standardized evaluation metrics aligned with its target capabilities. For example, ASR uses word error rate (WER) or character error rate (CER); translation and captioning tasks rely on text similarity metrics such as BLEU and ROUGE; attribute and sound classification tasks use accuracy; and audio generation and codec tasks are evaluated with ASR-based WER or dedicated speech quality metrics.

By organizing tasks and metrics in this way, UltraEval-Audio provides a reusable and extensible evaluation taxonomy that offers methodological guidance for future evaluation design, ensuring consistent and comparable coverage of capabilities across audio domains and model types.

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

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 & 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)

Benchmark Instantiation: How Capabilities Are Measured

Building on the task taxonomies, UltraEval-Audio instantiates each evaluated capability through a carefully selected set of widely adopted benchmarks. As summarized in Table 2, the framework integrates 36 benchmarks covering 14 tasks and 10 languages. Each benchmark is explicitly mapped to a predefined task category, ensuring consistency between capability definitions and specific evaluation details.

Within each task, multiple benchmark datasets are aggregated to enable multidimensional and robust assessment of model capabilities. For the ASR task, we support over ten datasets. These datasets span clear read speech (e.g., LibriSpeech) to complex noisy scenarios (e.g., WenetSpeech), and cover single-language (e.g., AISHELL-1 for Chinese) as well as multilingual conditions (e.g., MLS, FLEURS). This design allows evaluation not only of absolute recognition accuracy (WER/CER) but also of models' generalization and robustness across accents, domains, noise levels, and languages. Similarly, for Speech QA, we incorporate multiple knowledge-based and instruction-based benchmarks, including SpeechTriviaQA and SpeechAlpacaEval.

Across all tasks, the selected benchmarks cover multiple languages and diverse acoustic domains, and are evaluated using metrics specified in the task taxonomy and tailored to each task. This instantiation strategy ensures that the evaluation framework is comprehensive, extensible, and tightly aligned with the underlying capability definitions, providing a clear foundation for systematic and reproducible assessment.

3.2 Codec Evaluation

Audio codecs are a fundamental component of audio foundation models, as their design directly affects the fidelity of audio representations and overall model performance. Existing evaluation approaches, however, rely on diverse metrics with inconsistent standards, and lack a systematic, comparable assessment framework. To address this, we propose a three-dimensional codec evaluation methodology encompassing **semantics**, **timbre fidelity**, and **acoustic quality**.

For semantics, we measure how well reconstructed audio preserves the original content using WER. Specifically, the reconstructed audio is transcribed by high-performance ASR models and compared to the original transcript. We employ Whisper-large-v3 for English and Paraformer-zh for Chinese. For timbre fidelity, we

extract audio embeddings using WavLM-large fine-tuned on speaker verification, and compute the cosine similarity between embeddings of the original and reconstructed audio. This quantifies the codec's ability to retain speaker characteristics. For acoustic quality, we assess the naturalness and perceptual comfort of audio using a combination of UTMOS (Saeki et al., 2022) to predict overall naturalness, alongside DNSMOS P.835 (Reddy et al., 2022) and P.808 (Reddy et al., 2021) to evaluate speech quality in noisy environments.

This three-dimensional evaluation framework provides complementary perspectives across content, speaker characteristics, and perceptual quality, enabling a more comprehensive and reliable characterization of codec performance and potential limitations, and offering a fine-grained diagnostic tool for model development and optimization.

3.3 Chinese Benchmarks

The development of audio foundation models shift speech evaluation from traditional low-level metrics like WER, toward higher-level capabilities including complex semantic understanding and knowledge reasoning. However, existing high-level speech benchmarks are largely focused on English, such as SpeechTriviaQA, SpeechWebQuestions, and SpeechAlpacaEval, leaving a relative scarcity of resources in the Chinese context. To address this gap, we introduce two new benchmarks: *SpeechCMMLU* evaluates models' ability to comprehend and reason over Chinese world-knowledge tasks, while *SpeechHSK* measures Chinese language proficiency.

SpeechCMMLU extends the widely recognized Chinese knowledge reasoning benchmark CMMLU (Li et al., 2023a) into the audio modality through systematic speech synthesis. The construction and quality assurance process is as follows:

1. **Text instruction construction**. Each multiple-choice question is formatted as a unified instruction, combining the question stem, options, and answer constraints into a standardized prompt.
2. **Audio instruction synthesis**. Considering the high cost of manual recording and the large scale of the dataset (11,583 items), we employ the high-quality TTS model CosyVoice2 to generate the speech prompts. This model demonstrates consistent performance in Chinese pronunciation and specialized terminology, making it suitable for large-scale speech dataset creation.
3. **Automated quality control**. To ensure semantic fidelity, all synthesized audio is transcribed via a ASR process. Only samples whose transcription exactly matches the original text (CER = 0%) are retained, effectively eliminating potential pronunciation errors or terminology deviations introduced during TTS synthesis.

Following this procedure, the released SpeechCMMLU dataset contains 3,519 high-quality speech samples spanning diverse academic subjects, suitable for evaluating models' Chinese knowledge comprehension and reasoning capabilities in professional domains.

SpeechHSK leverages the Hanyu Shuiping Kaoshi (HSK), the standardized global Chinese proficiency test established by the Chinese Ministry of Education, widely recognized for its authority and scientific rigor. SpeechHSK converts the listening comprehension portion of HSK into an audio benchmark, providing a hierarchical and practical measure of models' Chinese language skills.

- **Data sources and construction**. The benchmark draws from official HSK listening comprehension questions. Each sample consists of a question and four options, where the question audio uses the original exam recordings, while the options are originally provided in text form. To adapt textual options for speech evaluation, native Chinese speakers re-record them as audio, ensuring naturalness of the audio.
- **Difficulty-level structure**. The dataset follows HSK's six proficiency levels (Level 1: beginner; Level 6: advanced), with an increasing number of questions at higher levels, totaling 170 samples. This structure supports progressive, diagnostic assessment of models' Chinese language comprehension.

Together, SpeechCMMLU and SpeechHSK form a multi-level evaluation suite for Chinese speech. SpeechCMMLU focuses on knowledge reasoning, while SpeechHSK emphasizes general language proficiency. Both benchmarks are constructed via automated pipelines with rigorous quality control, ensuring reproducibility and scalability.

4 UltraEval-Audio Framework

This section presents the engineering architecture of UltraEval-Audio. To address the integration challenges in the evaluation of audio foundation models arising from diversity in input modalities and model interfaces, the

framework adopts a decoupled modular design together with a configuration-driven workflow. This design enables components to be developed independently, extended flexibly, and reused efficiently. All datasets, models, and evaluation pipelines are declaratively specified through unified YAML configuration files, enabling end-to-end evaluation without manual scripting.

As illustrated in Figure 1, the framework consists of three main modules: data preparation, model deployment, and evaluation. The following subsections provide a detailed description of each component. Section 4.1 details data loader and prompt management, section 4.2 presents the model deployment mechanisms, and section 4.3 describes the evaluation pipeline and automated scoring modules.

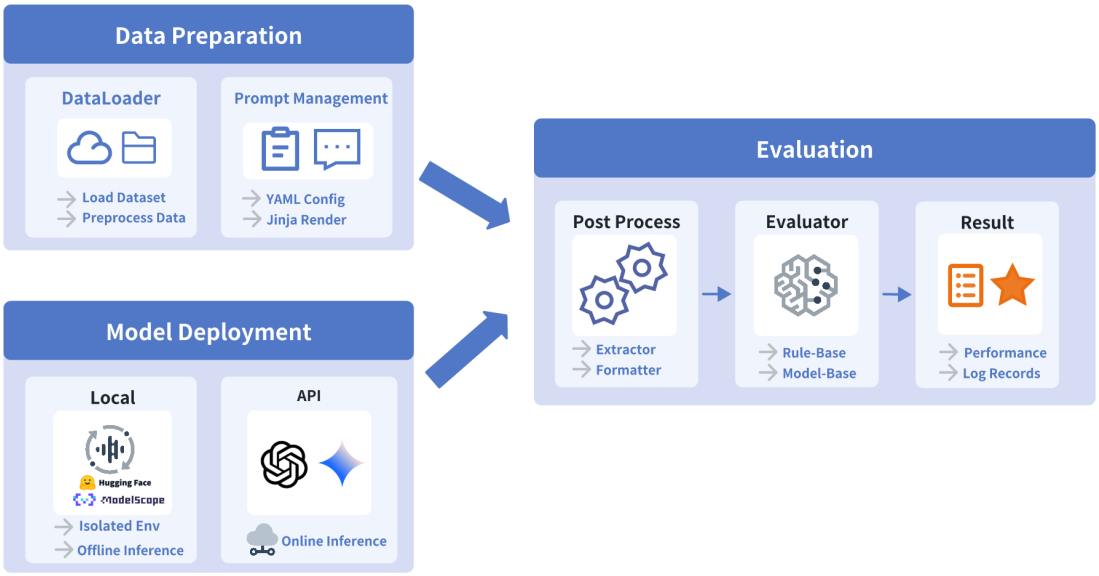

Figure 1: Overview of the UltraEval-Audio framework

4.1 Data Preparation

Data loader. Data loading and management form the foundational infrastructure of audio evaluation. Unlike textual datasets, which typically have simple structures, audio datasets involve a strong coupling between audio signals and textual annotations. This coupling introduces significant data management challenges, including managing dispersed file paths, handling diverse audio encoding formats, and normalizing inconsistent sampling rates.

To address these challenges, UltraEval-Audio adopts an audio-centric dataset organization scheme. In this design, a reserved field *audio* is used to store audio content. The framework provides built-in automated pipelines that dynamically manage audio resource acquisition during the loading, including automatic downloading and caching from open-source repositories such as HuggingFace, as well as decoding and format standardization.

The framework decouples complex multimodal preprocessing into a standardized data interface, significantly simplifying the preparation workflow and ensuring efficient execution of subsequent model inference and evaluation.

Prompt management. Prompts play a critical role in guiding models to produce relevant task outputs. Prompt design not only relate to the format of model input but also directly affects the stability and comparability of evaluation results. However, due to differences in training strategies, models are often highly sensitive to prompt, such as the use of special tokens or role definitions, and this sensitivity can vary substantially across tasks. The combination of model-specific sensitivity and task-dependent variation makes hard-coded prompts difficult to reuse and undermines fair comparison across models.

To address this issue, UltraEval-Audio introduces a configuration-driven prompt management mechanism. Prompt structures are defined using YAML-based configurations and combined with the Jinja templating engine to support variable injection and conditional logic. This design enables flexible prompt construction at the task, model, and even sample level. Consequently, researchers can define or replace prompt templates

for different tasks without modifying any code, greatly enhancing the maintainability and extensibility of the evaluation workflow.

In practice, the prompt management module in UltraEval-Audio comprises three components:

- **Prompt Registry**, which registers and indexes prompt templates associated with different tasks.
- **YAML Parser**, which interprets declarative configuration files and loads prompt definitions for specific models or tasks.
- **Template Renderer**, which leverages Jinja-based rendering to inject dataset fields (e.g., {{audio}}, {{question}}, {{choice_a}}) into templates and produce the final model inputs.

This mechanism supports not only fixed templates but also more complex prompt constructions involving dynamic variables and conditional branches. The following is an example of a MiniCPM-o 2.6 ASR prompt:

```
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}}'
```

the {{audio}} placeholder is dynamically replaced at inference time with the audio field from the corresponding data sample. To adapt the prompt for different ASR models, users only need to define a model-specific prompt and select it at runtime via the *--prompt $prompt* argument, without modifying any code.

For tasks that require sample-dependent prompt structures, such as multiple-choice questions with a variable number of options, the template can incorporate conditional rendering logic. An example is shown below:

```
single_choice_extended:
  class: audio_evals.prompt.base.Prompt
  args:
    template:
      - role: user
        contents:
          - type: audio
            value: "{{audio}}"
          - type: text
        value: "Choose the most suitable answer from options A, B, C, D{% if choice_e is defined
            and choice_e %}, and E{% endif %} to respond to the question below.
            You may only choose A, B, C, or D{% if choice_e is defined and choice_e %},
            or E{% endif %}.
              {{question}}
              A. {{choice_a}}
              B. {{choice_b}}
              C. {{choice_c}}
              D. {{choice_d}}{% if choice_e is defined and choice_e %}
              E. {{choice_e}}{% endif %}"
```

The prompt management mechanism achieves unified and automated evaluation through configuration-driven design, template reuse, and dynamic rendering. On the one hand, it fully decouples prompt design from the core evaluation logic, reducing integration overhead across tasks and models. On the other hand, it ensures that prompt definitions are reproducible, extensible, and easily shareable, laying a solid foundation for standardized evaluation of audio foundation models.

4.2 Model Deployment

Model deployment is a critical link between data and evaluation. Audio models are typically deployed in one of two forms: (1) Remote API models, which perform inference via official SDKs or HTTP API endpoints, and (2) Locally deployed models, which rely on local hardware resources for inference. Differences in environment dependencies, hardware requirements, and execution modes between these two types pose significant engineering challenges for a unified evaluation framework.

This challenge is especially pronounced for locally deployed models, where dependency conflicts are a common issue. Different models may require incompatible versions of deep learning frameworks, drivers, or audio processing libraries. For instance, audio quality evaluation modules (such as UTMOS, SIM, DNSMOS) and ASR-WER evaluators often depend on specific environment versions that may conflict with the dependency stacks of the models under evaluation. Mixing multiple models in the same runtime environment may lead to contamination or inference failures, compromising the reproducibility of evaluation results.

To address this challenge, UltraEval-Audio introduces an **Isolated Runtime** mechanism ensuring environment independence and safe execution at the system level. The core design features are as follows:

1. **Environment-level isolation**: Each evaluated model is allocated an independent virtual runtime environment with only the dependencies it requires. This setup is completely isolated from the main evaluation process, eliminating dependency conflicts at the source.
2. **Subprocess-based model execution**: Models run as independent subprocesses in a daemonized manner, maintaining a persistent loaded state to support continuous inference and significantly reduce the overhead associated with frequent model loading.
3. **Inter-process communication (IPC)**: The main evaluation process communicates with each model subprocess via system pipes, exchanging data and inference results securely and with low latency.

At the design level, this mechanism effectively implements a microservice-style inference architecture. Users do not need to manually manage or install additional dependencies, and the system significantly reduces the engineering complexity of large-scale, multi-model evaluation. Furthermore, to standardize the interface across different model types, the framework provides a unified *.inference()* method at the deployment layer, which accepts inputs from the prompt management module and returns the inference results.

Through this design, UltraEval-Audio achieves full decoupling and automated management of model execution at the system level. On one hand, the isolated runtime eliminates dependency conflicts and environment contamination between models, ensuring reproducibility and consistency of evaluation results. On the other hand, the unified inference interface abstracts away underlying model implementations, lowering the technical barrier for cross-model comparison and framework extension. Overall, this module provides a stable, scalable, and transparent runtime foundation for large-scale audio model evaluation.

4.3 Evaluation

Following model inference, UltraEval-Audio proceeds to the evaluation phase, in which raw model outputs are converted into structured and quantifiable results to enable objective performance assessment across models. The evaluation workflow comprises two core components: post-processing and evaluator. The post-processing module standardizes and semantically aligns the model outputs, while the evaluator computes metrics according to task. This two level structure enables a fully automated and standardized pipeline from output generation to metric calculation.

Post-processing. In multimodal audio evaluation, model outputs often include extraneous prefixes or suffixes, formatting artifacts, or unstructured text unrelated to the task objectives, which can compromise accurate metric computation. To ensure consistent and well-structured inputs for evaluation, UltraEval-Audio employs a flexible post-processing mechanism that parses and standardizes outputs at multiple levels.

This mechanism adopts a modular and composable design, enabling adaptive workflows based on task specifications. The framework includes several built-in post-processing modules—such as *Option Extraction*, *Yes/No Parser*, and *JSON Field Parser*—which can be sequentially combined to form multi-step pipelines for handling complex tasks.

Evaluator. The evaluator module processes model outputs that are standardized via post-processing and is responsible for computing the final performance metrics. UltraEval-Audio categorizes evaluators into two types:

Table 3: Overview of audio foundation models participating in the evaluation

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

- **Rule-based Evaluators**: For tasks with reference answers, classical algorithmic metrics are applied. The framework includes WER for ASR tasks, BLEU for AST, accuracy for classification tasks as well as other similar measures.
- **Model-based Evaluators**: For audio generation quality or open-ended generative tasks, pretrained evaluation models are employed to approximate human judgment. This includes Speaker Similarity (SIM) for assessing timbre or voice cloning fidelity, UTMOS and DNSMOS for evaluating naturalness and perceptual quality of speech, and GPT-based open-domain QA scorers (LLM-as-a-Judge).

All evaluators are registered and orchestrated through a unified interface, with their outputs automatically aggregated into a consolidated results summary. This design ensures compatibility across metrics from different tasks while facilitating rapid integration of new evaluators, rendering the framework highly modular and maintainable.

By employing standardized post-processing and a multidimensional evaluation framework, UltraEval-Audio establishes an end-to-end automated pipeline from model outputs to quantifiable metrics. This design enhances consistency, transparency, and reproducibility, providing a systematic foundation for objective comparison and ongoing development of multimodal audio foundation models.

5 Evaluation Results

UltraEval-Audio provides a unified solution for systematically assessing the performance of audio-processing models in various testing environments. In this section, we carefully select representative evaluation tasks and build three leaderboards: audio understanding, audio generation, and audio codec. Then we evaluate 13 leading audio foundation models and 9 audio codecs, introduced in Section 5.1. We present the leaderboards of audio understanding, audio generation, and audio codec in Section 5.2, 5.3 and 5.4 respectively.

5.1 Evaluated Models

We evaluate emerging audio foundation models and audio codecs to build leaderboards. The evaluated audio foundation models are shown in Table 3, including leading proprietary models such as GPT-4o-Realtime and Gemini-2.5-Pro, as well as the latest open-source models like Qwen3-Omni-30B-A3B-Instruct and Kimi-Audio-7B-Instruct. Audio codecs include Encodec (Défossez et al., 2022), ChatTTS-DVAE[4], the Mimi (Défossez et al., 2024) family, WavTokenizer-large-speech-75token (denoted as WavTokenizer-large-75)[5] , WavTokenizer-large-unify-40token (denoted as WavTokenizer-large-40)[6] (Ji et al., 2024) and Spark (Wang et al., 2025b).

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

Several models design different prompts and parameters for specific tasks, for example, Kimi-Audio-7B-Instruct has specific prompts for ASR tasks. To replicate these results, we run the tasks using the prompts and parameters provided by the publisher. For tasks where the publisher has not provided these configurations, we use the official inference parameters and prompts for evaluation. Furthermore, we do not perform additional prompt or parameter optimizations, ensuring a fair and consistent evaluation protocol.

5.2 Audio Understanding

Table 4: 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.

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

Notes: WER/CER values can be greater than 100 when the total number of recognition errors exceeds the number of reference words/characters.

For audio understanding, we select well-established and widely cited benchmarks that are commonly used in existing papers. Specifically, we use **LibriSpeech** (en), **TED-LIUM** (en), **Common Voice 15** (en/zh), **AISHELL-1** (zh), **FLEURS** (zh), **WenetSpeech-test-net** (zh) for ASR, **covost2-en2zh**, **covost2-zh2en** for AST, and **MELD** for emotion recognition (EMO). The final results are shown in Table 4, from which we can make 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, Kimi-Audio-7B-Instruct, MiniCPM-o 2.6 as well as proprietary models like Gemini-2.5-Pro, achieving superior performance in the evaluation. A key reason for this is GPT-4o-Realtime's relatively underwhelming performance on Chinese ASR benchmarks, especially when evaluated on datasets such as **Wenet-test-net** and **Common Voice 15** (CV-15).

(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 improvement room in the latter area.

(3) For models from Gemini family, we observe that each generation's Flash model performs weaker than the Pro model. However, compared to Gemini 1.5, the Flash model of Gemini 2.5 performs better, while the Pro models show similar performance.

5.3 Audio Generation

To build the audio generation leaderboard, we select **SpeechWebQuestions**, **SpeechTriviaQA**, and **SpeechAlpacaEval** for English capability evaluation. Our self-proposed benchmarks **SpeechCMMLU** and **SpeechHSK** are also used to assess Chinese audio generation capabilities. We first evaluate the models' performance on each dataset by answer accuracy (Acc.), then assess the acoustic quality of the generated audio.

To ensure better alignment with existing evaluation results, we adopt different evaluation settings for different datasets. Specifically:

Table 5: 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.

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

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.

- For **SpeechWebQuestions** and **SpeechTriviaQA**, we follow the evaluation approach in (Nachmani et al., 2023) and (Défossez et al., 2024). 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 **SpeechAlpacaEval**, 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 further evaluate acoustic values of the generated audio responses from the aforementioned benchmarks, employing UTMOS, DNSMOS P.835, and DNSMOS P.808 as metrics. The performance of all models is summarized in Table 5, with key findings as follows:

(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.

5.4 Audio Codec

We evaluate audio codecs using clean speech corpora, including **LibriSpeech-dev-clean** (en), **LibriSpeech-test-clean** (en), and **AISHELL-1** (zh). For each dataset, we use the ASR-WER/CER metric to measure the codec accuracy, SIM (short for similarity) to measure timbre fidelity and acoustic scores to measure the acoustic quality. The results are shown in Table 6, from which we observe the following:

(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 codec accuracy and timbre fidelity, indicating that its token representation effectively captures both linguistic and timbral information from raw audio. However, its acoustic scores lag behind Spark and WavTokenizer-large-75, suggesting that improvements could be made to its decoder.

(3) We further propose Figure 2 to compare Mimi codec variants. Comparing Mimi (default 32-bit) with Mimi (8bit), the performance drop is most pronounced in timbre fidelity, while ASR-WER 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 acoustic quality.

Table 6: The audio codec leaderboard. 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			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**

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 $100 - \text{val}$. 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.

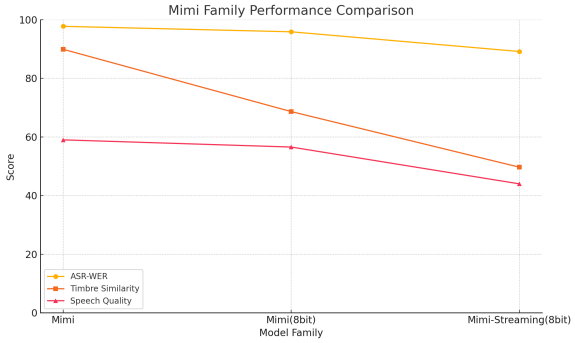

Figure 2: Mimi family performance comparison. Note that ASR-WER is normalized as $(100 - \text{WER})$, and speech quality scores (acoustic scores) are scaled by a factor of 20 for visualization.

(4) ChatTTS-DVAE, WavTokenizer-large-40, and Mimi-Streaming (8bit) underperform on the AISHELL-1 dataset, indicating a need for improved handling of Chinese-language audio.

6 Conclusion

In this paper, we introduce UltraEval-Audio, the first unified evaluation framework for comprehensive assessment of audio foundation models. We first construct a complete audio evaluation taxonomy encompassing tasks and benchmarks. To enrich the evaluation, we develop a systematic methodology for assessing audio codecs and introduce two Chinese benchmarks: SpeechCMMLU and SpeechHSK. 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, imporve 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, et al. Seed-tts: A family of high-quality versatile speech generation models. *arXiv preprint arXiv:2406.02430*, 2024.

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.

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

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.

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.

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

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.

Chandan KA Reddy, Vishak Gopal, and Ross Cutler. 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)*, pp. 6493–6497. IEEE, 2021.

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.

Gemini Team, Petko Georgiev, Ving Ian Lei, and ... others. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. URL https://arxiv.org/abs/2403.05530.

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.

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

Xinsheng Wang, Mingqi Jiang, Ziyang Ma, Ziyu Zhang, Songxiang Liu, Linqin Li, Zheng Liang, Qixi Zheng, Rui Wang, Xiaoqin Feng, Weizhen Bian, Zhen Ye, Sitong Cheng, Ruibin Yuan, Zhixian Zhao, Xinfa Zhu, Jiahao Pan, Liumeng Xue, Pengcheng Zhu, Yunlin Chen, Zhifei Li, Xie Chen, Lei Xie, Yike Guo, and Wei Xue. Spark-tts: An efficient llm-based text-to-speech model with single-stream decoupled speech tokens, 2025b. URL https://arxiv.org/abs/2503.01710.

Detai Xin, Xu Tan, Shinnosuke Takamichi, and Hiroshi Saruwatari. Bigcodec: Pushing the limits of low-bitrate neural speech codec. *arXiv preprint arXiv:2409.05377*, 2024.

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.

Dongchao Yang, Songxiang Liu, Rongjie Huang, Jinchuan Tian, Chao Weng, and Yuexian Zou. Hifi-codec: Group-residual vector quantization for high fidelity audio codec. *arXiv preprint arXiv:2305.02765*, 2023.

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.

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.

Neil Zeghidour, Alejandro Luebs, Ahmed Omran, Jan Skoglund, and Marco Tagliasacchi. Soundstream: An end-to-end neural audio codec. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 30: 495–507, 2021.

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.