

Fun-Audio-Chat Technical Report

Tongyi Fun Team, Alibaba Group

Abstract

Recent advancements in joint speech-text models have demonstrated great potential for seamless voice interactions. However, existing models face critical challenges: the temporal resolution mismatch between speech tokens (typically 25Hz) and text tokens (approximately 3Hz) dilutes semantic information, incurs high computational costs limit practical deployment, and leads to catastrophic forgetting of text LLM knowledge during multimodal training. In this work, we introduce **Fun-Audio-Chat**, a Large Audio Language Model (LALM) that addresses these limitations by adopting two key innovations from our previous work DrVoice. First, we employ the **Dual-Resolution Speech Representations (DRSR)** architecture: the Shared LLM backbone processes audio at an efficient 5Hz frame rate (achieved through speech token grouping), while the Speech Refined Head (SRH) generates high-quality speech tokens at 25Hz resolution. This dual-resolution design effectively balances computational efficiency (reducing GPU hours by nearly 50%) and speech generation quality. Second, we adopt the **Core-Cocktail Training strategy** in full supervised fine-tuning, a two-stage training approach with intermediate model merging that mitigates catastrophic forgetting. After Core-Cocktail training, we introduce **Multi-Task DPO Training** to enhance robustness, audio understanding, instruction-following and voice empathy capabilities. This multi-stage post-training paradigm enables Fun-Audio-Chat to effectively retain knowledge of the original text LLM while gaining powerful audio understanding, reasoning, and generation skills. Different from the majority of recent LALMs that rely on both large-scale audio-text pre-training and post-training to develop audio capabilities, **Fun-Audio-Chat only leverages pre-trained models and utilizes extensive post-training**. Fun-Audio-Chat dense 8B and MoE 30B-A3B models achieve competitive performance on Speech-to-Text and Speech-to-Speech generation tasks, ranking Top among models of similar scales across multiple Spoken Question Answering benchmarks. It also achieves competitive to superior performance on Audio Understanding, Speech Function Calling, Speech Instruction-Following and Voice Empathy benchmarks. We further develop **Fun-Audio-Chat-Duplex**, a full-duplex variant that achieves strong performance on Spoken Question Answering benchmarks and full-duplex interactions. We open-source the Fun-Audio-Chat-8B model checkpoint with its training and inference code, and provide an interactive demo.

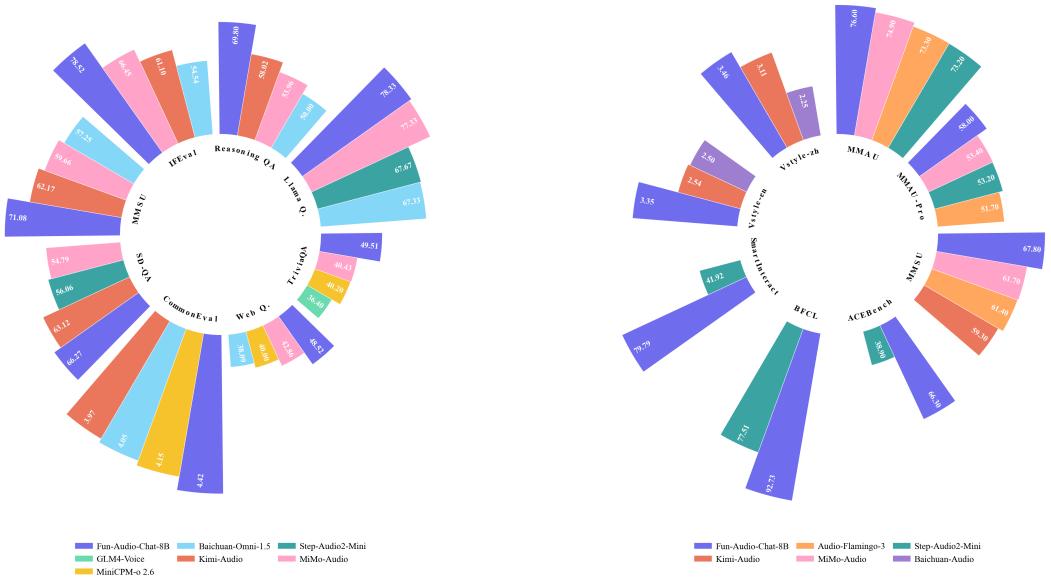


- | | |
|--------------------|-------------------------------------------------------------------------------------------------------------------------|
| GitHub | https://github.com/FunAudioLLM/Fun-Audio-Chat |
| HuggingFace | https://huggingface.co/FunAudioLLM/Fun-Audio-Chat-8B |
| ModelScope | https://modelscope.cn/FunAudioLLM/Fun-Audio-Chat-8B |
| Demo Page | https://funaudiollm.github.io/funaudiochat |

1 Introduction

The development of spoken dialogue systems is critical to human-computer interaction, as natural human communication inherently relies on verbal exchanges. Recently, Large Language Model (LLM) based spoken dialogue systems, exemplified by systems like GPT-4o (OpenAI, 2024b), demonstrate great

✉ FunAudioLLM@list.alibaba-inc.com



(a) Performance comparison on Spoken QA tasks.

(b) Performance comparison on other tasks.

Figure 1: Performance comparison between our Fun-Audio-Chat-8B and previous ~8B-scale state-of-the-art (SOTA) models across multiple benchmarks. (a) illustrates results on Spoken Question Answering benchmarks (Speech-to-Speech SQA on LlamaQ, TriviaQ, WebQ in UltraEvalAudio; Speech-to-Text SQA on ReasoningQA in OpenAudioBench; Speech-to-Text SQA on CommonEval, SD-QA, MMSU, and IFEval in VoiceBench), while (b) presents results on Audio Understanding (MMAU, MMAU-Pro, MMSU), Speech Function Calling (Speech-ACEBench, Speech-BFCL, Speech-SmartInteract), Speech Instruction-Following and Voice Empathy (VStyle, English and Mandarin subsets) benchmarks. Detailed evaluations are presented in Section 3.

potential for seamless and natural voice interactions with users. LLM-based spoken dialogue systems can be generally categorized into **cascaded** and **end-to-end (E2E)** systems, with the distinction lying in whether the backbone LLM can *directly* comprehend speech representations and generate speech outputs.

Many recent E2E models focus on **Joint Speech-Text Models** (Défossez et al., 2024; Chen et al., 2024a; KimiTeam et al., 2025), where LLMs take speech representations as input and generate *both text tokens and speech tokens simultaneously*. However, existing joint speech-text models face critical challenges: (1) the temporal resolution mismatch between speech tokens (typically 25Hz) and text tokens (approximately 3Hz) (Chen et al., 2024a) dilutes semantic information and hinders the full utilization of the LLM’s core capabilities; (2) continual pre-training and post-training the text-LLM backbone into multimodal models often lead to catastrophic forgetting of the text LLM’s knowledge; (3) high computational costs due to high audio frame rates (typically 12.5Hz or 25Hz), limiting practical deployment.

In this work, we present **Fun-Audio-Chat**, a parallel large audio language model (LALM) that extends our previous work DrVoice (Tan et al., 2025) by adopting the **Dual-Resolution Speech Representations (DRSR)** architecture and scaling it up to significantly larger training datasets of millions of hours of diverse audio data and larger model scales (dense 8B and MoE 30B-A3B¹).

For speech comprehension in Fun-Audio-Chat, we employ a grouping mechanism that maps 25Hz audio tokens to 5Hz speech representations, enabling the shared LLM backbone to process audio at an efficient 5Hz frame rate. During generation, the hidden states from the shared LLM layer are passed in parallel to

¹30B-A3B denotes a Mixture-of-Experts (MoE) model with 30B total parameters and 3B active parameters.

a Text Head for text token prediction and a Speech Refined Head (SRH) to generate high-quality speech tokens at 25Hz resolution. This dual-resolution design effectively balances computational efficiency (reducing GPU hours by nearly 50%) and speech generation quality.

The majority of recent open-source LALMs and Omni-language-models rely on both large-scale audio-text pre-training (e.g., audio/text unimodal pre-training, audio-text mapping and interleaving pre-training tasks) and post-training to develop strong audio capabilities, such as Kimi-Audio ([KimiTeam et al., 2025](#)), Step-Audio 2 ([Wu et al., 2025](#)), MiMo-Audio ([Xiaomi, 2025](#)), and Longcat-Flash-Omni ([Team, 2025b](#)). In contrast, Fun-Audio-Chat leverages pre-trained models and is trained with a multi-stage post-training paradigm, without large-scale audio-text pre-training (similarly, Audio-Flamingo-3 ([Goel et al., 2025](#)) also does not use large-scale audio-text pre-training). After initialization from text-based or vision-language LLMs, the **Pre-alignment** stage updates the audio encoder, the adapter, and the Speech Refined Head using large-scale speech-text paired data. We then adopt the **Core-Cocktail Training strategy** proposed in our earlier work DrVoice ([Tan et al., 2025](#)), to address catastrophic forgetting in multimodal training. Core-Cocktail Training is a two-stage approach that involves: (1) Stage 1: fine-tuning with high learning rate to rapidly adapt the model, (2) intermediate model merging of the Stage-1 model and the original pre-trained LLM backbone to preserve knowledge, and (3) Stage 2: fine-tuning with low learning rate for stable optimization. Following Core-Cocktail Training, we conduct **Multi-Task DPO Training** to boost the robustness to real speech data and the capabilities of speech instruction-following, audio understanding, and voice empathy. This multi-stage post-training paradigm enables Fun-Audio-Chat to retain the original text-LLM’s capabilities while gaining powerful audio understanding, reasoning, and generation skills.

Our contributions can be summarized as follows:

- **Large-Scale Post-Training and Model Scaling.** Fun-Audio-Chat scales up the two key innovations of **Dual-Resolution Speech Representations (DRSR)** architecture and **Core-Cocktail Training strategy** in our earlier work DrVoice ([Tan et al., 2025](#)) to significantly larger data scales of millions of hours of diverse audio data and larger model scales, including **dense 8B and MoE 30B-A3B parameters**. This work verifies that the two key innovations in DrVoice demonstrate excellent scalability: DRSR, with its efficient 5Hz processing for the backbone LLM and 25Hz generation head, retains high computational efficiency (**approximately 50% reduction in training GPU hours**) at larger scales; and Core-Cocktail Training strategy, with its two-stage training using different learning rates and intermediate model merging, effectively mitigates catastrophic forgetting in both 8B and 30B-A3B models. The large-scale post-training enables Fun-Audio-Chat to achieve superior performance across multiple benchmarks while maintaining the computational efficiency advantages from the dual-resolution design.
- **Multi-Task DPO Training for Enhancing Robustness and Generalizability.** Following Core-Cocktail Training, we introduce Multi-Task DPO Training to enhance the capabilities of Fun-Audio-Chat in multiple dimensions: robustness to real speech data, capabilities of instruction-following, audio understanding, and voice empathy. This training approach enables the model to better align with human preferences and improve performance on real-world conversational scenarios, distinguishing Fun-Audio-Chat from previous works that primarily rely on supervised fine-tuning. Through Multi-Task DPO training, Fun-Audio-Chat acquires advanced capabilities beyond basic speech-text interaction, including speech function calling, speech instruction-following, and voice empathy (recognizing and reasoning over user’s emotional states and generating empathetic responses), enabling the model to understand and respond to complex voice interactions with appropriate emotional intelligence and functional execution.
- **Comprehensive Evaluation and Strong Performance.** Extensive evaluations demonstrate that Fun-Audio-Chat 8B and 30B-A3B achieve superior performance on Spoken Question Answering (on both Speech-to-Text and Speech-to-Speech generation tasks), ranking top among models of similar scales. It also demonstrates competitive capabilities of Audio Understanding, Speech Function Calling, Speech Instruction-Following, and Voice Empathy, as demonstrated across a wide variety of commonly

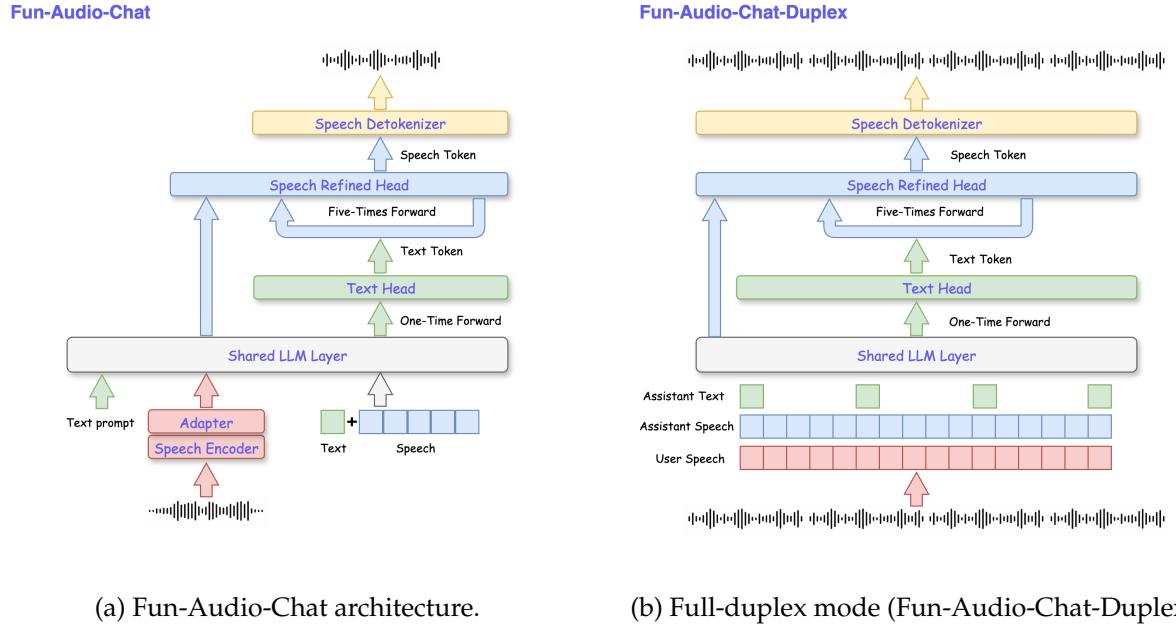


Figure 2: Overview of Fun-Audio-Chat. (a) User speech inputs are tokenized, grouped, and encoded by the MLLM for autoregressive text token prediction by a **Text Head** and speech token prediction by a **Speech Refined Head (SRH)**. The MLLM comprises the **Shared LLM Layer**, the Text Head, and SRH. The generated speech tokens are then converted to speech waveform by the speech detokenizer. Note that SRH generates 5 speech tokens through 5 autoregressive forward passes, where 5 is the grouping factor. (b) Full-duplex communication mode of Fun-Audio-Chat.

used benchmarks including OpenAudioBench², VoiceBench (Chen et al., 2024b), UltraEval-Audio³, MMAU (Sakshi et al., 2025), MMAU-Pro (Kumar et al., 2025), MMSU (Wang et al., 2025a), multiple speech function calling benchmarks, and VStyle (Zhan et al., 2025). Detailed evaluation results are presented in Section 3.

- **Full-Duplex Voice Interaction.** We extend Fun-Audio-Chat to a full-duplex variant, **Fun-Audio-Chat-Duplex**, which supports simultaneous two-way communications. This model achieves competitive performance on Spoken Question Answering benchmarks, suggesting excellent intelligence, and strong performance in full-duplex interaction metrics (Section 3), demonstrating superior capabilities in natural conversation and turn-taking.
- **Open-source Contribution and Interactive Demo.** To promote research in this field, we open-source the dense Fun-Audio-Chat-8B model, making the model checkpoint and the training and inference code publicly available so that researchers can build upon our work. Additionally, we provide an interactive demo that showcases Fun-Audio-Chat’s voice conversation capabilities.

2 Methodology

Figure 2 provides an architectural overview of Fun-Audio-Chat and its full-duplex variant Fun-Audio-Chat-Duplex. The framework of Fun-Audio-Chat comprises three primary modules: (1) For audio inputs, Speech Encoder and Speech Tokenizer transform raw audio waveforms into structured representations for both User and Assistant sides; (2) a Multimodal Large Language Model (MLLM) integrates a Shared LLM backbone with specialized Text Head and Speech Refined Head (SRH) components for generating tokens; and (3) the Speech Detokenizer reconstructs audio waveforms from the generated speech tokens.

²<https://huggingface.co/datasets/baichuan-inc/OpenAudioBench>

³<https://github.com/OpenBMB/UltraEval-Audio>

The architecture facilitates unified audio-text encoding and synchronized speech-text generation. At inference time, either text or audio inputs are converted into a common semantic representation space, which the MLLM processes to simultaneously generate both speech and text outputs via the SRH and the Text Head.

2.1 Speech Tokenization and Detokenization

To achieve robust audio comprehension, Fun-Audio-Chat employs Whisper-Large-v3 (Radford et al., 2022) as the **Speech Encoder** to derive continuous representations from user audio inputs. An **Adapter** module is then applied to reduce the temporal resolution of these features and match their dimensionality to the LLM’s hidden space. Given the demonstrated effectiveness of semantic tokens for speech representations (Zhang et al., 2023a; Borsos et al., 2023), particularly their strong correspondence with textual content (Zhang et al., 2023b), we adopt S3Tokenizer (Du et al., 2024a,b; 2025) as the **Speech Tokenizer** to transform audio waveforms into discrete semantic token sequences $\mathbf{S} = [s_0, s_1, \dots, s_{T-1}]$ (T denotes the sequence length) for the assistant’s output. In the reverse process, the Speech Detokenizer leverages speaker-specific embeddings that encode acoustic characteristics like timbre. The Flow Matching model (Lipman et al., 2023) generates Mel-spectrogram representations from these tokens, which are then converted back to audio waveforms using the HiFi-GAN vocoder (Kong et al., 2020).

2.2 Dual-Resolution Speech Representations (DRSR)

To maintain the text capabilities of pretrained text LLMs while supporting cross-modal functionality, Fun-Audio-Chat adopts the **Dual-Resolution Speech Representations (DRSR)** architecture from our earlier work DrVoice (Tan et al., 2025). This architecture effectively addresses the temporal resolution mismatch between speech tokens (typically 25Hz) and text tokens (approximately 3Hz), improves computational efficiency, and achieves high-quality speech generation.

Speech Token Grouping. To bridge the temporal resolution discrepancy, we apply a **grouping** technique from DrVoice (Tan et al., 2025) that reduces 25Hz speech tokens to 5Hz representations for the Shared LLM backbone. The grouping transformation is expressed as follows:

$$\mathbf{g}_i = \text{Linear} \left(\text{Concat}_{j=ik}^{(i+1)k-1} (\mathbf{s}_j) \right) \in \mathbb{R}^{d_{\text{text}}} \quad (1)$$

where \mathbf{s}_j represents individual speech tokens, Concat indicates concatenation, and $k = 5$ is the grouping factor based on the ratio of speech token frequency (25Hz) to the desired LLM processing frequency (5Hz). This mechanism reduces sequence length from T to T/k , allowing the Shared LLM to operate at a 5Hz frame rate, which substantially reduces computational overhead (yielding approximately 50% reduction in training GPU hours) while retaining the semantic reasoning abilities of the LLM.

Speech Refined Head (SRH). Although grouping facilitates efficient processing, it sacrifices fine-grained acoustic information essential for natural speech synthesis. To compensate this limitation, Fun-Audio-Chat integrates a specialized **Speech Refined Head (SRH)** that generates speech tokens at the complete 25Hz resolution. The SRH executes an **ungrouping** operation: the final hidden state from the Shared LLM, $\mathbf{h}_L^{\text{[SLLM]}}$, is initially transformed into group-sized embeddings through linear projection:

$$\mathbf{h}_{ug} = \mathbf{W}_p \mathbf{h}_L^{\text{[SLLM]}} \quad \text{where} \quad \mathbf{W}_p \in \mathbb{R}^{d_g \times d_h}, \quad (2)$$

which is followed by decomposition into k segments:

$$\mathbf{H} = \text{Split}_k(\mathbf{h}_{ug}) = [\mathbf{h}_{ug}^{(1)}, \mathbf{h}_{ug}^{(2)}, \dots, \mathbf{h}_{ug}^{(k)}], \quad (3)$$

where $\mathbf{h}_{ug}^{(i)} \in \mathbb{R}^{d_{ug}/k}$. The resulting \mathbf{H} provides conditional context for SRH, which generates speech tokens autoregressively at 25Hz. The training objective optimizes speech token prediction:

$$\mathcal{L}_{\text{SRH}} = - \sum_{i=1}^T \log P(s_i | s_{<i}, \mathbf{H}_{<i}), \quad (4)$$

where s_i denotes the i -th speech token. This dual-resolution framework allows Fun-Audio-Chat to simultaneously achieve computational efficiency (5Hz processing in the Shared LLM Layer) and high-fidelity speech synthesis (25Hz generation through SRH), following the design principles established in DrVoice (Tan et al., 2025).

2.3 Multimodal Large Language Model (MLLM)

The MLLM architecture extends pretrained text-LLMs to support *unified audio-text processing*, enabling the model to handle either speech or text inputs and generate *simultaneous* speech and text outputs.

Fun-Audio-Chat is a **Parallel Joint Speech-Text Model**. Following the approach in Moshi (Défossez et al., 2024), we integrate explicit text streams to provide semantic guidance for speech generation. Our design concentrates modality alignment **solely on the assistant side**, reflecting the inherent **asymmetry** in human-computer dialogue: *Users typically provide single-modality inputs (text or speech), while assistants can deliver coordinated multimodal responses (that is, joint speech-text response or text-only response)*.

The model exploits the autoregressive nature of LLMs by iteratively incorporating both speech tokens s_t and text tokens t_t into the Shared LLM Layer at each step. These token embeddings are combined through addition to create a unified input representation. The composite embedding c_t at step t is formulated as:

$$c_t = E_{\text{speech}}(s_t) + E_{\text{text}}(t_t) \quad (5)$$

where E_{speech} and E_{text} represent the embedding functions for speech and text tokens, respectively. To handle the length mismatch between speech and text sequences, we pad the shorter sequence with a special silence token <|SIL|> for each utterance.

The generation follows an autoregressive pattern:

$$P(y_t | y_{<t}, x) = \prod_{i=1}^t P(y_i | y_{<i}, x) \quad (6)$$

where x denotes the input and $y_t = (s_t, t_t)$ represents the combined speech-text output at step t . This formulation **unifies speech and text generation within one autoregressive process**.

2.4 Post-Training

Fun-Audio-Chat leverages existing pre-trained models and is trained with a multi-stage post-training pipeline, utilizing millions of hours of diverse speech data that encompasses diverse domains and tasks, including conversational and multilingual speech, audio for understanding tasks, ensuring comprehensive coverage of various scenarios and use cases. The training data combines open-source data following the training setup of DrVoice (Tan et al., 2025) and Audio-Flamingo-3 (Goel et al., 2025), along with in-house text, ASR, TTS, audio understanding, speech instruction-following, and voice empathy data.

The multi-stage training pipeline includes: (1) **Pre-alignment** uses large-scale speech-text paired data for aligning the Speech Encoder, the Adapter, and the Speech Refined Head; (2) **Core-Cocktail Training**, for supervised full fine-tuning, employs high quality speech data synthesized from billions of text tokens using CosyVoice 3 (Du et al., 2025) and selected by thresholding on synthesis Word Error Rate (WER); (3) **Multi-Task DPO Training** employs diverse real speech data for robustness enhancement, audio understanding and ASR data for comprehension capabilities, instruction-following data (including emotion, style, and prosody control) for speech instruction-following capabilities, and voice empathy data for emotion understanding and empathetic response generation capabilities. This training pipeline is carefully designed to progressively enhance the model’s audio comprehension, reasoning, and generation capabilities, while retaining the text capabilities of the backbone LLM.

Pre-Alignment. The training process begins with proper initialization of model components. The Speech Encoder is initialized with the weights of Whisper-Large-v3 (Radford et al., 2022; Xu et al.,

2025), providing robust voice understanding capabilities. The Shared LLM Layer is initialized using Qwen3-30B-A3B (Yang et al., 2025) or alternatively from Vision-language base models Qwen3-VL-8B (Bai et al., 2025), leveraging the strong semantic understanding capabilities of the pre-trained text LLMs. The pre-trained Speech Tokenizer and Detokenizer from CosyVoice 3 (Du et al., 2025) are employed and kept frozen throughout the entire training process of Fun-Audio-Chat. To establish effective alignment between audio and text modalities, we perform **Pre-alignment** training using large-scale speech-text pair data to align the Speech Encoder, the Adapter, and the Speech Refined Head before the main training stages. During this pre-alignment stage, the Shared LLM Layer is kept frozen to preserve its pre-trained capabilities.

Core-Cocktail Training. We find that multimodal model training faces a fundamental learning rate trade-off: high learning rates risk degrading the MLLM performance and exacerbating catastrophic forgetting of the base text-LLM’s knowledge, while low learning rates cause slow convergence and training stagnation. To address this optimization dilemma and prevent knowledge loss, we utilize the **Core-Cocktail Training** methodology introduced in our earlier work DrVoice (Tan et al., 2025), which employs a two-phase training procedure.

Stage 1: Fine-tuning with High Learning Rate. In this initial phase, we perform full fine-tuning on all MLLM parameters, the Audio Encoder, and Adapter using an elevated learning rate. For Fun-Audio-Chat, the learning rate is decayed from 1×10^{-4} to 1×10^{-5} in Stage 1, utilizing a cosine annealing schedule. This stage aims to quickly shift model parameters toward regions of the loss surface that are more conducive to multimodal learning, facilitating rapid task adaptation.

Intermediate Model Merging. To mitigate potential MLLM degradation from the intensive Stage 1 training phase, we implement an intermediate model merging operation. Following Xiao et al. (2024), we combine the Stage-1-trained MLLM parameters (M_1) with those of the original pretrained LLM (M_0) through weighted interpolation, producing a merged model M_r :

$$M_r \leftarrow \alpha M_1 + (1 - \alpha) M_0 \quad (7)$$

where α controls the interpolation balance. This merging operation reintroduces the foundational knowledge from the base LLM, safeguarding the original text understanding capabilities. Lower α values favor stronger retention of the base LLM’s knowledge. In our implementation, α is set to 0.5.

Stage 2: Refinement with Low Learning Rate. Stage 2 applies full fine-tuning to the merged model M_r with a reduced learning rate. For Fun-Audio-Chat, the learning rate is decayed from 1×10^{-5} to 1×10^{-6} in Stage 2, also utilizing a cosine annealing schedule. This enables stable, precise optimization that improves model performance without the instability associated with high learning rates. The Core-Cocktail Training strategy successfully reconciles fast adaptation with knowledge retention, substantially mitigating catastrophic forgetting while promoting effective multimodal learning. Fun-Audio-Chat supports a maximum context length of 2048 tokens (approximately 6 minutes of speech), sufficiently facilitating typical conversational interactions.

Multi-Task DPO Training. Following the Core-Cocktail training, we further conduct **Multi-Task DPO Training** (Rafailov et al., 2023) to enhance the model’s robustness to real speech data, audio understanding abilities, speech instruction-following and voice empathy capabilities. The Multi-Task DPO Training stage incorporates multiple preference learning objectives: (1) **robustness preference**: preferring responses that maintain quality under noisy or diverse speech inputs; (2) **instruction-following preference**: preferring responses that accurately follow voice instructions, including emotion, style, and prosody control; (3) **audio understanding preference**: preferring responses that demonstrate accurate comprehension of audio content; and (4) **voice empathy preference**: preferring responses that show appropriate emotional understanding and empathetic responses. The DPO training loss is computed across these multiple preference dimensions, allowing the model to learn a unified preference signal that balances all these

capabilities. This multi-task DPO training stage enables the model to better align with human preferences and improve performance on real-world conversational scenarios, distinguishing Fun-Audio-Chat from previous works that primarily rely on supervised fine-tuning.

Full-duplex Interaction Training. To enable real-time full-duplex voice interaction, we introduce a parallel speech-text input stream architecture and extend Fun-Audio-Chat to a full-duplex variant, **Fun-Audio-Chat-Duplex**, which can support natural human-like conversations with seamless two-way communication. Specifically, the parallel speech-text input stream architecture allows the model to accept user speech when the assistant is generating speech, effectively utilizing the time slots that would otherwise be idle. The parallel input stream is designed to handle both user and assistant speech inputs simultaneously, enabling the model to process overlapping speech segments and maintain conversation context. The Full-duplex Interaction Training continues from the checkpoint resulting from the Core-Cocktail Training stage, building upon the multimodal capabilities that the model already acquires. Full-duplex training uses full-duplex conversation data synthesized by augmenting high-quality half-duplex dialogue datasets with simulated full-duplex interaction behaviors, following the data synthesis approach in OmniFlatten (Zhang et al., 2025). This approach transforms traditional turn-based text dialogues into concurrent dual-stream interactions, in which both user and assistant can speak simultaneously. Full-duplex training allows the model to learn natural turn-taking, interruption handling, and backchanneling behaviors.

3 Experiments

3.1 Experimental Setup

Evaluation Datasets. Following prior works (Yao et al., 2024; KimiTeam et al., 2025), we evaluate the performance of Fun-Audio-Chat comprehensively on the widely used benchmarks:

- **Speech-To-Text ($S \rightarrow T$) Evaluation.** We use two types of Spoken Question Answering benchmarks to evaluate the model’s ability to understand speech inputs and generate both text and speech responses, including $S \rightarrow T$ Evaluation and $S \rightarrow S$ Evaluation. For $S \rightarrow T$ evaluation, we use VoiceBench (Chen et al., 2024b) and OpenAudioBench⁴. **VoiceBench** encompasses AlpacaEval, CommonEval, SD-QA, MMSU, OpenBookQA, IFEval, and AdvBench, providing comprehensive evaluation across instruction-following, general knowledge, safety alignment, and robustness to real-world variations. In contrast, **OpenAudioBench** includes multiple sub-tasks including AlpacaEval (Li et al., 2023), Llama Q., Reasoning QA, TriviaQA, and Web Q., covering diverse Spoken Question Answering scenarios, with more focuses on general knowledge and reasoning and less emphases on robustness.
- **Speech-to-Speech ($S \rightarrow S$) Evaluation.** We use UltraEval-Audio⁵, which includes AlpacaEval, Llama Q., TriviaQA, and Web Q. for end-to-end Speech-to-Speech Question Answering evaluation.
- **Audio Understanding.** We evaluate on **audio understanding benchmarks** including MMAU (Sakshi et al., 2025), MMAU-Pro (Kumar et al., 2025), and MMSU (Wang et al., 2025a) for comprehensive audio comprehension capabilities. These benchmarks focus on different aspects of audio understanding: **MMAU** is a generalist benchmark covering the “Big Three” audio domains (Speech, Music, Sound) with a focus on complex reasoning; **MMAU-Pro** is an advanced-scenario benchmark that stresses models with “wild” conditions like long-form audio, spatial audio, and overlapping sounds; **MMSU** is a speech specialist benchmark grounded in linguistic theory, focusing deeply on the nuances of spoken language (intonation, emotion, prosody) rather than general environmental sounds or music.
- **Speech Recognition.** We evaluate ASR performance on the widely used Librispeech (Panayotov et al., 2015) for English (EN) ASR and Common Voice (Ardila et al., 2020) for English and Mandarin (ZH) ASR.

⁴<https://huggingface.co/datasets/baichuan-INC/OpenAudioBench>

⁵<https://github.com/OpenBMB/UltraEval-Audio>

-
- **Speech Function Calling.** We evaluate on Speech-ACEBench, Speech-BFCL, and Speech-SmartInteract⁶ for evaluating the model’s ability to execute function calls based on speech instructions. These three benchmarks focus on different aspects of speech function calling: **Speech-ACEBench** is derived from the text-based ACEBench (Chen et al., 2025) and contains Mandarin speech recorded by human speakers. It covers both single and parallel function calling scenarios, with particular emphasis on cases where functions take *nested (deep)* object-type arguments. **Speech-BFCL** is derived from BFCL (Patil et al., 2025) and consists of English data synthesized with TTS. It also targets single and parallel function calling, but focusing on TTS-generated English interactions. **Speech-SmartInteract** is a purpose-built TTS-synthesized Mandarin speech dataset designed specifically for speech-first interactive use; rather than merely voicing a text-based benchmark, it better reflects the characteristics of real spoken interactions in practical voice assistant settings.
 - **Speech Instruction-Following and Voice Empathy.** We use the VStyle benchmark (Zhan et al., 2025) to evaluate the model’s ability to understand and execute voice instructions for controlling speech generation attributes such as emotion, speaking style, speed, pitch, and volume. We also use an internal test set to assess the model’s speech instruction-following and voice empathy capabilities, including understanding emotional context and responding with appropriate empathetic expressions.

Evaluation Metrics. Evaluations adhere to the established protocols for each respective benchmark. For $S \rightarrow T$ and $S \rightarrow S$ evaluations on **Spoken Question Answering benchmarks**, we use different metrics depending on the task type: (1) **Accuracy** is used for close-ended QA tasks including Llama Q., Reasoning QA, TriviaQA, Web Q., SD-QA, MMSU, OpenBookQA, and IFEval; (2) **G-Eval** (Liu et al., 2023) is used for open-ended QA tasks including AlpacaEval (Li et al., 2023) and CommonEval, which employs LLM-based evaluation to assess response quality; (3) **Refusal Rate** is reported for AdvBench to measure safety compliance.

Additionally, for **Speech Quality evaluation**, the generated speech is transcribed using Whisper-v3-large model (Radford et al., 2022), then **ASR-WER** (Word Error Rate of the ASR-transcripts against the model-generated text) is used to assess the alignment between the generated speech and text. **UTMOS** (Saeki et al., 2022) is used to evaluate the overall speech quality,

For **Audio Understanding tasks** (MMAU, MMAU-Pro, MMSU), **Accuracy** is used to measure the model’s comprehension capabilities across diverse audio understanding scenarios. For **Speech Recognition tasks** (Librispeech, Common Voice), **Word Error Rate (WER)** is reported.

For **Speech Function Calling tasks** (Speech-ACEBench, Speech-BFCL, Speech-SmartInteract), **Accuracy** is used to measure the percentage of correctly executed function calls.

For **Speech Instruction-Following and Voice Empathy tasks**, for the VStyle benchmark, we use **Large Audio Language Model (LALM) evaluation scores** on a 1-5 scale across multiple dimensions: acoustic attributes (age, speed, gender, emotion, pitch, volume), instruction following (emotion, style, variation), role-play (scenario, character), and empathy (anger, sadness, anxiety, joy). For evaluations on our internal test set, similar to the VStyle benchmark, we use LALM as Judge to evaluate the model’s performance on Speech Instruction-Following, Semantics-based Empathy, and Paralinguistic-Cue-based Empathy. **Semantics-based Empathy** refers to the empathy capability that can be judged solely based on text semantics, while **Paralinguistic-Cue-based Empathy** refers to the empathy capability that requires using Paralinguistic Cues to judge and cannot be judged solely from text semantics.

For **Full-Duplex Interaction evaluation**, we use **S2M-T** (the text output accuracy in multimodal response) and **S2M-S** (the speech output accuracy in multimodal response) to measure the knowledge understanding performance, and the **Turn-taking Success Rate** to measure the percentage of interactions where the model correctly handles turn-taking in full-duplex scenarios.

⁶<https://huggingface.co/datasets/FunAudioLLM/SpeechFCEval>

Baselines. We select representative and competitive models as baselines to ensure comprehensive comparisons across different model sizes and model architectures. For **around-8B dense models**, we compare Fun-Audio-Chat-8B with open-source Large Audio Language Models (LALMs) including GLM-4-Voice (9B) (Zeng et al., 2024), MiniCPM-o 2.6 (7B) (Yao et al., 2024), Baichuan-Omni-1.5 (7B) (Li et al., 2025b), Kimi-Audio (7B) (KimiTeam et al., 2025), Step-Audio2-Mini (7B) (Wu et al., 2025), and MiMo-Audio (7B) (Xiaomi, 2025). For **large-scale models**, we compare Fun-Audio-Chat-30B-A3B with the open-source Longcat-Flash-Omni-Instruct (560B-A27B) (Team, 2025b) and the closed-source GPT-Audio (OpenAI, 2024b) and Gemini-2.5-Pro (Team, 2025a). For **audio understanding tasks**, we additionally compare with the open-source Audio-Flamingo-3 (Goel et al., 2025) alongside Kimi-Audio, Step-Audio2-Mini, and MiMo-Audio. For **speech instruction-following and voice empathy tasks** (VStyle benchmark), we compare with the open-source Baichuan-Audio (Li et al., 2025a) and Kimi-Audio, and add the closed-source GPT-4o (OpenAI, 2024b) and Doubao⁷ into baselines for comprehensive comparison with both open-source and commercial models. For **full-duplex interaction evaluation**, we compare Fun-Audio-Chat-Duplex with the open-source Moshi (Défossez et al., 2024) and FreezeOmni (Wang et al., 2025b).

In summary, the selected baselines cover diverse modeling paradigms (Text-Driven vs. Joint Speech-Text, interleaved vs. parallel architectures) and model scales, enabling systematic comparisons across mainstream speech-text modeling strategies and providing comprehensive evaluation of Fun-Audio-Chat’s capabilities across different task categories.

3.2 Spoken Question Answering

Accuracy. Fun-Audio-Chat demonstrates strong performance on spoken question answering tasks. Table 1 compares Fun-Audio-Chat-30B-A3B with large-scale baselines, including GPT-Audio, Gemini-2.5-Pro, and Longcat-Flash-Omni-Instruct. Table 2 compares Fun-Audio-Chat-8B with Kimi-Audio, Step-Audio2-Mini, and other similarly-scaled open-source models. As shown in Table 1 and Table 2, Fun-Audio-Chat achieves competitive performance among similarly-scaled models (8B and 30B-A3B parameters). Specifically, **Fun-Audio-Chat-8B achieves the best overall performance on OpenAudioBench (76.61%) and VoiceBench (83.21%) among ~8B-scale models**, while **Fun-Audio-Chat-30B-A3B achieves competitive results compared to large-scale baselines, including top-tier closed-source models**.

Speech Quality. We evaluate the speech quality of Fun-Audio-Chat-8B on UltraEval-Audio using UTMOS for the overall speech quality and ASR-WER for alignment between the generated speech and text. On the Llama Q. test set, Fun-Audio-Chat-8B achieves a UTMOS score of 4.37, indicating excellent overall speech quality, and an ASR-WER of 4.32%, demonstrating strong alignment between the generated speech and the corresponding text outputs. These results demonstrate that the dual-resolution architecture maintains high-quality speech generation despite operating at the efficient 5Hz frame rate, validating the effectiveness of the Dual-Resolution Speech Representations (DRSR) architecture in balancing efficiency and speech quality.

3.3 Audio Understanding

Table 3 demonstrates that **Fun-Audio-Chat achieves the best performance on comprehensive audio understanding benchmarks including MMAU, MMAU-Pro, and MMSU, over strong open-source baselines**, including Kimi-Audio (KimiTeam et al., 2025), Audio-Flamingo-3 (Goel et al., 2025), MiMo-Audio (Xiaomi, 2025), and Step-Audio2-Mini (Wu et al., 2025). On MMAU⁸, Fun-Audio-Chat-30B-A3B achieves the best performance (77.9%) among all evaluated models, followed by Fun-Audio-Chat-8B (76.6%). On MMAU-Pro⁹, Fun-Audio-Chat-30B-A3B achieves the best result (59.9%), with Fun-

⁷<https://www.doubao.com/>

⁸MMAU v05.15.25 test-mini.

⁹Only one audio test case included.

Table 1: Performance comparison on **Spoken Question Answering** benchmarks for **large-scale models**. The best result in each row is in **bold**. **Frame Rate-In** denotes the input speech frame rate (Hz), and **Frame Rate-Out** denotes the output (speech + text) frame rate (Hz) for the LLM backbone.

	GPT-Audio	Gemini-2.5-Pro	Longcat-Flash -Omni-Instruct	Fun-Audio-Chat -30B-A3B
LLM Size	–	–	560B-A27B	30B-A3B
Frame Rate-In	–	–	12.5	5
Frame Rate-Out	–	–	16.67	5
<i>OpenAudioBench (S2T)</i>				
AlpacaEval	83.37	76.58	75.43	88.89
Llama Q.	90.67	83.00	83.33	85.00
Reasoning QA	74.75	80.30	79.71	75.25
TriviaQA	92.20	90.20	86.20	76.00
Web Q.	83.70	80.90	76.00	77.80
Overall	84.94	82.20	80.13	80.59
<i>VoiceBench (S2T)</i>				
AlpacaEval	4.84	4.70	4.94	4.82
CommonEval	4.47	4.11	4.32	4.49
SD-QA	89.72	83.54	82.46	72.87
MMSU	83.25	88.32	81.95	75.31
OpenBookQA	92.53	95.16	93.41	88.57
IFEval	79.12	77.83	77.99	77.25
AdvBench	99.62	97.69	100	99.23
Overall	90.06	88.39	88.72	85.63
<i>UltraEval-Audio (S2S)</i>				
AlpacaEval	73.38	–	–	64.49
Llama Q.	89.00	–	–	78.67
TriviaQA	72.85	–	–	54.20
Web Q.	55.41	–	–	51.18
Overall	72.66	–	–	62.14

Audio-Chat-8B achieving the second-best performance (58.0%). On MMSU, Fun-Audio-Chat-30B-A3B achieves 70.1%, the highest result among all models, followed by Fun-Audio-Chat-8B (67.8%). For speech recognition tasks, Fun-Audio-Chat achieves competitive WERs across multiple datasets in both English (EN) and Mandarin (ZH), demonstrating robust audio comprehension capabilities across diverse domains and languages.

3.4 Speech Function Calling

Table 4 presents the performance of Fun-Audio-Chat on speech function calling benchmarks. **Fun-Audio-Chat-30B-A3B achieves the highest overall score (79.63%) among all evaluated models**, with particularly strong performance on Speech-ACEBench (Single: 76.40%) and Speech-SmartInteract (84.13%). The model demonstrates strong capabilities in understanding speech-based function calling instructions and executing them accurately, which is crucial for building practical voice-controlled applications. **The performance on parallel function calling scenarios (54.50% on ACEBench-Parallel and 87.63% on BFCL-Parallel by Fun-Audio-Chat-8B) further highlights Fun-Audio-Chat’s ability to handle complex, multi-step instructions in voice interactions, with Fun-Audio-Chat-8B outperforming the top tier closed-source GPT-Audio and Gemini-2.5-Pro on BFCL-Parallel.**

3.5 Speech Instruction-Following and Voice Empathy

Table 5 and Table 6 demonstrate that **Fun-Audio-Chat achieves strong performance on Speech Instruction-Following and Voice Empathy tasks**. As shown in Table 5, Fun-Audio-Chat-30B-A3B and Fun-Audio-Chat-8B demonstrate competitive performance on Speech Instruction-Following across multiple dimensions, including acoustic attributes, instruction following, role-play, and empathy capabilities, in

Table 2: Performance comparison on **Spoken Question Answering** benchmarks for $\sim 8\text{B}$ -scale **dense models**. The best result in each row is in **bold**. **Frame Rate-In** denotes the input speech frame rate (Hz), and **Frame Rate-Out** denotes the output (speech + text) frame rate (Hz) for the LLM backbone. τ denotes the average number of text tokens per second of speech.

	GLM4 -Voice	MiniCPM -o 2.6	Baichuan -Omni-1.5	Kimi -Audio	Step-Audio2 -Mini	MiMo -Audio	Fun-Audio-Chat -8B
LLM Size	9B	7B	7B	7B	7B	7B	8B
Frame Rate-In	12.5	25	12.5	12.5	12.5	6.25	5
Frame Rate-Out	$12.5+\tau$	τ	$12.5+\tau$	12.5	$25+\tau$	$6.25+\tau$	5
<i>OpenAudioBench (S2T)</i>							
AlpacaEval	57.89	64.10	77.90	75.73	59.60	85.43	88.94
Llama Q.	76.00	78.00	78.50	79.33	75.00	79.67	83.33
Reasoning QA	47.43	38.60	50.00	58.02	46.04	53.96	69.80
TriviaQA	51.80	63.00	57.20	62.10	57.70	52.80	68.10
Web Q.	55.40	69.20	59.10	70.20	65.10	55.40	72.90
Overall	57.70	62.58	64.54	69.08	60.69	65.45	76.61
<i>VoiceBench (S2T)</i>							
AlpacaEval	3.97	4.42	4.50	4.46	4.17	4.60	4.80
CommonEval	3.42	4.15	4.05	3.97	3.00	3.77	4.42
SD-QA	36.98	50.72	43.40	63.12	56.06	54.79	66.27
MMSU	39.75	54.78	57.25	62.17	52.18	59.66	71.08
OpenBookQA	53.41	78.02	74.51	83.52	64.18	73.41	83.52
IFEval	52.80	49.25	54.54	61.10	38.01	66.45	78.52
AdvBench	88.08	97.69	97.31	100.00	93.08	96.73	98.65
Overall	59.83	71.69	71.14	76.93	63.84	74.06	83.21
<i>UltraEval-Audio (S2S)</i>							
AlpacaEval	51.00	51.00	58.69	44.20	51.72	61.46	61.87
Llama Q.	50.00	61.00	67.33	57.33	67.67	77.33	78.33
TriviaQA	36.40	40.20	30.57	35.71	33.50	40.43	49.51
Web Q.	32.00	40.00	38.09	33.90	34.65	42.86	48.52
Overall	42.35	48.05	48.67	42.79	46.89	55.52	59.56

Table 3: Performance comparison on **Audio understanding** (top section) on MMAU, MMAU-Pro, and MMSU, and **Speech Recognition** (bottom) on Librispeech and Common Voice. The best result in each row is in **bold**.

	Kimi -Audio	Audio -Flamingo-3	Step-Audio2 -Mini	MiMo -Audio	Fun-Audio-Chat -30B-A3B	Fun-Audio-Chat -8B
<i>Audio Understanding</i>						
MMAU	69.6	73.3	73.2	74.9	77.9	76.6
MMAU-Pro	46.6	51.7	53.2	53.4	59.9	58.0
MMSU	59.3	61.4	56.8	61.7	70.1	67.8
<i>Speech Recognition</i>						
Librispeech clean	1.28	1.57	1.33	3.56	1.64	1.71
Librispeech other	2.42	3.13	2.86	16.22	3.73	4.13
Common Voice-EN	10.31	7.4	6.76	62.05	7.79	8.88
Common Voice-ZH	7.21	–	5.38	44.11	5.88	6.16

both English and Chinese, substantially outperforming open-source models including Baichuan-Audio and Kimi-Audio while remaining competitive with commercial models. (1) In terms of the **Overall** performance, Fun-Audio-Chat-8B achieves scores 3.35 and 3.46 for English and Mandarin respectively, substantially outperforming the open-source Baichuan-Audio (2.50/2.25) and Kimi-Audio (2.54/3.11) in both languages, while remaining competitive with commercial models. (2) Specifically, for **acoustic attributes**, Fun-Audio-Chat-8B shows strong performance in emotion control (4.13/4.00 for en/zh) and volume control (3.95/3.70), demonstrating effective acoustic attribute manipulation capabilities. Notably,

Table 4: Performance comparison on **Speech Function Calling**. The best result in each row is in **bold**.

	GPT-Audio	Gemini -2.5-Pro	Step-Audio2 -Mini	Fun-Audio-Chat -30B-A3B	Fun-Audio-Chat -8B
Speech-ACEBench (Single)	68.30	68.30	38.90	76.40	66.30
Speech-ACEBench (Parallel)	60.20	53.40	4.50	59.10	54.50
Speech-BFCL (Single)	88.58	88.41	77.51	92.21	92.73
Speech-BFCL (Parallel)	83.60	80.91	49.73	86.29	87.63
Speech-SmartInteract (Single)	66.77	79.19	41.92	84.13	79.79
Overall	73.49	74.04	42.51	79.63	76.19

Fun-Audio-Chat-8B achieves the best performance on Age control in English (4.04), and achieves a score of 4.20 on speed control in Mandarin, ranking second only to Doubao (4.35). (3) In **instruction-following** tasks, Fun-Audio-Chat-8B achieves moderate performance with scores of 4.09 and 3.14 for style control in English and Mandarin, indicating room for improvements in complex instruction understanding. (4) For **role-play** capabilities, Fun-Audio-Chat-8B performs better in Mandarin (3.42/3.30 for scenario/character) compared to English (2.50/3.06), suggesting stronger contextual understanding in Mandarin scenarios.

We further evaluate Speech Instruction-Following and Voice Empathy capabilities on our internal test set, as shown in Table 6. Notably, Fun-Audio-Chat achieves superior performance over GPT-Audio in terms of both Semantics-based Empathy and Paralinguistic-Cue-based Empathy, demonstrating **the model’s strong ability to understand emotional context and respond with appropriate empathetic expressions**.

3.6 Full-Duplex Interaction

We evaluate the full-duplex variant Fun-Audio-Chat-Duplex on two key aspects: knowledge understanding in full-duplex scenarios and objective full-duplex interaction metrics.

Full-Duplex Knowledge Understanding. Table 7 shows the full-duplex knowledge understanding performance of Fun-Audio-Chat-Duplex. The results demonstrate that **Fun-Audio-Chat-Duplex maintains strong knowledge understanding capabilities in full-duplex conversation scenarios**. Fun-Audio-Chat-Duplex-30B-A3B achieves the highest average performance on both S2M-T (54.89%) and S2M-S (49.28%) metrics, significantly outperforming Moshi (33.17%/29.86%) and FreezeOmni (Wang et al., 2025b) (47.58%/34.49%). On individual benchmarks, Fun-Audio-Chat-Duplex-30B-A3B achieves the highest results on Llama Q. (81.00%/71.33%), AlpacaEval (68.23%/59.65%), and TriviaQA (41.70%/40.04%) for both text and speech outputs. This indicates that the full-duplex architecture successfully preserves the model’s knowledge comprehension abilities while enabling simultaneous two-way communication, allowing the system to maintain context and understanding even when processing overlapping speech inputs and outputs.

Full-Duplex Interaction. Table 7 also presents the turn-taking success rates for full-duplex voice interactions. Fun-Audio-Chat-Duplex-30B-A3B achieves perfect turn-taking success rate (100.00%), outperforming both Moshi (99.77%) and FreezeOmni (Wang et al., 2025b) (93.87%). Fun-Audio-Chat-Duplex-8B achieves 99.94%, also demonstrating excellent turn-taking capabilities. These results indicate that Fun-Audio-Chat-Duplex successfully enables natural and efficient full-duplex voice interactions, with the model’s ability to handle simultaneous speech and maintain appropriate conversation flow, closely mirroring the dynamics of human-human conversations.

3.7 Computational Efficiency

A key advantage of Fun-Audio-Chat is its computational efficiency, highlighted in Table 1 and Table 2. As shown in the **Frame Rate-In/Frame Rate-Out** rows, Fun-Audio-Chat operates at a frame rate of **5/5 Hz**, indicating that the LLM backbone processes only 5 audio tokens per second for both input and output. This represents a $1.25 \times$ to $5 \times$ reduction in input frame rate compared to other models, which operate

Table 5: Performance comparison on **Speech Instruction-Following** on the VStyle benchmark (Zhan et al., 2025). The best result in each row for each language is in **bold**.

	Lang	GPT-Audio	GPT-4o	Doubaao	Baichuan-Audio	Kimi-Audio	Fun-Audio-Chat-30B-A3B	Fun-Audio-Chat-8B
Overall	en	3.78	4.05	3.63	2.50	2.54	3.31	3.35
	zh	3.75	3.84	4.10	2.25	3.11	3.68	3.46
<i>Acoustic Attributes</i>								
Age	en	3.67	3.67	3.75	2.71	2.79	3.79	4.04
	zh	3.67	3.42	3.88	2.67	3.33	3.42	3.50
Speed	en	4.05	3.45	3.55	2.20	2.45	3.55	3.45
	zh	3.65	3.10	4.35	2.45	3.45	4.47	4.20
Gend.	en	3.75	2.79	3.46	3.83	2.54	2.96	3.33
	zh	4.08	3.50	3.25	3.08	2.25	3.26	3.08
Emot.	en	4.50	4.00	3.38	2.58	3.04	4.25	4.13
	zh	4.42	3.83	4.65	2.29	3.75	4.08	4.00
Pitch	en	3.30	3.60	3.25	2.05	1.55	2.95	3.20
	zh	3.05	3.35	4.35	2.00	2.95	3.00	2.75
Vol.	en	4.20	4.10	4.05	2.05	3.00	3.90	3.95
	zh	4.25	3.90	4.70	2.80	3.25	4.35	3.70
Comp.	en	3.73	3.27	3.13	2.55	2.33	3.36	3.17
	zh	3.47	3.22	3.77	2.58	3.17	3.60	3.37
<i>Instruction</i>								
Emot.	en	4.13	3.93	3.52	2.23	2.19	3.88	3.70
	zh	3.73	3.37	3.90	1.71	2.66	3.66	3.42
Style	en	4.51	4.23	3.67	2.21	2.41	3.79	4.09
	zh	4.07	3.51	3.96	1.72	2.74	4.01	3.14
Vari.	en	4.03	4.07	2.90	1.88	2.33	3.47	3.06
	zh	3.48	3.11	2.88	1.69	2.43	2.96	2.94
<i>Role-Play</i>								
Scen.	en	2.65	3.89	3.27	2.08	1.73	2.65	2.50
	zh	3.69	3.89	4.45	2.29	3.01	4.02	3.42
Char.	en	3.37	3.83	2.56	2.33	1.72	2.48	3.06
	zh	3.65	3.90	3.79	1.95	2.23	3.95	3.30
<i>Empathy</i>								
Anger	en	4.25	4.95	4.89	2.41	3.59	2.84	3.64
	zh	3.80	4.75	4.59	2.11	3.86	3.64	3.73
Sad.	en	3.80	4.90	5.00	3.43	3.97	4.00	4.10
	zh	3.62	4.83	4.72	2.55	3.86	3.83	3.93
Anx.	en	4.23	5.00	4.81	2.74	3.65	3.61	2.90
	zh	4.33	4.67	4.80	2.20	3.80	2.90	4.03
Joy	en	3.97	4.54	4.94	3.91	3.46	3.54	3.69
	zh	3.91	4.80	4.83	3.51	4.57	3.71	3.77

Table 6: Performance comparison on **Speech Instruction-Following and Voice Empathy** on our internal test set. The best result in each row is in **bold**.

	GPT-Audio	Fun-Audio-Chat-30B-A3B	Fun-Audio-Chat-8B
Speech Instruction-Following	4.53	4.31	3.98
Semantics-based Empathy	4.73	4.80	4.80
Paralinguistic-Cue-based Empathy	3.20	3.55	3.85

at frame rates ranging from 6.25Hz (MiMo-Audio) to 25Hz (MiniCPM-o 2.6), with most models using 12.5Hz. For output frame rates, Fun-Audio-Chat’s 5Hz is significantly lower than other models, which operate at rates of 12.5Hz, 16.67Hz, 25Hz, or higher when including text token generation, e.g., $12.5 + \tau$ for GLM-4-Voice and Baichuan-Omni-1.5, $25 + \tau$ for Step-Audio2-Mini, where τ denotes the average number of text tokens per second of speech. **The dual-resolution design significantly reduces computational requirements and potential latency, with empirical measurements showing approximately 50% reduc-**

Table 7: Performance comparison on Knowledge Understanding (in terms of **S2M-T** (text output in multimodal response) and **S2M-S** (speech output in multimodal response)) and Full-duplex Interaction (in terms of Turn-taking Success Rate) on the full-duplex variant of the UltraEvalAudio benchmark. The best result for each metric on each dataset is in **bold**.

	Moshi		FreezeOmni		Fun-Audio-Chat-Duplex-30B-A3B		Fun-Audio-Chat-Duplex-8B	
	S2M-T	S2M-S	S2M-T	S2M-S	S2M-T	S2M-S	S2M-T	S2M-S
Llama Q.	65.67	57.00	74.00	58.00	81.00	71.33	72.33	64.33
AlpacaEval	25.51	25.08	47.39	32.78	68.23	59.65	68.03	57.32
TriviaQA	18.46	16.31	30.08	20.31	41.70	40.04	29.59	27.73
Web Q.	23.03	21.06	38.83	26.87	28.64	26.08	26.18	24.36
Avg.	33.17	29.86	47.58	34.49	54.89	49.28	49.03	43.44
Turn-taking Success Rate	99.77		93.87		100.00		99.94	

tion in GPU hours during training compared to models operating at higher frame rates. Importantly, this efficiency is achieved without compromising speech quality, as demonstrated by the high-quality speech generation results.

4 Conclusion

This report introduces Fun-Audio-Chat, a large-scale Large Audio Language Model (LALM) designed to overcome the limitations of existing joint speech-text models for seamless voice interaction. Fun-Audio-Chat extends our previous work DrVoice (Tan et al., 2025) by adopting one key innovation, Dual-Resolution Speech Representations (DRSR) architecture, at significantly larger scales. The DRSR architecture enables the Shared LLM backbone to process audio at an efficient 5Hz frame rate (Frame Rate-In/Frame Rate-Out: 5/5 Hz) while the Speech Refined Head generates high-quality speech tokens at 25Hz resolution. This dual-resolution design effectively balances computational efficiency (reducing GPU hours by nearly 50%) and speech generation quality.

To address the catastrophic forgetting challenge in multimodal learning, we adopt the Core-Cocktail Training strategy introduced in DrVoice (Tan et al., 2025), a two-stage approach with intermediate parameter merging. Subsequently, we enhance the model through Multi-Task DPO Training to strengthen the robustness to real speech data, capabilities of speech instruction-following, audio understanding, and voice empathy. The multi-stage post-training paradigm enables Fun-Audio-Chat to retain the original text-LLM’s capabilities while gaining powerful multimodal skills.

Trained on millions of hours of diverse speech data and scaled to larger model sizes (dense 8B and MoE 30B-A3B parameters), Fun-Audio-Chat achieves strong performance on Spoken Question Answering (Speech-to-Text and Speech-to-Speech generation) tasks, ranking Top among models of the same sizes. It also achieves competitive results on audio understanding, speech function calling, speech instruction-following, and voice empathy tasks, as demonstrated across comprehensive benchmarks including OpenAudioBench, VoiceBench, UltraEvalAudio, MMAU, MMAU-Pro, MMSU, Speech-ACEBench, Speech-BFCL, Speech-SmartInteract, and VStyle. Furthermore, we develop Fun-Audio-Chat-Duplex, a full-duplex variant that achieves strong performance on Spoken Question Answering benchmarks and full-duplex interactions.

We open-source Fun-Audio-Chat-8B model, including the model checkpoint and its training and inference code, and provide an interactive demo, encouraging researchers and practitioners to experience and build upon our work. We believe that Fun-Audio-Chat represents a significant advancement in the field of voice interaction systems, demonstrating that carefully designed large-scale post-training and architectural innovations can significantly enhance the audio comprehension, reasoning, and speech

generation capabilities of LALMs while achieving high computational efficiency.

5 Limitations

While Fun-Audio-Chat demonstrates strong performance across multiple benchmarks, several limitations remain to be addressed in future work. First, for complex question answering in multi-turn conversations, the model occasionally exhibits memory loss of context, where information from earlier turns may not be consistently retained. This limitation is particularly noticeable in scenarios requiring long-context comprehension and complex reasoning across multiple turns.

Second, speech instruction-following capabilities show some instability in expressiveness. While the model generally performs strongly on voice instruction tasks, there are cases where the generated speech may not fully capture the intended emotional nuances, speaking styles, or prosodic variations specified in the instructions. This variability in expressiveness can affect the naturalness and appropriateness of voice responses in certain contexts.

Third, the voice empathy capabilities demonstrate some instability in performance. Although Fun-Audio-Chat achieves competitive results on empathy evaluation benchmarks (including both Semantics-based Empathy and Paralinguistic-Cue-based Empathy), the model’s ability to consistently recognize and respond with appropriate emotional empathy can vary across different scenarios and emotional contexts. This inconsistency may impact the reliability of empathetic response generation in real-world applications where emotional understanding is critical.

These limitations highlight important directions for future research, including improving long-term context management in multi-turn conversations, enhancing the stability and expressiveness of speech instruction-following, and developing more robust and consistent voice empathy capabilities across diverse emotional scenarios.

6 Contributions and Acknowledgments

All contributors of Fun-Audio-Chat are listed in alphabetical order by their last names.

Core contributors: Qian Chen, Luyao Cheng, Chong Deng, Xiangang Li, Jiaqing Liu, Chao-Hong Tan, Wen Wang, Junhao Xu, Jieping Ye, Qinglin Zhang, Qiquan Zhang, Jingren Zhou

Contributors: Zhifu Gao, Weiqin Li, Mengge Liu, Xiang Lv, Yukun Ma, Gang Qiao, Hui Wang, Chong Zhang, Han Zhao, Tianyu Zhao

References

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis M. Tyers, and Gregor Weber. Common voice: A massively-multilingual speech corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, and Stelios Piperidis (eds.), *Proceedings of The 12th Language Resources and Evaluation Conference, LREC 2020, Marseille, France, May 11-16, 2020*, pp. 4218–4222. European Language Resources Association, 2020. URL <https://aclanthology.org/2020.lrec-1.520/>.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-vl technical report, 2025. URL <https://arxiv.org/abs/2511.21631>.

Zalán Borsos, Raphaël Marinier, Damien Vincent, Eugene Kharitonov, Olivier Pietquin, Matthew Sharifi, Dominik Roblek, Olivier Teboul, David Grangier, Marco Tagliasacchi, and Neil Zeghidour. Audioltm: A language modeling approach to audio generation. *IEEE ACM Trans. Audio Speech Lang. Process.*, 31:2523–2533, 2023. doi: 10.1109/TASLP.2023.3288409. URL <https://doi.org/10.1109/TASLP.2023.3288409>.

Chen Chen, Xinlong Hao, Weiwen Liu, Xu Huang, Xingshan Zeng, Shuai Yu, Dexun Li, Shuai Wang, Weinan Gan, Yuefeng Huang, Wulong Liu, Xinzhi Wang, Defu Lian, Baoqun Yin, Yasheng Wang, and Wu Liu. Acebench: Who wins the match point in tool learning? *CoRR*, abs/2501.12851, 2025. doi: 10.48550/ARXIV.2501.12851. URL <https://doi.org/10.48550/arXiv.2501.12851>.

Wenxi Chen, Ziyang Ma, Ruiqi Yan, Yuzhe Liang, Xiquan Li, Ruiyang Xu, Zhikang Niu, Yanqiao Zhu, Yifan Yang, Zhanxun Liu, et al. Slam-omni: Timbre-controllable voice interaction system with single-stage training. *arXiv preprint arXiv:2412.15649*, 2024a.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. Voicebench: Benchmarking llm-based voice assistants. *CoRR*, abs/2410.17196, 2024b. doi: 10.48550/ARXIV.2410.17196. URL <https://doi.org/10.48550/arXiv.2410.17196>.

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. *CoRR*, abs/2410.00037, 2024. doi: 10.48550/ARXIV.2410.00037. URL <https://doi.org/10.48550/arXiv.2410.00037>.

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>.

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. Cosyvoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *CoRR*, abs/2407.05407, 2024a. doi: 10.48550/ARXIV.2407.05407. URL <https://doi.org/10.48550/arXiv.2407.05407>.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen

Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *CoRR*, abs/2412.10117, 2024b. doi: 10.48550/ARXIV.2412.10117. URL <https://doi.org/10.48550/arXiv.2412.10117>.

Zhihao Du, Changfeng Gao, Yuxuan Wang, Fan Yu, Tianyu Zhao, Hao Wang, Xiang Lv, Hui Wang, Chongjia Ni, Xian Shi, Keyu An, Guanrou Yang, Yabin Li, Yanni Chen, Zhifu Gao, Qian Chen, Yue Gu, Mengzhe Chen, Yafeng Chen, Shiliang Zhang, Wen Wang, and Jieping Ye. Cosyvoice 3: Towards in-the-wild speech generation via scaling-up and post-training. *CoRR*, abs/2505.17589, 2025. doi: 10.48550/ARXIV.2505.17589. URL <https://doi.org/10.48550/arXiv.2505.17589>.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, and Bryan Catanzaro. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *CoRR*, abs/2507.08128, 2025. doi: 10.48550/ARXIV.2507.08128. URL <https://doi.org/10.48550/arXiv.2507.08128>.

KimiTeam, Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, Zhengtao Wang, Chu Wei, Yifei Xin, Xinran Xu, Jianwei Yu, Yutao Zhang, Xinyu Zhou, Y. Charles, Jun Chen, Yanru Chen, Yulun Du, Weiran He, Zhenxing Hu, Guokun Lai, Qingcheng Li, Yangyang Liu, Weidong Sun, Jianzhou Wang, Yuzhi Wang, Yuefeng Wu, Yuxin Wu, Dongchao Yang, Hao Yang, Ying Yang, Zhilin Yang, Aoxiong Yin, Ruibin Yuan, Yutong Zhang, and Zaida Zhou. Kimi-audio technical report, 2025. URL <https://arxiv.org/abs/2504.18425>.

Jungil Kong, Jaehyeon Kim, and Jaekyoung Bae. Hifi-gan: Generative adversarial networks for efficient and high fidelity speech synthesis. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin (eds.), *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. URL <https://proceedings.neurips.cc/paper/2020/hash/c5d736809766d46260d816d8dbc9eb44-Abstract.html>.

Sonal Kumar, Simon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plicka, Miroslav Hlaváček, William Fineas Ellingwood, Sathvik Udupa, Siyuan Hou, Allison Ferner, Sara Barahona, Cecilia Bolaños, Satish Rahi, Laura Herrera-Alarcón, Satvik Dixit, Rupali S. Patil, Soham Deshmukh, Lasha Koroshinadze, Yao Liu, Leibny Paola Garcia Perera, Eleni Zanou, Themos Stafylakis, Joon Son Chung, David Harwath, Chao Zhang, Dinesh Manocha, Alicia Lozano-Diez, Santosh Kesiraju, Sreyan Ghosh, and Ramani Duraiswami. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *CoRR*, abs/2508.13992, 2025. doi: 10.48550/ARXIV.2508.13992. URL <https://doi.org/10.48550/arXiv.2508.13992>.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, Jianhua Xu, Haoze Sun, Zenan Zhou, and Weipeng Chen. Baichuan-audio: A unified framework for end-to-end speech interaction. *CoRR*, abs/2502.17239, 2025a. doi: 10.48550/ARXIV.2502.17239. URL <https://doi.org/10.48550/arXiv.2502.17239>.

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 2023.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matthew Le. Flow matching for generative modeling. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL <https://openreview.net/forum?id=PqvMRDCJT9t>.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. G-eval: NLG evaluation using gpt-4 with better human alignment. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pp. 2511–2522. Association for Computational Linguistics, 2023. doi: 10.18653/V1/2023.EMNLP-MAIN.153. URL <https://doi.org/10.18653/v1/2023.emnlp-main.153>.

OpenAI. Hello GPT-4o, 2024b. URL <https://openai.com/index/hello-gpt-4o/>.

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 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pp. 5206–5210. IEEE, 2015. doi: 10.1109/ICASSP.2015.7178964. URL <https://doi.org/10.1109/ICASSP.2015.7178964>.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, Charlie Cheng-Jie Ji, Vishnu Suresh, Ion Stoica, and Joseph E. Gonzalez. The berkeley function calling leaderboard (BFCL): from tool use to agentic evaluation of large language models. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=2GmDdhBdDk>.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision, 2022. URL <https://arxiv.org/abs/2212.04356>.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. In Alice Oh, Tristan Naumann, Amir Globerson, Kate Saenko, Moritz Hardt, and Sergey Levine (eds.), *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*, 2023. URL http://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html.

Takaaki Saeki, Detai Xin, Wataru Nakata, Tomoki Koriyama, Shinnosuke Takamichi, and Hiroshi Saruwatari. UTMOS: utokyo-sarulab system for voicemos challenge 2022. In Hanseok Ko and John H. L. Hansen (eds.), *23rd Annual Conference of the International Speech Communication Association, Interspeech 2022, Incheon, Korea, September 18-22, 2022*, pp. 4521–4525. ISCA, 2022. doi: 10.21437/INTERSPEECH.2022-439. URL <https://doi.org/10.21437/Interspeech.2022-439>.

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. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL <https://openreview.net/forum?id=TeVAZXr3yv>.

Chao-Hong Tan, Qian Chen, Wen Wang, Chong Deng, Qinglin Zhang, Luyao Cheng, Hai Yu, Xin Zhang, Xiang Lv, Tianyu Zhao, Chong Zhang, Yukun Ma, Yafeng Chen, Hui Wang, Jiaqing Liu, Xiangang Li, and Jieping Ye. Drvoice: Parallel speech-text voice conversation model via dual-resolution speech representations, 2025. URL <https://arxiv.org/abs/2506.09349>.

Gemini Team. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *CoRR*, abs/2507.06261, 2025a. doi: 10.48550/ARXIV.2507.06261. URL <https://doi.org/10.48550/arXiv.2507.06261>.

Meituan LongCat Team. Longcat-flash-omni technical report. *CoRR*, abs/2511.00279, 2025b.

Dingdong Wang, Jincenzi Wu, Junan Li, Dongchao Yang, Xueyuan Chen, Tianhua Zhang, and Helen Meng. MMSU: A massive multi-task spoken language understanding and reasoning benchmark. *CoRR*, abs/2506.04779, 2025a. doi: 10.48550/ARXIV.2506.04779. URL <https://doi.org/10.48550/arXiv.2506.04779>.

Xiong Wang, Yangze Li, Chaoyou Fu, Yike Zhang, Yunhang Shen, Lei Xie, Ke Li, Xing Sun, and Long Ma. Freeze-omni: A smart and low latency speech-to-speech dialogue model with frozen LLM. In *Forty-second International Conference on Machine Learning, ICML 2025, Vancouver, BC, Canada, July 13-19, 2025*. OpenReview.net, 2025b. URL <https://openreview.net/forum?id=s1EImzs5Id>.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, Mingrui Chen, Peng Liu, Wang You, Xiangyu Tony Zhang, Xingyuan Li, Xuerui Yang, Yayue Deng, Yechang Huang, Yuxin Li, Yuxin Zhang, Zhao You, Brian Li, Changyi Wan, Hanpeng Hu, Jiangjie Zhen, Siyu Chen, Song Yuan, Xuelin Zhang, Yimin Jiang, Yu Zhou, Yuxiang Yang, Bingxin Li, Buyun Ma, Changhe Song, Dongqing Pang, Guoqiang Hu, Haiyang Sun, Kang An, Na Wang, Shuli Gao, Wei Ji, Wen Li, Wen Sun, Xuan Wen, Yong Ren, Yuankai Ma, Yufan Lu, Bin Wang, Bo Li, Changxin Miao, Che Liu, Chen Xu, Dapeng Shi, Dingyuan Hu, Donghang Wu, Enle Liu, Guanzhe Huang, Gulin Yan, Han Zhang, Nie Hao, Haonan Jia, Hongyu Zhou, Jianjian Sun, Jiaoren Wu, Jie Wu, Jie Yang, Jin Yang, Junzhe Lin, Kaixiang Li, Lei Yang, Liying Shi, Li Zhou, Longlong Gu, Ming Li, Mingliang Li, Mingxiao Li, Nan Wu, Qi Han, Qinyuan Tan, Shaoliang Pang, Shengjie Fan, Siqi Liu, Tiansheng Cao, Wanying Lu, Wenqing He, Wuxun Xie, Xu Zhao, Xueqi Li, Yanbo Yu, Yang Yang, Yi Liu, Yifan Lu, Yilei Wang, Yuanhao Ding, Yuanwei Liang, Yuanwei Lu, Yuchu Luo, Yuhe Yin, Yumeng Zhan, and Yuxiang Zhang. Step-audio 2 technical report. *CoRR*, abs/2507.16632, 2025. doi: 10.48550/ARXIV.2507.16632. URL <https://doi.org/10.48550/arXiv.2507.16632>.

Shitao Xiao, Zheng Liu, Peitian Zhang, and Xingrun Xing. Lm-cocktail: Resilient tuning of language models via model merging. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, pp. 2474–2488. Association for Computational Linguistics, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.145. URL <https://doi.org/10.18653/v1/2024.findings-acl.145>.

LLM-Core-Team Xiaomi. Mimo-audio: Audio language models are few-shot learners, 2025. URL <https://github.com/XiaomiMiMo/MiMo-Audio>.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *CoRR*, abs/2503.20215, 2025. doi: 10.48550/ARXIV.2503.20215. URL <https://doi.org/10.48550/arXiv.2503.20215>.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengan Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jian Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report. *CoRR*, abs/2505.09388, 2025. doi: 10.48550/ARXIV.2505.09388. URL <https://doi.org/10.48550/arXiv.2505.09388>.

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.

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. *CoRR*, abs/2412.02612, 2024. doi: 10.48550/ARXIV.2412.02612. URL <https://doi.org/10.48550/arXiv.2412.02612>.

Jun Zhan, Mingyang Han, Yuxuan Xie, Chen Wang, Dong Zhang, Kexin Huang, Haoxiang Shi, DongXiao Wang, Tengtao Song, Qinyuan Cheng, Shimin Li, Jun Song, Xipeng Qiu, and Bo Zheng. Vstyle: A benchmark for voice style adaptation with spoken instructions. *CoRR*, abs/2509.09716, 2025. doi: 10.48550/ARXIV.2509.09716. URL <https://doi.org/10.48550/arXiv.2509.09716>.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pp. 15757–15773. Association for Computational Linguistics, 2023a. doi: 10.18653/V1/2023.FINDINGS-EMNLP.1055. URL <https://doi.org/10.18653/v1/2023.findings-emnlp.1055>.

Qinglin Zhang, Luyao Cheng, Chong Deng, Qian Chen, Wen Wang, Siqi Zheng, Jiaqing Liu, Hai Yu, Chao-Hong Tan, Zhihao Du, and ShiLiang Zhang. OmniFlatten: An end-to-end GPT model for seamless voice conversation. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar (eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14570–14580, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.709. URL <https://aclanthology.org/2025.acl-long.709/>.

Xin Zhang, Dong Zhang, Shimin Li, Yaqian Zhou, and Xipeng Qiu. Speechtokenizer: Unified speech tokenizer for speech large language models. *CoRR*, abs/2308.16692, 2023b. doi: 10.48550/ARXIV.2308.16692. URL <https://doi.org/10.48550/arXiv.2308.16692>.