Step-Audio-R1.5 Technical Report

StepFun-Audio Team

StepAudio R1.5 Official Github Page

Abstract

Recent advancements in large audio language models have extended Chain-of-Thought (CoT) reasoning into the auditory domain, enabling models to tackle increasingly complex acoustic and spoken tasks. To elicit and sustain these extended reasoning chains, the prevailing paradigm—driven by the success of text-based reasoning models—overwhelmingly relies on Reinforcement Learning with Verified Rewards (RLVR). However, as models are strictly optimized to distill rich, continuous auditory contexts into isolated, verifiable text labels, a fundamental question arises: **are we fostering true audio intelligence, or merely reducing a continuous sensory medium into a discrete puzzle?** We identify this as the **"verifiable reward trap."** While RLVR yields remarkable scores on standardized objective benchmarks, it systematically degrades the real-world conversational feel of audio models. By prioritizing isolated correctness over acoustic nuance, RLVR reduces dynamic interactions to mechanical "answering machines," severely compromising prosodic naturalness, emotional continuity, and user immersion, particularly in long-turn dialogues. To bridge the gap between mechanical objective verification and genuine sensory empathy, we introduce Step-Audio-R1.5. Marking a paradigm shift toward Reinforcement Learning from Human Feedback (RLHF) in audio reasoning. Comprehensive evaluations demonstrate that Step-Audio-R1.5 not only maintains robust analytical reasoning but profoundly transforms the interactive experience, redefining the boundaries of deeply immersive long-turn spoken dialogue.

1 Introduction

Chain-of-Thought reasoning has substantially advanced large language models. By decomposing complex problems into explicit intermediate steps, models such as OpenAI o1 [1] and DeepSeek-R1 [2] have achieved human-level performance in mathematical olympiads, competitive programming, and scientific inquiry. Central to this progress is Reinforcement Learning with Verified Rewards (RLVR) [2], a training paradigm that reinforces extended reasoning chains using binary, automatically checkable correctness signals, thereby bypassing the need for a learned reward model.

Efforts to transplant this recipe into the auditory domain are accelerating. A growing body of large audio language models [3–5] applies CoT reasoning to speech, music, and environmental sound, with early RLVR-trained variants [6–8] reporting strong results on objective tasks such as speech question answering and acoustic scene tagging. However, these benchmarks share a critical structural limitation: the temporally extended audio input is ultimately reduced to a single

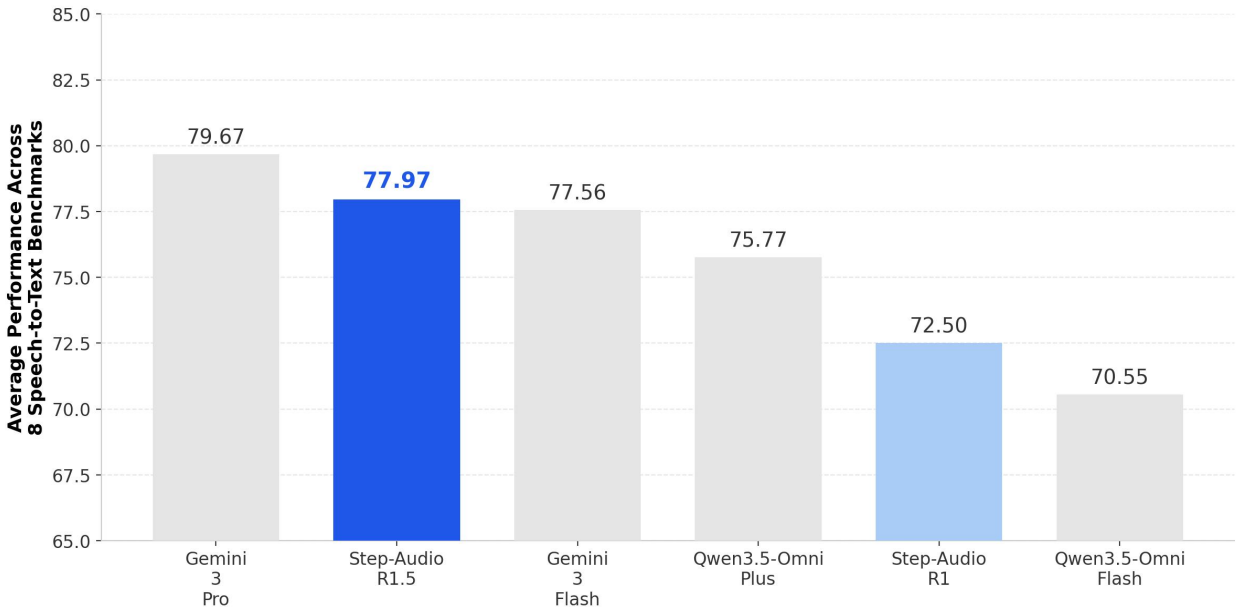

Figure 1: **Aggregate Performance across Speech-to-Text Benchmarks.** The average score represents the holistic capabilities of each model computed over 8 distinct reasoning and perception benchmarks, including Audio MultiChallenge, Big Bench Audio, MMSU, MMAU, Spoken MQA, Step-Caption, Step-DU, and Step-SPQA. Step-Audio-R1.5 substantially outperforms its predecessor and remains highly competitive with state-of-the-art commercial systems such as Gemini 3 Pro.

discrete label—a category, a number, or a short factual string. Consequently, RLVR can only reward the model for producing that specific label, leaving it structurally blind to prosodic naturalness, emotional continuity, and conversational coherence. We term this the **verifiable reward trap**: the optimization objective strictly selects for isolated answer accuracy while ignoring the nuanced qualities that determine user experience in real-world deployment.

The empirical consequence of this trap is consistent and reproducible. Under prolonged RLVR training, models become increasingly accurate on held-out test sets yet increasingly unnatural to interact with; responses grow terse, mechanical, and emotionally flat. In multi-turn spoken dialogues, where users expect not merely correct answers but genuine conversational flow, the model often degenerates into a literal "answering machine"—technically accurate but experientially hollow. This stems from a fundamental mismatch between what RLVR optimizes (**what to say**) and what users value (**how to say it**). While factual correctness is a necessary condition, it is not sufficient for high-quality audio interaction.

To bridge this gap, we introduce **Step-Audio-R1.5**, which complements RLVR with Reinforcement Learning from Human Feedback (RLHF). Rather than relying solely on binary correctness checks, we train a reward model on holistic human preference judgments over end-to-end interactions. This approach distills correctness, fluency, and emotional resonance into a unified supervisory signal, enabling the policy to escape the reward trap and optimize for overall response quality rather than isolated factual accuracy.

Comprehensive evaluations confirm that Step-Audio-R1.5 preserves the analytical reasoning cultivated by RLVR while substantially improving multi-turn interaction quality. On traditional reasoning benchmarks, the model remains highly competitive. We further evaluate its conversational capabilities on the AudioMultiChallenge [9] benchmark, which rigorously tests four key dimen-

sions of spoken dialogue — Inference Memory, Instruction Retention, Self Coherence, and Voice Editing — under naturalistic multi-turn conditions. In this demanding setting, Step-Audio-R1.5 demonstrates robust capabilities that rival or exceed those of leading commercial systems such as Gemini-2.5-Flash in key interaction dimensions. To our knowledge, Step-Audio-R1.5 is the first audio reasoning model to systematically integrate RLHF, demonstrating that the verifiable reward trap is not an inherent limitation of audio CoT, but an artifact of an impoverished reward signal that human feedback can effectively resolve.

2 Architecture

Building upon the structural foundation established by Step-Audio-R1, Step-Audio-R1.5 employs a streamlined architecture explicitly tailored for extended audio-based reasoning. The model comprises three primary components: an audio encoder, an audio adaptor, and a Large Language Model (LLM) decoder.

The acoustic front-end utilizes the Qwen2 audio encoder [10], which is extensively pretrained on diverse speech and audio understanding tasks. Operating at a frame rate of 25 Hz, the encoder is kept strictly frozen throughout the training pipeline to preserve its robust auditory perception. To bridge the continuous acoustic modality with the discrete textual space, an audio adaptor applies a temporal downsampling rate of 2. This effectively compresses the latent representations to 12.5 Hz, mitigating sequence length explosion [11–14] during complex, multi-turn interactions.

The core reasoning engine is an LLM decoder initialized from Qwen2.5 32B [15]. It directly ingests the downsampled audio features to generate purely textual outputs. To support sophisticated Chain-of-Thought (CoT) reasoning, the generation process is structurally partitioned: the decoder is prompted to first synthesize explicit intermediate reasoning traces before auto-regressively generating the final reply. This decoupling of internal analysis and external response is critical, as it forms the architectural basis for seamlessly integrating Reinforcement Learning from Human Feedback (RLHF).

3 Training Method

3.1 Audio-Centric Mid-Training

Given the base audio-language model π_{θ_0}, we perform an audio-centric mid-training stage to strengthen audio understanding, audio-grounded reasoning, and general deliberative capability before post-training alignment. The training objective combines audio-grounded reasoning data with auxiliary text-only reasoning data under a unified supervised objective:

$$\mathcal{L}_{\text{mid}} = \mathbb{E}_{(x,q,r,y)\sim\mathcal{D}_{\text{audio}}} \left[\log \pi_\theta(r, y \mid x, q)\right] + \mathbb{E}_{(q,r,y)\sim\mathcal{D}_{\text{text}}} \left[\log \pi_\theta(r, y \mid q)\right], \tag{1}$$

where (x, q, r, y) denotes audio-grounded samples with input audio x, associated textual context q, reasoning trace r, and response y, while (q, r, y) denotes text-only samples with context q, reasoning trace r, and response y. Audio-grounded supervision is drawn from diverse, high-quality audio-centric data, allowing the model to build broad perceptual coverage and robust reasoning capability over acoustically grounded contexts. Complementarily, auxiliary text-only supervision provides high-quality reasoning traces and long-form deliberative structure, facilitating the transfer of these reasoning patterns to audio-grounded understanding and inference.

3.2 Cold-start Supervised Fine-tuning

We perform a cold-start supervised fine-tuning (SFT) stage to initialize the model for interaction-oriented alignment. Although mid-training improves audio-domain knowledge, perceptual capability, and general reasoning ability, it does not directly optimize the model for high-quality multi-turn interaction. As a result, strong audio understanding alone is insufficient to ensure natural, coherent, and instruction-sensitive dialogue behavior.

Rather than further expanding domain knowledge, cold-start SFT provides a supervised initialization for interaction-oriented behavior prior to preference-based optimization. Concretely, this stage emphasizes four aspects of interaction behavior: *(1) multi-turn dialogue continuity*, the ability to maintain context and user constraints across turns; *(2) instruction following*, the ability to respond consistently under user-specified requirements on content, format, and style; *(3) response naturalness*, the ability to produce coherent and conversationally appropriate responses; and *(4) interaction awareness*, the ability to respond robustly to follow-up questions, clarification requests, interruptions, and user-side revisions.

To support these objectives, cold-start SFT is constructed from instruction-rich, multi-turn conversational data that encourages the model to organize responses in a user-oriented manner rather than as isolated task outputs. This stage provides a stronger conversational initialization for the subsequent RLHF stage, allowing preference optimization to focus on refining holistic interaction quality rather than correcting basic dialogue behavior.

3.3 RLHF with Rubric-based Generated Reward Model

Multi-turn spoken interaction exhibits substantially heterogeneous optimization targets. Some behaviors are governed by explicit and localized constraints, such as content requirements, formatting specifications, persona settings, and instruction retention across turns. Others are inherently preference-driven and only weakly specifiable, including conversational naturalness, coherence under follow-up interaction, appropriateness of tone, and overall dialogue fluency. These objectives differ not only in form, but also in how they should be evaluated: some admit relatively clear criteria, whereas others are better captured through comparative preference judgments over complete responses.

To accommodate this heterogeneity, we adopt a unified RLHF framework based on a generated reward model that jointly supports rubric-guided evaluation and ordinary preference comparison. For samples with explicit evaluation criteria, the reward model conditions on task-specific rubrics to assess whether the response satisfies the intended requirements. For samples without such criteria, the model instead performs standard pairwise preference judgment against a reference response. Formally, let $\mathcal{H}_{1:T} = \{h_t\}_{t=1}^{T}$ denote a multi-turn dialogue history up to turn T, where each h_t represents the full interaction context at turn t. Given $\mathcal{H}_{1:T}$, a policy response y, a reference response y^{ref}, and an optional rubric c, the generated reward model produces a relative quality judgment

$$g = \mathcal{R}(\mathcal{H}_{1:T}, y, y^{\mathrm{ref}}; c), \qquad c \in \mathcal{C} \cup \{\varnothing\}, \tag{2}$$

where $c = \varnothing$ corresponds to ordinary pairwise preference comparison, while $c \neq \varnothing$ denotes rubric-conditioned evaluation. The judgment g is then mapped to a scalar reward

$$r = \phi(g), \tag{3}$$

which is used for subsequent policy optimization. We optimize the policy by maximizing a PPO-style objective,

$$\mathcal{L}_{\mathrm{RLHF}}(\theta) = \mathbb{E}_t \left[\min \left(\rho_t(\theta) \hat{A}_t,\ \mathrm{clip}(\rho_t(\theta), 1-\epsilon, 1+\epsilon)\, \hat{A}_t \right) \right] - \beta\, D_{\mathrm{KL}}(\pi_\theta(\cdot \mid \mathcal{H}_{1:T}, c) \,\|\, \pi_{\mathrm{ref}}(\cdot \mid \mathcal{H}_{1:T}, c)) \tag{4}$$

where

$$\rho_t(\theta) = \frac{\pi_\theta(y_t \mid \mathcal{H}_{1:T}, c)}{\pi_{\theta_{\mathrm{old}}}(y_t \mid \mathcal{H}_{1:T}, c)}, \tag{5}$$

$\hat{A}_t$ is the advantage estimated from the generated reward, and π_{ref} denotes the reference policy used for regularization. These two forms of supervision are optimized jointly rather than in separate stages, since their optimization directions can differ substantially; empirically, decoupled training tends to induce non-trivial forgetting, where later optimization on one interaction regime degrades behaviors acquired in the other. Joint optimization therefore provides a more stable route to aligning both instruction-sensitive and preference-sensitive aspects of multi-turn dialogue within a single policy.

Within this unified RLHF framework, supervision is instantiated through a generated reward model based on relative comparison. Instead of assigning an absolute quality score to each response, the reward model compares the policy response against a reference response under the same multi-turn dialogue context and produces a preference judgment according to their comparative quality. This relative reward formulation is better suited to spoken dialogue alignment, where many important aspects of interaction quality are difficult to calibrate with a single absolute score. By representing reward as a fine-grained relative preference signal with multiple ordinal levels, the model can capture different degrees of response quality beyond binary distinction, yielding a more discriminative supervision signal for policy optimization.

4 Evaluation

4.1 Benchmarks

To comprehensively evaluate Step-Audio-R1.5's reasoning and perception capabilities, we employ a suite of speech-to-text (S2T) benchmarks. S2T evaluation isolates the model's ability to understand and reason over acoustic signals by requiring text-based responses, enabling direct comparison with state-of-the-art large language models.

AudioMultiChallenge (Audio MC). AudioMultiChallenge [9] is a multi-turn benchmark that evaluates spoken dialogue systems on natural human interaction patterns, including interruptions, hesitations, and mid-utterance repairs. It measures performance across four dimensions: Inference Memory, Instruction Retention, Self Coherence, and Voice Editing, providing a comprehensive assessment of a model's ability to handle long-context dialogue, follow instructions over multiple turns, and maintain consistency under real-world conversational noise.

Step-Caption. Step-Caption is a newly proposed benchmark designed to evaluate the model's fine-grained audio description capability. The test set consists of 905 carefully curated audio samples sourced from YouTube and Bilibili, covering both single-speaker and multi-speaker scenarios primarily in Chinese and English. Each sample is annotated by human experts across 16 dimensions, including gender, age, speaking rate, rhythm, pitch, timbre, emotion, accent, and other paralinguistic features. The model is required to generate a natural language paragraph that comprehensively

describes the speaker's vocal characteristics, with the prompt explicitly requesting analysis of all 16 dimensions. This benchmark specifically measures the model's ability to perceive and articulate acoustic attributes such as timbre, age, gender, and emotional state from raw audio.

Step-Dialogue-Understanding (Step-DU). While Step-Caption focuses on a comprehensive acoustic description, Step-Dialogue-Understanding evaluates the model's ability to answer specific questions about paralinguistic features in a conversational context. The test set consists of 87 samples recorded by diverse speakers, each directly asking about their own vocal characteristics, such as age, gender, speaking rate, or rhythm. The model must infer the correct answer solely from the acoustic signal, testing its perception and reasoning of paralinguistic cues in an interactive dialogue setting.

StepEval-Audio-Paralinguistic (Step-SPQA). StepEval-Audio-Paralinguistic was originally introduced as an AQAA (Audio Query–Audio Answer) benchmark in Step-Audio 2 [5]. To ensure consistent text-based evaluation across all models in this work, we have converted it to the AQTA (Audio Query–Text Answer) format, while preserving the original audio understanding tasks.

Additional Public Benchmarks. In addition to the proposed benchmarks, we also report results on widely adopted public benchmarks to enable broad comparison with existing models. These include MMSU [16] and MMAU [17] for expert-level audio understanding and reasoning, Big Bench Audio[1] for complex multi-step logical reasoning from audio, and Spoken MQA [18] for mathematical reasoning with verbally expressed problems.

4.2 Experimental Results

To ensure a fair and consistent comparison, we evaluated all baseline models using their official APIs through our own unified evaluation framework, rather than relying on previously reported numbers. This approach guarantees that all results are directly comparable under identical conditions. The baseline models include the Gemini family (Gemini 3 Flash and Gemini 3 Pro)[2] and the Qwen family (qwen3.5-omni-flash and qwen3.5-omni-plus) [19].

Table 1: Performance comparison on speech-to-text benchmarks. Avg. is calculated over all benchmarks for each model. Best results in bold, second-best underlined.

Model	**Avg.**	**Audio MC**	**Big Bench**	**MMSU**	**MMAU**	**Spoken MQA**	**Step-Caption**	**Step-DU**	**Step-SPQA**
Gemini 3 Flash	77.56	<u>56.42</u>	96.80	76.64	75.90	95.37	65.12	80.46	73.80
Gemini 3 Pro	**79.67**	**66.37**	**99.40**	**83.70**	**79.80**	**96.56**	**75.55**	72.41	63.60
qwen3.5-omni-flash	70.55	25.44	59.59	72.50	77.20	93.39	73.57	<u>83.91</u>	78.80
qwen3.5-omni-plus	75.77	39.38	73.03	<u>82.74</u>	<u>79.60</u>	<u>96.03</u>	<u>74.93</u>	**85.63**	<u>74.80</u>
Step-Audio-R1	72.50	24.61	98.29	75.68	77.00	95.06	70.60	64.37	74.36
Step-Audio-R1.5	<u>77.97</u>	41.15	<u>98.30</u>	79.03	77.90	93.74	71.48	82.76	**79.40**

As shown in Table 1, Step-Audio-R1.5 achieves an average score of 77.97, ranking second among all evaluated models and demonstrating competitive performance against much larger proprietary models. Notably, despite having only 32B parameters, Step-Audio-R1.5 attains 41.15 on Audio MC,

[1] https://huggingface.co/datasets/ArtificialAnalysis/big_bench_audio

[2] https://blog.google/technology/developers/gemini-3-pro-vision/

a highly competitive result that trails only the Gemini family models. Across all benchmarks, Step-Audio-R1.5 maintains balanced performance, achieving a significant average score improvement of 5.47 points over its predecessor Step-Audio-R1 (72.50). This gain is primarily driven by substantial advances on complex tasks requiring multi-turn and long-context understanding, as evidenced by the Audio MC benchmark, while its performance on perceptual benchmarks also shows broad improvements: substantial gains on Step-DU (+18.39) and Step-SPQA (+5.04), alongside a modest gain on Step-Caption (+0.88). These results collectively validate the effectiveness of our architecture and training pipeline.

5 Conclusion

The mechanical, emotionally flat responses observed in early audio reasoning models are not an inherent limitation of the Chain-of-Thought process, but rather an artifact of the verifiable reward trap. In this work, we demonstrate that heavily optimizing for isolated semantic correctness via RLVR structurally blinds models to the multidimensional nuances of genuine human interaction. Step-Audio-R1.5 breaks this trade-off by systematically integrating Reinforcement Learning from Human Feedback (RLHF), leveraging a decoupled generation architecture and a rubric-guided preference reward model. By realigning the optimization objective from merely *what to say* to holistically *how to say it*, Step-Audio-R1.5 substantially improves multi-turn conversational quality while preserving analytical rigor. This work provides a critical insight for the evolution of audio language models: as acoustic understanding matures, the next frontier of artificial audio intelligence lies not in reducing continuous sensory inputs to discrete factual puzzles, but in aligning model behavior with the rich, empathetic dynamics of natural spoken dialogue.

6 Contributors

Core Contributors: Yuxin Zhang[1,4], **Xiangyu Tony Zhang**[3], **Daijiao Liu**[1,3], **Fei Tian**[1,*,†], **Yayue Deng**[1], **Jun Chen**[1], **Qingjian Lin**[1]

Contributors: Haoyang Zhang[1,2], **Yuxin Li**[1,2], **Jinglan Gong**[1], **Yechang Huang**[1], **Liang Zhao**[1], **Chengyuan Yao**[1], **Hexin Liu**[2], **Eng Siong Chng**[2], **Xuerui Yang**[1], **Gang Yu**[1], **Xiangyu Zhang**[1], **Daxin Jiang**[1]

[1]StepFun [2]Nanyang Technological University [3]University of New South Wales [4]Shanghai Jiao Tong University

[*]Corresponding authors: `tianfei@stepfun.com` [†]Project Leader

References

[1] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024.

[2] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Peiyi Wang, Qihao Zhu, Runxin Xu, Ruoyu Zhang, Shirong Ma, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

[3] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

[4] Dong Zhang, Gang Wang, Jinlong Xue, Kai Fang, Liang Zhao, Rui Ma, Shuhuai Ren, Shuo Liu, Tao Guo, Weiji Zhuang, et al. Mimo-audio: Audio language models are few-shot learners. *arXiv preprint arXiv:2512.23808*, 2025.

[5] Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*, 2025.

[6] Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

[7] Andrew Rouditchenko, Saurabhchand Bhati, Edson Araujo, Samuel Thomas, Hilde Kuehne, Rogerio Feris, and James Glass. Omni-r1: Do you really need audio to fine-tune your audio llm? In *IEEE Automatic Speech Recognition and Understanding Workshop*, 2025.

[8] Fei Tian, Xiangyu Tony Zhang, Yuxin Zhang, Haoyang Zhang, Yuxin Li, Daijiao Liu, Yayue Deng, Donghang Wu, Jun Chen, Liang Zhao, et al. Step-audio-r1 technical report. *arXiv preprint arXiv:2511.15848*, 2025.

[9] Advait Gosai, Tyler Vuong, Utkarsh Tyagi, Steven Li, Wenjia You, Miheer Bavare, Arda Uçar, Zhongwang Fang, Brian Jang, Bing Liu, and Yunzhong He. Audio multichallenge: A multi-turn evaluation of spoken dialogue systems on natural human interaction, 2025. URL `https://arxiv.org/abs/2512.14865`.

[10] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

[11] Tri Dao. FlashAttention-2: Faster attention with better parallelism and work partitioning. In *International Conference on Learning Representations (ICLR)*, 2024.

[12] Xiangyu Zhang, Qiquan Zhang, Hexin Liu, Tianyi Xiao, Xinyuan Qian, Beena Ahmed, Eliathamby Ambikairajah, Haizhou Li, and Julien Epps. Mamba in speech: Towards an alternative to self-attention. *IEEE Transactions on Audio, Speech and Language Processing*, 2025.

[13] Xiangyu Zhang, Jianbo Ma, Mostafa Shahin, Beena Ahmed, and Julien Epps. Rethinking mamba in speech processing by self-supervised models. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

[14] Xiangyu Zhang, Daijiao Liu, Tianyi Xiao, Cihan Xiao, Tünde Szalay, Mostafa Shahin, Beena Ahmed, and Julien Epps. Auto-landmark: Acoustic landmark dataset and open-source toolkit for landmark extraction. In *Proc. Interspeech 2025*, pages 4263–4267, 2025.

[15] Qwen Team et al. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*, 2(3), 2024.

[16] 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. *arXiv preprint arXiv:2506.04779*, 2025.

[17] 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.

[18] Chengwei Wei, Bin Wang, Jung-jae Kim, and Nancy F Chen. Towards spoken mathematical reasoning: Benchmarking speech-based models over multi-faceted math problems. *arXiv preprint arXiv:2505.15000*, 2025.

[19] Qwen Team. Qwen3. 5-omni technical report. *arXiv preprint arXiv:2604.15804*, 2026.