Omni-R1: Do You Really Need Audio to Fine-Tune Your Audio LLM?

Andrew Rouditchenko[1], Saurabhchand Bhati[1,*], Edson Araujo[2,*],
Samuel Thomas[3,4], Hilde Kuehne[2,4,5], Rogerio Feris[3,4], James Glass[1]
[1]*MIT CSAIL* [2]*Goethe University of Frankfurt*
[3]*IBM Research AI* [4]*MIT-IBM Watson AI Lab* [5]*Tuebingen AI Center/University of Tuebingen*
roudi@mit.edu

***Abstract*—We propose Omni-R1 which fine-tunes a recent multi-modal LLM, Qwen2.5-Omni, on an audio question answering dataset with the reinforcement learning method GRPO. This leads to new State-of-the-Art performance on the recent MMAU benchmark. Omni-R1 achieves the highest accuracies on the sounds, music, speech, and overall average categories, both on the Test-mini and Test-full splits. To understand the performance improvement, we tested models both with and without audio and found that much of the performance improvement from GRPO could be attributed to better text-based reasoning. We also made a surprising discovery that fine-tuning without audio on a text-only dataset was effective at improving the audio-based performance.**

***Index Terms*—Audio Large Language Models (LLMs)**

I. Introduction

Reinforcement Learning (RL) has recently been shown to improve the reasoning capabilities of Large Language Models (LLMs) [1]. We are motivated by these advancements to improve the capabilities of Audio LLMs - models which take in audio input and text and can perform tasks such as question answering. Building on Qwen2.5-Omni [2], a State-of-the-Art (SOTA) multi-modal LLM, we introduce Omni-R1 using a fine-tuning pipeline based on the RL method Group Relative Policy Optimization (GRPO) [3]. Using a simple prompt, our model forgoes complex chain-of-thought or structured reasoning outputs and instead directly outputs answer choices. We use the recent MMAU [4] benchmark to test our models. Fine-tuning on the audio and human-annotated questions from the AVQA [5] dataset boosts Qwen2.5-Omni's average accuracy on MMAU Test-mini from 65.9% to 68.6%, and on the Test-full split from 68.4% to 70.8%.

We also propose to automatically generate audio question answering datasets. We prompt ChatGPT with several audio captions from an audio LLM to generate questions and answer choices for the 40k audios in AVQA and 182k audios in VGGSound. Scaling up the training data with our automatically generated questions results in further gains, achieving new SOTA of 71.3% on Test-mini and 71.2% on Test full.

In probing where these improvements arise, we made two surprising discoveries. First, text-only fine-tuning - dropping all audio inputs and training solely on question–answer text - results in significant gains in audio performance (e.g., Qwen2.5-Omni's Test-mini average improves from 65.9% to 68.2% after fine-tuning on text-only science questions). Second, evaluation with audio withheld at inference shows that much of the performance boost originates from improved text-based reasoning rather than enhanced audio processing.

In summary, our contributions are:

- Omni-R1, a streamlined GRPO fine-tuning on Qwen2.5-Omni that achieves new SOTA on MMAU without any complex prompts or explicit reasoning.
- Automatically generated audio question answering datasets which scale question–answer pairs across 182k VGGSound clips to further boost performance.
- Analysis of text-only fine-tuning, demonstrating that improving an audio LLM's text reasoning yields larger-than-expected gains on audio benchmarks.

We plan to release our code, models, and datasets publicly.

II. Related Work

The MMAU benchmark [4] is a recent large-scale audio question answering dataset designed to test audio LLMs. It is a fixed answer choice dataset providing one correct answer and three incorrect answers for each question. It contains questions about sounds, speech, and music of varying difficulty. Some of the questions require external world-based knowledge that is not provided in the questions themselves.

GRPO is a Reinforcement Learning (RL) method proposed for instilling better reasoning capabilities in LLMs [1], [3]. Towards better Audio LLMs, R1-AQA [6] proposes to use GRPO by fine-tuning Qwen2-Audio [7] on the AVQA [5] dataset of audio question and answers. This method achieved the previous SOTA on MMAU, and our approach is inspired by them. Using GRPO, we fine-tune Qwen2.5-Omni-7B [2], a recent multi-modal LLM proposed to handle both audio and video inputs, on AVQA. Doing so, we were able to achieve the new SOTA on MMAU. Further, we propose to automatically generate audio question and answer training data, which leads to even better performance.

SARI [8] is a concurrent method which fine-tunes Qwen2.5-Omni using RL. However, their setup is more complex than ours: they fine-tune the model with a schedule of Supervised Fine-Tuning (SFT) and RL, and use both structured reasoning and unstructured reasoning. In contrast, we have a much

*Equal Contribution.

simpler pipeline: we only fine-tune the model with RL, and we don't use any explicit reasoning.

Finally, we investigated how GRPO improves the Audio LLMs' performance by testing them with text-only inputs. While recent work already tested audio LLMs with text-only inputs [4], [9], we go a step further and fine-tune the models without audio (just text). We made a surprising discovery that fine-tuning the models with text-only inputs could work almost as well as fine-tuning with audio. It explains that much of the performance improvement from GRPO is due to improving the text-only reasoning capabilities.

III. Method

A. Audio LLM and Prompt

We propose Omni-R1 which fine-tunes Qwen2.5-Omni-7B [2] using GRPO. During fine-tuning and inference, we use the prompt: *"`<question>` Please choose the answer from the following options: `<choices>`"*. We chose this prompt since it is simple and the model learns to directly output one of the answer choices. This is especially useful since we have limited GPU memory. Thanks to this simple prompt, we were able to perform full-finetuning on GPUs with only 48GB GPU memory, in contrast to recent works which perform fine-tuning on GPUs with 80GB GPU memory [6]. Although recent works [6], [8], [10] propose more complex prompts with reasoning, our simple prompt was sufficient for our models to achieve the new SOTA on MMAU.

B. Reinforcement Learning

Our training uses Group Relative Policy Optimization (GRPO) [3], [6], an adaptation of Proximal Policy Optimization (PPO) [11]. GRPO is designed to mitigate the substantial memory overhead and the challenges of training an accurate per-token value function, common in standard PPO implementations for LLMs. GRPO achieves this by eliminating the explicit value function. Instead, it derives advantage estimates by comparing outputs within a group generated for the same input prompt, effectively using the average reward of sampled responses as a baseline.

The GRPO process begins by sampling G distinct outputs, $\{o_1, o_2, \ldots, o_G\}$, for a given input question q using a recent policy version, $\pi_{\theta_{old}}$. Each output o_i is assigned a scalar reward r_i by a reward model r_ϕ. For outcome-based supervision, these rewards $\mathbf{r} = \{r_1, \ldots, r_G\}$ are normalized across the group to define the advantage. Specifically, for an output o_i, the advantage $\hat{A}_{i,t}$ is computed as:

$$\hat{A}_{i,t} = \widetilde{r}_i = \frac{r_i - \text{mean}(\mathbf{r})}{\text{std}(\mathbf{r})}, \tag{1}$$

meaning this same advantage $\widetilde{r}_i$ is applied to all tokens t within that specific output o_i.

The specific reward r_i used in Equation (1) is derived from a rule-based reward function that evaluates model responses based on correctness:

$$r_i = r_{acc} \tag{2}$$

r_{acc} is an accuracy reward: it is 1 if the model's response contains the correct answer, and 0 otherwise. We do not include any other rewards such as formatting rewards.

The learning update for the current policy π_θ involves maximizing the GRPO objective function. The expectation is taken over the distribution of input questions $P(Q)$ and the outputs sampled from the old policy for each question:

$$\mathcal{J}_{GRPO}(\theta) = \mathbb{E}_{\mathcal{D}}\left[\frac{1}{G}\sum_{i=1}^{G}\frac{1}{|o_i|}\sum_{t=1}^{|o_i|}\left\{\min\left[\rho_{i,t}\hat{A}_{i,t},\right.\right.\right.$$
$$\left.\left.\left.\text{clip}\left(\rho_{i,t}, 1-\epsilon, 1+\epsilon\right)\hat{A}_{i,t}\right] - \beta\mathbb{D}_{KLi,t}\left[\pi_\theta||\pi_{ref}\right]\right\}\right], \tag{3}$$

where $\rho_{i,t} = \frac{\pi_\theta(o_{i,t}|q,o_{i,<t})}{\pi_{\theta_{old}}(o_{i,t}|q,o_{i,<t})}$ is the probability ratio between the current policy π_θ and the policy $\pi_{\theta_{old}}$ used for sampling and $\mathcal{D} = [q \sim P(Q), \{o_i\}_{i=1}^{G} \sim \pi_{\theta_{old}}(O|q)]$ denote the sampling distribution over questions and their corresponding outputs. The $\text{clip}(x, min\_val, max\_val)$ function limits the value of x to be within the range $[min\_val, max\_val]$. The hyperparameter ϵ defines the PPO clipping range. The per-token KL divergence $\mathbb{D}_{KLi,t}\left[\pi_\theta||\pi_{ref}\right]$ between the current policy π_θ and a reference policy π_{ref} is weighted by the hyperparameter β:

$$\mathbb{D}_{KLi,t}\left[\pi_\theta||\pi_{ref}\right] \approx \frac{\pi_{ref}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - \log\frac{\pi_{ref}(o_{i,t}|q,o_{i,<t})}{\pi_\theta(o_{i,t}|q,o_{i,<t})} - 1. \tag{4}$$

C. Automatic Q/A Dataset Generation with ChatGPT

We propose to generate audio question answer datasets automatically. Given an audio dataset, we first acquire audio-based text captions from an audio LLM. Specifically, we use the captions generated by Qwen-2 Audio [7] in the AudioSetCaps dataset [17], which is a compilation of captions generated for many open-source audio datasets. We use multiple captions which describe the overall acoustic scene, the speech present in the audio, and the music present in the audio. Next, we use the captions and prompt ChatGPT to generate questions and answers using the provided captions. We instruct the model to generate a correct answer and three other answers which are similar but clearly incorrect. We reviewed the generated questions and answers and found the quality to be reasonable, however, we noticed that the original captions feature hallucinations especially about music not present in the audio. The generated questions could then ask about some music or sounds not actually present in the audio. We also manually randomized the order of the correct answer within the four choices to ensure a balance.

To verify the potential of such automatically generated training data, we first generated questions for the 40k audios in the AVQA training dataset and name our dataset AVQA-GPT. Using AVQA-GPT, we are able to compare the performance of our models trained on our automatically curated questions vs the original, ground truth human annotated questions. Since the two datasets share the same audio files but have different text question and answers, it would verify the quality of our

TABLE I
ACCURACIES (%) ON MMAU. SARI† CALCULATES SCORES USING AN LLM AS A JUDGE INSTEAD OF THE USUAL STRING MATCHING.

Model	Method	MMAU Test-mini (1k audios)				MMAU Test-full (9k audios)			
		Sound	Music	Speech	Avg.	Sound	Music	Speech	Avg.
Baselines:									
—	Human [4]	86.3	78.2	82.2	82.2	-	-	-	-
Audio Flamingo 2	Direct Inference [12]	61.6	**74.0**	30.9	55.5	65.1	**72.9**	40.3	59.4
Phi-4 Multi-Modal	Direct Inference [13]	61.0	52.9	52.8	55.6	-	-	-	-
GPT-4o Voice Model	Direct Inference [14]	63.4	60.8	53.2	59.1	-	-	-	-
Qwen2-Audio-7B-Instruct	Direct Inference [7]	55.0	51.0	42.0	49.2	45.9	53.3	45.9	52.5
Qwen2-Audio-7B-Instruct	Reproduced [7]	61.6	54.5	42.0	52.7	-	-	-	-
Qwen2-Audio-7B-Instruct	Zero-Shot-CoT [15]	61.9	56.3	55.3	57.8	-	-	-	-
Qwen2-Audio-7B-Instruct	CoTA [10]	60.1	64.3	60.7	61.7	-	-	-	-
Qwen2-Audio-7B-Instruct	R1-AQA [6]	68.8	64.3	63.7	65.6	69.8	61.4	62.7	64.4
Kimi-Audio	Direct Inference [16]	**73.3**	61.7	60.7	65.2	-	-	-	-
Qwen2.5-Omni-7B	Direct Inference [2]	67.9	69.2	59.8	65.6	-	-	-	-
Qwen2.5-Omni-7B	Reproduced	69.4	66.8	**61.6**	65.9	**71.6**	67.1	**66.5**	**68.4**
Qwen2.5-Omni-7B	SARI† [8]	72.8	67.2	61.3	**67.1**	-	-	-	-
Ours:									
Qwen2.5-Omni-7B	Omni-R1 (AVQA)	70.9	70.1	64.9	68.6	73.6	68.6	**70.1**	70.8
Qwen2.5-Omni-7B	Omni-R1 (AVQA-GPT)	72.4	73.1	64.3	69.9	**74.3**	70.2	69.2	**71.2**
Qwen2.5-Omni-7B	Omni-R1 (VGGS-GPT)	**73.6**	**74.3**	**66.1**	**71.3**	74.1	**70.8**	68.7	**71.2**

dataset. In Section IV-B, we show that training on our AVQA-GPT dataset can lead to even better performance than training on the AVQA dataset. Since this approach shows promise, we scaled our method to 182k audio files in the VGGSound training dataset [18] and name this dataset VGGS-GPT. Note that the audio data in AVQA is a subset of the audio data in VGGSound. Due to GPU memory limitations, we filtered the dataset by length of the question and answer pairs, and only used 54k samples from our VGGS-GPT dataset for training. As shown in Section IV-B, training on this extra data could lead to even better performance.

IV. EXPERIMENTS

A. Experimental Setup

To train our models, we use a node with 4 A6000 GPUs (48GB) and 500 GB of RAM. The batch size per GPU is 1 with gradient accumulation steps of 2 for a total effective batch size of 8. We train for 1000 steps on AVQA and AVQA-GPT and 2000 steps on VGGS-GPT. We use a learning rate of 1×10^{-6}, temperature of 1.2, 4 responses per GRPO step, and a KL coefficient β of 0.04.

B. Main Results

Table I shows the main results on MMAU [4]. For the baselines, we show the recent methods which achieve State-of-the-Art (SOTA) performance. We refer readers to MMAU [4] for the older baselines [7], [19]–[27]. Note that all of the methods were evaluated on the Test-mini split with 1k audio samples, while only a few of them evaluated on the Test-full split with 9k audio samples.

Qwen2-Audio [7] was one of the strongest baselines and several methods are based on it. For example, Zero-Shot-CoT [15] and CoTA [10] try to enhance the model's reasoning capabilities by encouraging the model to output more of the thinking process. Qwen2.5-Omni [2] is a more recent model which reaches a significantly better average score. Although Kimi-Audio [16] and Audio-Flamingo 2 [12] obtain the best scores on the sound and music categories, Qwen2.5-Omni obtains the best average score of 65.6% (65.9% via our reproduction). SARI [8] adapts Qwen2.5-Omni and improves the average score to 67.1%, which is the current SOTA. However, they calculate the score using an LLM as a judge instead of the usual method based on string matching with the correct answer, therefore their scores might not be directly comparable with the other results. On the larger Test-full set of 9k audios, Audio-Flamingo 2 performs the best on music, while Qwen2.5-Omni obtains the highest scores on the sound, speech, and overall average (68.4%) categories.

Next, we show our Omni-R1 models which fine-tune Qwen2.5-Omni with GRPO. First, compared to the base model, Omni-R1 fine-tuned on AVQA [5] improves the average performance on Test-mini from 65.9% to 68.6% and on Test-full from 68.4% to 70.8%. Next, we show the result of fine-tuning on our proposed AVQA-GPT dataset. It uses the same audio as AVQA but the questions are generated by ChatGPT using audio-based text captions from Qwen-2 Audio [17]. Omni-R1 fine-tuned on AVQA-GPT improves the

TABLE II
ABLATION OF RL FINE-TUNING WITH VS. WITHOUT AUDIO ACROSS DATASETS. ACCURACIES (%) ON MMAU MINI AT INFERENCE TIME.

Model	RL FT: w/ audio?	FT Dataset	Inference	
			w/ audio	w/o audio
Qwen2-Audio	–	–	52.7	30.5
Qwen2-Audio	✓	AVQA	**63.2**	**44.6**
Qwen2-Audio	✗	AVQA	58.8	<u>42.4</u>
Qwen2-Audio	✗	ARC–Easy	<u>60.2</u>	42.2
Qwen2.5-Omni	–	–	65.9	<u>49.3</u>
Qwen2.5-Omni	✓	AVQA	**68.6**	**51.7**
Qwen2.5-Omni	✗	AVQA	65.6	49.2
Qwen2.5-Omni	✗	ARC–Easy	<u>68.2</u>	**51.7**

average performance on Test-mini to 69.9% and on Test-full from to 71.2%. Although AVQA and AVQA-GPT use the same audio files, they use different questions. It verifies the quality of the questions generated by ChatGPT and shows that they are potentially more useful than human generated questions for fine-tuning. Finally, we show the result of fine-tuning on our proposed VGGS-GPT, a larger dataset of questions using the same question generation technique as AVQA-GPT, but scaling to more audio from the VGGSound dataset. It achieves the best average performance of **71.3%** on Test-mini and **71.2%** on Test-full. Omni-R1 also achieves the highest scores on the sound, music, and speech categories comparing all previous models both on Test-mini and Test-full. It shows that scaling the fine-tuning data can lead to improvements. Compared to the base model, Omni-R1 improved the average scores by 5.4% absolute and 2.8% absolute on Test-mini and Test-full. Finally, compared to SARI which is the most related method and also builds on the same base model, our average Test-mini scores are 4.2% absolute higher.

C. Fine-Tuning Audio LLMs without Audio

In this section, we investigate how the models improve with GRPO fine-tuning. In particular, we wanted to understand why GRPO improved Qwen2.5-Omni's performance less than Qwen2.5-Omni's. Comparing the result of fine-tuning Qwen2.5-Omni with GRPO (Omni-R1, ours) and Qwen2-Audio with GRPO (R1-AQA [6]) on AVQA, Qwen2.5-Omni's average score improved from 65.9% to 68.6% (2.7% absolute) while Qwen2-Audio's average score improved from 52.7% to 65.6% (12.9% absolute). To investigate this, we tested the models without audio to understand their text reasoning capabilities. We then fine-tune the models without audio (just text) and measure the resulting performance.

Table II shows the result of performing inference both with audio and without audio. The scores reported are the averages on MMAU Test-mini split. For inference without audio, we simply drop the audio tokens from the input. Without access to the input audio, Qwen2-Audio and Qwen2.5-Omni achieve 30.5% and 49.3% respectively. While the performance for Qwen2-Audio is around chance (25%), it is surprising that Qwen2.5-Omni performs so well without audio. It shows that many of the MMAU questions can be answered without audio and that Qwen2.5-Omni has good text-based knowledge about audio. We suspected that GRPO helped Qwen2-Audio more significantly because its text-based reasoning was worse than Qwen2.5-Omni's. Therefore, we checked the performance after fine-tuning both models with GRPO on AVQA. Qwen2-Audio's performance without audio increased significantly from 30.5% to 44.6%, while Qwen2.5-Omni's performance without audio increased only slightly from 49.3% to 51.7%. This shows that fine-tuning with GRPO significantly improved Qwen2-Audio's text-based reasoning, but not Qwen2.5-Omni since the improvement was smaller. This is reasonable since Qwen2.5-Omni's performance without audio was already high.

Since fine-tuning the models with GRPO could improve the performance for inference both with and without audio, we wondered how well it work if we fine-tuned the models *without audio* on AVQA. Therefore, we fine-tuned the models on AVQA with GRPO to answer the questions just based on the provided text questions and answers. Despite fine-tuning without audio, this strategy could significantly improve the performance for Qwen2-Audio. The performance without audio improved from 30.5% to 42.4%. Surprisingly, the performance for inference with audio improved from 52.7% to 58.8%. This is surprising because we expected that fine-tuning on just text would make the performance worse as it could lead to catastrophic forgetting with respect to handling the audio input. This actually happened for Qwen2.5-Omni: the performance slightly decreased compared to the base model. Overall, these results are consistent with our hypothesis that much of Qwen2-Audio's improvement with GRPO could be attributed to improving the text-based reasoning. Meanwhile, Qwen2.5-Omni's stronger base text knowledge leads to smaller improvements with GRPO. Note that fine-tuning *with audio* on AVQA still performed better both for inference with and without audio.

Since fine-tuning the models without audio on AVQA could improve the base model performance, we decided to fine-tune on a text-based Q/A dataset without audio. Therefore, we fine-tuned on the ARC-Easy dataset [28], which contains Q/A questions about science in the same format as MMAU (one question and four answer choices). Interestingly, fine-tuning on this dataset resulted in even better performance than fine-tuning on AVQA without audio. Notably, Qwen2.5-Omni's performance with audio after training on text-only ARC-Easy (68.2%) nearly matches the performance of training on AVQA with audio (68.6%). However, the performance was not as good as fine-tuning on AVQA with audio. This provides further evidence that the GRPO fine-tuning is mainly improving the text-based reasoning of the models.

Overall, our results show: 1.) Qwen2.5-Omni is a much stronger text-based reasoner than Qwen2-Audio 2.) Most of the improvement from fine-tuning Qwen2-Audio with GRPO can be attributed to the improvement in text-based reasoning 3.) Fine-tuning Qwen2.5-Omni with GRPO results in a smaller performance boost since the base text-based reasoning

is already strong 4.) Fine-tuning the models on text-only Q/A datasets is surprisingly effective, but fine-tuning on Q/A datasets with audio still works better.

V. Conclusion

We propose Omni-R1, an Audio LLM based on fine-tuning the multi-modal LLM Qwen2.5-Omni with GRPO for better audio question answering. We use a simple yet effective prompt which is straight to the point and makes training and testing efficient. On the MMAU benchmark, Omni-R1 sets new State-of-the-Art results across sounds, music, speech, and overall accuracy on both Test-mini and Test-full. To scale up, we also introduced two large, auto-generated audio question answering datasets (AVQA-GPT and VGGS-GPT), which further boost performance. Through our experiments of testing the models with and without audio, we showed that much of the gains come from better text-based reasoning. Surprisingly, fine-tuning with only text (no audio) also delivers strong improvements in audio question answering.

On the one hand, since multi-modal LLMs usually start with a text-only LLM and add multi-modal encoders, it's reasonable that fine-tuning Audio LLMs on text-only datasets and improving the base knowledge would be helpful for the multi-modal abilities. On the other hand, it's surprising that fine-tuning on a text-only dataset could help the audio-based performance so much. This suggests future work is necessary on better text-only datasets to help multi-modal LLMs and on curriculum training of text-only and audio-text datasets.

Finally, collecting transcribed audio or human-annotated audio can be expensive. Our findings suggest that Audio LLMs can be steered towards better capabilities using text-only data. Moreoever, our best models were fine-tuned on audio data with questions automatically generated without human annotators. These results could help reduce the costs of building intelligent audio-based agents.

References

[1] D. Guo, D. Yang, H. Zhang, J. Song, R. Zhang, R. Xu, Q. Zhu, S. Ma, P. Wang, X. Bi *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[2] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang *et al.*, "Qwen2. 5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[3] Z. Shao, P. Wang, Q. Zhu, R. Xu, J. Song, X. Bi, H. Zhang, M. Zhang, Y. Li, Y. Wu *et al.*, "Deepseekmath: Pushing the limits of mathematical reasoning in open language models," *arXiv preprint arXiv:2402.03300*, 2024.

[4] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *The Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=TeVAZXr3yv

[5] P. Yang, X. Wang, X. Duan, H. Chen, R. Hou, C. Jin, and W. Zhu, "Avqa: A dataset for audio-visual question answering on videos," in *Proceedings of the 30th ACM international conference on multimedia*, 2022, pp. 3480–3491.

[6] G. Li, J. Liu, H. Dinkel, Y. Niu, J. Zhang, and J. Luan, "Reinforcement learning outperforms supervised fine-tuning: A case study on audio question answering," *arXiv preprint arXiv:2503.11197*, 2025.

[7] Y. Chu, J. Xu, Q. Yang, H. Wei, X. Wei, Z. Guo, Y. Leng, Y. Lv, J. He, J. Lin *et al.*, "Qwen2-audio technical report," *arXiv preprint arXiv:2407.10759*, 2024.

[8] C. Wen, T. Guo, S. Zhao, W. Zou, and X. Li, "Sari: Structured audio reasoning via curriculum-guided reinforcement learning," *arXiv preprint arXiv:2504.15900*, 2025.

[9] Y. Zang, S. O'Brien, T. Berg-Kirkpatrick, J. McAuley, and Z. Novack, "Are you really listening? boosting perceptual awareness in music-qa benchmarks," *arXiv preprint arXiv:2504.00369*, 2025.

[10] Z. Xie, M. Lin, Z. Liu, P. Wu, S. Yan, and C. Miao, "Audio-reasoner: Improving reasoning capability in large audio language models," *arXiv preprint arXiv:2503.02318*, 2025.

[11] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[12] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," *arXiv preprint arXiv:2503.03983*, 2025.

[13] A. Abouelenin, A. Ashfaq, A. Atkinson, H. Awadalla, N. Bach, J. Bao, A. Benhaim, M. Cai, V. Chaudhary, C. Chen *et al.*, "Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras," *arXiv preprint arXiv:2503.01743*, 2025.

[14] Y.-X. Lin, C.-K. Yang, W.-C. Chen, C.-A. Li, C.-y. Huang, X. Chen, and H.-y. Lee, "A preliminary exploration with gpt-4o voice mode," *arXiv preprint arXiv:2502.09940*, 2025.

[15] Z. Ma, Z. Chen, Y. Wang, E. S. Chng, and X. Chen, "Audio-cot: Exploring chain-of-thought reasoning in large audio language model," *arXiv preprint arXiv:2501.07246*, 2025.

[16] D. Ding, Z. Ju, Y. Leng, S. Liu, T. Liu, Z. Shang, K. Shen, W. Song, X. Tan, H. Tang *et al.*, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025.

[17] J. Bai, H. Liu, M. Wang, D. Shi, W. Wang, M. D. Plumbley, W.-S. Gan, and J. Chen, "Audiosetcaps: An enriched audio-caption dataset using automated generation pipeline with large audio and language models," *arXiv preprint arXiv:2411.18953*, 2024.

[18] H. Chen, W. Xie, A. Vedaldi, and A. Zisserman, "Vggsound: A large-scale audio-visual dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 721–725.

[19] S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," *Advances in Neural Information Processing Systems*, 2023.

[20] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," in *International Conference on Learning Representations*, 2024.

[21] Y. Gong, A. H. Liu, H. Luo, L. Karlinsky, and J. Glass, "Joint audio and speech understanding," in *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 2023.

[22] Z. Deng, Y. Ma, Y. Liu, R. Guo, G. Zhang, W. Chen, W. Huang, and E. Benetos, "Musilingo: Bridging music and text with pre-trained language models for music captioning and query response," in *Findings of the Association for Computational Linguistics: NAACL 2024*, 2024.

[23] S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," in *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2024.

[24] ——, "M^{2}ugen: Multi-modal music understanding and generation with the power of large language models," *arXiv preprint arXiv:2311.11255*, 2023.

[25] S. Ghosh, S. Kumar, A. Seth, C. K. R. Evuru, U. Tyagi, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities," in *Conference on Empirical Methods in Natural Language Processing*, 2024.

[26] C. Tang, W. Yu, G. Sun, X. Chen, T. Tan, W. Li, L. Lu, Z. Ma, and C. Zhang, "Salmonn: Towards generic hearing abilities for large language models," in *International Conference on Learning Representations*, 2024.

[27] Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2311.07919*, 2023.

[28] P. Clark, I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord, "Think you have solved question answering? try arc, the ai2 reasoning challenge," *arXiv preprint arXiv:1803.05457*, 2018.