DCASE 2025 CHALLENGE TASK 5 TECHNICAL REPORT

Technical Report

Minjun Chen, Jun Shao, Yangyang Liu, Bo Peng, Jie Chen
Samsung Research China-Nanjing, Nanjing, China
{minjun.chen, jun.shao, yang17.liu, b.peng, ada.chen}@samsung.com

ABSTRACT

We describe our submitted systems for DCASE2025 Task 5 in this technical report: Audio Question Answering. Our proposed systems focus on training a Large Audio Language Model (LALM) with carefully curated training datasets and training sessions, based on a carefully chosen multi-modality baseline model. We choose Qwen 2.5 Omni 7B, which has shown impressive performance on audio and vision related tasks, as the base to initialize the audio encoder and LLM component of the proposed systems. We collect and transform multiple audio-text datasets for the training, the total number of samples reached 800K, covering multiple audio related tasks, such as closed Audio QA, opened Audio QA, audio caption, and audio temporal understanding and reasoning. We curated a multi-stage training procedure to help the model learning to focus on different aspects of the data, and to learn from easy to hard during the training process. In the post-training stage, we adopt different training methods, including fine-tune (SFT) and GRPO, to take advantage of their different abilities in generalization and memory. With these carefully considered designs, our model not only learn to answer the questions correctly in content, but also in the required format specified in prompts, which simplify the post-process procedure for evaluation. Our proposed systems achieve top-1 accuracy 81.3% on the DCASE Task 5 development set.

***Index Terms*—** LALM, LLM, Audio, Multi-modality

1. INTRODUCTION

In recent years, the rapid development of artificial intelligence has led to significant breakthroughs in natural language processing (NLP) and general artificial intelligence (GAI). Due to the advancements in Large Language Models (LLMs) [7, 8, 9, 11, 12, 13, 14], multi-modality LMs have also made a significant progress, such as Vision LMs [10, 19, 20, 26], Audio LMs [5, 6, 23, 27], and Omni LMs [25]. Among these developments, Large Audio Language Models (LALMs) have emerged as a transformative technology capable of understanding, generating, and manipulating audio data with human-like proficiency. LALMs extend the capabilities of traditional text-based large language models (LLMs) by incorporating multimodal learning, enabling them to process and synthesize speech, music, and environmental sounds. The core architecture of LALMs leverages deep neural networks, particularly transformer-based architectures, trained on vast datasets of audio and corresponding textual information. These trainings enables LALMs to perform tasks such as speech-to-text transcription, text-to-speech synthesis, voice cloning, and even contextual dialogue understanding in audio formats.

Even though significant progress has made in several domains, especially in the speech-text understand and conversion, such as ASR, TTS, STT, and Translation, the research of general LALMs are still lag behind LLM and Vision LM, especial in the domains of general or ambient environment sounds understood. As an open-source automatic speech recognition (ASR) model, Whisper [22] is designed to convert spoken audio into text with high accuracy and robustness. It was trained on an extensive dataset of approximately 680,000 hours of multilingual and multitask supervised data, enabling it to handle diverse languages, accents, and environments effectively. Due to its impressive performance and transformer-based architecture, several LALMs has choose it as the audio encoder, such as Qwen2 audio, Qwen 2.5 Omni, and Kimi-Audio. These LALMs combined an audio encoder with a pre-trained LLM, and may also connect an audio decoder for generating final audio output. The components of LALMs are then thoroughly trained on a vast and diversity audio-text or multi-modality datasets. These LALMs have shown powerful performance not only in speech-text related tasks, but also in general or ambient sounds understanding and reasoning.

Audio Question Answering (AQA) is task focuses on advancing question-answering capabilities in the realm of "interactive audio understanding," covering both general acoustic events and knowledge-heavy sound information within a single track. It is one of ideal application areas of LALMs. The AQA not only requires understanding and reasoning to the input audio and question, but also requires the LALMs to integrate external knowledge, such as common sense and math capability to resolve the question. Because of the generative nature of large language models, for open AQA, it may requires semantic metrics to evaluate its performance, such as BERT-Score [28], GPT-Eval [29], or even human judgment due to the answers' subjectivity. For the closed AQA, the instruction fine-tuned LALMs may generate specified answer according to the prompt, so the accuracy may be as the evaluation metric easily. The DCASE 2025 Task 5 [1, 2, 3, 4]: Audio Question Answering encourages participants to develop systems that can accurately interpret and respond to complex audio-based questions (i.e., (A), (B), or (C) option), requiring models to process and reason across diverse audio types. In this technical report, we describe our submitted systems for this task.

This technical report is organized as follows: Section 2 details the architecture of our systems, the audio-text datasets and the training procedure. In Section 3, we demonstrate the experimental setting and results of our proposed scheme.

2. PROPOSED METHOD

This section introduces the model architecture, the training datasets and the training courses arrange for our systems.

2.1. Architecture

Our systems adopt the similar structure with the popular multi-modality model, Qwen 2.5 Omni [25], except that we do not have the vision encoder, and without the audio talker component for generating audio output. As illustrated in Figure 1. It contains below three components: (1) an audio encoder that convert the Mel-spectrogram of an audio into hidden features. It is similar with the Whisper-large-v3 model, and is originally taken from Qwen2 Audio, with about 600M parameters. (2) An aligner layer for projecting the audio feature into LLM's sematic space, it is a simple linear layer. (3) An LLM model for generating response according to the concatenated model input， which based on Qwen 2.5 7B model [30]. The parameters for the whole system is about 8.3B. These three components are initialized from the pre-trained weight of Qwen 2.5 Omni 7B directly. We continue training the system with all the collected audio-text datasets for couple of stages, keeping different components be frozen or trainable during different stages.

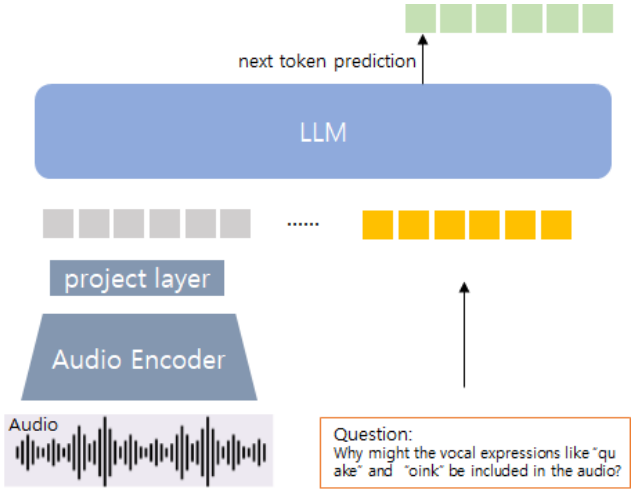

Figure1: The system architecture. The Audio encoder is a Whisper model, the aligner layer is a linear layer, and the LLM is based on QWen 2.5 LM, all of them are initialized from the corresponding components of the pre-trained weights of QWen 2.5 Omni 7B model.

2.2. Datasets

We trained and evaluated our model on the DCASE 2025 task 5 official datasets [3, 4], and collected a couple of external datasets. We details these datasets in following segments.

AVQA: a dataset [21] for audio-visual question answering on videos, which are sourced from the VGG-Sound dataset [31]. We only used the audios extracted from the videos, paired with the corresponding text questions and answers.

Clothov-AQA: is a dataset [24] for Audio question answering consisting of 1991 audio files each between 15 to 30 seconds in duration selected from the Clotho dataset. For each audio file, there are six different questions and corresponding answers, which are collected by crowdsourcing using Amazon Mechanical Turk.

Audioset-Strong: this audio dataset [18, 32] contains manually labelled audio events from AudioSet clips; each sound event is annotated with start-time and end-time timestamps. We believe these timestamps would help to improve the system's temporal reasoning capability. We apply multiple manually written templates to convert these timestamps to instruction-response format. The instructions are constructed with consideration of the occurrence sequence of different sound events and their inter-relationships, the responses are derived from the timestamps and corresponding sound event types. For each audio, we generate three samples by random chosing three templates, with each sample focus on different aspects of the temporal sound events occurring in the audio. We filter out the audios having 'music' in their labels, because the Audioset-Strong dataset does not annotate the specific music types or musical instruments.

AudioCaps: this is a large-scale dataset [16] of about 46K audio clips. All audio clips were labelled with human-written text pairs collected via crowdsourcing on the AudioSet dataset. We apply templates to convert the captions to instruction-response format, where the instruction is manually written such as "Descript the audio in details.", "What happening in the audio?", and the caption is used as the model response.

WavCaps: the captions are generated by ChatGPT, and the audio clips are extracted from several sources, include Freesound, BBC Sound Effects, SoundBible, and AudioSet Strongly-labelled Subset [34]. We convert the captions to instruction-response format using the similar method as with AudioCaps.

FSD50K: it contains over 51k audio clips (~100 hours) manually labeled using 200 classes drawn from the AudioSet Ontology [17]. For each audio clip, the caption generated by prompting ChatGPT (GPT-4) with its sound event tags. The captions are convert to instruction-response format with above method.

CompA: the CompA [27, 38], is a collection of two expert annotated benchmarks with a majority of real-world audio samples, to evaluate compositional reasoning in ALMs. CompA-order evaluates how well an ALM understands the order or occurrence of acoustic events in audio, and CompA-attribute evaluates attribute-binding of acoustic events.

TACOS: contains 12,358 audio recordings, and corresponding 47,748 temporally strong audio captions [33]. Each audio file is additionally paired with a weak caption, which was automatically generated from the strong captions using OpenAI's gpt-4o-mini-2024-07-18. We use a set of manually written templates to convert these timestamps to instruction-response format.

MMAU: is a dataset designed to evaluate multi-modal audio understanding models on tasks requiring expert-level knowledge and complex reasoning [4]. It comprises 10k carefully curated audio clips paired with human-annotated natural language questions and answers spanning speech, environmental sounds, and music. It features 27 diverse tasks, including 12 information-retrieval types and 15 reasoning types, challenging models to perform at the level of human experts in complex, multimodal audio understanding. The MMAU dataset have no answers public for the corresponding questions. Therefore, we adopt an iterative training-labeling-training procedure to utilize this dataset. We first train the model on others datasets and a tiny set of MMAU: the MMAU-test-mini, which has the answers available. Then we use the trained model to answer these questions, at the same time, another open source model Kimi-Audio [23] is used to predict the answers, we use the common answers as the labels to training the model again, and then update the labels iteratively.

For maintaining the diversity of natural language-based instructions/prompts, we use multiple templates to convert these captions to instruction-response format, and we also add a format requirement in the instruction, which ask the model output the answer in a pair of <answer></answer> tags. We filter out some audios, which are too short or too long from the datasets; those audios may harm the training stability or consume too much GPU memory. Finally we got a dataset of about 800k audio-text pair samples (some audios may appeared multiple times), covering multiple areas, such as instruction following, closed AQA, and temporal sound events understanding & reasoning and so on.

2.3. Training procedure

All the datasets are processed as a unified user-assistant chat format that is suitable for supervised training and RL (e.g., GRPO). These datasets cover a diversity areas of audio related tasks, such as audio caption, sound types identification, sound scenes classification, sound event temporal reasoning and so on. We prepared a three-stage train course for training our systems on these datasets.

Stage 1: only the parameters of LLM is set to be trainable, the parameters of other components are frozen. We apply parameter effective fine-tune (PEFT) for saving the GPU memory, all the linear layers of the LLM are updated during this stage. And all the datasets are used for this stage. Because the audio encoder and the projector layer are already trained during the training procedure of Qwen 2.5 Omni, the objective for this stage is to learn the model for answering the questions in the specified format. This would make the training of following stages and the evaluation more easy and accurate.

Stage 2: the parameters for all of the components are updated during this stage. We still apply parameter effective fine-tune (PEFT) for saving the GPU memory, and all the datasets are used during this stage. We believe by this setting, the model would learn to better extract the audio features and make a tie connection between the audio and the text (question/instruction).

Stage 3: the parameters for all of the components are updated as stage 2, but we use only the closed QA datasets. These closed QA datasets contains the questions and the corresponding options, the model should answer the question with one of the provided options. These datasets are more close with the DCASE Task5's settings, and we could apply the GRPO method [35, 36] for fine-tune the model, because the reward functions could be derived from the provided options easily. We hope the GRPO method would make the model generalize better with smaller datasets.

In the stage 3, we tried with different settings and different methods to train multiple systems. In addition to GRPO, we also applied SFT in this stage to compare the effects of different post-training methods.

3. EXPERIMENTS AND RESULT

In this section, we details the experiment settings and the results for our submitted systems in DCASE 2025 Task 5.

3.1. Data preprocessing and Hyper-parameters

The datasets are pre-processed in a similar manner as the QWen2.5-Omni. All the text (questions, options, and answers) are tokenized using Qwen's tokenizer [14], which applies byte-level byte-pair encoding with a vocabulary comprising 151,643 regular tokens. For the audios, all of the audios are resampled to a frequency of 16 kHz. We extract 128-channel Mel-spectrogram features from the raw audios with the window size setting to 25ms and the hop size setting to 10ms. The audio features are encoded as a sequence of frames by the audio encoder after then, with each frame corresponds to 40ms origin audio signal.

For the hyper-parameters, we adopt the cosine learning rate (LR) scheduler, the max LR was set to10e-5, and the warmup steps were set to 0.05 ratio of the total steps. For each stage, we trained about 3-4 epochs. The LORA rank was set to 8, and the LORA alpha was set to 32. These hyper-parameters were basicly the same during all stages.

3.2. Results

We trained multiple systems with some differences on the final training stage, e.g., training with SFT or GRPO, and some changes on the datasets arrange. Below we report the metric: top-1 accuracy, on the DCASE 2025 task 5 development set for these systems.

We designated the system, which trained only with SFT method as the system_1, and the system that is trained with GROP method in stage 3 as system_2. As according to [37], the SFT may more easily to memory the samples rather than learning the inner patterns, while GRPO could generalize better on unseen data, we designed these two different systems for comparing their differences. To validate the validation of our scheme, we also used the same datasets and procedure to learn a system_3, based on the Qwen 2 audio 7B instruct [36], and a lightweight system_4, based on the raw Qwen 2.5 omni 3B.

As can be seen from Table 1, the system_1 get the best accuracy 81.3% on the development set. We believe the development set is similar with the training set, and SFT could fit better with the training set if the training samples is big enough. The system_2 get an accuracy of 81.1%, which is lower but very close to System_1, we hope it could generalize better. The system_3 get 73.5% accuracy, and system_4 get 69.9% accuracy on the Task 5 development set. All the systems' performance get a significantly improvement than the baselines.

System	Stage 1	Stage 2	Stage 3	Acc %
Baselines				
Qwen2 Audio 7B Instruct	-	-	-	49.9
Qwen2.5 Omni 3B	-	-	-	54.6
Qwen2.5 Omni 7B	-	-	-	50.8
Ours				
system_1	SFT	SFT	SFT	81.3
system_2	SFT	SFT	GRPO	81.1
system_3	SFT	SFT	GRPO	73.5
system_4	SFT	SFT	GRPO	70.1

Table 1: The accuracy of the systems on DCASE 2025 Task 5 development set. The table details the accuracy of the four systems on the dev set, trained through three stages with different methods. We also list the accuracy for Qwen2 audio 7B instruct and Qwen 2.5 omni 7B / 3B, evaluated with the same method. Please note that in these baselines, the raw Qwen2.5 Omni 3B even get a higher score than raw Qwen2.5 Omni 7B, but after training with our datasets and methods, the systems based on Qwen2.5 Omni 7B get much higher accuracy.

The results show that the scheme we proposed is efficiency: With carefully prepared datasets, designing a progressive multi-stages training session, from easy to hard, and applying different training methods during different stages, could boost the system's performance on the target tasks.

4. REFERENCES

[1] https://dcase.community/challenge2025/.

[2] Chao-Han Huck Yang, Sreyan Ghosh, Qing Wang, Jaeyeon Kim, Hengyi Hong, Sonal Kumar, Guirui Zhong, Zhifeng Kong, S Sakshi, Vaibhavi Lokegaonkar, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha, Gunhee Kim, Jun Du, Rafael Valle, and Bryan Catanzaro. Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge. 2025. URL: https://arxiv.org/abs/2505.07365, arXiv: 2505.07365.

[3] Jaeyeon Kim, Heeseung Yun, Sang Hoon Woo, Chao-Han Huck Yang, and Gunhee Kim. Wow-bench: evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations. 2025.

[4] 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 ICLR. 2025.

[5] Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen2-audio technical report. 2024. URL: https://arxiv.org/abs/2407.10759, arXiv: 2407.10759.

[6] Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: an audio-language model with long-audio understanding and expert reasoning abilities. 2025. URL: https://arxiv.org/abs/2503.03983, arXiv: 2503.03983.

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In NeurIPS, 2020.

[8] OpenAI. ChatML, 2022. URL: https://github.com/openai/openai-python/blob/e389823ba013a24b4c32ce38fa0bd87e6bccae94/chatml.md.

[9] OpenAI. GPT4 technical report. CoRR, abs/2303.08774, 2023.

[10] OpenAI. Gpt-4v(ision) system card, 2023. URL: https://openai.com/research/gpt-4v-system-card.

[11] OpenAI. Hello GPT-4o, 2024. URL: https://openai.com/index/hello-gpt-4o/.

[12] Gemini Team. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. Technical report, Google, 2024. URL: https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf.

[13] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, Binyuan Hui, Luo Ji, Mei Li, Junyang Lin, Runji Lin, Dayiheng Liu, Gao Liu, Chengqiang Lu, Keming Lu, Jianxin Ma, Rui Men, Xingzhang Ren, Xuancheng Ren, Chuanqi Tan, Sinan Tan, Jianhong Tu, Peng Wang, Shijie Wang, Wei Wang, Shengguang Wu, Benfeng Xu, Jin Xu, An Yang, Hao Yang, Jian Yang, Shusheng Yang, Yang Yao, Bowen Yu, Hongyi Yuan, Zheng Yuan, Jianwei Zhang, Xingxuan Zhang, Yichang Zhang, Zhenru Zhang, Chang Zhou, Jingren Zhou, Xiaohuan Zhou, and Tianhang Zhu. Qwen technical report. CoRR, abs/2309.16609, 2023a.

[14] An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, et al. Qwen2 technical report. arXiv: 2407.10671, 2024a.

[15] Chao-Han Huck Yang, Yile Gu, Yi-Chieh Liu, Shalini Ghosh, Ivan Bulyko, and Andreas Stolcke. Generative speech recognition error correction with large language models and task-activating prompting. In ASRU, 1–8. IEEE, 2023.

[16] C. Dongjoo Kim, B. Kim, H. Lee, and Gunhee Kim, "AudioCaps: Generating Captions for Audios in The Wild," in NAACL-HLT, 2019.

[17] E. Fonseca, X. Favory, J. Pons, F. Font, and Xavier Serra, "FSD50K: An Open Dataset of Human-Labeled Sound Events," arXiv preprint arXiv: 2010.00475, 2020.

[18] J. F. Gemmeke, D. P. W. Ellis, D. Freedman, A. Jansen, W. Lawrence, R. C. Moore, M. Plakal, and M. Ritter, "Audio

Set: An ontology and human-labeled dataset for audio events," in Proc. IEEE Int. Conf. Acoustic., Speech and Signal Process., ICASSP, 2017.

[19] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, Jingren Zhou. Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond. URL: https://arxiv.org/abs/2308.12966, arXiv: 2308.12966.

[20] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Xuehui Wang, Yue Cao, Yangzhou Liu, Xingguang Wei, Hongjie Zhang, Haomin Wang, Weiye Xu, Hao Li, Jiahao Wang, Nianchen Deng, Songze Li, Yinan He, Tan Jiang, Jiapeng Luo, Yi Wang, Conghui He, Botian Shi, Xingcheng Zhang, Wenqi Shao, Junjun He, Yingtong Xiong, Wenwen Qu, Peng Sun, Penglong Jiao, Han Lv, Lijun Wu, Kaipeng Zhang, Huipeng Deng, Jiaye Ge, Kai Chen, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, Wenhai Wang. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. URL: https://arxiv.org/abs/2504.10479, arXiv: 2504.10479.

[21] Pinci Yang, Xin Wang, Xuguang Duan, Hong Chen, Runze Hou, Cong Jin, Wenwu ZhuAuthors. AVQA: A Dataset for Audio-Visual Question Answering on Videos. In Proceedings of the 30th ACM International Conference on Multimedia. ACM, 3480--3491(2022).

[22] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, Ilya Sutskever. Robust Speech Recognition via Large-Scale Weak Supervision. URL: https://arxiv.org/abs/2212.04356, arXiv: 2212.04356.

[23] Kimi Team. Kimi-Audio Technical Report. URL: https://arxiv.org/html/2504.18425v1, arXiv: 2504.18425v1.

[24] Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, Tuomas Virtanen. Clotho-AQA: A Crowdsourced Dataset for Audio Question Answering. URL: https://arxiv.org/abs/2204.09634, arXiv: 2204.09634.

[25] Qwen Team. Qwen2.5-Omni Technical Report URL: https://arxiv.org/pdf/2503.20215, arXiv: 2503.20215.

[26] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Qiang Liu, et al. Language is not all you need: Aligning perception with language models. arXiv: 2302.14045, 2023.

[27] Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. URL: https://arxiv.org/pdf/2406.11768, arXiv: 2406.11768.

[28] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, Yoav Artzi. BERTScore: Evaluating Text Generation with BERT. URL: https://arxiv.org/abs/1904.09675, arXiv: 1904.09675.

[29] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, Chenguang Zhu. G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. URL: https://arxiv.org/abs/2303.16634, arXiv: 2303.16634.

[30] Qwen Team. Qwen2.5 Technical Report. URL: https://arxiv.org/abs/2412.15115, arXiv: 2412.15115.

[31] Honglie Chen, Weidi Xie, Andrea Vedaldi, Andrew Zisserman. VGGSound: A Large-scale Audio-Visual Dataset. URL: https://arxiv.org/abs/2004.14368, arXiv: 2004.14368.

[32] Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, Manoj Plakal. The Benefit Of Temporally-Strong Labels In Audio Event Classification. URL: https://arxiv.org/abs/2105.07031, arXiv: 2105.07031.

[33] Paul Primus, Florian Schmid, Gerhard Widmer. TACOS: Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining. URL: https://arxiv.org/abs/2505.07609, arXiv: 2505.07609.

[34] X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, Mark D. Plumbley, Y. Zou, and W. Wang, "WavCaps: A ChatGPT-Assisted Weakly-Labelled Audio Captioning Dataset for Audio-Language Multimodal Research," arXiv preprint arXiv: 2303.17395, 2023.

[35] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y.K. Li, Y. Wu1, Daya Guo. DeepSeekMath: Pushing the Limits of Mathematical Reasoning in Open Language Models. URL: https://arxiv.org/pdf/2402.03300, arXiv: 2402.03300.

[36] Gang Li, Jizhong Liu, Heinrich Dinkel, Yadong Niu, Junbo Zhang, Jian Luan. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering. URL: https://arxiv.org/abs/2503.11197, arXiv: 2503.11197.

[37] Tianzhe Chu, Yuexiang Zhai, Jihan Yang, Shengbang Tong, Saining Xie, Dale Schuurmans, Quoc V. Le, Sergey Levine, Yi Ma. SFT Memorizes, RL Generalizes: A Comparative Study of Foundation Model Post-training. URL: https://arxiv.org/abs/2501.17161, arXiv: 2501.17161.

[38] Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Evuru, S. Ramaneswaran, S. Sakshi, Oriol Nieto, Ramani Duraiswami, Dinesh Manocha. COMPA: ADDRESSING THE GAP IN COMPOSITIONAL REASONING IN AUDIO-LANGUAGE MODELS. URL: https://openreview.net/pdf?id=86NGO8qeWs. In ICLR 2024.