TAKE IT WITH A GRAIN OF SALT: IMPROVING AUDIO QUESTION ANSWERING WITH LARGE LANGUAGE MODELS

Technical Report

Juliusz Wójtowicz-Kruk[1], *Piotr Masztalski*[1,2], *Bartłomiej Zgórzyński*[1], *Michal K. Grzeszczyk*[1]
[1] Samsung R&D Institute Poland, Warsaw, Poland
[2] AGH University of Krakow, Poland
m.grzeszczyk@samsung.com

ABSTRACT

In this report, we present our solution for DCASE 2025 Task 5: Audio Question Answering. We explore two distinct architectures: audio encoder trained with a text decoder model, and the R1-AQA model fine-tuned for the challenge tasks. Our original solution utilizes the PaSST-S audio encoder and the Qwen2.5-1.5B-Instruct model. The model was pre-trained on captioning, tagging, and question answering tasks, followed by fine-tuning for each of the three challenge tasks. The R1-AQA model underwent fine-tuning across all challenge tasks using LoRA. Through experimentation with various datasets and training methodologies, including a few-shot approach, our best trained-from-scratch model achieved an accuracy of 0.61, while the fine-tuned R1-AQA model reached an accuracy of 0.71 on the challenge development split.

***Index Terms*—** Audio Question Answering, Large Language Models

1. INTRODUCTION

Audio Question Answering (AQA) is an emerging task that combines audio comprehension with natural language understanding, leveraging the capabilities of large deep learning models. While transformer-based architectures [1] have driven remarkable advances in large language models (LLMs), commercial-grade systems with robust audio understanding remain relatively scarce. Nevertheless, a few notable approaches have begun to explore this intersection — most prominently LTU [2], SALMONN [3], and GAMA [4]. Only recently have major players in the AI space started integrating audio processing capabilities into their LLM systems. This shift is reflected in the emergence of new models such as Qwen2-Audio [5], AudioFlamingo2 [6], and Gemini 2.0 [7], which now serve as baseline for this AQA task.

For our submissions to this challenge [8] we decided to utilize two architectures: our original audio-language system employing PaSST-S audio encoder [9] trained with the frozen Qwen-2.5-1.5B-Instruct LLM model [10], and the state-of-the-art R1-AQA model [11] pretrained for audio understanding and fine-tuned to multi-question answering task. Detailed information on the architectures can be found in section 2. Comprehensive details about the data are provided in section 3. An explanation of the training steps is given in section 4, followed by a description of the submission and the final metrics in section 5.

2. ARCHITECTURE

We propose two architectures to solve the AQA task. Firstly, we develop our custom system initially used for general audio understanding and adapt it to answering multichoice questions. Secondly, we fine-tune the Large Audio-Language model (R1-AQA) on audio question answering datasets.

2.1. Original system

The original system consists of three main components: an audio encoder, a text decoder, and a connection module. The overall architecture is inspired by the LTU [2] and GAMA [4] models.

For the audio encoder, we employ PaSST-S (Patchout faSt Spectrogram Transformer Small) [9], a state-of-the-art spectrogram-based transformer model introduced in 2022 by the Institute of Computational Perception and the LIT AI Lab at Johannes Kepler University Linz. PaSST-S processes audio by converting it into spectrograms, which are then divided into patches. During training, a subset of these patches is randomly dropped using a technique called patchout, which improves generalization while reducing computational overhead. PaSST-S has been trained on large-scale audio datasets and achieves competitive results across several audio benchmarks. Including the projection layer, the encoder comprises approximately 87 million parameters and produces audio embeddings with a dimensionality of 768.

For the text decoder, we use the Qwen 2.5 Instruct model with 1.5 billion parameters [10]. Released in 2024 by Alibaba Cloud, the Qwen 2.5 series includes models ranging in size from 0.5B to 72B parameters. The selected 1.5B Instruct variant features 28 transformer layers and supports a context length of up to 32,768 tokens. This model size was chosen primarily due to computational constraints, balancing performance with efficiency to ensure feasibility within the available hardware resources.

For the connection module, we initially experimented with more complex architectures, including Q-Former module. However, we ultimately adopted a simpler design: a single linear projection layer that maps the 768-dimensional embeddings from the PaSST-S audio encoder to the 1536-dimensional input space of the Qwen 2.5 1.5B Instruct model. This decision was based on initial results — no significant performance gains were observed when using more sophisticated connection modules. We attribute this to the fact that the audio encoder remained fully trainable (unfrozen) during training, allowing the system to adapt effectively without additional architectural complexity.

2.2. R1-AQA

The R1-AQA model is a state-of-the-art AQA LLM developped by Xiaomi Reasearch [11] and is based on the Qwen2-Audio7B-Instruct architecture built by Alibaba Group [5]. The Qwen2-Audio model consists of audio encoder (Whisper-large-v3 [12]) and large language model Qwen-7B [13]. As the input, Qwen LLM receives the embedded 40ms frames of the audio signal created with the audio encoder as well as embeddings of text tokens to autoregressively predict next text tokens. The R1-AQA model is Qwen2-Audio7B-Instruct model with the Group Relative Policy Optimization (GRPO) algorithm [14] applied on the AVQA dataset [15]. GRPO is the Reinforcement Learning method and has proven to improve performance on AQA tasks in comparison to full model fine-tuning.

3. DATA

In this section we describe datasets used for the development of our models. We use several datasets for pretraining and fine-tuning our models to solve the AQA task. Subsections 3.1 through 3.7 describe the datasets employed during pretraining of our original system. Following that, subsections 3.8 to 3.10 detail the datasets used for fine-tuning both architectures.

3.1. AudioCaps

AudioCaps is a high quality audio captioning dataset with captions collected via crowdsourcing. The dataset we used for training our model consists of 43,698 audio samples with one caption and 1,293 audio samples with five captions each. Altogether, the dataset of 50,161 audio-caption pairs was used for training.

3.2. Clotho v2.1

Clotho is a high quality audio captioning dataset. The official version comprises 6,974 audio samples and 34,870 captions, with each audio sample being paired with five captions. The dataset we utilized consists of 5,925 audio samples. This includes 3,839 samples in the development split and 1,045 samples in the validation split used for training, and 1,045 samples in the evaluation split used for validation. In total, the dataset provided us with 24,420 audio-caption pairs for training and 5,225 pairs for validation.

3.3. WavCaps

WavCaps is a large-scale, weakly-labeled audio captioning dataset. The authors utilized the GPT-3.5-turbo model to process and refine raw captions for audio samples collected from various sources. Consequently, WavCaps stands as the largest open-source audio-caption dataset, comprising over 400,000 audio-caption pairs. However, the quality of the captions is notably lower compared to those found in the Clotho and AudioCaps datasets. The dataset we used for training consists of 401,112 audio-caption pairs.

3.4. AudioSet

AudioSet is possibly the largest open-source audio tagging dataset. The original version consists of 2,084,320 10-second sound segments from YouTube videos labelled with 632 audio classes. Our version of the AudioSet used for training consists of 1,834,466 audio-label pairs, and 17,795 used for validation.

3.5. VGGSound

VGGSound is a large-scale audio-visual dataset designed for audio event recognition and related tasks. It contains approximately 200,000 video clips sourced from YouTube, each lasting around 10 seconds and annotated with one of 309 sound event classes. For training, we use version of the dataset consisting of 174,574 audio-label pairs, and 14,602 pairs for validation.

3.6. FSD-50k

FSD-50K is an open-source dataset for audio event tagging. It comprises approximately 51,197 audio clips, each ranging from a few seconds to about 10 seconds in length, annotated with 200 sound classes. For training, we use version of the dataset consisting of 40,966 audio samples paired with annotated classes, and 10,231 for validation.

3.7. OpenAQA

OpenAQA is a large-scale dataset developed by the authors of the LTU model [2] specifically for training audio question answering models. The dataset was created using ChatGPT 3.5 Turbo to generate extensive question-answer pairs based on existing audio tagging and captioning datasets. The resulting corpus comprises approximately 1.9 million close-ended and 3.7 million open-ended audio-question-answer triplets. For our training, we utilized only the open-ended portion of the dataset.

3.8. Task 5 Data

The main dataset used within this challenge consists of three parts. Part 1 is a set of almost a thousand data samples sourced from the Watkins Marine Mammal Sound Database [16]. This subset requires the model to classify the species and the vocalization types. Part 2 involves 1.6 thousand samples revolving around temporal soundscapes questions like the order, number, or timestamps of sounds. The largest subset is Part 3 which includes more than 8 thousand audio samples combined with complex questions requiring reasoning over multiple levels of audio understanding [17].

3.9. MMAU

MMAU Mini is a subset of the A Massive Multi-Task Audio Understanding and Reasoning Benchmark (MMAU) dataset [17], designed for evaluating audio understanding and reasoning capabilities in multimodal models. The publicly available labelled portion includes 1,000 audio segments, each paired with human-annotated multiple-choice questions offering four possible answers. This dataset emphasizes high-level reasoning over audio content.

3.10. TACOS

Temporally-aligned Audio CaptiOnS for Language-Audio Pretraining (TACOS) [18] is a dataset specifically designed for temporal audio understanding. It consists of approximately 12,000 audio recordings from Freesound annotated with a single-sentence caption describing the audio combined with almost 48,000 temporal annotations representing onset and offset of the specific sound. We adapt this dataset to the AQA task by generating four types of questions from each caption aligned with Part 2 questions. The questions are:

- What is the start time and end time of <caption>?
- What is the onset of <caption>?
- What is the offset of <caption>?
- What is the duration of <caption>?

We randomly generate incorrect answers, ensuring that the generated timestamps are not closer than 0.3s to the correct timestamps.

4. TRAINING

This section describes training procedures of our systems. For training of the models implemented in PyTorch library we utilize a single NVIDIA A100 40GB GPU.

4.1. Original system

Training of the original system was inspired by LTU training curriculum [2] and was conducted in three stages: (1) pretraining using audio captioning and tagging datasets, (2) fine-tuning on a combination of captioning, tagging, and open-ended question answering datasets, and (3) final fine-tuning on the Task 5 challenge dataset. After experimenting with various hyperparameter settings, we settled on a unified configuration for all three stages. In each phase, the audio encoder and the connection module are left unfrozen and updated during training, while the text decoder remain fully frozen to reduce computational overhead and avoid overfitting. We use the AdamW optimizer with a batch size of 8. A cosine decay learning rate scheduler with warmup is applied, with a maximum learning rate of 1e-4 and a minimum of 1e-7. Model checkpoints are selected based on validation loss.

4.2. Fine-tuned R1

Due to compute limitations we fine-tune R1-AQA model with Low-Rank Adaptation (LoRA) method which is a parameter efficient fine-tuning approach [19]. We apply LoRA to query and value matrices and set rank to 8. This approach adapted to Qwen2-Audio yields 5.5 million trainable parameters. We fine-tune the model on a batch size of 2 for 10 epochs with 2700 steps per epoch and a cosine decay learning rate scheduler with warmup. The initial learning rate is 1e-4 with minimum learning rate set to 1e-7. When fine-tuning on TACOS dataset we increase the number of steps to 7000 due to size of the dataset. In this case the ratio of TACOS, Part 2 and Part 3 samples used for training is 4, 1, 2, respectively.

5. SUBMISSION

For the DCASE 2025 task 5: Audio Question Answering, we prepared four submissions. For submission 1, we utilized our original system consisting of the PaSST-S audio encoder and the Qwen2.5-1.5B-Instuct model. This model consists of approximately 1.6 billion parameters.

The rest of submitted systems are based on the R1-AQA model with LoRA. For submission 2, we utilized only the training dataset from task 5 to fine-tune the model. For submission 3, we additionally included samples from the MMAU dataset. Including the MMAU dataset increased the performance on the Part 1 and Part 3 subsets but decreased the accuracy on Part 2. Therefore, for submission 4, we utilized answers for Parts 1 and 3 from submission 2 and we specifically fine-tuned additional model on Part 2, Part 3 data samples and TACOS dataset tweaked for AQA task. We used this model to generate answers for Part 2. Such an approach yielded the best performance on the development dataset. The results achieved by all submitted systems are presented in Table 1. In all submissions we used the Levenshtein distance between the model output and the possible answers. The answer minimizing the Levenshtein distance was used as the output of our systems.

Submission	Part 1	Part 2	Part 3	Total accuracy
SRPOL_1	64.29	39.74	68.69	61.11
SRPOL_2	67.86	48.77	79.98	71.17
SRPOL_3	68.75	44.66	80.77	70.76
SRPOL_4	68.75	55.01	80.77	73.32

Table 1: Part 1, Part 2, Part 3 and total accuracy [%] of the proposed systems on the Task 5 development dataset.

6. REFERENCES

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," 2023. [Online]. Available: https://arxiv.org/abs/1706.03762

[2] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," 2024. [Online]. Available: https://arxiv.org/abs/2305.10790

[3] 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," 2024. [Online]. Available: https://arxiv.org/abs/2310.13289

[4] 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," 2024. [Online]. Available: https://arxiv.org/abs/2406.11768

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

[6] 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," 2025. [Online]. Available: https://arxiv.org/abs/2503.03983

[7] G. Team, R. Anil, and S. B. et al., "Gemini: A family of highly capable multimodal models," 2025. [Online]. Available: https://arxiv.org/abs/2312.11805

[8] C.-H. H. Yang, S. Ghosh, Q. Wang, J. Kim, H. Hong, S. Kumar, G. Zhong, Z. Kong, S. Sakshi, V. Lokegaonkar, O. Nieto, R. Duraiswami, D. Manocha, G. Kim, J. Du, R. Valle, and B. Catanzaro, "Multi-domain audio question answering toward acoustic content reasoning in the dcase 2025 challenge," 2025. [Online]. Available: https://arxiv.org/abs/2505.07365

[9] K. Koutini, J. Schlüter, H. Eghbal-zadeh, and G. Widmer, "Efficient training of audio transformers with patchout," in *Interspeech 2022*. ISCA, 2022. [Online]. Available: http://dx.doi.org/10.21437/Interspeech.2022-227

[10] A. Yang, B. Yang, B. Zhang, and B. H. et al., "Qwen2.5 technical report," 2025. [Online]. Available: https://arxiv.org/abs/2412.15115

[11] 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," 2025. [Online]. Available: https://arxiv.org/abs/2503.11197

[12] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. Mcleavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *Proceedings of the 40th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato, and J. Scarlett, Eds., vol. 202. PMLR, 23–29 Jul 2023, pp. 28 492–28 518. [Online]. Available: https://proceedings.mlr.press/v202/radford23a.html

[13] J. Bai, S. Bai, Y. Chu, Z. Cui, K. Dang, X. Deng, Y. Fan, W. Ge, Y. Han, F. Huang, *et al.*, "Qwen technical report," *arXiv preprint arXiv:2309.16609*, 2023.

[14] 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.

[15] 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.

[16] J. Kim, H. Yun, S. H. Woo, C.-H. H. Yang, and G. Kim, "Wow-bench: Evaluating fine-grained acoustic perception in audio-language models via marine mammal vocalizations," 2025.

[17] 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," 2024. [Online]. Available: https://arxiv.org/abs/2410.19168

[18] P. Primus, F. Schmid, and G. Widmer, "Tacos: Temporally-aligned audio captions for language-audio pretraining," *arXiv preprint arXiv:2505.07609*, 2025.

[19] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, *et al.*, "Lora: Low-rank adaptation of large language models." *ICLR*, vol. 1, no. 2, p. 3, 2022.