Enhancing Temporal Understanding in Audio Question Answering for Large Audio Language Models

Arvind Krishna Sridhar
Qualcomm Technologies Inc.
San Diego, CA
arvisrid@qti.qualcomm.com

Yinyi Guo
Qualcomm Technologies Inc.
San Diego, CA
yinyig@qti.qualcomm.com

Erik Visser
Qualcomm Technologies Inc.
San Diego, CA
evisser@qti.qualcomm.com

Abstract

The Audio Question Answering (AQA) task includes audio event classification, audio captioning, and open-ended reasoning. Recently, AQA has garnered attention due to the advent of Large Audio Language Models (LALMs). Current literature focuses on constructing LALMs by integrating audio encoders with text-only Large Language Models (LLMs) through a projection module. While LALMs excel in general audio understanding, they are limited in temporal reasoning, which may hinder their commercial applications and on-device deployment. This paper addresses these challenges and limitations in audio temporal reasoning. First, we introduce a data augmentation technique for generating reliable audio temporal questions and answers using an LLM. Second, we perform a further fine-tuning of an existing baseline using curriculum learning strategy to specialize in temporal reasoning without compromising performance on fine-tuned tasks. We demonstrate the performance of our model using state-of-the-art LALMs on public audio benchmark datasets. Third, we implement our AQA model on-device locally and investigate its CPU inference for edge applications.

1 Introduction

Multimodal Question Answering (MQA) involves generating relevant answers for multimedia inputs such as images, audio, and video, in response to user queries (Pan et al., 2024). Following the success of large pretrained transformer models for MQA, audio-specialized question answering has gained traction. Audio Question Answering (AQA) is an audio-to-text task where, given an audio file and a question, the model produces an answer by analyzing the audio content.

Audio Question Answering: Recent literature (Gong et al., 2023; Ghosh et al., 2024a; Tang et al., 2024; Deshmukh et al., 2023) in AQA develops end-to-end pretrained transformer-based architectures known as Large Audio Language Models (LALMs). Figure 1 provides a general framework for our AQA model architecture (Gong et al., 2023). It comprises three components: an audio encoder, a projection module, and a text decoder. The Audio Spectrogram Transformer (AST) (Gong et al., 2021) encodes the input audio clip into spectrogram feature representations. The projection module converts these audio feature representations into text-equivalent embeddings for the text decoder. The LLaMA model serves as the text LLM decoder, taking the converted audio feature embedding and the question as input. During training, we add metadata as an optional input that is generated by the proposed data augmentation in Section 2.1. It helps provide extra guidance to the LLM decoder along with the text projections of the audio clip and aids in the overall audio-text representation learning. The GAMA model (Ghosh et al., 2024a) follows a similar architecture to LTU (Gong et al., 2023), combining multiple types of audio features, including activations from multiple layers of AST, Audio Q-former, and a soft prompt that provides audio events information. In this paper, we intend to discuss a few problems and limitations that we discovered in the process of developing a LALM for commercial edge devices and explain our proposed techniques to overcome them. We chose LTU as the base model for our experiments over GAMA due to the ease of on-device implementation.

Use Case Motivation: Although LALMs excel at general audio understanding and have shown good overall performance in audio captioning, classification tasks, and open-ended reasoning tasks, there is a significant gap between LALM research and real-world product requirements. First, LALMs fine-tuned end-to-end with millions of audio-text samples do not capture fine-grained audio understanding well. Their performance isn't impressive on specialized reasoning tasks that require fine-grained understanding, such as temporal reasoning

Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Industry Track), pages 1026–1035

(Gong et al., 2023). Audio temporal reasoning is the ability to understand the temporal context and relationship between events in the input media. Specialized audio temporal understanding has significant potential across various sectors for commercial adoption. In healthcare, it can be used for continuous monitoring and analysis of heartbeat and respiration over a period of time and provide useful analysis and recommendations to the user. In smart homes, it can enable advanced security monitoring with privacy protection by capturing and analyzing the sequence of events in live stream audio coming from sensors located in multiple areas. (Gong et al., 2023) explains that the lack of fine-grained understanding in LALMs might be due to performing temporal downsampling at the audio encoder-projection module juncture, which is a trade-off to save computational efficiency and limited training data for temporal analysis. In this paper, we address both these limitations while also keeping in mind the limitations in commercial LALMs, including low memory footprint, ease of on-device implementation, reliability, and minimal training compute. Due to the difficulty in procuring large amounts of pretraining data, expensive compute power, and time constraints, it is painstakingly difficult to retrain an LALM from scratch for improving a particular skill. On top of that, the large memory requirements of LALMs make it difficult to run them on low-compute edge devices.

Existing Work on Temporal Reasoning in AQA: In this paper, we focus on optimal training pipeline strategies to improve audio temporal understanding. Before the pre-trained transformers era, DAQA (Fayek and Johnson, 2020) and ClothoAQA (Lipping et al., 2022) proposed a synthetic rule based and crowd sourced audio temporal reasoning datasets respectively. (Ghosh et al., 2024b) published an annotated benchmark to evaluate the audio encoders on compositional reasoning including order or occurrence of acoustic events. (Yuan et al., 2024) discuss the limitations of CLAP encoder in capturing temporal information and propose a data augmentation strategy to improve the same.

Motivation for Deploying AQA on Edge: With the large memory requirements of LALMs scaling billions of parameters, the inference becomes expensive to run on cloud GPUs (Desislavov et al., 2023). For commercial audio understanding use cases, such as smart home Internet of Things (IoT) and industrial IoT, where we can capture streams of audio from various sources such as machinery, front door, kitchen, etc., using a simple audio receiver, we need the AQA model on an always-on low-powered edge device for reasonable inference cost and preserving privacy by performing computation of audio on a self-contained edge CPU.

Contributions: To the best of our knowledge, we are the first to investigate the problem and limitations of audio temporal understanding in LALMs and address them from a commercialization perspective. Our contributions in this paper are as follows: First, we propose a data augmentation technique to reliably generate audio temporal question and answer pairs using GPT-4. Second, we show that fine-tuning the baseline checkpoint via curriculum learning helps improve the model's temporal awareness and reasoning without losing its original performance. Finally, we implement the AQA to run on CPU locally for commercial edge applications.

2 Methodology

We divide our proposed methodology into two sections. First, we explain the data augmentation strategy for generating temporal reasoning data. Second, we discuss our temporal fine tuning strategy.

2.1 Temporal Reasoning Data Augmentation

In order to improve the temporal reasoning capability of generalist LALMs, we developed a data augmentation technique that ensures the questions are intuitive to human temporal perception. We prompt GPT-4 (OpenAI et al., 2024) with the audio clip's ground truth, such as audio event labels, audio captions, and their metadata comprising timestamps of audio events. For Audioset-SL, we use audio clips containing fewer than seven types of foreground sound events with a total occurrence number of less than ten and consider sounds with a duration longer than nine seconds as background sound. We state the temporal information of the sound events in natural language and use that as part of the prompt for GPT-4 to generate question-answer pairs with rationale. In the prompt, we include few-shot paired examples of temporal reasoning tasks, including temporal relationships, chronological ordering, duration comparison, and occurrence counting. Table 1 showcases the different types of metadata-question-answer pairs generated by our temporal data augmentation technique. We refer to the data generated by this pipeline as TemporalQA. We preprocessed the question-answer pairs to filter

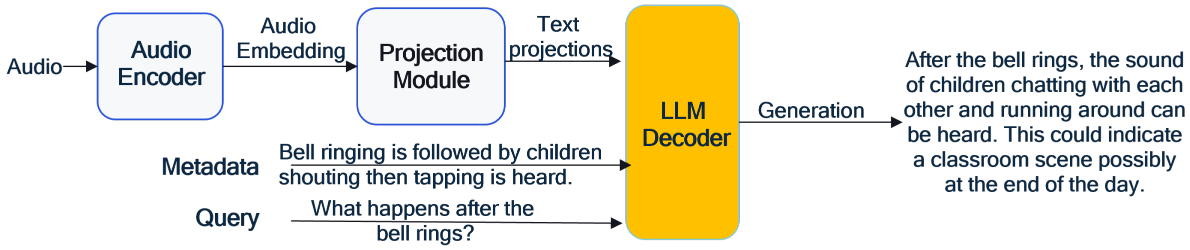

Figure 1: Our Proposed Framework for Audio Question Answering (AQA) model architecture

Figure 2: Data Generation for Temporal Reasoning using GPT-4

Type	Metadata	Question	Answer
Chronological order	The background of the audio is Music.The audio starts with Cough followed by Ratchet then Male singing.	What's the sequence of the audio events in the clip?	The chronological order of the sound events in the audio clip is as follows: First, a Cough is heard. This is immediately followed by a Ratchet sound. Finally, Male singing is heard. Hence, the sequence is cough, ratchet, and then male singing.
Counting	The audio starts with Pig followed by Mechanisms then Pig followed by Mechanisms then Sound effect followed by Pig.	How many times does the Pig oink?	The Pig oink is heard 3 times with mechanisms sound occurring in between. This order forms a pattern where the pig sound reappears intermittently.
Before/After	The background of the audio is Static.The audio starts with Male singing followed by Choir then Male singing followed by Choir.	What sequence of sound events follows the first choir singing?	Following the first choir singing, the sequence cycles back to male singing again, creating a repetitive pattern which is then followed by another choir performance.
Duration	The background of the audio is Sawing.The audio starts with Speech.	What ambient sound persists throughout the audio clip?	The ambient sound that persists throughout the audio is sawing. This backdrop of sawing is mentioned as the continuous sound in the background.
Temporal pattern detection	The background of the audio is Music.The audio starts with Child singing followed by Breathing then Child singing followed by Breathing then Child singing.	Which sound recurs after each instance of breathing?	The child's singing recurs after each instance of breathing. The pattern repeats multiple times in the audio.

Table 1: Types of temporal questions in TemporalQA

out wrongly formatted data. To assess the quality of GPT-4 generated question-answer pairs, we conducted a human evaluation to score on correctness, reasoning quality and hallucination as shown in Table 2. Correctness measure checks if the answer is correct for the given question. Reasoning quality ensures that the accompanying reason is meaningful and helps in arriving at the answer. Hallucination refers to audio events/phrases present in the answer that are not mentioned in the metadata. In the Table 2, the near perfect scores for correctness and reasoning quality and low hallucination rate of the generated question-answer pairs reflects the high quality of generated temporal reasoning data.

Metrics	Score
Correctness	4.98
Reasoning Quality	4.99
Hallucination	0.02

Table 2: Human evaluation of the GPT-4 generated question answer pairs. All the metrics score range from 0 to 5. For correctness and reasoning quality, higher score is preferred while for hallucination, a lower score is optimal.

2.2 Temporal Finetuning via Curriculum Learning

In this section, we outline the training strategy employed to integrate temporal reasoning capabilities into a Large Audio Language Model (LALM) designed and finetuned for general audio understanding. To learn temporal reasoning skill on an already finetuned AQA model, we adopt a curriculum learning approach that merges TemporalQA with a few core finetuned Audio Question Answering (AQA) tasks, including audio classification and audio captioning. We conducted an empirical investigation to determine the optimal types of AQA tasks and the appropriate ratio of new skills (temporal reasoning) to existing skills. Based on our analysis and hyperparameter tuning, we observed that a 50:50 ratio of temporal reasoning to core AQA tasks—comprising audio event tagging, audio label classification, and audio captioning—combined with a learning rate ten times lower than that of the original finetuning, is optimal for learning temporal reasoning skills without significantly compromising the model's original performance. We refer to our temporal finetuned model with and without metadata on LTU base as AQA+Temp-M and AQA+Temp, respectively.

$$T_{\text{total}} = T_{\text{temporal}} + T_{\text{core AQA}}, \quad (1)$$

Where T refers to training data and the + operation combines both operand datasets with a random shuffle. We also provide metadata of audio, such as audio events and background noise information, in natural language in the text prompt as guidance to mitigate the information bottleneck at the projection module.

3 Experiments

3.1 Datasets

We choose the LTU model (Gong et al., 2023) as our baseline. We adopt a similar training dataset accruing strategy to (Gong et al., 2023). Our initial stages of curriculum learning focus on training the audio encoder and projection model with a combination of audio event classification public datasets, including Audioset, FSD50k, VGGSound, and Freesound, and audio captioning public datasets, such as Clotho and Audiocaps (Gong et al., 2023). We use Audioset-strong labelled (Hershey et al., 2021) and FSD50k datasets to synthetically generate 20k temporal reasoning data using the data augmentation strategy explained in Sec 2.1. TemporalQA has an 80:20 train-test split. We adopt the inference style of (Gong et al., 2023), including the generation of audio descriptions for the FSD dataset. All audio clips are truncated to 10s to fit the audio encoder context window.

3.2 Experiment Setup

We train the AQA architecture from scratch with four-stage curriculum learning as described in (Gong et al., 2023). For temporal reasoning finetuning, we perform model parallelism-based distributed training on 8 A100 GPUs for 2 epochs with a learning rate of 1e-4 and cross-entropy as the loss function. We found that a batch size of 24 and a micro-batch size of 1 work best for specializing the model further on a single task, as opposed to a batch size of 256 and a micro-batch size of 16 for fine-tuning from scratch. The low-rank adaptors (LoRA) hyperparameters alpha and r are set to 16 and 8, respectively. We set α_1 and α_2 to 1 while keeping α_3 at 0 to provide equal weightage for answer and reason and for a fair comparison with conventional metrics.

3.3 On-device Implementation

To run the AQA model on CPU, we perform 16-bit and 8-bit post-training quantization as mentioned

Question	Groundtruth Answer	LTU	AQA-Temp (Ours)
Does the whispering begin after the sound of the drum kit?	Yes, the whispering starts after the drum kit is heard.	Yes, it starts at 10.2 seconds.	Yes, the whispering starts after the sound of the drum kit. This suggests that the whispering is a separate event that occurs after the initial musical performance has begun. It's possible that the whispering is a response to the music or a way for the performers to engage with the audience in a more intimate setting.
What is the sequence of sound events following the first Speech?	Following the first instance of Speech, the sequence of sound events is as follows: Male speech, then another instance of Speech, followed by Male speech, and finally, Speech again.	The sequence includes a tap, followed by generic impact sounds and then more speech. This could suggest a casual conversation or activity involving objects being used or moved around in an outdoor setting	Following the initial speech, there are two instances of male speech. This suggests that the conversation continues with multiple participants, possibly a dialogue between two or more individuals.

Table 3: Demonstration examples showcasing temporal understanding after fine-tuning. LTU and AQA-Temp refers to the baseline and temporal reasoning finetuned checkpoints.

Model	Clotho		AudioCaps		FSD		TemporalQA	
	SPIDER	FENSE	SPIDER	FENSE	SPIDER	FENSE	SPIDER	FENSE
LTU	0.19	0.56	0.31	0.67	**0.08**	**0.47**	0.27	0.57
GAMA	0.04	0.41	0.09	0.55	0.05	0.42	0.22	0.65
AQA+Temp(w/o CL)	0.04	0.46	0.04	0.49	0.06	0.33	0.52	0.64
AQA+Temp-M(w/o CL)	0.04	0.36	0.04	0.48	0.07	0.32	0.43	0.63
AQA+Temp (Ours)	0.24	0.61	0.38	0.71	0.06	0.44	0.48	0.66
AQA+Temp-M (Ours)	**0.31**	**0.62**	**0.43**	**0.73**	0.07	0.43	**0.70**	**0.73**

Table 4: Comparison of performance on LTU baseline with proposed finetuning on temporal reasoning. Temp refers to temporal finetuning and Temp-M refers to temporal finetuning with meta data information. w/o CL refers to training AQA on temporal reasoning data without curriculum learning.

in llama.cpp. We implement the AQA architecture on top of the C++ implementation of LLaMA in the llama.cpp framework. First, we merge the LoRA weights into the LLaMA model of AQA+Temp and convert the checkpoint to gguf format. Second, we implement the audio encoder and projection module in onnxruntime to combine their outputs with the LLaMA in C++. We perform the experiment to measure inference speed on 100 randomly sampled questions from our test set of AQA described in 3.1 and report the average.

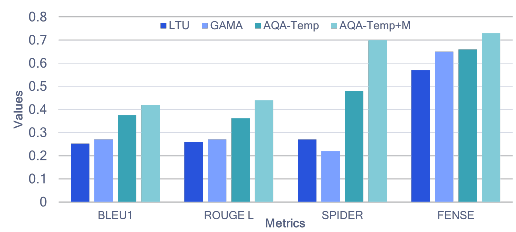

Figure 3: Barplot of LTU and GAMA baseline and temporal finetuned checkpoints for temporal dataset.

Model Name	Size	Accuracy(%)
Random Guess	-	26.72
Most Frequent Choice	-	27.02
Human (test-mini)	-	86.31
Pengi	323 M	6.1
Audio Flamingo Chat	2.2B	23.42
M2UGen	7B	3.6
LTU	7B	22.52
LTU AS	7B	23.35
MusiLingo	7B	23.12
MuLLaMA	7B	40.84
GAMA	7B	41.44
GAMA-IT	7B	**43.24**
Qwen-Audio-Chat	8.4B	55.25
Qwen2-Audio	8.4B	7.5
Qwen2-Audio-Instruct	8.4B	54.95
SALAMONN	13B	41
Gemini Pro v1.5	-	56.75
GPT4o + weak cap.	-	39.33
GPT40 + strong cap.	-	57.35
Llama-3-Instruct + weak cap.	8B	34.23
Llama-3-Instruct + strong cap.	8B	50.75
AQA+Temp (Ours)	7B	28.83
AQA+Temp-M (Ours)	7B	**32.73**

Table 5: Results on MMAU Test-Mini Sound Split

4 Results

4.1 Quantitative Analysis of Temporal Finetuning

Table 4 shows the performance of the proposed temporal fine-tuning for temporal reasoning with LTU as the base model. For a fair evaluation, during inference, we do not provide metadata to the models. After temporal fine-tuning, there is a considerable increase in all the metrics across datasets except for FSD. This might be due to differences in the format, adopted from LTU (Gong et al., 2023), of FSD dataset's groundtruth and LALM's response. FSD is an audio classification dataset while the other datasets in evaluation are natural language description based datasets. FSD has a list of audio events as label while the LALMs generate an audio caption style answer. For example, the ground truth FSD label is "Electric guitar; Guitar; Plucked string instrument; Musical instrument; Music" while the generated audio caption is "Music is playing with a plucked string instrument and a bass guitar, creating a rich and dynamic soundscape.". For a reliable accuracy, in future, we can convert the audio event labels of FSD into natural language sentence using an off-the-shelf LLM and train our LALM on uniform response format. The significant improvement of Spider and FENSE metrics for AQA+Temp-M over LTU shows that we can offset the information bottleneck at the projection layer to some extent with extra textual guidance. It is notable that our AQA+Temp and AQA+Temp-M models performs better than the GAMA baseline, which has a sophisticated audio encoding. This emphasizes the need for good data augmentation in addition to architectural improvements. From the reasonable improvement in scores across all the datasets of AQA+Temp-M compared to AQA+Temp, we infer that providing metadata during training helps in better detection of audio events and improved audio-text representation mapping. In Fig 3, our proposed models show consistent improvements over the baseline, indicating the effectiveness of temporal finetuning. Table 5 presents the performance of various models on the MMAU Test-Mini Sound split benchmark (Sakshi et al., 2024). Based on our organization's guidelines, we use the test-mini instead of the full test set as the latter requires us to upload our model's generations to the MMAU webpage. Our proposed method, AQA+Temp-M, performs better than the baseline LTU by a significant margin of 10.21. This shows the efficacy of our proposed data augmentation and temporal finetuning. Hence, the proposed method improves temporal reasoning in the baseline LALM while maintaining previously learned skills, as illustrated quantitatively in Table 4 and 5.

4.2 Qualitative Analysis of Temporal Finetuning

From Table 3, it is evident that temporal finetuning with temporal reasoning data augmentation, as described in Section 2.2, results in the generation of rationale with temporal commonsense knowledge compared to the baseline. In the first example, although the baseline's answer is correct, the reasoning is wrong since the model is only provided with 10 seconds of audio clip content. In the second example, the baseline model states incorrect audio events—tap and generic impact sounds—and continues to use them in the rationale. On the other hand, the AQA+Temp generates the correct temporal answer along with a plausible explanation as rationale. This illustrates a qualitative improvement in our proposed method's answer generation over the baseline.

4.3 Ablation Study on Meta data and Curriculum Learning

We conduct an ablation study on the design choices, namely, providing meta data information and learn-

FP (bits)	Model Size (GiB)	Load Time (ms)	Prompt Eval Rate (TPS)	Eval Rate (TPS)
16	12.55	10925.84	6.95	7.35
8	6.67	2690.79	13.57	13.16
4	3.56	1395.71	15.79	19.64

Table 6: Comparison of inference speed for AQA across different floating point (FP) precision on-device. FP and TPS refers to floating point precision and tokens per second respectively.

ing with curriculum learning. In Table 4, the LTU model shows the baseline performance. The second section comprising of AQA+Temp (w/o CL) and AQA+Temp-M (w/o CL) reflects our model's performance without curriculum learning while the last two rows, AQA+Temp (ours) and AQA+Temp-M (ours) uses curriculum learning. Without curriculum learning, the AQA+Temp and AQA+Temp-M models perform poorly on all the datasets except TemporalQA. This is expected as the model forgets it's base checkpoint finetuning and overfits to temporal reasoning. Another interesting observation is that AQA+Temp-M performs better than AQA+Temp only when trained with curriculum learning. This could be due to better learning of the audio-text embedding due to a combination of multiple audio tasks - audio tagging, audio captioning and audio question answering. This analysis emphasizes the joint importance of curriculum learning and meta data information.

4.4 Insight on On-device AQA Inference

Table 6 presents the model loading time and inference speed of AQA for different floating point precisions. The load time denotes the time taken to load the model into the CPU. Prompt Eval Rate measures the number of user query prompt tokens encoded relative to the time taken for performing audio and prompt encoding. Eval rate refers to the time taken to generate the response. User prompts should usually be encoded quicker than the response generation because user prompts can be encoded as a batch of tokens while a response is generated auto-regressively, word by word. Despite this, for the 4-bit and 16-bit models, we see a lower Prompt Eval Rate than Eval Rate. This could be due to the audio encoding overhead, which needs to be kept in mind for improving overall inference latency.

5 Conclusion

In this work, we proposed a novel data augmentation strategy to generate temporal reasoning QA pairs using LLMs. Next, we finetuned a SOTA AQA model on the generated temporal reasoning data and showcased quantitative improvements across evaluation metrics. Finally, we showcased our implementation of the AQA model on-device and studied its performance. In the future, we will reduce the memory footprint of our AQA model to be able to fit into low-powered devices. This will also significantly reduce the active RAM usage and boost encoding and decoding speeds. Also, we plan to investigate quantization-aware fine-tuning techniques and study the generation quality vs. quantization trade off. We plan to introduce an evaluation metric that can appropriately select the facts from the answer and compare them against the ground truth. We can use the metric as a loss term during fine-tuning of the AQA model to prioritize the learning of specialized skills reliably.

References

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems*, volume 36, pages 18090–18108. Curran Associates, Inc.

Radosvet Desislavov, Fernando Martínez-Plumed, and José Hernández-Orallo. 2023. Trends in ai inference energy consumption: Beyond the performance-vs-parameter laws of deep learning. *Sustainable Computing: Informatics and Systems*, 38:100857.

Haytham M. Fayek and Justin Johnson. 2020. Temporal reasoning via audio question answering. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:2283–2294.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024a. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *Preprint*, arXiv:2406.11768.

Sreyan Ghosh, Ashish Seth, Sonal Kumar, Utkarsh Tyagi, Chandra Kiran Reddy Evuru, Ramaneswaran S, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024b. Compa: Addressing the gap in compositional reasoning in audio-language models. In *The Twelfth International Conference on Learning Representations*.

Yuan Gong, Yu-An Chung, and James Glass. 2021. AST: Audio Spectrogram Transformer. In *Proc. Interspeech 2021*, pages 571–575.

Yuan Gong, Hongyin Luo, Alexander H Liu, Leonid Karlinsky, and James Glass. 2023. Listen, think, and understand. *arXiv preprint arXiv:2305.10790*.

Shawn Hershey, Daniel P W Ellis, Eduardo Fonseca, Aren Jansen, Caroline Liu, R Channing Moore, and Manoj Plakal. 2021. The benefit of temporally-strong labels in audio event classification. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 366–370.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa dataset.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, and Lama Ahmad et al. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Zhenyu Pan, Haozheng Luo, Manling Li, and Han Liu. 2024. Chain-of-action: Faithful and multimodal question answering through large language models. *arXiv preprint arXiv:2403.17359*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *Preprint*, arXiv:2410.19168.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *The Twelfth International Conference on Learning Representations*.

Yi Yuan, Zhuo Chen, Xubo Liu, Haohe Liu, Xuenan Xu, Dongya Jia, Yuanzhe Chen, Mark D. Plumbley, and Wenwu Wang. 2024. T-clap: Temporal-enhanced contrastive language-audio pretraining. *Preprint*, arXiv:2404.17806.

A OnDevice Graphical User Interface Examples

Figure 4 and 5 shows the GUI and an example sample for AQA running on edge CPU.

B System Prompts

The System Prompts used for generating temporal question answering data and for on-device inference are shown in Table 7.

C Sample Conversation with AQA

Figure 6 shows a sample conversation with AQA on an audio file recorded in an industrial setting.

D Device Specifications for the on-device demo

The Device has an ARM-based Snapdragon(R) X Elite processor with 32.0 GB RAM (31.6 GB usable). The CPU has 3.42 GHz clock speed operating on a 64-bit operating system.

E Additional Annotation Details

For the human evaluation to assess the quality of GPT-4 generated question answer pairs, we recruited 2 annotators through advertisement inside the department. We randomly sampled 100 metdata-question-answer pairs and provided to the consented annotators in the form of a double blindfolded survey. Therefore, not required by our IRB to be reviewed by them. The authors of this work are not lawyers. However, this opinion is based on the United States Federal regulation 45 CFR 46, under which this study qualifies for exemption via 46.104 exempt research.

Stage	System Prompt
Temporal Data Generation	Generate 5 questions and answer pairs along with metadata from the following information about the audio. The questions are used for temporal audio question answering task. Assume the audio description and audio event time information as the audio file itself. Do not ask questions whose answers are not present in the description. Write the answers in a more explanatory and human friendly manner. You can add some common senses or facts whenever it is possible along with the answer. Format each question in a single line as a JSON dictionary with keys - "id", "question", "answer", "metadata". Some examples of questions you could ask are : What sound events occurs first? What sound comes after the male speech at the beginning? (if male speech is present in the description) What event happens before the engine running sound? Which event occurs towards the end ? Is the door bell sound after the dog barking? Answer true or false and provide your reasoning steps. Can you hear footsteps before the baby cries? Answer true or false and provide your reasoning steps. What is the chronological order of the sound events? What is the background sound if there's any? Please generate diverse questions with paraphrasing.
AQA On-device Inference	A chat between a curious user and an audio question answering artificial intelligence assistant. The assistant gives helpful, detailed, and polite answers to the user's questions. You are given an audio clip and a question from the user. Do not generate false audio events or hallucinations that are not there in the audio clip. Do not contradict yourself without proper evidence.

Table 7: Systemp Prompts for Temporal Data Generation and AQA On-device Inference.

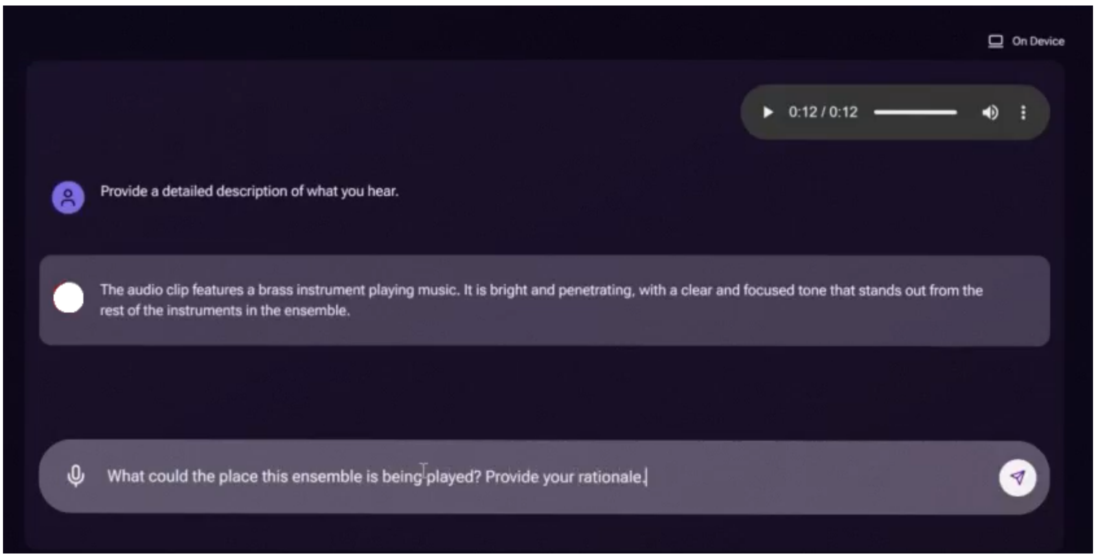

Figure 4: I: Graphical User Interface (GUI) of Audio Question Answering on ARM CPU.

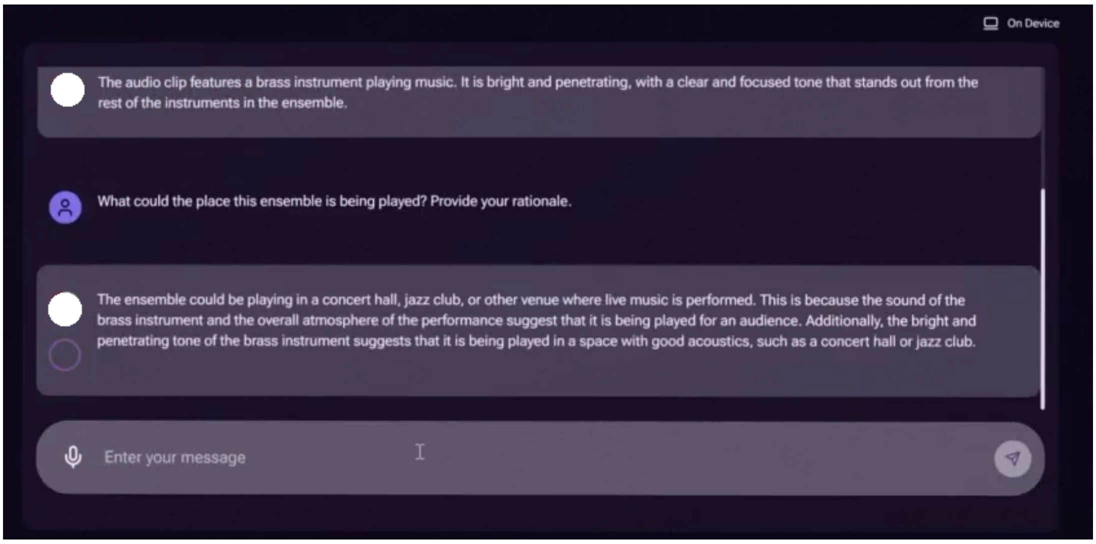

Figure 5: II: Graphical User Interface (GUI) of Audio Question Answering on ARM CPU.

Context: Audio file comprises a 10 second recording from an industry manufacturing area. Vehicle sounds are heard initially and later during the middle part of the audio. It is followed by speech, explosion and fire alarm.

Conversation

User: Describe the audio in detail.

AQA: The audio features a vehicle starting up, followed by speech, another instance of the vehicle sound, an explosion, more vehicle sounds, and finally a fire alarm.

User: Where do you think this audio was recorded?

AQA: Based on the audio content, it seems like the audio was recorded at a construction site where vehicles are being used for transportation and heavy machinery is also present.

User: When do you hear people talking and fire alarm sound?

AQA: The people start speaking at around 0.5 seconds into the audio, while the fire alarm sound can be heard at around 3.2 seconds.

User: Do you hear any babies?

AQA: No, I don't hear any babies in the audio. It primarily focuses on the sounds of vehicles and machinery with occasional speech and alarms.

User: Guess the emotion of the scene.

AQA: The scene likely has a sense of urgency due to the presence of the explosions and fire alarm, which could indicate danger or emergency situations. However, the presence of people speaking, and vehicles suggests some level of activity and productivity despite these challenging circumstances.

Figure 6: Example of a conversation with AQA-Temp proposed model