MERaLiON-AudioLLM: Advancing Speech and Language Understanding for Singapore

Yingxu He[*], **Zhuohan Liu**[*], **Geyu Lin**[*], **Shuo Sun**[*],
Bin Wang[*], **Wenyu Zhang**[*], **Xunlong Zou**[*], **Nancy F. Chen**, **Ai Ti Aw**
Institute for Infocomm Research (I^{2}R), A*STAR, Singapore
{sun_shuo,wang_bin}@i2r.a-star.edu.sg

Abstract

We introduce MERaLiON-AudioLLM, the first general-purpose multitask audio-based large language model designed to understand Singlish, a colloquial and code-switched variety of English spoken in Singapore. Trained on 62 million multimodal instruction samples spanning over 260,000 hours of audio, MERaLiON-AudioLLM exhibits strong performance across diverse tasks including automatic speech recognition, spoken question answering, speech translation, and paralinguistic analysis. We benchmark MERaLiON-AudioLLM across a broad range of multilingual and multitask scenarios, and it demonstrates competitive performance against existing open-source models. The model achieves significant gains in local speech recognition and task-specific understanding, underscoring its utility for region-specific AI applications. We develop an interactive demo interface to enable user-friendly access, supported by a back-end with custom caching and load-balancing mechanisms. The interactive demos, model weights and video are publicly available for both the first release of MERaLiON-AudioLLM[1] and the recent second release of MERaLiON-2[2]. This paper focuses exclusively on the development and evaluation of the first release.

1 Introduction

Large Language Models (LLMs) have rapidly advanced, showcasing exceptional capabilities in understanding and generating human-like text. Recent progress in transformer-based LLMs, pretrained on web-scale text corpora, has significantly improved their linguistic comprehension and generation abilities (Minaee et al., 2024; Cui et al., 2023). However, while these models excel in text-based tasks, their effectiveness in spoken language understanding remains limited, particularly in scenarios with non-standard accents, code-switching, and culturally specific linguistic patterns. This limitation presents a major challenge in multilingual regions such as Singapore, where speech-based AI systems must handle mixed languages and diverse accents.

AudioLLMs (Fang et al., 2025; Défossez et al., 2024; Gong et al., 2024; Ghosh et al., 2024; Chu et al., 2024, 2023; Tang et al., 2024; Hu et al., 2024; Lu et al., 2024; Nguyen et al., 2024) incorporate speech processing capabilities into the LLM framework, enabling the seamless integration of speech and text. AudioLLMs facilitate applications such as spoken dialogue systems, speech-based translation, and audio-driven reasoning. However, existing AudioLLMs are predominantly optimized for high-resource languages and struggle with regional linguistic adaptations, leading to suboptimal performance in real-world speech applications.

To address this challenge, we introduce MERaLiON-AudioLLM (**M**ultimodal **E**mpathetic **R**easoning and **L**earning in **O**ne **N**etwork), a speech-text model designed to enhance speech recognition and language understanding in Singapore's multilingual and multicultural environment. Developing a model that accurately understands local accents and contextual nuances is essential to create more inclusive and effective AI systems. To support multimodal LLM training, we have built a robust distributed data pipeline capable of processing more than 30 TB of speech-text datasets and scalable training workflows deployed across high-performance H100 GPU clusters. Given the challenges of low-resource datasets, particularly in spoken question answering and dialogue summarization, we have enhanced our pipeline with synthesized and augmented data to improve linguistic diversity. These innovations enable MERaLiON-AudioLLM to balance computational efficiency and task-specific accuracy within a scalable 10-

[*]Equal contributions, listed in alphabetical order by last name.

[1]MERaLiON-AudioLLM: Demo, Model Card, Video

[2]MERaLiON-2: Demo, Model Card

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pages 22–30

billion-parameter architecture. Our key contributions are as follows:

- Regionally adapted speech-text model: MERaLiON-AudioLLM is specifically designed for multilingual and accent-adaptive speech understanding. By leveraging large-scale speech-text data with synthesized and augmented samples, the model effectively handles regional accents, code-switching, and culturally specific linguistic patterns. The model weights are open-sourced to encourage further research and development.
- State-of-the-art performance across multiple tasks: MERaLiON-AudioLLM achieves state-of-the-art results in local speech recognition and spoken language understanding, reducing word error rates (WER) and improving semantic alignment for regional accents.
- Interactive demo system for real-time exploration: We present an interactive demo that enables seamless real-time interaction with MERaLiON-AudioLLM, allowing researchers and developers to evaluate its performance across diverse linguistic scenarios.

2 Overview of Interactive Demo System

To enable rapid experimentation, we designed and deployed an interactive demo on HuggingFace. We adhered to the conventional model-view-controller (MVC) design paradigm, dividing the system into three key components: 1) a user-friendly front-end interface built with Streamlit **(view)**, a backend powered by vLLM for efficient language model inference with MERaLiON-AudioLLM **(model)**, and a carefully designed interaction pipeline to manage the complex logic between the user and the model **(controller)**.

2.1 MERaLiON-AudioLLM Playground

The landing page of our demo system is *MERaLiON-AudioLLM Playground*, which provides an interactive and intuitive interface that allows users to upload audio clips, inspect and listen to the audio content, and interact with the MERaLiON-AudioLLM backend in real time.

As shown in Figure 1, the interface includes a navigation panel that enables users to explore different configurations. For example, the cascade system channels the output of MERaLiON-AudioLLM to other text-based LLMs for further inference, while the voice chat feature allows users to engage with the system through spoken interaction, eliminating the need for text prompts. To enhance the user experience, the interface offers a variety of speech samples, including standard English, Singapore-accented English, and Singlish. Users can also choose from multiple variants of the MERaLiON-AudioLLM model. We would update the selection progressively as new models become available.

2.2 Model-Serving Backend

To enhance the efficiency of processing audio inputs, we have integrated MERaLiON-AudioLLM with vLLM (Kwon et al., 2023), a LLM fast inference framework that leverages PagedAttention to optimize memory allocation and minimize latency. It supports Audio Input Processor and provides the flexibility to integrate customised model architectures. By developing a custom vLLM integration plugin, MERaLiON-AudioLLM can now handle up to 16 concurrent requests. As is illustrated in Table 1, running on NVIDIA H100 GPU, performance benchmarks show a Time-To-First-Token of 0.149 seconds, with an Inter-Token Latency of 16 milliseconds. This results in a throughput of 867 tokens per second. The model weights, together with the vLLM plugin, are fully open-sourced on our Hugging Face page.

2.3 Interaction Pipeline

When a user submits an audio clip and text prompt by clicking the send button, the web interface transmits the inputs via HTTP connections to our backend infrastructure, which is hosted on a GPU server and managed by a FastAPI application. We have implemented carefully designed logics to dynamically route incoming user requests to multiple model instances and orchestrate the complex processes required for our AI system. The core component, MERaLiON-AudioLLM, processes the audio and text inputs, generating appropriate responses that are sent back through HTTP connections to the frontend for display to the user.

3 Model Architecture

MERaLiON-AudioLLM is designed to take a pair of inputs (audio, text) and generate text outputs. As shown in Figure 2, MERaLiON-AudioLLM consists of three components: 1) an **audio encoder** that transforms speech or audio inputs into sequences of vector representations; 2) an **adapter module**

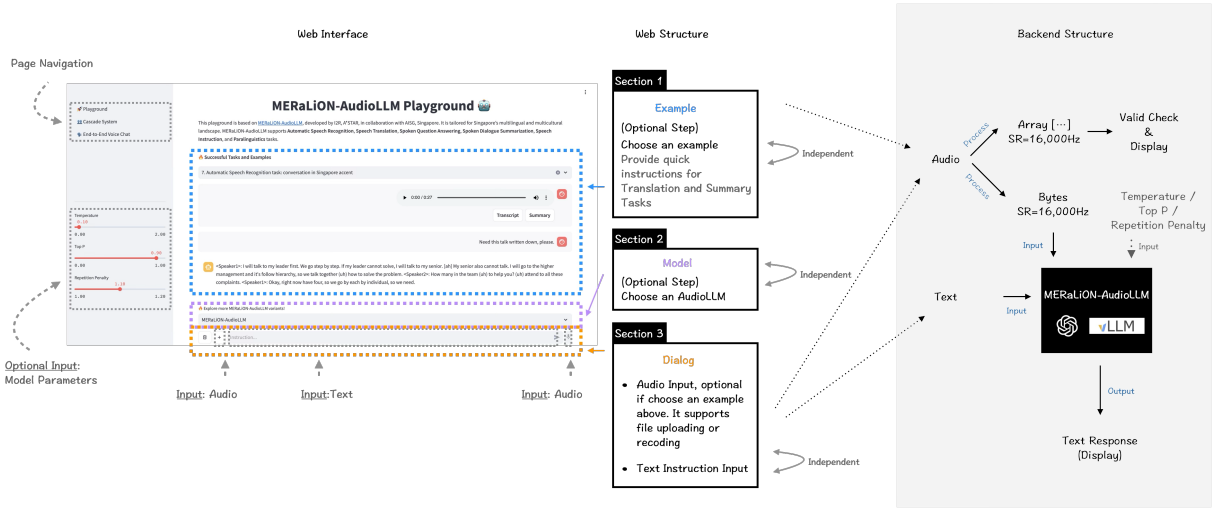

Figure 1: Demo Workflow: 1) Audio Input: Users can either upload an audio clip or use sample audio clips; 2) Text Input; 3) Multimodal Understanding: The audio and direct text input are processed together to understand user intent; 4) Output: The model generates a text response based on the user's inputs.

Concurrent Requests	30s Audio		1min Audio		2mins Audio	
	TTFT (ms)	ITL (ms)	TTFT (ms)	ITL (ms)	TTFT (ms)	ITL (ms)
1	85.8	9.9	126.4	9.6	214.5	9.7
4	96.9	11.4	159.6	11.1	258.1	11.2
8	109.6	13.0	206.5	12.7	261.9	13.0
16	149.9	16.3	236.7	16.2	299.0	16.8

Table 1: vLLM Performance benchmark for MERaLiON-AudioLLM running on a single H100 GPU. We report average **Time To First Token** (TTFT, unit: ms) together with **Inter-Token Latency** (ITL, unit: ms), over 120 trials for each input audio length and concurrency combination.

to align the speech or audio embeddings with the embedding size of the text decoder; 3) and a **text decoder** that interprets and responds to natural language instructions.

3.1 Audio Encoder

The audio encoder of MERaLiON-AudioLLM is initialized from the encoder of Whisper-large-v2 (Radford et al., 2022), which has demonstrated strong performance across various speech recognition tasks, to develop our in-house MERaLiON-Whisper. To adapt Whisper to local accents and linguistic contexts, we further fine-tune the model using a mixture of publicly available and in-house automatic speech recognition (ASR) datasets.

3.2 MLP Adapter Module

Since the output dimension of the audio encoder (1280) is significantly smaller than the embedding size of the text decoder (3584), we employ a two-layer MLP adapter module, referred to as the MLP-100 adapter, to align the speech (or audio) embeddings with the text instruction embedding space. The adapter module consists of two hidden layers. The first layer transforms the sequence of encoder outputs into 100 audio token embeddings, while the second layer upscales the hidden size of these token embeddings to match the dimensionality of the text decoder. Our experiments show that this simple adapter module outperforms other alternatives, such as the Q-former (Tang et al., 2024) and ConvMLP (Li et al., 2021).

3.3 LLM Decoder

The text decoder of MERaLiON-AudioLLM ingests a concatenated sequence of audio context tokens and text instruction tokens, and then generates a text-based response. For this purpose, we leverage on SEA-LION V3 (Singapore, 2024), a state-of-the-art localized large language model for the Southeast Asia region. SEA-LION V3 was built upon the 9B version of Google's Gemma

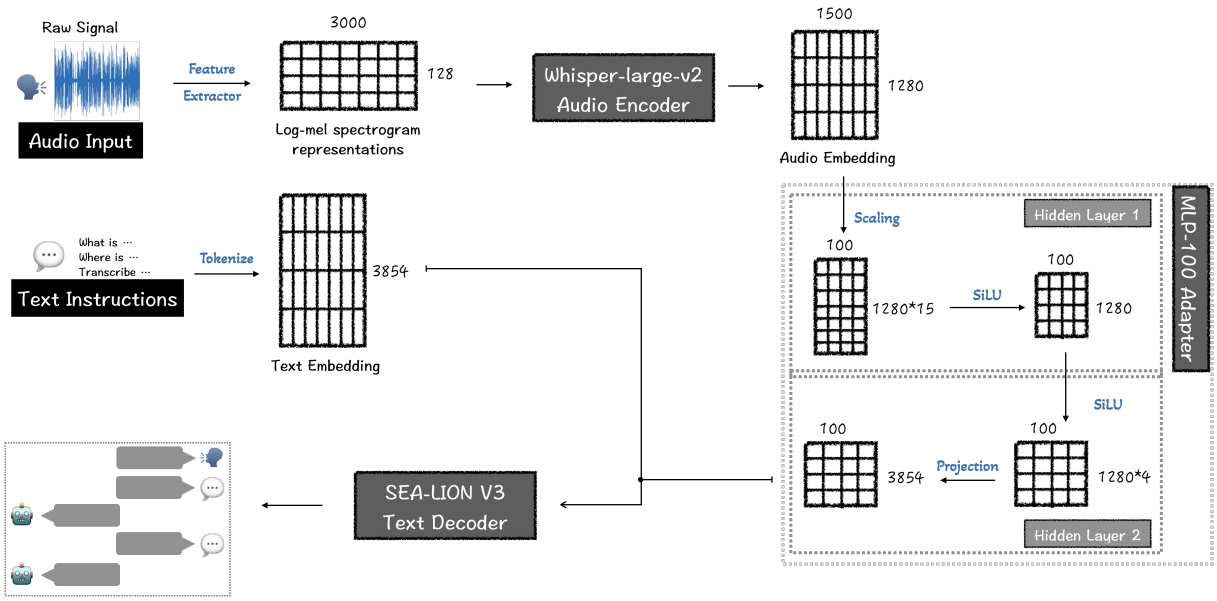

Figure 2: Architecture of MERaLiON-AudioLLM: 1) Audio Encoder: Fine-tuned Whisper-large-v2 encoder on a collection of local dataset; 2) Text Decoder: SEA-LION V3; 3) MLP-100 Adaptor Module: Consists of two hidden layers that reshape and project the audio embedding to match the dimension size of the text decoder.

2 (Team et al., 2024) by continual pre-training it on an additional 200 billion tokens sourced from diverse datasets. We use the instruct version of SEA-LION V3,[3] which was further fine-tuned on approximately 500,000 English instruction-tuning pairs and approximately 1 million instruction tuning pairs in various ASEAN languages.

3.4 Training Data

We curated an extensive collection of speech-text instruction-tuning pairs totaling 260,000 hours of data. A significant portion of this dataset is derived from IMDA's National Speech Corpus (NSC) (Koh et al., 2019), which is licensed under the Singapore Open Data License.[4] To enhance the diversity of the collection, we further augmented it with both in-house and open-source datasets, covering a wide range of audio tasks, including Automatic Speech Recognition (ASR), Spoken Dialogue Summarization (SDS), Speech Translation (ST), Spoken Question Answering (SQA), Audio Question Answering (AQA), Audio Captioning (AC), Speech Instruction (SI), and Paralinguistic Question Answering (PQA). We standardized all training samples into a unified schema consisting of an audio context, a text instruction, and a corresponding text answer. Examples of the datasets are illustrated in Figure 3.

[3] https://huggingface.co/aisingapore/gemma2-9b-cpt-sea-lionv3-instruct

[4] https://data.gov.sg/open-data-licence

ASR:{'context': [-0.0201416, ..., 0.02240472], 'instruction': "Please transcribe.", 'answer': "Groves started writing songs when she was four years old."}
SQA:{'context': [-1.22070312e-04, ..., -0.07333374], 'instruction': "Why does the woman buy a new bike?", 'answer': "The old one is broken."}

Figure 3: Examples of our training data

As the National Speech Corpus contains mislabelled data, we polished the dataset by performing extensive data cleaning and filtering. Additionally, we expanded it by synthesizing examples for various tasks, such as Speech Question Answering (SQA) and Gender Recognition (GR). The final dataset, which we named Multitask National Speech Corpus (MNSC), has been released for open access (Wang et al., 2025).

3.5 Training Strategy

This speech-text instruction-tuning supports multiple tasks and facilitates multimodal instruction fine-tuning, enabling MERaLiON-Whisper and SEA-LION V3 to perform cross-modal reasoning and achieve improved task-specific performance.

With a global batch size of 640, we train the current release of MERaLiON-AudioLLM for around 200,000 steps, which took 2 days to complete using 128 H100 GPUs. During the training, we minimize the autoregressive loss function that measures the difference between the predicted and ground truth

Models	ASR-PART1/2	ASR-PART3/4/5/6	SQA-PART3/4/5/6	SDS-PART3/4/5/6	Accent	Gender
Cascade Model	20.0	29.7	66.9	53.2	16.8	23.0
SALMONN-7B	25.8	50.8	42.2	14.4	1.3	51.25
WavLLM	18.2	69.6	51.2	39.5	1.5	47.9
Qwen2-Audio-7B	13.2	35.6	46.7	35.3	1.8	65.0
MERaLiON-AudioLLM	**4.5**	**20.0**	**59.2**	**53.6**	**42.8**	**80.0**

Table 2: Results for Singlish understanding datasets, reported as unweighted averages across subsets. The best result for each dataset is underlined, while the top-performing end-to-end AudioLLM is highlighted in **bold**.

sequences. The model predicts for the output sequence $\mathbf{y}_{i,j} = \left\{\mathbf{y}_{i,j}^{(1)}, \mathbf{y}_{i,j}^{(2)}, \ldots, \mathbf{y}_{i,j}^{(L)}\right\}$ autoregressively, where L is the output sequence length. The autoregressive loss for a sample is formulated as:

$$\mathcal{L}_{i,j} = \sum_{\ell=1}^{L} -\log P\left(\mathbf{y}_{i,j}^{(\ell)} \mid \mathbf{y}_{i,j}^{(<\ell)}, \mathbf{x}_{i,j}^{audio}, \mathbf{x}_{i,j}^{text}\right) \tag{1}$$

where $\mathbf{y}_{i,j}^{(<\ell)}$ represents the output tokens before the current prediction token. This loss encourages the model to accurately predict each token in the output sequence, conditioned on the prior output tokens and the multimodal input representations.

Besides, we fully fine-tune the audio encoder and adaptor module, while partially fine-tuning the SEA-LION V3 text decoder by adding LoRA (Low-Rank Adaptation) (Hu et al., 2022) layers with a rank of 8 to all MLP layers. We used the fused AdamW optimizer in PyTorch, along with a linear learning rate scheduler that includes 100 warm-up steps and a peak learning rate of 5e-5. To mitigate overfitting to artifacts in the input audio log-Mel spectrograms, we find it helpful to apply spectrogram augmentation (Park et al., 2019) by randomly masking a sequence of 20 time steps with a probability of 5%.

4 Performance Evaluation

To systematically evaluate the performance of AudioLLMs, we incorporated the AudioBench (Wang et al., 2024) evaluation framework and evaluated tasks covering speech, audio, and paralinguistic tasks (Achiam et al., 2023). Additionally, we use the MMAU (Sakshi et al., 2024) dataset as a general performance evaluator for audio understanding and reasoning tasks.

For comparison, we include end-to-end models that present a comprehensive understanding of audio content and cascaded models that provide a strong baseline for speech semantic tasks. The included AudioLLMs comprise recent and widely adopted models including Qwen2-Audio-7B (Chu et al., 2024), WavLLM (Hu et al., 2024), and SALMONN (Tang et al., 2024) as well as GPT4o-Audio (Achiam et al., 2023) and Gemini-1.5-Flash (Team et al., 2023). For the cascaded model, we feed the transcriptions recognized by Whisper-large (Radford et al., 2022) along with the instruction prompt to Gemma2-9B-CPT-SEA-LIONv3-Instruct model. For ASR tasks, we report the Whisper-large outputs for cascaded models.

Dataset	MERaLiON	Qwen2-Audio-7B	Cascaded Model
Automatic Speech Recognition (↓)			
LibriSpeech-Test-Clean	**0.03**	0.03	0.03
LibriSpeech-Test-Other	**0.05**	0.06	0.05
Common-Voice-15-En-Test	**0.10**	0.11	0.11
Earnings21-Test	**0.17**	0.19	0.11
Earnings22-Test	**0.20**	0.24	0.14
Speech Translation (↑)			
CoVoST 2 En → Id	**32.6**	16.3	27.6
CoVoST 2 En → Zh	**38.0**	25.8	35.3
CoVoST 2 Id → En	**37.1**	6.3	46.8
CoVoST 2 Zh → En	15.0	**16.5**	15.2
Spoken Question Answering (↑)			
CN-College-Listen-Test	**85.0**	74.5	91.9
Singapore-Public-Speech-SQA	**60.3**	58.3	73.1
SLUE-SQA-5	**82.9**	80.1	88.6
Spoken-SQuAD	**70.3**	64.9	88.6
Speech Instruction (↑)			
OpenHermes-Audio	**71.4**	44.8	72.2
Alpaca-GPT4-Audio	**73.4**	52.6	73.8
Paralinguistics (↑)			
VoxCeleb-Gender-Test	**99.5**	99.1	35.3
VoxCeleb-Accent-Test	**46.4**	29.2	24.6
MELD-Sentiment-Test	42.3	**53.5**	56.7
MELD-Emotion-Test	30.2	**40.5**	47.4

Table 3: Detailed experimental results on general-purpose evaluation datasets. The best result for each dataset is underlined, while the top-performing end-to-end AudioLLM is highlighted in **bold**.

4.1 Singlish Spoken Understanding

For Singapore-Accented English datasets, we leveraged the standard benchmark from MNSC datasets (Wang et al., 2025) where we evaluated multiple speech and voice understanding tasks including ASR, spoken question answering, spoken dialogue summarization, and paralinguistic question answering tasks.

The results are shown in Table 2. For ASR tasks, we observe that Singlish exhibits many unique words and usage patterns that deviate from stan-

Models	MMAU-Mini	Speech	Music	Sound
Cascade Model	55.6	60.7	44.0	53.5
GPT4o-Audio	40.6	54.4	29.0	38.4
Gemini-1.5-Flash	58.2	57.1	58.7	58.9
SALMONN-7B	48.4	38.1	56.0	51.1
Phi-4-Multimodal-Instruct	59.4	44.7	**68.9**	64.6
Qwen2-Audio-7B	58.9	53.5	60.2	63.1
MERaLiON-AudioLLM	**64.6**	**59.2**	64.4	**70.3**

Table 4: Results for MMAU dataset. The best result for each dataset is underlined, while the top-performing end-to-end AudioLLM is highlighted in **bold**.

dard English. As a hybrid of multiple languages and dialects, it presents significant challenges for conventional models. Without proper adaptation, both ASR systems and multitask AudioLLMs struggle to interpret the content accurately. In contrast, MERaLiON-AudioLLM has undergone careful fine-tuning on both general English data and a Singlish corpus, enabling it to adapt effectively to this linguistic domain and deliver reliable transcriptions in both sentence-level and dialogue contexts.

For SQA and SDS tasks, we observe that MERaLiON achieves performance comparable to cascaded models when trained on synthesized data. This suggests that the alignment-based approach is capable of reasoning directly over speech tokens, eliminating the need for ASR-based conversion to text. Moreover, the end-to-end model enables broader capabilities, such as paralinguistic analysis, which can be challenging for cascaded systems to handle holistically. This is evident in the results for accent and gender recognition tasks.

4.2 General Speech and Audio Understanding

Beside Singlish spoken understanding, we also include a series of other tasks to benchmark the general capability of our model. The results are shown in Table 3 and the detailed experimental setup follows Wang et al. (2024). The ASR capabilities of our model outperform other audio-based LLMs and are comparable to strong ASR systems like Whisper. However, performance drops on long-audio transcriptions, as the model is optimized for inputs under 30 seconds and may introduce errors due to unnatural truncation; further optimization is needed for handling longer audio more effectively. In speech translation, our model outperforms Qwen2-Audio in Indonesian, likely because Qwen2-Audio is primarily optimized for Chinese and English. MERaLiON also demonstrates strong capabilities in speech understanding tasks, such as spoken question answering and speech instruction following. At the same time, cascaded models establish solid baselines in these tasks, benefiting from high ASR accuracy and the instruction-following strengths of text-based LLMs. Additionally, cascaded systems excel in gender and accent recognition—tasks that remain challenging for current end-to-end models. Emotion recognition, however, continues to be a difficult area for AudioLLMs, largely due to limitations in data quality and availability and encoder's capabilities.

4.3 MMAU Evaluation

Table 4 shows the results on MMAU (mini) datasets which contains 1000 multiple choices questions covering speech, sound and music understanding (Sakshi et al., 2024). From the results, we observe that MERaLiON-AudioLLM achieves the highest average performance, outperforming both closed-source and open-source models. GPT-4o-Audio tends to abstain from answering when uncertain, which negatively impacts its final score. A similar pattern is observed in cascaded models for speech tasks, which, despite their generally strong performance, also experience penalties due to non-responses. Although MERaLiON is not specifically fine-tuned for music tasks, its performance in music understanding ranks just behind the Phi-4 model. This suggests that multitask training and broad coverage in the training data can significantly enhance a model's zero-shot capabilities.

5 Conclusion and Future Work

We introduce MERaLiON-AudioLLM, the first audio-centric large language model tailored specifically for speech and audio comprehension within Singapore's local context. Utilizing multitask learning, it demonstrates impressive performance across a range of speech and audio-related tasks. This advancement underscores the effectiveness of combining large-scale multimodal datasets with sophisticated model architectures.

For future work, we plan to expand AudioLLM to support Singapore's other official languages — Chinese, Malay, and Tamil — along with additional languages from the Southeast Asia region. We are also exploring methods to enhance the instruction-following capabilities of these models while preserving their performance in core audio tasks, such as ASR. Future updates to the model will be progressively rolled out through our Hugging Face page and interactive demo system.

Limitations

Context Length. Our demo is optimal for 30 seconds of audio context and can handle audios up to 2 minutes. We plan to enhance its ability to handle long-range dependencies in both conversational speech and complex narratives. Additionally, we are improving its capacity for multi-turn interactions and processing interleaved text and audio inputs.

Safety Considerations. Our demo is not specifically fine-tuned for safety alignment; instead, its safety characteristics are inherited from the integrated pre-trained LLMs, which may be impacted during fine-tuning. Enhancing multimodal safety alignment remains a promising direction for future work.

Instruction Following. Fine-tuning AudioLLM end-to-end for tasks like speech recognition and translation has caused certain level of catastrophic forgetting, reducing its ability to follow text instructions. To address this, we are exploring mitigations by incorporating more diverse multimodal datasets and better alignment strategies.

Multilingualism and Empathetic Reasoning. While the model and demo can handle non-English speech and non-speech tasks. It is still limited to the pre-trained capability from Whisper's multilingual encoder. We believe its performance can be improved with more data sources, especially for low-resource languages. We are actively exploring strategies to scale up data collection efficiently.

Acknowledgement

We extend our sincere gratitude to Jeremy H. M. Wong, Kye Min Tan, Hardik B. Sailor, Qiongqiong Wang, Muhammad Huzaifah, Xin Huang, Tarun K. Vangani, Nattadaporn Lertcheva, Xi Wang, Kui Wu, Jayden Lum, and Xue Cong Tey for their invaluable contributions to data management, human annotations, logistics, insightful discussions, and future work explorations.

This research is supported by the National Research Foundation, Singapore and Infocomm Media Development Authority, Singapore under its National Large Language Models Funding Initiative. Any opinions, findings, conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore and Infocomm Media Development Authority, Singapore.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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. 2024. Qwen2-audio technical report. *Preprint*, arXiv:2407.10759.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *Preprint*, arXiv:2311.07919.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, Tianren Gao, Erlong Li, Kun Tang, Zhipeng Cao, Tong Zhou, Ao Liu, Xinrui Yan, Shuqi Mei, Jianguo Cao, Ziran Wang, and Chao Zheng. 2023. A survey on multimodal large language models for autonomous driving. *Preprint*, arXiv:2311.12320.

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *Preprint*, arXiv:2410.00037.

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. Llama-omni: Seamless speech interaction with large language models. *Preprint*, arXiv:2409.06666.

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

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James Glass. 2024. Listen, think, and understand. *Preprint*, arXiv:2305.10790.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Lingwei Meng, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, and Furu Wei. 2024. Wavllm: Towards robust and adaptive speech large language model. *Preprint*, arXiv:2404.00656.

Jia Xin Koh, Aqilah Mislan, Kevin Khoo, Brian Ang, Wilson Ang, Charmaine Ng, and Ying-Ying Tan.

2019. Building the Singapore English National Speech Corpus. In *Proc. Interspeech 2019*, pages 321–325.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Jiachen Li, Ali Hassani, Steven Walton, and Humphrey Shi. 2021. Convmlp: Hierarchical convolutional mlps for vision.

Ke-Han Lu, Zhehuai Chen, Szu-Wei Fu, He Huang, Boris Ginsburg, Yu-Chiang Frank Wang, and Hung yi Lee. 2024. Desta: Enhancing speech language models through descriptive speech-text alignment. *Preprint*, arXiv:2406.18871.

Shervin Minaee, Tomas Mikolov, Narjes Nikzad, Meysam Chenaghlu, Richard Socher, Xavier Amatriain, and Jianfeng Gao. 2024. Large language models: A survey. *Preprint*, arXiv:2402.06196.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R. Costa-jussa, Maha Elbayad, Sravya Popuri, Christophe Ropers, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, Mary Williamson, Gabriel Synnaeve, Juan Pino, Benoit Sagot, and Emmanuel Dupoux. 2024. Spirit lm: Interleaved spoken and written language model. *Preprint*, arXiv:2402.05755.

Daniel S. Park, William Chan, Yu Zhang, Chung-Cheng Chiu, Barret Zoph, Ekin D. Cubuk, and Quoc V. Le. 2019. Specaugment: A simple data augmentation method for automatic speech recognition. In *Interspeech 2019*, interspeech_2019. ISCA.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint*.

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. In *The Thirteenth International Conference on Learning Representations*.

AI Singapore. 2024. Sea-lion (southeast asian languages in one network): A family of large language models for southeast asia. https://github.com/aisingapore/sealion.

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. *Preprint*, arXiv:2310.13289.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, Johan Ferret, Peter Liu, Pouya Tafti, Abe Friesen, Michelle Casbon, Sabela Ramos, Ravin Kumar, Charline Le Lan, Sammy Jerome, Anton Tsitsulin, Nino Vieillard, Piotr Stanczyk, Sertan Girgin, Nikola Momchev, Matt Hoffman, Shantanu Thakoor, Jean-Bastien Grill, Behnam Neyshabur, Olivier Bachem, Alanna Walton, Aliaksei Severyn, Alicia Parrish, Aliya Ahmad, Allen Hutchison, Alvin Abdagic, Amanda Carl, Amy Shen, Andy Brock, Andy Coenen, Anthony Laforge, Antonia Paterson, Ben Bastian, Bilal Piot, Bo Wu, Brandon Royal, Charlie Chen, Chintu Kumar, Chris Perry, Chris Welty, Christopher A. Choquette-Choo, Danila Sinopalnikov, David Weinberger, Dimple Vijaykumar, Dominika Rogozińska, Dustin Herbison, Elisa Bandy, Emma Wang, Eric Noland, Erica Moreira, Evan Senter, Evgenii Eltyshev, Francesco Visin, Gabriel Rasskin, Gary Wei, Glenn Cameron, Gus Martins, Hadi Hashemi, Hanna Klimczak-Plucińska, Harleen Batra, Harsh Dhand, Ivan Nardini, Jacinda Mein, Jack Zhou, James Svensson, Jeff Stanway, Jetha Chan, Jin Peng Zhou, Joana Carrasqueira, Joana Iljazi, Jocelyn Becker, Joe Fernandez, Joost van Amersfoort, Josh Gordon, Josh Lipschultz, Josh Newlan, Ju yeong Ji, Kareem Mohamed, Kartikeya Badola, Kat Black, Katie Millican, Keelin McDonell, Kelvin Nguyen, Kiranbir Sodhia, Kish Greene, Lars Lowe Sjoesund, Lauren Usui, Laurent Sifre, Lena Heuermann, Leticia Lago, Lilly McNealus, Livio Baldini Soares, Logan Kilpatrick, Lucas Dixon, Luciano Martins, Machel Reid, Manvinder Singh, Mark Iverson, Martin Görner, Mat Velloso, Mateo Wirth, Matt Davidow, Matt Miller, Matthew Rahtz, Matthew Watson, Meg Risdal, Mehran Kazemi, Michael Moynihan, Ming Zhang, Minsuk Kahng, Minwoo Park, Mofi Rahman, Mohit Khatwani, Natalie Dao, Nenshad Bardoliwalla, Nesh Devanathan, Neta Dumai, Nilay Chauhan, Oscar Wahltinez, Pankil Botarda, Parker Barnes, Paul Barham, Paul Michel, Pengchong Jin, Petko Georgiev, Phil Culliton, Pradeep Kuppala, Ramona Comanescu, Ramona Merhej, Reena Jana, Reza Ardeshir Rokni, Rishabh Agarwal, Ryan Mullins, Samaneh Saadat, Sara Mc Carthy, Sarah Cogan, Sarah Perrin, Sébastien M. R. Arnold, Sebastian Krause, Shengyang Dai, Shruti Garg, Shruti Sheth, Sue Ronstrom, Susan Chan, Timothy Jordan, Ting Yu, Tom Eccles, Tom Hennigan, Tomas Kocisky, Tulsee Doshi, Vihan Jain, Vikas Yadav, Vilobh Meshram, Vishal Dharmadhikari, Warren Barkley, Wei Wei, Wenming Ye, Woohyun Han, Woosuk Kwon, Xiang Xu, Zhe Shen, Zhitao Gong, Zichuan Wei, Victor Cotruta, Phoebe Kirk, Anand Rao, Minh Giang, Ludovic Peran, Tris Warkentin,

Eli Collins, Joelle Barral, Zoubin Ghahramani, Raia Hadsell, D. Sculley, Jeanine Banks, Anca Dragan, Slav Petrov, Oriol Vinyals, Jeff Dean, Demis Hassabis, Koray Kavukcuoglu, Clement Farabet, Elena Buchatskaya, Sebastian Borgeaud, Noah Fiedel, Armand Joulin, Kathleen Kenealy, Robert Dadashi, and Alek Andreev. 2024. Gemma 2: Improving open language models at a practical size. *Preprint*, arXiv:2408.00118.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2024. Audiobench: A universal benchmark for audio large language models. *Preprint*, arXiv:2406.16020.

Bin Wang, Xunlong Zou, Shuo Sun, Wenyu Zhang, Yingxu He, Zhuohan Liu, Chengwei Wei, Nancy F. Chen, and AiTi Aw. 2025. Advancing singlish understanding: Bridging the gap with datasets and multimodal models. *Preprint*, arXiv:2501.01034.