The Interspeech 2026 Audio Encoder Capability Challenge for Large Audio Language Models

1 Introduction

Building on the success of the inaugural ICME 2025 Audio Encoder Capability Challenge [1], we propose its second iteration for Interspeech 2026. The previous challenge aimed at establishing a vital benchmark for audio representation learning by evaluating encoders' performance across dozens of discrete classification tasks, from speech and music to environmental sounds. However, the landscape of audio processing is undergoing a seismic shift, driven by the rapid advancement of Large Audio Language Models (LALMs). These models are revolutionizing human-computer interaction, but their progress is currently hampered by a critical bottleneck: an overwhelming dependence on a limited set of frontend encoders [2, 3, 4, 5], primarily derived from Whisper [6], with few works utilizing different audio-encoders [7, 8]. This dependency presents a challenge to architectural innovation and to realizing the full potential of LALMs.

This challenge represents a considered evolution from its predecessor, strategically leveraging the established resources and evaluation framework from the previous edition while fundamentally reengineering the assessment paradigm. Whereas the inaugural challenge focused on measuring encoder capabilities through linear fine-tuning on classification tasks, we now shift to evaluating encoders specifically as frontend modules within a complete LALM architecture for generative modeling. To support this refined focus, we have developed a comprehensive new benchmark named "XARES-LLM" [1] including a specialized test set specifically designed for generative audio-language tasks, along with a completely re-optimized testing system that ensures fair and reproducible evaluation of how different encoders impact overall LALM performance. Furthermore, to encourage broader participation, we are providing commercial-grade datasets that are immediately available to all registered participants at no cost. We hope to address the current bottleneck in frontend encoder diversity and stimulate innovative approaches that move beyond the dominant Whisper-based paradigms, encourage the community to explore alternatives to monolithic encoders and open up new avenues for research in audio AI.

2 Challenge Design

2.1 Overview

The goal of this challenge is to evaluate the capability of pre-trained audio encoders serving as front-end encoders for LALMs, with a focus on their ability to understand and represent audio semantics in complex scenarios. The challenge adopts a unified end-to-end generative evaluation framework. Participants only need to submit a pre-trained encoder model, while the downstream task training and evaluation are completed by the organizers.

2.2 Evaluation System

The organizers provide XARES-LLM, an open-source evaluation system. XARES-LLM trains a typical LALM using the audio encoder provided by the user. The system automatically downloads training data, trains the LALM then tests various downstream tasks, providing scores for each. The XARES-LLM system is depicted in Figure 1.

[1] `https://github.com/xiaomi-research/xares-llm.git`

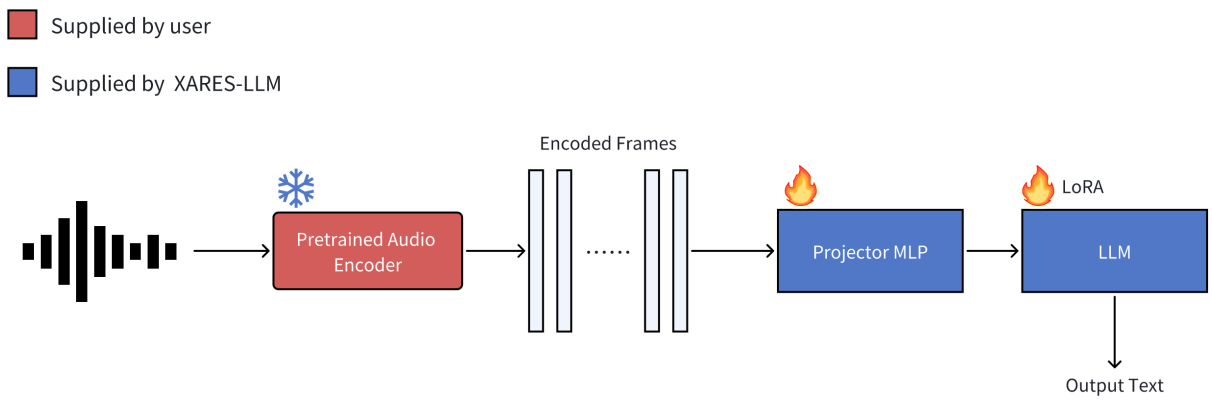

Figure 1: The XARES-LLM system. Users provide a pretrained audio encoder that processes input audio into encoded frames. XARES-LLM then trains a projector ontop of a pretrained LLM, and automates dataset downloading as well as evaluation.

Since XARES-LLM is open-source, participants can run the system themselves to preview the scores of their submitted models on a select number of evaluation tasks. However, this operation is not mandatory. The only mandatory task for all participants is to properly encapsulate their encoder according to the framework requirements. The XARES-LLM system provides a tool to check if the participant's encapsulation is valid and guides them to correctly encapsulate the encoder. Unlike the XARES [9] system used in the previous challenge, which trained separate models (MLPs) for each downstream task, XARES-LLM trains only a single model and then tests and scores it simultaneously on all downstream tasks to evaluate the model's generalization performance. In order to control the model output, we append a unique task identifier as a prompt into the LLM.

The XARES-LLM evaluation framework is structured around two primary components: Closed-Set Recognition and Open-Set Generation. Both are designed for a comprehensive, speech-centric capability assessment, and models are jointly trained on a predefined mixture of all datasets. The Closed-Set Recognition part follows a traditional classification paradigm, requiring models to map audio to predefined labels. As detailed in Table 1, this task category features a rich set of speaking-focused challenges (e.g., keyword spotting, speaker/language ID, emotion/intent recognition) and also includes tasks in music and general sound classification. Evaluation metrics are chosen accordingly: Accuracy (Acc) is used for single-label classification, while mean Average Precision (mAP) is used for multi-label tasks.

The Open-Set Generation part offers a novel, dual-faceted evaluation of a model's ability to generate text from audio, with specific tasks shown in Table 2. The two task categories are:

1. Precise Transcription: This category focuses on tasks like Automatic Speech Recognition (ASR), where objective accuracy is paramount. Performance is measured using the inverted Word Error Rate (iWER), defined as:

$$\text{iWER}(\text{hyp}, \text{target}) = \max(0, 1 - \text{WER}(\text{hyp}, \text{target})).$$

2. Descriptive Summarization: This includes holistic tasks like cross-domain Audio Captioning, where models generate natural language summaries of complex acoustic scenes (e.g., speech, music, environmental sounds). Performance is assessed using the FENSE metric [10] for sound and music captions, and the novel DATE metric [11] for general audio captions.

Training is done by jointly training on all datasets using a predefined mixture, by balancing each task respectively. All datasets have been organized into the WebDataset [12] format and uploaded to Hugging Face [2] and will be downloaded automatically during runtime. Users are not allowed to change the content or proportion of the training set.

[2] https://huggingface.co/datasets/mispeech/xares_llm_data

2.3 Baseline System

A challenge baseline is provided using the Whisper encoder [6]. The baseline implementation and evaluation results are available in the XARES-LLM repository.

Table 1: Classification tasks in XARES-LLM. The preprocessed version of the datasets are provided on Zenodo. The number of labels # for each dataset are displayed.

Domain	Dataset	Task Type	Metric	#
Speech	Speech Commands [13]	Keyword spotting	Acc	30
	LibriCount [14]	Speaker counting	Acc	11
	VoxLingua107 [15]	Language identification	Acc	33
	VoxCeleb1-Binary [16]	Binary speaker identification	Acc	2
	LibriSpeech [17]	Gender classification	Acc	2
	Fluent Speech Commands [18]	Intent classification	Acc	248
	VocalSound [19]	Non-speech sounds	Acc	6
	CREMA-D [20]	Emotion recognition	Acc	5
	ASV2015 [21]	Spoofing detection	Acc	2
Sound	ESC-50 [22]	Environment classification	Acc	50
	FSD50k [23]	Sound event detection	mAP	200
	UrbanSound 8k [24]	Urban sound classification	Acc	10
	FSD18-Kaggle [25]	Sound event detection	mAP	41
Music	GTZAN Genre [26]	Genre classification	Acc	10
	NSynth-Instruments [27]	Instruments Classification	Acc	11
	NSynth-Pitch [27]	Pitches Classification	Acc	128
	Free Music Archive Small [28]	Music genre classification	Acc	8

2.4 Participant Training Dataset

The challenge places a significant emphasis on data collection and utilization, which is a crucial component of the competition. The organizers do not prescribe a specific training dataset for each participant. Instead, participants are free to use any data for training, as long as it meets the following conditions:

- All training data must be publicly accessible, or in Table 3.
- Data in Table 1 and Table 2 must be excluded from training.

To further support participants, the organizers are offering free access to a subset of eight commercial datasets. These datasets are typically paid, but upon registration for the challenge, participants can download them at no cost. Participants may choose to incorporate these datasets into their training pipeline, but their use is entirely optional. The details of these commercial datasets are provided in Table 3.

Note that the provided subset may not include the full datasets due to licensing restrictions, but a representative sample will be made available.

Table 2: Open-ended tasks in XARES-LLM. The preprocessed version of the datasets are provided on Zenodo.

Dataset	Task Type	Metric
LibriSpeech-100h [17]	Speech recognition	iWER
AISHELL-1-100h [29]	Speech recognition	iWER
Clotho [30]	Sound Caption	FENSE
The Song Describer Dataset [31]	Music Caption	FENSE
MECAT [11]	General Caption	DATE

Table 3: Commercial datasets provided for the challenge. Participants can access a subset of these datasets for free after registration.

Dataset Name	Description
King-ASR-457	Real scenario noise corpus with 27 environments
King-ASR-610	Far-field English TV programs noise, 3-channel
King-ASR-719	Environmental noise, water, footstep, and TV noise
King-ASR-829	Beijing subway broadcast noise
King-ASR-862	Howling noise speech database
King-ASR-876	In-car, out-car, and other environmental noises
King-ASR-955	In-car noise database
King-ASR-958	Multi-scenario noise corpus with 11 scenarios

Additionally, participants are permitted to utilize publicly available pre-trained models as starting points for their encoder development. Any pre-trained audio model is allowed to be used, which fulfill the conditions denoted before, such as Whisper [6], Dasheng [32], HuBERT [33] etc.

2.5 Evaluation and Ranking

The task metrics are specifically designed such that for each task higher is better. An overall performance is then computed by a weighted average algorithm. For each task T_i, we first normalize each task-specific metric M_i (e.g., accuracy, iWER , mAP). To normalize these metrics, we use the following formula:

$$\hat{M}_i = \frac{M_i - M_i^{\min}}{M_i^{\max} - M_i^{\min}} \tag{1}$$

where $\hat{M}_i$ is the normalized metric for task T_i, and M_i is the raw metric value for task T_i. $M_i^{\min}$ and $M_i^{\max}$ are the worst and best possible values of the metric M_i, respectively.

The final score for each participant is calculated as the weighted average of the normalized metrics across all tasks, where the weight n_i is determined by the committee. Since the open-ended tasks are more difficult than classification tasks, we aim to set the weights such that there is a larger importance on these open-ended tasks.

$$S = \frac{\sum_{i=1}^{N_{\text{task}}} n_i \hat{M}_i}{\sum_{i=1}^{N_{\text{task}}} n_i} \tag{2}$$

where N_{task} is the total number of tasks applicable to the respective task, n_i is the task-specific weight and $\hat{M}_i$ is the normalized metric for task T_i.

3 Submission Guide

Participants are required to submit a pre-trained model encapsulated within the specified API. The model should accept a single-channel audio signal, represented as a PyTorch tensor with shape $[B, T]$, where B denotes the batch size and T represents the number of samples in the time domain. The model should output a frame-level prediction of shape $[B, T', D]$, where T' can be different from the input T and D is the embedding dimension defined by the participant.

While there are no strict limitations on model size, submitted models must be able to be run successfully in a Google Colab T4 environment, where the runtime is equipped with a 16 GB NVIDIA Tesla T4 GPU, 12GB RAM.

Participants are also required to submit a technical report along with their submission.

The submission steps are as follows:

1. Clone the audio encoder template from the GitHub repository[3].

[3] `https://github.com/jimbozhang/xares-llm-template.git`

2. Implement your own audio encoder following the instructions in `README.md` within the cloned repository. The implementation must pass all checks in `audio_encoder_checker.py` provided in the repository.
3. Before the submission deadline, email the organizers [4] a ZIP file containing the complete repository. Additionally, please attach a technical report paper (PDF format) not exceeding 6 pages describing your implementation. Pre-trained model weights can either be included in the ZIP file or downloaded automatically from external sources (e.g., Hugging Face) during runtime. If choosing the latter approach, please implement the automatic downloading mechanism in your encoder implementation.

4 Challenge Schedule

The Challenge will follow this schedule:

- November 21, 2025: Challenge announcement
- February 12 11:59 AM AoE, 2026: Submissions Deadline
- February 20, 2026: Final Ranking Announced
- February 25 11:59 PM AoE, 2026: Paper Submission Deadline

5 Challenge Organizers

This Challenge is organized by teams from four institutions: Xiaomi Corporation, the University of Surrey, Tsinghua University and Dataocean AI Inc.

Xiaomi Corporation is a renowned technology company established in 2010. It is widely known for its diverse product range including smartphones, cars, tablets, laptops, wearables, and smart home devices, to form a platform of more than 800 million active devices. The company emphasizes innovation and user experience, is dedicated to fundamental technologies, blends into open-source. AI has been fully integrated into to reinforce Xiaomi's machie intelligence and service efficiency, ranging from user interaction, imaging, auto pilot, to internet sales, delivery, and service.

Dr. Junbo Zhang is an AI Research Scientist at Xiaomi Corporation. He earned his Ph.D. from the Institute of Acoustics at the Chinese Academy of Sciences. With years of experience in developing acoustic and speech algorithms, Dr. Zhang has made significant contributions to various fields, including speech recognition, pronunciation evaluation, speech synthesis, audio tagging, sound separation, and noise reduction. He has authored over 30 papers in prestigious journals and top-tier conferences. As a code contributor to the open-source project Kaldi, he also wrote the book "Kaldi Speech Recognition Practice", which has sold more than ten thousand copies. At Xiaomi, he was instrumental in developing and launching the company's initial speech recognition system, the wake word detection for "Xiao Ai" (Xiaomi's AI assistant), and the voiceprint recognition system. Currently, he leads several pioneering projects in the large model technology domain, pushing the boundaries of what is possible in consumer electronics.

Dr. Heinrich Dinkel is an Algorithm Engineer at Xiaomi Corporation. He recieved his Ph.D. from Shanghai Jiao Tong University, supervised by Kai Yu and Mengyue Wu, in 2020. His research interest are mainly focused around general sound understanding and its applications in consumer facing real-world scenarios. He has written over 50 papers, which were published in prestigious journals and conferences. He developed multiple models at Xiaomi that have been deployed on a variety of products, ranging from smart speakers, intelligent lights, phones and smart cars.

Dr. Yadong Niu is currently an Algorithm Engineer at Xiaomi Corporation. He received his Ph.D. from Peking University in 2024. His research interests primarily revolve around the applications of deep learning and signal processing in the audio domain, including areas such as speech understanding, speech enhancement, and acoustic modeling.

University of Surrey The Machine Audition Lab within the Centre for Vision Speech and Signal Processing at the University of Surrey, led by Prof Wenwu Wang, is a leading research lab in audio signal processing and machine learning, consisting more than 30 researchers. They have developed several widely

[4] `2026interspeech-aecc@dataoceanai.com`

used audio representation models such as PANNs, AudioLDM, AudioLDM 2, AudioSep, etc. They have been contributing to the activities in Detection and Classification of Acoustic Scenes and Events (DCASE) challenges and workshops since 2013, including the organisation of two tasks of the DCASE 2024 Challenges, i.e. Task 6 - Automated Audio Captioning and Task 9 - Language-Queried Audio Source Separation.

Dr. Wenwu Wang is a Professor in Signal Processing and Machine Learning, University of Surrey, UK. He is also an AI Fellow at the Surrey Institute for People Centred Artificial Intelligence. His current research interests include signal processing, machine learning and perception, artificial intelligence, machine audition (listening), and statistical anomaly detection. He has (co)-authored over 300 papers in these areas. He has been recognized as a (co-)author or (co)-recipient of more than 15 accolades, including the 2022 IEEE Signal Processing Society Young Author Best Paper Award, ICAUS 2021 Best Paper Award, DCASE 2020 and 2023 Judge's Award, DCASE 2019 and 2020 Reproducible System Award, and LVA/ICA 2018 Best Student Paper Award. He is an Associate Editor (2020-2025) for IEEE/ACM Transactions on Audio Speech and Language Processing, and an Associate Editor (2024-2026) for IEEE Transactions on Multimedia. He was a Senior Area Editor (2019-2023) and Associate Editor (2014-2018) for IEEE Transactions on Signal Processing. He is the elected Chair (2023-2024) of IEEE Signal Processing Society (SPS) Machine Learning for Signal Processing Technical Committee, a Board Member (2023-2024) of IEEE SPS Technical Directions Board, the elected Chair (2025-2027) and Vice Chair (2022-2024) of the EURASIP Technical Area Committee on Acoustic Speech and Music Signal Processing, an elected Member (2021-2026) of the IEEE SPS Signal Processing Theory and Methods Technical Committee. He has been on the organising committee of INTERSPEECH 2022, IEEE ICASSP 2019 & 2024, IEEE MLSP 2013 & 2024, and SSP 2009. He is Technical Program Co-Chair of IEEE MLSP 2025. He has been an invited Keynote or Plenary Speaker on more than 20 international conferences and workshops.

Tsinghua University Human-Computer Speech Interaction Lab at Tsinghua University (THUHCSI) targets at cutting-edge research in intelligent speech interaction technologies, including audio foundation models, expressive and controllable speech generation, digital human generation, natural language processing, and machine learning. Over the years, THUHCSI has undertaken and contributed to major national and international research programs, yielding internationally recognized research achievements. THUHCSI maintains strong collaborations with both academia and industry, including ModelBest, Tencent, Microsoft, ByteDance, Alibaba and Xiaomi. Many of our research outputs have been widely transferred and applied in education, intelligent customer service, smart hardware, and digital human applications.

Dr. Zhiyong Wu is a professor of Tsinghua Shenzhen International Graduate School, Tsinghua University. He is also a coordinator with the Tsinghua-CUHK Joint Research Center for Media Sciences, Technologies and Systems. His research interests cover the areas of intelligent speech interactions, specifically, audio-visual bimodal modeling, text-to-audio-visual expressive speech synthesis, and natural language understanding and generation. He has authored more than 150 papers on peer-reviewed top international conferences and journals, including IEEE/ACM TASLP, IEEE TMM, IEEE TPAMI, NeurIPS, ICLR, AAAI, IJCAI, CVPR, ACM Multimedia, ICASSP, INTERSPEECH), etc. His research achievements have been recognized with multiple awards, including the Ministry of Education (MoE) Science and Technology Progress Awards (2009, 2016), the Beijing Science and Technology Progress Award (2021), and the Shenzhen Science and Technology Progress Award (2023). He won the first prize in the "Spoofing Attack Task" in the GeekPwn 2017 Shanghai Contest (2017), and the first place in all challenge tracks of the ICASSP 2023 Speech Signal Improvement Challenge (2023). He was also awarded the Best Student Paper in INTERSPEECH 2023 (2023). He is currently the member of China Computer Federation (CCF), Chinese Association for Artificial Intelligence (CAAI), Institute of Electrical and Electronics Engineers (IEEE), and International Speech Communication Association (ISCA). He serves as the Standing Committee Member and Deputy Secretary-General of the Speech Dialogue and Auditory Processing Technical Committee, China Computer Federation (CCF TFSDAP),. He is the Publication Chair of ISCSLP 2012, Special Session Area Chair of INTERSPEECH 2020, Local Arrangement Chair of SLT 2020, Local Chair (China) of ICASSP 2022, and serves as the reviewer for IEEE/ACM TASLP, Speech Communications, MTAP, ICASSP, INTERSPEECH, IJCNLP, COLING, AAAI, NeurIPS, ISCSLP, etc.

DataOcean AI Inc. is a global data collection and labeling services provider that combines technology with a diverse network of millions data contributors, scientists, and engineers. The company delivers cutting-edge data solutions across multiple domains, including text, audio, image, and multimodal for foundation models or GenAI applications. With over 1,600 off-the-shelf datasets and a proven track record of delivering

thousands of customized data projects, DataOcean AI have been trusted by of over 1,000 global AI leading enterprises and institutions. The company cover more than 200 languages around the world. Its self-developed data platform ensures precision and efficiency in tasks such as collection, cleansing, labeling and evaluation. With nearly two decades of experience, DataOcean AI has established itself as a trusted partner in the AI ecosystem, consistently delivering excellence and earning global recognition.

Dr. Yufeng Hao obtained his PhD from Southeast University in 2004 and currently serves as Vice President and Chief Linguist of Beijing Haitian Ruisheng Technology Co., Ltd. With long-term engagement in the R&D of intelligent speech technology, basic research on multilingual linguistics, design of high-quality datasets, and quality control & evaluation, he has so far developed phonological systems for over 200 languages and hundreds of datasets. He has also published more than 20 academic papers, been granted over 30 patents, and participated in the compilation and release of 2 national standards, 1 industrial standard, 2 corporate standards, and 2 industry monographs.

6 List of Recommended Expert Reviewers

The following is a list of recommended expert reviewers for the challenge.

- Helen Meng, hmmeng@se.cuhk.edu.hk, The Chinese University of Hong Kong
- Tan Lee, tanlee@ee.cuhk.edu.hk, The Chinese University of Hong Kong
- Haizhou Li, haizhouli@cuhk.edu.cn, The Chinese University of Hong Kong, Shenzhen
- Hung-yi Lee, hungyilee@ntu.edu.tw, National Taiwan University
- Jianhua Tao, jhtao@tsinghua.edu.cn, Tsinghua University
- Kai Yu, kai.yu@sjtu.edu.cn, Shanghai Jiao Tong University
- Dong Yu, dongyu@ieee.org, Tencent
- Kong-aik Lee, kongaik.lee@ieee.org, The Hong Kong Polytechnic University
- Jinyu Li, jinyli@microsoft.com, Microsoft
- Zhijian Ou, ozj@tsinghua.edu.cn, Tsinghua University
- Zhenhua Ling, zhling@ustc.edu.cn, University of Science and Technology of China
- Yuexian Zou, zouyx@pku.edu.cn, Peking University Shenzhen Graduate School
- Mark Plumbley, mark.plumbley@kcl.ac.uk, King's College London
- Emmanouil Benetos, emmanouil.benetos@qmul.ac.uk, Queen Mary University of London
- Mohsen Naqvi, Mohsen.Naqvi@newcastle.ac.uk, Newcastle University
- Danilo Comminiello, danilo.comminiello@uniroma1.it, Sapienza University of Rome
- Mads Christensen, mgc@es.aau.dk, Aalborg University
- Noboru Harada, harada.noboru@ntt.com, NTT Communication Science Laboratories
- Xin Wang, wangxin@nii.ac.jp, National Institute of Informatics
- Lei Xie, lxie@nwpu.edu.cn, Northwestern Polytechnical University

References

[1] J. Zhang, H. Dinkel, Q. Song, H. Wang, Y. Niu, S. Cheng, X. Xin, K. Li, W. Wang, Y. Wang *et al.*, "The icme 2025 audio encoder capability challenge," *arXiv preprint arXiv:2501.15302*, 2025.

[2] J. Bai, X. Chen, Y. Zhou, C.-C. Liu, Y. Chen, S. Huang, K. Chen, J.-F. Li, H. Lin, H. Zhou, L. Yang, Z. Li, Y. Wang, J. Lin, Y.-H. Zheng, Y. Chen, C. Zhang, X. Lu, X. Xu, X. Zhao, W. Han, C. Wang, Y. Hu, J. Lu, H. Chen, P. Lv, W. Liu, W. Dai, and M. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," *arXiv preprint arXiv:2403.02422*, 2024. [Online]. Available: https://qwen-audio.github.io/Qwen-Audio/

[3] J. Bai, X. Chen, X. Lu, J. Lin, Y. Chen, J.-F. Li, Y. Wang, X. Xu, C. Zhang, Y.-H. Zheng, C. Wang, W. Han, J. Lu, H. Chen, P. Lv, W. Liu, W. Dai, and M. Zhou, "Qwen2.5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025. [Online]. Available: https://arxiv.org/abs/2503.20215

[4] M. AI, "Kimi-audio technical report," *arXiv preprint arXiv:2504.18425*, 2025. [Online]. Available: https://arxiv.org/abs/2504.18425

[5] A. Goel, S. Ghosh, J. Kim, S. Kumar, Z. Kong, S.-g. Lee, C.-H. H. Yang, R. Duraiswami, D. Manocha, R. Valle *et al.*, "Audio flamingo 3: Advancing audio intelligence with fully open large audio language models," *arXiv preprint arXiv:2507.08128*, 2025.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International conference on machine learning*. PMLR, 2023, pp. 28 492–28 518.

[7] Z. Kong, A. Goel, R. Badlani, W. Ping, R. Valle, and B. Catanzaro, "Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities," *arXiv preprint arXiv:2402.01831*, 2024.

[8] H. Dinkel, G. Li, J. Liu, J. Luan, Y. Niu, X. Sun, T. Wang, Q. Xiao, J. Zhang, and J. Zhou, "Midashenglm: Efficient audio understanding with general audio captions," *arXiv preprint arXiv:2508.03983*, 2025.

[9] J. Zhang, H. Dinkel, Y. Niu, C. Liu, S. Cheng, A. Zhao, and J. Luan, "X-ares: A comprehensive framework for assessing audio encoder performance," in *Interspeech 2025*, 2025.

[10] Z. Zhou, Z. Zhang, X. Xu, Z. Xie, M. Wu, and K. Q. Zhu, "Can audio captions be evaluated with image caption metrics?" in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 981–985.

[11] Y. Niu, T. Wang, H. Dinkel, X. Sun, J. Zhou, G. Li, J. Liu, X. Liu, J. Zhang, and J. Luan, "Mecat: A multi-experts constructed benchmark for fine-grained audio understanding tasks," *arXiv preprint arXiv:2507.23511*, 2025.

[12] A. Perlmutter, "Webdataset: A library for efficient loading of large-scale datasets," https://github.com/webdataset/webdataset, 2021, accessed: [current date]. [Online]. Available: https://github.com/webdataset/webdataset

[13] P. Warden, "Speech commands: A dataset for limited-vocabulary speech recognition," *arXiv preprint arXiv:1804.03209*, 2018.

[14] F.-R. Stöter, S. Chakrabarty, E. Habets, and B. Edler, "Libricount, a dataset for speaker count estimation," 2018.

[15] J. Valk and T. Alumäe, "Voxlingua107: a dataset for spoken language recognition," in *2021 IEEE Spoken Language Technology Workshop (SLT)*. IEEE, 2021, pp. 652–658.

[16] A. Nagrani, J. S. Chung, W. Xie, and A. Zisserman, "Voxceleb: Large-scale speaker verification in the wild," *Computer Speech & Language*, vol. 60, p. 101027, 2020.

[17] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[18] L. Lugosch, M. Ravanelli, P. Ignoto, V. S. Tomar, and Y. Bengio, "Speech model pre-training for end-to-end spoken language understanding," *arXiv preprint arXiv:1904.03670*, 2019.

[19] Y. Gong, J. Yu, and J. Glass, "Vocalsound: A dataset for improving human vocal sounds recognition," in *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2022, pp. 151–155.

[20] H. Cao, D. G. Cooper, M. K. Keutmann, R. C. Gur, A. Nenkova, and R. Verma, "Crema-d: Crowd-sourced emotional multimodal actors dataset," *IEEE transactions on affective computing*, vol. 5, no. 4, pp. 377–390, 2014.

[21] T. Kinnunen, Z. Wu, E. Nicholas Evans, and J. Yamagishi, "Automatic speaker verification spoofing and countermeasures challenge (asvspoof 2015) database," 2018.

[22] K. J. Piczak, "Esc: Dataset for environmental sound classification," in *Proceedings of the 23rd ACM international conference on Multimedia*, 2015, pp. 1015–1018.

[23] E. Fonseca, X. Favory, J. Pons, F. Font, and X. Serra, "Fsd50k: an open dataset of human-labeled sound events," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 30, pp. 829–852, 2021.

[24] J. Salamon, C. Jacoby, and J. P. Bello, "A dataset and taxonomy for urban sound research," in *Proceedings of the 22nd ACM international conference on Multimedia*, 2014, pp. 1041–1044.

[25] E. Fonseca, M. Plakal, F. Font, D. P. Ellis, X. Favory, J. Pons, and X. Serra, "General-purpose tagging of freesound audio with audioset labels: Task description, dataset, and baseline," *arXiv preprint arXiv:1807.09902*, 2018.

[26] B. L. Sturm, "The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use," *arXiv preprint arXiv:1306.1461*, 2013.

[27] J. Engel, C. Resnick, A. Roberts, S. Dieleman, D. Eck, K. Simonyan, and M. Norouzi, "Neural audio synthesis of musical notes with wavenet autoencoders," 2017.

[28] M. Defferrard, K. Benzi, P. Vandergheynst, and X. Bresson, "Fma: A dataset for music analysis," *arXiv preprint arXiv:1612.01840*, 2016.

[29] B. Hui, D. Jiayu, N. Xingyu, W. Bengu, and Z. Hao, "Aishell-1: An open-source mandarin speech corpus and a speech recognition baseline," in *Oriental COCOSDA 2017*, 2017, p. Submitted.

[30] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2020, pp. 736–740.

[31] I. Manco, B. Weck, S. Doh, M. Won, Y. Zhang, D. Bogdanov, Y. Wu, K. Chen, P. Tovstogan, E. Benetos, E. Quinton, G. Fazekas, and J. Nam, "The song describer dataset: a corpus of audio captions for music-and-language evaluation," in *Machine Learning for Audio Workshop at NeurIPS 2023*, 2023.

[32] H. Dinkel, Z. Yan, Y. Wang, J. Zhang, Y. Wang, and B. Wang, "Scaling up masked audio encoder learning for general audio classification," in *Interspeech 2024*, 2024.

[33] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, K. Lakhotia, R. Salakhutdinov, and A. Mohamed, "Hubert: Self-supervised speech representation learning by masked prediction of hidden units," *IEEE/ACM transactions on audio, speech, and language processing*, vol. 29, pp. 3451–3460, 2021.