Synthetic Conversation Dataset Using Large Language Models

Dhivya Nagasubramanian *

Posted Date: 27 October 2025

doi: 10.20944/preprints202510.2025.v1

Keywords: audio-augmented LLMS; text-to-speech; multi-turn dialogues; instruction-tuning; large language models

 doi:10.20944/preprints202510.2025.v1

Article

Synthetic Conversation Dataset Using Large Language Models

Dhivya Nagasubramanian

Independent Researcher, Minneapolis, MN, 55446, USA; nagas021@alumni.umn.edu

Abstract

Most existing open-source datasets are designed around single-turn interactions, such as question-answering tasks, or are based on monotonic speech datasets taken from audiobooks, where a single speaker talks throughout. These datasets do not fully capture the dynamic nature of real-world conversations, which involve multiple speakers, shifting tones, and diverse dialects. Unlike ImageNet, which has played a key role in advancing image recognition, the speech AI research community currently lacks a comprehensive, diverse, multilingual dataset for conversational speech. To fill this gap, we introduce the Multi-Lingual Dialogue Dataset (MLDD), consisting of 200,000 multi-turn dialogue samples. The topic for the conversation generation is derived from the New York Times annotated corpus, and we enhance it by making the dataset multilingual using large language model (LLM) capabilities. Additionally, the emotion of the conversation is set through LLM prompting, and the pitch and talking speed of the dialogues are set through text-to-speech models to mimic real-world conversations. The Multi-Lingual Dialogue dataset (MLDD) is generated by prompting the LLM with article titles and summaries from the New York Times and providing emotional tone as input to produce engaging multi-turn conversations. To demonstrate the utility and complexity of the MLDD, we evaluate it using audio-augmented large language models. Our results show the practical applications of this dataset for more interactive and nuanced dialogue.

Keywords: audio-augmented LLMS; text-to-speech; multi-turn dialogues; instruction-tuning; large language models

---

Multilingual dialogue, a cornerstone of human interaction, encompasses the linguistic diversity across languages and the complexities of culture, context, and communication styles. In recent years, the development of multilingual dialogue systems has become a critical area of research in natural language processing (NLP). Dialogue systems, which facilitate communication between humans and machines, have predominantly been designed for a single language. However, multilingual capabilities have become essential for widespread adoption with increasing globalization and the need for more inclusive technologies. This has led to the creation of large-scale multilingual dialogue datasets such as MultiWOZ 2.0 and OPUS-MT, which serve as foundational resources for training and evaluating dialogue models across multiple languages. These datasets have enabled significant strides in developing systems that can handle cross-lingual interactions effectively.

Advances in pre-trained language models, such as BERT and T5, have accelerated progress in multilingual dialogue systems. These models leverage transfer learning to adapt to various linguistic contexts, enabling dialogue systems to generate coherent, context-aware responses in multiple languages. Additionally, the application of cross-lingual transfer learning has proven valuable in improving performance for low-resource languages by transferring knowledge from high-resource languages, as demonstrated by Ji et al. In parallel, techniques like dialogue state tracking have enhanced the ability of multilingual systems to maintain context and coherence over long conversations.

Despite these advancements, several challenges remain. Issues such as data silos, inconsistent data quality, and the scalability of models across languages continue to hinder the development of

 doi:10.20944/preprints202510.2025.v1

truly universal multilingual systems. Efforts to address these challenges include improving the quality of training data (the area where the research paper is focused on), leveraging AI-powered multilingual tools, and optimizing data preprocessing strategies. These advancements are vital to ensuring the development of multilingual dialogue systems that are not only scalable but also capable of providing seamless, context-aware interactions in various languages.

While models trained on multilingual dialogue datasets have demonstrated significant potential in enhancing communication and overcoming language barriers worldwide, a notable lack of comprehensive multilingual conversational speech datasets remains. The research community requires a diverse, multilingual conversation dataset to improve models on speaker diarization and ASR (Automatic Speech Recognition). Just like ImageNet is for image recognition, we need a multilingual conversation dataset that could be used to train and build a highly scalable, generalized speech recognition model. Though some publicly available conversation datasets exist, they are all in American English accents and will not work in training speech recognition models for other languages. LibriSpeech and TED-LIUM are suitable for automatic speech recognition (ASR) tasks but primarily consist of isolated utterances or formal speech. Switchboard and Fisher are focused on telephone conversations in English but lack multilingual and emotionally rich diversity.

To address these limitations, in this paper, we propose the Multi-lingual Dialogue Dataset, a dialogue dataset with multilingual, multi-turn dialogues from the topics and summaries derived from the New York Times dataset. Similar to how instruction-tuning datasets are generated to train vision and language assistants, we use a prompting-based approach to develop a multilingual, multi-turn dialogue dataset using a pre-trained LLM.

The recent advancements in large language models (LLMs) have led to impressive performance across various natural language processing (NLP) benchmarks. These models, trained on vast amounts of unsupervised text data, significantly benefit numerous downstream text generation tasks. LLMs can be fine-tuned by utilizing instruction tuning to address a wide range of NLP challenges better. Furthermore, LLMs exhibit in-context learning capabilities, meaning they can adapt and learn from a few examples within a given context, even if those examples were not present during training. These characteristics make LLMs highly appealing for other modalities, such as speech. Given their success in NLP, there is growing interest in leveraging LLMs to enhance speech modeling as well.

For creating the synthetic dataset, we utilize New York Times article topics and the synopsis of the article to guide the dialogue generation process using GPT-4 Additionally, we implement data filtration strategies to filter out any harmful opinionated dialogues and noisy synthetic dialogues, therefore promoting the retention of the most reliable ones. Our proposed dataset comprises 200K samples, each containing between four and eight dialogue turns. Our main contributions are: 1) a dataset pipeline for generating a multilingual dialogue dataset for any language from any existing structured dataset, and 2) the evaluation of existing audio-augmented large language models on our proposed dataset.

Related Work

Instruction-following Large Language Models (LLMs) have exhibited exceptional performance in zero-shot and few-shot tasks within the language domain, including machine translation, summarization, and other related tasks. The concept of creating models that can follow instructions has since been expanded to other domains, including vision and audio.

LLaVA made the first attempt at generating instruction-following data involving visual content using GPT-4. Specifically, they leverage image captions and bounding box localization as metadata for the image, which is then used as a query for the language model. In total, they gather 158k samples for language-image instruction-following data.

Since then, increasing interest has been in creating instruction-following datasets like VALLEY, Macaw-LLM, and Video-ChatGPT. In the audio domain, the LTU dataset was generated using GPT to make an open-ended question-answering dataset to assess general knowledge and reasoning about everyday sounds. However, LTU's dataset is limited in several ways: it only includes single-turn

 doi:10.20944/preprints202510.2025.v1

conversations, lacks complex inter-conversational context, and does not feature strong correlations between rounds. On the other hand, Qwen-Audio has curated a 20k audio-based instruction-following dataset, but there is minimal discussion regarding its curation process or the specifics of the dataset itself. Although some conversations exist today, such as Switchboard and DailyDialogue, for audio LLM training, these datasets are particular to the English language, and the opportunity to scale the models to multilingual becomes limited due to the unavailability of multilingual datasets.

Our Multi-Lingual Dialogue Dataset addresses all the above-mentioned limitations by generating multi-lingual, multi-turn conversations for article topics and summaries from the New York Times dataset. Compared to existing datasets, Multi-lingual Dialogue Datasets have multi-turn dialogues in various languages with strong correlations between rounds through the presence of pronouns (e.g., he, she, it), follow-up questions based on the previous answer, and complex context.

Recent research has focused on advancing audio foundation models [that leverage large language models (LLMs) to understand audio content. These models typically use an audio encoder to convert audio into tokens, which are then processed with textual instructions by an LLM to generate a final response. Pretrained on various tasks, such as audio captioning, emotion recognition, sound event classification, speech recognition, and music understanding, these models have demonstrated significant improvements in zero-shot and few-shot performance when using a unified architecture. While these models show strong audio comprehension, recent work like Audio Flamingo has introduced methods such as in-context learning and retrieval-augmented generation to further enhance their ability to follow instructions, achieved through fine-tuning with interleaved audio-text pairs. To assess the relevance of our proposed Multi-lingual Dialogue Dataset, we evaluate the performance of audio foundation models, including LTU, Qwen-Audio, and Audio Flamingo, on multi-turn dialogues.

Methodology

This section will discuss the data generation pipeline illustrated in Figure 1. The Multi-Lingual Dialogue Dataset (MLDD) is constructed using The New York Times dataset.

Data Pipeline

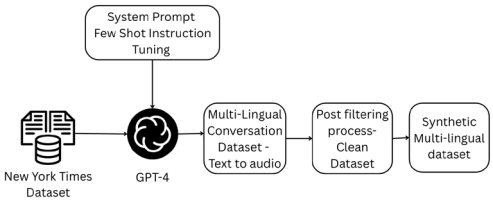

Figure 1. Illustration of Multi-Lingual Dialogue Dataset Generation Pipeline.

The structured NYT dataset consists of 1.8 million articles published between January 1, 1987, and June 19, 2007. The article metadata is provided by the NYT Newsroom in the English language. The dataset contains 1.8M records in XML file format.

Out of the 1.8M files, we randomly selected 500K documents for our dataset creation process. The New York dataset metadata comprises the date and time published, title, synopsis of the article, tags, etc. Articles are tagged for persons, places, organizations, titles, and topics using a controlled vocabulary applied consistently across articles. Tags are leveraged as filtering criteria to filter out articles on categories such as sports and places. Further, the title of the article and the summary of the article are considered for prompting the LLM, GPT-4 in this case, to generate 4-8 turn conversations.

Further, the LLM is prompted carefully to not exhibit any political opinion or biases in the conversation that it is being generated. LLM is instructed to take an emotional tone to the topic and

 doi:10.20944/preprints202510.2025.v1

make the conversation sound as natural as possible by adding natural fillers, sighs, emotional additives, breaks, etc. LLM is also instructed in the language of the conversation.

The annotated conversations are then passed to the TTS (text-to-speech) model to convert the text conversation to audio with varying pitch and tone based on the emotion label generated through LLM.

Prompt

To generate conversation on different topics, we leverage intensive LLM prompting. We design specific prompt templates to generate

1) Multiturn dialogues and

2) emotional tones of the conversation.

LLMs are prompted to adapt while coming up with the conversation content. Necessary jargon specific to expressing the emotion of the conversation is used. For example, for an angry emotion, conversations like "Huh! This is disappointing ", and for a happy emotion, the conversation will start with "I am so thankful and glad that…"

LLMs are carefully prompted to limit the conversation generation to between 4-8 turns, and guardrails are established to ensure LLMs do not become opinionated in the conversations generated. The language in which the conversation is generated is decided dynamically through random selection. The system prompt would look like in Figure 2. We adopted a few-shot prompting to instruct the LLM model on the expected outcomes. We observed that in this case, few-shot prompting worked better than zero-shot prompting.

You are a conversational AI that generates realistic dialogue based on the following parameters:

Language: {{language}} (e.g., English, Spanish, French, etc.)

Emotional Tone: {{emotional tone}} (e.g., happy, sad, angry, neutral, excited, etc.)

Topic: {{topic}} (e.g., technology, movies, sports, travel, etc.)

Summary: {{summary}} (A brief description of the context or scenario for the conversation)

Number of Turns: {{Num turns}} (e.g., 4, 5, 6 turns)

Instructions: Generate a conversation with a specified number of turns between two people where each person responds to the other's statements. Ensure the dialogue aligns with the language, emotional tone, topic, and summary provided.

Add fillers like "um," "uh," "you know," "well," "I mean," etc., to make the conversation feel more authentic, as if the speakers are pausing to think or reflect. 2. Include hesitations and pauses (e.g., "...", "I... I don't know," "It's just..."), especially in moments where the speakers feel uncertain or emotional.

Maintain a natural conversational flow, using appropriate emotions, vocabulary, and style based on the emotional tone and topic. The conversation should avoid any offensive words,

Figure 2. Illustration of prompt engineering to generate a conversational dataset.

 doi:10.20944/preprints202510.2025.v1

Data Filtration

Our data generation pipeline in the methodology section has two filters—one before passing on to the LLM and the other after the LLM output is generated. In the pre-processing step, we filter out contents from the New York Times dataset on topics such as sports, music, fashion, health & wellness, technology, science & environment, education, lifestyle, culture, arts, and travel. We restricted ourselves to generic topics to make sure the conversation generated did not become opinionated and biased.

The second filtration is done on the LLM output (conversation dataset) to ensure the conversation is aligned with the topic and the summary being passed to the LLM. Any conversation record that does not adhere to this rule will be filtered out as part of the final dataset. This is done by computing the cosine similarity between the QA pair in each dialogue and the article synopsis, which was the input to the prompt

LLM, GPT-4, in this case. The dialogues are ranked in order based on the cosine similarity score, and the top 200,000 records are considered for the final processing of converting text to audio.

Text to Speech for Audio Generation

The final step in the data pipeline is the text-to-speech conversion. The conversation dataset containing the dialogues and the emotion tag is passed through text to the speech processor, which converts each dialogue into individual audio segments. The emotional aspect of the speech is handled by manipulating the pitch and rate of speech. An example of how pitch and rate are manipulated through emotions is illustrated in Figure 3. To create a conversation that reflects the real world, we also attempted to adjust the speed throughout the conversation dynamically. Random noise was artificially appended to the existing audio to replicate background noise throughout the conversations.

Emotion to Pitch		
Emotional Tone	**Pitch**	**Description**
Happy	1.2	Increase pitch for happy
Sad	0.8	Lower pitch for sad
Angry	1.4	Increase pitch and stress for angry
Neutral	1	Normal pitch for neutral

Emotion to Rate		
Emotional Tone	**Rate**	**Description**
Happy	1.5	Slightly faster rate for happy
Sad	0.7	Slower rate for sad
Angry	1.2	Slightly faster rate for angry
Neutral	1	Normal rate for neutral

Figure 3. Illustration of the manipulation of the rate and pitch of the speech to replicate real-world audio.

Datasets

Real-world conversational datasets are a rare find. Even the few examples stated in the introduction section of this paper are limited in size and are constrained to English conversations. No single dataset exists today that is a multilingual conversational dataset that mimics real-world discussions that could be used for training state-of-the-art ASR. The dataset created through the pipeline discussed in the methodology section would address the current gap in the speech analytics field. The dialogue dataset is split into a training dataset that contains 140,000 conversation samples, each between 4 and 8 turns of dialogue, and the testing dataset has the remaining 60,000 samples. The dataset was randomly sampled, and no filtering mechanism was applied to develop a training and test group.

While the Multi-Lingual Dialogue Dataset addresses the gap of a multilingual conversation dataset for researchers looking to build state-of-the-art models in the speech domain, we see opportunities for improvement in the existing process. For example, the post-filtering process that calculates the cosine similarity between the summary input and each conversation within the dialogue could be fully automated through the reinforcement learning technique, where an LLM can

 doi:10.20944/preprints202510.2025.v1

be assigned as an assessor or grader and, therefore, do an iterative prompt tuning for better conversation generation.

Experiments & Results

In this section, we give researchers ideas on how the dataset can be evaluated to understand if the dataset is rich and diverse in the context and if it represents real-world audio dialogues. We evaluated three recent audio understanding LLMs on our dialogues: Qwen-Audio and Audio-Flamingo. We use the pre-trained models as they are for inferencing and measurement on the Multi-lingual Dialogue Dataset. The results are in Figure 5. We first did a zero-shot evaluation. We convert the audio generated from the pipeline back to a text transcript and compare it to the original transcript created through the LLM. We picked CIDEr because it is more tolerant of variations and paraphrasing, so it is better at recognizing that different phrasings can still convey the same meaning.

We also fine-tuned Audio-Flamingo and LTU on the Multi-Lingual Dialogue Dataset and compared the zero-shot results vs. fine-tuned results. The fine-tuned model (identified with ^ in below Figure 4) produced better results compared to zero-shot inferencing with the pre-trained model. As Audio Flamingo is trained with retrieval and in-context learning, it shows better performance and can use context better than LTU. This shows fine-tuning models on multilingual dialogues enables an audio-understanding LLM to have much stronger dialogue capabilities.

Dataset	Model	CIDEr	Bleu4
Multi-Lingual Dialogue Dataset	Audio-Flamingo	0.6	0.1
	Audio-Flamingo Δ	1.2	0.3
	LTU	0.48	0.17
	LTU Δ	0.67	0.25
	Qwen-Audio	0.53	0.03

Figure 4. Evaluation of LTU [17], Qwen-Audio [16], and Audio Flamingo [29] on the subsets of Multi-lingual dialogue test sets. We report the following metrics: CIDEr [48], Bleu4 [49]. Scores improve for all models fine-tuned on the synthetic dataset and it is marked as Δ.

Conclusion

This paper introduces the Multi-Lingual Dialogue Dataset, specifically for multi-turn dialogues, covering a broad spectrum of topics ranging from education, lifestyle, sports, etc. By leveraging a prompting-based approach through few-shot learning, we generate a substantial volume of high-quality dialogues suitable for training and evaluating ASR models. This paper addresses the multilingual aspect of the dialogues and attempts to lay out the flow of how the conversation generated can represent the real world by manipulating the audio generated for pitch, tone, speed, adding random noise, etc. To represent the real world, this paper also addressed ways to add modulations, fillers, and breaks to dialogues generated through LLM, aligning with the emotion of the conversation in the multi-turn dialogue. Although this pipeline addresses the current gap in the audio analytics research field, one limitation is the lack of quality assurance of the LLM-generated conversation. Although we do the cosine similarity to ensure the content generated is like the topic in the prompt, that does not provide the quality and the policy guardrails. The future paper will discuss applying reinforcement learning to self-correct the conversation content generated through LLMs.

Dhivya: Nagasubramanian holds a master's degree in Business Analytics from the University of Minnesota. Her research interests include agentic AI, Generative AI, intelligent document processing, and real-time AI systems. She is a member of Women in AI and the AI Frontier Network. Contact her at nagas021@alumni.umn.edu.

References

 doi:10.20944/preprints202510.2025.v1

1. P. Budzianowski, T.-H. Wen, R. El Asri, and M. Gasic, "MultiWOZ 2.0: A consolidated multi-domain dialogue dataset," Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 451–457, 2018.
2. S. Ruder, R. Johnson, and P. R. Kumar, "Dialogue datasets for multilingual conversational AI: A survey," Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1-10, 2020.
3. L. M. Rojas-Barahona and P. A. Perez, "Universal dialogue state tracking with a pretrained language model," Proceedings of the 2020 Conference on Natural Language Learning (CoNLL), pp. 65-75, 2020.
4. Y. Ji, D. H. Kim, and D. Lee, "Cross-lingual transfer learning for dialogue generation," Proceedings of the 2019 Conference on Neural Information Processing Systems (NeurIPS), pp. 1124-1135, 2019..
5. J. Tiedemann and S. Thottingal, "OPUS-MT: A massive multilingual machine translation system," Proceedings of the 2020 Workshop on Neural Generation and Translation (WNGT), pp. 19-23, 2020.
6. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," Proceedings of NAACL-HLT 2019, pp. 4171-4186, 2019
7. C. Raffel, N. Shazeer, A. Roberts, et al., "T5: Exploring the limits of transfer learning with a unified text-to-text transformer," Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 4171-4186, 2020.
8. L. Li, Y. Song, Z. Wang, et al., "Building multilingual dialogue systems with pretrained language models," Proceedings of the 2020 International Conference on Learning Representations (ICLR), pp. 125-136, 2020..
9. Y. Zhang, S. Chen, and X. Zeng, "Towards multilingual text generation for dialogue systems," Proceedings of the 2020 Conference on Natural Language Processing (COLING), pp. 2205-2214, 2020.
10. R. Singh, D. N. D. M. and T. K. Chakraborty, "Exploring multilingual conversational agents: Challenges and approaches," Proceedings of the 2020 Workshop on Multilingual Natural Language Processing (MLNLP), pp. 34-42, 2020.
11. X. Mei, C. Meng, H. Liu, Q. Kong, T. Ko, C. Zhao, M. D. Plumbley, Y. Zou, and W. Wang, "Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research," arXiv preprint arXiv:2303.17395, 2023.
12. 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.
13. C. D. Kim, B. Kim, H. Lee, and G. Kim, "Audiocaps: Generating captions for audios in the wild," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), 2019, pp. 119–132.
14. Y. Gong, S. Khurana, L. Karlinsky, and J. Glass, "Whisper-at: Noise-robust automatic speech recognizers are also strong general audio event taggers," arXiv preprint arXiv:2307.03183, 2023.
15. S. Deshmukh, B. Elizalde, R. Singh, and H. Wang, "Pengi: An audio language model for audio tasks," arXiv preprint arXiv:2305.11834, 2023.
16. Y. Chu, J. Xu, X. Zhou, Q. Yang, S. Zhang, Z. Yan, C. Zhou, and J. Zhou, "Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models," arXiv preprint arXiv:2311.07919, 2023.
17. Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," arXiv preprint arXiv:2305.10790, 2023.
18. S. Liu, A. S. Hussain, C. Sun, and Y. Shan, "Music understanding llama: Advancing text-to-music generation with question answering and captioning," arXiv preprint arXiv:2308.11276, 2023.
19. A.-M. Oncescu, A. Koepke, J. F. Henriques, Z. Akata, and S. Albanie, "Audio retrieval with natural language queries," arXiv preprint arXiv:2105.02192, 2021.
20. R. Huang, M. Li, D. Yang, J. Shi, X. Chang, Z. Ye, Y. Wu, Z. Hong, J. Huang, J. Liu et al., "Audiogpt: Understanding and generating speech, music, sound, and talking head," arXiv preprint arXiv:2304.12995, 2023.
21. L. Salewski, S. Fauth, A. Koepke, and Z. Akata, "Zero-shot audio captioning with audio-language model guidance and audio context keywords," arXiv preprint arXiv:2311.08396, 2023.

 doi:10.20944/preprints202510.2025.v1

22. A. Agostinelli, T. I. Denk, Z. Borsos, J. Engel, M. Verzetti, A. Caillon, Q. Huang, A. Jansen, A. Roberts, M. Tagliasacchi et al., "Musiclm: Generating music from text," arXiv preprint arXiv:2301.11325, 2023.
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. H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale et al., "Llama 2: Open foundation and fine-tuned chat models," arXiv preprint arXiv:2307.09288, 2023.
25. J. Achiam, S. Adler, S. Agarwal, L. Ahmad, I. Akkaya, F. L. Aleman, D. Almeida, J. Altenschmidt, S. Altman, S. Anadkat et al., "Gpt-4 technical report," arXiv preprint arXiv:2303.08774, 2023.
26. A. Adigwe, N. Tits, K. E. Haddad, S. Ostadabbas, and T. Dutoit, "The emotional voices database: Towards controlling the emotion dimension in voice generation systems," arXiv preprint arXiv:1806.09514, 2018.
27. B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," arXiv preprint arXiv:2304.03277, 2023.
28. P. K. Rubenstein, C. Asawaroengchai, D. D. Nguyen, A. Bapna, Z. Borsos, F. d. C. Quitry, P. Chen, D. E. Badawy, W. Han, E. Kharitonov et al., "Audiopalm: A large language model that can speak and listen," arXiv preprint arXiv:2306.12925, 2023.
29. 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.
30. B. Elizalde, S. Deshmukh, and H. Wang, "Natural language supervision for general-purpose audio representations," 2023. [Online]. Available: https://arxiv.org/abs/2309.05767
31. J. Gardner, S. Durand, D. Stoller, and R. M. Bittner, "Llark: A multimodal foundation model for music," arXiv preprint arXiv:2310.07160, 2023.
32. S. Lipping, P. Sudarsanam, K. Drossos, and T. Virtanen, "Clotho-aqa: A crowdsourced dataset for audio question answering," in 2022 30th European Signal Processing Conference (EUSIPCO). IEEE, 2022, pp. 1140–1144.
33. Z. Yang, W. Ping, Z. Liu, V. Korthikanti, W. Nie, D.-A. Huang, L. Fan, Z. Yu, S. Lan, B. Li et al., "Re-vilm: Retrieval-augmented visual language model for zero and few-shot image captioning," in EMNLP, 2023.
34. H. Liu, C. Li, Q. Wu, and Y. J. Lee, "Visual instruction tuning," arXiv preprint arXiv:2304.08485, 2023.
35. J.-B. Alayrac, J. Donahue, P. Luc, A. Miech, I. Barr, Y. Hasson, K. Lenc, A. Mensch, K. Millican, M. Reynolds et al., "Flamingo: a visual language model for few-shot learning," Advances in Neural Information Processing Systems, vol. 35, pp. 23 716–23 736, 2022.
36. Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. Glass, "Listen, think, and understand," arXiv preprint arXiv:2305.10790, 2023.
37. H. Wang, H. Wu, Z. He, L. Huang, and K. W. Church, "Progress in machine translation," Engineering, vol. 18, pp. 143–153, 2022.
38. W. S. El-Kassas, C. R. Salama, A. A. Rafea, and H. K. Mohamed, "Automatic text summarization: A comprehensive survey," Expert systems with applications, vol. 165, p. 113679, 2021.
39. R. Luo, Z. Zhao, M. Yang, J. Dong, M. Qiu, P. Lu, T. Wang, and Z. Wei, "Valley: Video assistant with large language model enhanced ability," arXiv preprint arXiv:2306.07207, 2023.
40. C. Lyu, M. Wu, L. Wang, X. Huang, B. Liu, Z. Du, S. Shi, and Z. Tu, "Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration," arXiv preprint arXiv:2306.09093, 2023.
41. M. Maaz, H. Rasheed, S. Khan, and F. S. Khan, "Video-chatgpt: Towards detailed video understanding via large vision and language models," arXiv preprint arXiv:2306.05424, 2023.
42. 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.
43. C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," Language resources and evaluation, vol. 42, pp. 335–359, 2008.

 doi:10.20944/preprints202510.2025.v1

44. P. Barros, N. Churamani, E. Lakomkin, H. Siqueira, A. Sutherland, and S. Wermter, "The omg-emotion behavior dataset," in 2018 International Joint Conference on Neural Networks (IJCNN). IEEE, 2018, pp. 1–7.
45. Z. Rafii, A. Liutkus, F.-R. St¨oter, S. I. Mimilakis, and R. Bittner, "Musdb18-hq - an uncompressed version of musdb18," Aug. 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3338373
46. S. Hershey, D. P. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporallystrong labels in audio event classification," in ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2021, pp. 366–370. [47] Y. Wu, K. Chen, T. Zhang, Y. Hui, T. Berg-Kirkpatrick, and S. Dubnov, "Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation," in ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 2023, pp. 1–5.
47. R. Vedantam, C. Lawrence Zitnick, and D. Parikh, "Cider: Consensus-based image description evaluation," in Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 4566–4575.
48. K. Papineni, S. Roukos, T. Ward, and W.-J. Zhu, "Bleu: a method for automatic evaluation of machine translation," in Proceedings of the 40th annual meeting of the Association for Computational Linguistics, 2002, pp. 311–318. [50] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," in Text summarization branches out, 2004, pp. 74–81.