A THREE-LEVEL EVALUATION PROTOCOL FOR ACOUSTIC SCENE UNDERSTANDING OF LARGE LANGUAGE AUDIO MODELS

Dilip Harish[1], *Jakob Abeßer*[2,1]

[1]Semantic Music Technologies Group, Fraunhofer IDMT, Ilmenau, Germany
[2]Computational Humanities, University of Bamberg, Germany

***Abstract*—Reaching a semantic understanding of complex acoustic scenes requires computational models to capture the temporal-spatial sound source composition as well as individual sound events. This is a great challenge for computational models due to the large variety of everyday sound events and the extensive temporal-spectral overlap in real-life acoustic scenes. In this work, we aim to evaluate the acoustic scene understanding capabilities of two large audio-language models (LALMs). As a challenging scenario, we use the USM dataset, which features synthetic urban soundscapes with 2-6 overlapping sound sources per mixture. Our main contribution is a novel three-layer evaluation protocol, which includes four analysis tasks for low-level sound event understanding (sound event tagging), mid-level understanding and reasoning (sound polyphony estimation, sound source loudness ranking), as well as high-level scene understanding (audio captioning). We apply standardized metrics to assess the models' performances for each task. The proposed multi-layer protocol allows for a fine-grained analysis of model behavior across soundscapes of various complexity levels. Our results indicate that despite their remarkable controllability using textual instructions, the ability of state-of-the-art LALMs to understand acoustic scenes is still limited as the performance on individual analysis tasks degrades with increasing sound polyphony.**

***Index Terms*—audio captioning, large audio-language models, sound event tagging, sound polyphony estimation, sound source loudness ranking**

1. INTRODUCTION

Large audio language models (LALMs) are multi-modal neural networks that learn joint representations of audio and text data. This class of models represents an important milestone on the path towards general audio intelligence, as they demonstrated state-of-the-art performance in several machine listening tasks. To this day, a major challenge for LALMs is understanding complex acoustic scenes in everyday life scenarios. Such scenes are shaped not only by the tonal diversity and overlap of sounds but also by the spatio-temporal relationships among individual sound sources.

Previous studies on evaluating the reasoning skills of audio language models have focused mainly on audio reasoning tasks, such as compositional reasoning and attribute binding by comparing audio captions. While compositional reasoning looks at how effectively a model creates new meanings from existing concepts like understanding the order of occurrence between multiple acoustic events, attribute binding focuses on its precision in matching attributes to specific acoustic events [1]. In the MMAU (Massive Multi-Task Audio Understanding) benchmark [2], the following audio reasoning types were tested: *temporal reasoning* involves inferring the timing and duration of individual sound events, *acoustic-source inference* focuses on identifying sound sources for each sound event, *eco-acoustic knowledge* involves inferring the overall environmental setting from ambient and sound event cues, *ambient sound interpretation* relates to understanding background sounds from the entire soundscape recording, *event-based sound reasoning* involves identifying causal relationships between sound events, and *sound-based event recognition* focuses on inferring high-level scenes or activities from multiple sound events. While MMAU evaluates audio recognition and reasoning tasks through multiple choice question answers, other benchmarks, such as CMM (The Curse of Multi-Modalities) have examined hallucinations in large multi-modal models to assess the gap between the factual multi-modal input and the generated text [3]. Other LALM evaluation studies use discriminative tasks to study sound object hallucinations [4]. Although these benchmarks progress in general audio reasoning evaluation, there is a current lack of studies on the abilities of LALMs to understand complex acoustic scenes.

Table 1: Semantic Levels in Acoustic Scene Understanding

Semantic Level	**Task(s)**	**Objective(s)**
Low	Sound event tagging (T_1)	Identify audible sound sources
Medium	Sound polyphony estimation (T_2)	Count audible sound sources
	Ranking-by-loudness (T_3)	Rank sound sources by loudness
High	Audio captioning (T_4)	Describe an acoustic scene as a text caption

As our main contribution, we propose a novel three-level evaluation protocol for acoustic scene understanding of LALMs as shown in Table 1. We focus on three semantic levels of acoustic scene understanding through four tasks: sound event tagging (low-level), sound polyphony estimation, and the ranking of sound sources by loudness (mid-level), as well as audio captioning (high-level). We implement this protocol using the example of the USM dataset [5], which includes five-second long polyphonic soundscapes generated by systematically mixing isolated sound recordings from the FSD50k dataset [6]. Due to its tonal diversity and detailed annotations of audible sound sources, the dataset allows us to examine the performance of LALMs in various tasks to be examined depending on the sound polyphony, i. e., the number of audible sound sources.

As a concrete use case, we select two top-performing LALMs from the MMAU benchmark and analyze in detail their acoustic scene understanding capabilities. We highlight the different strengths and unexpected pitfalls, including how evaluation metrics fluctuate due to hallucinated predictions. The correlation of the metrics computed for different tasks reveals notable performance trends and inconsistencies depending on the specific evaluation task. We believe these insights can aid in developing more robust and reliable audio-language understanding systems.

2. RELATED WORK

LALMs enhance language models with auditory capabilities. Several benchmarking studies on LALMs highlighted their strengths and limitations in audio perception. AIR-Bench [7] is the first hierarchical benchmark for evaluating tasks across all audio types (speech, music, sound). It includes a foundation benchmark with 19 audio tasks and over 19k single-choice questions, and a chat benchmark with over 2k curated open-ended audio questions. The CompA [1] benchmarks

test the compositional reasoning of LALMs. CompA-order assesses comprehension of event sequences and CompA-attribute evaluates attribute-binding of acoustic events.

Although supervised audio classification with pre-defined labels is well-studied, using audio language models for tasks like counting audio events with structured reasoning remains underexplored. Recent benchmarks like MMAU [2] have introduced structured question-answering tasks that require LALM models to reason with audio beyond just classification. These tasks are pivotal for enhancing machine listening algorithms to interpret sounds in context rather than as isolated events.

Audio captioning generates natural language descriptions of acoustic scenes [8]. It is considered a subtask in Audio Question Answering within LALMs [9]. Recent methods use encoder-decoder models trained on datasets like AudioCaps [10] and Clotho [11] for rich annotations of diverse soundscapes. Similarly, weak label generation [12] [13] transforms metadata or sound class labels into pseudo-captions, aiding multi-modal pretraining and task transfer.

3. METHODOLOGY

3.1. Dataset

Throughout this paper, we use the Urban Sound Monitoring (USM) dataset [14], which includes 24,000 five-second-long two-channel soundscape recordings. The audio clips have been synthesized by mixing isolated sound samples from the FSD50k dataset [6]. Although the dataset does not provide strong labels, i. e., precise time stamps of sound events, it includes weak labels for sound event tagging, focusing on 26 sound classes relevant for urban soundscapes. These sound classes come from six categories of sound: miscellaneous sounds, climate sounds, animal sounds, human-made sounds, construction site sounds, and vehicle sounds. The mixing process of the USM dataset involves defining a random sound polyphony between two and six sound sources, assigning each source a foreground or background role, and assigning it a random sound level between -20 dB and -8 dB for the background sounds and -6 dB to 0 dB for the foreground sounds. The USM dataset's synthetic nature provides precise loudness information for audio events, useful for evaluation. However, it lacks temporal grounding due to missing start and end timestamps, preventing event ordering, which is a limitation. In this work, we use the USM validation subset, which includes 2,000 audio clips.

3.2. Automatic Caption Generation

As a ground-truth for audio caption evaluation, we generate pseudo-natural language captions for the USM dataset following the method described in [13]. As a large language model (LLM), we use the Qwen2.5:7B model [15] motivated by its strong instruction-following and structured output capabilities. We access the model using the open source local LLM serving and research platform Ollama [16]. The model processes the existing metadata of the USM dataset, including the list of foreground and background events and their dynamic level and stereo positioning in the audio clip, provided as a JSON-like input. We found that generated captions such as "The sharp sound of sawing cuts through the air, while in the distance, the rhythmic hum of a train adds a subtle background noise to the scene." well reflect the complex acoustic composition of the generated soundscapes. We will publish the captions generated for the USM data set as a free resource for future audio captioning research.[1]

[1]https://github.com/diliprobbi/lalm-acoustic-scene-understanding

3.3. Evaluation Protocol Task Design

As the first task, *sound event tagging* (SET) extracts a low-level semantic description of an acoustic scene by identifying all audible sound sources. The USM dataset consists of polyphony levels between two and six sources and therefore allows to evaluate how well LALMs can disentangle and correctly identify multiple concurrent sound sources. A closely related task to SET is sound event detection (SED), where the time stamps and durations of individual sound events, emitted by different sound sources must be identified [17]. In addition to a number of existing classical and deep learning-based SED algorithms, methods involving language models have been proposed in [18] [19], enabling more flexible handling of audio classes, including open-vocabulary that predict temporal locations of each class.

For a mid-level semantic description, we incorporate the two tasks *sound polyphony estimation* (SPE) and *ranking-by-loudness* (RBL) to test the fundamental understanding of LALMs' ability to deduce relationships between different sound sources in a soundscape recording. We adopt the SPE task from [20], where it is defined as estimating the number of audible sound sources in a short five-second long acoustic scene recording. The RBL task is used to measure the ability of LALMs to differentiate between loudness levels of sound events in complex soundscapes and thus to distinguish between salient foreground sound events and quieter background noise. This aspect of scene understanding is currently underexplored in LALM research. We consider RBL to be a reasoning task, as it involves detecting, comparing and ordering sound events based on perceptual attributes, demonstrating higher cognitive and inferential abilities beyond basic classification or detection.

Finally, the *audio captioning* (AP) task offers a high-level perspective, as it involves creating a textual description of soundscapes, summarizing all sound sources, events, and their complex tonal and dynamic relationships.

3.4. Evaluated Large Audio Language Models (LALMs)

We select two state-of-the-art LALMs from the MMAU benchmark leaderboard to evaluate their acoustic scene understanding capabilities based on the proposed evaluation protocol.

Qwen2.5-Omni-7B [21] is an instruction-tuned multi-modal model capable of processing text, vision, and audio inputs. It is trained with an architecture optimized for general-purpose multi-modal reasoning tasks. The model incorporates a novel position embedding method entitled Time-aligned multi-modal RoPE (TMRoPE) to synchronize the timestamps of video and audio inputs. We just use the model for audio input.

Audio Flamingo 2 [22] is a cross-modal architecture that uses audio as a primary modality. The model combines a pre-trained audio encoder with a frozen large language model via cross-attention to enable high-quality audio captioning and multi-turn dialogue grounded in sound. Audio Flamingo 2 achieved strong performance on audio-language tasks including AudioCaps and Clotho and is capable of processing audio clips from 30 seconds up to 5 minutes duration. Audio Flamingo 2 was trained on the AudioSkills dataset [22], which is a high-quality skill-specific synthetic dataset with approximately 4.2 million question-answer pairs, designed to enhance expert-level reasoning in LALMs. The targeted skills include temporal reasoning, attribute identification, counting the occurrences of specific sounds, contextual sound event reasoning, contextual speech event reasoning, information extraction, and general reasoning.

Both models prioritize complex, reasoning-intensive questions and achieved top-tier performance on the MMAU benchmark, making

Task	Example Prompt
Sound Event Tagging (T_1)	"Analyze the audio and identify all the audio events by class. Provide just the class names corresponding to each event. Choose from the following options: {string of all classes}. Just give the class names as comma separated values string, e.g., 'class1', 'class2'. Do not include any other text or explanations."
Sound Polyphony Estimation (T_2)	"Analyze how many unique sound events are present in the audio? Just give the number of events in number format. eg. 1, 2, 3"
Sorting-by-loudness (T_3)	"Analyze the audio and return the classes in the order of loudest to softest sounds: {classes}. Just give the class names from loudest to softest. Do not include any other text."
Audio Captioning (T_4)	"Describe the audio as a caption in English in detail, including all the sound events present."

Table 2: Example prompts used for the four tasks towards acoustic scene understanding.

them suitable for the scenario of complex polyphonic soundscapes targeted in our work.

4. EVALUATION AND RESULTS

4.1. Metrics

In the four evaluation tasks T_1 to T_4, we use mainly well-established evaluation metrics. Table 2 lists example prompts we used for each task.

For the SET task (T_1), we compute precision, recall and F1 score as they are standard metrics in multi-label tagging tasks. More specifically, we compute the sample-based F1 score for different subsets of the test set files, for instance, grouped by polyphony level. Therefore, we do not use micro-level or macro-level averaging. After analyzing an audio clip for SET, a LALM outputs a comma-separated list of predicted sound classes, we map these to the 26 sound classes of the USM dataset to create a multi-hot vector. We prompt the model with the entire set of 26 sound class labels of the USM dataset and evaluate its instruction-following ability for the SET task. This approach stands in contrast to other evaluation studies such as [23], where sound classification was devised as multiple binary classification tasks to evaluate the presence or absence of specific sounds.

In the SPE task (T_2), we compute the mean absolute error (MAE) between the true and estimated number of sound sources.

In the RBL task (T_3), we apply the Normalized Discounted Cumulative Gain (nDCG) metric [24] with linear gain, which is a widely used metric for evaluating the quality of ranking systems. Due to the positional sensitivity of this metric, systems that accurately rank the most relevant item in the list are rewarded more. In our case, relevance refers to the loudness of sound sources. We sort the sound sources in the descending order from the loudest to the softest sounds in an audio clip. The highest relevance score rel_i is assigned to the loudest sound event and the lowest relevance is given to the softest sound event. During the evaluation in task T_3, we first prompt the model with the list of ground truth sound sources in a given clip and query the LALM to rank them from loudest to softest. We compute the nDCG metric using the following steps:

Given an audio clip with 4 sound sources, the relevance scores are set to $rel := \{4, 3, 2, 1\}$ to emphasize the importance of louder sounds at the first rank positions. Then, we compute the DCG (Discount Cumulative Gain), IDCG (Ideal Discount Cumulative Gain) for perfect ranking order of events and WDCG (Worst Discount Cumulative Gain) for reversed ranking order of events as:

		Sound Event Tagging			Sound Polyphony Estimation	Ranking By Loudness	Audio Captioning
Model	P_G	Precision ↑	Recall ↑	F1 ↑	MAE ↓	nDCG ↑	FENSE ↑
Audio Flamingo 2	2	0.114	0.440	0.140	1.748	0.510	0.372
	3	0.180	0.452	0.190	1.577	0.508	0.379
	4	0.189	0.418	0.197	2.032	0.457	0.384
	5	0.220	0.424	0.220	2.503	0.439	0.389
	6	0.242	0.448	0.247	3.125	0.428	0.391
	Mean	0.191	0.433	0.200	2.131	0.467	0.384
Qwen2.5 Omni	2	0.446	0.298	0.346	0.431	0.551	0.504
	3	0.506	0.239	0.314	0.660	0.550	0.470
	4	0.507	0.196	0.274	1.335	0.498	0.458
	5	0.536	0.162	0.239	2.260	0.486	0.464
	6	0.525	0.148	0.225	3.142	0.449	0.445
	Mean	0.509	0.204	0.278	1.500	0.507	0.466

Table 3: Performance metrics of Audio Flamingo 2 and Qwen 2.5 Omni across Polyphony Level (P_G).

$$\text{DCG@k} = \sum_{i=1}^{k} \frac{rel_i}{\log_2(i+1)}, \quad \text{IDCG@k} = \sum_{i=1}^{k} \frac{rel_{\sigma(i)}}{\log_2(i+1)},$$

$$\text{WDCG@k} = \sum_{i=1}^{k} \frac{rel_{\tau(i)}}{\log_2(i+1)}$$

where rel_i, $rel_{\sigma(i)}$, and $rel_{\tau(i)}$ are relevance scores at position i in the actual, ideal (descending), and worst-case (ascending) rankings, respectively. Finally, we apply min-max normalization while calculating nDCG to ensure that worst ranking scores are scaled down to zero across different lengths of polyphony levels as:

$$\text{nDCG@k}_j = \frac{\text{DCG@k} - \text{WDCG@k}}{\text{IDCG@k} - \text{WDCG@k}}$$

where k is the predicted ranked list length for the j-th audio sample ($j = 1, \ldots, 2000$). This approach scales the nDCG metric to a range between 0 and 1, similar to the approach in [25].

Finally, we use the FENSE (Fluency ENhanced Sentence-bert Evaluation) metric [26] for the AC task (T_4). This metric is commonly used in audio captioning as it assesses the faithfulness of a generated caption to the audio content and its linguistic quality by leveraging pretrained audio-text and language models.

4.2. Results

Table 3 shows the different evaluation metrics computed across test files of different polyphony levels for the two LALMs under comparison. The results show contrasting performance in polyphonic SET, with Qwen2.5 Omni's F1 Score declining as polyphonic complexity increases. Audio Flamingo 2 (AF2) shows higher recall at each polyphony level but at a lower overall precision (0.191) compared to Qwen2.5 Omni (0.509), indicating AF2 recognizes more audio events but is more prone to false positives, potentially hallucinating audio events not present.

In the SPE task, Qwen2.5 Omni shows superior performance with a mean MAE of 1.500, compared to Audio Flamingo 2's 2.131, suggesting that it more accurately estimates audio sources across all P_G levels. Qwen2.5 Omni consistently achieves lower errors, especially at lower levels of P_G (e. g. 0.431 at $P_G = 2$ compared to 1.748 for Audio Flamingo 2), highlighting its effectiveness in simpler cases. As the number of simultaneous sound sources increases in an audio scene, LALMs show higher MAE, indicating the challenge of counting sources in complex, overlapping soundscapes.

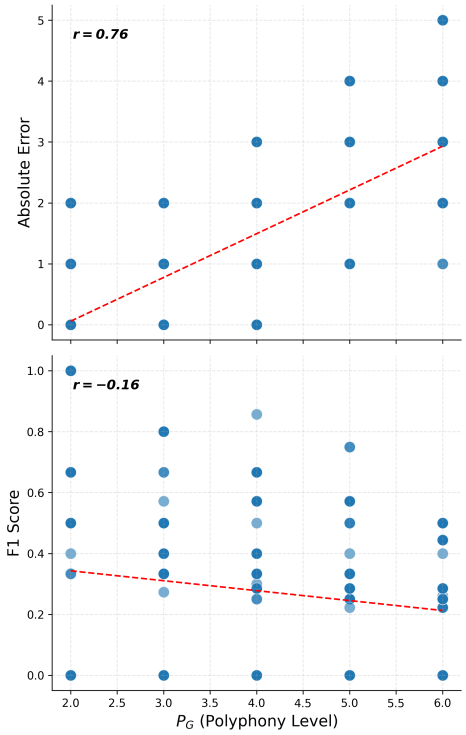

Fig. 1: Pairwise scatter plots for Qwen 2.5 Omni, illustrating the pearson correlation between the computed metrics (F_1, Absolute Error) and Polyphony level (P_G) both with p-value < 0.001.

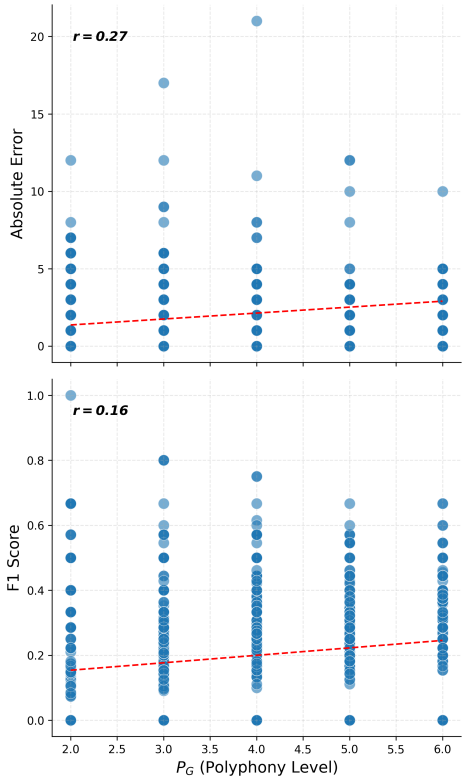

Fig. 2: Pairwise scatter plots for Audio Flamingo 2, illustrating the Pearson correlation between the computed metrics (F_1, Absolute Error) and Polyphony Level (P_G) both with p-value < 0.001.

A correlation analysis between P_G and $F1$ in Figures 1 and 2 shows a positive correlation for Absolute Error, as observed its value increases with higher polyphony levels. This underscores the difficulty in source estimation task. The F1 score in Fig. 1 exhibits a negative correlation as polyphony increases, while Table 3 show improving precision indicate lower number of hallucinations with increasing polyphony. Audio Flamingo 2 as in Fig. 2 shows a slightly positive correlation in F1 score for audio class predictions with lower values of precision that indicate greater hallucinations.

The nDCG correlation scatter plots show very weak negative correlation values r = -0.09 (AF2) and r = -0.08 (Qwen 2.5 Omni). While FENSE score correlation values are r = 0.03 (AF2) and r = -0.09 (Qwen 2.5 Omni). The plots are omitted here, and further investigation is needed to ascertain if these metrics are statistically independent of increasing polyphony. Table 3 indicates no clear relationship in loudness ranking in complex polyphonic scenes, with nDCG values not effectively distinguishing event loudness between audio events. FENSE scores weakly correlate with polyphony, suggesting FENSE may emphasize sentence coherence and semantic similarity than on audio event accuracy in captions.

5. CONCLUSION

This study evaluates LALM for understanding the acoustic scene based on the USM dataset, which focuses on urban soundscapes with overlapping sound sources. We proposed a novel three-stage evaluation protocol which includes four machine listening tasks selected to measure scene understanding on different semantic levels. Despite achieving high scores on the MMAU benchmark, the two LALMs evaluated need further improvements in training procedures to understand complex acoustic scenes. Future direction aims to enhance evaluation procedures through metrics that account for hallucinations and omissions.

6. LIMITATIONS

Synthetic datasets like USM replicate natural environmental soundscapes and alleviate the laborious task of labeling datasets with loudness values for various sound events, even if such datasets were available. Moreover, they offer precise manipulation of the number, timing and loudness of overlapping events. However, the dataset employs random mixing and stereo placement, which do not adequately represent the spatial and temporal dynamics of real-world environments, such as moving sources or gradual loudness changes. Additionally, USM's definition of polyphony is more lenient, as it counts the number of active sounds within a short segment instead of the exact count of simultaneous overlapping events, which may not faithfully depict the intricacies of natural polyphonic soundscapes. Models struggle to caption or classify audio classes with very similar acoustic characteristics like car, bus, motorcycle accurately, generalizing them at times to a common sound class as "engine".

The technique of generating captions for weakly labeled soundscapes, with tags and metadata like the loudness of each audio event and their spatial positioning (foreground versus background), does not ensure that the resulting captions will accurately reflect the soundscape. There are still cases of misinterpreted audio event attributes in the reference captions.

ACKNOWLEDGMENTS

This research was funded by the Federal Ministry of Education and Research in Germany (BMBF) within the project news-polygraph (funding code 03RU2U151D).

REFERENCES

[1] S. Ghosh, A. Seth, S. Kumar, U. Tyagi, C. K. R. Evuru, R. S, S. Sakshi, O. Nieto, R. Duraiswami, and D. Manocha, "Compa: Addressing the gap in compositional reasoning in audio-language models," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=86NGO8qeWs

[2] S. Sakshi, U. Tyagi, S. Kumar, A. Seth, R. Selvakumar, O. Nieto, R. Duraiswami, S. Ghosh, and D. Manocha, "MMAU: A massive multi-task audio understanding and reasoning benchmark," in *Proceedings of the Thirteenth International Conference on Learning Representations*, 2025. [Online]. Available: https://openreview.net/forum?id=TeVAZXr3yv

[3] S. Leng, Y. Xing, Z. Cheng, Y. Zhou, H. Zhang, X. Li, D. Zhao, S. Lu, C. Miao, and L. Bing, "The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio," *arXiv preprint arXiv:2410.12787*, 2024.

[4] C.-Y. Kuan, W.-P. Hsu, and H.-y. Lee, "Understanding sounds, missing the questions: The challenge of object hallucination in large audio-language models," in *Proceedings of Interspeech 2024*, Kos, Greece, 2024, pp. 4144–4148.

[5] J. Abeßer, "Classifying Sounds in Polyphonic Urban Sound Scenes," in *Proceedings of the 152nd Audio Engineering Society (AES) Convention*, Online, 2022.

[6] 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, 2022.

[7] Q. Yang, J. Xu, W. Liu, Y. Chu, Z. Jiang, X. Zhou, Y. Leng, Y. Lv, Z. Zhao, C. Zhou, and J. Zhou, "AIR-bench: Benchmarking large audio-language models via generative comprehension," in *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, L.-W. Ku, A. Martins, and V. Srikumar, Eds. Bangkok, Thailand: Association for Computational Linguistics, August 2024, pp. 1979–1998. [Online]. Available: https://aclanthology.org/2024.acl-long.109/

[8] K. Drossos, S. Adavanne, and T. Virtanen, "Automated audio captioning with recurrent neural networks," in *Proceedings of the 2017 IEEE Workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA 2017), New Paltz, NY, USA, October 15-18, 2017*. IEEE, 2017, pp. 374–378. [Online]. Available: https://doi.org/10.1109/WASPAA.2017.8170058

[9] A. K. Sridhar, Y. Guo, and E. Visser, "Enhancing temporal understanding in audio question answering for large audio language models," in *Proceedings of the 2025 Conference of the North American Chapter of the Association for Computational Linguistics: Industry Track*. Mexico City, Mexico: Association for Computational Linguistics, 2025, pp. 799–809. [Online]. Available: https://aclanthology.org/2025.naacl-industry.78.pdf

[10] 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)*, J. Burstein, C. Doran, and T. Solorio, Eds. Minneapolis, Minnesota: Association for Computational Linguistics, June 2019, pp. 119–132. [Online]. Available: https://aclanthology.org/N19-1011/

[11] K. Drossos, S. Lipping, and T. Virtanen, "Clotho dataset," October 2019. [Online]. Available: https://doi.org/10.5281/zenodo.3490684

[12] S. Hershey, D. P. W. Ellis, E. Fonseca, A. Jansen, C. Liu, R. C. Moore, and M. Plakal, "The benefit of temporally-strong labels in audio event classification," in *Proceedings of the 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2021, pp. 366–370.

[13] 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," *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, pp. 1–15, 2024.

[14] J. Abeßer, S. Grollmisch, and M. Müller, "How robust are audio embeddings for polyphonic sound event tagging?" *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, vol. 31, pp. 2658–2667, 2023.

[15] Q. Team, "Qwen2.5: A party of foundation models," Website. https://qwenlm.github.io/blog/qwen2.5/, September 2024.

[16] Ollama contributors, "Ollama," 2025, version 2025, accessed July 9, 2025. [Online]. Available: https://ollama.com

[17] A. Mesaros, T. Heittola, T. Virtanen, and M. D. Plumbley, "Sound Event Detection: A tutorial," *IEEE Signal Processing Magazine*, vol. 38, no. 1, pp. 67–83, 2021.

[18] H. Wang, J. Mao, Z. Guo, J. Wan, H. Liu, and X. Wang, "Leveraging language model capabilities for sound event detection," in *Proceedings of Interspeech 2024*, 2024, pp. 4803–4807.

[19] Y. Gong, H. Luo, A. H. Liu, L. Karlinsky, and J. R. Glass, "Listen, think, and understand," in *Proceedings of the Twelfth International Conference on Learning Representations*, 2024. [Online]. Available: https://openreview.net/forum?id=nBZBPXdJlC

[20] J. Abeßer, A. Ullah, S. Ziegler, and S. Grollmisch, "Human and machine performance in counting sound classes in single-channel soundscapes," *Journal of the Audio Engineering Society (JAES)*, vol. 71, no. 12, pp. 860–872, 2023.

[21] J. Xu, Z. Guo, J. He, H. Hu, T. He, S. Bai, K. Chen, J. Wang, Y. Fan, K. Dang, B. Zhang, X. Wang, Y. Chu, and J. Lin, "Qwen2.5-omni technical report," *arXiv preprint arXiv:2503.20215*, 2025.

[22] S. Ghosh, Z. Kong, S. Kumar, S. Sakshi, J. Kim, W. Ping, R. Valle, D. Manocha, and B. Catanzaro, "Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities," in *Proceedings of the Forty-second International Conference on Machine Learning*, 2025. [Online]. Available: https://openreview.net/forum?id=xWu5qpDK6U

[23] C.-Y. Kuan and H.-y. Lee, "Can large audio-language models truly hear? tackling hallucinations with multi-task assessment and stepwise audio reasoning," in *Proceedings of the 2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[24] K. Järvelin and J. Kekäläinen, "Cumulated gain-based evaluation of ir techniques," *ACM Transactions on Information Systems*, vol. 20, no. 4, pp. 422–446, 2002.

[25] L. Gienapp, M. Fröbe, M. Hagen, and M. Potthast, "The impact of negative relevance judgments on ndcg," in *Proceedings of the 29th ACM International Conference on Information Knowledge Management*. Virtual Event, Ireland: ACM, October 2020. [Online]. Available: https://dl.acm.org/doi/10.1145/3340531.3412123

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