AudioRAG: A Challenging Benchmark for Audio Reasoning and Information Retrieval

Jingru Lin JINGRULIN@U.NUS.EDU
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore

Chen Zhang E0397123@U.NUS.EDU
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore

Tianrui Wang WANGTIANRUI@TJU.EDU.CN
College of Intelligence and Computing,
Tianjin University, China

Haizhou Li HAIZHOU.LI@NUS.EDU.SG
Department of Electrical and Computer Engineering,
National University of Singapore, Singapore
SRIBD, School of Data Science,
The Chinese University of Hong Kong, Shenzhen, Guangdong

Editors: Tatsuya Komatsu, Keisuke Imoto, Xiaoxue Gao, Nobutaka Ono, Nancy F. Chen

Abstract

Due to recent advancements in Large Audio-Language Models (LALMs) that demonstrate remarkable performance across a range of sound-, speech- and music-related tasks, there is a growing interest in proposing benchmarks to assess these models. Existing benchmarks generally focus only on reasoning with internal knowledge, neglecting real-world scenarios that require external information grounding. To bridge this gap, we introduce AudioRAG, a novel benchmark designed to evaluate audio-based reasoning augmented by information retrieval in realistic web environments. This benchmark comprises both LLM-generated and manually curated question-answer pairs. Our evaluations reveal that even the state-of-the-art LALMs struggle to answer these questions. We therefore propose an agentic pipeline that integrates audio reasoning with retrieval-augmented generation, providing a stronger baseline for future research.

Keywords: Audio Reasoning, Information Retrieval

1. Introduction

Advancements in Large Audio-Language Models (LALMs) highlight their potential to unify speech, sound and music understanding within a single multimodal framework Chu et al. (2024); Kong et al. (2024); Ghosh et al. (2025). These advanced LALMs exhibit impressive performance across a wide range of understanding and generation tasks involving speech, sound, and music, such as speech recognition Ma et al. (2025b), music genre classification Meguenani et al. (2025), audio captioning Chen et al. (2025), etc. While these tasks assess foundational audio understanding, they largely emphasize perceptual recognition rather than complex reasoning, which characterizes more sophisticated forms of intelligence.

Recently, there has been a growing research interest in exploring the capacity of LALMs to perform multi-hop or deep reasoning, and several benchmarks have been proposed for this purpose Yang et al. (2025b); Ma et al. (2025a); Sakshi et al. (2025). For example, SAKURA Yang et al. (2025b) evaluates multi-hop reasoning capability. MMAR Ma et al. (2025a) evaluates the LALMs' deep reasoning capabilities, including expert-level perceptual understanding and multi-step inference. However, these benchmarks primarily assess reasoning over internal, parameterized knowledge only.

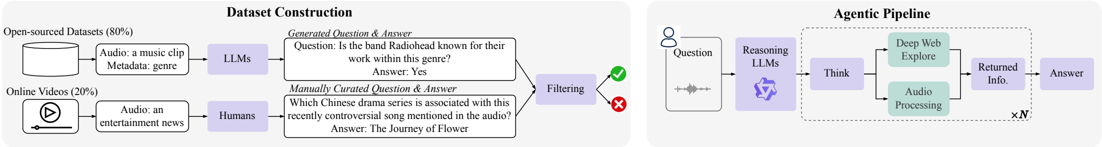
In real-world scenarios, users often ask questions that extend beyond a model's internal knowledge. For example, a user might inquire about the final score of a soccer match mentioned in a recent news broadcast. If such information is absent from the model's training data, the model may generate an unfaithful response. This problem is often recognized as *hallucination* in large language model (LLM) research Huang et al. (2025b). To alleviate *hallucination*, retrieval-augmented generation (RAG) has been widely adopted, as it grounds the model outputs in retrieved and verifiable knowledge Izacard et al. (2023); Gao et al. (2023). By coupling generation with retrieval, RAG effectively enhances factual accuracy and reduces reliance on internal memorization. Nevertheless, existing audio reasoning benchmarks have not yet accounted for scenarios where audio-based reasoning is combined with retrieval.

To address this gap, we introduce AudioRAG, a new benchmark featuring challenging questions tailored for audio reasoning with information seeking under real-world web environments. AudioRAG is constructed based on both open-sourced datasets and human-collected data. For open-sourced datasets, we take one attribute of the audio, e.g., genre of the music, species of the animal, from the metadata and prompt LLMs to generate questions based on that attribute. The generated questions are designed to require both audio processing and information retrieval steps. To further diversify the dataset, we manually collect videos from online sources and extract their audio tracks. Based on the extracted audios, we curate multi-hop questions similar to the ones generated by LLMs from the open-sourced datasets.

Based on the benchmark, we evaluate several state-of-the-art LALMs. Our results show that existing LALMs struggle to answer these audio-based multi-hop reasoning questions. This observation is consistent with findings from previous studies Ma et al. (2025a); Yang et al. (2025b). Moreover, when lacking the necessary knowledge to answer the question, the LALMs provide hallucinated responses. To better address such challenges, we propose an agentic pipeline capable of performing both audio reasoning and information retrieval, serving as a strong baseline for future research.

Overall, our contributions are: (1) Introducing AudioRAG, the first benchmark for systematically evaluating models' multi-hop audio reasoning and information retrieval capabilities; (2) We evaluate LALMs on the benchmark, revealing that current LALMs struggle to answer the questions; (3) We develop an agentic pipeline that shows a relative improvement of up to 24.9%.

Figure 1: Data construction process (left) and the agentic pipeline (right).


Dataset Construction
Open-sourced Datasets (80%)
Audio: a music clip
Metadata: genre
LLMs
Generated Question & Answer
Question: Is the band Radiohead known for their work within this genre?
Answer: Yes
Online Videos (20%)
Audio: an entertainment news
Humans
Manually Curated Question & Answer
Which Chinese drama series is associated with this recently controversial song mentioned in the audio?
Answer: The Journey of Flower
Filtering
Agentic Pipeline
Question
Reasoning LLMs
Think
Deep Web Explore
Audio Processing
Returned Info.
Answer
×N

2. Methods

Each question in our dataset consists of a multi-hop question based on an audio context and the corresponding answer. The question is either in text or audio format, while the context is always in audio format. The multi-hop question is either generated based on open-source datasets or manually curated by humans. The following sections give details about the constructed dataset.

2.1. Dataset Statistics

We first examine many open-source datasets and select the following for our dataset construction:

- **Speech/Sound/Music**: MMAU Sakshi et al. (2025), CinePile Rawal et al. (2024), Multitask-National-Speech-Corpus Wang et al. (2025), FMA Defferrard et al. (2017), Jazznet Adegbija (2023), MusicNet Thickstun et al. (2017), iNaturalist Chasmai et al. (2025)

- **Text**: CHEER Huang et al. (2025a)

On top of these open-source datasets, we also collect additional audio from online sources, then manually curate multi-hop questions based on the collected audio.

Overall, we collect and generate 500 samples and release them at `https://github.com/jingru-lin/AudioRAG`[1].

2.2. Question Generation

We utilize GPT-4o to generate the multi-hop question-answer pairs. Specifically, we take one attribute of the audio from metadata of the open-source datasets, for example, genre of music, species of animals heard in the audio, transcriptions of speech, and prompt GPT-4o to generate multi-hop questions that might require external knowledge to answer. Few-shot examples are provided in the prompts. Different templates and few-shot examples are needed for the audio samples from different datasets. Here, we only give an example of the prompt to illustrate the idea, as shown in Table 1. The generated question should resemble a true user question after hearing the audio. Figure 1 shows an example.

In addition, we manually collect some data from online sources. These questions contain timely information and are less likely to be exposed to data contamination. Therefore, current models lack sufficient parametric knowledge about the information required to answer

[1] Details about the original sample sources are also provided.

Table 1: Prompts for the LLM to generate the multi-hop questions. Here, [ATTRIBUTE] refers to an attribute of the audio, such as the genre of music, the source of a sound etc.

A prompt example to generate the multi-hop questions
You are good at creating multi-hop questions. Your task is to generate multi-hop questions based on the given information about [ATTRIBUTE]. **Requirements**: 1. Do not mention the given [ATTRIBUTE] directly. You should assume an audio is given and the [ATTRIBUTE] must be derived from the audio. You can refer to it indirectly (e.g., using pronouns). 2. If possible, create a question that might involve an information-retrieval step to answer the question. 3. The final question must arrive at a clear correct answer. 4. Phrase the question naturally so it reads like a real user query. 5. Provide an answer to the questions. 6. No questions about the recognition of [ATTRIBUTE] should be generated. Examples: [FEW-SHOT_EXAMPLES]

the question, which makes retrieving external information highly important. An example question is shown in Figure 1.

2.3. Data Filtering

We employ two filters: a) Question Validity Filter; b) Answer Correctness Filter. The first ensures the quality of the generated questions. Specifically, we employ both LLM and annotators to verify whether the question has one unique answer. Questions lacking a unique answer are removed. The second filter focuses on verifying the answer correctness. To ensure the correctness of answers generated, we employ LLM-based agents with a search tool that allows for up-to-date information retrieval when answering the questions. We provide the ground-truth audio attributes (text) that were used to generate the questions for agents to answer the question. In such a way, it reduces the noise introduced when processing the audio, and the agents can focus exclusively on answer correctness. One example is shown in Appendix A. When the agents' answers differ from the originally generated ones, human annotators review those cases and the corresponding questions are either revised or discarded.

3. Agentic Pipeline

Our initial results show that the most open-source Large Audio Language Models (LALMs) struggle to answer these multi-hop questions. There are two main reasons. Firstly, the LALMs, while capable of handling audio data, show weak text capabilities, including instruction following, question interpretation, etc. Therefore, most LALMs struggle to break down the non-straightforward multi-hop questions into solvable subtasks. Secondly, the

LALMs lack the knowledge, especially timely information, to answer the question. To tackle these problems, we propose an LLM-based agentic pipeline. Our pipeline makes use of a text-based LLM to handle user queries and drive tool usage. This pipeline is integrated with two tools: an audio processing tool and a search tool. While the former compensate for the text capabilities that the large language models lack, the latter complement the model with external knowledge. We will introduce the details of the pipeline in the following sections.

3.1. Pipeline Details

Our pipeline is based on WebThinker Li et al. (2025a), but we enhance it with audio-processing capabilities. Given a user query, consisting of a question q that is either in text or audio form and an audio context c, the pipeline generates a solution for answering the user's query regarding the audio context, guided by an instruction I. To answer the complex, multi-hop queries, the pipeline implements an autonomous **Think-Call-Answer** strategy, which iteratively thinks upon the current status, calls relevant tools, and generates intermediate/final responses. The final solution comprises a reasoning chain $\mathcal{R}$ and a final output y. The reasoning LLM in the pipeline orchestrates the whole process and autonomously invokes tools from an available set $\mathcal{T}$ during its reasoning process. The whole process can be formalized as:

$$P(R, y \mid I, q, \mathcal{T}) = \underbrace{\prod_{t=1}^{T_r} P(R_t \mid R_{<t}, I, q, \{O_\tau\}_{\tau<t})}_{\text{Reasoning with Tools}} \cdot \underbrace{\prod_{t=1}^{T_y} P(y_t \mid y_{<t}, R, I, q)}_{\text{Final Output Generation}}, \tag{1}$$

where $\mathcal{T}_r$ is the number of tokens in the reasoning sequence $\mathcal{R}$. The token at position t is $\mathcal{R}_t$, and $\mathcal{R}_{<t}$ represents all tokens generated before position t. Similarly, T_y is the length of the output sequence y, with y_t being the token at position t and $y_{<t}$ indicating all generated output tokens before position t. $\{O_\tau\}_{\tau<t}$ denotes the outputs of all tool calls made before position t. The tool set consists of two tools $\mathcal{T} = \{\mathcal{T}_{\text{exp}}, \mathcal{T}_a\}$. $\mathcal{T}_{\text{exp}}$ is a deep web explorer tool that can retrieve information online and $\mathcal{T}_a$ is an audio processing tool that can extract necessary information from the c. During the reasoning process, the reasoning LLM needs to iteratively call one of the tools and generate intermediate outputs.

3.2. Tools

When the LLM's internal knowledge is not enough to answer the user query, it can invoke the search tool $\mathcal{T}_{\text{exp}} \in \mathcal{T}$ to extract external knowledge. The deep web explorer tool is directly adopted from the WebThinker Li et al. (2025a), which conducts search actions and web browsing actions[2]. To extract relevant information from the audio, an audio

[2] Interested audiences may refer to the original paper for details.

Table 2: Main results. Since the agentic pipeline consists of audio and text LLMs, the size is broken down into size of (audio) + (text). The size of Gemini-2.5-Flash is unknown as it is a closed-source model. The rest are open-source models.

Model	Size	Accuracy (%)
Raw Model		
Qwen2.5-Omni Xu et al. (2025a)	7B	32.2
Audio Flamingo 3 Goel et al. (2025)	8.3B	28.8
Audio-Reasoner Xie et al. (2025)	8.4B	20.2
Baichuan-Omni Li et al. (2025b)	11B	24.4
Qwen3-Omni Xu et al. (2025b)	30B	37.0
Gemini-2.5-Flash Comanici et al. (2025)	-	**45.0**
Agentic Pipeline		
Qwen2.5-Omni + Qwen3-8B Yang et al. (2025a)	7B + 8B	39.5
Qwen3-Omni + Qwen3-8B Yang et al. (2025a)	30B + 8B	**46.2**

processing tool is needed. When an audio processing step is needed, the reasoning LLM will generate an audio-related query that extracts information from the audio. The query is wrapped in <begin_audio_analysis>[AUDIO-RELATED QUERY]<end_audio_analysis>. Upon receiving the query, the pipeline will initiate the audio processing tool, which will then output a response O_a based on the instruction.

4. Experiments

4.1. Raw Models

We evaluate several open- and closed-sourced models, including Qwen2.5-Omni Xu et al. (2025a), Audio Flamingo 3 Goel et al. (2025), Audio-Reasoner Xie et al. (2025), Baichuan-Omni Li et al. (2025b), Qwen3-Omni Xu et al. (2025b) and Gemini-2.5-Flash Comanici et al. (2025).

4.2. Agentic Pipeline

We use Qwen3-8B Yang et al. (2025a) as the reasoning LLM to drive the whole pipeline. The search tool $\mathcal{T}_{\text{exp}}$ used is the Google search engine API. The audio processing tool is Qwen2.5-Omni Xu et al. (2025a) / Qwen3-Omni Xu et al. (2025b). Both reasoning LLM and audio processing tool are served with vLLM Kwon et al. (2023), with four A100 (40GB) GPUs each.

4.3. Main Results

Table 2 shows the performance of both raw models and agentic pipelines on the benchmark. The reported metric is accuracy (%), which records to number of correct questions answered. Each question is provided with an answer and GPT-4o is used to evaluate the generated

Table 3: Case study comparing outputs from a raw model and agentic pipeline. Reasoning error is highlighted in red.

Audio: an audio clip about "The Wizard of Oz"
Question: How is the movie mentioned in this audio related to Wicked?

Answer by Qwen3-Omni:
The movie mentioned is **The Wizard of Oz**. The audio does not mention the movie **Wicked** at all.
Answer by Agentic Pipeline with Qwen3-Omni:
Wicked is a prequel to The Wizard of Oz, exploring...
Ground-truth Answer:
Wicked is a prequel to The Wizard of Oz

answer. The evaluation prompt is shown in Appendix B. The results in Table 2 are averaged over three runs.

From Table 2, the strongest raw model is the closed-source Gemini-2.5-Flash, achieving 45% accuracy, followed by open-sourced Qwen3-Omni with 37.0%. All other open-sourced models perform notably worse, highlighting a clear performance gap between open-source and closed-source models. This is likely due to the latter's stronger internal knowledge or agentic abilities. Among all open-sourced models, the Qwen family consistently outperforms others, demonstrating stronger audio reasoning capabilities.

Compared to the raw models, the agentic pipeline consistently delivers superior performance. Specifically, Qwen2.5-Omni and Qwen3-Omni achieve accuracies of 32.2% and 37.0%, respectively. When integrated into the agentic pipeline with Qwen3-8B as the reasoning model, their accuracies rise to 39.5% and 46.2%, corresponding to relative improvements of 22.7% and 24.9%.

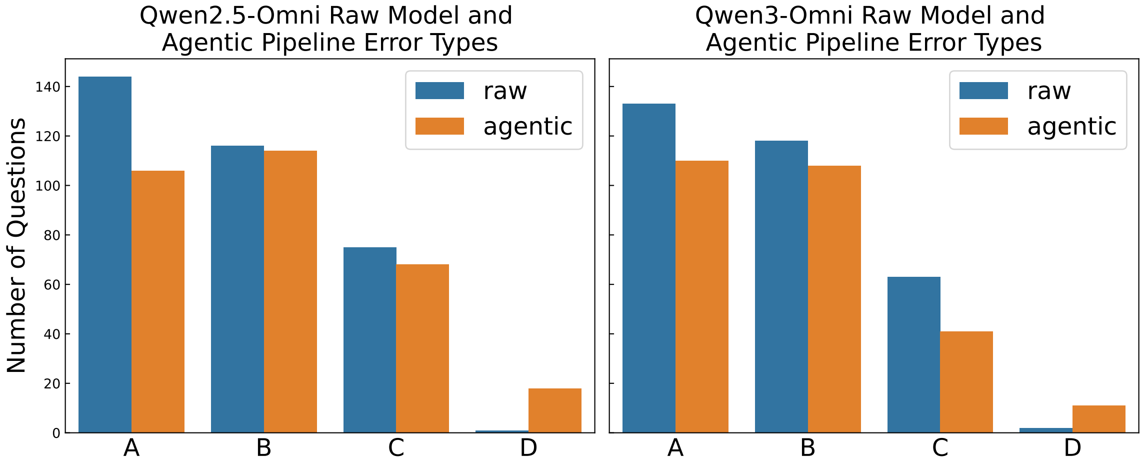
Table 3 shows an example of actual outputs from a raw model and an agentic pipeline. The question in this example requires reasoning beyond the information directly available in the audio content. The raw model focuses purely on the audio input, correctly identifying the mentioned movie but failing to infer the link between the two movies. In contrast, the agentic pipeline demonstrates a clearer understanding of the question and effectively retrieves external knowledge to identify the correct relationship between the two movies.

Table 2 and 3 both show that the agentic pipeline bridges the gap in reasoning and knowledge integration between raw models and human-like multimodal understanding. While the raw models demonstrate perceptual capabilities, their performance is limited by a lack of contextual reasoning and external knowledge grounding. The agentic pipeline addresses these limitations.

4.4. Analysis on Error Types

In this section, we analyze the different error types made by the raw models and the agentic pipelines. We summarise four error types: a) Reasoning Error, where the model fails to understand the question or apply correct reasoning to arrive at the answer; b) Audio Processing Error, where the question requires to recognise a specific attribute of the audio but

Figure 2: Error types breakdown for incorrect answers. The x-axis is the category of errors. A refers to Reasoning Error, B refers to Audio Processing Error, C refers to Knowledge Error and D refers to Invalid Answer.



the model wrongly recognise it; c) Knowledge Error, where the question requires external knowledge beyond what is in the audio but the model gives the wrong answer due to lack of or incorrect knowledge; d) Invalid Answer: the model gives no response or incomplete output. GPT-4o is used to analyze the answers and categorize the incorrect answers into one of these categories. Figure 2 shows the number of questions belonging to each error category. The evaluation prompt is shown in Appendix B. This plot is based on outputs from raw models Qwen2.5-7B and Qwen3-30B and their corresponding agentic pipeline coupled with Qwen3-8B.

From Figure 2, we observe that the agentic pipeline reduces Reasoning, Audio Processing and Knowledge Error, but tends to produce more Invalid Answer. The largest reduction is in Reasoning Error. This indicates that raw models often struggle to interpret complex multi-hop questions but the structured multi-step reasoning of the agentic pipeline significantly improves comprehension. The second major improvement lies in Knowledge Error. This is due to the knowledge retrieval component in the agentic pipeline which enhances the model with up-to-date and factually accurate information. Improvements in Audio Processing Error are smaller, as the pipeline still relies on the raw model to process audio. Invalid Answer increases, likely because the more complex multi-step reasoning can lead the pipeline into infinite logical loops.

5. Conclusion

In this paper, we introduce AudioRAG, the first benchmark designed to evaluate models' multi-hop audio reasoning and information retrieval capabilities. Our experiments show that current LALMs perform poorly on this challenging benchmark, while an agentic pipeline that integrates audio reasoning with information retrieval establishes a stronger baseline for future research.

6. Acknowledgements

This work is supported by Shenzhen Stability Science Program 2023, Shenzhen Key Lab of Multi-Modal Cognitive Computing, Shenzhen Science and Technology Program ZDSYS20230626091302006 and the Program for Guangdong Introducing Innovative and Entrepreneurial Teams, Grant No. 2023ZT10X044.

References

Tosiron Adegbija. jazznet: A dataset of fundamental piano patterns for music audio machine learning research. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

Mustafa Chasmai, Alexander Shepard, Subhransu Maji, and Grant Van Horn. The inaturalist sounds dataset, 2025. URL https://arxiv.org/abs/2506.00343.

Wenxi Chen, Ziyang Ma, Xiquan Li, Xuenan Xu, Yuzhe Liang, Zhisheng Zheng, Kai Yu, and Xie Chen. Slam-aac: Enhancing audio captioning with paraphrasing augmentation and clap-refine through llms. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2025.

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, et al. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*, 2024.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Michaël Defferrard, Kirell Benzi, Pierre Vandergheynst, and Xavier Bresson. FMA: A dataset for music analysis. In *18th International Society for Music Information Retrieval Conference (ISMIR)*, 2017. URL https://arxiv.org/abs/1612.01840.

Yunfan Gao, Yun Xiong, Xinyu Gao, Kangxiang Jia, Jinliu Pan, Yuxi Bi, Yixin Dai, Jiawei Sun, Haofen Wang, and Haofen Wang. Retrieval-augmented generation for large language models: A survey. *arXiv preprint arXiv:2312.10997*, 2(1), 2023.

Sreyan Ghosh, Zhifeng Kong, Sonal Kumar, S Sakshi, Jaehyeon Kim, Wei Ping, Rafael Valle, Dinesh Manocha, and Bryan Catanzaro. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=xWu5qpDK6U.

Arushi Goel, Sreyan Ghosh, Jaehyeon Kim, Sonal Kumar, Zhifeng Kong, Sang-gil Lee, Chao-Han Huck Yang, Ramani Duraiswami, Dinesh Manocha, Rafael Valle, et al. Audio flamingo 3: Advancing audio intelligence with fully open large audio language models. *arXiv preprint arXiv:2507.08128*, 2025.

Chen Huang, Junkai Luo, Xinzuo Wang, Wenqiang Lei, and Jiancheng Lv. Can large language models understand internet buzzwords through user-generated content. In *Annual Meeting of the Association for Computational Linguistics*, 2025a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025b.

Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Atlas: Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 24(251):1–43, 2023.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. Audio flamingo: A novel audio language model with few-shot learning and dialogue abilities. In *International Conference on Machine Learning*, pages 25125–25148. PMLR, 2024.

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

Xiaoxi Li, Jiajie Jin, Guanting Dong, Hongjin Qian, Yutao Zhu, Yongkang Wu, Ji-Rong Wen, and Zhicheng Dou. Webthinker: Empowering large reasoning models with deep research capability. *CoRR*, abs/2504.21776, 2025a. doi: 10.48550/ARXIV.2504.21776. URL https://doi.org/10.48550/arXiv.2504.21776.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*, 2025b.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, et al. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*, 2025a.

Ziyang Ma, Guanrou Yang, Yifan Yang, Zhifu Gao, Jiaming Wang, Zhihao Du, Fan Yu, Qian Chen, Siqi Zheng, Shiliang Zhang, et al. Speech recognition meets large language model: Benchmarking, models, and exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 24840–24848, 2025b.

Mohamed El Amine Meguenani, Alceu de Souza Britto, and Alessandro Lameiras Koerich. Music genre classification using large language models. In *2025 IEEE Symposium on*

Computational Intelligence in Image, Signal Processing and Synthetic Media (CISM), pages 1–7. IEEE, 2025.

Ruchit Rawal, Khalid Saifullah, Ronen Basri, David Jacobs, Gowthami Somepalli, and Tom Goldstein. Cinepile: A long video question answering dataset and benchmark. *arXiv preprint arXiv:2405.08813*, 2024.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. Mmau: A massive multi-task audio understanding and reasoning benchmark. In *The Thirteenth International Conference on Learning Representations*, 2025.

John Thickstun, Zaid Harchaoui, and Sham M. Kakade. Learning features of music from scratch. In *International Conference on Learning Representations (ICLR)*, 2017.

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

Zhifei Xie, Mingbao Lin, Zihang Liu, Pengcheng Wu, Shuicheng Yan, and Chunyan Miao. Audio-reasoner: Improving reasoning capability in large audio language models. *arXiv preprint arXiv:2503.02318*, 2025.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025a.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025b.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Chih-Kai Yang, Neo Ho, Yen-Ting Piao, and Hung yi Lee. SAKURA: On the Multi-hop Reasoning of Large Audio-Language Models Based on Speech and Audio Information. In *Interspeech 2025*, pages 1788–1792, 2025b. doi: 10.21437/Interspeech.2025-839.

Appendix A. Data Filtering

In this section, we give an example of how questions from our benchmark are converted to text-only questions and how the Question Validity Filter and Answer Correctness Filter are performed in Table 4. This is an example of a prompt for questions generated from FMA. The prompt can be adapted to other questions generated from other datasets.

Table 4: An example of the data filtering prompt.

You are good at assessing the quality of a question. Please use the following criteria to determine the validity of the question:
a. If identifying the specific genre is essential to answer, output “yes”. If the question can be answered directly from other information provided (without the genre), output “no”.
b. The question is invalid if the question is vague or too general (e.g. refers only to “an award-winning artist”)
After determining validity, check whether the provided answer correctly addresses the question. If correct, output “no”; if not, output “yes”.

Examples
Example 1:
Genre: Hip-hop
Question: What award-winning artist, known for his lyrics and storytelling, released an album in 2022 that belongs to this genre?
Answer: Kendrick Lamar
Validity: no
Correctness: yes
[MORE EXAMPLES]

Your Turn:
Question: question
Answer: answer
Ensure you strictly follow the above output format for “Validity” and “Correctness” (do not need to output “Question” and “Answer”)

Appendix B. Evaluation Prompt

In this section, we present the evaluation prompt in Table 5. It is used to assess the correctness of the models’ answer and, at the same time, to categorize the incorrect responses into specific error types.

Table 5: Evaluation Prompt.

You are given a question, an attribute of the audio, a ground truth answer, and a model answer.
Your task is to categorize the error in the model's answer into one of the following four types:
a) Reasoning Error:
- the model fails to understand the question or fails to apply correct reasoning to arrive at the answer;
- the error is not due to misrecognizing the audio or lack of external knowledge, but rather due to incorrect logic or interpretation;

b) Audio Processing Error:
- the question requires recognizing the attribute of the audio but the model recognizes it wrongly;
- if the model misinterprets the question, this is a **not** an Audio Processing Error, it **should be** Reasoning Error;

c) Knowledge Error:
- the question requires external knowledge beyond what is present in the audio, and the model gives the wrong answer due to lack of or incorrect knowledge;

d) Invalid Answer:
- the model gives no response or a nonsensical/incomplete output.

Question:
{question}

Audio Attribute:
{audio_attr}

Ground Truth:
{gt_answer}

Model Answer:
{model_answer}

Output only the error type (a,b,c or d). No additional explanation is needed.