AudioBench: A Universal Benchmark for Audio Large Language Models

Anonymous ACL submission

Abstract

We introduce AudioBench, a universal benchmark designed to evaluate Audio Large Language Models (AudioLLMs). It encompasses 8 distinct tasks and 26 datasets, among which, 7 are newly proposed datasets. The evaluation targets three main aspects: speech understanding, audio scene understanding, and voice understanding (paralinguistic). Despite recent advancements, there lacks a comprehensive benchmark for AudioLLMs on instruction following capabilities conditioned on audio signals. AudioBench addresses this gap by setting up datasets as well as desired evaluation metrics. Besides, we also evaluated the capabilities of five popular models and found that no single model excels consistently across all tasks. We outline the research outlook for AudioLLMs and anticipate that our open-sourced evaluation toolkit, data, and leaderboard will offer a robust testbed for future model developments.[1]

1 Introduction

Foundation models, built upon large language models (LLMs), have demonstrated strong capability in handling diverse tasks and diverse modalities (Bommasani et al., 2021; Achiam et al., 2023; Team et al., 2023). There have been a series of benchmarks proposed recently to provide a holistic evaluation of LLMs (Hendrycks et al., 2021; Cobbe et al., 2021a; Wang et al., 2024; Rein et al., 2023), image-enhanced multimodal LLMs (Marino et al., 2019; Yue et al., 2023; Padlewski et al., 2024), and even video-enhanced multimodal LLMs (Xiao et al., 2021; Li et al., 2023). However, the audio large language models (AudioLLMs) lag, and the performance of their audio interpretation capability is often intractable and not systematically compared in different tasks. The existing evaluation regime does not cover the breadth of their possible use cases.

Although instruction-following audio-language models have attracted significant interest from both industry and research communities, their reported evaluation datasets are notably different. For instance, the Qwen-Audio-Chat (Chu et al., 2023) model has tested on 12 datasets, whereas the SALMONN (Tang et al., 2024) model used 15 datasets, with only two datasets in common. Although WavLLM (Hu et al., 2024a) has been compared with some earlier models, the scope of evaluation tasks remains limited, which makes the comparison less justified. Another limitation arises from the evaluation set; similar to text-based LLMs (Chiang et al., 2024), evaluations predominantly rely on previous datasets and metrics (Hendrycks et al., 2021). Given that these models are expected to respond flexibly following the instructions, conventional evaluation metrics are not well prepared for this purpose. Consequently, it is crucial to develop innovative benchmarks for AudioLLMs. Since their ultimate goal is to handle diverse audio inputs and respond accurately to user queries, the evaluation set should include not only traditional speech tasks but also new datasets and metrics that better reflect their particular nature and assess the instruction-following capability of models, such as zero-shot speech question answering, etc.

In this study, we introduce AudioBench as a comprehensive evaluation benchmark specifically designed for general instruction-following Audio Large Language Models, which we refer to as AudioLLMs. These models are designed to comprehend general audio inputs and user queries. The tasks should include but are not limited to, understanding speech semantics, interpreting environmental sounds, and recognizing paralinguistic features, as well as combinations of these elements. Therefore, as shown in Figure 1, we incorporate 8 tasks and 26 datasets covering these aspects, with 7 of the datasets being newly adapted or collected to

[1]To be appear upon acceptance, toolkit and public leaderboard.

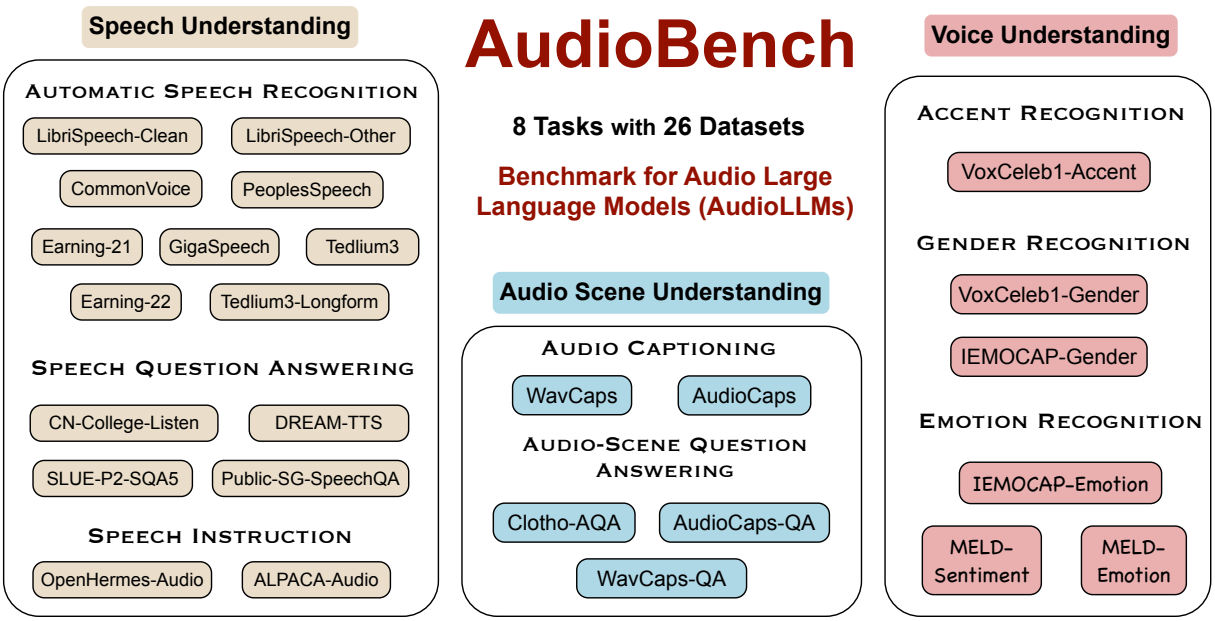

Figure 1: Overview of **AudioBench** datasets.

address gaps in the existing datasets.

For a more comprehensive assessment of the models, our benchmark incorporates a variety of prompt templates to evaluate model compatibility with different instructions. This is because we found that some models are more robust to seen instruction while do not generalize to unseen ones which could lead to robustness issues for deployment. Additionally, we vary the input lengths from seconds to minutes to better assess model performance on longer audio sequences. Our test suite includes 400+ hours of audio, distributed across 100k+ samples. Our **AudioBench** toolkit is available, enabling the evaluation of future AudioLLMs by simply implementing a customized model inference program.

We also explored and decided what evaluation metrics to leverage, particularly for open-ended generation of AudioLLMs. Inspired by text-based and vision-based large language models (LLMs), which can be accessed with multiple-choice and open-ended questions, AudioLLMs are designed to manage diverse instruction and need specific metrics to cover the unconstrained generation space. This introduces difficulties in quantitatively assessing their outputs. Most practices go to model-as-judge methods (Padlewski et al., 2024; Yang et al., 2024a). However, cloud-sourced models, e.g. GPT-4, can be costly and subject to instability introduced by commercial version updates. Therefore, we investigate open-source model-as-judge methods and validate the best practices by correlation studies. Our findings indicate that LLaMA-3-70B-Instruct model demonstrates a higher correlation compared to judgment models like Prometheus-2 (Kim et al., 2024) which is specifically fine-tuned for evaluation purposes.

Finally, we evaluated 5 models to offer a comprehensive review of various solutions, including both end-to-end and cascaded models. Our benchmark results indicate that none of the models performed exceptionally well across all criteria, highlighting significant opportunities for future advancements in model capabilities. To guide these developments, we also include an in-depth discussion on future challenges and directions.

2 Related Work

Multimodal LLM Benchmark. Recent progress in both text-based and multimodal LLMs has spurred the creation of a series of benchmarks designed to monitor and direct their development efficiently. Our research is consistent with earlier benchmarks on multimodal LLMs, primarily concentrating on visual components. There is a lack of comprehensive benchmarks for multimodal LLMs that provide a holistic understanding of audio inputs including speech, ambient sounds, paralinguistic features, etc. When classifying the models targeted for evaluation, we can categorize them into several distinct types: Text-LLMs, Vision-LLMs, and Audio-LLMs.

Text-based LLMs present the most fundamental advancements in the past years and serve as the core for developing multimodal systems. Their evaluation is also multifaceted covering aspects such as reasoning (Zellers et al., 2019; Cobbe et al., 2021b; Shi et al., 2022), subject knowledge (Clark et al., 2018; Hendrycks et al., 2021; Huang et al., 2024b), safety (Zhang et al., 2023b; Li et al., 2024b), multilingual capability (Wang et al., 2024; Ahuja et al., 2023), and more (Chiang et al., 2024; Zheng et al., 2024; Dubois et al., 2024). Text-based LLMs have driven fundamental advancements in recent years (Jiang et al., 2023; Achiam et al., 2023; Mesnard et al., 2024) and have significantly contributed to the evolution of vision LLMs that integrate visual understanding with text queries (Team et al., 2023). Additionally, a series of multimodal benchmarks specifically tailored for vision LLMs have also been established.

The evaluation of Vision-LLMs is multifaceted, encompassing aspects such as multidisciplinary understanding (Yue et al., 2023; Zhang et al., 2024b; Hu et al., 2024b), perception tests (Yu et al., 2023; Liu et al., 2023), mathematical reasoning (Li et al., 2024a; Zhang et al., 2024a), and video understanding (Li et al., 2023; Ning et al., 2023; Liu et al., 2024; Fu et al., 2024). Notably, video understanding benchmarks assess modalities including both visual and audio elements. However, these benchmarks cannot be directly applied to audio understanding, as they predominantly rely on visual inputs with audio serving merely as a supplementary feature. Consequently, there is an urgent need to standardize benchmarks for AudioLLMs that specifically measure model performance with audio context. Such benchmarks are essential for advancing research and improving the capabilities of models that focus on audio processing and understanding.

Advancements in AudioLLMs. Although no comprehensive evaluation benchmarks currently exist, we observe various multimodal foundation models that enhance speech and audio understanding, utilizing either cascaded methods or integrated multitask optimization. Some models specialize in Automatic Speech Recognition (ASR) (Shu et al., 2023; Ma et al., 2024) or other non-speech audio tasks (Kong et al., 2024). However, ideal speech-text foundation models often employ multitasking approaches. Notable examples include AudioGPT (Huang et al., 2023), SpeechGPT (Zhang et al., 2023a), SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), AudioPALM (Rubenstein et al., 2023), LTU-AS (Gong et al., 2023), VioLA (Wang et al., 2023), LLaSM (Shu et al., 2023), Pengi (Deshmukh et al., 2023), WavLLM (Hu et al., 2024a), UniverSLU (Arora et al., 2024), SpeechVerse (Das et al., 2024), and Qwen2-Audio (Chu et al., 2024) each focusing on a subset of tasks tailored to their specific use cases. In this study, we focus on four models—WavLLM, Qwen-Audio-Chat, SALMONN, and Qwen2-Audio-Instruct—that represent a broader range of task-handling capabilities.

3 AudioLLM Benchmark (AudioBench)

3.1 Core Concept

The design of AudioLLMs is to interpret audio content and respond flexibly to user queries, handling tasks from basic speech recognition to complex reasoning. To evaluate these models effectively, we should assess their ability to understand not only speech but also environmental sounds and paralinguistic features such as emotion, gender, and accent. Therefore, our evaluation involves designing benchmarks that cover three key aspects: 1) speech understanding, focusing on the interpretation of the semantic content within the audio; 2) audio scene understanding, focusing on non-human audio sounds; and 3) voice understanding, focusing on non-speech but human-related information and including paralinguistic tasks such as emotion, accent, and gender. For each task, we compile existing datasets and develop new ones tailored to the requirements for comprehensive understanding and reasoning. Figure 1 is an illustration of the tasks and datasets used in our setup.

In AudioBench, the focus is on audio understanding of English semantics, where most models have not yet demonstrated satisfactory results. Meanwhile, multilingual understanding, which includes speech translation and code-switching, represents a useful application that could be integrated with future work.

3.2 Evaluation Setup

From a use case perspective, we expect LLMs to respond to user queries and generate natural responses. Therefore, we expect the model's response to adhere to user queries and the generation style can adapt accordingly. Most conventional speech models and tasks have a constrained output

Category	Dataset Name	#Samples	Hours	Avg.L/Min.L/Max.L(s)	Metrics
Speech Understanding					
ASR	LibriSpeech-Clean	2.6k	5.40	7.43 / 1.28 / 34.96	$\text{WER}_{(\downarrow)}$
	LibriSpeech-Other	2.9k	5.34	6.55 / 1.47 / 34.51	
	CommonVoice-15	16k	26.95	5.93 / 1.34 / 105.67	
	PeoplesSpeech	32k	59.20	6.54 / 1.00 / 99.91	
	GigaSpeech	18k	35.09	6.77 / 1.00 / 22.03	
	Tedlium3	1.1k	2.61	8.24 / 1.07 / 32.55	
	Tedlium3-Longform	11	2.61	0.8k / 0.3k / 1.6k	
	Earning-21	44	39.26	3.2k / 1k / 5.7k	
	Earning-22	125	119.88	3.4k / 0.87k / 7.4k	
SQA	CN-College-Listen	2.2k	13.3	21.09 / 5.76 / 137.82	$\text{M.J.}_{(\uparrow)}$
	SLUE-P2-SQA5	408	4.5	39.85 / 13.00 / 40.0	
	DREAM-TTS	1.9k	18.1	34.14 / 3.14 / 261.85	
	Public-SG-SpeechQA	688	7.6	39.86 / 15.78 / 95.71	
SI	OpenHermes-Audio	100	0.16	5.95 / 2.04 / 15.77	$\text{M.J.}_{(\uparrow)}$
	ALPACA-Audio	100	0.12	4.32 / 1.80 / 8.85	
Audio Scene Understanding					
AQA	Clotho-AQA	2.2k	14.1	22.59 / 15.03 / 29.97	$\text{M.J.}_{(\uparrow)}$
	WavCaps-QA	304	0.87	10.28 / 1.0 / 30.62	
	AudioCaps-QA	313	0.86	9.86 / 3.27 / 10.00	
AC	WavCaps	1.7k	4.9	10.22 / 1.00 / 30.97	$\text{M.J.}_{(\uparrow)}$ & $\text{METEOR}_{(\uparrow)}$
	AudioCaps	4.4k	12.1	9.86 / 1.74 / 10.0	
Voice Understanding					
ER	IEMOCAP-Emotion	1k	1.3	4.51 / 0.75 / 24.12	$\text{M.J.}_{(\uparrow)}$
	MELD-Sentiment	2.6k	2.4	3.35 / 0.13 / 304.9	
	MELD-Emotion	2.6k	2.4	3.35 / 0.13 / 304.9	
AR	VoxCeleb1-Accent	4.8k	11.2	8.27 / 3.96 / 69.04	$\text{M.J.}_{(\uparrow)}$
GR	VoxCeleb1-Gender	4.8k	11.2	8.27 / 3.96 / 69.04	$\text{M.J.}_{(\uparrow)}$
	IEMOCAP-Gender	1k	1.26	4.55 / 0.69 / 26.77	

Table 1: Statistics of AudioBench Dataset Statistics. Yellow refers to our newly expanded and collected datasets to accommodate the missing of suitable datasets. WER refers to Word-Error-Rate. METEOR is a common metric for audio captioning. M.J. refers to model-as-judge, where we deploy Llama-3-70B-Instruct in the current setup.

space. Given classification tasks as an example, output is typically restricted to a finite set of categories or classes. However, the generation space of AudioLLMs is much broader and more complex, which could lead to difficulties in evaluation.

Therefore, we employed the Model-as-Judge (M.J.) approach for most tasks except for automatic speech recognition, where word error rate (WER) is used as the sole metric, and audio captioning, where the METEOR score is utilized as an additional measure. In subsequent experiments, we thoroughly validated the effectiveness of our proposed evaluation approach. We rescaled the M.J. scores to a 100-point scale for easier comparison.

Besides, we found that models can be less robust to diverse instruction, which can negatively impact their applicability and overall user experience. This issue is also present in pioneering text-based LLMs (Zhu et al., 2023; Wang et al., 2024) but it is even more severe in AudioLLMs due to the complexity of processing multiple modalities during the fusion process. Therefore, we propose to leverage multiple instructions to evaluate a dataset, especially when the initial query lacks diversity. This approach is particularly relevant for tasks such as ASR, ER, AR, GR, and AC. The detailed analysis is presented in Section 4.

3.3 Tasks

In this section, we discuss the functionality and applicability of each task within this benchmark. For newly proposed datasets, we provide a detailed description of their development process. The detailed statistics are provided in Table 1.

3.3.1 Speech Understanding

Automatic Speech Recognition (ASR). As a fundamental task in speech processing, ASR aims to convert spoken content into text format. It measures the accuracy of speech-to-text conversion and requires robust algorithms capable of understanding and processing a wide range of linguistic nuances and dialects. ASR systems must be equipped to handle diverse environments, speech rates, accents, and background noises, ensuring that they deliver high-quality transcription under various conditions. To test these capabilities, we included 9 datasets, 3 of which contain audio in a long-form format. Current AudioLLMs struggle with long audio files, which may exceed 10 minutes in duration, presenting a potential area for enhancement in future models. For our current evaluations, we segment the long audio into smaller chunks and then reassemble them for assessment if the model naively does not support the original length. Examples can be found at Table 4.

Speech Question Answering (SQA). The task involves responding to questions based on speech-related audio content. However, there is a lack of suitable datasets. In this part, we involve both monologue understanding and dialogue understanding tasks. Three datasets are newly curated. Examples can be found in Table 5.

First, we collect questions from the English listening comprehension section of China's national college entrance examinations, aimed at assessing students' ability for listening comprehension in both academic and everyday situations. We manually compiled 271 questions and integrated them with 2000 questions sourced from Hu et al. (2024a), which includes questions from approximately 130 different exam sets. In line with the free-text QA format, we presented the questions without multiple-choice options. The dataset is named as *CN-College-Listen*.

The second dataset is DREAM-TTS, which is built upon text-based dialogue comprehension datasets DREAM (Sun et al., 2019). We leverage one SOTA TTS engine (Casanova et al., 2024) to convert the text input into the spoken format. We leverage 60 speakers in the generation and the genders are kept consistent with the dialogue content.

For Public-SG-SpeechQA dataset, we chose four public speaking videos from Singaporean politicians, accompanied by clean transcriptions. These transcriptions were manually segmented, and five questions per segment were generated using LLMs. Each question and its corresponding reference answer were manually reviewed to ensure their validity, where around 30% of the samples were discarded. Ultimately, this process resulted in the collection of 688 speech question-answer pairs for evaluation.

Speech Instruction (SI). This task evaluates whether the model can directly follow instructions provided through speech input, mirroring a natural human-computer interaction. Specifically, the question is delivered via audio, and the model is expected to understand and generate responses in the appropriate text format. This approach offers unique benefits. For instance, by incorporating paralinguistic information such as emotions, the model can adapt its responses based on the user's emotional state. However, finding suitable testing examples has been a challenge. In this study, we synthesized audio instruction (Casanova et al., 2024) from existing instruction-following datasets: ALPACA (Taori et al., 2023) and OpenHermes (Teknium, 2023). We then involve humans in selecting instances to ensure that 1) the speech is accurately synthesized and 2) the content is suitable as spoken instructions. As a result, around 10.5% samples are kept and the examples can be found in Table 8.

3.3.2 Audio Scene Understanding

Audio Question Answering (AQA). This task focus more on the understanding of environmental contexts. To assess the capability to follow instructions in these scenarios, we refined a test set from Clotho-AQA, retaining only the samples with high confidence levels. Simultaneously, we developed the WavCaps-QA and AudioCaps-QA datasets, each featuring over 300 diverse questions accompanied by human-verified questions and answers. Examples can be found in Table 6.

Audio Captioning (AC). This task involves generating captions (descriptions) for an audio clip. We incorporate two popular datasets to assess this task: WavCaps and AudioCaps. The examples can be found in Table 9. (Banerjee and Lavie, 2005)

3.3.3 Voice Understanding

Emotion Recognition (ER). Emotion is a critical paralinguistic element that can be conveyed through human speech or non-speech content (El Ayadi et al., 2011). It plays a vital role in making communications interpretable for listeners.

Dataset Name	AudioLLMs				Whisper+Llama3
	SALMONN	Qwen-Audio-Chat	WavLLM	Qwen2-Audio-Instruct	
Speech Understanding					
LibriSpeech-Clean$_{(\downarrow)}$	55.58	2.25	**2.10**	3.20	1.83
LibriSpeech-Other$_{(\downarrow)}$	41.80	**4.16**	4.80	6.07	3.71
CommonVoice-15$_{(\downarrow)}$	33.75	11.65	14.53	**11.44**	9.89
PeoplesSpeech$_{(\downarrow)}$	34.33	30.72	37.92	**22.32**	14.54
GigaSpeech$_{(\downarrow)}$	14.22	13.32	15.49	**11.89**	9.51
Tedlium3$_{(\downarrow)}$	8.56	**4.00**	6.62	6.39	3.81
Tedlium3-Longform$_{(\downarrow)}$	**18.39**	45.29	45.37	95.35	4.75
Earning-21$_{(\downarrow)}$	**26.87**	38.46	64.47	98.65	11.77
Earning-22$_{(\downarrow)}$	**36.38**	51.18	66.72	98.84	15.61
CN-College-Listen	50.51	60.85	65.43	**74.50**	85.25
SLUE-P2-SQA5	78.24	76.12	**83.92**	80.05	82.99
DREAM-TTS	55.93	57.76	64.56	**66.70**	86.09
Public-SG-SpeechQA	56.77	57.47	**58.55**	58.31	64.94
OpenHermes-Audio	19.20	11.00	22.40	**44.80**	63.0
ALPACA-Audio	12.40	9.60	21.60	**52.60**	70.8
Audio Scene Understanding					
Clotho-AQA	51.18	**58.20**	43.01	50.92	29.47
WavCaps-QA	**46.25**	38.68	26.25	44.47	17.38
AudioCaps-QA	47.03	**47.99**	29.84	45.75	16.71
WavCaps$_{(\text{M.J.})}$	21.16	29.25	6.40	**33.78**	3.45
AudioCaps$_{(\text{M.J.})}$	34.37	**47.99**	4.17	40.78	2.47
WavCaps$_{(\text{METEOR})}$	17.72	**24.02**	9.78	21.34	13.89
AudioCaps$_{(\text{METEOR})}$	21.20	**27.70**	6.70	19.89	7.95
Voice Understanding					
IEMOCAP-Emotion	21.56	27.34	45.91	**49.30**	34.43
MELD-Emotion	33.06	**50.57**	41.07	40.54	33.36
MELD-Sentiment	41.87	43.87	50.08	**53.49**	43.87
VoxCeleb1-Accent	28.06	**45.70**	37.65	29.19	39.33
VoxCeleb1-Gender	88.90	70.56	70.51	**99.12**	53.41
IEMOCAP-Gender	**51.60**	51.13	45.29	49.30	51.50

Table 2: Main results of four AudioLLMs and one cascade model. The word-error-rate (WER) for ASR tasks is the lower the better$_{(\downarrow)}$.

Three datasets are incorporated and the examples can be found in Table 7.

Accent Recognition (AR). Accent, a significant paralinguistic feature, has been overlooked in previous evaluations. We have included metadata from the VoxCeleb1 dataset and tasked the model with predicting the speaker's likely origin based on their accent. Examples can be found in Table 10.

Gender Recognition (GR). Gender could be recognized based on vocal characteristics. This task tests the model's sensitivity to gender-specific acoustic patterns. We have prepared two datasets and the examples can be found in Table 11.

4 Results and Analysis

4.1 Main Results

For comparative analysis, we selected four representative AudioLLMs: SALMONN (Tang et al., 2024), Qwen-Audio (Chu et al., 2023), WavLLM (Hu et al., 2024a) and Qwen2-Audio (Chu et al., 2024). These models are among the most capable multi-task speech-text models available, designed to manage various tasks simultaneously. Additionally, we include a cascade model named Whisper+Llama3, which processes data in a pipeline manner. First, transcriptions are extracted using the Whisper-larg-v3 (Huang and Tsai, 2023) model, and then these transcriptions along with user queries are input into the Llama-3-8B-Instuct model to generate responses. While this model cannot comprehend rich audio content, relying solely on transcriptions for context, it demonstrates strong performance in speech-intensive tasks and serves as a robust baseline.

Table 2 detailed the overall results across 26 datasets for 5 models, revealing that no single model consistently outperforms the others on all tasks. Specifically, SALMONN model is sensitive

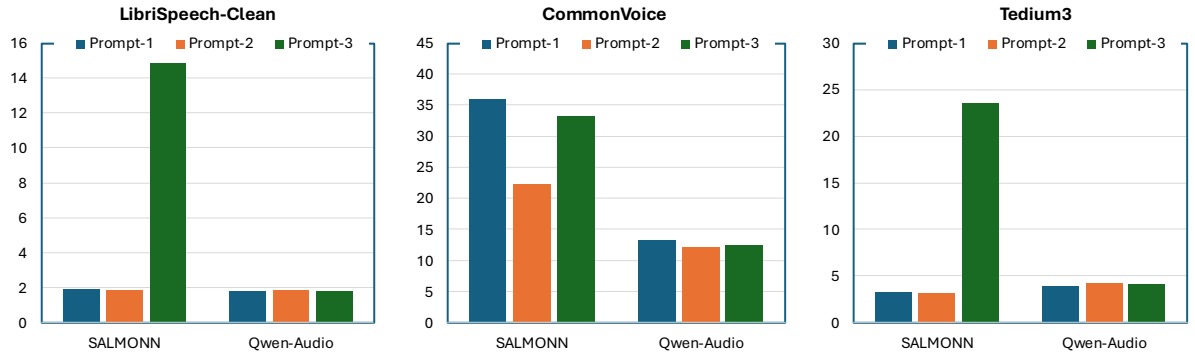

Figure 2: Sensitivity tests were conducted on three different prompts on the same datasets. Performance is measured by word error rate (WER)$_{(\downarrow)}$. The results state that SALMONN is the least robust when faced with varying prompt templates for the same task. Prompt 1-3: {"Turn the speech input into a text transcription", "Recognize the speech and give me the transcription", "Decode the audio and give me the transcription."}

to varying instructions for ASR tasks, which indicates that it is overfitting to certain audio features and ignores proper instructions. Qwen1/2-Audio and WavLLM demonstrate robust ASR capabilities. However, all AudioLLMs struggle with long-form ASR tasks. This could be because the models are mainly fine-tuned with limited audio length which makes them hard to generalize to arbitrary context lengths. Additionally, the training and pre-training datasets for AudioLLMs are generally smaller than those used for other task-specific models like Whisper, which limits their generalizability to unnatural truncation as well.

For speech-intensive tasks such as SQA and SI, the cascade model Whisper+Llama3 exhibits superior performance. This effectiveness stems from the Whisper model's robust recognition capabilities and Llama's strong reasoning abilities, which together efficiently handle queries primarily contained in the verbal content. In contrast, the modality fusion process in AudioLLMs may distort speech content, highlighting an area for future improvement.

For tasks involving paralinguistic features and non-verbal sounds, AudioLLMs generally outperform cascade models, although they do not always yield satisfactory results. An exception is found in sentiment and emotion recognition tasks, where some emotions can be directly inferred from the speech semantics. Overall, cascade models struggle with understanding non-verbal content, highlighting the need for the more robust development of AudioLLMs to handle these complexities better.

In comparing three AudioLLMs, we observed that models perform better when exposed to related tasks during training. A significant example is WavLLM, which specializes in speech content integration. It excels in SQA tasks but lacks exposure to non-spoken scenarios such as audio captioning. Consequently, it underperforms in audio scene understanding and vocal understanding tasks, demonstrating that its instruction-following capabilities are not readily generalizable without specifically aligned training samples.

4.2 Robustness Queries

In evaluation, we found that AudioLLMs could behave differently to different prompt templates. Even though this also applies to text-based LLMs, it is even more severe in AudioLLMs as the two modalities are not aligned in the first stage and aggressive alignment would hurt the generalizability of different prompts. Hu et al. (2024a) also spot such issues but only limited experimental studies are presented. Therefore, we did a more thorough experimental evaluation in this case with 2 models (SALMONN and Qwen-Audio) on 3 ASR datasets (LirbriSpeech-Clean, CommonVoice, Tedium3).

Figure 2 illustrates the results. We can witness that SALMONN can effectively perform ASR tasks and achieve performance comparable to the Qwen-Audio model on the LibriSpeech-Clean and Tedium3 tasks, provided the right prompts are used. However, prompt 3, "Decode the audio and give me the written transcripts," triggers the SALMONN model to conduct phoneme recognition on the LibriSpeech and Tedium3 datasets. For the CommonVoice dataset, SALMONN tends to perform speech translation on a significant number of samples, which adversely affects its performance on the Word Error Rate (WER) metric. This suggests that SALMONN may be overly tuned to speech

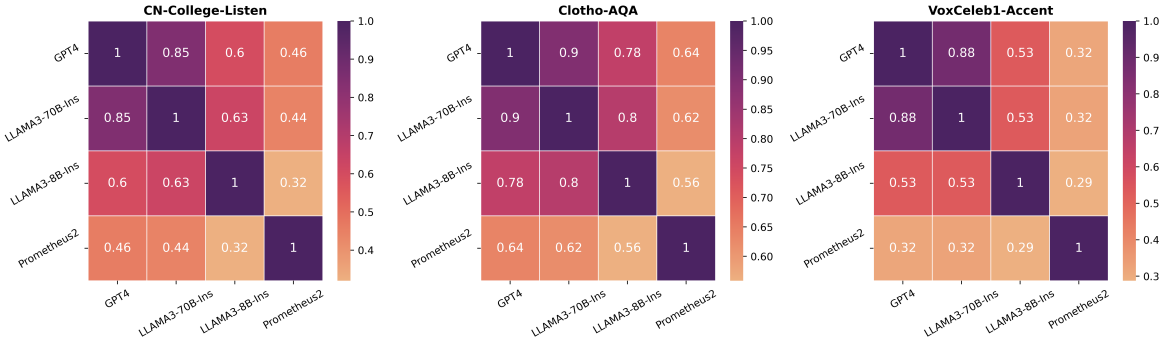

Figure 3: Correlation results for three benchmark datasets across three variations of the model-as-judge approach were analyzed using Spearman's rank correlation. The findings show that Llama-3-70B-Instruct has a very strong correlation with GPT-4, when serving as the judge.

features (tokens) and not sufficiently responsive to the prompts. In contrast, Qwen-Audio's performance remains stable across all three prompt templates. Consequently, to minimize variability and thoroughly test model robustness, we incorporate at least 20 diverse prompt templates into our evaluation framework for the tasks without diverse prompts originally.

4.3 Model-as-Judge Comparison

In many datasets, we use an open-ended generation style, which is considered the most natural method of interaction. However, this style introduces challenges for evaluation, as traditional metrics are not easily adaptable. To address this, we employ a model-as-judge approach to assess model responses against reference answers and questions. Yet, there are ongoing challenges in selecting appropriate models to ensure both accessibility and robustness.

The most frequently-used judgement model is GPT-4. However, GPT-4 is closed-source, subject to ongoing updates, and often incurs costs that may not be feasible for the wider community. To explore alternatives, we examined 3 open-source models: Llama-3-8B-Instruct, Llama-3-70B-Instruct, and Prometheus-2 (Kim et al., 2024). Prometheus-2 is specifically fine-tuned for response grading, while Llama-3-Instruct is a general-purpose open-source model.

To identify the most suitable open-source model for evaluation tasks, we use all samples from the following three datasets: CN-College-Listen, Clotho-AQA, and VoxCeleb1-Accent, to evaluate their replaceability of GPT-4. The three datasets cover speech understanding, audio scene understanding, and voice understanding. Initially, we get the model outputs from the SALMONN model. These model outputs, along with golden answers and questions, are then fed into the four judgment models to obtain judgment scores. We subsequently calculated the Spearman's rank correlation to compare the effectiveness of the different judgment models. The evaluation template can be found in Table 3. The results are detailed in Figure 3.

As a result, the Llama-3-70B-Instruct model exhibits the highest correlation with GPT-4-as-a-judge (gp4-turbo-0409). The correlation scores exceeding 0.85 across all three datasets indicate a "very strong correlation" with GPT-4. We attribute this to Llama-3-70B-Instruct's robust generalizability. In contrast, Prometheus-2, though fine-tuned from the Mistral model and further adapted for scoring purposes, does not effectively compensate for the limitations of its base model. Therefore, in experiments, we will use Llama-3-70B-Instruct as the judging model (scale from 0-5 or 0-1 followed by rescaling to 100-point scale for easier interpretation), as it provides transparency and offers ease of adaptability. It is important to note that using a model-as-judge is not an ultimate solution. The issue of accurately grading free-form generations remains an unresolved challenge in NLP.

5 Conclusion

In this work, we introduce the first comprehensive evaluation benchmark for AudioLLMs, named AudioBench, which includes 8 tasks and 26 datasets. Our report includes a detailed analysis of data selection, metric design, and the evaluation pipeline. We emphasize the critical factors for benchmark design and model development. We anticipate that our proposal will stimulate further progress in multimodal foundation models.

Limitations

First, the current AudioBench exclusively includes English datasets. However, multilingual capabilities and code-switching are crucial for comprehensive speech understanding and generation. We plan to expand the benchmark to incorporate these aspects in future iterations.

Second, similar to text-based LLMs, evaluating free-style generation is challenging and demands robust metrics or models to serve as judges. Traditional metrics fall short in zero-shot scenarios, and assessing the correctness of output responses is complex. Thus, the development of suitable evaluation metrics, particularly for audio inputs, is crucial.

Third, as end-to-end models, AudioLLMs typically involve large model sizes, which result in longer inference times. In this benchmark, our focus has been primarily on accuracy rather than efficiency. Moving forward, it will be important to consider inference speed and the deployment environment to provide a comprehensive evaluation of these models during the deployment process.

References

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

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Maxamed Axmed, et al. 2023. Mega: Multilingual evaluation of generative ai. *arXiv preprint arXiv:2303.12528*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Kohler, Josh Meyer, Michael Henretty, Reuben Morais, Lindsay Saunders, Francis Tyers, and Gregor Weber. 2020. Common voice: A massively-multilingual speech corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4218–4222, Marseille, France. European Language Resources Association.

Siddhant Arora, Hayato Futami, Jee weon Jung, Yifan Peng, and Yosuke Kashiwagi Roshan Sharma, Emiru Tsunoo, Karen Livescu, and Shinji Watanabe. 2024. UniverSLU: Universal spoken language understanding for diverse tasks with natural language instructions. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72.

Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

Edresson Casanova, Kelly Davis, Eren Gölge, Görkem Göknar, Iulian Gulea, Logan Hart, Aya Aljafari, Joshua Meyer, Reuben Morais, Samuel Olayemi, and Julian Weber. 2024. Xtts: a massively multilingual zero-shot text-to-speech model.

Guoguo Chen, Shuzhou Chai, Guanbo Wang, Jiayu Du, Wei-Qiang Zhang, Chao Weng, Dan Su, Daniel Povey, Jan Trmal, Junbo Zhang, et al. 2021. Gigaspeech: An evolving, multi-domain asr corpus with 10,000 hours of transcribed audio. *arXiv preprint arXiv:2106.06909*.

Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E Gonzalez, et al. 2024. Chatbot arena: An open platform for evaluating llms by human preference. *arXiv preprint arXiv:2403.04132*.

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

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

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021a. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro

Nakano, et al. 2021b. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nilaksh Das, Saket Dingliwal, S. Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, Xilai Li, Karel Mundnich, Monica Sunkara, Sundararajan Srinivasan, Kyu J Han, and Katrin Kirchhoff. 2024. SpeechVerse: A large-scale generalizable audio language model. *arXiv*.

Miguel Del Rio, Natalie Delworth, Ryan Westerman, Michelle Huang, Nishchal Bhandari, Joseph Palakapilly, Quinten McNamara, Joshua Dong, Piotr Zelasko, and Miguel Jetté. 2021. Earnings-21: A practical benchmark for asr in the wild. *arXiv preprint arXiv:2104.11348*.

Miguel Del Rio, Peter Ha, Quinten McNamara, Corey Miller, and Shipra Chandra. 2022. Earnings-22: A practical benchmark for accents in the wild. *arXiv preprint arXiv:2203.15591*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yann Dubois, Balázs Galambosi, Percy Liang, and Tatsunori B Hashimoto. 2024. Length-controlled alpacaeval: A simple way to debias automatic evaluators. *arXiv preprint arXiv:2404.04475*.

Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. 2011. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern recognition*, 44(3):572–587.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*.

Daniel Galvez, Greg Diamos, Juan Ciro, Juan Felipe Cerón, Keith Achorn, Anjali Gopi, David Kanter, Maximilian Lam, Mark Mazumder, and Vijay Janapa Reddi. 2021. The people's speech: A large-scale diverse english speech recognition dataset for commercial usage. *arXiv preprint arXiv:2111.09344*.

Yuan Gong, Alexander H. Liu, Hongyin Luo, Leonid Karlinsky, and James Glass. 2023. Joint audio and speech understanding. In *IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Proceedings of the International Conference on Learning Representations (ICLR)*.

Shujie Hu, Long Zhou, Shujie Liu, Sanyuan Chen, Hongkun Hao, Jing Pan, Xunying Liu, Jinyu Li, Sunit Sivasankaran, Linquan Liu, et al. 2024a. Wavllm: Towards robust and adaptive speech large language model. *arXiv preprint arXiv:2404.00656*.

Yutao Hu, Tianbin Li, Quanfeng Lu, Wenqi Shao, Junjun He, Yu Qiao, and Ping Luo. 2024b. Omnimedvqa: A new large-scale comprehensive evaluation benchmark for medical lvlm. *arXiv preprint arXiv:2402.09181*.

Chien-yu Huang, Ke-Han Lu, Shih-Heng Wang, Chi-Yuan Hsiao, Chun-Yi Kuan, Haibin Wu, Siddhant Arora, Kai-Wei Chang, Jiatong Shi, Yifan Peng, et al. 2024a. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 12136–12140. IEEE.

Rongjie Huang, Mingze Li, Dongchao Yang, Jiatong Shi, Xuankai Chang, Zhenhui Ye, Yuning Wu, Zhiqing Hong, Jia-Bin Huang, Jinglin Liu, Yixiang Ren, Zhou Zhao, and Shinji Watanabe. 2023. AudioGPT: Understanding and generating speech, music, sound, and talking head. In *AAAI Conference on Artificial Intelligence*.

Yi-Chin Huang and Ji-Qian Tsai. 2023. Whisper model adaptation for FSR-2023 Hakka speech recognition challenge. In *Proceedings of the 35th Conference on Computational Linguistics and Speech Processing (ROCLING 2023)*, pages 423–427, Taipei City, Taiwan. The Association for Computational Linguistics and Chinese Language Processing (ACLCLP).

Yuzhen Huang, Yuzhuo Bai, Zhihao Zhu, Junlei Zhang, Jinghan Zhang, Tangjun Su, Junteng Liu, Chuancheng Lv, Yikai Zhang, Yao Fu, et al. 2024b. C-eval: A multi-level multi-discipline chinese evaluation suite for foundation models. *Advances in Neural Information Processing Systems*, 36.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *NAACL-HLT*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models.

Zhifeng Kong, Arushi Goel, Rohan Badlani, Wei Ping, Rafael Valle, and Bryan Catanzaro. 2024. Audio Flamingo: A novel audio language model with few-shot learning and dialogue abilities. *arXiv*.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. 2023. Mvbench: A comprehensive multi-modal video understanding benchmark. *arXiv preprint arXiv:2311.17005*.

Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. 2024a. Multimodal arxiv: A dataset for improving scientific comprehension of large vision-language models. *arXiv preprint arXiv:2403.00231*.

Lijun Li, Bowen Dong, Ruohui Wang, Xuhao Hu, Wangmeng Zuo, Dahua Lin, Yu Qiao, and Jing Shao. 2024b. Salad-bench: A hierarchical and comprehensive safety benchmark for large language models. *arXiv preprint arXiv:2402.05044*.

Samuel Lipping, Parthasaarathy Sudarsanam, Konstantinos Drossos, and Tuomas Virtanen. 2022. Clotho-aqa: A crowdsourced dataset for audio question answering. In *2022 30th European Signal Processing Conference (EUSIPCO)*, pages 1140–1144. IEEE.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Rao Ma, Adian Liusie, Mark John Francis Gales, and Kate Knill. 2024. Investigating the emergent audio classification ability of asr foundation models. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Xinhao Mei, Chutong Meng, Haohe Liu, Qiuqiang Kong, Tom Ko, Chengqi Zhao, Mark D Plumbley, Yuexian Zou, and Wenwu Wang. 2023. Wavcaps: A chatgpt-assisted weakly-labelled audio captioning dataset for audio-language multimodal research. *arXiv preprint arXiv:2303.17395*.

Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, Pouya Tafti, Léonard Hussenot, and et al. 2024. Gemma. *Kaggle*.

Arsha Nagrani, Joon Son Chung, and Andrew Zisserman. 2017. Voxceleb: a large-scale speaker identification dataset. *arXiv preprint arXiv:1706.08612*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Piotr Padlewski, Max Bain, Matthew Henderson, Zhongkai Zhu, Nishant Relan, Hai Pham, Donovan Ong, Kaloyan Aleksiev, Aitor Ormazabal, Samuel Phua, et al. 2024. Vibe-eval: A hard evaluation suite for measuring progress of multimodal language models. *arXiv preprint arXiv:2405.02287*.

Vassil Panayotov, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. Librispeech: an asr corpus based on public domain audio books. In *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*, pages 5206–5210. IEEE.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. 2023. Gpqa: A graduate-level google-proof q&a benchmark. *arXiv preprint arXiv:2311.12022*.

Anthony Rousseau, Paul Deléglise, and Yannick Esteve. 2012. Ted-lium: an automatic speech recognition dedicated corpus. In *LREC*, pages 125–129.

Paul K Rubenstein, Chulayuth Asawaroengchai, Duc Dung Nguyen, Ankur Bapna, Zalán Borsos, Félix de Chaumont Quitry, Peter Chen, Dalia El Badawy, Wei Han, Eugene Kharitonov, et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, et al. 2022. Language models are multilingual chain-of-thought reasoners. *arXiv preprint arXiv:2210.03057*.

Suwon Shon, Siddhant Arora, Chyi-Jiunn Lin, Ankita Pasad, Felix Wu, Roshan Sharma, Wei-Lun Wu, Hung-Yi Lee, Karen Livescu, and Shinji Watanabe. 2022. Slue phase-2: A benchmark suite of diverse spoken language understanding tasks. *arXiv preprint arXiv:2212.10525*.

Yu Shu, Siwei Dong, Guangyao Chen, Wenhao Huang, Ruihua Zhang, Daochen Shi, Qiqi Xiang, and Yemin Shi. 2023. LLaSM: Large language and speech model. *arXiv*.

Kai Sun, Dian Yu, Jianshu Chen, Dong Yu, Yejin Choi, and Claire Cardie. 2019. Dream: A challenge data set and models for dialogue-based reading comprehension. *Transactions of the Association for Computational Linguistics*, 7:217–231.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun MA, and Chao Zhang. 2024. SALMONN: Towards generic hearing abilities for large language models. In *International Conference on Learning Representations (ICLR)*.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca.

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

Teknium. 2023. Openhermes.

Bin Wang, Zhengyuan Liu, Xin Huang, Fangkai Jiao, Yang Ding, Ai Ti Aw, and Nancy F Chen. 2024. Seaeval for multilingual foundation models: From cross-lingual alignment to cultural reasoning. *NAACL*.

Tianrui Wang, Long Zhou, Ziqiang Zhang, Yu Wu, Shujie Liu, Yashesh Gaur, Zhuo Chen, Jinyu Li, and Furu Wei. 2023. VioLA: Unified codec language models for speech recognition, synthesis, and translation. *arXiv*.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024a. 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)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics.

Shu-wen Yang, Heng-Jui Chang, Zili Huang, Andy T Liu, Cheng-I Lai, Haibin Wu, Jiatong Shi, Xuankai Chang, Hsiang-Sheng Tsai, Wen-Chin Huang, et al. 2024b. A large-scale evaluation of speech foundation models. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. 2023. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. *arXiv preprint arXiv:2311.16502*.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. SpeechGPT: Empowering large language models with intrinsic cross-modal conversational abilities. In *Findings of the Association for Computational Linguistics: EMNLP*.

Renrui Zhang, Dongzhi Jiang, Yichi Zhang, Haokun Lin, Ziyu Guo, Pengshuo Qiu, Aojun Zhou, Pan Lu, Kai-Wei Chang, Peng Gao, et al. 2024a. Mathverse: Does your multi-modal llm truly see the diagrams in visual math problems? *arXiv preprint arXiv:2403.14624*.

Wenxuan Zhang, Mahani Aljunied, Chang Gao, Yew Ken Chia, and Lidong Bing. 2024b. M3exam: A multilingual, multimodal, multilevel benchmark for examining large language models. *Advances in Neural Information Processing Systems*, 36.

Zhexin Zhang, Leqi Lei, Lindong Wu, Rui Sun, Yongkang Huang, Chong Long, Xiao Liu, Xuanyu Lei, Jie Tang, and Minlie Huang. 2023b. Safety-bench: Evaluating the safety of large language models with multiple choice questions. *arXiv preprint arXiv:2309.07045*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

A Appendix

A.1 Discussion on Dataset

In Figure 4, we present the structure of the tasks and datasets included in our benchmark. In this section, we aim to provide a detailed description of the newly curated datasets along with their construction processes.

CN-College-Listen. This is a zero-shot testing set compiled from English comprehension questions extracted from college exam papers. Initially, we downloaded audio files and corresponding exam papers from public websites. These audio files were segmented according to the questions in the exam papers. Typically, an audio segment may relate to multiple questions depending on the sections of the exam paper. Each segmented audio is paired with its corresponding questions to formulate the test items. Originally, the questions in the exam papers are formatted as multiple-choice; however, for the application with AudioLLMs, it is more natural to frame them as open-ended questions. We retain the correct choice as the reference answer. In total, we collected 271 audio-question-answer triples and supplemented these with another 2000 questions from (Hu et al., 2024a) to create the final dataset. This dataset requires a precise capture of audio content, including both monologues and dialogues.

DREAM-TTS. The evaluation set for spoken dialogue understanding remains limited. DREAM is a dataset featuring human-annotated questions and answers related to a dialogue. We used text-to-speech (TTS) technology to convert the original text dialogues into spoken format, adhering to gender information to closely mimic real-life scenarios. The original questions and answers were retained to create dialogue-question-answer triples.

Public-SG-SpeechQA. To evaluate real speech content, we selected four videos from public speeches in Singapore. These speeches vary in length from 20 minutes to an hour, and clean transcripts are available online. We manually segmented the speeches based on topic transitions. Using the Llama-2-7B-Chat model, we generated question-and-answer pairs for each segment. These pairs underwent a subsequent round of human review to ensure they were concrete and relevant. Ultimately, we compiled 688 speech-question-answer triples, with each speech segment corresponding to multiple questions and answers.

OpenHermes-Audio and **ALPACA-Audio**. Since effective speech instruction datasets are lacking, we utilized an open-source instruction tuning set designed for text-based LLMs and converted these instructions into spoken format using text-to-speech (TTS) technology. Human reviewers then verified these converted samples, filtering out approximately 90% and retaining only 10% for our test set. This filtering process ensures the readability, naturalness of the spoken instructions, and correctness of the reference answers. Figure 5 and 6 show the annotation platform we developed using Streamlit for data annotation for these two datasets.

WavCaps QA and **AudioCaps QA**. Figure 7 illustrates the data annotation platform we developed for the WavCaps and AudioCaps datasets. Initially, we used the Llama-3-8B-Instruction model to generate questions and answers from the provided captions. Each generated sample then went through another round of human annotation to refine the questions and ensure the validity of the test set. Answers were also revised as needed. Through this process, we compiled over 300 questions for each dataset to assess the models' performance effectively.

A.2 Model-as-Judge Template

Table 3 the scoring templates used for the model-as-judge approach. The scoring range is from 0 to 5, except for Prometheus2, where the range is from 1 to 5. For some tasks that require binary judgment, the score will be either 0-incorrect or 1-correct. In practice, we rescaled all scores to a 100-point scale for easier comparison. For some tasks,

A.3 Comparison with AIR-Bench

AIR-Bench is one concurrent work with AudioBench (Yang et al., 2024a). Even though both focus on evaluating audio-based instructions with LLMs. The selected datasets and evaluation preferences are different. First, the biggest difference comes from the coverage of datasets. In AudioBench, we proposed 6 new datasets to accommodate what is missing in the evaluation. Meantime, we have multiple ASR datasets, speech question answer datasets, and speech instruction datasets that are not covered by AIR-Bench. Second, the AudioBench evaluation toolkit takes care of prompt variants and considers the model's robustness in following instructions. Third, we present a holistic study on the choice of evaluation metrics and provide stable and affordable solutions for future benchmarking purposes. Some merits can also be seen in AIR-Bench which also includes Music datasets. Nevertheless, AudioBench servers a good foundation for evaluating AudioLLMs in three aspects: speech understanding, audio scene understanding, and voice understanding, and are expandable toolkit for more suitable scenarios.

A.4 Why SUPERB and Dynamic-SUPERB is not Ideal

For instance, a pair ($Audio$, $Query$) is presented to the model to elicit a ($Model_Answer$), which is then compared with a reference answer for relevant tasks. This approach does not align well with previous speech benchmarks like SUPERB (Yang et al., 2024b) and Dynamic-SUPERB (Huang et al., 2024a). The design for SUPERB is for the evaluation of a self-supervised speech encoder where a supervised fine-tuning step is normally conducted for the evaluation. Dynamic-SUPERB can measure the instruction following tasks in a zero-shot fashion. However, it is an open collection with crowd-sourced contributions that lacks a specific focus for AudioLLMs as also discussed in Yang et al. (2024a). Therefore, AudioBench is introduced to focus primarily on evaluating AudioLLMs, offering new datasets and evaluation pipelines to address the existing gaps.

B Research Outlook

Long Audio Processing and Understanding: The current benchmarks primarily assess the understanding capabilities of AudioLLMs using audio clips of limited duration (within minutes). However, extending the capability to process longer audio can open up broader applications such as meeting summarization and sequential event understanding. In text-based LLMs, long-sequence processing has advanced rapidly. By embedding speech content as tokens, this capability could be effectively explored.

Multi-round Query Handling: The ability to manage multi-round queries is still limited in open-source models. Enhancing this feature would allow for more dynamic interactions where each query could involve different modalities such as images or audio, making the models more versatile in practical applications.

Multilingual Capabilities, Code-Switching, and Dialects: Expanding the linguistic capabilities of AudioLLMs to handle multiple languages, code-switching, and various dialects is crucial. This enhancement would improve the models' applicability across diverse linguistic and cultural contexts, making them more effective for global use.

Speech Generation: Developing more sophisticated speech generation capabilities within AudioLLMs would enable more natural and engaging human-computer interactions. This includes improving the models' ability to generate speech that not only conveys information but also mimics human-like intonations and rhythms.

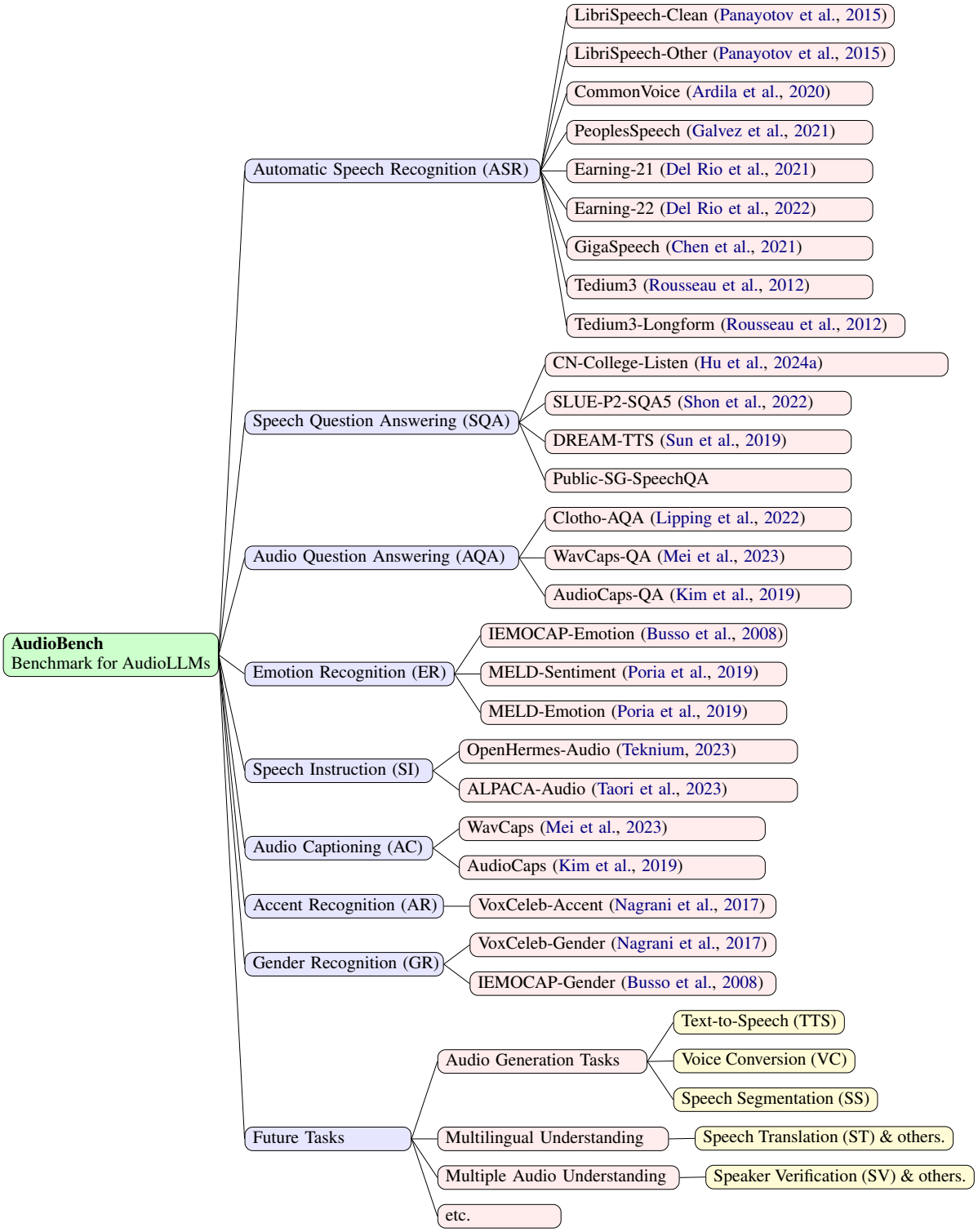

Figure 4: Structure of AudioBench datasets.

Judgement Model	Template
Llama-3-70B-Instruct & **Llama-3-8B-Instruct** & **GPT4**	*[Reference Answer]* *{reference}* *[Model Answer]* *{prediction}* *[Question]* *{question}* *[Task]* *Rate the model's answer based on its alignment with the reference answer, focusing on accuracy and relevance to the reference provided. Please be critical on the details.* *Criteria: Assess if the model's response mirrors the reference in terms of content, accuracy, and relevance.* *Score0: The answer is completely misaligned, providing incorrect or irrelevant information compared to the reference.* *Score1: The answer shows minimal alignment, often misunderstanding or providing irrelevant details unrelated to the reference.* *Score2: The answer recognizes the topic but diverges significantly from the reference in accuracy or relevance.* *Score3: The answer aligns with the reference generally but lacks detail or precise accuracy in some aspects.* *Score4: The answer is mostly accurate and relevant, closely following the reference but could be clearer or more detailed.* *Score5: The answer is highly accurate, detailed, and matches the reference answer perfectly, capturing its essence and detail.* *Your response should be formatted as follows:* *Explanation: (Provide a concise explanation of your rating, comparing the reference answer with the model's response. "The reference answer is [XXX], while the model's answer is [YYY]. I think ...")* *Rating: (int)"""*
Prometheus2	*"criteria": "Does the model provide accurate, relevant, and contextually appropriate responses to user inquiries?"* *"score1_description": "The model frequently fails to understand or address the core of the user's inquiries, providing inaccurate, irrelevant, or inappropriate responses."* *"score2_description": "The model occasionally recognizes the topic of inquiry but often provides responses that are not sufficiently accurate, detailed, or contextually relevant."* *"score3_description": "The model usually understands the question and attempts to provide a relevant answer, yet the responses may sometimes lack detail, accuracy, or context."* *"score4_description": "The model consistently understands and appropriately addresses the questions, providing accurate and relevant responses. However, there may still be minor inaccuracies or instances where additional context could enhance clarity."* *"score5_description": "The model excels in understanding user inquiries and consistently delivers accurate, detailed, and contextually appropriate responses that thoroughly address the user's needs."*

Table 3: Template for Model-as-Judge.

Dataset	Context	Instruction (Example)	Answer
LibriSpeech-Clean	No, I wasn't thinking of that.	Turn the speech input into a text transcription.	No, I wasn't thinking of that.
LibriSpeech-Other	The history of the house is plain now.	Decode the audio and give me the written transcription.	The history of the house is plain now.
CommonVoice	This is Jon Davis.	Process the audio speech and provide the text output.	This is Jon Davis.
PeoplesSpeech	that's where you have a lot of windows in ...	Convert the audio speech into a text transcript.	that's where you have a lot of windows in ...
GigaSpeech	I believe that fast fashion has made it possible for everyone to have access to aesthetically thoughtful clothing.	Transform the speech into a text document.	I believe that fast fashion has made it possible for everyone to have access to aesthetically thoughtful clothing.
Earning-21	Good morning, everyone, and welcome to the NextEra Energy Inc. and NextEra Energy Partners...	Convert the audio speech into a text transcript.	Good morning, everyone, and welcome to the NextEra Energy Inc. and NextEra Energy Partners...
Earning-22	Good day, ladies and gentlemen, and welcome to the CD Projekt Group financial results for H1 2021 conference call. Today's call...	Transform the speech into a text document.	Good day, ladies and gentlemen, and welcome to the CD Projekt Group financial results for H1 2021 conference call. Today's call...
Tedium3	One day, Los Angeles Times columnist Steve Lopez was walking along the streets...	Turn the speech input into a text transcription.	One day, Los Angeles Times columnist Steve Lopez was walking along the streets...
Tedium3-Longform	I'd like to share with you a discovery I made a few months ago while writing an article for Italian Wired. I always keep my thesaurus handy whenever...	Process the audio speech and provide the text output.	I'd like to share with you a discovery I made a few months ago while writing an article for Italian Wired. I always keep my thesaurus handy whenever...

Table 4: Examples from Automatic Speech Recognition (ASR) datasets. Capitalization normalized for display purposes.

Dataset	Context	Instruction	Answer
CN-College-Exam-Listening	F: Excuse me, this is the address, where do I find it? M: All right, you need a street map, here is one, and I will show you where it is.	Question: What does the woman want to do? Choices: (A) Find a place. (B) Buy a map. (C) Get an address.	(A) Find a place.
SLUE-P2-SQA5	1 Climate Stephens City is located in the humid subtropical climate zone (Köppen climate classification : Cfa) ...	Which regions have temperate climates?	mid-atlantic
DREAM-TTS	F: The movie next Tuesday has been canceled due to a lack of interest. M: what do you mean? F: Well by last night only a few tickets been sold.	Question: What can we conclude about the movie? Choices: (A) They want to buy the tickets for the movie. (B) The tickets for the movie were sold. (C) The movie will not be shown.	(C) The movie will not be shown.
Public-SG-SpeechQA	Today, speaking to a roomful of economists, I am inclined to confine myself to talk about markets and price, dollars ...	How can economics help solve complex healthcare challenges, as mentioned by the speaker?	Economics can help solve complex healthcare ...

Table 5: Examples from Speech Question Answering (SQA) datasets

Dataset	Context	Instruction	Answer
Clotho-AQA	(Wave sound)	Are there waves?	yes
WavCaps-QA	(Electronic Music playing)	What type of sound is being played?	The sound being played is music.
AudioCaps-QA	(Mechanical vibration sound)	What type of object or equipment is likely to produce a constant rattling noise and sharp vibrations?	A loose or worn-out bolt or screw on a machine or equipment is likely to produce a constant rattling noise and sharp vibrations.

Table 6: Examples from Audio Question Answering (AQA) datasets

Dataset	Context	Instruction (Sample)	Answer
IEMOCAP-Emotion	Thank you.	Can you interpret the emotions in the speaker's speech (frustration, anger, excited, neutral, happiness, surprise, sad)?	From the speaker's speech, it seems they are in a sad state.
MELD-Sentiment	Yeah, I'm not in that.	What sentiment signals can you hear in the speaker's speech (neutral, positive, negative)?	From the speaker's speech, it seems they are in a neutral sentiment state.
MELD-Emotion	Yeah, I'm not in that.	How does the speaker's speech reflect their emotional state (neutral, joy, disgust, sadness, surprise, anger, fear)?	Based on the speaker's speech patterns, it seems like they are feeling neutral.

Table 7: Examples from Emotion Recognition (ER) datasets

Dataset	Context	Instruction (Sample)	Answer
OpenHermes-Audio	Pretend to be Ghost, expressing frustration to Soap that they're no closer to finding the elusive enemy leader after weeks of searching.	Please follow the instruction in the speech	It feels like we've been chasing ghosts, Soap. After all these weeks, we're still no closer to finding the bastard. It's bloody infuriating.
ALPACA-Audio	identify the type of the sentence: she can play guitar.	Please follow the instruction in the speech	The type of sentence is a declarative sentence.

Table 8: Examples from Speech Instruction (SI) datasets

Dataset	Context	Instruction (Sample)	Answer
WavCaps	(Electronic Music playing)	Detail the ambient sounds included in the audio.	A sound is playing.
AudioCaps	(Mechanical vibration sound)	Describe the environmental sounds in the audio.	Constant rattling noise and sharp vibrations

Table 9: Examples from Audio Captioning (AC) datasets

Dataset	Context	Instruction (Sample)	Answer
VoxCeleb-Accent	because I was trying to get away from her, but where did you go? what did you do?	Based on their accent, where is the speaker most likely from?	From the audio, I guess the speaker is from USA.

Table 10: Examples from Accent Recognition (AR) datasets

Dataset	Context	Instruction (Sample)	Answer
VoxCeleb-Gender	because I was trying to get away from her, but where did you go? what did you do?	Can you determine the gender of the speaker from the audio (Male or Female)?	The speaker is a female.
IEMOCAP-Gender	For God's sake, Augie. It's- Grow up, we're not going to see the grunion.	From the audio, can you determine the speaker's gender (Male or Female)?	Based on the auditory cues, it sounds like the speaker is a male.

Table 11: Examples from Gender Recognition (GR) datasets

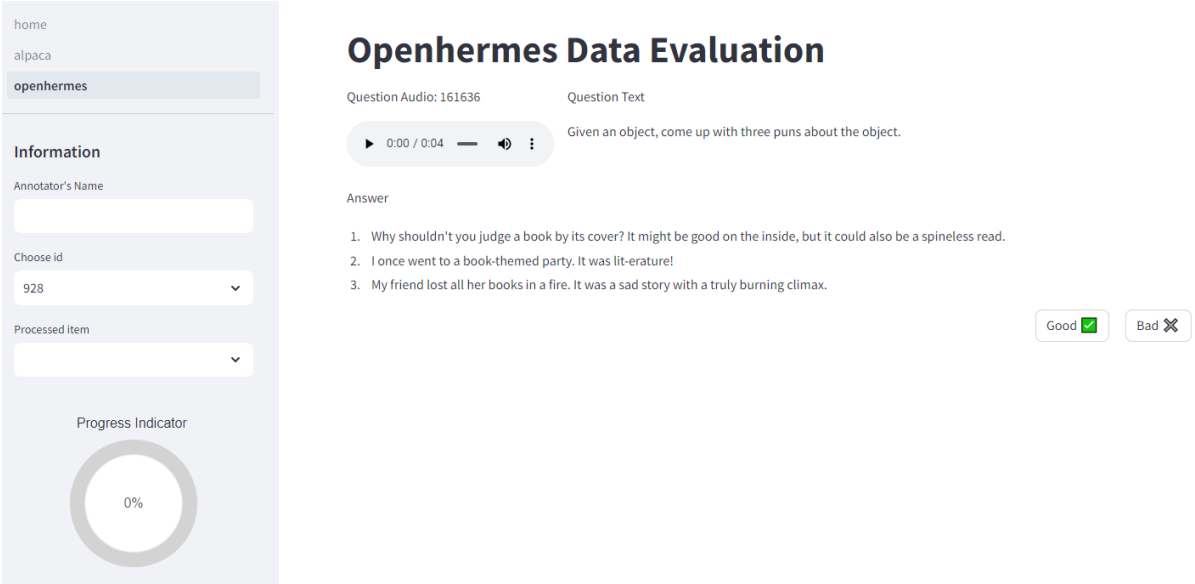

Figure 5: Our Labeling Platform for OpenHermes-Audio.

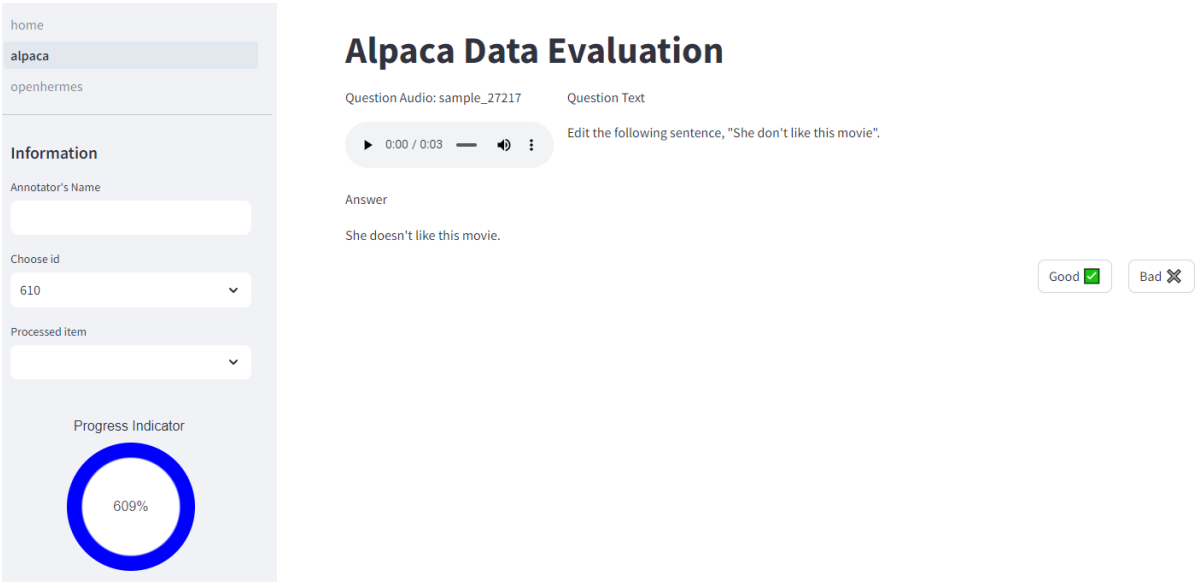

Figure 6: Our Labeling Platform for ALPACA-Audio.

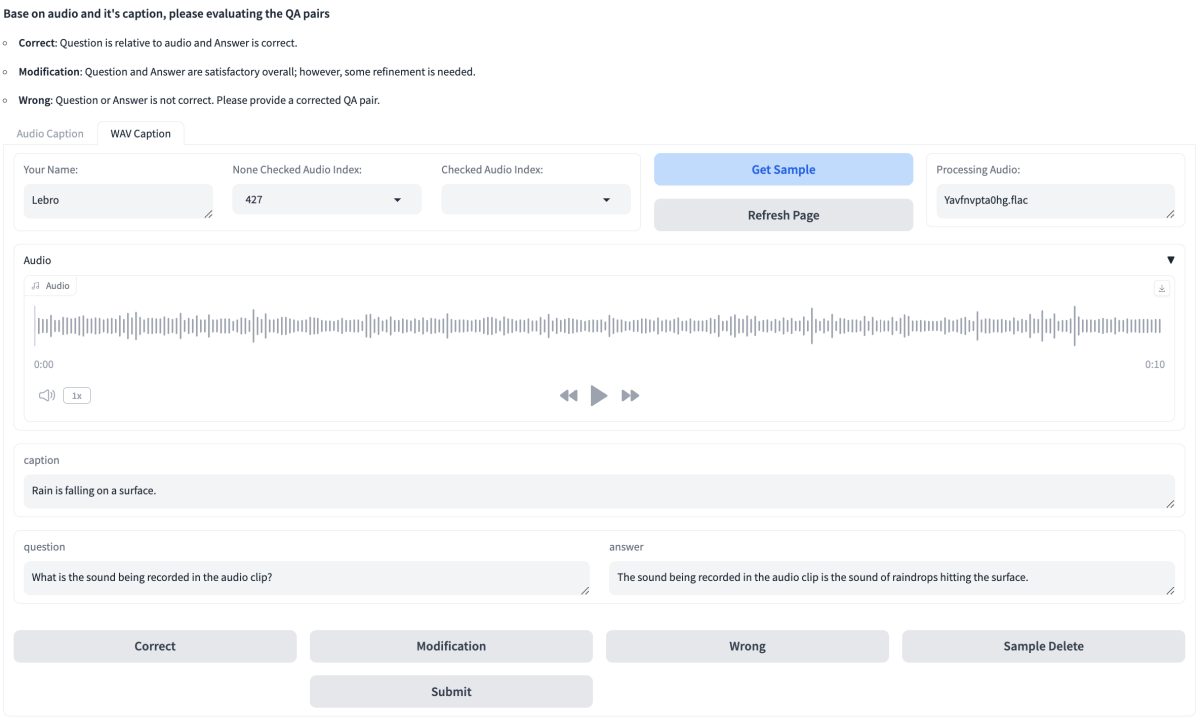

Figure 7: Our Labeling Platform for AudioCaps QA and WavCasp QA.