

# Evaluation on Audio-LLMs and Beyond

**Wang Bin**

Mar. 7, 2025

Event: Lorong AI

Address: WeWork @ 22 Cross Street, Singapore 048421, Level 3,  
Event Space

# LLM - Can do a lot of things in just **One Model**

- LLM solves problems by following human instructions
  - Why does the sky appear blue during the day? -> **QA**
  - Summarize the following article in one paragraph. -> **Summarization**
  - What is the sentiment of the above sentence? -> **Sentiment Analysis**
  - Write a Python function that returns the factorial of a number. -> **Coding**

# LLM - Can do a lot of things in just **One Model**

- LLM solves problems by following human instructions
  - Why does the sky appear blue during the day? -> **QA**
  - Summarize the following article in one paragraph. -> **Summarization**
  - What is the sentiment of the above sentence? -> **Sentiment Analysis**
  - Write a Python function that returns the factorial of a number. -> **Coding**

But before this LLM era, we usually need one model for one task.

# LLM - Can do a lot of things in just **One Model**

- LLM solves problems by following human instructions
  - Why does the sky appear blue during the day? -> **QA**
  - Summarize the following article in one paragraph. -> **Summarization**
  - What is the sentiment of the above sentence? -> **Sentiment Analysis**
  - Write a Python function that returns the factorial of a number. -> **Coding**

But before this LLM era, we usually need one model for one task.

**What is next? One model for all tasks? An intelligent agent?**

# An intelligent system takes all the surrounding information, E.g. as humans

- Information

- Textual data
- Audio and speech
- Image and video
- Sensor data
- Numerical or structured data
- Location and geospatial data
- Knowledge, reasoning, and others

Towards AGI –  
Artificial General Intelligence



# An intelligent system takes all the surrounding information, E.g. as humans

- Information

- Textual data
- Audio and speech
- Image and video
- Sensor data
- Numerical or structured data
- Location and geospatial data
- Knowledge, reasoning, and others



Towards AGI –  
Artificial General Intelligence

- Our Focus: **AudioLLM: Audio-based Large Language Models**

# Agenda

- What is AudioLLM? (in comparison with previous speech models)
- Design of new evaluation methods.
- AudioBench: A Universal Benchmark for Audio Large Language Models – NAACL 2025
- What is next?

# Conventional Audio / Speech Solutions

## “One task - one model” paradigm

- Automatic Speech Recognition – Whisper, Conformer, Wav2Vec 2.0
- Speech Translation – Whisper, SeamlessM4T, Translatotron 2
- Audio Captioning – CLAP, PANNs, AudioCLIP
- Sentiment Recognition – EmoReact, emotion2vec

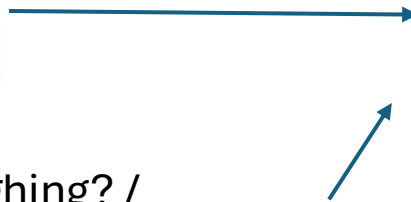
## Characteristics:

- Usually supervised learning
- Predefined prediction space



# AudioLLM – **One Model** for all audio/speech understanding tasks

Audio + Human Instruction = Answer



Audio LLM

The man is laughing because xx. /  
It was xx's speech given in 1989. /  
...

Why there is a man laughing? /  
When and where was this speech given? /  
Can you help me transcribe the speech? /  
Is it an indoor or outdoor environment? / ...

## Characteristics:

- Complete the task by following human instructions
- Can generalize to zero-shot / unsupervised scenarios
- Response in human language

# Before Model Development: Evaluation

- We need the performance measure.
  - To know what is the ideal case.
  - To monitor the progress of model training.
  - To compare models and methods.
  - ...

# How do researcher evaluate LLMs?

- Multiple choices questions
  - Unnatural for its common use cases
- Open-ended questions
  - How can we judge answer A is better than B?
- Human rating
  - Troublesome, no immediate feedback



Comments from X.com for GPT4.5 Release  
March 3, 2025

**TLDR: It is hard.**

# How do researcher evaluate AudioLLMs?

- Status: No unified paradigm yet.
- The new evaluation should incorporate the major paradigm shift.
  1. One model that handles (any) audio/speech tasks
  2. Complete task by human instructions
  3. Open-ended answer generation

# How do researcher evaluate AudioLLMs?

- Status: No unified paradigm yet.
- The new evaluation should incorporate the major paradigm shift.
  1. One model that handles (any) audio/speech tasks → The more, the better
  2. Complete task by human instructions → Instruction following measurement
  3. Open-ended answer generation → Judgement by LLMs with/without references

# AudioBench: A Universal Benchmark for AudioLLMs

- Task categories: Speech Understanding, Audio-Scene Understanding, Voice Understanding (Paralinguistic), Music Understanding, Singlish Understanding, ...

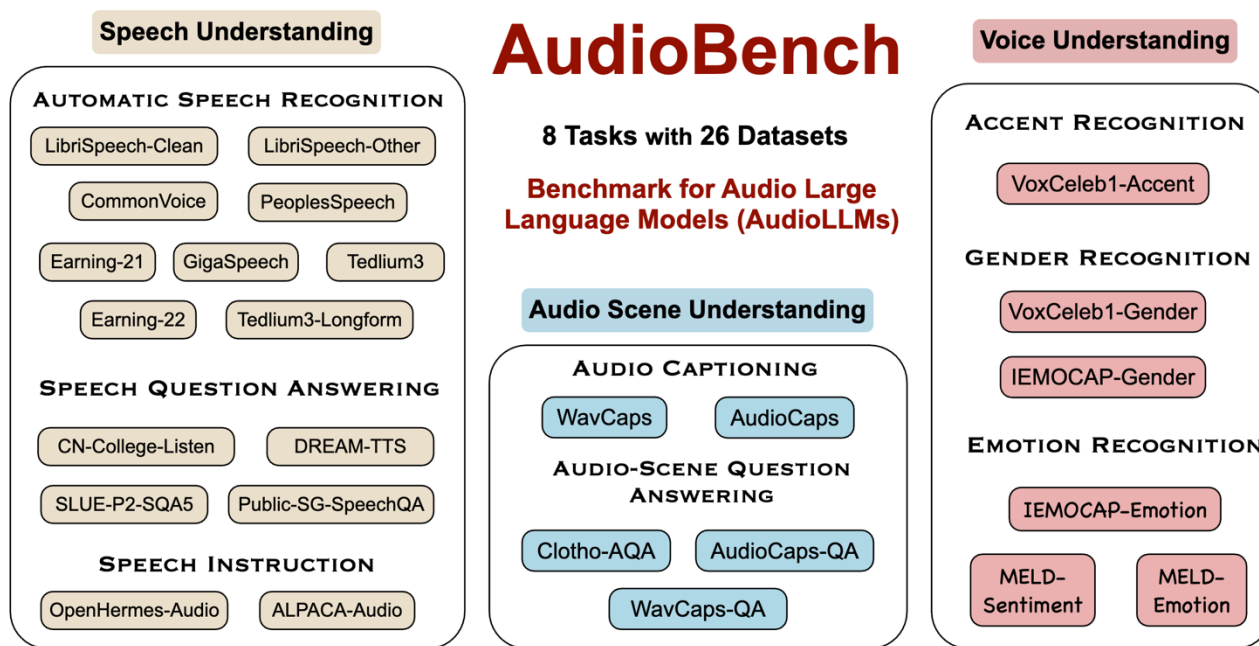
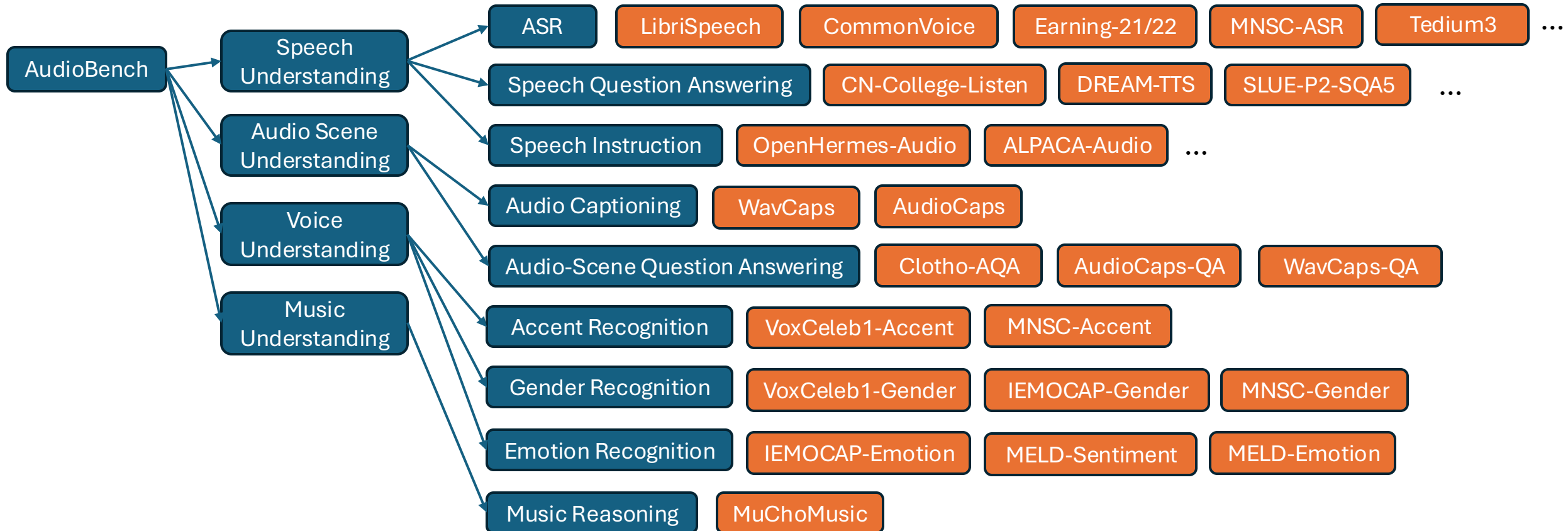


Figure 1: Overview of **AudioBench** datasets.

# AudioBench

- Collections of Evaluation Tasks



# AudioBench

- Evaluation Measurement
  - ASR – Word Error Rate
  - Speech Translation – BLEU score
  - Speech QA? -> **Model-as-a-judge (LLAMA-3-70B-Instruct)**



AudioLLM  
(SALMONN)

What emotions do you detect in the speaker's voice (frustration, anger, excited, neutral, happiness, surprise, sad)?

Model Answer:  
The speaker is expressing **frustration**  
**and anger** in their speech.

Reference Answer:  
Based on the speaker's speech patterns, it  
seems like they are **feeling anger**.

Judgement  
LLM Model

Judgement model:  
Explanation: The reference answer is "anger",  
while the model's answer is "frustration and  
anger". I think the model's answer is accurate and  
relevant to the reference, as it not only mentions  
anger but also adds frustration, which is a closely  
related emotion.

**Rating: 1**





# AudioBench

- Evaluation of AudioLLMs and Cascade Models

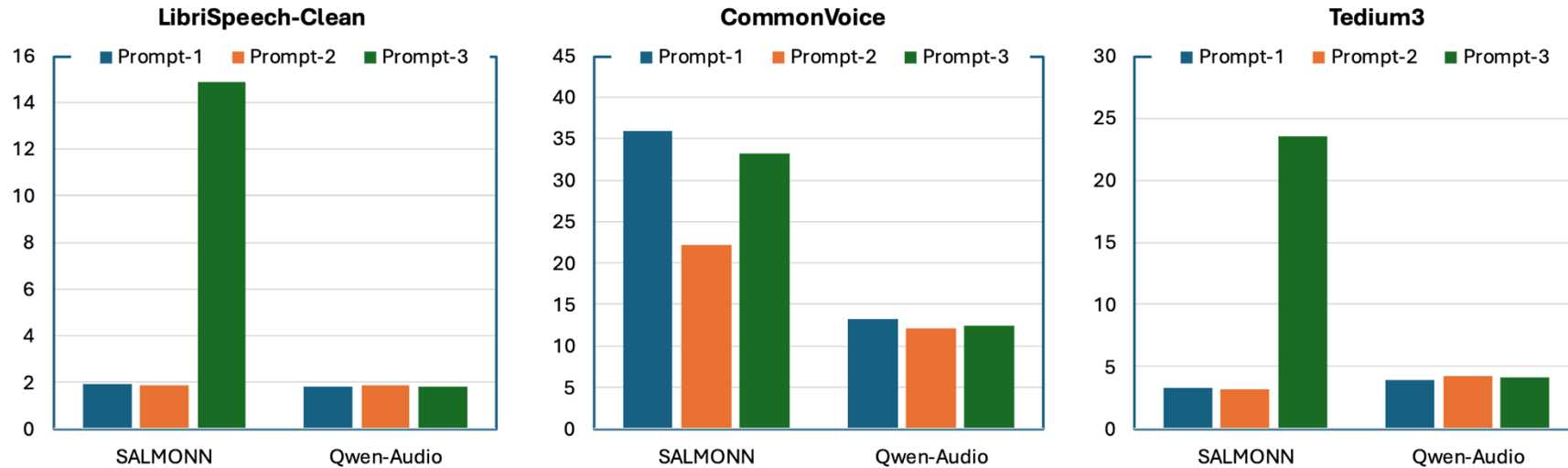
Dataset Name	AudioLLMs				Whisper+Llama
	SALMONN	Qwen-Audio-Chat	WavLLM	Qwen2-Audio-Instruct	
Speech Understanding					
LibriSpeech-Clean <sub>(↓)</sub>	55.58	2.25	2.10	3.20	1.83
LibriSpeech-Other <sub>(↓)</sub>	41.80	4.16	4.80	6.07	3.71
CommonVoice-15 <sub>(↓)</sub>	33.75	11.65	14.53	11.44	9.89
PeoplesSpeech <sub>(↓)</sub>	34.33	30.72	37.92	22.32	14.54
GigaSpeech <sub>(↓)</sub>	14.22	13.32	15.49	11.89	9.51
Tedlium3 <sub>(↓)</sub>	8.56	4.00	6.62	6.39	3.81
Tedlium3-Longform <sub>(↓)</sub>	18.39	45.29	45.37	95.35	4.75
Earning-21 <sub>(↓)</sub>	26.87	38.46	64.47	98.65	11.77
Earning-22 <sub>(↓)</sub>	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 <sub>(M.J.)</sub>	21.16	29.25	6.40	33.78	3.45
AudioCaps <sub>(M.J.)</sub>	34.37	47.99	4.17	40.78	2.47
WavCaps <sub>(METEOR)</sub>	17.72	24.02	9.78	21.34	13.89
AudioCaps <sub>(METEOR)</sub>	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

# AudioBench

- What about diverse prompts?

E.g. ASR Instructions

- "Please help me transcribe the speech into text.",
- "Transcribe the spoken words into written form.",
- "Listen to the speech and provide the text version.",
- "Transform the speech into a text document as transcriptions.",
- ...



Most recent models are becoming robust towards diverse prompts.

# AudioBench

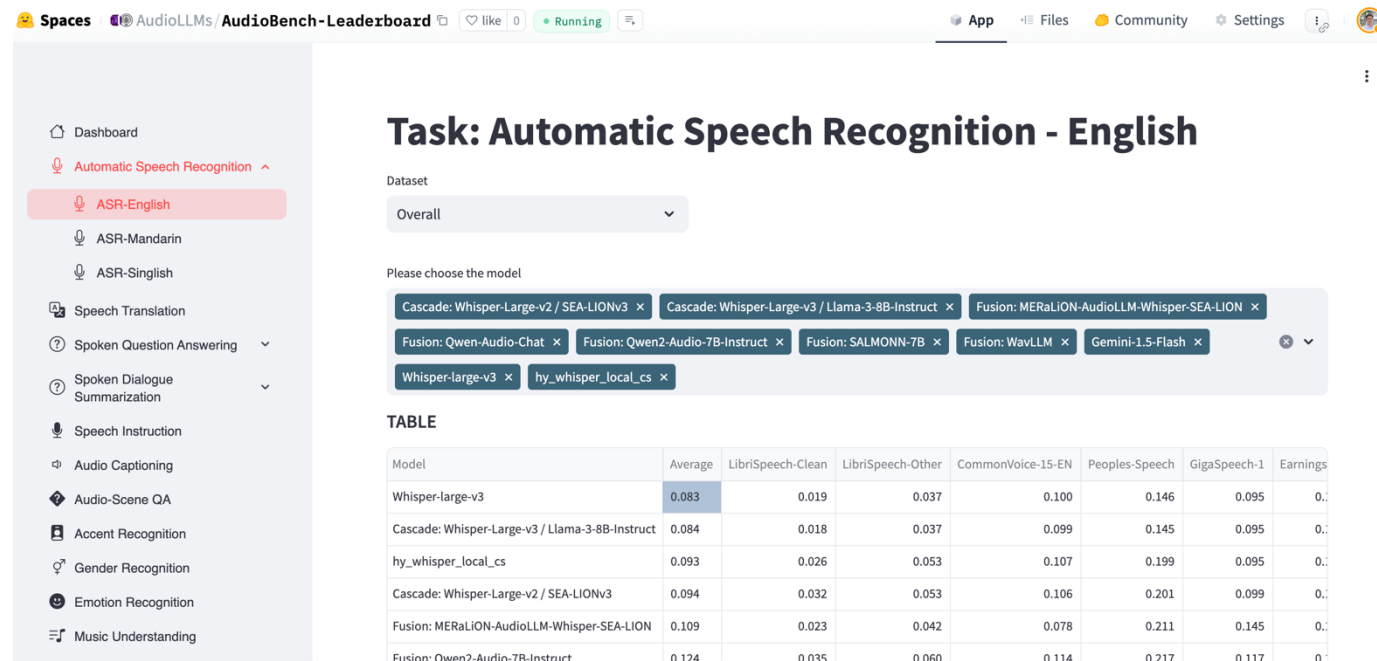


Correlation between judgement models

- What judgement model should we use?
  - LLAMA-3-8B-Instruct
    - Not have enough capability
  - GPT4o
    - Costly and will have version updates
  - LLAMA-3-70B-Instruct
    - Can fit in 1 H100 GPU after int4 quantization
    - Good instruction following and judgement quality
    - High correlation with GPT4o (especially for reference-based judgements)

# AudioBench - Resources

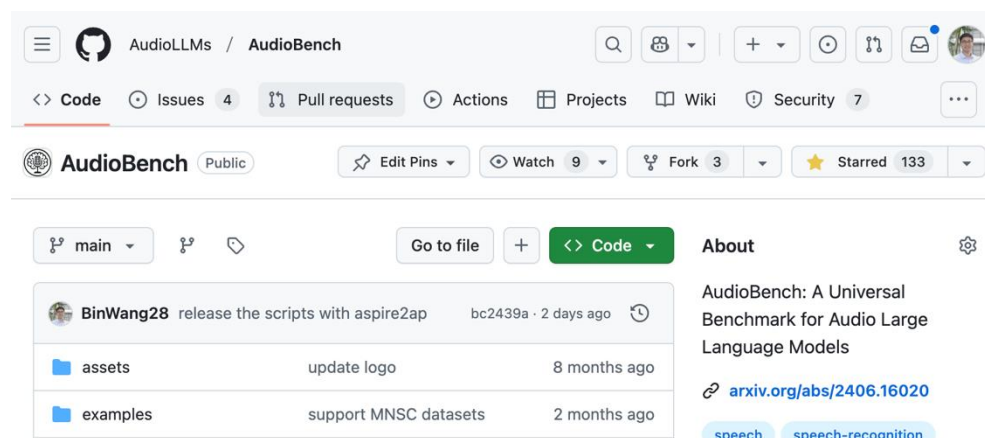
- Paper: <https://arxiv.org/pdf/2406.16020> (NAACL 2025)
- Leaderboard: <https://huggingface.co/spaces/AudioLLMs/AudioBench-Leaderboard>



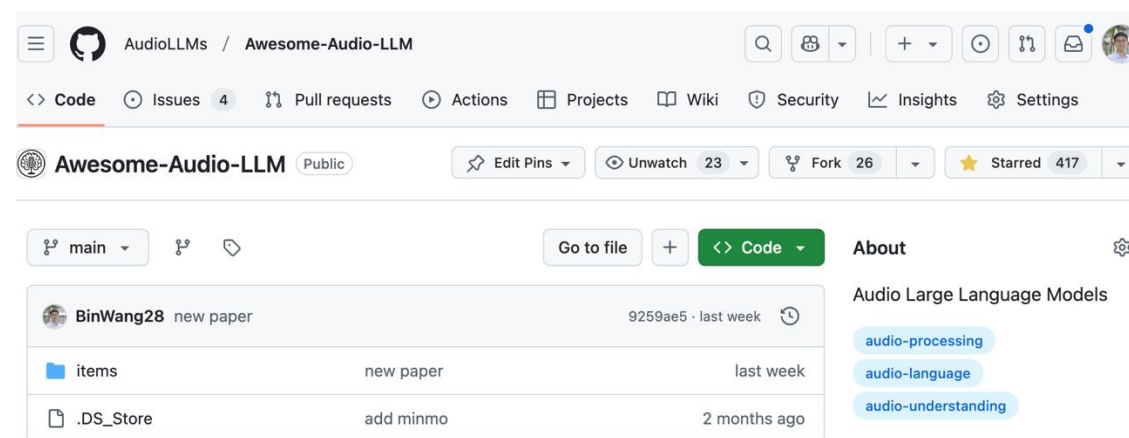
Wang, Bin, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. "Audiobench: A universal benchmark for audio large language models." NAACL, 2025

# AudioBench - Resources

- Paper: <https://arxiv.org/pdf/2406.16020> (NAACL 2025)
- Leaderboard: <https://huggingface.co/spaces/AudioLLMs/AudioBench-Leaderboard>
- Toolkit: <https://github.com/AudioLLMs/AudioBench>



AudioBench Toolkit



Collection of AudioLLM Publications

# What should we do next? AudioBench v2

- Instruction following of AudioLLMs
- Multilingual Support
- Multi-round Evaluation
- Evaluation on Speech Generation (Can hear and speak!)

# Next Talks

- Our Models!
- MOWE-Audio (ICASSP 2025)
- MERaLiON-AudioLLM (Technical Report, Open-Sourced)