An In-depth Investigation on the LLM-based Speech Understanding

Anonymous ACL submission

Abstract

The success of large language models (LLMs) has prompted efforts to integrate speech and audio data, aiming to create general foundation models capable of processing both textual and non-textual inputs. Recent advances, such as GPT-4o, highlight the potential for end-to-end Speech LLMs, which preserves non-semantic information and world knowledge for deeper speech understanding. By analyzing the development of Speech LLMs, this paper outlines distinct levels for Speech LLMs, ranging from basic automatic speech recognition (ASR) to advanced superhuman models capable of integrating non-semantic information with abstract acoustic knowledge for complex tasks. Moreover, we design a benchmark, **DitingBench**, that standardizes critical aspects across various tasks in these five levels, uncovering challenges in using abstract acoustic knowledge and completeness of capability. Our findings reveal gaps in handling paralinguistic cues and abstract acoustic knowledge, and we offer future directions. [1]

1 Introduction

Paradigms to process *language* have been reshaped thanks to LLMs and its scaling law. Given the success of LLMs, one may expect to integrate extensive data in *speech* and *audio* modality into LLMs (similar to visual language models (Liu et al., 2023; Li et al., 2023) [2]), resulting in a more general foundation model. Towards this path, the exploration on speech foundation models recently brings new research insights from the perspectives of multi-task and multi-lingual processing (Radford et al., 2023; Bapna et al., 2021; Zhang et al., 2023c; Seamless Communication et al., 2023; Pratap et al., 2024). A remarkable event is the release of GPT-4o, which is notable for its ability in open-ended speech-to-speech dialogue. Its performance in speech understanding, speech synthesis, and system latency has reached new levels, leading to a wave of studies on Speech LLMs. The next question is, *where are we now and where should we go?* To answer this, we begin by analyzing the benefits and potential advancements of using LLMs to understand speech.

Processing Speech using LLMs Compared to the traditional approach of feeding ASR-transcribed text (Radford et al., 2023) into text-only language models, unified speech-language models process raw audio or speech directly in an end-to-end fashion. The *benefits* for using LLMs to process speech are mainly two-fold.

I) Preservation of Non-semantic Information: Processing raw speech directly through language models allows for the preservation of paralinguistic information (Schuller et al., 2013) , such as emphasis, speaker identity, background sounds, emotions, and feelings, to the greatest extent possible. At the same time, it becomes easier to perceive the surrounding environment (Hu et al., 2020) and events happening nearby, such as alarm sounds, the hustle and bustle of an airport terminal, or other diverse forms of information that are difficult to enumerate exhaustively. For clarity, we classify these two types of information, which cannot be captured through ASR-transcribed text, as non-semantic information.

II) Acoustic Knowledge Retention: LLMs store vast amounts of world knowledge compared to traditional models (Roberts et al., 2020; Li et al., 2021; Srivastava et al., 2022). Therefore, starting with an LLM as the foundation for speech processing allows the model to have greater potential to achieve abstraction of expert speech/acoustic knowledge (e.g., inferring from cough and melody in some applications).

[1] The data and code are available at Anonymous GitHub

[2] There exists lighweight solutions for adapting language models to process data beyond text (e.g., visual or auditory), such as: 1) using a lightweight encoder and alignment process, and 2) discretizing data into tokens, which supports the autoregressive objectives of LLMs.

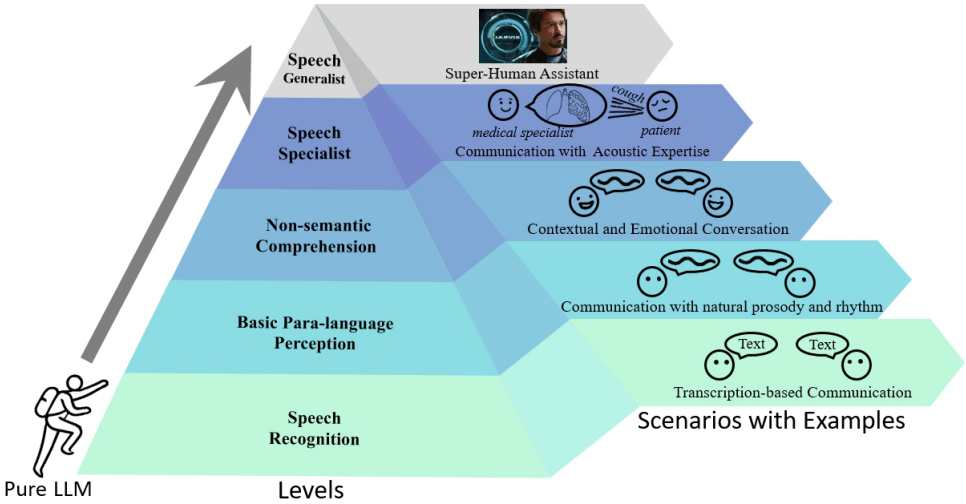

Figure 1: Levels of speech understanding using LLMs.

Five-level Speech Understanding These two benefits highlight the potential of Speech LLMs, achieving of which requires the models to perceive complete speech information and achieve abstraction of expert speech/acoustic knowledge. To this regards, we define five levels (see Fig. 1.) as below:

- **Basic Level** At the most basic level (**Level 1**), Speech LLMs should be able to recognize speech as text. The rationale for defining automatic speech recognition as the foundational level is that it serves as the basis for directly interacting with LLMs through speech. However, these capabilities at the basic level (e.g., speech recognition) offer limited additional benefits for ASR-equipped cascade paradigm to understand human speech as it is somehow equivalent to a combination with a ASR model and a text-only LLM.

- **Non-semantic Information Perception Levels** More advanced models (at **Level 2**) are expected to directly perceive basic paralinguistic information such as tone, pitch, and loudness; it (at **Level 3**) further enables comprehension of paralinguistic cues like emotions (e.g., sarcasm) and other types of non-semantic information, such as the surrounding environment.

- **Abstract Acoustic Knowledge Levels** At a higher level (at **Level 4**), models can integrate speech with specialized speech/audio knowledge to perform vertical tasks, such as medical assessments, pronunciation correction, or music appreciation. The ultimate goal (at **Level 5**) extends beyond a single type of abstract acoustic knowledge, aiming to generalize across diverse and complex speech tasks.

Benchmark However, these levels remain insufficiently intuitive. Therefore, we have preliminarily developed a benchmark to concretize and exemplify these capability levels. We designed the **DitingBench** [3] to evaluate Speech LLMs across various tasks that typically represent the characteristics of each level.

The benchmark covers a wide range of tasks, including speech recognition, language distinction, volume perception, emotion recognition, and more, with each task corresponds to a specific level of capability within Speech LLMs. The reliability of these evaluation sets was verified using human test, open-source and custom-trained models, demonstrating that the tasks are feasible and can be accomplished. The benchmark aims to comprehensive, tiered evaluate Speech LLMs' capabilities, and exploration of their ability to apply abstract acoustic knowledge.

Findings In the experiment, we found the following: **Human** was generally strong in tasks from Level 1 to 3. However, at higher levels, human performance was limited due to a lack of abstract acoustic knowledge, which Speech LLMs may start to outperform in certain tasks.**The current Speech LLMs**, though capable of surpassing human performance in a few areas, still fall short in terms

[3]The benchmark is named TingdiBench in honor of *Diting*, a mythical creature from Chinese folklore, notably featured in tales like *Journey to the West*. *Diting* is renowned for its extraordinary ability to detect subtle, non-semantic nuances in human speech that are imperceptible to ordinary beings. Inspired by this mythological figure, TingdiBench evaluates LLMs' capacity to understand nuanced, non-semantic cues in speech that goes beyond literal meaning.

of task diversity and comprehensiveness. Most models struggle with even basic paralinguistic information processing, highlighting the need for further improvement. We analyzed four reasons for the performance deficiency of Speech LLMs : 1) limited types of training data, 2) inability to comprehensively perceive acoustic information, 3) inadequate instruction following, and 4) weak LLM backbones.

The **contributions** of this paper are as follows: We outlining *five distinct levels* to better characterize the current state of speech language models. Additionally, we design a *benchmark* aligned with this five distinct levels, supplementing existing benchmarks with a variety of tasks. Finally, we present key *findings* from the benchmark, based on evaluations of both Speech LLMs and humans, and conduct a comprehensive *analysis* of the factors behind their suboptimal performance, offering insights and guidance for future model and architecture development.

2 Five Levels for Advanced Understanding Speech

To outline distinct levels for Speech LLMs towards an advanced understanding of speech. we first analyzed the development process of Speech LLMs in the past (in Sec. 2.1). Following that, we present our philosophy of the five distinct levels in Sec. 2.2.

2.1 The Background

Current Speech LLMs are mainly divided into two types: the cascade paradigm and the end-to-end paradigm. Below, we will focus on analyzing these two paradigms.

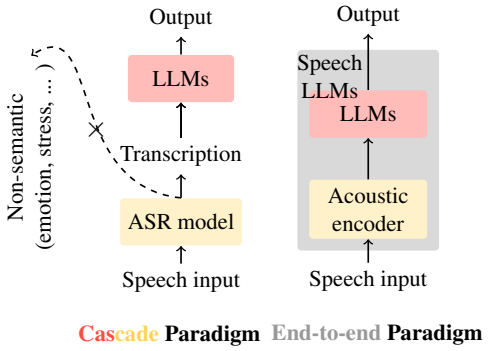

Figure 2: *Cascade* and *End-to-end* paradigms.

Cascade Paradigm A straightforward approach to understanding speech using LLMs is to feed speech transcriptions (in text format) into LLMs. This is known as the *cascade* paradigm (see the left in Fig. 2). While this method allows for basic speech understanding, it lacks the ability to perceive non-semantic information (e.g., emotion, stress) within LLMs. This hinders a deeper understanding of the spoken content as its non-semantic information is often crucial for fully grasping the intent or nuances in speech.

End-to-end Paradigm In contrast, an *end-to-end* Speech LLM can process both semantic and non-semantic information simultaneously within a single model. This approach not only retains more detailed information within the LLM but also allows the world knowledge embedded in the LLM to interact directly with speech data. Note that this end-to-end speech paradigm introduces additional complexity, as it requires LLMs to handle raw speech data, which operates at a lower level compared to textual inputs.

In summary, the end-to-end solution enables LLMs to directly handle non-semantic information, such as emotions. Additionally, due to its stronger perceptual capabilities, it holds greater potential for understanding and applying abstract acoustic knowledge. As a result, end-to-end solution can be considered the future direction for the development of Speech LLMs.

2.2 The Philosophy of the Five Levels

With the rise of speech LLMs, there is an increasing demand to understand information beyond text, particularly speech and sound that involves environmental sound, speaker information and paralinguistic speech information. The core idea is that speech conveys richer information than text alone, positioning ASR (Automatic Speech Recognition) as a foundational level. End-to-end Speech LLMs can begin with ASR capabilities to directly leverage the capabilities of text LLMs. And then, it progressively incorporate more advanced comprehension of non-semantic features. Finally it contains the ability to retain and apply abstract acoustic knowledge. This progress can be described as evolving through the following five levels:

Level 1. *Speech Recognition Level* *At the most basic level, a speech language model should be capable of recognizing text.*

These tasks form the most fundamental requirements for interacting with large models using speech. However, at Level 1, the model offers limited advantages over a traditional cascade paradigm (e.g., feeding ASR-transcribed text into LLMs).

Table 1: Levels of speech understanding using LLMs

Level		Semantic Information	Non-Semantic Information	Abstract Acoustic Knowledge	Remark
-	**Pure LLM**	-	-	-	Without speech input.
L1	**Basic ASR**	✓	✗	✗	Recognizing Speech as texts.
L2	**Paralinguistic Perception**	✓	only paralinguistic	✗	Perceiving direct paralinguistic *information* like tone, pitch, loudness, rhythm, and speech rate.
L3	**Non-semantic Comprehension**	✓	✓	✗	Comprehending non-semantic *information* like speaker identity, gender, age, emotional state, and environmental sounds.
L4	**Speech Specialist**	✓	✓	specialist	Understanding speech with *specific* acoustic *knowledge*.
L5	**Speech Generalist**	✓	✓	generalist	Understanding speech with *general* acoustic *knowledge*.

The real benefits of Speech LLMs begin to emerge at the next level, with the ability to capture non-semantic features such as paralinguistic information.

Level 2. ***Basic Paralinguistic Perception Level*** *At this level, Speech LLMs gain the ability to perceive basic paralinguistic features in speech, such as tone, pitch, volume, rhythm, and speech rate.*

Banse and Scherer (1996), Shafran et al. (2003), Schuller et al. (2013), and Wang et al. (2015) highlight the importance of basic paralinguistic features, such as volume changes, pitch, and speech rate, in identifying emotions, age, gender, and accent. These features serve as the foundation for higher-level paralinguistic information and provide distinct advantages over pure text-based models (or Speech LLMs at Level 1). While this lays the foundation for more advanced capabilities, the insights derived at this level are still relatively shallow. For a deeper understanding, we must move to Level 3, where a model comprehends a broader range of non-semantic information.

Level 3. ***Non-semantic Comprehension Level*** *At this stage, the Speech LLM extends beyond basic paralinguistic features and is capable of comprehending and interpreting more complex non-semantic information, such as emotions, sarcasm, and heightened states like pride.*

For example, emotions are higher-level human experiences that involve cognitive functions, distinguishing them from basic paralinguistic information. Interestingly, even some higher animals, like pet dogs, can perceive these types of non-semantic information. To fundamentally distinguish humans from animals from the perspective of speech understanding, we designed Level 4 by leveraging human strengths in higher-level cognitive capabilities.

Level 4. ***Speech Specialist Level*** *Speech Specialist can integrate expert-level abstract acoustic knowledge to handle a few specific, complex tasks.*

This requires integrating abstract acoustic knowledge which are advanced knowledge derived from acoustic information. This goes beyond mere recognition and comprehension at Level 1 and Level 2, requiring the model to apply higher-order thinking skills (such as analysis, evaluation, and creation) based on acoustic information [4], according to Bloom's cognitive taxonomy (Krathwohl, 2002). Despite these abilities, the model at this level remains domain-specific, which leads to the need for a fully generalized Speech LLM, as defined by Level 5.

Level 5. ***Speech Generalist level*** *The ultimate level, Speech Generalist, represents a comprehensive speech model that functions as a generalist.*

This vision of Speech Generalist represents the culmination of speech understanding, combining domain expertise, adaptability, and the capacity to

[4]This capability benefits a range of tasks, e.g., 1) using cough sounds to identify the type and origin of the cough, 2) pronunciation correction, 3) music appreciation, 4) stethoscope auscultation, 5) early screening for depression and Parkinson's disease, and 6) understanding animal vocalizations.

exceed human performance in speech-based tasks. Speech Generalist's potential to outperform humans probably stems from its ability to scale learning time and superior memory retention compared to humans.

3 Benchmarking

3.1 The Benchmark: DitingBench

To implement the five levels (Sec. 2), we aim to build a benchmark to concretes these levels. Though previous benchmarks have contributed significantly, they focus mainly on the first and third levels, neglecting foundational paralinguistic and abstract acoustic knowledge. Our benchmark completes the tasks at the 2nd, 4th, and 5th levels, and supplements the missing tasks in the 1st and 3rd levels. It also addresses issues such as data class imbalance in the original benchmark and the potential for models to bypass true speech understanding by inferring sentiment from text. For details, please refer to App. A.

Selection for Five-level Tasks DitingBench is structured to align with the five levels of speech understanding[5], see overview in Tab. 9. The tasks are organized into five levels: **Level 1** focuses on testing the **recognition capabilities** of Speech LLMs, including ASR, lyrics transcription, and term recognition tasks. **Level 2** evaluates **foundational perception** abilities, such as pitch and volume perception for tasks like age, gender, and emotion recognition. **Level 3** assesses **non-semantic comprehension**, incorporating tasks like emotion-integrated translation, environment perception, and emotional intensity recognition. **Level 4** explores the application of **abstract acoustic knowledge**, specifically focusing on medical-related contexts. Finally, **Level 5** represents a **Speech Generalist** with multidisciplinary knowledge, capable of fostering creativity and diverse thinking. Note that each subsequent level builds upon the foundation of the preceding ones.

3.2 Benchmarked Objects

Humans To conduct an initial evaluation of human performance, we created evaluation subsets by randomly selecting 80 samples per label for the objective multiple-choice tasks, and 80 samples in total for the other tasks. Four students (two males and two females) with strong English proficiency completed the assessments. The results are recorded in Tab. 2. The participant information and consistency test is in App. C.1.

Speech LLMs There are four types of Speech LLMs, see more details in Sec. 5. We selected an open-source model for each type, except for video LLMs, where the performance on audio-only tasks is not stable. For speech-related models, we chose Qwen2-Audio (Chu et al., 2024) for its strong performance. We selected Mu-llama (Liu et al., 2024a) for the music model and GAMA (Ghosh et al., 2024b) for the audio model. Additionally, we tested SALMONN (Tang et al., 2023) as a mixed audio and speech model. We further test GPT-4o advanced speech mode. Because only some models supports the speech instruction, we utilize the text instruction to ensure fair comparison. For more details on model replication, please refer to App. C.2.

3.3 Benchmarking Results

Humans As seen in Tab. 2, human performs generally well from Level 1 to 3. However, it becomes worse at higher levels due to a lack of acoustic knowledge. On the other side, speech understanding for humans are generally better than speech language models.

Take-away 1. ***Human performance:*** *Human generally performs well in speech understanding from Level 1 to 3, but fails to reach a higher level probably due to a lack of abstract acoustic knowledge.*

Speech LLMs As shown in Tab. 2, Speech LLMs exhibit a significant weakness in Level 2 which consists of basic listening abilities of the human. These models are currently focused on directly addressing high-level tasks while neglecting basic paralinguistic information perception, thereby the model fails to shows generalization at higher level. Furthermore, most models do not fully satisfy the requirements at any given level, highlighting a lack of consideration for both task diversity and comprehensiveness. Notably, Qwen2-Audio has outperformed humans in tasks like emotion recognition. This suggests that Speech LLMs have the potential to detect subtle changes in speech, even beyond human capabilities.

Take-away 2. ***Speech LLMs:*** *Speech LLMs still struggle with non-semantic perception and comprehension from Level 1 to Level 3, despite excelling*

[5]The current version lacks a complete set of tasks for these levels, which we reserve for future work.

Table 2: Performance of Speech LLMs on DitingBench.

Level	Task	Human Baseline	Models				
			GPT-4o	MuLLaMA	GAMA	SALMONN	Qwen2-Audio
L1	Language Identification	×	88.50%	8.48%	×	35.17%	96.44%
	Auto-Speech Recognition	15.49*	10.24*	×	×	5.45*	4.63*
	ASR for Legal Terms	98.50%	26.47%	×	×	×	81.04%
	ASR for Medical Terms	97.50%	41.87%	×	×	×	53.86%
	Auto-Lyrics Transcription	26.88*	×	×	×	77.12*	32.48*
	- Hallucination Rate	3.00%	×	×	×	29.26%	38.21%
L2	Volume Perception	100.00%	×	50.00%	11.98%	53.22%	48.96%
	Pitch Perception	96.25%	29.33%	33.78%	41.50%	50.00%	50.00%
	Binaural Effect Perception	100.00%	41.38%	×	×	49.88%	×
	Loudness Assessment	85.63%	×	49.77%	×	×	50.13%
	Speech Rate Assessment	76.25%	×	50.00%	×	×	44.93%
	Speech Pause Detection	91.88%	×	50.00%	49.97%	×	51.70%
L3	Ambient Noise Detection	91.88%	45.27%	50.00%	60.17%	49.88%	50.00%
	Acoustic Scene Classification	90.28%	16.36%	5.07%	12.05%	20.74%	27.67%
	Speaker's Age Prediction	52.59%	13.43%	33.60%	×	36.87%	38.55%
	Speaker's Gender Recognition	97.50%	×	50.00%	×	48.12%	79.60%
	Speech Emotion Recognition	50.71%	16.77%	9.20%	3.68%	10.93%	79.51%
	Cappella Emotion Recognition	62.25%	21.50%	12.42%	7.08%	14.62%	62.38%
	Emotion Intensity Perception	97.50%	72.67%	50.00%	50.00%	49.29%	50.00%
	Emotion Translation†	3.68	0.32	×	×	0.27	0.31
	Singing Detection	99.38%	53.11%	50.00%	64.82%	56.47%	50.22%
L4	COVID-19 Risk Detection	60.63%	×	×	×	50.00%	14.17%
	Cough Type Classification	52.50%	40.33%	50.16%	44.17%	49.17%	43.39%
	Cough Origin Diagnosis	32.19%	×	×	×	4.01%	25.65%
	Cough Severity Assessment	45.42%	24.12%	30.85%	28.50%	38.24%	33.86%
	Lung Risk Screening	49.38%	×	47.62%	×	×	50.16%
L5	Spoken English Coach†	1.39	0.15	1.29	0.44	0.48	0.54
	Voice Detective†	1.20	×	0.84	0.83	0.86	1.24

"×" indicates that the model fails to follow the instruction. "*" denotes that the metric is Word Error Rate (WER) and similar metrics, for which lower values indicate better performance. "†" indicates that the task is evaluated by GPT-4, with a score ranging from 1 to 4.

in some tasks, limiting their performance on more complex tasks at higher levels.

GPT-4o The results indicate that GPT-4o tends to reject audio-related tasks. Compared to other models, GPT-4o shows merit in emotion-related tasks but fails to demonstrate overwhelming advantages in understanding ability. We suppose its strength lies in its interaction capability. Therefore, we tested its ability to follow speech instructions, which directly evaluates its interaction skills. We also tested Qwen2-Audio, one of the few models that support speech instructions.

The performance is detailed in Tab. 3. Compared to the results with text instructions, GPT-4o performs better with speech instructions, while Qwen2-Audio loses most of its capabilities. However, there remains a significant gap compared to the best results achieved using text instructions.

Take-away 3. ***GPT-4o:*** *GPT-4o demonstrates clear advantages in following speech instructions, but there is still significant room for improvement.*

Future Prospects We observe that abstract acoustic knowledge presents a common bottleneck for both humans and Speech LLMs in reaching higher performance levels. Given superior capabilities of LLMs in knowledge acquisition, meanwhile, the deficiencies in diversity and completeness of capabilities can be ameliorated by incorporating additional training data. we contend:

Take-away 4. *Speech LLMs have the potential to exceed human capabilities, yet they currently fall short in addressing the full scope of tasks and integrating abstract acoustic knowledge.*

4 In-depth Analysis on Performance Deficiency

In this section, we discuss reasons for performance deficiency in the benchmark. We first consider composition of training data (in Sec. D). Then we

Table 3: Comparison of performance based on text instructions and speech instructions. We selected tasks in which at least one model performed well under text instruction conditions. Details about the speech instruction can be found in App. C.4.

Task	Text instructions		Speech instructions	
	GPT-4o	Qwen2-Audio	GPT-4o	Qwen2-Audio
Language Identification	88.50%	**93.01%**	**91.45%**	18.64%
Auto-Speech Recognition	10.24	**4.63**	**14.65**	22.39
Speech Emotion Recognition	16.77%	**79.51%**	**23.46%**	×
Emotion Intensity Perception	**72.67%**	50.00%	**10.84%**	×

analyse the model from three perspectives: 1) perception of acoustic information (in Sec. 4.1), and 2) capacity of LLM backbone (in Sec. 4.2). At the same time, we also noticed that the ability of instruction following is problematic (in Sec. E).

Table 4: Comparison of task-specific model and LLMs. The *Small* model uses Transformer with 10M parameters.

Task	Model	Result vs LLMs
Language Identi.	Whisper	91.45%vs**96.62%**
Auto-Speech Recog.	Whisper	**2.44** vs 4.63
Auto-Lyrics Trans.	Whisper	**22.10** vs 32.48
ASR for Legal	Whisper	33.33%vs**81.04%**
ASR for Medical	Whisper	34.98%vs**53.86%**
Volume Perception	Small model	**100.00%**vs53.22%

4.1 Inability to Comprehensively Perceive Acoustic Information

The current end-to-end paradigm universally adopts the stacking paradigm. However, the stacking paradigm may suffer from two types of information loss: 1) the latent representation produced by the acoustic encoder does not fully capture or convey the necessary information, and 2) the acoustic encoder fails to transfer all the information to the downstream LLMs.

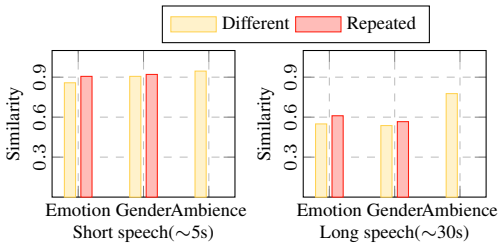

Figure 3: Representation cosine similarity of different speeches. Each speech pair has the same content but is spoken in a different style. The representation is generated by the Whisper encoder.

We first investigate whether the representation loses information. We compare the speech features generated from the same text content, which are spoken by different genders and with different emotions. The results, shown in Fig. 3, indicate that there is no significant difference between different speech samples. This suggests that emotion and gender information is lost during the acoustic encoder process. This could explain why some Speech LLMs perform poorly on certain tasks.

We then assess whether information is lost during the transfer from the acoustic encoder to downstream LLMs. We select cases from the ASR task where the WER is higher than 20%, as shown in Tab. 5. We found that the error types is different between the Whisper and Speech LLMs. Considering that Qwen2-Audio is built on Whisper, the results confirm that LLMs cannot correct errors from the acoustic model. A notable difference between Whisperand Speech LLMs is the tendency of the latter to produce overlong outputs, which is a form of hallucination.

Table 5: Two types of recognized errors. Two recognition errors are defined: "truncation" (outputs >20% shorter) and "over-long" (outputs >20% longer) than the reference.

Model	Total	Truncation	Over-long
Whisper	64	3	0
Qwen-Audio	68	5	6
Qwen2-Audio	149	89	3
SALMONN	251	154	5

Another notable phenomenon is that almost 60% of errors are due to truncation, which is due to the LLM's propensity to generate semantically complete sentences. Additionally, we observed that the Speech LLMs sometimes omits the start of a sentence, which does not happen with Whisper. This proves that Speech LLMs suffer the loss of information transfer between the LLMs and the acoustic encoder. The current stacked paradigm often tunes base on LLMs with most parameters frozen, which requires the acoustic features to fit the LLMs' rep-

resentation space. This requirement hinders the seamless transmission of acoustic information to the LLMs, leading to premature termination of the generation process.

Take-away 5. *LLMs in current solutions fail to encode complete acoustic information.*

4.2 Weak LLM Backbones

Most Speech LLMs stack acoustic models with text LLMs, requiring the latter to process audio-like tokens. This raises the question of whether text LLMs can handle cross-modal tasks. We designed a direct task of converting a phoneme sequence into a complete sentence. The phoneme represents pronunciation in text format, thus understanding phonemes can demonstrate the model's potential to process audio. We designed three different tasks, as shown in Tab. 17.

Table 6: Results of LLMs processing phonemes. The results are assessed using the WER. In instances where LLMs generate hallucinations or decline to provide a response, the WER is recorded as 100%.

Model	**Seq.** ↓	**Token** ↓ zero-shot	**Token** ↓ one-shot
GPT-4o	17.5	8.3	8.3
Mixtra-7B	99.5	98.9	97.7
Qwen2-7B	99.3	98.3	95.8
Llama3-7B	97.5	89.6	87.9
Llama3.1-8B	94.0	83.7	78.0
Mixtra 8x7B	98.2	95.1	92.6
Qwen2-72B	93.4	75.4	73.5
Llama3.1-70B	80.5	51.1	46.9

We evaluate the most commonly used LLMs for building Speech LLMs, and the results are shown in Tab. 6. We found that the closed-source GPT-4o demonstrates a surprising ability to process phonemes, proving that it can easily be converted into a powerful Speech LLM. On the other hand, all open-source models fail to show potential in handling audio. And increasing model size does little to improve performance.

One explanation is that open-source models overlook potential audio-related tasks, which is quite unlike GPT-4o. This leads to a significant gap between the two types of models. A piece of evidence supporting this is that Llama 3.1, which emphasizes multi-modal capabilities (Dubey et al., 2024), shows a noticeable improvement in WER in token-level tasks and delivers robust performance with 70B parameters. Overall, open-source foundation models still have substantial room for improvement in their ability to handle audio-related tasks.

Take-away 6. *The used LLM backbone is relatively weak for current Speech LLMs.*

5 Related Work

Speech language models have seen a surge in development following the advent of LLMs. These outstanding works can generally be categorized into four main types.

Categorization of Speech LLMs Some works aim to build **universal multi-modal LLMs** (Su et al., 2023; Zhan et al., 2024; Wu et al., 2023b; Lyu et al., 2023; Zhang et al., 2023b; Shukor et al., 2023). Several studies focus on enhancing **music understanding**, an important area that has not yet received enough attention (Deshmukh et al., 2023; Zhan et al., 2024; Liu et al., 2024b). Most Speech LLMs aim to improve **speech-to-text tasks** and **multi-turn dialogue capabilities** (Fathullah et al., 2024; Shu et al., 2023; Wang et al., 2023b; Pan et al., 2023; Rubenstein et al., 2023; Zhang et al., 2023a; Bai et al., 2024; Wu et al., 2023a; Maiti et al., 2024; Wang et al., 2023a; Chu et al., 2024; Dubey et al., 2024). Some works utilize audio codec models to enhance audio processing performance (Chen et al., 2023; Kong et al., 2024; Nguyen et al., 2024; Das et al., 2024; Gong et al., 2023). Inspired by these efforts, several studies (Tang et al., 2023; Ghosh et al., 2024a; Hu et al., 2024) combine acoustic and semantic codecs to integrate audio and speech processing capabilities into a single model.

6 Conclusion

In this paper, we explored the development of LLMs in speech understanding, outlining five distinct levels for Speech LLMs, from basic ASR to advanced models integrating non-semantic information and abstract acoustic knowledge. We designed a benchmark for consistent performance evaluation and identified current limitations in speech understanding by humans and LLMs. Our evaluation of GPT-4o's speech capabilities, as well as other recent Speech LLMs, highlights challenges in following speech instructions and structural flaws in existing models, particularly in acoustic information transfer and foundational LLM potential. This work offers a structured evaluation approach and valuable insights for Speech LLMs.

Limitation

The current work has the following limitations:

- It primarily focuses on speech understanding, without considering conversation-related data.
- The audio data for some tasks is insufficient, and we plan to include more data in the future.
- Our task setup for the fourth and fifth levels is not yet diverse enough.

We will continue to dedicate human resources to focus on the evaluation of large speech models, addressing the aforementioned shortcomings, and striving to achieve a more comprehensive assessment of Speech LLM. During the review period, we have already uploaded all our data and evaluation code to an anonymous GitHub repository, and we welcome everyone to join our discussions through it. After the review process is completed, we will release our data on Hugging Face, and we also hope to gather your opinions and suggestions on more tasks. We will promptly organize and publish additional evaluation tasks.

Ethical Considerations

Potential Risks

Our research includes human evaluation data, and if released, this data may pose uncontrolled privacy and preference leakage risks. Therefore, we will carefully consider various ethical and moral risks before deciding whether to release the details of this human evaluation data.

Artifact Use Consistent With Intended Use

We only used these data in the research, including appropriate human-annotated data and data processed using Speech LLM models, and did not use the data for any other purposes.

We only used the tools to construct our benchmark data and did not use them for any other purposes.

Documentation Of Artifacts

The use of our benchmark

Our benchmark is used to assess the level of understanding that Speech LLMs have of speech, evaluate their performance, and help related research identify their issues and shortcomings.

Privacy protection during the data collection process

All the data we collected come from publicly available datasets and strictly comply with the relevant privacy protection protocols.

Privacy protection during the data acquisition process

We did not collect any real human speech data, only synthesizing some data using TTS tools, and the synthesis process strictly adhered to the relevant privacy protocols.

Privacy protection of volunteers

We discuss our participants anonymously in both Appendix Section B.5 and Appendix Section C.1.1.

Information About Use Of AI Assistants

We utilized AI models primarily for grammar checking to ensure linguistic accuracy and clarity. Additionally, in certain instances, we leveraged AI-assisted tools for code generation, enhancing efficiency and precision in the development process. The AI models were not used for any other purposes.

References

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *arXiv preprint arXiv:2406.13340*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Ye Bai, Jingping Chen, Jitong Chen, Wei Chen, Zhuo Chen, Chen Ding, Linhao Dong, Qianqian Dong, Yujiao Du, Kepan Gao, et al. 2024. Seed-asr: Understanding diverse speech and contexts with llm-based speech recognition. *arXiv preprint arXiv:2407.04675*.

Rainer Banse and Klaus R Scherer. 1996. Acoustic profiles in vocal emotion expression. *Journal of personality and social psychology*, 70(3):614.

Ankur Bapna, Yu-an Chung, Nan Wu, Anmol Gulati, Ye Jia, Jonathan H Clark, Melvin Johnson, Jason Riesa, Alexis Conneau, and Yu Zhang. 2021. Slam: A unified encoder for speech and language modeling via speech-text joint pre-training. *arXiv preprint arXiv:2110.10329*.

Gunvant Chaudhari, Xinyi Jiang, Ahmed Fakhry, Asriel Han, Jaclyn Xiao, Sabrina Shen, and Amil Khanzada. 2020. Virufy: Global applicability of crowdsourced and clinical datasets for ai detection of covid-19 from cough. *arXiv preprint arXiv:2011.13320*.

Qian Chen, Yunfei Chu, Zhifu Gao, Zerui Li, Kai Hu, Xiaohuan Zhou, Jin Xu, Ziyang Ma, Wen Wang, Siqi Zheng, et al. 2023. Lauragpt: Listen, attend, understand, and regenerate audio with gpt. *arXiv preprint arXiv:2310.04673*.

Chung-Ming Chien, Jheng-Hao Lin, Chien-yu Huang, Po-chun Hsu, and Hung-yi Lee. 2021. Investigating on incorporating pretrained and learnable speaker representations for multi-speaker multi-style text-to-speech. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8588–8592.

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*.

Ondřej Cífka, Constantinos Dimitriou, Cheng-i Wang, Hendrik Schreiber, Luke Miner, and Fabian-Robert Stöter. 2023. Jam-ALT: A formatting-aware lyrics transcription benchmark. *arXiv preprint arXiv:2311.13987*.

Nilaksh Das, Saket Dingliwal, Srikanth Ronanki, Rohit Paturi, David Huang, Prashant Mathur, Jie Yuan, Dhanush Bekal, Xing Niu, Sai Muralidhar Jayanthi, et al. 2024. Speechverse: A large-scale generalizable audio language model. *arXiv preprint arXiv:2405.08295*.

Soham Deshmukh, Benjamin Elizalde, Rita Singh, and Huaming Wang. 2023. Pengi: An audio language model for audio tasks. *Advances in Neural Information Processing Systems*, 36:18090–18108.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Simon Durand, Daniel Stoller, and Sebastian Ewert. 2023. Contrastive learning-based audio to lyrics alignment for multiple languages. In *IEEE International Conference on Acoustics, Speech and Signal Processing ICASSP 2023, Rhodes Island, Greece, June 4-10, 2023*, pages 1–5. IEEE.

Yassir Fathullah, Chunyang Wu, Egor Lakomkin, Ke Li, Junteng Jia, Yuan Shangguan, Jay Mahadeokar, Ozlem Kalinli, Christian Fuegen, and Mike Seltzer. 2024. Audiochatllama: Towards general-purpose speech abilities for llms. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5522–5532.

Mohammad Fraiwan, Luay Fraiwan, Basheer Khassawneh, and Ali Ibnian. 2021. A dataset of lung sounds recorded from the chest wall using an electronic stethoscope. *Data in Brief*, 35:106913.

Ming Gao, Hang Chen, Jun Du, Xin Xu, Hongxiao Guo, Hui Bu, Jianxing Yang, Ming Li, and Chin-Hui Lee. 2024. Enhancing voice wake-up for dysarthria: Mandarin dysarthria speech corpus release and customized system design. *arXiv preprint arXiv:2406.10304*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, S Sakshi, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024a. Gama: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

Sreyan Ghosh, Sonal Kumar, Ashish Seth, Chandra Kiran Reddy Evuru, Utkarsh Tyagi, Sakshi Singh, Oriol Nieto, Ramani Duraiswami, and Dinesh Manocha. 2024b. GAMA: A large audio-language model with advanced audio understanding and complex reasoning abilities. *arXiv preprint arXiv:2406.11768*.

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

François Hernandez, Vincent Nguyen, Sahar Ghannay, Natalia A. Tomashenko, and Yannick Estève. 2018. TED-LIUM 3: Twice as much data and corpus repartition for experiments on speaker adaptation. In *Speech and Computer - 20th International Conference, SPECOM 2018, Leipzig, Germany, September 18-22, 2018, Proceedings*, volume 11096 of *Lecture Notes in Computer Science*, pages 198–208. Springer.

Di Hu, Zheng Wang, Haoyi Xiong, Dong Wang, Feiping Nie, and Dejing Dou. 2020. Curriculum audiovisual learning. *arXiv preprint arXiv:2001.09414*.

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

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. 2024. 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.

Javier Iranzo-Sánchez, Joan Albert Silvestre-Cerdà, Javier Jorge, Nahuel Roselló, Adrià Giménez, Albert Sanchís, Jorge Civera, and Alfons Juan. 2020. Europarl-st: A multilingual corpus for speech translation of parliamentary debates. In *2020 IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2020, Barcelona, Spain, May 4-8, 2020*, pages 8229–8233. IEEE.

Keith Ito and Linda Johnson. 2017. The lj speech dataset. https://keithito.com/LJ-Speech-Dataset/.

Shengpeng Ji, Jialong Zuo, Minghui Fang, Ziyue Jiang, Feiyang Chen, Xinyu Duan, Baoxing Huai, and Zhou Zhao. 2024. Textrolspeech: A text style control speech corpus with codec language text-to-speech models. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 10301–10305. IEEE.

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 preprint arXiv:2402.01831*.

David R Krathwohl. 2002. A revision of bloom's taxonomy: An overview. *Theory into practice*, 41(4):212–218.

Belinda Z. Li, Maxwell I. Nye, and Jacob Andreas. 2021. Implicit representations of meaning in neural language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1813–1827. Association for Computational Linguistics.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024a. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2024, Seoul, Republic of Korea, April 14-19, 2024*, pages 286–290. IEEE.

Shansong Liu, Atin Sakkeer Hussain, Chenshuo Sun, and Ying Shan. 2024b. Music understanding llama: Advancing text-to-music generation with question answering and captioning. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 286–290. IEEE.

Steven R Livingstone and Frank A Russo. 2018. The ryerson audio-visual database of emotional speech and song (ravdess): A dynamic, multimodal set of facial and vocal expressions in north american english. *PloS one*, 13(5):e0196391.

Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. 2023. Macaw-llm: Multi-modal language modeling with image, audio, video, and text integration. *arXiv preprint arXiv:2306.09093*.

Soumi Maiti, Yifan Peng, Shukjae Choi, Jee-weon Jung, Xuankai Chang, and Shinji Watanabe. 2024. Voxtlm: Unified decoder-only models for consolidating speech recognition, synthesis and speech, text continuation tasks. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 13326–13330. IEEE.

Tu Anh Nguyen, Benjamin Muller, Bokai Yu, Marta R Costa-Jussa, Maha Elbayad, Sravya Popuri, Paul-Ambroise Duquenne, Robin Algayres, Ruslan Mavlyutov, Itai Gat, et al. 2024. Spirit-lm: Interleaved spoken and written language model. *arXiv preprint arXiv:2402.05755*.

OpenAI. 2023. Gpt-4: Largest language model ever with 100 trillion parameters. Accessed: 2023-04-01.

Lara Orlandic, Tomas Teijeiro, and David Atienza. 2021. The coughvid crowdsourcing dataset, a corpus for the study of large-scale cough analysis algorithms. *Scientific Data*, 8(1):156.

Jing Pan, Jian Wu, Yashesh Gaur, Sunit Sivasankaran, Zhuo Chen, Shujie Liu, and Jinyu Li. 2023. Cosmic: Data efficient instruction-tuning for speech in-context learning. *arXiv preprint arXiv:2311.02248*.

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 2015, South Brisbane, Queensland, Australia, April 19-24, 2015*, pages 5206–5210. IEEE.

Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, et al. 2024. Scaling speech technology to 1,000+ languages. *Journal of Machine Learning Research*, 25(97):1–52.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 28492–28518. PMLR.

Srijith Radhakrishnan, Chao-Han Huck Yang, Sumeer Ahmad Khan, Rohit Kumar, Narsis A Kiani, David Gomez-Cabrero, and Jesper N Tegner. 2023.

Whispering llama: A cross-modal generative error correction framework for speech recognition. *arXiv preprint arXiv:2310.06434*.

Chandan KA Reddy, Ebrahim Beyrami, Jamie Pool, Ross Cutler, Sriram Srinivasan, and Johannes Gehrke. 2019. A scalable noisy speech dataset and online subjective test framework. *Proc. Interspeech 2019*, pages 1816–1820.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5418–5426. Association for Computational Linguistics.

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*.

Björn W. Schuller, Stefan Steidl, Anton Batliner, Felix Burkhardt, Laurence Devillers, Christian A. Müller, and Shrikanth S. Narayanan. 2013. Paralinguistics in speech and language - state-of-the-art and the challenge. *Comput. Speech Lang.*, 27(1):4–39.

Seamless Communication, Loïc Barrault, Yu-An Chung, Mariano Coria Meglioli, David Dale, Ning Dong, Mark Duppenthaler, Paul-Ambroise Duquenne, Brian Ellis, Hady Elsahar, Justin Haaheim, et al. 2023. Seamless: Multilingual expressive and streaming speech translation. *arXiv preprint arXiv:2312.05187*.

Izhak Shafran, Michael Riley, and Mehryar Mohri. 2003. Voice signatures. In *2003 IEEE workshop on automatic speech recognition and understanding (IEEE Cat. No. 03EX721)*, pages 31–36. IEEE.

Sheng Shen, Le Hou, Yanqi Zhou, Nan Du, Shayne Longpre, Jason Wei, Hyung Won Chung, Barret Zoph, William Fedus, Xinyun Chen, et al. 2023. Mixture-of-experts meets instruction tuning: A winning combination for large language models. *arXiv preprint arXiv:2305.14705*.

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 preprint arXiv:2308.15930*.

Mustafa Shukor, Corentin Dancette, Alexandre Rame, and Matthieu Cord. 2023. Unified model for image, video, audio and language tasks. *arXiv preprint arXiv:2307.16184*.

Tongyi SpeechTeam. 2024. Funaudiollm: Voice understanding and generation foundation models for natural interaction between humans and llms. *arXiv preprint arXiv:2407.04051*.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Yixuan Su, Tian Lan, Huayang Li, Jialu Xu, Yan Wang, and Deng Cai. 2023. Pandagpt: One model to instruction-follow them all. *arXiv preprint arXiv:2305.16355*.

Changli Tang, Wenyi Yu, Guangzhi Sun, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, and Chao Zhang. 2023. Salmonn: Towards generic hearing abilities for large language models. *arXiv preprint arXiv:2310.13289*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Cassia Valentini-Botinhao et al. 2017. Noisy speech database for training speech enhancement algorithms and tts models. *University of Edinburgh. School of Informatics. Centre for Speech Technology Research (CSTR)*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Jinliang Lu, Junhong Wu, Yuchen Liu, Chengqing Zong, and Jiajun Zhang. 2023a. Blsp: Bootstrapping language-speech pre-training via behavior alignment of continuation writing. *arXiv preprint arXiv:2309.00916*.

Kunxia Wang, Ning An, Bing Nan Li, Yanyong Zhang, and Lian Li. 2015. Speech emotion recognition using fourier parameters. *IEEE Trans. Affect. Comput.*, 6(1):69–75.

Mingqiu Wang, Wei Han, Izhak Shafran, Zelin Wu, Chung-Cheng Chiu, Yuan Cao, Nanxin Chen, Yu Zhang, Hagen Soltau, Paul K Rubenstein, et al. 2023b. Slm: Bridge the thin gap between speech and text foundation models. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

S. Weinberger. 2013. Speech accent archive.

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023a. On decoder-only architecture for speech-to-text and large language model integration. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 1–8. IEEE.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2023b. Next-gpt: Any-to-any multimodal llm. *arXiv preprint arXiv:2309.05519*.

Junichi Yamagishi, Christophe Veaux, and Kirsten MacDonald. 2019. CSTR VCTK Corpus: English multi-speaker corpus for CSTR voice cloning toolkit (version 0.92).

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, et al. 2024. Air-bench: Benchmarking large audio-language models via generative comprehension. *arXiv preprint arXiv:2402.07729*.

Jun Zhan, Junqi Dai, Jiasheng Ye, Yunhua Zhou, Dong Zhang, Zhigeng Liu, Xin Zhang, Ruibin Yuan, Ge Zhang, Linyang Li, et al. 2024. Anygpt: Unified multimodal llm with discrete sequence modeling. *arXiv preprint arXiv:2402.12226*.

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. *arXiv preprint arXiv:2305.11000*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Junbo Zhang, Zhiwen Zhang, Yongqing Wang, Zhiyong Yan, Qiong Song, Yukai Huang, Ke Li, Daniel Povey, and Yujun Wang. 2021. speechocean762: An open-source non-native english speech corpus for pronunciation assessment. In *22nd Annual Conference of the International Speech Communication Association, Interspeech 2021, Brno, Czechia, August 30 - September 3, 2021*, pages 3710–3714. ISCA.

Yu Zhang, Wei Han, James Qin, Yongqiang Wang, Ankur Bapna, Zhehuai Chen, Nanxin Chen, Bo Li, Vera Axelrod, Gary Wang, et al. 2023c. Google usm: Scaling automatic speech recognition beyond 100 languages. *arXiv preprint arXiv:2303.01037*.

A The Contribution of DitingBench

To better present our contributions, we will first analyze the shortcomings of existing benchmarks, and then describe the contributions we have made.

A.1 Existing Benchmark for Speech LLMs

Tab. 7 summarizes the coverage of existing benchmarks across different levels of speech model tasks, We will outline the shortcomings of the existing benchmark according to the distinct levels we have proposed.

- **Level 1** Everyone has overlooked the Auto-Lyrics Transcription task, as well as the validation of whether large language models can better recognize proper nouns, as mentioned by Radhakrishnan et al. (2023). Additionally, there is a class imbalance in Language Identification; specifically, among the 1,000 test samples in AIRBench, only 2 are in Japanese.

- **Level 2** Only AIR-Bench (Yang et al., 2024) provides support, but it focuses on music rather than speech.

- **Level 3** Tasks related to non-semantic comprehension, such as Emotion, Environment, and Speaker Gender/Age, have already been explored in previous work. However, there are still some gaps, such as Cappella Emotion Recognition, Singing Detection, Emotion Intensity Perception, and Emotion Translation. Additionally, the existing benchmark suffers from issues like class imbalance. For example, in Speaker's Age Prediction in AIR-Bench (Yang et al., 2024), 65.3% of the data falls under the 'teens to twenties' category, or there are issues with a limited number of categories, such as in SD-Eval (Ao et al., 2024), where Speaker's Age Prediction only has two categories. Furthermore, in Speech Emotion Recognition, Speech LLMs may infer emotion from text content, rather than genuinely deriving the emotion from speech itself.

- **Level 4** The previous benchmark did not take this level into account.

- **Level 5** The previous benchmark did not take this level into account.

This underscores the urgent need to build a more comprehensive benchmark that addresses the gaps

Table 7: Existing benchmarks across Levels. L2, L4 and L5 have not received enough attention yet.

Level	Task	Dynamic-SUPERB	AIR-Bench	SD-Eval
L1	Speech ASR	✓	✓	✗
	Intent Classification	✓	✓	✗
	Language Identification	✓	✓	✗
L2	Music Pitch and Velocity	✗	✓	✗
L3	Emotion	✓	✓	✓
	Environment	✓	✓	✓
	Accent	✓	✗	✓
	Speaker Gender/Age	✗	✓	✓
	Noise Detection	✓	✗	✗
	Speaker Verification	✓	✓	✗
	Sarcasm Detection	✓	✗	✗
	Stress Detection	✓	✗	✗
	How Far Are You	✓	✗	✗
	Spoof Detection	✓	✗	✗
	Synthesized Voice Detection	✗	✓	✗
L4	No Related Work	✗	✗	✗
L5	No Related Work	✗	✗	✗

in Level 2, Level 4, and Level 5, Modify and expand some tasks of Level 1 and Level 3. Ensuring more robust evaluation across all levels of speech model tasks. Additionally, the existing benchmark still has two shortcomings, which we list below:

- The existing benchmark lacks exploration of noise data interference and interference from other languages, so we will also consider these issues.
- Dynamic-SUPERB (Huang et al., 2024) uses an excessive number of evaluation metrics, making it difficult to simultaneously reflect the performance of Speech LLMs across multiple tasks.

A.2 Our Contribution

From a macro perspective, our DitingBench fills the gaps in the previous benchmarks at the second, fourth, and fifth levels, while also providing more supplements and adjustments to the tasks at the first and third levels. We have adjusted the class imbalance and addressed the potential for model shortcuts in the data.

At the same time, we discussed the situation of noise interference and used only three evaluation metrics to facilitate the comparison of performance across multiple tasks (specific details are provided in App. C.3).

From a more detailed perspective, we have outlined the differences and contributions of each of our tasks compared to previous work in Tab. 8. Additionally, we discuss the impact of noise on Speech LLM performance in App H.

B Details of Benchmark Construction

The overall construction principles are provided in Sec. B.1. The data and tools used are detailed in Sec. B.2. The composition structure of the data is outlined in Sec. B.3. Detailed construction details for each task are available in Sec. B.4.The credibility verification of synthesized speech is provided in Sec. B.5.3.

B.1 General Principles of Data Construction

B.1.1 Question Construction

For objective multiple-choice questions, we guide large models by including multiple-choice options within the questions to facilitate the generation of final results. For subjective response questions, we specified the main aspects around which the questions revolve and set suggested answers, although these do not require the model to produce results that are exactly identical, illustrated in Fig. 4.

B.1.2 Uniform Sampling Rate

Considering the potential introduction of extraneous factors due to varying sampling rates of audio data, this paper standardizes all datasets to the one with the lowest sampling rate. Consequently, all test data is downsampled to 16,000 Hz.

B.1.3 Uniform number of audio channels

To standardize the format of the input audio, we converted all audio files for the tasks into mono channel, except for those in the Binaural Effect Perception task.

Table 8: Related work for each task.

Level	Task	Dataset
L1	Language Identification	Compared to Dynamic-SUPERB and AIRBench, we have made the classes balanced.
	Auto-Speech Recognition	Lack of consideration for noisy interference.
	ASR for Legal Terms*	Our proposed in Speech LLM evaluation.
	ASR for Medical Terms*	Our proposed in Speech LLM evaluation.
	Auto-Lyrics Transcription	Our proposed in Speech LLM evaluation.
L2	Volume Perception	Our proposed in Speech LLM evaluation.
	Pitch Perception	Our proposed in Speech LLM evaluation.
	Binaural Effect Perception	Our proposed in Speech LLM evaluation.
	Loudness Assessment	Our proposed in Speech LLM evaluation.
	Speech Rate Assessment	Our proposed in Speech LLM evaluation.
	Speech Pause Detection	Our proposed in Speech LLM evaluation.
L3	Ambient Sound Detection	Our proposed in Speech LLM evaluation.
	Acoustic Scene Classification	Expanded the scene categories of other benchmark designs.
	Speaker's Age Prediction	Compared to AIRBench, we have made the classes balanced.
	Speaker's Gender Recognition	Compared to AIRBench, we have made the classes balanced.
	Speech Emotion Recognition	Eliminate the possibility of inferring results from the text.
	Cappella Emotion Recognition	Our proposed in Speech LLM evaluation.
	Emotional Intensity Perception	Our proposed in Speech LLM evaluation.
	Emotion Translation*	Our proposed in Speech LLM evaluation.
	Singing Detection	Our proposed in Speech LLM evaluation.
L4	COVID-19 Risk Detection	Our proposed in Speech LLM evaluation.
	Cough Type Classification	Our proposed in Speech LLM evaluation.
	Cough Origin Diagnosis	Our proposed in Speech LLM evaluation.
	Cough Severity Assessment	Our proposed in Speech LLM evaluation.
	Lung Risk Screening	Our proposed in Speech LLM evaluation.
L5	Spoken English Coach	Our proposed in Speech LLM evaluation.
	Voice Detective	Our proposed in Speech LLM evaluation.

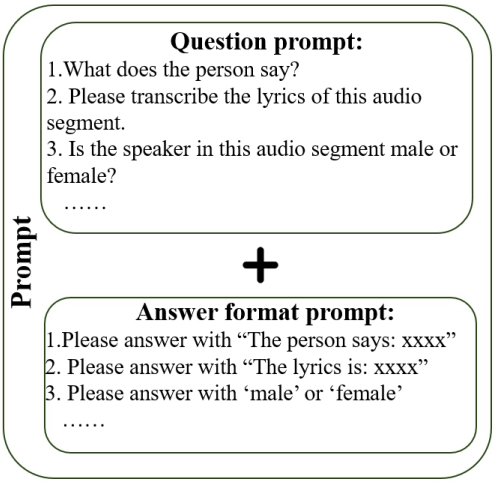

Figure 4: The method to generate text instructions for the problems.

B.1.4 Uniform Audio Duration

Most Speech LLMs (Chu et al., 2023, 2024; Liu et al., 2024a; Tang et al., 2023) utilize the encoder from (Radford et al., 2023), which limits their maximum audio processing duration to 30 seconds. To ensure fairness, we have restricted the lengths of the audio inputs to a maximum of 30 seconds.

B.1.5 Uniform Option Ratio

For the multiclass classification problem, we performed data balancing. Taking binary classification tasks as an example, due to some limitations in the current models, they might always choose one option in binary classification tasks. If the data were unbalanced, such as 40% for one option and 60% for the other, different models that always pick the same option could yield very different results, even though their capabilities are similar. This is not what we want, so we balanced the data for all multiclass classification tasks. Please refer to Tab. 19 for detailed information.

B.2 Datasets and Tools Utilized

We used the following 10 datasets. The licenses for these datasets can be found in Tab. 10. It should be noted that the JamendoLyrics MultiLang dataset and TED-LIUM dataset includes some ND-restricted data, so we won't provide the related test entries in the open-source release, but will offer the corresponding data processing scripts instead.

These datasets include:

Europarl-ST (Iranzo-Sánchez et al., 2020)

Table 9: Overview of the levels and the corresponding tasks.

Level	Task	Dataset	Utterances
L1	Language Identification	Europarl-ST (Iranzo-Sánchez et al., 2020)	2505
	Auto-Speech Recognition	LibriSpeech (Panayotov et al., 2015)	2791
	ASR for Legal Terms*	Made of CosyVoice (SpeechTeam, 2024)	102
	ASR for Medical Terms*	Made of CosyVoice (SpeechTeam, 2024)	203
	Auto-Lyrics Transcription	Jam-Lyrics (Durand et al., 2023)	868
L2	Volume Perception	Made of LJSpeech (Ito and Johnson, 2017)	1024
	Pitch Perception	Made of SpeechAccentArchive (Weinberger, 2013)	600
	Binaural Effect Perception	Our proposed method	600
	Loudness Assessment	Made of TextrolSpeech (Ji et al., 2024)	1000
	Speech Rate Assessment	Made of TextrolSpeech (Ji et al., 2024)	1000
	Speech Pause Detection	Made of TED-LIUM (Hernandez et al., 2018)	1000
L3	Ambient Sound Detection	Noisy speech (Valentini-Botinhao et al., 2017)	1648
	Acoustic Scene Classification	Made of MS-SNSD (Reddy et al., 2019)	2775
	Speaker's Age Prediction	Made of AIR-Bench (Yang et al., 2024) & SpeechAccentArchive (Weinberger, 2013)	990
	Speaker's Gender Recognition	Made of VCTK (Yamagishi et al., 2019)	2820
	Speech Emotion Recognition	Selected from RAVDESS (Livingstone and Russo, 2018)	1300
	Cappella Emotion Recognition	Selected from RAVDESS (Livingstone and Russo, 2018)	828
	Emotional Intensity Perception	Made of RAVDESS (Livingstone and Russo, 2018)	286
	Emotion Translation*	Made of RAVDESS (Livingstone and Russo, 2018) & CosyVoice (SpeechTeam, 2024)	325
	Singing Detection	Made of RAVDESS (Livingstone and Russo, 2018)	2024
L4	COVID-19 Risk Detection	Made of Virufy (Chaudhari et al., 2020)	120
	Cough Type Classification	Made of COUGHVID (Orlandic et al., 2021)	600
	Cough Origin Diagnosis	Made of COUGHVID (Orlandic et al., 2021)	798
	Cough Severity Assessment	Made of COUGHVID (Orlandic et al., 2021)	510
	Lung Risk Screening	Made of Lung Sound (Fraiwan et al., 2021)	210
L5	Spoken English Coach	Made of speechocean762 (Zhang et al., 2021)	1009
	Voice Detective	Made of SpeechAccentArchive (Weinberger, 2013)	2134

"*" denotes that utterances are synthesized, and the credibility verification is provided in Appendix B.5.3. The specific proportion of each category in these classification tasks can be found in Table 19.

, LibriSpeech (Panayotov et al., 2015), JamendoLyrics MultiLang dataset (Durand et al., 2023), LJSpeech (Ito and Johnson, 2017), Noisy speech (Valentini-Botinhao et al., 2017), SpeechAccentArchive (Weinberger, 2013) , VCTK (Yamagishi et al., 2019), RAVDESS((Livingstone and Russo, 2018), AISHELL-MDSC (Gao et al., 2024), speechocean762 (Zhang et al., 2021), TED-LIUM (Hernandez et al., 2018), TextrolSpeech (Ji et al., 2024), Lung Sounds (Fraiwan et al., 2021).

We utilized two open-source tools:

cosyVoice (SpeechTeam, 2024) and MS-SNSD (Reddy et al., 2019).

B.2.1 Datasets and their License Types

The licenses for these datasets can be found in Tab. 10

Table 10: Datasets and their License Types

Dataset	License Type
Europarl-ST	CC BY-NC 4.0
LibriSpeech	CC BY 4.0
JamendoLyrics MultiLang	CC BY-NC-SA-ND
LJSpeech	Public Domain
Noisy speech	CC BY 4.0
SpeechAccentArchive	CC BY-NC-SA 2.0
VCTK	CC BY 4.0
RAVDESS	CC BY-NC-SA 4.0
AISHELL-MDSC	CC BY-NC-SA 4.0
speechocean762	CC BY 4.0
TED-LIUM	CC BY-NC-ND 3.0
TextrolSpeech	CC BY-NC 4.0
Lung Sounds	CC BY 4.0

B.3 Data Structure of Benchmark

Data samples are represented as (P, Q, A, D), where P denotes the audio path, Q represents the question, A corresponds to the answer, and D provides additional explanations to aid researchers in understanding the data.

B.4 Details of Each Task

B.4.1 Language Identification

We used Europarl-ST (Iranzo-Sánchez et al., 2020) to construct our evaluation dataset. Europarl-ST is a multilingual speech translation corpus containing paired audio-text samples for speech translation. It was constructed using debates held in the European Parliament between 2008 and 2012. We selected five commonly used languages: German, English, French, Spanish, and Italian. The task was set as: "What language is spoken in this audio segment? Please choose from the German, English, French, Spanish and Italian options."

B.4.2 Automatic Speech Recognition

We constructed our evaluation dataset based on LibriSpeech (Panayotov et al., 2015). Inspired by (Radford et al., 2023), we used the test-clean and test-other splits as our test sets, comprising a total of 2791 data entries. Since we addressed specific aspects within our metric C.3.1, we did not perform any additional processing when constructing the dataset. The task was set as: "What does the person say? Please answer with 'The person says: xxxx'."

B.4.3 ASR for Legal Terms

We selected 27 offenses defined under Chinese criminal law and combined them with four templates to generate 108 sentences, which were synthesized using cosyVoice (SpeechTeam, 2024). After manual screening (detailed in Sec. B.5.4), 102 utterances remained. The task was set as: "What does the person say? Please answer with 'The person says: xxxx'." This approach is consistent with ASR, as we believe that this ability should be demonstrated automatically during the ASR process without the need for additional prompts.

B.4.4 ASR for Medical Terms

We selected 62 medical terms referring to specific locations and combined them with four templates to generate 248 sentences, which were synthesized using cosyVoice (SpeechTeam, 2024). After manual screening (detailed in Sec. B.5.4), 203 utterances remained. The task was set as: "What does the person say? Please answer with 'The person says: xxxx'." This approach is consistent with ASR, as we believe that this ability should be demonstrated automatically during the ASR process without the need for additional prompts.

B.4.5 Automatic Lyrics Transcription

We utilized the JamendoLyrics MultiLang dataset (Durand et al., 2023) for our research. We acknowledge that a revised version of this dataset has been released as the Jam-Alt dataset (Cífka et al., 2023). However, in accordance with the constraints outlined in Sec. B.1.4, we were required to resegment the audio files. Given that the Jam-Alt dataset, as described by its authors, exhibits certain deviations in its timestamps, we elected to employ the JamendoLyrics MultiLang dataset as our primary dataset for construction purposes. During the construction process, we manually selected the segmentation points and employed code to segment the audio files, thereby obtaining our final dataset.The task was set as: "Please transcribe the lyrics of this audio segment.Please answer with: 'The lyrics is: xxxx'."

B.4.6 Volume Perception

We constructed our evaluation dataset based on LJSpeech (Ito and Johnson, 2017). Following the data split of (Chien et al., 2021), we used 512 test samples. We set up two scenarios: one where the volume gradually increases from 0 to its original level, and another where it decreases from the original level to 0. We tasked the model with determining whether the volume is increasing or decreasing. The task was set as: "Is the volume of this audio segment gradually increasing or decreasing?"

B.4.7 Pitch Perception

We used the SpeechAccentArchive (Weinberger, 2013) dataset to construct our test set. During this process, we first identified the frequency ranges with the highest proportion of fundamental frequency (F0). Ultimately, we selected the ranges (80, 150) Hz and (180, 250) Hz for our experiments. We framed the problem as follows: "In the following audio segment, into which range does more than 70% of the fundamental frequency content fall? Please choose from the following two ranges: (80, 150) Hz and (180, 250) Hz." We calculated the proportion of F0 content falling within these two ranges for each audio segment and selected the corresponding data. During the process, we ranked all the data, prioritizing those segments with a higher proportion.

B.4.8 Binaural Effect Perception

We generated random sounds using four methods: sine wave, square wave, triangle wave, and noise.

These sounds are heard only in the left ear or the right ear. For more details, please refer to our public code. The model is used to determine which ear hears these sounds. The task was set as: “In this audio segment, does the sound appear in the left ear or the right ear? Please answer with ‘left’ or ‘right’."

B.4.9 Loudness Assessment

We constructed our data set using TextrolSpeech (Ji et al., 2024). In the original data set, the loudness was classified into three classes: loud, soft, and normal. To make the distinction more pronounced, we selected only the loud and soft categories, ensuring an equal gender ratio during the selection process. We frame the problem as follows: “Please determine whether the following audio clip has a loud or soft sound. Please respond with ’loud’ or ’soft’."

B.4.10 Speech Rate Assessment

We constructed our data set using TextrolSpeech (Ji et al., 2024). In the original data set, the speech rate was classified into three classes: rapid, slow, and normal. To make the distinction more pronounced, we selected only the rapid and slow categories, ensuring an equal gender ratio during the selection process. We frame the problem as follows: “How do you feel about the current pace of the speech? Please respond with ’rapid’ or ’slow’."

B.4.11 Speech Pause Detection

We use Python code to determine the durations of the pause and the number of pauses based on the energy signals of the speech. Since the TED-LIUM dataset (Hernandez et al., 2018) inherently includes pause annotations, it allows for secondary verification, making it an ideal choice for constructing our benchmark dataset. We frame the problem as follows: “Please determine if there are noticeable pauses in this audio. Answer with ’yes’ or ’no.’"

B.4.12 Ambient Noise Detection

We constructed the evaluation dataset using Noisy speech (Valentini-Botinhao et al., 2017). The Noisy speech dataset contains the corresponding pairs of clean and noisy data. The purpose of the data set is to explore methods for speech enhancement.We selected the entire test set from this dataset, which includes 824 clean audio clips and 824 audio clips with ambient noise. We used all of these data and the task was set as: “Is there any ambient noise in this audio segment, in addition to the speaker voice? Please answer with yes or no."

B.4.13 Acoustic Scene Classification

We used MS-SNSD (Reddy et al., 2019) to synthesize these test datasets.MS-SNSD is a tool to synthesize speech with environmental noise, aimed at advancing research in speech enhancement. We selected 51 environmental noise samples from its test set to synthesize 6,105 test samples, and the task was set as: “What is the ambient noise of this audio segment? Please choose from the [’Babble’, ’CopyMachine’, ’Neighbor’, ’ShuttingDoor’, ’AirportAnnouncements’, ’Munching’, ’Typing’, ’AirConditioner’, ’VacuumCleaner’] options?"

B.4.14 Speaker’s Age Prediction

We have observed that there are relatively few datasets specifically aimed at speaker age recognition. We noted that the AIR Bench (Yang et al., 2024) has done an excellent job in addressing this task,We followed their approach of categorizing age into four groups but noticed that their data distribution was not balanced, specifically: teens to twenties: 653, thirties to forties: 268, fifties to sixties: 64, seventies to eighties: 15. Therefore, we used the SpeechAccentArchive (Weinberger, 2013) to balance the age distribution. Unfortunately, we found it difficult to obtain sufficient data for the seventies to eighties category, so we retained only three categories: teens to twenties, thirties to forties, and fifties to sixties. And the task was set as: “Which age range do you believe best matches the speaker’s voice? Please choose from the [‘teens to twenties’, ‘thirties to forties’, ‘fifties to sixties’] options?"

B.4.15 Speaker’s Gender Recognition

We constructed the evaluation dataset using VCTK (Yamagishi et al., 2019).To balance the number of males and females in the benchmark, considering there are 61 female speakers and 47 male speakers in the VCTK dataset, we selected the top 47 female speakers along with all the male speakers. For each speaker, we chose the first 30 audio recordings. The task was set as: “Is the speaker in this audio segment male or female?Please answer with ‘male’ or ‘female’."

B.4.16 Speech Emotion Recognition

In a genuine sense, understanding emotions in models should not solely depend on interpreting text. Emotions do not have a one-to-one correspondence

with sentences; the same sentence can express various emotional tones depending on the speaker's emotional state. Therefore, it is crucial to advocate for models to move beyond mere textual content of sentences when inferring emotions and to delve into the non-textual information within the speech. Accordingly, in the evaluation set for emotion recognition, we employed a dataset unrelated to both the emotions and the sentence content—the RAVDESS dataset (Livingstone and Russo, 2018). The task is then defined as: "What emotion does this audio clip convey? Please answer by single word select from ['neutral', 'happy', 'sad', 'angry', 'fearful', 'disgust', 'surprised']."

To demonstrate that the emotions in our constructed dataset are independent of the textual content, we used a combination of the Whisper-v3-large (Radford et al., 2023) model and the gpt-4-o (OpenAI, 2023) model to predict the emotions in the audio files of the dataset. The experimental results can be found in the Tab. 11

B.4.17 Cappella Emotion Recognition

We also used RAVDESS (Livingstone and Russo, 2018) to construct the evaluation set for singing emotion detection.The task is then defined as: "What emotion does this audio clip convey? Please answer by single word select from ['neutral', 'happy', 'sad', 'angry', 'fearful', 'disgust', 'surprised']."

B.4.18 Emotional Intensity Perception

We used the RAVDESS (Livingstone and Russo, 2018) dataset to construct the evaluation set for Emotional Intensity Perception. Since most models accept only a single audio input, we merged two audio segments and tasked the model with analyzing which part of the combined audio segment exhibits stronger emotional intensity. Specifically, we defined the problem as follows: "In this audio segment, a sentence is repeated twice. Is the emotion in the 'former' stronger or the 'latter' stronger? Please answer with 'former' or 'latter'." To balance the proportion between the two options, we alternated the placement of the stronger emotion, sometimes positioning it at the former and other times at the latter when synthesizing the data.

B.4.19 Emotion Translation

We believe that translations should reflect different expressions based on the emotional context. For example, the phrase "What are you doing?" can convey various meanings depending on the emotion—whether it's anger, surprise, sadness, or neutrality. In an angry context, it expresses strong disapproval or questioning of the person's actions; in a surprised context, it conveys disbelief about what the other person is doing; and in a sad context, it should reflect disappointment. Therefore, translations should be adjusted accordingly to better capture these nuances.

We observed that cosyVoice (SpeechTeam, 2024) demonstrates excellent zero-shot capabilities, effectively mimicking the tone and style of the input speech prompt. Therefore, we used cosyVoice to emulate the sentences with strong emotions from the RAVDESS (Livingstone and Russo, 2018) dataset to generate speech with corresponding emotions. After synthesis, we had five native speakers review the generated speech. If any of the native speakers felt that the synthesized speech did not convey the intended emotion, that segment was discarded. Ultimately, we obtained xxx valid speech samples. The task was set as: "Please translate the following sentence into the most appropriate Chinese, based on the emotion and content of this audio segment."

B.4.20 Singing Detection

We aim for singing detection to go beyond simply identifying background music or relying on lyrics to determine whether singing is occurring. Instead, we seek to differentiate singing from normal speech by recognizing the distinct rhythm and melody of singing. To achieve this, we constructed our singing detection dataset using RAVDESS ((Livingstone and Russo, 2018)), which consists entirely of a cappella performances where the context is unrelated to the singing. The task is then defined as: "Is there singing in this audio clip?Please answer by yes or no.".

B.4.21 COVID-19 Risk Detection

We use the Virufy COVID-19 Open Cough Dataset (Chaudhari et al., 2020) to construct our evaluation set. We classify the samples with positive test results as COVID-19 at risk, while those with negative results are classified as not at risk. And the task was set as: "Please listen to the following cough sound and determine whether the person is at risk of having a COVID-19 infection. Respond with yes or no."

Table 11: emotion detection evaluation set Supplementary experiments

	First repetition	Second repetition	Third repetition
Accuracy	10.53%	9.33%	9.73%

B.4.22 Cough Type Classification

We use the COUGHVID (Orlandic et al., 2021) dataset to construct our evaluation set. We only utilize the data that has been assessed by experts, which falls into two categories: evaluations by four experts and evaluations by one expert. We prioritize samples where three out of four experts agree, and then we use samples rated as "good" by the single expert. In this task, we ask the model to distinguish whether the cough is a wet cough or a dry cough. And the task was set as: “Please help me determine whether the cough in this audio segment is a dry cough or a wet cough. Please respond with ‘wet’ or ‘dry’."

B.4.23 Cough Origin Diagnosis

We use the COUGHVID (Orlandic et al., 2021) dataset to construct our evaluation set. We only utilize the data that has been assessed by experts, which falls into two categories: evaluations by four experts and evaluations by one expert. We prioritize samples where three out of four experts agree, and then we use samples rated as "good" by the single expert. In this task,The origins we tested include‘COVID-19’, ‘healthy cough’, ‘lower infection’, or ‘upper infection’. And the task was set as: “Please help me determine the infection origin of the cough in the following audio segment. Choose from ‘COVID-19’, ‘healthy cough’, ‘lower infection’, or ‘upper infection’."

B.4.24 Cough Severity Assessment

We use the COUGHVID (Orlandic et al., 2021) dataset to construct our evaluation set. We only utilize the data that has been assessed by experts, which falls into two categories: evaluations by four experts and evaluations by one expert. We prioritize samples where three out of four experts agree, and then we use samples rated as "good" by the single expert. In this task, the severity levels we tested include: ‘pseudocough’, ‘mild’, or ‘severe’. And the task was set as: “Please help me assess the severity of the cough in the audio segment. Choose from ‘pseudocough’, ‘mild’, or ‘severe’."

B.4.25 Spoken English Coach

We used speechocean762 (Zhang et al., 2021) to construct our evaluation set.In selecting our evaluation set, we aimed to include a wide variety of pronunciation errors by prioritizing sentences with poorer pronunciation quality. Here is how we built our sentence collection: We started by selecting 207 sentences based on word stress errors (score == 5). Next, we chose 6 sentences with incomplete sentences or error-containing words (score < 10). Then, we selected 332 sentences with poor fluency (score <= 5). Following that, we picked 85 sentences with poor rhythm (score <= 5). Subsequently, we chose 179 sentences with low accuracy (score <= 5). Finally, we selected 40 sentences from each accuracy score level where the scores were higher. This process resulted in a final set of 1009 sentences. When constructing the ground truth for the answer output, we adopted the descriptions used in the original project for dataset scoring, and by concatenating these descriptions, we formed the final answer.

B.4.26 Voice Detective

When constructing the Voice Detective evaluation set, we used the SpeechAccentArchive dataset (Weinberger, 2013). The primary reason for choosing this dataset is the difficulty in obtaining a large amount of similar data, which significantly reduces the risk of data leakage. This constraint also compels researchers to focus more on factors such as the age and background of the users within the dataset.

B.5 Credibility Verification

B.5.1 Instructions Given To Participants

All participants were informed to select data from the current set that did not meet the expected question-answer criteria or lacked natural fluency.

B.5.2 Recruitment And Payment

We provided all participants with a one-time research stipend of 500 RMB, which is more than sufficient for the region they are located in.

B.5.3 Data Consent

We explicitly informed these participants that we would use their evaluation results as a quality check for our data and that their data would not be used for any other purposes.

These participants have authorized us to use their data for research purposes. As the data has not been authorized for public release, we will not make it publicly available.

B.5.4 ASR for Legal Term

Since the legal vocabulary we selected, can be found in open-source code, is not complex, we introduced only one evaluator with a background in legal education, who is a native Mandarin speaker. The remaining three evaluators are regular native Mandarin speakers, making a total of four evaluators. If any one of the evaluators deems the speech quality insufficient, the corresponding speech will be discarded. The specific details of the evaluators are as follows:

Evaluator 1: 24 years old, male, graduated with a bachelor's degree from China University of Political Science and Law and is currently a master student at China University of Political Science and Law. Native Mandarin speaker.

Evaluator 2: 20 years old, female, currently an undergraduate student at Hubei University of Technology. Native Mandarin speaker.

Evaluator 3: 20 years old, female, currently an undergraduate student at Wuchang Shouyi University. Native Mandarin speaker.

Evaluator 4: 26 years old, male, high school graduate. Native Mandarin speaker.

B.5.5 ASR for Legal Medical

Due to the involvement of some medical terminology, this paper selected two evaluators with a medical background, along with two additional evaluators without a medical background. All of them are native Mandarin speakers. Similarly, if any one of the evaluators finds an abnormality in the speech, it will be discarded. The specific details of the evaluators are as follows:

Evaluator 1: 33 years old, female, graduated with a bachelor's degree from Hebei Medical University and has since been working in a medical-related field. Native Mandarin speaker.

Evaluator 2: 26 years old, female, completed an eight-year integrated program (continuously pursued both bachelor's and master's degrees) at Hebei Medical University and continues to work in a medical-related field. Native Mandarin speaker.

Evaluator 3: 25 years old, male, graduated with a bachelor's degree from Beijing Forestry University and is currently a graduate student at Beijing University of Posts and Telecommunications. Native Mandarin speaker.

Evaluator 4: 54 years old, male, graduated from a technical secondary school. Native Mandarin speaker.

B.5.6 Emotion Translation

We selected four evaluators and recorded their English proficiency. Similarly, if any one of the evaluators finds an abnormality in the speech, it will be discarded. The specific details of the evaluators are as follows:

Evaluator 1: 25 years old, female, graduated with a bachelor's degree from China Jiliang University and a master's degree from Beijing University of Posts and Telecommunications. English proficiency: CET-6.

Evaluator 2: 25 years old, female, graduated with both a bachelor's and a master's degree from Beijing University of Posts and Telecommunications. English proficiency: CET-6.

Evaluator 3: 23 years old, male, graduated with a bachelor's degree from Beijing Institute of Technology and is currently a PhD student at The Chinese University of Hong Kong, Shenzhen. English proficiency: IELTS Academic score: 6.5.

Evaluator 4: 28 years old, male, graduated with a bachelor's degree from Beijing University of Posts and Telecommunications and is a PhD student at Beijing University of Posts and Telecommunications. English proficiency: CET-6.

C Experiment Details

Below, we will divide the experiment details into four parts: details of human evaluation in Sec. C.1, details of model evaluation in Sec. C.2, and metric details in Sec. C.3.

C.1 Humans Evaluation Details

In this section, we will introduce the participant information of our humans performance evaluation in Sec. C.1.4 and present the results of the consistency test for the result in Sec. C.1.5.

C.1.1 Instructions Given To Participants

We informed the participants to select the answer they deemed correct based on their first impression

of the question displayed on the webpage and the audio they heard.

To clearly demonstrate our testing process, we present the details of our web-based testing in Fig 5. It should be noted that each webpage contains only one type of task for testing, and I am using Loudness Assessment as an example here.

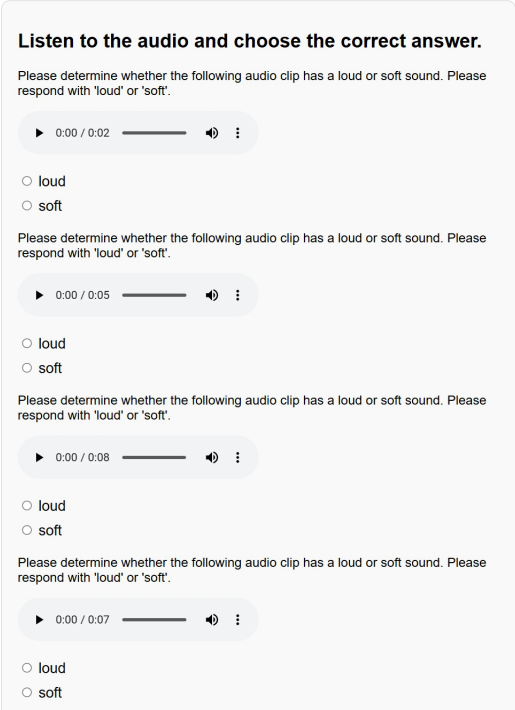

Figure 5: Website for human testing.

C.1.2 Recruitment And Payment

We provided all participants with a research stipend of 1000 RMB, which is more than sufficient for the region they are located in.

C.1.3 Data Consent

We have explicitly informed all participants that their evaluation data will be used in our research, including but not limited to reflecting the accuracy of human performance on our benchmark.

All participants have given informed consent, acknowledging that their data may be used for academic research purposes and potentially published. However, this data will only be used for academic research purposes.

C.1.4 Participant Information

Evaluator 1: Female, 28 years old, graduated with a bachelor's degree from East China Normal University, PhD from the Institute of Physics CAS.

Evaluator 2: Female, 26 years old, graduated with a bachelor's degree from Beijing Normal University, master's degree from Shanghai Jiao Tong University.

Evaluator 3: Male, 29 years old, graduated with a bachelor's degree from Beijing University of Chemical Technology, PhD from Beijing University of Posts and Telecommunications.

Evaluator 4: Male, 27 years old, graduated with a bachelor's degree from Xidian University, currently pursuing a PhD at Singapore University of Technology and Design.

C.1.5 Consistency Test

To verify the consistency of the humans evaluation, We focus on objective multiple-choice questions. we calculated the proportion of questions where all three volunteers selected the same option, as well as the proportion where all four volunteers chose the same option, relative to the total number of questions. These proportions are shown in Tab. 12.

It is also important to note that, since our testers are only proficient in English, they were unable to complete the Language Identification task.

C.1.6 Deficiency in Humans Evaluation.

During the Humans Evaluation process, we were unable to find a native English speaker, but all participants involved in the evaluation are proficient English users. We also could not find individuals who are proficient in multiple languages, which made it difficult to conduct a Humans Evaluation for the Language Identification task.

C.2 Models Evaluation Details

We divide our experimental details into two sections: the model replication platform in Sec. C.2.1, and the model replication details in Sec. C.2.2.

C.2.1 Experimental Platform

In this paper's experiments, all servers used are equipped with an Intel® Xeon® Platinum 8358 CPU @ 2.60GHz as the core processor. Each server is loaded with eight NVIDIA A800-SXM4-80GB graphics cards, and each model runs with exclusive use of one A800 card.

C.2.2 Models Replication Details

In this paper, we aim to select the 7B-level versions of various models wherever possible. However, due to the differences between various models, it is difficult to ensure that their parameter counts are exactly the same.

Table 12: Consistency for Humans Evaluation

Task	Accuracy	Num of Questions	Proportion (3 Evaluators Same)	Proportion (4 Evaluators Same)
Volume Perception	100.00%	160	100.00%	100.00%
Pitch Perception	96.25%	160	100.00%	95.00%
Binaural Effect Perception	100.00%	160	100.00%	100.00%
Ambient Noise Detection	91.88%	160	100.00%	87.50%
Acoustic Scene Classification	90.28%	720	97.22%	93.89%
Speaker's Age Prediction	52.59%	240	76.67%	46.67%
Speaker's Gender Recognition	97.50%	160	100.00%	100.00%
Speech Emotion Recognition	50.71%	560	94.29%	85.71%
Cappella Emotion Recognition	62.25%	400	92.00%	68.00%
Emotion Intensity Perception	97.50%	160	100.00%	95.00%
Singing Detection	98.13%	160	100.00%	97.50%
COVID-19 Risk Detection	60.63%	160	70.00%	17.50%
Cough Type Classification	52.50%	160	77.50%	22.50%
Cough Origin Diagnosis	32.19%	320	28.75%	2.50%
Cough Severity Assessment	45.42%	240	45.00%	11.67%

Table 13: Gap Between Professionals and Non-Professionals

Task	Non-Professionals Accuracy	Professionals Accuracy
COVID-19 Risk Detection	60.63%	c
Cough Type Classification	52.50%	d
Cough Origin Diagnosis	32.19%	a
Cough Severity Assessment	45.42%	b

GPT-4o For the GPT-4o model, we reproduced the model by calling its API.

Mu-LLaMA In the process of implementing the model Mu-LLaMA (Liu et al., 2024a) , this paper used the LLama2-7B-chat (Touvron et al., 2023) checkpoint to maintain consistency with the original paper, and utilized the open-source MU-LLaMA checkpoint provided.

GAMA Since the primary focus of this paper is to test the audio understanding capabilities of the GAMA model (Ghosh et al., 2024b), we consulted with the authors and selected the 'state4epoch2' checkpoint over the 'state5epoch2' checkpoint, as it has superior audio comprehension abilities

SALMONN For the SALMONN model (Tang et al., 2023), we tested the model using its open-source code.

Qwen2-Audio For the Qwen2-Audio model (Chu et al., 2024), we reproduced the model using the 7B version of its open-source code.

C.3 Matrix

We have designed three metrics: WER, the accuracy for objective multiple-choice questions, and GPT-4o scoring, specifically targeting ASR tasks, objective multiple-choice questions, and subjective responses. This section will provide detailed explanations. For an overview, please refer to the following Tab. 14.

C.3.1 WER for ASR

The Word Error Rate (WER), a key metric for gauging the effectiveness of Automatic Speech Recognition (ASR) systems, quantifies the divergence between an ASR system's output and a reference transcript. It assesses the total error rate by tallying the number of insertion, deletion, and substitution operations needed to align the ASR output with the true reference text.

While computing the WER, certain variances in word usage, like "I am" compared to "I'm," may be seen as semantically equivalent by human standards but are flagged as errors by computational algorithms. Thus, a standardization process is essential prior to WER calculation to make both texts directly comparable. The methodology for this standardization, akin to what is employed in the Whisper (Radford et al., 2023) framework, has been detailed in a related research paper. It has been demonstrated that this approach exerts negligible influence on the assessment of WER outcomes when tested against the LibriSpeech (Panayotov et al., 2015) dataset, which was utilized in our paper.

For cases where the error rate exceeds 100% (i.e.,

Table 14: Metrics for Each task

Task	Metric
Language Identification	5-Categories Acc
Speech ASR	WER
Song ASR	WER
Volume Perception	2-Categories Acc
Binaural Effect Perception	2-Categories Acc
Ambient Noise Detection	2-Categories Acc
Speaker's Age	3-Categories Acc
Speaker's Gender	2-Categories Acc
Sound Event Classification	9-Categories Acc
Singing Detection	2-Categories Acc
Speech Emotion Recognition	7-Categories Acc
Song Emotion Recognition	5-Categories Acc
Emotion Intensity Perception	2-Categories Acc
Disorder Detection	2-Categories Acc
Speech Disorders Detection	2-Categories Acc
COVID-19 Risk Detection	2-Categories Acc
ALS Detection	2-Categories Acc
Accent Detection	11-Categories Acc
Emotion Translation	GPT Score
Spoken English Coach	GPT Score
Voice Detective	GPT Score

WER is over 1), we mark them in our experimental records as having significant recognition errors. Such data will not be included in the calculation of the final average WER. In the final record of the experiment, we will focus on two key metrics: first, the ASR completion rate, which is the percentage of data with a WER less than 1; second, the mean WER of the completed portion, which is the average WER of data with a WER less than 1. If the mean WER of the completed portion does not decrease to below 0.8, we will conclude that the model lacks effective automatic speech recognition (ASR) capabilities and document this finding in detail in the experimental results.

The implementation details regarding WER (Word Error Rate) can be found in our publicly available code.

C.3.2 Accuracy for objective multiple-choice questions

A selection is considered correct only if the model chooses the correct answer and no other options. If the model selects two or more options, even if the correct one is included, it will be deemed incorrect.

C.3.3 Accuracy for ASR on Terms

Since in these tasks we primarily assess the ability of Speech LLMs to transcribe terms, we consider a response correct as long as the correct term is included in the speech transcription, without focusing on the accuracy of other parts of the sentence.

C.3.4 Scoring for Subjective Response Questions

In our experiments, we used GPT-4o to assist in evaluating the results. The specific prompt used is as follows.

Prompt for Emotion Translation

I currently need your assistance in evaluating some translations. The most suitable translations should incorporate the corresponding emotions appropriately. The scoring ranges from 0 to 4. I will provide you with the original English sentence, the associated emotional label, and the suggested translation, allowing you to score them based on the context.

Here are some examples:

[Here are some scoring examples. Due to space limitations, we have omitted them in this section. You can find the details in the code we have made available.]

Now Answer:[ANSWER]

Label:The original sentence is: <emotion>[SENTENCE] The suggested translation is: [SUGGESTION].

Please provide your score. **Prompt for Spoken English Coach**

I now need you to help me evaluate some Answers for accuracy. You need to evaluate and score in the order of overall pronunciation, fluency, prosody, words that are mispronounced, and words that have incorrect stress. The score ranges from 0 to 4. Here are the specific scoring rules: You need to first check if the evaluation of overall pronunciation in the Answer matches the Label. If they do not match, give a score of 0 and continue with the evaluation; if there is no relevant description, also give a score of 0 and continue with the evaluation; if it is correct, add 1 point and continue with the evaluation.

For fluency and prosody in the Answer compared to the Label, award up to 1 point for each if completely correct, a partial score for partially correct, and no points if there is no relevant expression. Finally, check the descriptions in the Answer and Label regarding words that are mispronounced and words that have incorrect stress. Award 1 point

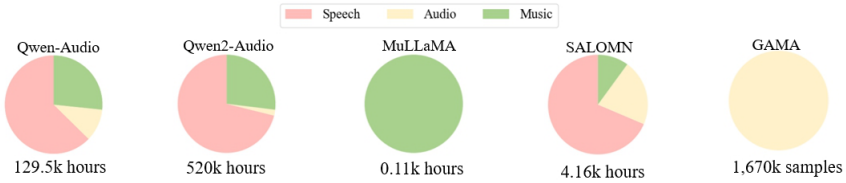

Figure 6: Distribution of three types of training data used by various models

only if all are correct. If part of the descriptions are correct, you can give a partial score, such as 0.33 points for one out of three correct descriptions. Here are some examples:

[Here are some scoring examples. Due to space limitations, we have omitted them in this section. You can find the details in the code we have made available.]

Now Answer:[ANSWER]

Label:[LABEL]

Please provide your score.

C.3.5 Prompt for Voice Detective

I now need you to help me evaluate some Answers for accuracy. You should focus on whether the information about gender, place of birth, age, and native language in the Answer matches the Label, and provide a final rating. Award 1 point for each correct piece of information, with no points for incorrect information. Please give your score on a scale of 0 to 4. Here are some examples:

[Here are some scoring examples. Due to space limitations, we have omitted them in this section. You can find the details in the code we have made available.]

Now Answer:[ANSWER]

Label:[LABEL]

Please provide your score.

C.4 Speech Instruction

When adopting the speech instruction, we use Google Translate's text-to-speech tool to convert the text instruction into speech, which is then merged with the original audio segment and fed into the Speech LLMs.

D Now Speech LLM Limited Types of Training Data

We observed in Tab. 2 that certain tasks, particularly those in Level 2, are easy for humans but challenging for Speech LLMs. We first analyzed the composition of the training data for Speech LLMs, as shown in Fig. 6. We found that most Speech LLMs tend to disregard audio data except

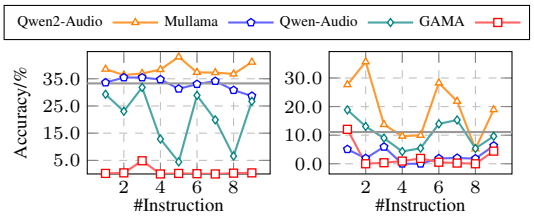

Figure 7: Performance of speech LLMs with different instructions on speaker age task (left) and scene classification task (right). Gray line shows random selection accuracy. Details about the instructions and results are shown in App. F.

for GAMA, whereas GAMA focuses primarily on audio. This indicates distinct data biases among different Speech LLMs, leading to variations in task preferences.

To further examine the influence of task preference, we compared the performance of various Speech LLMs with Whisper V3 (trained with ∼5,000k hours), as shown in Tab. 4. We found that Whisper still outperforms other models on the Lyrics Transcription task due to its the massive training data. On the other hand, with the help of the learned knowledge, Speech LLMs perform significantly better at recognizing certain terms. This demonstrates that Speech LLMs have great potential compared to traditional speech models. Notably, we also tested a *Small* model trained exclusively on an audio dataset. This *Small* model achieved 100% accuracy, while Speech LLMs struggled with the task.

Take-away 7. ***Current insufficient diversity and completeness of training data could not help Speech LLMs reach a higher level.***

E Now Speech LLM Inadequate Instruction Following

We observed that some models exhibit poor instruction following in Tab. 2. Two reasons can lead to these results: 1) the models do not understand the instructions, and 2) the instruction fails to help the models comprehend the speech.

We classify the cause by observing changes in

performance after perturbing the prompt. If the model is insensitive to different perturbed prompts, it indicates that the model cannot understand the prompt. On the other hand, if the models show significantly better performance with a properly structured prompt, it suggests that the model could understand the task, while requires the specific instruction. We choose the two Level 3 tasks (Age prediction and Ambient Noise Detection) where the instruction following ability is crucial, and the results shown in Fig. 7.

For the result of Fig. 7, we can find the Mullama is not sensitive about the instruction. This prove the model can not figure out this task. Further, the performance of most Speech LLMs highly related with the specific prompt, this shows models are sensitive with the format of instruction. Comparing with the text LLMs which are robust with diverse instruction, the Speech LLMs need much effect to guarantee instruction following.

Take-away 8. ***Current Speech LLMs follow instructions poorly.***

F Instruction Follow Experiment Prompt

F.1 Speaker's Age Prediction

The instructions used in the experiment are as follows:

- **Instruction variation I** In which age group do you think the speaker's voice belongs?
- **Instruction variation II** What age category do you believe the speaker's voice fits into best?
- **Instruction variation III** Which age bracket do you feel corresponds to the speaker's voice?
- **Instruction variation IV** How old do you think the speaker sounds, based on their voice?
- **Instruction variation V** Which age range would you assign to the speaker's voice?
- **Instruction variation VI** What age range do you associate with the speaker's voice?
- **Instruction variation VII** Which age group do you think best describes the speaker's vocal characteristics?
- **Instruction variation VIII** What do you believe is the age range of the speaker judging by their voice?

The experimental results are recorded in Tab. 15.

F.2 Acoustic Scene Classification

- **Instruction variation I** How would you detect the background sound in this audio clip?
- **Instruction variation II** What kind of ambient noise can be heard in this segment?
- **Instruction variation III** Can you describe the environmental sounds present in this audio?
- **Instruction variation IV** What background audio elements are featured in this segment?
- **Instruction variation V** What atmosphere is created by the sounds in this audio segment?
- **Instruction variation VI** Can you identify the ambient sound in this clip?
- **Instruction variation VII** What noises are occurring in the background of this audio?
- **Instruction variation VIII** What type of surrounding sound is present in this recording?

The experimental results are recorded in Tab. 16.

G Phonemic Processing Ability Experiment

The specific details of the experimental tasks are presented in Tab. 17.

The citations for the models used in this experiment are as follows: Mixture (Shen et al., 2023), Qwen2 (Bai et al., 2023), Llamma (Dubey et al., 2024)

H Experiment on the Impact of Noise.

H.1 Experimental Settings

In the experiment, we selected the original data and added MS-SNSD data to it according to a random signal-to-noise ratio, while keeping everything else unchanged. All tests were conducted using the same testing method.

H.2 Result

Our results are presented in Tab. 18. As can be seen, after adding noise interference, the overall performance deteriorated, especially for tasks closely related to speech transcription in Level 1.

Table 15: The impact of different prompts on age detection

Prompt	Qwen-Audio	Qwen2-Audio	MuLLama	GAMA
Our benchmark instruction	29.29%	38.55%	33.60%	0.2%
Instruction variation I	23.03%	36.36%	35.45%	0.4%
Instruction variation II	31.82%	36.97%	35.45%	4.85%
Instruction variation III	12.83%	38.38%	34.75%	0.0%
Instruction variation IV	4.44%	43.03%	31.31%	0.2%
Instruction variation V	28.89%	37.37%	33.03%	0.1%
Instruction variation VI	19.90%	37.27%	34.14%	0.0%
Instruction variation VII	6.57%	36.77%	30.81%	0.3%
Instruction variation VIII	26.77%	41.11%	28.67%	0.4%

Table 16: The impact of different prompts on acoustic scene classification

Prompt	Qwen-Audio	Qwen2-Audio	MuLLama	GAMA
Our benchmark instruction	18.84%	27.67%	5.07%	12.05%
Instruction variation I	13.05%	35.68%	1.91%	0.00%
Instruction variation II	8.97%	13.73%	5.91%	0.36%
Instruction variation III	4.29%	9.66%	0.00%	0.94%
Instruction variation IV	5.43%	9.95%	0.00%	1.87%
Instruction variation V	13.95%	28.29%	1.87%	0.54%
Instruction variation VI	15.32%	21.87%	2.02%	0.25%
Instruction variation VII	5.37%	5.23%	1.8%	0.00%
Instruction variation VIII	9.62%	18.92%	6.31%	4.32%

Table 17: Three tasks for assessing phonemic processing ability

Task	Prompt
Sequence-level	Given a phone sequence, "M AA0 R K IH0 Z ...", what sentence does it represent?
Token-level	Given a tokenized phone sequence, "[M AA0 R K] [IH0 Z] ...", what sentence does it represent?
Token-level with one shot	Given a tokenized phone sequence, "[M AA0 R K] [IH0 Z] ...", what sentence does it represent? For example, if the phone sequence is "[F AO0 R] [F AY0 V], [S IH0 K S] [S EH1 V N] [EY0 T]" the sentence can be: "four five six seven eight nine".

Table 18: Experimental results with noise interference added.

	Qwen2-Audio	MuLLaMA	GAMA
Language Identification	96.44%	8.48%	×
Language Identification(sound)	90.55%	×	×
Auto-Speech Recognition	4.63	×	×
Auto-Speech Recognition(sound)	48.74	×	×
ASR for Legal Terms	81.04%	×	×
ASR for Legal Terms(sound)	41.18%	×	×
ASR for Medical Terms	53.86%	×	×
ASR for Medical Terms(sound)	26.6%	×	×
Auto-Lyrics Transcription	32.48	×	×
Auto-Lyrics Transcription(sound)	34.37	×	×
Volume Perception	48.96%	50.00%	11.98%
Volume Perception(sound)	50.02%	50.00%	×
Pitch Perception	50.00%	33.78%	41.50%
Pitch Perception(sound)	50.01%	14.00%	41.67%
Binaural Effect Perception	×	×	×
Binaural Effect Perception_sound	×	×	×
Loudness Assessment	50.13%	49.77%	×
Loudness Assessment(sound)	50.10%	49.80%	×
Speech Rate Assessment	44.93%	50.00%	×
Speech Rate Assessment(sound)	40.00%	50.00%	×
Speech Pause Detection	51.70%	50.00%	49.97%
Speech Pause Detection(sound)	55.00%	50.00%	50.00%
Speaker's Age Prediction	38.55%	33.60%	×
Speaker's Age Prediction(sound)	36.46%	33.94%	×
Speaker's Gender Recognition	79.60%	50.00%	×
Speaker's Gender Recognition(sound)	76.13%	50.00%	×
Speech Emotion Recognition	79.51%	9.20%	3.68%
Speech Emotion Recognition(sound)	29.08%	8.85%	7.31%
Cappella Emotion Recognition	62.38%	12.42%	7.08%
Cappella Emotion Recognition(sound)	34.30%	11.96%	10.63%
Emotion Intensity Perception	50.00%	50.00%	50.00%
Emotion Intensity Perception(sound)	50.00%	50.00%	×
Singing Detection	50.22%	50.00%	64.82%
Singing Detection(sound)	56.18%	50.00%	59.88%

Table 19: Utterances for Each Task

Task	**Utterances**
Language Identification	German: 501, Spanish: 501, English: 501, French: 501, Italian: 501
Auto-Speech Recognition	English:2791
ASR for Legal Terms	Chinese:102
ASR for Medical Terms	Chinese:203
Auto-Lyrics Transcription	English: 868
Volume Perception	Increasing: 512, Decreasing: 512
Pitch Perception	(80-150)Hz: 300, (180-250)Hz: 300
Binaural Effect Perception	Left ear: 400, Right ear: 400
Loudness Assessment	loud: 500, soft: 500
Speech Rate Assessment	rapid: 500, slow: 500
Speech Pause Detection	Yes: 500, No: 500
Ambient Noise Detection	Yes: 824, No: 824
Acoustic Scene Classification	Babble: 310, Copy Machine: 310, Neighbor: 310, Shutting Door: 315, Airport Announcements: 305, Munching: 300, Typing: 310, Air-Conditioner: 305, Vacuum Cleaner: 310
Speaker's Age	Teens to Twenties: 330, Thirties to Forties: 330, Fifties to Sixties: 330
Speaker's Gender	Female: 1410, Male: 1410
Speech Emotion Recognition	Happy: 200, Disgust: 200, Fearful: 200, Sad: 200, Surprised: 200, Angry: 200, Neutral: 100
Cappella Emotion Recognition	Angry: 184, Sad: 184, Happy: 184, Fearful: 184, Neutral: 92
Emotion Intensity Perception	Former: 143, Latter: 143
Emotion Translation	English: 325
Singing Detection	Singing: 1012, Speech: 1012
COVID-19 Risk Detection	Yes:56, No:64
Cough Type Classification	Wet: 300 , Dry: 300
Cough Origin Diagnosis	COVID-19: 198, Healthy Cough: 200, Lower Infection: 200,Upper Infection: 200
Cough Severity Assessment	Pseudocough: 170, Mild: 170, Severe: 170
Spoken English Coach	English: 1009
Voice Detective	English: 2134