Benchmarking Open-ended Audio Dialogue Understanding for Large Audio-Language Models

Kuofeng Gao[1], **Shu-Tao Xia**[1,3†], **Ke Xu**[1], **Philip Torr**[2], **Jindong Gu**[2†]
[1]Tsinghua University, [2] University of Oxford, [3] Peng Cheng Laboratory
gkf21@mails.tsinghua.edu.cn, xiast@sz.tsinghua.edu.cn
xuke@tsinghua.edu.cn, {philip.torr,jindong.gu}@eng.ox.ac.uk

Abstract

Large Audio-Language Models (LALMs), such as GPT-4o, have recently unlocked audio dialogue capabilities, enabling direct spoken exchanges with humans. The potential of LALMs broadens their applicability across a wide range of practical scenarios supported by audio dialogues. However, given these advancements, a comprehensive benchmark to evaluate the performance of LALMs in the open-ended audio dialogue understanding remains absent currently. To address this gap, we propose an **A**udio **D**ialogue **U**nderstanding **Bench**mark **(ADU-Bench)**, which consists of 4 benchmark datasets. They assess the open-ended audio dialogue ability for LALMs in 3 general scenarios, 12 skills, 9 multilingual languages, and 4 categories of ambiguity handling. Notably, *we firstly propose the evaluation of ambiguity handling* in audio dialogues that expresses different intentions beyond the same literal meaning of sentences, *e.g.*, “Really!?” with different intonations. In summary, ADU-Bench includes over 20,000 open-ended audio dialogues for the assessment of LALMs. Through extensive experiments on 16 LALMs, our analysis reveals that existing LALMs struggle with mathematical symbols and formulas, understanding human behavior such as roleplay, comprehending multiple languages, and handling audio dialogue ambiguities from different phonetic elements, such as intonations, pause positions, and homophones. The benchmark is available at https://adu-bench.github.io/.

1 Introduction

Large Audio-Language Models (LALMs) (Chu et al., 2023; Tang et al., 2024; Wu et al., 2023; Kong et al., 2024; Lin et al., 2024; Xie and Wu, 2024; Fu et al., 2024) have received attention for their abilities to handle various audio-related tasks. In particular, LALMs recently unlock unprecedented capabilities for interactive audio dialogues with humans. These dialogues are defined as a direct exchange of spoken language between LALMs and humans, which fosters a more dynamic mode of communication. Recent advances, such as GPT-4o (OpenAI, 2024), have enabled LALMs to engage in back-and-forth dialogues with humans and can observe various audio characteristics, which broadens their applicability across diverse real-world situations that rely on interactive audio dialogues.

However, given these advancements, there is currently no comprehensive benchmark to evaluate LALMs’ performance in handling open-ended audio dialogue understanding. Previous benchmarks on LALMs predominantly focus on their performance in multiple fundamental tasks (Huang et al., 2024b,a), audio question answering with text-based instructions (Yang et al., 2024; Deshmukh et al., 2024; Sakshi et al., 2024; Wang et al., 2025) or audio dialogues in general scenarios (Ao et al., 2024; Chen et al., 2024). The absence of a comprehensive benchmark for evaluating LALMs in open-ended audio dialogues has led to suboptimal comparisons between different LALMs.

Open-ended audio dialogues, where users can directly engage with LALMs through audio, constitute a significant portion of real-world interactions. These dialogues can encompass many subjects, such as helpful and daily questions, domain-specific skills, and multiple different languages. Additionally, the variations in intonations or pause positions can allow speakers to express different intentions beyond the same literal meaning of sentences, adding further complexity to the dialogues. Therefore, the ability to handle open-ended audio dialogues effectively is crucial for LALMs to be truly useful in real-world applications.

In this work, we propose an **A**udio **D**ialogue **U**nderstanding **Bench**mark **(ADU-Bench)**, a benchmark to evaluate the open-ended audio dialogue understanding for LALMs, which comprises

[†]Corresponding authors.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4763–4784
July 27 - August 1, 2025 ©2025 Association for Computational Linguistics

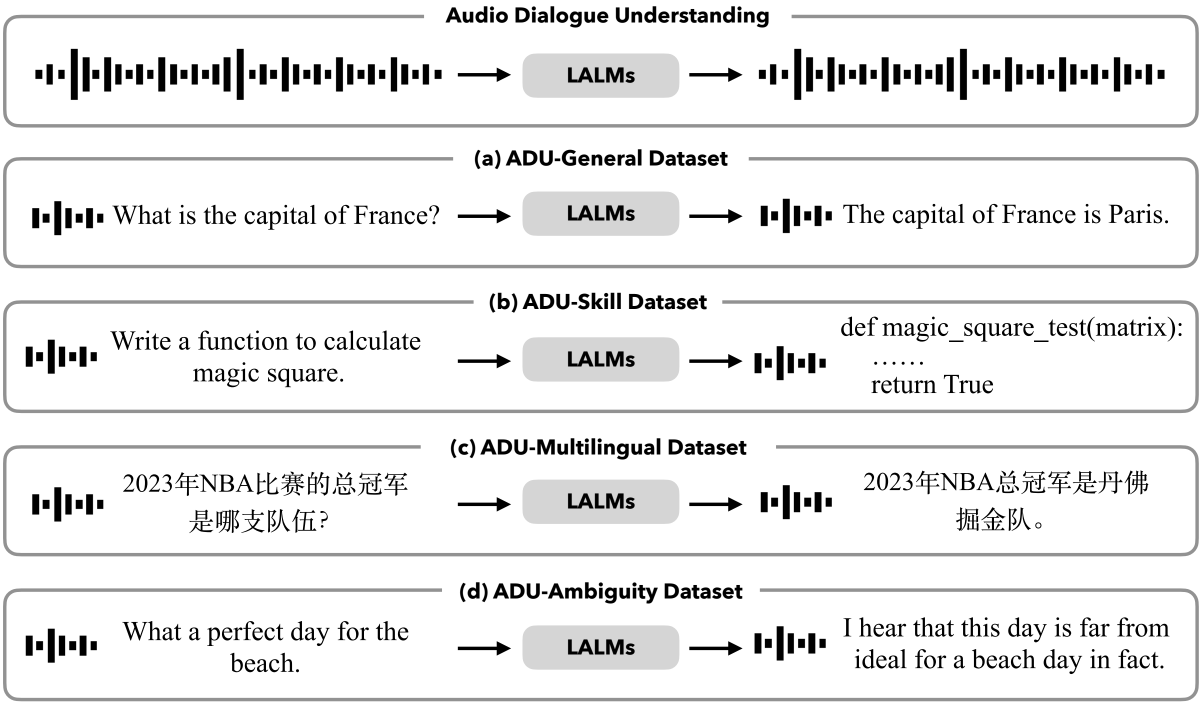

Audio Dialogue Understanding
LALMs
(a) ADU-General Dataset
What is the capital of France?
LALMs
The capital of France is Paris.
(b) ADU-Skill Dataset
Write a function to calculate magic square.
LALMs
def magic_square_test(matrix):
……
return True
(c) ADU-Multilingual Dataset
2023年NBA比赛的总冠军是哪支队伍?
LALMs
2023年NBA总冠军是丹佛掘金队。
(d) ADU-Ambiguity Dataset
What a perfect day for the beach.
LALMs
I hear that this day is far from ideal for a beach day in fact.

Figure 1: ADU-Bench evaluates the open-ended audio dialogue understanding for LALMs, where users interact with LALMs directly through audio. Our ADU-Bench consists of 4 datasets, including (a) ADU-General dataset, (b) ADU-Skill dataset, (c) ADU-Multilingual dataset, and (d) ADU-Ambiguity dataset. In total, it encompasses 20,715 open-ended audio dialogues, comprising over 8,000 real-world recordings alongside synthetic audio samples.

4 benchmark datasets as follows. (1) The ADU-General dataset assesses the general dialogue understanding of LALMs, including 3 scenarios, *i.e.*, helpful questions to query search engines, daily questions happening among human dialogues, and daily statements without rich contexts. (2) The ADU-Skill dataset evaluates the skill-based dialogue ability, encompassing 12 different skills such as mathematics, physics, coding, *etc*. (3) The ADU-Multilingual dataset tests the multilingual dialogue understanding, covering 9 languages, including English, French, and Chinese, *etc*. (4) The ADU-Ambiguity dataset is designed to evaluate the audio dialogue ambiguity handling ability from different phonetic elements, including intonation-based, pause-based, homophone-based, and repetition-based ambiguity. Notably, *we firstly analyze the ambiguity within audio dialogues,* specifically addressing the challenge of different intentions that share the same literal sentence, such as the word "`Really!?`" spoken with different intonations. In total, ADU-Bench comprises over 20,000 open-ended audio dialogues for LALMs. An overall example of ADU-Bench is shown in Fig. 1.

For the evaluation, LALMs are first queried with user audio inputs and generate corresponding textual responses directly or convert audio responses into a textual format. Then, we primarily use GPT-4 (Achiam et al., 2023) or manual annotation to generate references (expected ground truths) based on the textual transcriptions of each audio. Subsequently, following Zheng et al. (2023); Bai et al. (2024); Yang et al. (2024), we include the transcriptions of audio, references, and responses into an evaluation prompt and use this prompt to query GPT-4 (Achiam et al., 2023), which generates a score for evaluating the quality of generated responses. However, the order in which the references and responses are presented in the evaluation prompt can influence the scores generated by GPT-4, leading to position bias (Zheng et al., 2023). To eliminate position bias, we conduct a second scoring by swapping the positions of the references and responses during evaluation. In addition, to eliminate bias from the GPT-4 based evaluation, we have included more LLMs for evaluation, such as LLaMA-3-70B-Instruct (MetaAI, 2024) and Qwen-2-72B-Instruct (Chu et al., 2023).

We benchmark 16 popular LALMs on our ADU-Bench and analyze the results. Our analysis reveals: (1) There is still considerable room for improvement in the audio dialogue understanding of existing open-sourced LALMs. (2) LALMs face challenges when dealing with skills, such as Mathematics and Coding, which involve mathematical symbols and formulas. (3) LALMs exhibit limitations in handling tasks related to Common Sense and Roleplay, as they lack a deeper understanding of human behavior. (4) Existing LALMs struggle to comprehend different meanings of audio dialogues that have the same transcriptions, but differ in phonetic elements, such as intonations, pause positions, and homophones. We include some demonstrations of our audio dialogues on our project page.

2 Related Work

Large Audio-language Models. Large audio-language models (LALMs) (Chu et al., 2023; Tang et al., 2024; Wu et al., 2023; Kong et al., 2024) typically integrate audio modalities into large language models (LLMs) (Touvron et al., 2023; OpenAI, 2024; Gao et al., 2024c,b,a; Fang et al., 2025; Zou et al., 2025; Kong et al., 2025) to extend their capabilities for general-purpose audio and language understanding. LALMs can be broadly classified into two types: end-to-end LALMs and cascaded LALMs. End-to-end LALMs can be further divided into two categories: (1) End-to-end LALMs specialize in audio understanding, which focus on integrating audio modality into LLMs, such as SpeechGPT (Zhang et al., 2023), BLSP (Wang et al., 2023), SALMONN (Tang et al., 2024), and Qwen-Audio (Chu et al., 2023). (2) End-to-end LALMs extend their capabilities beyond audio understanding, which align various modalities into a single LLM, such as PandaGPT (Su et al., 2023) and NExT-GPT (Wu et al., 2024). Another approach involves cascaded LALMs like the combination of an automatic speech recognition model, such as Whisper-large (Radford et al., 2023), and an LLM, such as GPT-4 (Achiam et al., 2023), to process a wide range of audio types. Our ADU-Bench aims to evaluate their performance in audio dialogue understanding across different domains.

Benchmarks for LALMs. Existing benchmarks for audio-related tasks can be broadly categorized into three areas: (1) fundamental audio tasks, (2) audio question answering with text-based instructions, and (3) audio dialogues. For benchmarks focusing on fundamental audio tasks (Huang et al., 2024b,a), evaluations are typically centered around specific objectives such as speech-to-text translation or emotion recognition. In audio question answering with text-based instructions (Yang et al., 2024; Deshmukh et al., 2024; Sakshi et al., 2024; Wang et al., 2025), models are required to interpret input audio and respond to input text-based instructions. In contrast, benchmarks for audio dialogues evaluate models to directly respond to audio queries without text-based instructions. While several established benchmarks (Ao et al., 2024; Chen et al., 2024) exist for audio dialogues, they predominantly focus on general scenarios, leaving a comprehensive benchmark unexplored. To bridge this gap, we propose ADU-Bench, which concentrates on evaluating LALMs in a wide range of audio dialogue scenarios.

3 ADU-Bench

3.1 Overall

ADU-Bench is a comprehensive evaluation benchmark designed to assess the open-ended audio dialogue understanding of LALMs in scenarios where LALMs directly respond to user audio inputs. ADU-Bench consists of 4 datasets, including ADU-General dataset, ADU-Skill dataset, ADU-Multilingual dataset, and ADU-Ambiguity dataset. During data collection, our ADU-Bench contains 20,715 open-ended audio dialogues, comprising over 8,000 real-world recordings alongside synthetic audio samples. The generation details of synthetic audio samples are in Appendix A. The dataset details for ADU-Bench are in Table 1. Each data point within these datasets is presented as a tuple consisting of (*audio queries*, *textual references*). The audio queries serve as the input for LALMs, while the textual references function as the expected ground truths. The generation of textual references involves inputting the corresponding textual transcriptions of audio queries into GPT-4 or employing human annotation for ambiguity types. A textual format is chosen for the data construction because ADU-Bench focuses on the understanding of audio dialogues instead of generation.

3.2 Data Construction

The ADU-General dataset is constructed to evaluate the general dialogue understanding capabilities of LALMs. This dataset comprises 12,000 open-ended audio dialogues, specifically designed to reflect a wide array of inquiries and remarks commonly encountered in life. It covers 3 scenarios as follows. (1) Helpful questions: These are typically aimed at eliciting useful responses from search engines, such as “Who won the most gold medals in the Olympics?”. (2) Daily questions: These repre-

Table 1: Data collection and statistics on 4 datasets in ADU-Bench, including dataset domains, dataset source, and dataset number. In total, ADU-Bench consists of 20,715 open-ended audio dialogues.

Datasets	Domains	Source	Number
ADU-General	Helpful Question Daily Question Daily Statement	Alpaca, NQ-Bench WebGLM, Slue HVB Common Voice	12,000
ADU-Skill	Mathematics, Physics Chemistry, Biology Computer Science, Code, Law Finance, Common Sense Writing, Roleplay, Medicine	GSM8K, MATH WizardLM, ShareGPT MBPP, MMLU HotpotQA, StrategyQA	3,725
ADU-Multilingual	Arabic, Chinese, English French, German, Japanese Korean, Russian, Spanish	Alpaca, NQ-Bench WebGLM	3,600
ADU-Ambiguity	Intonation-based Pause-based, Homophone-based Repetition-based	Phonetics and phonology books	1,390

sent casual questions that arise in real-life conversations, for example, "What are you doing on this fine day?". (3) Daily statements: These include everyday remarks, such as "One today is worth two tomorrows.". In particular, daily questions and statements are relatively casual without rich contextual information to represent real-world situations. The construction of this dataset draws from various sources including Alpaca (Taori et al., 2023), NQ-Bench (Kwiatkowski et al., 2019), WebGLM (Liu et al., 2023), Slue HVB (Shon et al., 2022), and Common Voice (Ardila et al., 2019). To eliminate queries that do not align with the aforementioned categories, we implement a filtering process combining GPT-4 and human inspection.

The ADU-Skill dataset is specifically designed to assess the domain-specific skills of LALMs. This dataset comprises 3,750 audio dialogues and encompasses 12 different domains, including Mathematics, Physics, Chemistry, Biology, Computer Science, Coding, Law, Finance, Common Sense, Writing, Roleplay, and Medicine. To cover these diverse domains, we collect sources for these dialogues from GSM8K (Cobbe et al., 2021), MATH (Hendrycks et al., 2021), WizardLM (Xu et al., 2023), ShareGPT (Chiang et al., 2023), MBPP (Austin et al., 2021), MMLU (Hendrycks et al., 2020), HotpotQA (Yang et al., 2018), and StrategyQA (Geva et al., 2021). Notably, in certain domains, particularly Mathematics, Physics, and Coding, some queries involve a high volume of Latex formulas or Python code, which can be challenging to comprehend when transformed into audio. Therefore, we employ GPT-4 and human inspection to filter out queries with an excessive number of Latex formulas or Python code.

The ADU-Multilingual dataset aims to evaluate the multilingual dialogue understanding abilities, covering 9 languages: Arabic, Chinese, English, French, German, Japanese, Korean, Russian, and Spanish. This dataset consists of 3,600 audio dialogues. For generation, we first randomly choose 400 different queries in English from ADU-General dataset. Subsequently, these queries are then translated into the other 8 languages using GPT-4. By including multiple languages, this dataset tests LALMs to understand the audio dialogues in various linguistic contexts. Furthermore, the design of this dataset allows for future expansion, enabling the inclusion of additional languages as needed.

The ADU-Ambiguity dataset is specifically designed to evaluate the robustness of LALMs in addressing ambiguity from different phonetic elements present in audio dialogues. It is important to note that ambiguity refers to instances where the textual transcriptions alone, without the accompanying audio or contexts, can lead to confusion. However, when considering the phonetic elements or contextual information provided by the audio, these ambiguities can be resolved, leading to a standard, unambiguous response for humans. Concretely, this dataset consists of 1,390 audio dialogues, which can be classified into 4 types of ambiguous situations, as described below. (1) Intonation-based ambiguity: In this case, expressing the same sentence with different intonations leads to different interpretations. For instance, "What a perfect day for the beach." can convey different meanings depending on the intonation used. An uplifting intonation indicates that it is indeed a perfect day, while a disappointed intonation signifies that the conditions are far from ideal for a beach day. (2) Pause-based ambiguity: The placement of pauses can alter the meaning of a sentence. For example, consider the phrase "professional reviewers and authors." Depending on where the pause is placed, it can imply that both the reviewers and authors are smart, or that only the reviewers are smart while the authors are not. The ambiguity arises from the different ways in which pauses can be inserted into the sentence, leading to contrasting interpretations. (3) Homophone-based ambiguity: These are sentences containing words that sound almost the same when spoken but have completely different meanings due to variations in word spelling. For example, the words "weight" and "wait" sound almost the same but convey different meanings. (4) Repetition-based ambiguity: These sentences contain multiple occurrences of the same word, often leading to confusion. An example of this is, "I saw a man saw a saw with a saw." The construction of the ADU-Ambiguity dataset is achieved manually, drawing upon research studies (McMahon, 2002; Carr, 2019) related to phonetics. To annotate textual references, we employ a combination of GPT-4 and manual inspection, ensuring the accuracy and relevance of the references.

4 Evaluation Method

Given recent studies (Zheng et al., 2023; Yang et al., 2024) have demonstrated that the evaluation with LLMs exhibits better alignment with human preferences, we propose to adopt the advanced LLM, GPT-4, to evaluate the quality of the responses generated by LALMs. Concretely, LALMs first are queried with audio queries and generate textual responses directly, or convert audio responses into textual format. Subsequently, we present the textual transcriptions of audio queries, textual references (expected ground truths) generated by GPT-4, and textual responses generated by LALMs into the GPT-4 evaluator. Finally, the GPT-4 evaluator assigns an overall score on a scale of 0 to 10 for each data point. A higher score indicates the better LALMs' capabilities in handling open-ended audio dialogues. The evaluation prompt templates are in Appendix B. To eliminate the position bias arising from the order of references and responses, we perform a second scoring by swapping their positions and report the average results. Moreover, to avoid bias from GPT-4, we also use LLaMA-3-70B-Instruct and Qwen-2-72B-Instruct for evaluation.

5 Results and Analysis

5.1 Experimental Settings

To benchmark the audio dialogue understanding of existing LALMs, we evaluate 16 foundational models with audio understanding capabilities. These models include PandaGPT-7B (Su et al., 2023), NExT-GPT-7B (Wu et al., 2024), Qwen-Audio-7B (Chu et al., 2023), Qwen-Audio-Chat-7B (Chu et al., 2023), Mini-Omni-0.5B (Xie and Wu, 2024), SpeechGPT-7B (Zhang et al., 2023), Moshi-7B (Défossez et al., 2024), SALMONN-7B (Tang et al., 2024), SALMONN-13B (Tang et al., 2024), BLSP-7B (Wang et al., 2023), Step-Audio-Chat-130B (Huang et al., 2025), Whisper-large-v3 (Radford et al., 2023) with LLaMA-2-7B-Chat (Touvron et al., 2023), with LLaMA-3-8B-Instruct (MetaAI, 2024), with LLaMA-3-70B-Instruct (MetaAI, 2024), with GPT-4 (`gpt-4o-0613`) (Achiam et al., 2023), and GPT-4o (`gpt-4o-audio-preview-2024-12-17`) (OpenAI, 2024). Unless stated otherwise, the hyperparameters and setups used during the evaluation process remain consistent with those specified in the original papers of the respective models. For evaluation, we obtain two evaluation scores by swapping references and responses in the prompts for the GPT-4 evaluator and finally report the average scores for each model in Table 2. In addition, to avoid the bias of evaluation only using GPT-4, we apply various open-sourced LLMs for such evaluations, including LLaMA-3-70B-Instruct (MetaAI, 2024) and Qwen-2-72B-Instruct (Chu et al., 2023). In addition, we conduct a direct human evaluation on randomly selected 140 audio dialogues. Each sample is assessed by three human testers, who rate the generated responses. More details about experimental settings and human evaluation are in Appendix C and Appendix D, respectively.

5.2 Main Results

We report the experimental results for the performance of 16 different LALMs on audio dialogue understanding in Table 2 and provide a comprehensive analysis of them. Firstly, it can be observed that PandaGPT, NExT-GPT, and Qwen-Audio exhibit the lowest performances, with an average score value of about 1.00. It illustrates that although PandaGPT and NExT-GPT are end-to-end LALMs capable of processing a wide range of

Table 2: The average evaluation scores for audio dialogue understanding under 16 LALMs in our ADU-Bench.

Models	Size	ADU-Bench				Average	Human Evaluation
		General	Skill	Multilingual	Ambiguity		
PandaGPT	7B	1.02	0.98	0.98	0.50	0.87	-
NExT-GPT	7B	1.07	1.03	1.02	0.52	0.91	-
Qwen-Audio	7B	1.32	1.08	1.07	0.61	1.02	-
Mini-Omni	0.5B	2.31	1.96	1.55	1.67	1.87	-
SALMONN	7B	2.47	2.01	1.83	1.73	2.01	-
Qwen-Audio-Chat	7B	2.34	2.46	1.58	1.93	2.08	-
SpeechGPT	7B	3.99	3.56	1.42	2.25	2.81	-
Moshi	7B	4.37	3.08	1.49	2.81	2.94	-
SALMONN	13B	4.07	3.12	3.25	1.86	3.08	-
BLSP	7B	4.66	4.49	2.89	3.37	3.85	-
Step-Audio-Chat	130B	6.37	7.31	2.45	4.72	5.21	-
Whisper+LLaMA-2	7B	6.30	6.26	4.92	4.39	5.47	6.43
Whisper+LLaMA-3	8B	6.94	7.88	6.27	4.92	6.50	6.85
Whisper+LLaMA-3	70B	7.26	8.03	6.12	5.13	6.64	7.46
Whisper+GPT-4	-	8.42	8.62	8.07	5.54	7.66	8.02
GPT-4o	-	8.64	8.97	8.16	6.87	8.16	8.58

modalities, their performances on audio dialogue understanding are relatively lower. As for Qwen-Audio, a pre-trained base LALM, its weak capabilities in audio dialogue indicate a potential necessity for more specialized training to enhance its understanding of audio dialogues.

Compared to them, Mini-Omni-0.5B, SALMONN-7B and Qwen-Audio-Chat show somewhat superior performance. This can be attributed to the fact that Mini-Omni-0.5B, SALMONN-7B, and Qwen-Audio-Chat have been developed under audio instruction tuning, making them suitable for a variety of audio-oriented scenarios. Moreover, SpeechGPT, Moshi, SALMONN-13B, BLSP, and Step-Audio-Chat have demonstrated even higher proficiency, as reflected in their average scores all about or exceeding 3.00. Among these, BLSP stands out with the highest average score of 3.85 among all LALMs. As SALMONN increases in size from 7B to 13B, its audio dialogue understanding capabilities also show improvement. In addition, both SpeechGPT and BLSP enable audio dialogue with LLMs using speech and exhibit impressive dialogue capabilities. Therefore, it can achieve enhanced performance when using the targeted audio dialogue tuning for end-to-end LALMs.

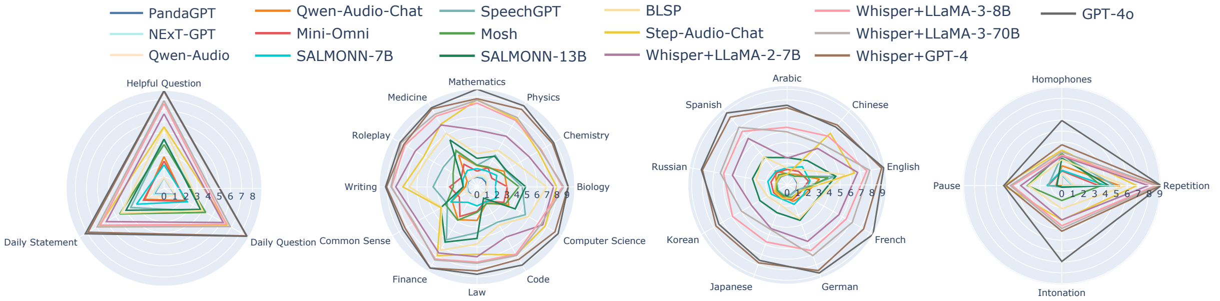
Furthermore, cascaded LALMs, including LLaMA-2-7B, LLaMA-3-8B, LLaMA-3-70B, and GPT-4 with a Whisper model, obtain higher scores in audio dialogue understanding. Therein, GPT-4 leads the pack with a high score of 7.66. Following it, LLaMA-3 (including LLaMA-3-8B and LLaMA-3-70B) ranks second, outperforming its predecessor, LLaMA-2. The improved performance of LLaMA-3 to LLaMA-2 highlights the effectiveness of updates in the LLaMA series.

Notably, for the advanced proprietary LALM, GPT-4o, achieves the highest average score of 8.16, which indicates that it is the best-performing model among the evaluated LALMs.

In addition, experimental results reveal that the GPT-4 evaluator demonstrates a significantly higher correlation with human evaluations, as shown in Table 2. Furthermore, we also conduct another human evaluation study, detailed in Section 5.4. These human evaluations verify the alignment between GPT-4 evaluator and human judgments.

5.3 Results on Each Dataset

Results on ADU-General dataset. The ADU-General dataset aims to evaluate the proficiency in general dialogue understanding, with results across 3 scenarios shown in Fig. 2(a). Our analysis reveals that LALMs perform better in helpful questions compared to daily questions and daily statements. Helpful questions typically seek specific information, whereas daily questions and daily statements represent everyday communication between humans, often characterized by a lack of rich contextual information. This finding suggests that LALMs are more adept at handling audio dialogues that require precise information retrieval, while their performance in everyday dialogues remains an area for improvement. In summary, existing open-sourced LALMs understand helpful questions better than daily questions and statements, highlighting the continued development in LALMs



(a) ADU-General dataset (b) ADU-Skill dataset (c) ADU-Multilingual dataset (d) ADU-Ambiguity dataset

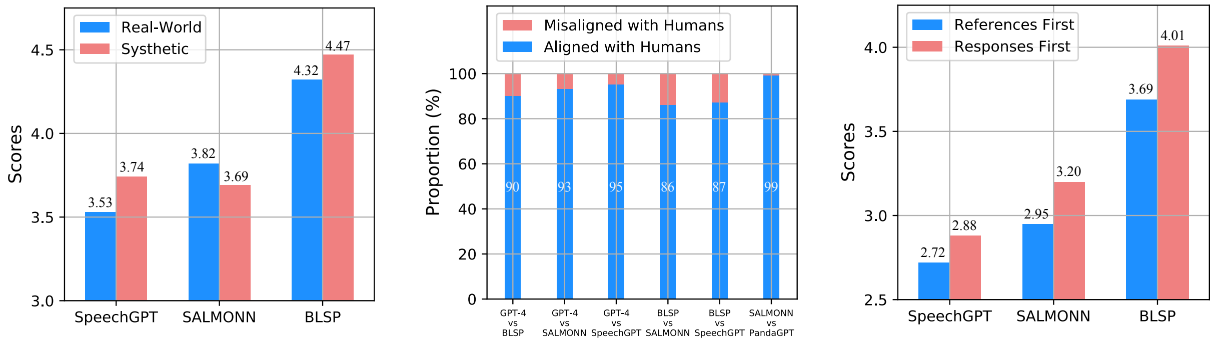
Figure 2: The average scores across each domain for 4 datasets within ADU-Bench under 16 LALMs.

to better address everyday human interactions.

Results on ADU-Skill dataset. The ADU-Skill dataset is designed to evaluate the skill capabilities of LALMs during audio dialogue and the results across 12 domains are shown in Fig. 2(b). Among all these domains, LALMs demonstrate a relative proficiency in handling topics such as Biology, Computer Science, Law, Finance, Writing, and Medicine. This observation suggests that LALMs possess a certain knowledge foundation in these domains. Meanwhile, these tasks primarily involve language understanding and generation, which align well with the core capabilities of LALMs. Moreover, LALMs exhibit weaker performance when dealing with subjects like Mathematics, Physics, Chemistry, and Coding. This can be attributed to the fact that they all involve mathematical symbols and formulas or programming languages so that LALMs struggle to effectively understand these domain-specific challenges they present. Additionally, LALMs display limitations in areas related to Common Sense and Roleplay. These domains usually require a deeper understanding of human behavior and LALMs lack the ability to infer implicit meanings or cultural nuances that are crucial for accurately understanding and responding to them. In summary, existing open-sourced LALMs have knowledge backgrounds in some domains but they face challenges in subjects involving mathematical notations or programming languages, as well as areas requiring a deeper understanding of human behavior.

Results on ADU-Multilingual dataset. The ADU-Multilingual dataset aims to evaluate multilingual capabilities of LALMs during audio dialogues, with results across 9 languages depicted in Fig. 2(c). It can be observed that all LALMs perform best in English due to the massive amount of training data in English. Subsequently, the performance is followed by German, Spanish, French, and Russian. We conjecture that this is because these languages all belong to the Indo-European languages that LALMs can understand to a certain extent. As for other languages, LALMs exhibit weaker performance which illustrates that they need to be incorporated into the development of LALMs. In conclusion, existing open-sourced LALMs struggle with their multilingual capabilities, highlighting further research to consider various linguistic contexts when developing LALMs.

Results on ADU-Ambiguity dataset. The ADU-Ambiguity dataset is designed to assess how well LALMs handle 4 types of ambiguity during audio dialogue, including intonation-based, pause-based, homophone-based, and repetition-based ambiguity, with results in Fig. 2(d). Overall, LALMs exhibit relatively better performance in handling repetition-based ambiguity, while their performance in managing other types of ambiguities is weaker. This observation suggests that LALMs can more effectively resolve ambiguities that do not involve phonetic elements, such as repetition-based ambiguity, which only has multiple words in an audio. However, when it comes to the other three types of ambiguities, including intonation-based, pause-based, and homophone-based, LALMs struggle to handle them effectively. For homophone-based ambiguity, it is difficult for LALMs to distinguish the words that have almost the same pronunciation. For the other two types of ambiguity, LALMs can not perceive the variations in intonations or pause positions, which can lead to expressing different intentions beyond the same literal meaning of sentences. When faced with these ambiguities, relatively advanced LALMs like GPT-4o can achieve an average score of 5.22 and 6.05 for pause-based and homophone-based ambiguity. The results show that GPT-4o often generates responses that encompass multiple possible interpretations, which is unable to effectively distinguish between the different



(a) Real-world and synthetic audio (b) Human evaluation study (c) Position bias study

Figure 3: Ablation study on ADU-Bench. (a) Real-world and synthetic audio can both serve as evaluation sources. (b) GPT-4 evaluator is aligned with human evaluation. (c) Scoring twice is necessary to eliminate the position bias.

meanings based on phonetic elements. In summary, existing LALMs, including GPT-4o, display limitations in handling the audio dialogue ambiguity in different phonetic elements.

5.4 Ablation Study

Effect of LALMs' size. We compare the audio dialogue understanding capabilities of SALMONN and LLaMA-3 with a Whisper model with different sizes. As shown in Table 2, it indicates a trend of improved average scores with increasing model sizes. However, it is noted that SALMONN-7B outperforms its larger counterpart, SALMONN-13B on Code within ADU-Skill dataset. Similarly, LLaMA-3-8B achieves superior performance than LLaMA-3-70B on Common Sense within ADU-Skill dataset and non-Indo-European languages within ADU-Multilingual dataset. These observations suggest that while a larger model size generally contributes to better overall audio dialogue understanding performance, it can also introduce performance losses in certain domains.

Effect of real-world and synthetic audio. For the audio dialogues difficult to obtain directly, we choose to adopt a synthetic algorithm to generate corresponding audios, as detailed in Appendix A. To demonstrate that the use of synthetic audio is a feasible approach compared to real-world audio when evaluating LALMs, we randomly sample 1,000 real-world audio dialogues and generate synthetic audio from their transcriptions. The comparison between the real-world audio and the synthetic audio with the same transcriptions is presented in Fig. 3(a). We observe that there is no considerable difference in the performance of LALMs when processing real-world and synthetic audio. In conclusion, both real-world audio and synthetic audio can effectively serve as evaluation sources for audio dialogue understanding.

Human evaluation study. For evaluation, we choose to adopt GPT-4 as the evaluator. To evaluate the consistency between the evaluations of GPT-4 and human judgments, we conduct a human evaluation study as follows. Given the challenge of human testers directly assigning a score on a scale of 0 to 10, we adopt a pairwise comparison approach for models, following (Touvron et al., 2023). Specifically, human testers first listen to the audio queries, then compare two textual responses from two models, finally indicate their preference as "A is better", "B is better", or "Both are equal". We then convert the GPT-4 scores into the same preference-based rating as the human testers. Finally, we evaluate the consistency between the two sets of results, as shown in Fig. 3(b). Our analysis reveals that the pairwise preference consistency achieves a score above 85%, indicating that GPT-4 evaluation aligns well with human judgments. The details are in Appendix D. We provide the evaluation results by LLaMA-3-70B-Instruct and Qwen-2-72B-Instruct and the corresponding human evaluation study in Appendix F.

Position bias study. To mitigate potential biases from the order of references and responses in the evaluation GPT-4 prompt, we query the GPT-4 evaluator to generate two scores by adjusting their positions. Subsequently, we report the average score for each model. To validate the necessity of scoring twice, we compare the differences between the two scores, presented in Fig. 3(c). We observe that despite using the same references and responses, the GPT-4 evaluator generates different scores after adjusting the positions. This suggests the existence of a positional bias, particularly when responses are placed before the references. The observation highlights the importance of conducting a second scoring to address this bias.

6 Conclusion

In conclusion, we present ADU-Bench, a comprehensive benchmark designed to evaluate the audio dialogue understanding of LALMs. It encompasses 4 benchmark datasets including ADU-General dataset for 3 general scenarios, ADU-Skill dataset for 12 skills, ADU-Multilingual dataset for 9 languages, and ADU-Ambiguity for 4 ambiguity types, providing over 20,000 open-ended audio dialogues for the LALM assessment. Our extensive experiments on 16 LALMs reveal that there is still significant room for improvement in their audio dialogue understanding. Notably, LALMs face challenges in processing mathematical symbols and formulas, comprehending human behavior like roleplay, understanding multiple languages, and handling audio dialogue ambiguities arising from various phonetic elements.

Limitations

The main limitation of this work is that the analysis is on a limited number of LALMs due to the availability of usable code, model weights, and massive experiments. Potential future work includes investigating more diverse LALMs and designing more domains about audio dialogues to make our ADU-Bench up-to-date.

Ethics Statement

Our ADU-Bench has been carefully curated to ensure that it does not contain any words or content that discriminate against any individual or group. The prompts used in our experiments, as detailed in Appendix B, have been meticulously reviewed to emphasize that none of them contain any discriminatory language or themes. Moreover, we have taken the necessary precautions to ensure that the prompts used in our work do not negatively impact anyone's safety or well-being. Furthermore, all the codes comply with the MIT License. This commitment to ethical considerations (Deng et al., 2024) in our research contributes to the responsible development and advancement of LALMs.

Acknowledgement

This work is supported in part by the National Natural Science Foundation of China under Grant 62171248 and Shenzhen Science and Technology Program (JCYJ20220818101012025).

References

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

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. In *NeurIPS*.

Rosana Ardila, Megan Branson, Kelly Davis, Michael Henretty, Michael Kohler, Josh Meyer, Reuben Morais, Lindsay Saunders, Francis M Tyers, and Gregor Weber. 2019. Common voice: A massively-multilingual speech corpus. *arXiv preprint arXiv:1912.06670*.

Jacob Austin, Augustus Odena, Maxwell Nye, Maarten Bosma, Henryk Michalewski, David Dohan, Ellen Jiang, Carrie Cai, Michael Terry, Quoc Le, et al. 2021. Program synthesis with large language models. *arXiv preprint arXiv:2108.07732*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. In *ACL*.

Philip Carr. 2019. *English phonetics and phonology: An introduction*. John Wiley & Sons.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, march 2023.

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*.

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

Alexandre Défossez, Laurent Mazaré, Manu Orsini, Amélie Royer, Patrick Pérez, Hervé Jégou, Edouard Grave, and Neil Zeghidour. 2024. Moshi: a speech-text foundation model for real-time dialogue. *arXiv preprint arXiv:2410.00037*.

Chengyuan Deng, Yiqun Duan, Xin Jin, Heng Chang, Yijun Tian, Han Liu, Yichen Wang, Kuofeng Gao, Henry Peng Zou, Yiqiao Jin, et al. 2024. Deconstructing the ethics of large language models from long-standing issues to new-emerging dilemmas: A survey. *arXiv preprint arXiv:2406.05392*.

Soham Deshmukh, Shuo Han, Hazim Bukhari, Benjamin Elizalde, Hannes Gamper, Rita Singh, and Bhiksha Raj. 2024. Audio entailment: Assessing deductive reasoning for audio understanding. *arXiv preprint arXiv:2407.18062*.

Hao Fang, Jiawei Kong, Tianqu Zhuang, Yixiang Qiu, Kuofeng Gao, Bin Chen, Shu-Tao Xia, Yaowei Wang, and Min Zhang. 2025. Your language model can secretly write like humans: Contrastive paraphrase attacks on llm-generated text detectors. *arXiv preprint arXiv:2505.15337*.

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, et al. 2024. Vita: Towards open-source interactive omni multimodal llm. *arXiv preprint arXiv:2408.05211*.

Kuofeng Gao, Yang Bai, Jindong Gu, Shu-Tao Xia, Philip Torr, Zhifeng Li, and Wei Liu. 2024a. Inducing high energy-latency of large vision-language models with verbose images. In *ICLR*.

Kuofeng Gao, Huanqia Cai, Qingyao Shuai, Dihong Gong, and Zhifeng Li. 2024b. Embedding self-correction as an inherent ability in large language models for enhanced mathematical reasoning. *arXiv preprint arXiv:2410.10735*.

Kuofeng Gao, Tianyu Pang, Chao Du, Yong Yang, Shu-Tao Xia, and Min Lin. 2024c. Denial-of-service poisoning attacks against large language models. *arXiv preprint arXiv:2410.10760*.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? a question answering benchmark with implicit reasoning strategies. *Transactions of the Association for Computational Linguistics*, 9:346–361.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2020. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. 2021. Measuring mathematical problem solving with the math dataset. In *NeurIPS*.

Ailin Huang, Boyong Wu, Bruce Wang, Chao Yan, Chen Hu, Chengli Feng, Fei Tian, Feiyu Shen, Jingbei Li, Mingrui Chen, et al. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Chien-yu Huang, Wei-Chih Chen, Shu-wen Yang, Andy T Liu, Chen-An Li, Yu-Xiang Lin, Wei-Cheng Tseng, Anuj Diwan, Yi-Jen Shih, Jiatong Shi, et al. 2024a. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

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. 2024b. Dynamic-superb: Towards a dynamic, collaborative, and comprehensive instruction-tuning benchmark for speech. In *ICASSP*.

Jiawei Kong, Hao Fang, Xiaochen Yang, Kuofeng Gao, Bin Chen, Shutao Xia, Yaowei Wang, and Min Zhang. 2025. Wolf hidden in sheep's conversations: Toward harmless data-based backdoor attacks for jailbreaking large language models. *arXiv preprint arXiv:2505.17601*.

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*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, et al. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *SOSP*.

Guan-Ting Lin, Cheng-Han Chiang, and Hung-yi Lee. 2024. Advancing large language models to capture varied speaking styles and respond properly in spoken conversations. *arXiv preprint arXiv:2402.12786*.

Xiao Liu, Hanyu Lai, Hao Yu, Yifan Xu, Aohan Zeng, Zhengxiao Du, Peng Zhang, Yuxiao Dong, and Jie Tang. 2023. Webglm: Towards an efficient web-enhanced question answering system with human preferences. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4549–4560.

April MS McMahon. 2002. *An introduction to English phonology*, volume 22. Edinburgh University Press Edinburgh.

MetaAI. 2024. Llama-3 technical report. *https://llama.meta.com/llama3/*.

Microsoft. 2024. Speech synthesis markup language service in microsoft. *https://azure.microsoft.com/*.

OpenAI. 2024. Gpt-4o technical report. *https://openai.com/index/hello-gpt-4o/*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *ICML*.

S Sakshi, Utkarsh Tyagi, Sonal Kumar, Ashish Seth, Ramaneswaran Selvakumar, Oriol Nieto, Ramani Duraiswami, Sreyan Ghosh, and Dinesh Manocha. 2024. Mmau: A massive multi-task audio understanding and reasoning benchmark. *arXiv preprint arXiv:2410.19168*.

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

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. 2024. Salmonn: Towards generic hearing abilities for large language models. In *ICLR*.

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

Paul Taylor and Amy Isard. 1997. Ssml: A speech synthesis markup language. *Speech communication*, 21(1-2):123–133.

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*.

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

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

Jian Wu, Yashesh Gaur, Zhuo Chen, Long Zhou, Yimeng Zhu, Tianrui Wang, Jinyu Li, Shujie Liu, Bo Ren, Linquan Liu, et al. 2023. On decoder-only architecture for speech-to-text and large language model integration. In *ASRU*.

Shengqiong Wu, Hao Fei, Leigang Qu, Wei Ji, and Tat-Seng Chua. 2024. Next-gpt: Any-to-any multimodal llm. In *ICML*.

Zhifei Xie and Changqiao Wu. 2024. Mini-omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv:2408.16725*.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, and Daxin Jiang. 2023. Wizardlm: Empowering large language models to follow complex instructions. *arXiv preprint arXiv:2304.12244*.

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*.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhutdinov, and Christopher D Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *EMNLP*.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Qingsong Zou, Jingyu Xiao, Qing Li, Zhi Yan, Yuhang Wang, Li Xu, Wenxuan Wang, Kuofeng Gao, Ruoyu Li, and Yong Jiang. 2025. Making them a malicious database: Exploiting query code to jailbreak aligned large language models. In *ACL*.

A Generation Details for Synthetic Datasets

Our ADU-Bench contains 20,715 open-ended audio dialogues, comprising over 8,000 real-world recordings alongside synthetic audio samples. In this section, we introduce the generation details for the synthetic datasets.

To generate synthetic datasets for ADU-General dataset, ADU-Skill dataset, and ADU-Multilingual dataset, we first adopt GPT-4 and human inspection to obtain the related textual dialogues for each dataset. Then, enclose them in the Speech Synthesis Markup Language (SSML) (Taylor and Isard, 1997) by human coding, where SSML is an XML-based markup language specifically designed for speech synthesis applications. Subsequently, execute the program code using Python interpreter with public SSML service (Microsoft, 2024) provided by Microsoft Azure to convert them into audio dialogues. Furthermore, to emulate real-world scenarios, we consider a wide array of variables for synthetic audio. They include 2 genders (male and female), 4 different speakers (2 men and 2 women), 4 emotions (calm, excited, angry, and sad), 3 speech rates (standard and $\pm 10\%$), 3 pitch levels (standard and $\pm 10\%$), and 3 volume levels (standard and $\pm 10\%$). During the generation of each dataset, a combination of these audio generation characteristics is randomly selected to create each audio data, ensuring diversity in the audio dialogues. Therefore, this generation method not only provides a scalable solution for generating synthetic audio datasets but also ensures a rich diversity that closely mirrors real-world audio dialogues.

To construct the ADU-Ambiguity dataset, we first identify four types of ambiguity handling from phonetics and phonology books (McMahon, 2002; Carr, 2019). These include ambiguity stemming from intonation, pause positions, homophones, and repetition. Based on the examples and principles outlined in these references, we then manually craft or use GPT-4 to generate many textual data instances representing these ambiguity types.

To convert these textual instances into audio samples, we leverage the Speech Synthesis Markup Language (SSML) (Taylor and Isard, 1997) and use a publicly available SSML service(Microsoft, 2024). Specifically:

- For intonation-based ambiguity, we use the SSML tags `<prosody>` to adjust the intonation elements of the audio.
- For pause-based ambiguity, we use the SSML tags `<break>` to insert pauses within the audio.
- For homophone-based and repetition-based ambiguity, we are able to directly generate the audio samples without the need for specialized SSML markup.

Finally, we conduct a manual validation process to ensure the quality and correctness of the generated audio samples. This involves having human annotators listen to the samples and verify that the intended ambiguity is successfully conveyed through the audio.

B Prompts for Evaluation

The score judgment is based on criteria including helpfulness, relevance, accuracy, and comprehensiveness, comparing the references and generated responses. The evaluation prompt for *the first scoring* is as follows.

System Prompt

You are a helpful and precise assistant for checking the quality of the answer.

Prompts for Evaluation in ADU-Bench

Please evaluate the following LALMs' response for the user query and a reference is provided.

Query: *Textual Transcriptions*
Reference: *Textual References*
Response: *Textual Responses*

Please rate the helpfulness, relevance, accuracy, and comprehensiveness of the LALMs' response. Please provide an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Do not provide any other output text or explanation. Only provide the score.

Output:

The evaluation prompt for *the second scoring* is as follows. To eliminate the position bias, we swap

Table 3: Association between human judgment and each dataset in ADU-Bench of GPT-4 evaluation.

	GPT-4 vs BLSP	GPT-4 vs SALMONN	GPT-4 vs SpeechGPT	BLSP vs SALMONN	BLSP vs SpeechGPT	SALMONN vs PandaGPT
ADU-General	86.7%	80.0%	93.3%	86.7%	93.3%	100%
ADU-Skill	86.7%	93.3%	93.3%	83.3%	88.3%	100%
ADU-Multilingual	95.6%	95.6%	97.8%	86.7%	86.7%	97.8%
ADU-Ambiguity	90.0%	95.0%	95.0%	90.0%	90.0%	100%
ADU-Bench	90.0%	92.9%	95.0%	85.7%	87.1%	99.3%

the position between responses and references in the evaluation prompt.

System Prompt

You are a helpful and precise assistant for checking the quality of the answer.

Prompts for Evaluation in ADU-Bench

Please evaluate the following LALMs' response for the user query and a reference is provided.

Query: *Textual Transcriptions*
Response: *Textual Responses*
Reference: *Textual References*

Please rate the helpfulness, relevance, accuracy, and comprehensiveness of the LALMs' response. Please provide an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. Do not provide any other output text or explanation. Only provide the score.

Output:

The evaluation pipeline is shown in Fig. 4. We choose GPT-4 as the default evaluation LLM. We also include LLaMA-3-70B-Instruct and Qwen-2-72B-Instruct to provide the evaluation score. The results are shown in Appendix F.

C Details of Experimental Settings

To benchmark the audio dialogue understanding of existing LALMs, we assess the performance of 16 LALMs across all 4 datasets within ADU-Bench. Unless stated otherwise, the hyperparameters and setups used during the evaluation process remain consistent with those specified in the original papers of the respective models. For the evaluation, LLaMA-2-7B-Chat, LLaMA-3-8B-Instruct, and LLaMA-3-70B-Instruct are run on 8 NVIDIA A100 40GB GPUs with vLLM library (Kwon et al., 2023), while other open-sourced models are run on a single NVIDIA A100 40GB GPU. By default, our evaluation method employs `gpt-4-0613` as the GPT-4 evaluator by calling the API.

D Human Evaluation Study Details

We conduct a direct human evaluation on randomly selected 140 audio dialogues from ADU-Bench. Each sample is assessed by three human testers, who rate the generated responses. Human testers should provide an overall score on a scale of 0 to 10, where a higher score indicates better overall performance. The results are shown in Table 2. Besides, we conduct a human evaluation study to evaluate the consistency between the evaluations of GPT-4 and human judgments. We show each pair of samples for ten human testers. The results are demonstrated in Fig. 3(b) and Table 3.

For the evaluation datasets, we randomly choose 5 audio queries from each domain in ADU-Bench, and finally obtain 140 audio queries. Since ADU-Multilingual contains multiple languages, it is difficult for human testers to understand each language. Hence, we provide the textual transcriptions and allow them to use the translation tools for evaluation. we carefully consider the ethical aspects and potential risks associated with the research involving human subjects. The information we collect is only the preference results and does not involve any personal information.

When selecting participants, there are no requirements for their qualifications, experience, or technical abilities; all participants are adults capable of giving informed consent. We clearly inform the participants of the experiment's content and corresponding compensation before the experiment begins, and we will not cause them any physiological or psychological harm. We randomly select participants within the university campus, informing them of the experiment content, purpose, compensation, and other information. Participants voluntarily de-

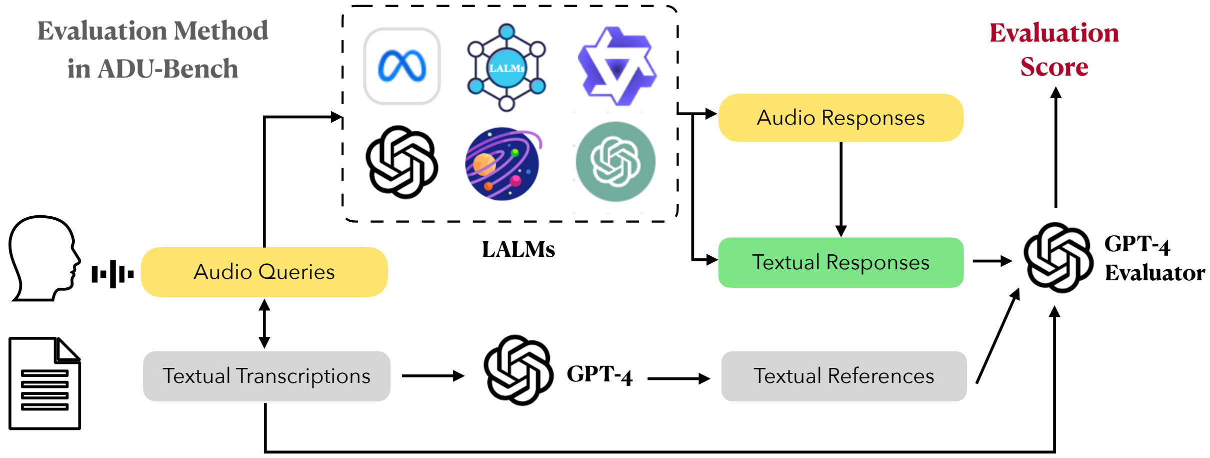


Figure 4: The evaluation method in ADU-Bench. To benchmark open-ended audio dialogue understanding for LALMs, we adopt a GPT-4 evaluator to provide evaluation scores as the metric. We also adopt LLaMA-3-70B-Instruct and Qwen-2-72B-Instruct as the evaluator to provide evaluation scores.

cide whether to participate in the experiment after reading the Ethics Informed Consent Form and Ethics Study Information Sheet. The compensation we provide to the participants is 1.5 times the local minimum hourly wage standard.

The instructions given to participants in Table 2 are as follows:

Welcome to our human evaluation study! Your participation is crucial in helping us assess the performance of Large Audio-Language Models (LALMs) in audio dialogue understanding.

In this study, you will be presented with a total of 140 audio clips, each accompanied by one textual response. For audio in foreign languages, we will provide textual transcriptions and translation tools to assist you.

Your task is as follows:

1. Listen to the audio queries carefully.

2. Based on the criteria of helpfulness, relevance, accuracy, and comprehensiveness, provide an overall score on a scale of 0 to 10 for the response, where a higher score indicates better overall performance.

We appreciate your time and effort in participating in this study. Your valuable insights will significantly contribute to the development and improvement of LALMs, enhancing their ability to understand and respond to audio dialogues effectively. Thank you for your participation!

The instructions given to participants in Fig. 3(b) and Table 3 are as follows:

Welcome to our human evaluation study! Your participation is crucial in helping us assess the performance of Large Audio-Language Models (LALMs) in audio dialogue understanding.

In this study, you will be presented with a total of 140 audio clips, each accompanied by two textual responses. For audio in foreign languages, we will provide textual transcriptions and translation tools to assist you.

Your task is as follows:

1. Listen to the audio queries carefully.

2. Compare the two textual responses provided for each audio.

3. Based on the criteria of helpfulness, relevance, accuracy, and comprehensiveness, indicate your preference. You can choose from the following options: "A is better", "B is better", or "Both are equal".

We appreciate your time and effort in participating in this study. Your valuable insights will significantly contribute to the development and improvement of LALMs, enhancing their ability to understand and respond to audio dialogues effectively. Thank you for your participation!

E Discussions

E.1 Real-world and Synthetic Audio in ADU-Bench

Our ADU-Bench includes both real-world and synthetic audio. As stated in Section 3, the data collection involves a combination of synthetically generated dialogues and real-world audio samples. Specifically, our ADU-Bench contains over 8000 audio samples from the real world. A reason prevents us from using real-world audio only is the challenges and costs of the collection process. In particular, the collection of professional technical terms or languages can be difficult, as it requires humans who are familiar with them. Without proper familiarity, the use of these terms or languages in audio samples may sound unnatural. To address this issue, we propose a synthetic method for audio generation in Appendix A. By leveraging it, we can easily expand a scalable ADU-Bench without incurring substantial expenses. Besides, we have conducted an ablation study to investigate the effects of real-world and synthetic audio on the performance of our benchmark, as detailed in Section 5.4. It can be observed that there is no significant difference in the performance of LALMs in the areas our ADU-Bench covers. It illustrates that these synthetic audios can also benchmark audio dialogue understanding.

E.2 Evaluation using Textual Response

In our evaluation process, we prompt audio queries to obtain audio responses and adopt their textual transcriptions with references to calculate a GPT-4 evaluation score. We have chosen this approach because *our primary focus is on how LALMs comprehend audio dialogue* and formulate appropriate replies. In our ADU-Bench, we emphasize *understanding ability* rather than generation quality of audio dialogues. Directly using audio for evaluation can be challenging, and evaluating generation quality is not within the scope of ADU-Bench. By opting for a textual format, we can concentrate on assessing LALMs' dialogue understanding abilities and their capacity to provide meaningful responses, without introducing the additional complexity of audio generation. Furthermore, our evaluation approach in ADU-Bench aligns with previous work (Yang et al., 2024).

E.3 Analysis for Weak Performance of LALMs

LALMs consist of two main components - audio feature extractors and base LLMs. For textual benchmarks such as GSM8K and MMLU, the base LLMs of LALMs are usually able to achieve effective mathematical and knowledge-based reasoning, which reflects the fundamental reasoning skills of the base LLMs. However, for our ADU-Bench, the LALMs exhibit weak performance and are unable to demonstrate the fundamental reasoning skills of their base LLM components. This observation leads us to conjecture that the poor performance of the LALMs on the ADU-Bench is primarily rooted in their audio comprehension abilities, rather than their core reasoning skills.

F Evaluation Results by LLaMA-3-70B-Instruct and Qwen-2-72B-Instruct

To avoid the bias of evaluation only using GPT-4, we apply various open-sourced LLMs for such evaluations, including LLaMA-3-70B-Instruct and Qwen-2-72B-Instruct. Our analysis shows that the evaluation scores obtained using these LLMs are mostly consistent with the conclusions drawn from the GPT-4 evaluation. The results are shown in Table 4 and Table 5.

Furthermore, we also include their corresponding human evaluation studies, which can be found in Table 6 and Table 7. All these results indicate that strong LLM evaluations, especially those involving GPT-4, align well with human judgments for audio dialogue understanding. Besides, note that GPT-4 based evaluation is shown to be effective in many areas (Zheng et al., 2023; Yang et al., 2024).

G Reproducibility Statement

We provide the code and data in the project page of our ADU-Bench.

H More Details of ADU-Bench

The details of ADU-Bench, including the number of each domain within ADU-Bench are in Table 8.

I More Overall Results

The overall results of the first and second scoring are shown in Table 9 and Table 10.

J More Results on Each Dataset

The results on each dataset of the first and second scoring are shown in Table 11, Table 12, Table 13, Table 14, Table 15, Table 16, Table 17, Table 18, Table 19, Table 20, Table 21, and Table 22.

Table 4: The average evaluation scores under 16 different LALMs on 4 datasets in our ADU-Bench. The evaluation is conducted by LLaMA-3-70B-Instruct.

Models	Size	ADU-Bench General	 Skill	 Multilingual	 Ambiguity	Average
PandaGPT	7B	1.00	1.00	1.00	1.00	1.00
NExT-GPT	7B	1.04	1.03	1.00	1.00	1.02
Qwen-Audio	7B	2.00	1.00	1.42	1.00	1.36
Mini-Omni	0.5B	2.12	1.26	1.49	1.27	1.54
SALMONN	7B	2.71	1.42	1.71	1.72	1.89
Qwen-Audio-Chat	7B	1.85	3.14	2.06	1.95	2.25
SpeechGPT	7B	3.71	3.57	1.94	2.42	2.91
Moshi	7B	3.96	3.15	2.04	2.64	2.95
SALMONN	13B	3.71	4.23	2.92	2.05	3.23
BLSP	7B	4.42	3.90	2.07	2.95	3.34
Step-Audio-Chat	130B	6.53	6.66	1.92	4.02	4.78
Whisper+LLaMA-2	7B	6.28	5.07	3.07	3.86	4.57
Whisper+LLaMA-3	8B	7.57	7.00	5.00	4.75	6.08
Whisper+LLaMA-3	70B	7.28	7.85	6.42	5.12	6.67
Whisper+GPT-4	-	8.57	7.92	8.50	5.46	7.61
GPT-4o	-	8.69	8.35	8.61	6.37	8.00

Table 5: The average evaluation scores under 16 different LALMs on 4 datasets in our ADU-Bench. The evaluation is conducted by Qwen-2-72B-Instruct.

Models	Size	ADU-Bench General	 Skill	 Multilingual	 Ambiguity	Average
PandaGPT	7B	1.00	1.00	1.00	1.00	1.00
NExT-GPT	7B	1.10	1.06	1.00	1.00	1.04
Qwen-Audio	7B	1.45	1.23	1.31	1.12	1.28
Mini-Omni	0.5B	1.74	1.49	1.53	1.31	1.52
SALMONN	7B	2.36	1.31	2.09	1.31	1.77
Qwen-Audio-Chat	7B	2.57	1.74	2.45	2.85	2.40
SpeechGPT	7B	4.09	4.13	2.35	2.64	3.30
Moshi	7B	4.14	3.35	2.36	2.85	3.18
SALMONN	13B	3.81	3.63	2.54	2.96	3.24
BLSP	7B	4.18	4.54	2.48	3.84	3.76
Step-Audio-Chat	130B	6.43	5.86	2.31	3.95	4.64
Whisper+LLaMA-2	7B	6.27	5.13	3.47	3.94	4.70
Whisper+LLaMA-3	8B	6.81	6.00	3.68	4.02	5.13
Whisper+LLaMA-3	70B	7.18	6.63	3.86	4.36	5.51
Whisper+GPT-4	-	8.45	8.09	6.63	4.87	7.01
GPT-4o	-	8.58	8.42	6.78	5.33	7.28

Table 6: Association between human judgment and each dataset in ADU-Bench of of LLaMA-3-70B-Instruct evaluation.

	GPT-4 vs BLSP	GPT-4 vs SALMONN	GPT-4 vs SpeechGPT	BLSP vs SALMONN	BLSP vs SpeechGPT	SALMONN vs PandaGPT
ADU-General	80.0%	86.7%	93.3%	80.0%	86.7%	100%
ADU-Skill	90.0%	86.7%	93.3%	85.0%	86.7%	98.3%
ADU-Multilingual	95.6%	97.8%	97.8%	82.2%	82.2%	100%
ADU-Ambiguity	90.0%	90.0%	90.0%	86.0%	86.0%	100%
ADU-Bench	90.7%	90.7%	94.3%	83.6%	85.0%	99.3%

Table 7: Association between human judgment and each dataset in ADU-Bench of Qwen-2-72B-Instruct evaluation.

	GPT-4 vs BLSP	GPT-4 vs SALMONN	GPT-4 vs SpeechGPT	BLSP vs SALMONN	BLSP vs SpeechGPT	SALMONN vs PandaGPT
ADU-General	80.0%	86.7%	86.7%	80.0%	80.0%	100%
ADU-Skill	86.7%	90.0%	96.0%	86.7%	80.0%	100%
ADU-Multilingual	93.3%	95.6%	95.0%	82.2%	85.0%	97.8%
ADU-Ambiguity	85.0%	90.0%	95.0%	86.0%	85.0%	100%
ADU-Bench	87.9%	91.4%	95.0%	84.3%	82.9%	99.3%

Table 8: The details of ADU-Bench, including the number of each domain within ADU-Bench.

Dataset	Domain	Number
ADU-General	Helpful Question	4,000
	Daily Question	4,000
	Daily Statement	4,000
ADU-Skill	Mathematics	1,000
	Physics	210
	Chemistry	180
	Biology	180
	Computer Science	115
	Code	1,000
	Law	325
	Finance	60
	Common Sense	500
	Writing	40
	Roleplay	20
	Medicine	95
ADU-Multilingual	Arabic	400
	Chinese	400
	English	400
	French	400
	German	400
	Japanese	400
	Korean	400
	Russian	400
	Spanish	400
ADU-Ambiguity	Intonation-based	395
	Pause-based	250
	Homophone-based	490
	Repetition-based	255

Table 9: The score for audio dialogue understanding performances under 16 different LALMs on 4 datasets in our proposed ADU-Bench. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Bench				Average
		General	Skill	Multilingual	Ambiguity	
PandaGPT	7B	0.98	0.97	0.97	0.49	0.85
NExT-GPT	7B	1.03	0.99	0.99	0.50	0.88
Qwen-Audio	7B	1.24	0.93	0.99	0.55	0.93
Mini-Omni	0.5B	2.20	1.87	1.49	1.51	1.77
SALMONN	7B	2.35	1.92	1.71	1.69	1.92
Qwen-Audio-Chat	7B	2.21	2.31	1.49	1.85	1.97
SpeechGPT	7B	3.91	3.40	1.39	2.18	2.72
Moshi	7B	4.31	3.00	1.45	2.76	2.88
SALMONN	13B	3.83	3.10	3.08	1.80	2.95
BLSP	7B	4.50	4.27	2.74	3.25	3.69
Step-Audio-Chat	130B	6.30	7.23	2.39	4.66	5.15
Whisper+LLaMA-2	7B	6.07	6.20	4.82	4.30	5.35
Whisper+LLaMA-3	8B	6.66	7.80	6.21	4.79	6.37
Whisper+LLaMA-3	70B	6.82	7.97	6.09	4.97	6.46
Whisper+GPT-4	-	8.33	8.54	8.04	5.43	7.59
GPT-4o	-	8.54	8.84	8.07	6.79	8.06

Table 10: The score for audio dialogue understanding performances under 16 different LALMs on 4 datasets in our proposed ADU-Bench. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Bench				Average
		General	Skill	Multilingual	Ambiguity	
PandaGPT	7B	1.06	0.98	0.98	0.50	0.88
NExT-GPT	7B	1.11	1.07	1.04	0.53	0.94
Qwen-Audio	7B	1.40	1.23	1.14	0.67	1.11
Mini-Omni	0.5B	2.42	2.06	1.61	1.84	1.98
SALMONN	7B	2.59	2.09	1.94	1.77	2.10
Qwen-Audio-Chat	7B	2.47	2.60	1.66	2.00	2.19
SpeechGPT	7B	4.06	3.71	1.44	2.32	2.88
Moshi	7B	4.43	3.16	1.52	2.86	2.99
SALMONN	13B	4.31	3.14	3.42	1.91	3.20
BLSP	7B	4.82	4.70	3.04	3.48	4.01
Step-Audio-Chat	130B	6.44	7.38	2.51	4.77	5.27
Whisper+LLaMA-2	7B	6.53	6.32	5.02	4.48	5.59
Whisper+LLaMA-3	8B	7.21	7.96	6.32	5.04	6.63
Whisper+LLaMA-3	70B	7.70	8.09	6.14	5.29	6.81
Whisper+GPT-4	-	8.51	8.70	8.09	5.64	7.74
GPT-4o	-	8.74	9.10	8.24	6.94	8.25

Table 11: The score for audio dialogue understanding performances under 16 different LALMs on ADU-General dataset. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-General		
		Helpful Question	Daily Question	Daily Statement
PandaGPT	7B	0.99	1.00	0.96
NExT-GPT	7B	1.00	1.10	1.00
Qwen-Audio	7B	0.90	1.23	1.58
Mini-Omni	0.5B	2.34	2.24	2.02
SALMONN	7B	2.05	2.34	2.66
Qwen-Audio-Chat	7B	2.77	2.00	1.86
SpeechGPT	7B	4.37	4.09	3.28
Moshi	7B	3.96	4.35	4.62
SALMONN	13B	4.19	3.59	3.70
BLSP	7B	5.33	3.91	4.27
Step-Audio-Chat	130B	5.56	6.61	6.73
Whisper+LLaMA-2	7B	6.69	5.88	5.64
Whisper+LLaMA-3	8B	7.65	6.12	6.22
Whisper+LLaMA-3	70B	7.71	6.34	6.42
Whisper+GPT-4	-	8.63	8.51	7.84
GPT-4o	-	8.76	8.65	8.20

Table 12: The score for audio dialogue understanding performances under 16 different LALMs on ADU-General dataset. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-General		
		Helpful Question	Daily Question	Daily Statement
PandaGPT	7B	0.99	1.17	1.03
NExT-GPT	7B	0.98	1.15	1.21
Qwen-Audio	7B	1.15	1.34	1.72
Mini-Omni	0.5B	2.56	2.43	2.26
SALMONN	7B	2.20	2.51	3.06
Qwen-Audio-Chat	7B	2.96	2.35	2.10
SpeechGPT	7B	4.39	4.12	3.66
Moshi	7B	4.12	4.41	4.75
SALMONN	13B	4.70	4.02	4.22
BLSP	7B	5.64	4.14	4.68
Step-Audio-Chat	130B	5.68	6.82	6.84
Whisper+LLaMA-2	7B	6.75	6.46	6.38
Whisper+LLaMA-3	8B	7.67	6.88	7.08
Whisper+LLaMA-3	70B	8.12	7.45	7.52
Whisper+GPT-4	-	8.86	8.67	8.00
GPT-4o	-	8.92	8.74	8.55

Table 13: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Skill dataset. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Skill (Part I)					
		Mathematics	Physics	Chemistry	Biology	Computer Science	Code
PandaGPT	7B	0.98	1.00	0.99	1.00	0.97	0.90
NExT-GPT	7B	0.99	1.02	1.14	0.98	0.99	0.96
Qwen-Audio	7B	1.03	1.19	1.04	0.86	0.89	0.84
Mini-Omni	0.5B	1.48	2.06	1.63	2.92	2.97	1.55
SALMONN	7B	1.78	1.73	2.26	1.87	2.09	1.66
Qwen-Audio-Chat	7B	1.99	2.06	2.96	2.79	3.62	1.74
SpeechGPT	7B	1.99	3.41	3.14	4.52	5.33	3.94
Moshi	7B	2.17	2.28	2.77	4.19	3.33	1.94
SALMONN	13B	3.15	3.24	3.09	4.76	4.31	1.31
BLSP	7B	2.99	3.94	4.39	6.91	5.76	4.31
Step-Audio-Chat	130B	8.70	7.81	7.06	7.96	7.85	7.85
Whisper+LLaMA-2	7B	5.65	5.59	5.86	7.59	7.41	5.78
Whisper+LLaMA-3	8B	8.21	7.65	7.35	8.58	7.12	7.73
Whisper+LLaMA-3	70B	8.63	7.93	7.38	8.62	7.21	7.84
Whisper+GPT-4	-	8.72	8.93	8.66	9.00	8.96	8.34
GPT-4o	-	9.53	9.06	8.67	8.98	9.21	8.84

Table 14: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Skill dataset. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Skill (Part II)					
		Law	Finance	Common Sense	Writing	Roleplay	Medicine
PandaGPT	7B	1.01	0.98	0.99	0.97	1.00	1.00
NExT-GPT	7B	0.98	1.12	1.00	0.99	1.05	0.99
Qwen-Audio	7B	0.88	0.77	0.84	1.36	1.10	0.83
Mini-Omni	0.5B	2.39	3.31	1.96	2.64	1.72	2.52
SALMONN	7B	1.84	1.82	2.81	1.39	1.56	1.85
Qwen-Audio-Chat	7B	3.20	3.65	2.65	1.19	0.80	3.49
SpeechGPT	7B	4.40	6.08	3.22	4.50	3.52	3.93
Moshi	7B	2.55	3.85	3.60	2.37	2.76	4.21
SALMONN	13B	4.93	6.09	3.90	1.44	1.65	5.23
BLSP	7B	5.52	7.10	3.87	6.63	5.07	5.97
Step-Audio-Chat	130B	6.74	7.95	4.16	7.41	6.12	7.23
Whisper+LLaMA-2	7B	6.87	7.60	6.77	8.20	6.68	7.05
Whisper+LLaMA-3	8B	7.44	8.35	7.26	8.42	8.24	8.10
Whisper+LLaMA-3	70B	7.59	8.46	7.16	8.55	8.64	8.26
Whisper+GPT-4	-	8.25	9.38	8.12	8.92	8.12	8.93
GPT-4o	-	8.41	9.25	8.32	8.71	8.14	8.98

Table 15: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Skill dataset. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Skill (Part I)					
		Mathematics	Physics	Chemistry	Biology	Computer Science	Code
PandaGPT	7B	0.98	1.00	1.00	0.98	1.01	0.95
NExT-GPT	7B	1.12	1.20	1.15	1.10	0.99	0.98
Qwen-Audio	7B	1.26	1.57	1.37	1.11	1.18	0.97
Mini-Omni	0.5B	1.65	2.33	1.79	3.14	3.22	1.77
SALMONN	7B	1.85	1.97	2.30	1.81	2.41	1.84
Qwen-Audio-Chat	7B	2.25	2.34	3.29	3.16	3.69	2.00
SpeechGPT	7B	2.31	3.87	3.40	4.52	5.82	4.22
Moshi	7B	2.44	2.46	2.85	4.22	3.52	2.27
SALMONN	13B	2.54	3.81	3.61	4.97	4.51	1.30
BLSP	7B	3.68	4.50	4.81	7.00	6.12	4.51
Step-Audio-Chat	130B	8.73	7.92	7.36	8.15	8.02	7.74
Whisper+LLaMA-2	7B	5.71	6.08	6.17	7.70	7.77	5.82
Whisper+LLaMA-3	8B	8.53	7.71	7.47	8.50	7.16	7.82
Whisper+LLaMA-3	70B	8.70	8.07	7.29	8.69	7.62	8.01
Whisper+GPT-4	-	8.90	8.94	8.72	9.21	9.03	8.41
GPT-4o	-	9.76	9.35	8.84	9.12	9.36	9.03

Table 16: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Skill dataset. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Skill (Part II)					
		Law	Finance	Common Sense	Writing	Roleplay	Medicine
PandaGPT	7B	1.02	1.00	0.99	1.00	1.05	1.00
NExT-GPT	7B	1.00	1.03	1.12	0.99	1.12	1.00
Qwen-Audio	7B	1.08	1.13	1.65	1.40	1.27	1.16
Mini-Omni	0.5B	2.52	3.53	2.11	2.82	1.93	2.68
SALMONN	7B	1.96	1.77	3.27	1.40	1.65	1.96
Qwen-Audio-Chat	7B	3.51	3.94	3.08	1.14	1.50	3.87
SpeechGPT	7B	4.78	6.14	3.61	4.28	4.29	4.21
Moshi	7B	2.77	4.02	3.85	2.53	2.84	4.13
SALMONN	13B	5.39	6.67	4.44	1.32	2.00	5.58
BLSP	7B	5.92	7.53	4.31	6.89	6.35	6.37
Step-Audio-Chat	130B	6.85	8.16	4.27	7.63	6.33	7.42
Whisper+LLaMA-2	7B	7.44	8.35	7.26	8.42	8.24	8.10
Whisper+LLaMA-3	8B	7.61	8.33	7.42	8.66	8.40	8.22
Whisper+LLaMA-3	70B	7.68	8.42	7.29	8.64	8.89	8.51
Whisper+GPT-4	-	8.54	9.36	8.46	9.16	8.78	9.07
GPT-4o	-	8.74	9.38	8.54	9.14	8.87	9.12

Table 17: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Multilingual dataset. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Multilingual (Part I)				
		Arabic	Chinese	English	French	German
PandaGPT	7B	0.98	0.98	0.96	0.98	0.97
NExT-GPT	7B	0.99	0.99	1.00	1.00	0.99
Qwen-Audio	7B	0.95	1.08	0.93	1.02	0.94
Mini-Omni	0.5B	1.37	1.57	1.99	1.38	1.40
SALMONN	7B	1.47	2.14	2.11	1.67	1.85
Qwen-Audio-Chat	7B	1.00	1.18	2.95	1.73	1.54
SpeechGPT	7B	0.98	1.04	4.48	1.01	1.00
Moshi	7B	1.07	1.08	3.94	1.30	1.27
SALMONN	13B	2.38	2.88	4.48	2.90	3.30
BLSP	7B	1.51	1.81	5.28	2.94	3.20
Step-Audio-Chat	130B	1.00	6.34	6.43	1.24	1.36
Whisper+LLaMA-2	7B	2.36	4.36	6.68	5.60	5.62
Whisper+LLaMA-3	8B	5.33	5.97	7.56	6.36	6.50
Whisper+LLaMA-3	70B	5.02	5.02	7.89	7.02	7.08
Whisper+GPT-4	-	7.26	7.34	8.99	8.32	8.60
GPT-4o	-	7.40	6.80	9.09	9.25	8.69

Table 18: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Multilingual dataset. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Multilingual (Part II)			
		Japanese	Korean	Russian	Spanish
PandaGPT	7B	0.98	0.98	0.98	0.96
NExT-GPT	7B	0.98	0.97	0.98	0.98
Qwen-Audio	7B	0.91	1.10	0.98	0.99
Mini-Omni	0.5B	1.36	1.41	1.49	1.47
SALMONN	7B	1.37	1.59	1.70	1.52
Qwen-Audio-Chat	7B	1.08	1.16	1.01	1.75
SpeechGPT	7B	1.00	1.01	1.03	1.00
Moshi	7B	1.08	1.07	1.03	1.24
SALMONN	13B	2.62	2.87	3.12	3.16
BLSP	7B	1.86	2.00	2.80	3.27
Step-Audio-Chat	130B	1.36	1.33	1.07	1.07
Whisper+LLaMA-2	7B	4.25	3.73	5.20	5.60
Whisper+LLaMA-3	8B	5.65	5.97	6.04	6.53
Whisper+LLaMA-3	70B	4.44	4.96	6.34	7.05
Whisper+GPT-4	-	7.81	7.68	8.07	8.31
GPT-4o	-	7.34	7.28	8.01	8.73

Table 19: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Multilingual dataset. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Multilingual (Part I)				
		Arabic	Chinese	English	French	German
PandaGPT	7B	0.99	0.97	0.98	0.98	0.98
NExT-GPT	7B	0.98	1.00	1.15	1.12	1.01
Qwen-Audio	7B	1.09	1.29	1.12	1.08	1.13
Mini-Omni	0.5B	1.55	1.76	2.12	1.47	1.52
SALMONN	7B	1.76	2.32	2.27	1.92	2.05
Qwen-Audio-Chat	7B	1.07	1.41	3.23	2.04	1.77
SpeechGPT	7B	1.00	1.10	4.68	1.04	1.03
Moshi	7B	1.16	1.23	4.21	1.48	1.43
SALMONN	13B	2.76	3.08	4.83	3.25	3.81
BLSP	7B	1.67	1.99	5.74	3.26	3.60
Step-Audio-Chat	130B	1.21	6.56	6.57	1.45	1.42
Whisper+LLaMA-2	7B	2.68	4.61	6.82	5.67	5.71
Whisper+LLaMA-3	8B	5.53	6.03	7.69	6.50	6.68
Whisper+LLaMA-3	70B	4.98	5.06	7.93	7.15	7.11
Whisper+GPT-4	-	7.28	7.39	9.10	8.33	8.60
GPT-4o	-	7.51	7.12	9.24	9.35	8.84

Table 20: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Multilingual dataset. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Multilingual (Part II)			
		Japanese	Korean	Russian	Spanish
PandaGPT	7B	0.98	0.98	0.99	0.98
NExT-GPT	7B	0.99	0.98	0.99	1.10
Qwen-Audio	7B	1.10	1.33	1.06	1.08
Mini-Omni	0.5B	1.43	1.53	1.60	1.53
SALMONN	7B	1.48	1.87	1.99	1.80
Qwen-Audio-Chat	7B	1.31	1.37	1.04	1.69
SpeechGPT	7B	1.02	1.05	1.05	1.02
Moshi	7B	1.24	1.21	1.16	1.42
SALMONN	13B	2.96	3.06	3.52	3.58
BLSP	7B	2.07	2.29	3.16	3.61
Step-Audio-Chat	130B	1.47	1.52	1.21	1.18
Whisper+LLaMA-2	7B	5.65	5.97	6.04	6.53
Whisper+LLaMA-3	8B	5.80	5.92	6.12	6.63
Whisper+LLaMA-3	70B	4.54	4.98	6.44	7.11
Whisper+GPT-4	-	7.84	7.81	8.13	8.37
GPT-4o	-	7.58	7.43	8.21	8.85

Table 21: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Ambiguity dataset. The textual reference is *before* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Ambiguity			
		Intonation-based	Pause-based	Homophone-based	Repetition-based
PandaGPT	7B	0	0	0.98	0.98
NExT-GPT	7B	0	0.01	0.99	0.99
Qwen-Audio	7B	0	0.04	1.13	1.03
Mini-Omni	0.5B	0.07	1.26	1.34	3.38
SALMONN	7B	0.08	1.31	1.35	4.00
Qwen-Audio-Chat	7B	0.01	0.52	1.70	5.18
SpeechGPT	7B	0.13	1.19	2.70	4.70
Moshi	7B	1.37	3.15	2.94	3.61
SALMONN	13B	0.14	0.40	2.41	4.26
BLSP	7B	2.01	2.92	2.38	5.70
Step-Audio-Chat	130B	3.17	5.34	3.25	6.87
Whisper+LLaMA-2	7B	3.02	3.65	2.65	7.86
Whisper+LLaMA-3	8B	3.40	4.44	2.76	8.56
Whisper+LLaMA-3	70B	3.64	4.56	2.92	8.76
Whisper+GPT-4	-	4.02	5.02	3.64	9.03
GPT-4o	-	6.96	5.11	5.97	9.10

Table 22: The score for audio dialogue understanding performances under 16 different LALMs on ADU-Ambiguity dataset. The textual reference is *after* the textual response in the evaluation prompt for the GPT-4 evaluator.

Models	Size	ADU-Ambiguity			
		Intonation-based	Pause-based	Homophone-based	Repetition-based
PandaGPT	7B	0	0	0.99	1.00
NExT-GPT	7B	0	0.02	1.10	1.00
Qwen-Audio	7B	0	0.02	1.27	1.38
Mini-Omni	0.5B	1.00	1.35	1.46	3.53
SALMONN	7B	0.08	1.42	1.46	4.10
Qwen-Audio-Chat	7B	0.02	0.57	1.94	5.48
SpeechGPT	7B	0.15	1.30	2.82	5.00
Moshi	7B	1.42	3.21	3.12	3.72
SALMONN	13B	0.16	0.52	2.62	4.35
BLSP	7B	2.22	3.23	2.39	6.09
Step-Audio-Chat	130B	3.26	5.47	3.45	6.90
Whisper+LLaMA-2	7B	3.27	3.88	2.75	8.02
Whisper+LLaMA-3	8B	3.98	4.66	2.86	8.64
Whisper+LLaMA-3	70B	4.23	4.87	3.26	8.81
Whisper+GPT-4	-	4.35	5.20	3.86	9.14
GPT-4o	-	7.11	5.32	6.12	9.20