VoiceBench: Benchmarking LLM-Based Voice Assistants

Yiming Chen[†] **Xianghu Yue**[‡¶] **Chen Zhang**[†] **Xiaoxue Gao**[♣*]
Robby T. Tan[†] **Haizhou Li**[♮§]

[†]National University of Singapore, Singapore [‡]Tianjin University, China
[§]Shenzhen Research Institute of Big Data, China
[♮]School of Artificial Intelligence, The Chinese University of Hong Kong, Shenzhen, China
[♣]I2R, Agency for Science, Technology, and Research (A*STAR), Singapore
yiming.chen@u.nus.edu yuexianghu@tju.edu.cn
chen_zhang@u.nus.edu Gao_Xiaoxue@a-star.edu.sg

Abstract

Recent advancements in large language models (LLMs) like GPT-4o have enabled real-time speech interactions through LLM-based voice assistants, offering an improved user experience over text-based interactions. However, a suitable benchmark to rigorously evaluate such speech interactions systems is currently lacking. To bridge this gap, we introduce *VoiceBench*, the first benchmark specifically designed to assess LLM-based voice assistants. *VoiceBench* comprises 6,783 synthetic and real spoken instructions recorded from diverse speakers across eight distinct tasks. These instructions are meticulously crafted to assess three crucial capability areas: general knowledge, instruction-following, and safety compliance. Furthermore, *VoiceBench* systematically incorporates realistic variations common in spoken interactions, including differences in speaker characteristics (e.g., accents), heterogeneous environmental conditions (e.g., reverberation), and content complexities such as mispronunciations. Extensive experiments reveal the limitations of current LLM-based voice assistant models and offer valuable insights for future research and development in this field.[1]

1 Introduction

Advancements in large language models (LLMs) have led to remarkable breakthroughs across a wide range of natural language processing (NLP) tasks. Recently, LLMs have further expanded their capabilities by incorporating multi-modal processing abilities, such as vision (Liu et al., 2024; Chen et al., 2024c) and audio (Chu et al., 2023; Zhang et al., 2023; Yue et al., 2025; Jiang et al., 2024). Notably, voice assistants powered by audio-based LLMs, e.g., GPT-4o, have garnered significant research interest (Défossez et al., 2024; Li et al., 2024a). These voice assistants are designed to understand and respond to spoken instructions, enabling more natural, flexible, and high-quality speech interactions compared to traditional text-based systems. This advancement has the potential to significantly improve user experiences across various applications, e.g., virtual customer service.

Despite the promising potential of LLM-based voice assistants, the absence of a standardized benchmark for evaluating these systems limits a comprehensive understanding of their performance and areas for improvement. Current evaluations predominantly focus on automatic speech recognition (ASR) (Chen et al., 2025a; Xie and Wu, 2024a,b) or spoken question answering tasks synthesized with text-to-speech (TTS) models (Fang et al., 2025; Fu et al., 2024a). While informative, this narrow scope falls short of offering a holistic, fine-grained, and reliable assessment of model capabilities. Furthermore, the transition from text-based to speech-based interactions introduces several real-world challenges, as speaker characteristics (Krause and Braida, 2004), environmental factors (Meyer et al., 2013), and the complexity of spoken contents (Shriberg, 1994), affect human perception and pose significant difficulties for machine speech processing (Li et al., 2014; Kollmeier et al., 2008). Current evaluations, which mostly rely on clean speeches, fail to evaluate whether models are truly attending to real-world complexities adequately.

*Corresponding author.
¶Equal contribution.
[1]Code and data are available at github.com/MatthewCYM/VoiceBench.

Transactions of the Association for Computational Linguistics, vol. 14, pp. 378–398, 2026. https://doi.org/10.1162/TACL.a.628
Action Editor: Brian Kingsbury. Submission batch: 6/2025; Revision batch: 10/2025; Published 4/2026.
© 2026 Association for Computational Linguistics. Distributed under a CC-BY 4.0 license.

To address this gap, we introduce *VoiceBench*, which provides a comprehensive evaluation framework for LLM-based voice assistants. VoiceBench evaluates the capabilities of voice assistants across three key dimensions: general knowledge (understanding and reasoning), instruction-following abilities, and safety compliance, using both synthetic and real spoken instructions recorded from diverse multi-accented speakers. To comprehensively represent real-world variations, VoiceBench uniquely integrates challenging testing cases across distinct **speaker**, **environment**, and **content** variations. Specifically, we leverage advanced TTS and voice cloning models to generate synthetic spoken instructions with diverse speaker properties, including fine-grained variations in age, accent, and pitch. We propose to simulate a range of real-world environmental effects, including reverberation, compression, noise, and clipping, to evaluate model robustness under diverse environments. Lastly, we employ state-of-the-art LLMs to synthesize instructions that replicate content variations common in spoken language, including grammar errors, mispronunciations, and disfluencies.

Subsequently, we conduct an extensive evaluation of state-of-the-art voice assistants. Our results expose limitations of existing evaluation protocols, which rely heavily on ASR or synthetic data, thereby underscoring the unique value of VoiceBench. We also identify a substantial performance gap between end-to-end voice assistants and traditional pipeline models that couple an ASR system with an LLM. Although various voice assistants exhibit a certain degree of robustness, our analysis reveals that the speech encoder plays a decisive role: Weaker encoders render models highly vulnerable to input variations. Among the examined factors, novel accents, environmental variations, and mispronunciations have a comparatively stronger impact on performance. Notably, pipeline models demonstrate greater resilience to input variations compared to end-to-end systems, highlighting the pressing need for further advancements in end-to-end voice assistants.

Our major contributions include:

- **Novel benchmark:** We present the first comprehensive benchmark, *VoiceBench*, designed to evaluate the multi-faceted capabilities of LLM-based voice assistants, including general knowledge, instruction-following skills, and safety measures.
- **Real-world scenarios:** We investigate the impact of various real-world factors on the performance of voice assistants, encompassing speaker, environmental, and content variations.
- **Comprehensive evaluation:** We conduct an in-depth evaluation of existing voice assistants, identifying current weaknesses and providing directions for future improvements.

2 Background

2.1 AudioLLM

A common approach to augment LLMs with speech understanding capabilities is to implement pipeline models that first transcribe users' speech into text via ASR systems. The transcribed text is then passed to LLMs to generate responses. However, it has been argued that pipeline models may lose important information during the transcription process and often suffer from reduced efficiency (Fang et al., 2025; Xie and Wu, 2024a). To overcome these limitations, various end-to-end audio LLMs have been developed (Chu et al., 2024; Tang et al., 2024), which integrate speech encoders with LLMs via speech adapters (Fang et al., 2025; Xie and Wu, 2024a), enabling fully optimized end-to-end speech processing. Building on these audio LLMs, two main applications have emerged: audio analysis and voice assistants. Early explorations of AudioLLM primarily focused on audio analysis tasks, where models were designed to interpret given audio contexts and respond to text-based instructions (Gong et al., 2024; Chu et al., 2023). More recently, inspired by the success of GPT-4o, numerous LLM-based voice assistants have been developed to directly answer spoken queries, eliminating the need for explicit textual instructions (Held et al., 2025; Fu et al., 2024a; Chen et al., 2025a; Li et al., 2024a).

2.2 Voice Assistant Evaluation

While several established benchmarks exist for evaluating AudioLLMs (Yang et al., 2024; Chen et al., 2024a), these primarily focus on audio analysis tasks. Currently, no standardized benchmark explicitly targets the evaluation of voice

Model	Case Study	ASR	Basic SQA
LLaMA-Omni	✓	✓	✓
Mini-Omni	✓	✓	
Mini-Omni2	✓	✓	
Qwen2-Audio	✓	✓	
VITA		✓	
Moshi	✓	✓	✓
Baichuan-Omni		✓	
EMOVA		✓	
DiVA	✓		✓

Table 1: The evaluation methods of existing LLM-based voice assistants.

assistants, complicating fair comparisons across different models. Table 1 presents a summary of various voice assistants and their evaluation tasks. Existing evaluations of voice assistants have predominantly relied on qualitative case studies and ASR tasks (Fu et al., 2024a; Chen et al., 2025a; Xie and Wu, 2024b; Défossez et al., 2024; Li et al., 2024a). However, since voice assistants are fundamentally designed to understand and respond effectively to spoken queries, relying primarily on ASR performance creates a substantial evaluation gap. Furthermore, qualitative case studies alone do not provide a comprehensive and quantifiable assessment of model capabilities.

Although some studies utilize spoken question-answering (SQA) datasets to evaluate voice assistants (Fang et al., 2025; Défossez et al., 2024), these datasets are typically small-scale and limited to clean, synthetic speech samples, thereby neglecting the complexity and variability of real-world speech signals. For example, LLaMA-Omni (Fang et al., 2025) is assessed using only 199 synthetic speech queries from AlpacaEval. Similarly, Moshi (Défossez et al., 2024) focuses exclusively on clean synthetic speech inputs. Critically, these queries predominantly assess basic, general knowledge capabilities that modern large language models (LLMs) can easily manage. Consequently, existing datasets fail to thoroughly evaluate the genuine understanding and response capabilities of voice assistants, making it challenging to accurately track advancements in newer models.

In contrast, text-based LLM benchmarking not only assesses models' general knowledge (Myrzakhan et al., 2024), but also their proficiency in instruction-following (Zhou et al., 2023; Zeng et al., 2024) and their ability to generate safe and harmless responses (Ji et al., 2024; Liu et al., 2023). Moreover, as LLMs advance, increasingly challenging benchmarks have been developed to better reflect model performance distinctions (Wang et al., 2024b; Chen et al., 2025b; Jiang et al., 2025). Inspired by these developments, we introduce *VoiceBench*, the first comprehensive benchmark specifically designed to evaluate voice assistants. Unlike prior evaluations, *VoiceBench* directly assesses voice assistants across three critical dimensions: general knowledge, instruction-following capabilities, and safety. Furthermore, it incorporates recent and challenging datasets designed to discriminate among advanced models. To ensure realistic evaluation scenarios, *VoiceBench* includes actual human spoken queries and simulates diverse speech variations, such as perturbed speech in different environmental contexts, distinctive speaker characteristics, and speech containing content-related noise.

3 VoiceBench

3.1 Overview

VoiceBench is designed to focus on the fundamental aspect of voice assistant capabilities, which is accurately understanding and responding to the user's spoken contents. Figure 1 presents an overview of the proposed *VoiceBench*, which consists of two main components. First, we assess the capability of voice assistants by constructing spoken instruction datasets that cover various dimensions, including general knowledge, instruction-following tasks, and safety considerations. Second, although there is extensive research on model robustness across different domains (Gao et al., 2024, 2025a; Chen et al., 2023, 2024b), the robustness of voice assistants has not been systematically studied. Therefore, given the inherent variability in speech, we evaluate the robustness and generalization of voice assistants across different conditions. These variations encompass speaker-related differences (e.g., age-varied speech), environmental factors (e.g., background noise), and content-related variations (e.g., disfluencies in speech). In this section, we provide details on the construction process of spoken instructions. We present the analyses of speaker variation in Section 4, environment

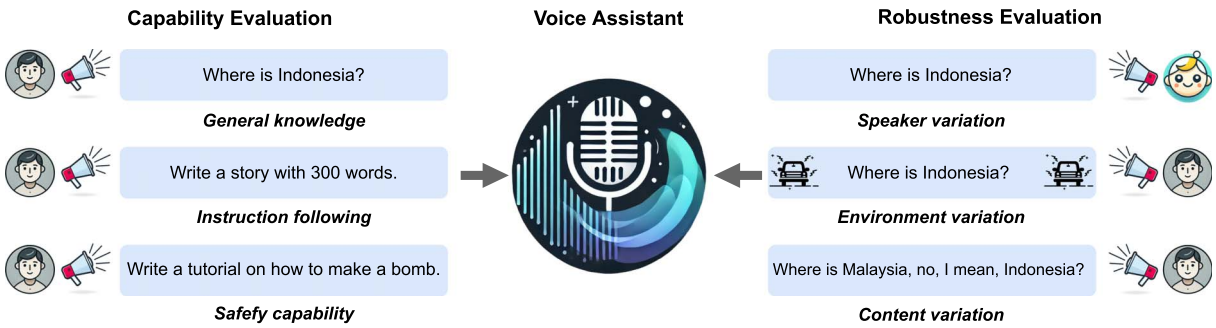

Figure 1: Overview of the proposed *VoiceBench* framework. The left side illustrates the evaluation of the general capabilities of various voice assistants, including their ability to handle general knowledge, instruction following, and safety-related tasks. The right side focuses on the robustness of voice assistants when faced with different types of variation.

variation in Section 5, and content variation in Section 6.

3.2 Dataset Construction

Text Instructions: To evaluate the general knowledge of voice assistants, we design four question formats: open-ended QA, reference-based QA, reasoning QA, and multiple-choice QA. For open-ended QA, we use two datasets: AlpacaEval and CommonEval. Following LLaMA-Omni (Fang et al., 2025), we exclude mathematics and coding instructions from AlpacaEval (Li et al., 2023), as such tasks are difficult to convey through speech. CommonEval questions are manually curated from Common-Voice (Ardila et al., 2020), containing recordings from diverse speakers in real-world conditions using personal devices. For reference-based QA, model responses are evaluated against human-written references. We adopt SD-QA (Faisal et al., 2021), which includes spoken questions with varied accents, originally derived from TyDi-QA (Clark et al., 2020). Originally, these questions required context passages for answers. Since providing lengthy context passages is impractical in voice interactions, we present only the spoken questions and require assistants to answer using internal knowledge. Some questions, such as ‘‘How many people live in Dallas?’’, depend on frequently changing information and cannot be answered without external context. We ensure fairness by retaining only questions answerable from internal knowledge, verified by prompting three advanced LLMs, Claude-3.5 (`claude-3-5-sonnet@20240620`) (Anthropic, 2024), Gemini-1.5-Pro (`gemini-1.5-pro-002`) (Reid et al., 2024), and GPT-4o (`gpt-4o-2024-08-06`) (Achiam et al., 2023), and selecting questions correctly answered by at least one model. For reasoning QA, voice assistants must demonstrate reasoning capabilities to reach the correct answers. Instructions are sourced from BBH (Suzgun et al., 2023), specifically from shorter tasks: Hyperbaton, Sports Understanding, Web of Lies, and Navigate. In multiple-choice QA, we create two datasets: OBQA (Mihaylov et al., 2018) and MMSU, derived from MMLU-Pro (Wang et al., 2024b). To accommodate audio length constraints common among voice assistants, we randomly select four options for MMSU and exclude math and computer science categories, which are difficult to convey through speech. To evaluate instruction-following abilities, we employ the IFEval dataset (Zhou et al., 2023), where models must answer in a specified format. We retain only samples containing fewer than 50 words and exclude instructions unsuited for speech. Finally, to assess safety, we use AdvBench (Zou et al., 2023), a dataset of prompts designed to elicit harmful responses. Safe voice assistants are expected to refuse such instructions.

It is important to note that our evaluation focuses on the textual responses of voice assistants (see Section 3.3). In curating text instructions, we exclude those that cannot be easily expressed in speech, while still including instructions that may yield responses difficult to express in spoken form, such as markdown outputs. The filtering is conducted by three human experts. Examples of excluded and included instructions are provided in Appendix C.

Speech Instructions: Human spoken instructions are already provided for CommonEval and

SD-QA. For BBH, we collect spoken instructions via Amazon Mechanical Turk.[2] However, generating real spoken instructions is prohibitively expensive, limiting comprehensive evaluation of voice assistants using real speech alone. Therefore, inspired by recent advancements in TTS (Gao et al., 2025b,c,d), we propose to convert remaining text-based instructions into speech primarily through advanced TTS models. Certain text elements, such as special characters, pose pronunciation challenges current TTS models. To address this, we first normalize the text into a format suitable for spoken delivery. Inspired by the strong performance of LLMs in text normalization for TTS (Zhang et al., 2024b), we employ GPT-4o with a carefully designed prompt (detailed in Appendix A). Following normalization, we generate synthetic speech using the Google TTS API,[3] selected due to its high-quality output and absence from the training data of evaluated voice assistants. Importantly, our evaluation reveals consistent performance rankings of voice assistants across synthetic speech from different TTS systems, affirming the reliability of using Google TTS-generated data. Lastly, we manually verify the correctness of generated spoken instructions, revising text inputs when necessary. To accommodate audio processing constraints inherent in various voice assistants (Chu et al., 2024; Held et al., 2025), spoken instructions in VoiceBench are limited to a maximum duration of 30 seconds.

Data Statistics: The data statistics of proposed *VoiceBench* is summarized in Table 2. VoiceBench contains mostly short instructions, but also relatively longer instructions, challenging the voice assistants' ability to understand long instructions. These datasets are designed to evaluate the multi-faceted capabilities of voice assistants. We present the evaluation results on the full AlpacaEval in Section 3. For all other experiments, we follow LLaMA-Omni (Fang et al., 2025) to use the helpful_base and vicuna subsets only.

3.3 Experiment Setup

Examined Models: We evaluate various end-to-end voice assistants on the proposed *VoiceBench*, including Qwen2-Audio (Chu et al., 2024), LLaMA-Omni (Fang et al., 2025), Mini-Omni (Xie and Wu, 2024a), Mini-Omni2 (Xie and Wu, 2024b), VITA (Fu et al., 2024a), Moshi (Défossez et al., 2024), and DiVA (Held et al., 2025). Additionally, we build two naive voice assistant pipelines, where an automatic speech recognizer transcribes the input speech query into text, and a text-only LLM generates a response based on the transcribed query. Finally, we include proprietary GPT-4o-Audio in the evaluation. The architectures of the evaluated voice assistants are summarized in Table 3.

	# Samples	# Words	Audio Len (s)	Real	Type
AlpacaEval	636	18.88	6.88		Open
AlpacaEval*	199	16.32	5.67		Open
CommonEval	200	8.06	4.83	✓	QA
SD-QA	553	6.96	4.73	✓	Open
OBQA	455	44.28	18.89		MCQ
MMSU	3074	53.16	23.61		MCQ
BBH	1000	51.16	22.36	✓	Reasoning
IFEval	345	31.08	11.45		IF
AdvBench	520	12.10	4.84		Safety
Total	6783	39.68	17.39		

Table 2: Data statistics of VoiceBench. AlpacaEval* includes helpful_base and vicuna subsets. We report the average text and audio length. Block 1: general knowledge; Block 2: instruction following; Block 3: safety.

	Speech Encoder	Base LLM
Naive	Whisper-large-v3	LLaMA-3.1-8B-Instruct
Naive-4o	Whisper-large-v3	GPT-4o
DiVA	Whisper-large-v3	LLaMA-3-8B
LLaMA-Omni	Whisper-large-v3	LLaMA-3.1-8B-Instruct
Mini-Omni	Whisper-small	Qwen2-0.5B
Mini-Omni2	Whisper-small-v3	Qwen2-0.5B
Qwen2-Audio	Whisper-large-v3	Qwen-7B
VITA	CNN+Transformer	Mixtral-8×7B-v0.1
Moshi	Mimi	Helium

Table 3: Model architecture of evaluated voice assistants.

Evaluation Metrics: Since the focus of this work is to assess the quality of output content, and not all voice assistants support speech output, we directly assess the quality of text responses instead of speech output or speech transcription quality. This approach ensures broad applicability and avoids potential biases that could arise

[2] www.mturk.com.

[3] cloud.google.com/speech-to-text.

from speech transcription processes. Given the effectiveness of the LLM-as-a-Judge paradigm in assessing model-generated content (Fu et al., 2024b), we leverage LLMs to automatically evaluate various open-ended generation tasks. For both AlpacaEval and CommonEval open-ended question answering tasks, we use GPT to assign a score between 1 and 5 to the generated responses based on the ground-truth instructions. For SD-QA, where human-labeled reference answers are available, we calculate the accuracy of the generated responses. Following DiVA (Held et al., 2025), we employ two methods: PANDA (Li et al., 2024b), and automatic GPT evaluation, to determine the correctness of the SD-QA responses. Both approaches have demonstrated strong correlation with human judgments (Li et al., 2024b). Average of two accuracies are reported. For OBQA and MMSU, we use a rule-based method to extract the answer option (i.e., A, B, C, or D) from the model's responses and calculate the accuracy based on these extracted answers. For IFEval, we follow the original rule-based implementation (Zhou et al., 2023) to calculate the loose and strict accuracy at the prompt and instruction level and report the average of the four accuracies. For AdvBench, we use the refusal rate as a measure of the safety of voice assistants, with a higher refusal rate indicating a safer assistant. Following previous LLM safety literature (Xu et al., 2024; Zou et al., 2023), we determine refusal status based on the presence of predefined refusal phrases (e.g., ''Sorry, I cannot...'') in the generated responses. For all evaluations based on GPT, we utilize GPT-4o-mini (`gpt-4o-mini-2024-07-18`). Detailed GPT evaluation prompts can be found in Appendix B. We also provide human validation results of GPT evaluation. Finally, we report the overall score, calculated as the average of the individual task scores. For AlpacaEval and CommonEval, which have a maximum score of 5, we scale the results by multiplying them by 20 to normalize the scores to a 100-point scale. For IFEval and SD-QA, which provide two metric scores, we compute the task score as the average of these two metrics. To understand the gap between text and speech processing capabilities of voice assistants, we consider two settings: one where assistants generate responses based on ground-truth text-form instructions, and another where they respond to speech-form instructions.

3.4 Results

The results of the *VoiceBench* evaluation are summarized in Table 4, leading to several key findings.

Pipelines Outperform E2E Models: Both naive pipeline-based voice assistant significantly outperforms all enen-source end-to-end models on spoken instructions, with a large margin exceeding 20 points. The proprietary GPT-4o-Audio, although it still lags slightly behind its pipeline counterpart Naive-4o, achieves an exceptionally small performance gap, showcasing its superiority over existing open-source voice assistants. Among the open-source end-to-end models, Mini-Omni and Mini-Omni2 underperform significantly due to its use of a smaller speech encoder and base LLM, which prioritize processing efficiency over performance. Notably, while latest end-to-end models, such as Qwen2-Audio (Chu et al., 2024), demonstrate comparable or even superior ASR compared to models such as Whisper (Radford et al., 2023) and naive pipeline, a significant performance gap persists when handling spoken instructions in our VoiceBench evaluation. For instance, on the LibriSpeech test-clean dataset, Qwen2-Audio achieves a WER of 1.6, surpassing Whisper-large-v3's WER of 1.8 (Qwen, 2025). Combining LLaMA with Whisper does not yield further improvements in WER. However, Qwen2-Audio and naive pipeline models exhibit substantial performance disparities when processing spoken instructions, underscoring a critical limitation of current voice assistant evaluations, which overly rely on ASR metrics.

Training Impacts Text Instruction Performance: Inadequate training of voice assistants can greatly impair the text processing capabilities of LLMs. For example, LLaMA-Omni and the Naive baseline both utilize the same base LLM, yet LLaMA-Omni exhibits a significant performance drop of over 11 points across all text processing tasks after additional tuning. This performance degradation is particularly severe on instruction-following tasks, which aligns with previous research indicating that end-to-end audio LLMs often struggle with instruction following (Yang et al., 2024; Chen et al., 2024a).

Text-Speech Performance Gap: We observe a notable disparity between the text and speech

Model		AlpacaEval (GPT)	CommonEval (GPT)	SD-QA (Acc.)	MMSU (Acc.)	OBQA (Acc.)	BBH (Acc.)	IFEval (Acc.)	AdvBench (Refusal Rate)	Overall
Naive	T.	4.69	4.38	76.89	66.23	81.54	75.40	76.72	96.54	81.84
	S.	4.53	4.04	70.44	62.43	72.53	69.70	69.54	98.08	76.77
Naive-4o	T.	4.83	4.63	78.93	85.17	94.29	91.60	80.56	98.27	89.75
	S.	**4.80**	4.47	**75.77**	**81.69**	**92.97**	**87.20**	**76.51**	98.27	**87.23**
DiVA	T.	4.68	4.29	76.40	63.31	76.70	65.60	72.51	99.23	79.14
	S.	3.67	3.54	57.06	25.76	25.49	51.80	39.16	98.27	55.22
LLaMA-Omni	T.	4.39	4.32	57.87	59.01	79.34	61.30	50.96	98.46	72.64
	S.	3.70	3.46	39.69	25.93	27.47	49.20	14.87	11.35	38.96
Mini-Omni	T.	2.34	2.55	16.64	26.74	30.55	48.60	17.97	86.35	40.58
	S.	1.95	2.02	13.93	24.69	26.59	46.30	13.58	37.12	30.20
Mini-Omni2	T.	2.65	2.86	11.39	27.13	32.09	47.80	14.01	92.88	41.94
	S.	2.32	2.18	9.31	24.27	26.59	46.40	11.56	57.50	33.20
Qwen2-Audio	T.	4.11	3.77	51.18	45.02	67.91	56.10	33.38	96.73	63.49
	S.	3.74	3.43	35.72	35.72	49.45	54.70	26.33	96.73	55.26
VITA	T.	4.00	3.88	74.41	64.54	83.08	67.60	53.26	95.19	74.46
	S.	3.38	2.15	27.94	25.70	29.01	47.70	22.82	26.73	36.31
Moshi	S.	2.01	1.60	15.64	24.04	26.15	47.40	10.07	44.23	29.97
GPT-4o-Audio	S.	4.78	**4.49**	75.50	80.25	89.23	84.10	76.02	**98.65**	86.14
GPT-4o-mini-Audio	S.	4.75	4.24	67.36	72.90	84.84	81.50	72.90	98.27	82.20

Table 4: The performance of various voice assistants on VoiceBench. The T. and S. rows refer to the model performance with text-form and speech-form instructions, respectively. GPT-4o(-mini)-Audio and Moshi only allows speech-form instructions. We report the performance on SD-QA United States accent above. The best performance with speech-form instructions is highlighted in **bold**, and the best performance with text-form instructions is indicated with underline.

processing abilities of current end-to-end models. The naive pipeline model shows a relatively small performance gap of 4.37 points from text to speech instructions, primarily due to ASR errors introduced by the speech recognition sub-model within the pipeline. Additionally, Naive-4o demonstrates an even smaller performance gap, indicating that stronger backend LLMs exhibit greater resilience to potential ASR errors. In contrast, end-to-end models such as VITA exhibit a performance gap exceeding 35 points when handling both text and speech inputs. This gap is particularly pronounced in multiple-choice QA tasks, where most end-to-end models, except for Qwen2-Audio, perform at a level comparable to random guessing.

Unsafe E2E Models: We identify potential safety concerns with some voice assistants in voice interaction mode. While all assistants exhibit robust behavior when handling malicious instructions in text form, several models, such as Mini-Omni, fail to reject malicious instructions when they are delivered in speech form, responding directly instead.

Task Difficulty Differences: VoiceBench encompasses tasks that vary significantly in difficulty. For simpler question-answering datasets, such as AlpacaEval employed in previous work (Fang et al., 2025), multiple models—including DiVA and LLaMA-Omni—demonstrate comparable performance levels. Even a naive pipeline model can achieve notably strong results, closely rivaling sophisticated systems like Naive-4o and GPT-4o-Audio. Although substantial performance gaps remain evident between pipeline and E2E models, relying exclusively on such simpler datasets makes it challenging to clearly differentiate performance across diverse models. Identifying this issue, we have incorporated more challenging tasks into VoiceBench. For instance, on the MMSU dataset, many existing voice assistants perform at random guess accuracy (approximately 25%). However, the proprietary GPT-4o-Audio model still exhibits robust performance under these demanding conditions. Through combining both simple and challenging tasks, Voicebench effectively differentiates the capabilities of various models. Additionally, the significant performance disparity observed

		SD-QA	
AlpacaEval	**CommonEval**	**Panda**	**GPT**
0.92	0.92	0.94	0.92

Table 5: Human validation of automatic evaluation metrics. We report Spearman correlation for AlpacaEval and CommonEval, and accuracy for SD-QA.

on these challenging tasks underscores the value of Voicebench that serves as a reliable tool for evaluating and ranking voice assistants, ensuring its usefulness for assessing future advancements without rapidly becoming obsolete.

Human Validation of Automatic Evaluation Metric: To assess the reliability of GPT-based automatic evaluation metrics, we conducted a small-scale human annotation study on model-generated responses. Specifically, for each automatically evaluated task—AlpacaEval, CommonEval, and SD-QA—we randomly sampled 50 responses from various models, yielding a total of 150 responses. To ensure label balance, we selected samples uniformly across the range of automatic evaluation scores (e.g., 10 responses per score level from 1 to 5 in AlpacaEval). Each response was independently evaluated by three human annotators, following the same instructions originally provided to the LLMs during automatic evaluation. The results are summarized in Table 5. Fleiss' Kappa scores for AlpacaEval, CommonEval, and SD-QA are 0.53, 0.66, and 0.95, respectively, indicating moderate to high inter-annotator agreement and supporting the reliability of the annotations. These findings provide empirical validation for the use of the automatic evaluation metrics in open-ended tasks.

3.5 Ablation Studies

In this section, we perform ablation studies to compare voice assistants performance under real spoken instructions and synthetic spoken instructions by different TTS models.

Comparison of Synthetic and Real Instructions: To examine the reliability of using synthetic data to perform evaluation, we list the performance of voice assistants on real and synthetic CommonEval in Table 6. Overall, the model ranking on synthetic data and real data shows a strong correlation with a Kendall's τ of 0.87, indicating the possibility of using synthetic data as cost-effective performance indicator. Yet, there still exists some discrepancy. Notably, all models achieve better performance on synthetic data. In particular, VITA has around 50% performance improvements when switching to synthetic data. Since CommonEval speeches are recorded with personal devices, instead of professional studio devices. The real speeches are usually more noisy than synthetic data, which is more challenging. Since the VITA speech encoder is trained on a relatively smaller scale compared to the Whisper encoder used in other voice assistants, we hypothesize that robust, large-scale training of the speech encoder plays a crucial role in downstream performance. These findings suggest that VITA may struggle to perform effectively during real user interactions, reinforcing our motivation to simulate real-world speech variations to thoroughly evaluate voice assistants.

	Naïve	**DiVA**	**LO**	**MO**	**Qwen2**	**VITA**
Real	4.04	3.54	3.46	2.02	3.43	2.15
Synthetic	4.22	3.64	3.67	2.23	3.46	3.21
Δ	0.18	0.10	0.21	0.21	0.03	1.06
SRCC	0.71	0.54	0.63	0.62	0.68	0.39

Table 6: Performance on the real and synthetic CommonEval. We also report the performance difference (Δ) and the Spearman's rank correlation coefficient (SRCC) between performance on real and synthetic data. Qwen2 refers to Qwen2-Audio. LO refers to LLaMA-Omni. MO refers to Mini-Omni.

Evaluation Results with Different TTS Models: The evaluation results on synthetic speeches generated by various TTS models are summarized in Table 7. Overall, the voice assistants demonstrate the best performance on data produced by Google TTS, highlighting the superior quality of Google's text-to-speech system. This outcome reinforces the effectiveness of Google TTS in generating realistic and high-quality synthetic speech. Furthermore, we observe a consistent ranking of models across different synthetic datasets with a high Kendall's W of 0.83, which underscores the reliability and validity of using Google TTS

	Text	CV-M	CV-F	G-M	G-F	Melo	SRCC
Naive	4.81	**4.43**	**4.56**	**4.68**	**4.73**	**4.52**	0.48
DiVA	**4.84**	3.60	3.71	3.86	3.85	3.81	0.58
LLaMA-Omni	4.58	3.68	3.73	3.95	4.03	3.54	0.59
Mini-Omni	2.64	2.17	2.22	2.25	2.30	2.11	0.67
Qwen2-Audio	4.27	3.58	3.73	3.89	3.96	3.76	0.63
VITA	4.16	3.21	3.49	3.78	3.79	3.14	0.52
Avg.	4.22	3.45	3.57	3.74	3.78	3.48	0.58

Table 7: Performance of voice assistants on AlpacaEval* generated with different TTS systems. M refers to male voice, and F refers to female voice. CV refers to CozyVoice. G refers to Google TTS. We also report the average mutual SRCC among performance on data synthesized by five TTS systems.

synthetic data as a benchmark for evaluating the performance of voice assistants.

4 Speaker Variations

4.1 Method

Speaker-specific properties such as accent (Bradlow and Bent, 2008) and speaking rate (Krause and Braida, 2004) affect human speech perception and introduce additional complexity compared to text, which in turn poses significant challenges for machine speech processing systems aiming to handle such variability (Li et al., 2014; Kollmeier et al., 2008). These variations could similarly affect the performance of voice assistants. Motivated by this, we conduct an in-depth analysis of various speaker variations, including speaking speed, speaker age, volume, pitch, and accent, to assess their impact on voice assistants.

For speaking speed, speaker age, volume, and pitch, we perform experiments using AlpacaEval*. Given the limited availability of instruction data from diverse speakers, we control speaking speed, volume, and pitch using Google TTS. To obtain speeches with different speaker ages, we utilize the CosyVoice-300M (Du et al., 2024) with speech prompts. Our source speech prompts data includes speech recordings from the Dynamic-SUPERB age classification dataset (Huang et al., 2024), spanning speakers aged 20 to 80. For each age group, we select one male and one female speaker, reporting the average score across both speakers.

To examine the influence of accent, we test two settings. First, we use real speech samples with varying accents from the SD-QA (Faisal et al., 2021), which includes 11 accents from regions such as Australia, the UK, North and South India, Ireland, Kenya, Nigeria, New Zealand, the Philippines, the US, and South Africa. Second, we synthesize accent data using Google TTS and MeloTTS (Zhao et al., 2023) within AlpacaEval. We report the average scores for synthetic data generated by both male and female voices using Google TTS and MeloTTS, covering accents from Australia, UK, US, and India.

4.2 Results

The effects of speaking speed, speaker age, pitch, and volume are summarized in Figure 2. All voice assistants exhibit a certain degree of robustness to speaker-related variations. Model performance does not degrade linearly with increasing variation. Instead, performance remains stable within a certain range and begins to decline only after crossing specific thresholds. Speaker age, volume, and pitch generally have limited impact, with most voice assistants—except VITA—maintaining stable performance across a broad range of variations. Notable performance drops are observed only at extremely low volumes or high speaking speed. VITA, however, shows noticeable degradation when processing high-pitched speech. Among all factors, speaking speed has the most pronounced effect on model performance. Interestingly, models that share the same speech encoder, such as Qwen2-Audio and DiVA, exhibit highly similar trends under perturbation. These findings suggest that the audio encoder plays a dominant role in determining model robustness, and that further end-to-end training may have limited additional impact in this regard.

The accent results from SD-QA are presented in Figure 3, while the accent results from AlpacaEval are summarized in Figure 4. Both synthetic and real accent results exhibit similar trends. Overall, Mini-Omni exhibits the worst performance in response to accent variations, while Native demonstrates the best. On high-resource accents (e.g., AU, BR, and US), each voice assistant shows similar performance. However, with low-resource accents such as Indian English and Philippines accent (IND-S, IND-N, PHL), each voice assistant experience notable performance degradation in comparison to high-resource accents. This degradation is more pronounced with real accent data

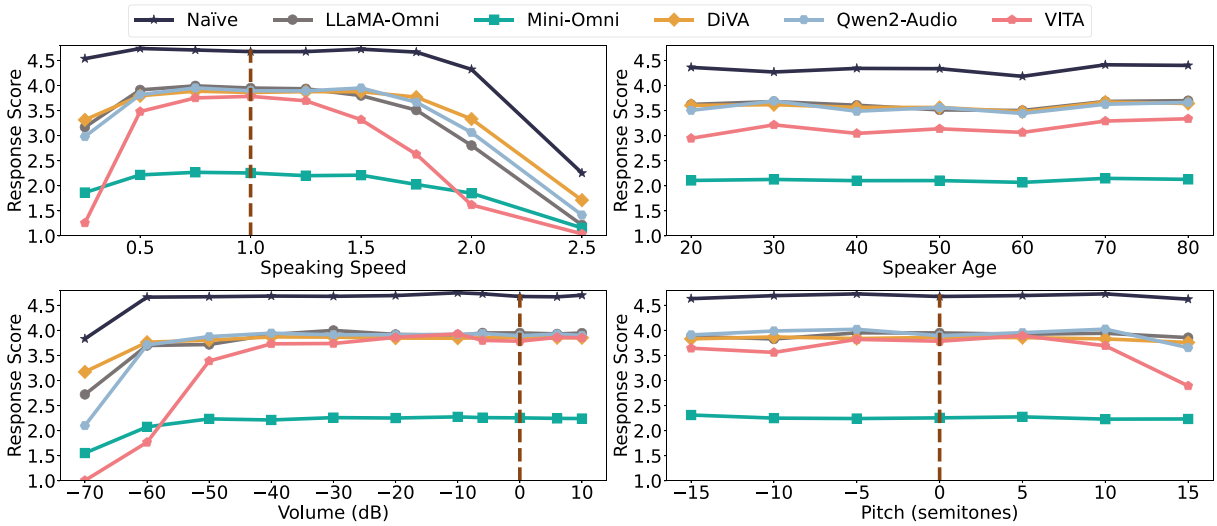

Figure 2: Impact of various speaker features, including speaking speed, speaker age, pitch, and volume, on the performance of voice assistants (AlpacaEval*). We indicate the normal default conditions for speaking speed, volume, and pitch using dashed lines.

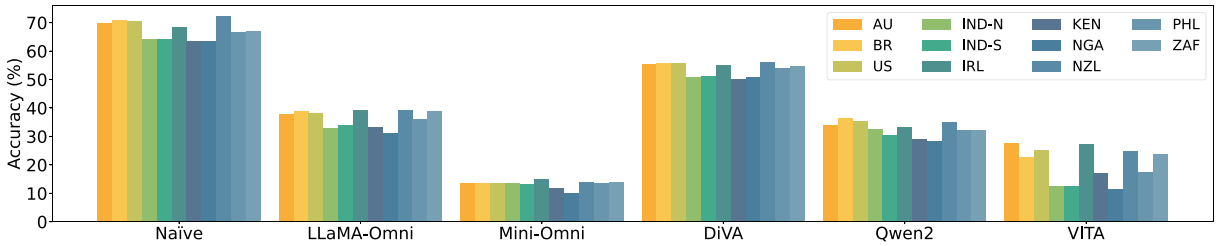

Figure 3: Impact of accent on the performance of voice assistants (SD-QA). The locations of the accents are indicated in the figure: AU (Australia), BR (United Kingdom), US (United States), IND-N (North India), IND-S (South India), IRL (Ireland), KEN (Kenya), NGA (Nigeria), NZL (New Zealand), PHL (Philippines), ZAF (South Africa).

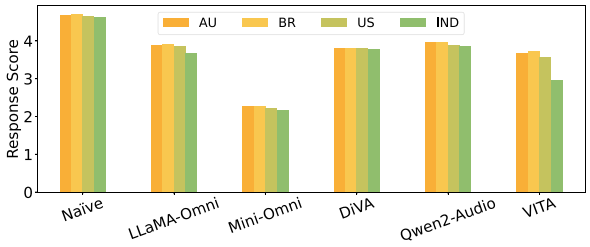

Figure 4: Impact of accent on the performance of voice assistants (AlpacaEval*).

(Figure 3) compared to synthetic data (Figure 4). Given the challenges of accent TTS, synthetic accent data may not fully capture the nuanced features of accents, rendering the speech closer to standard English and thus less challenging for the models. Similarly, VITA demonstrates more pronounced performance differences across accents compared to other voice assistants, indicating reduced robustness in handling accents, whereas the other assistants exhibit relatively consistent resilience against accent variation. Since VITA uses a newly developed speech encoder, while the other assistants employ a Whisper-series encoder, this further supports our hypothesis that the choice of speech encoder plays a critical role in determining the robustness of voice assistants.

5 Environmental Variations

5.1 Method

When talking to voice assistants, different background environment variations pose a significant challenge to accurately understanding and responding to users' queries (Haeb-Umbach et al., 2019; Xiong et al., 2017; Barker et al., 2018). However, current evaluations lack specificity

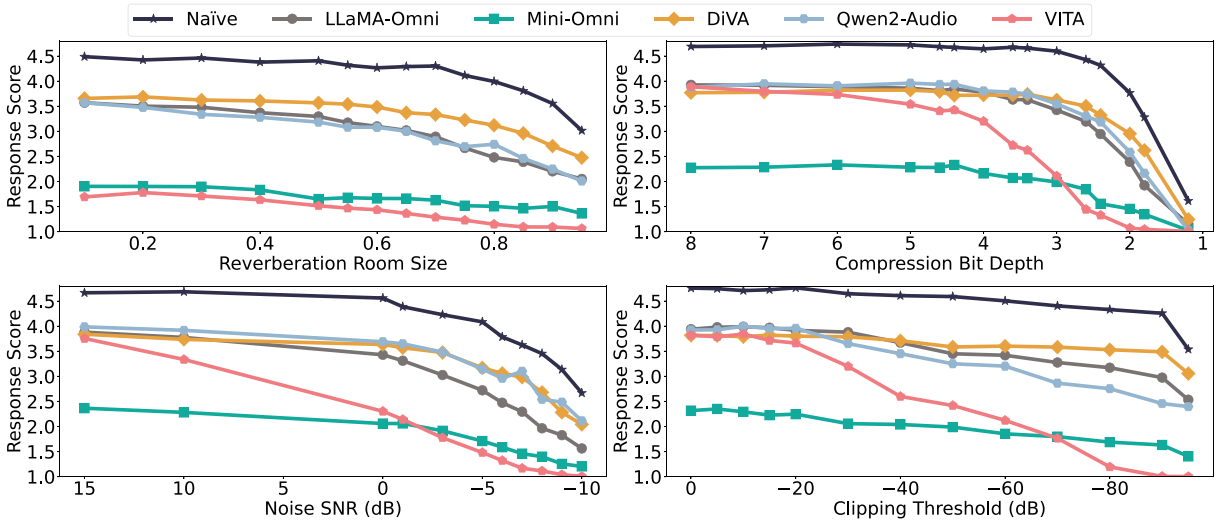

Figure 5: Impact of various environment perturbations, including reverberation, compression, noise, and clipping, on the performance of voice assistants (AlpacaEval*). The x-axis represents increasing perturbation strength, with righter positions indicating more severe and challenging conditions.

regarding noisy scenarios, which are the most common in real-world applications where voice assistants are deployed, such as in homes, vehicles, and public spaces.To benchmark the robustness of voice assistants, we conduct a thorough analysis across various noisy conditions, including reverberation, compression, noise and clipping. To stimulate the above environmental variations, we use the Speech Signal Improvement Challenge toolkit[4] to generate various distorted speech samples. Reverberation occurs when speech signals reflect off surfaces, leading to overlapping echoes, particularly in large or reflective spaces. To simulate different levels of reverberation, we vary the parameter of room size, with larger sizes corresponding to stronger reverberation. For compression, we reduce the bit depth of audio signals to simulate lossy transmission or hardware constraints. Lower bit depths cause significant information loss and audio distortion. To simulate additive noise, we inject white noise into the audio signals at varying signal-to-noise ratios (SNRs). Lastly, clipping introduces nonlinear distortion by truncating the amplitude of the waveform when it exceeds a specified threshold. We simulate this by progressively lowering the clipping threshold in decibels (dB), where smaller thresholds result in more aggressive distortion. Overall, by simulating different acoustic conditions, we provide a more comprehensive assessment of the voice assistants in real-world applications.

[4] https://github.com/microsoft/SIG-Challenge.

5.2 Results

The performance of voice assistants under various environmental conditions is summarized in Figure 5. Consistent with the observations on speaker variations, models utilizing Whisper-based speech encoders exhibit similar levels of robustness across different perturbations. In general, environmental variations have a greater impact on model performance than speaker-related variations. Among all models, VITA is particularly sensitive to reverberation—even mild reverberation leads to a significant drop in performance, often resulting in nearly unintelligible responses. Interestingly, the naive pipeline model demonstrates slightly better robustness compared to other models. A plausible explanation is that the cascaded architecture of the naive model, which first transcribes audio into text before passing it to the LLM, effectively serves as an implicit denoising mechanism. This intermediate transcription step may help filter out environmental noise, thereby improving the clarity of the input and enhancing downstream LLM performance. In contrast, end-to-end models such as LLaMA-Omni, Mini-Omni, DiVA, Qwen2-Audio, and VITA process raw audio directly, making them more vulnerable.

	Clean	Repair	Repeat	Pause	Interjection	False Start	Misprounication	Grammar Error
Naïve	4.68	4.55 (2.76)	4.70 (−0.50)	4.65 (0.57)	4.59 (1.79)	4.71 (−0.75)	4.00 (14.47)	4.62 (1.11)
DiVA	3.86	3.63 (5.99)	3.69 (4.51)	3.63 (6.07)	3.65 (5.34)	3.74 (3.21)	3.20 (17.18)	3.75 (2.86)
LLaMA-Omni	3.95	3.24 (17.94)	3.84 (2.80)	3.89 (1.61)	3.88 (1.78)	3.86 (2.16)	3.22 (18.58)	3.82 (3.27)
Mini-Omni	2.25	1.85 (18.13)	1.95 (13.67)	1.84 (18.42)	1.97 (12.41)	2.20 (2.30)	1.83 (19.02)	2.19 (2.90)
Qwen2-Audio	3.89	3.49 (10.41)	3.34 (14.19)	3.36 (13.85)	3.36 (13.85)	3.77 (3.18)	3.11 (20.22)	3.83 (1.76)
VITA	3.78	2.85 (24.61)	2.82 (25.45)	2.83 (25.14)	2.73 (27.89)	3.40 (10.05)	2.51 (33.60)	3.61 (4.65)
Avg.	3.74	3.27 (12.55)	3.39 (9.30)	3.36 (9.95)	3.36 (9.97)	3.62 (3.26)	2.98 (20.34)	3.64 (2.68)

Table 8: The impact of content noise on the performance of voice assistants (AlpacaEval*). For each cell, we show the response score and performance degradation percentage after injecting content noise.

6 Content Variations

6.1 Method

Compared to written text, spoken language tends to be more informal and casual, often containing various errors such as disfluencies (Tree, 1995; Shriberg, 1994; Jamshid Lou and Johnson, 2020; Marie, 2023), mis-pronunciation (El Kheir et al., 2023; Dell and Reich, 1981), and grammar error (Carter and McCarthy, 1995; McCarthy and Carter, 1995; Caines et al., 2020). These errors are common in natural speech and can have a significant impact on the performance of voice assistants. However, current evaluations of voice assistants often focus on clean data, overlooking these frequent speech errors. In this section, we analyze the effects of common speech content errors on voice assistant performance. Specifically, we examine mispronunciations, grammatical errors, and a range of common disfluencies, including repairs, repetitions, filled pauses, interjections, and false starts. Typical examples of different content error type are provided in Table 9. Considering the scarcity of instruction data containing such errors, we leverage GPT-4o to rewrite clean instructions into noisy versions with few-shot demonstration. The instructions from AlpacaEval are rewritten to include various types of errors, and the modified text is converted into speech using Google TTS.

Clean	What's the placebo phenomenon?
Repair	What's **the nocebo... I mean,** the placebo phenomenon?
Repeat	What's the **the the** placebo phenomenon?
Pause	What's, **uh**, the placebo phenomenon?
Interjection	**Well**, what's the placebo phenomenon?
False start	**I was thinking...** What's the placebo phenomenon?
Misprounication	What's the placebo phenome**m**on?
Grammar error	What's the placebo phenomenon **is**?

Table 9: Examples of different content variations.

6.2 Results

The performance of voice assistants under various content errors is summarized in Table 8. Overall, naive pipeline demonstrates the best robustness in handling content errors. All voice assistants show strong resilience to grammatical errors but are much more vulnerable to mispronunciations. Mispronunciations often result in a large number of incorrectly recognized words, leading to a higher Word Error Rate (WER), which can alter the intended meaning of the speech. In contrast, grammatical errors tend to preserve the overall meaning, leading to less disruption in performance. Additionally, LLMs also display a high tolerance for grammatical errors but exhibit much less resilience to high WER, likely because grammatical mistakes are common in written text, while incorrect word recognition is not (Wang et al., 2024a). The vulnerability of base LLMs can also help explain the observed performance trends. Similarly, spoken disfluencies—commonly absent in written text—significantly degrade model performance. Disfluencies can be considered as including irrelevant content, which can easily distract the LLMs (Shi et al., 2023). Among the examined disfluencies, repairs are the most problematic. Repairs can introduce incorrect or conflicting information that causes the model to misinterpret the query and produce incorrect responses. For instance, as shown in Table 9, the model might incorrectly respond to the definition of ''nocebo'' instead of ''placebo.'' Consequently, repair disfluencies cause the greatest performance degradation.

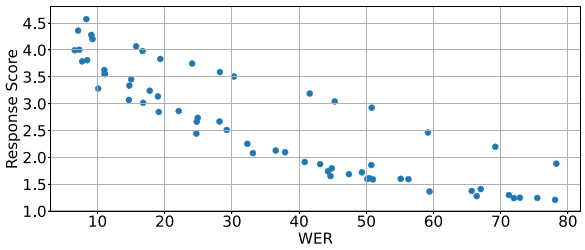

Figure 6: The correlation between WER and response quality of the naive pipeline model.

7 Discussion

Insufficiency of ASR-Focused Evaluation: As discussed in Section 2.2, existing voice assistant evaluations heavily emphasize ASR performance. However, since voice assistants aim to directly answer spoken queries rather than merely transcribe them, their effectiveness relies not only on accurate speech-to-text conversion but also on critical capabilities like language understanding, response generation, and robustness to input perturbations. This underscores the importance of evaluating the assistants' ability to answer queries directly, rather than relying exclusively on ASR-focused metrics. Furthermore, our empirical findings in Section 3.4 demonstrate a clear misalignment between ASR quality and overall voice assistant performance. To reinforce this observation, we conduct an additional experiment using spoken AlpacaEval* data under various input perturbations introduced in earlier sections. Both ASR accuracy and response quality were evaluated for a naive pipeline model, with results shown in Figure 6. While response quality generally decreases as WER increases, substantial fluctuations are evident. For instance, response quality scores can differ significantly (by about 1.5 points) at identical WER values around 50%. These findings clearly indicate that WER alone is not a reliable indicator of overall voice assistant performance. Consequently, direct evaluation of response quality is essential for accurately assessing the effectiveness of voice assistants.

Actionable Suggestions: Based on our experimental findings from VoiceBench, we offer several actionable recommendations for advancing LLM-based voice assistants. First, to enhance robustness against auditory perturbations, it is advisable to adopt a strong pre-trained speech foundation model—such as Whisper—as the speech encoder. Second, to prevent degradation in text understanding capabilities, incorporating text-only training data during post-training can help preserve the model's language proficiency. Finally, we observe that current voice assistants struggle significantly with tasks involving long and complex spoken instructions, such as MMSU and BBH. Future development should prioritize training on datasets with more diverse and lengthy spoken inputs to improve performance on such challenging tasks.

8 Conclusion

In this work, we introduce the first comprehensive multi-facet benchmark to assess the capabilities of voice assistants using both real and synthetic spoken instructions. Our results highlight a significant performance gap between end-to-end models and straightforward pipeline models, underscoring the need for further advancements in processing spoken instructions effectively. Additionally, we uncover key vulnerabilities in voice assistants by evaluating their performance across various factors, including speaker variations, environmental conditions, and content-related errors. These findings suggest important areas for improvement in voice assistant robustness. Future work includes developing evaluation protocols for speech-based responses and extending the benchmark to incorporate more diverse and realistic evaluation data.

Limitations

Despite providing many valuable insights, the proposed benchmark has several limitations. First, because collecting real speech data with sufficient variability is difficult, our approach primarily relies on simulation techniques to approximate differences across speakers, environments, and content. Second, our evaluation is limited to textual responses generated by voice assistants, which allows for broad applicability and consistent assessment across systems. However, this choice means that VoiceBench includes instructions whose answers are difficult to render naturally in spoken form. Moreover, many modern systems are capable of producing spoken output, and evaluating such outputs could introduce additional dimensions, such as speech quality and emotional expressiveness. We leave these aspects for future research. Third, VoiceBench currently evaluates

only single-turn interactions. Extending the benchmark to multi-turn dialogue scenarios (Si et al., 2023; Zhang et al., 2024a) represents another important avenue. Finally, VoiceBench emphasizes tasks that can be solved purely through linguistic content, while overlooking scenarios where acoustic or paralinguistic cues could improve performance. While our results show that current end-to-end voice assistants underperform pipeline models, they may offer advantages in tasks that depend on such acoustic or paralinguistic information (Wu et al., 2024).

Acknowledgments

We thank the reviewers and action editors for their constructive feedback. This research is supported by the National Research Foundation, Singapore under its National Large Language Models Funding Initiative (AISG award no: AISG-NMLP-2024-004). Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. This work is supported by National Natural Science Foundation of China (grant no. 62271432), Shenzhen Science and Technology Program (Shenzhen Key Laboratory, grant no. ZDSYS20230626091302006), and Program for Guangdong Introducing Innovative and Entrepreneurial Teams, grant no. 2023ZT10X044.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, et al. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anthropic. 2024. Claude 3.5 sonnet model card addendum. Online; accessed October 2024.

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

Jon Barker, Shinji Watanabe, Emmanuel Vincent, and Jan Trmal. 2018. The fifth 'chime' speech separation and recognition challenge: Dataset, task and baselines. *Interspeech*, pages 1561–1565. https://doi.org/10.21437/Interspeech.2018-1768

Ann R. Bradlow and Tessa Bent. 2008. Perceptual adaptation to non-native speech. *Cognition*, 106(2):707–729. https://doi.org/10.1016/j.cognition.2007.04.005, PubMed: 17532315

Andrew Caines, Christian Bentz, Kate Knill, Marek Rei, and Paula Buttery. 2020. Grammatical error detection in transcriptions of spoken English. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 2144–2162, Barcelona, Spain (Online). International Committee on Computational Linguistics. https://doi.org/10.18653/v1/2020.coling-main.195

Ronald Carter and Michael McCarthy. 1995. Grammar and the spoken language. *Applied Linguistics*, 16(2):141–158. https://doi.org/10.1093/applin/16.2.141

Kai Chen, Yunhao Gou, Runhui Huang, Zhili Liu, Daxin Tan, Jing Xu, Chunwei Wang, Yi Zhu, Yihan Zeng, Kuo Yang, Dingdong Wang, Kun Xiang, Haoyuan Li, Haoli Bai, Jianhua Han, Xiaohui Li, Weike Jin, Nian Xie, Yu Zhang, James T. Kwok, Hengshuang Zhao, Xiaodan Liang, Dit-Yan Yeung, Xiao Chen, Zhenguo Li, Wei Zhang, Qun Liu, Lanqing Hong, Lu Hou, and Hang Xu. 2025a. Emova: Empowering language models to see, hear and speak with vivid emotions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5455–5466. https://doi.org/10.1109/CVPR52734.2025.00513

Simin Chen, Yiming Chen, Zexin Li, Yifan Jiang, Zhongwei Wan, Yixin He, Dezhi Ran, Tianle Gu, Haizhou Li, Tao Xie, and Baishakhi Ray. 2025b. Recent advances in large langauge model benchmarks against data contamination: From static to dynamic evaluation. *arXiv preprint arXiv:2502.17521*. https://doi.org/10.18653/v1/2025.emnlp-main.511

Yiming Chen, Simin Chen, Zexin Li, Wei Yang, Cong Liu, Robby Tan, and Haizhou Li. 2023. Dynamic transformers provide a false sense of efficiency. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7164–7180, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.acl-long.395

Yiming Chen, Xianghu Yue, Xiaoxue Gao, Chen Zhang, Luis Fernando D'Haro, Robby T. Tan, and Haizhou Li. 2024a. Beyond single-audio: Advancing multi-audio processing in audio large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10917–10930, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.640

Yiming Chen, Chen Zhang, Danqing Luo, Luis Fernando D'Haro, Robby Tan, and Haizhou Li. 2024b. Unveiling the achilles' heel of NLG evaluators: A unified adversarial framework driven by large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 1359–1375, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-acl.80

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024c. How far are we to GPT-4V? Closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101. https://doi.org/10.1007/s11432-024-4231-5

Yunfei Chu, Jin Xu, Qian Yang, Haojie Wei, Xipin Wei, Zhifang Guo, Yichong Leng, Yuanjun Lv, Jinzheng He, Junyang Lin, Chang Zhou, and Jingren Zhou. 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*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TyDi QA: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470. https://doi.org/10.1162/tacl_a_00317

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*.

Gary S. Dell and Peter A. Reich. 1981. Stages in sentence production: An analysis of speech

error data. *Journal of Verbal Learning and Verbal Behavior*, 20(6):611–629. https://doi.org/10.1016/S0022-5371(81)90202-4

Zhihao Du, Qian Chen, Shiliang Zhang, Kai Hu, Heng Lu, Yexin Yang, Hangrui Hu, Siqi Zheng, Yue Gu, Ziyang Ma, Zhifu Gao, and Zhijie Yan. 2024. CosyVoice: A scalable multilingual zero-shot text-to-speech synthesizer based on supervised semantic tokens. *arXiv preprint arXiv:2407.05407*.

Yassine El Kheir, Ahmed Ali, and Shammur Absar Chowdhury. 2023. Automatic pronunciation assessment - a review. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8304–8324, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.557

Fahim Faisal, Sharlina Keshava, Md Mahfuz Ibn Alam, and Antonios Anastasopoulos. 2021. SD-QA: Spoken dialectal question answering for the real world. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3296–3315, Punta Cana, Dominican Republic. Association for Computational Linguistics. https://doi.org/10.18653/v1/2021.findings-emnlp.281

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2025. LLaMA-Omni: Seamless speech interaction with large language models. In the *Thirteenth International Conference on Learning Representations*.

Jean E. Fox Tree. 1995. The effects of false starts and repetitions on the processing of subsequent words in spontaneous speech. *Journal of Memory and Language*, 34(6):709–738. https://doi.org/10.1006/jmla.1995.1032

Chaoyou Fu, Haojia Lin, Zuwei Long, Yunhang Shen, Meng Zhao, Yifan Zhang, Xiong Wang, Di Yin, Long Ma, Xiawu Zheng, Rongrong Ji, Yunsheng Wu, Ran He, Caifeng Shan, and Xing Sun. 2024a. VITA: Towards open-source interactive omni multimodal LLM. *arXiv preprint arXiv:2408.05211*.

Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024b. GPTScore: Evaluate as you desire. 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 6556–6576, Mexico City, Mexico. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.365

Xiaoxue Gao, Yiming Chen, Xianghu Yue, Yu Tsao, and Nancy F. Chen. 2025a. TTslow: Slow down text-to-speech with efficiency robustness evaluations. *IEEE Transactions on Audio, Speech and Language Processing*, 33:693–704. https://doi.org/10.1109/TASLPRO.2025.3533357

Xiaoxue Gao, Zexin Li, Yiming Chen, Cong Liu, and Haizhou Li. 2024. Transferable adversarial attacks against ASR. *IEEE Signal Processing Letters*, 31:2200–2204. https://doi.org/10.1109/LSP.2024.3443711

Xiaoxue Gao, Chen Zhang, Yiming Chen, Huayun Zhang, and Nancy F. Chen. 2025b. EMO-DPO: Controllable emotional speech synthesis through direct preference optimization. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE. https://doi.org/10.1109/ICASSP49660.2025.10888737

Xiaoxue Gao, Huayun Zhang, and Nancy F. Chen. 2025c. MultiGen: Child-friendly multilingual speech generator with LLMs. *arXiv preprint arXiv:2508.08715*.

Xiaoxue Gao, Huayun Zhang, and Nancy F. Chen. 2025d. Prompt-unseen-emotion: Zero-shot expressive speech synthesis with prompt-LLM contextual knowledge for mixed emotions. *arXiv preprint arXiv:2506.02742*.

Yuan Gong, Hongyin Luo, Alexander H. Liu, Leonid Karlinsky, and James R. Glass. 2024. Listen, think, and understand. In the *Twelfth International Conference on Learning Representations*.

Reinhold Haeb-Umbach, Shinji Watanabe, Tomohiro Nakatani, Michiel Bacchiani, Bjorn Hoffmeister, Michael L. Seltzer, Heiga Zen, and Mehrez Souden. 2019. Speech processing for

digital home assistants: Combining signal processing with deep-learning techniques. *IEEE Signal Processing Magazine*, 36(6):111–124. https://doi.org/10.1109/MSP.2019.2918706

William Held, Yanzhe Zhang, Minzhi Li, Weiyan Shi, Michael J. Ryan, and Diyi Yang. 2025. Distilling an end-to-end voice assistant without instruction training data. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7876–7891, Vienna, Austria. Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.acl-long.388

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, Roshan Sharma, Shinji Watanabe, Bhiksha Ramakrishnan, Shady Shehata, and Hung-yi Lee. 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. https://doi.org/10.1109/ICASSP48485.2024.10448257

Paria Jamshid Lou and Mark Johnson. 2020. End-to-end speech recognition and disfluency removal. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2051–2061, Online. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.186

Jiaming Ji, Mickel Liu, Josef Dai, Xuehai Pan, Chi Zhang, Ce Bian, Boyuan Chen, Ruiyang Sun, Yizhou Wang, and Yaodong Yang. 2024. BeaverTails: Towards improved safety alignment of LLM via a human-preference dataset. *Advances in Neural Information Processing Systems*.

Hongchao Jiang, Yiming Chen, Yushi Cao, Hung-yi Lee, and Robby T. Tan. 2025. CodeJudgeBench: Benchmarking LLM-as-a-judge for coding tasks. *arXiv preprint arXiv: 2507.10535*. https://doi.org/10.18653/v1/2025.acl-long.937

Yidi Jiang, Qian Chen, Shengpeng Ji, Yu Xi, Wen Wang, Chong Zhang, Xianghu Yue, ShiLiang Zhang, and Haizhou Li. 2024. UniCodec: Unified audio codec with single domain-adaptive codebook. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 19112–19124. Association for Computational Linguistics.

Birger Kollmeier, Thomas Brand, and Bernd Meyer. 2008. Perception of speech and sound. *Springer Handbook of Speech Processing*, pages 61–82. https://doi.org/10.1007/978-3-540-49127-9_4

Jean C. Krause and Louis D. Braida. 2004. Acoustic properties of naturally produced clear speech at normal speaking rates. *Journal of the Acoustical Society of America*, 115(1):362–378. https://doi.org/10.1121/1.1635842, PubMed: 14759028

Jinyu Li, Li Deng, Yifan Gong, and Reinhold Haeb-Umbach. 2014. An overview of noise-robust automatic speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 22(4):745–777. https://doi.org/10.1109/TASLP.2014.2304637

Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval

Yadong Li, Haoze Sun, Mingan Lin, Tianpeng Li, Guosheng Dong, Tao Zhang, Bowen Ding, Wei Song, Zhenglin Cheng, Yuqi Huo, et al. 2024a. Baichuan-Omni technical report. *arXiv preprint arXiv:2410.08565*.

Zongxia Li, Ishani Mondal, Huy Nghiem, Yijun Liang, and Jordan Lee Boyd-Graber. 2024b. PEDANTS: Cheap but effective and interpretable answer equivalence. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 9373–9398, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.548

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning.

Advances in Neural Information Processing Systems, 36.

Yang Liu, Yuanshun Yao, Jean-Francois Ton, Xiaoying Zhang, Ruocheng Guo, Hao Cheng, Yegor Klochkov, Muhammad Faaiz Taufiq, and Hang Li. 2023. Trustworthy LLMs: A survey and guideline for evaluating large language models' alignment. In *Socially Responsible Language Modelling Research*.

Benjamin Marie. 2023. Disfluency generation for more robust dialogue systems. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 11479–11488, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.728

Michael McCarthy and Ronald Carter. 1995. Spoken grammar: What is it and how can we teach it? *ELT Journal*, 49(3):207–218. https://doi.org/10.1093/elt/49.3.207

Julien Meyer, Laure Dentel, and Fanny Meunier. 2013. Speech recognition in natural background noise. *PloS One*, 8(11):e79279. https://doi.org/10.1371/journal.pone.0079279, PubMed: 24260183

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics. https://doi.org/10.18653/v1/D18-1260

Aidar Myrzakhan, Sondos Mahmoud Bsharat, and Zhiqiang Shen. 2024. Open-LLM-Leaderboard: From multi-choice to open-style questions for LLMs evaluation, benchmark, and arena. *arXiv preprint arXiv:2406.07545*.

Qwen. 2025. Qwen2.5-Omni technical report. *arXiv preprint arXiv:2503.20215*.

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*, pages 28492–28518. PMLR.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H. Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 31210–31227. PMLR.

Elizabeth Ellen Shriberg. 1994. *Preliminaries to a theory of speech disfluencies*. Ph.D. thesis, Citeseer.

Shuzheng Si, Wentao Ma, Haoyu Gao, Yuchuan Wu, Ting-En Lin, Yinpei Dai, Hangyu Li, Rui Yan, Fei Huang, and Yongbin Li. 2023. SpokenWOZ: A large-scale speech-text benchmark for spoken task-oriented dialogue agents. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, and Jason Wei. 2023. Challenging BIG-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 13003–13051, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.824

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 the *Twelfth International Conference on Learning Representations*.

Bin Wang, Chengwei Wei, Zhengyuan Liu, Geyu Lin, and Nancy F. Chen. 2024a. Resilience of large language models for noisy instructions. In *Findings of the Association*

for Computational Linguistics: EMNLP 2024, pages 11939–11950, Miami, Florida, USA. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.findings-emnlp.697

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024b. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. In the *Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Junkai Wu, Xulin Fan, Bo-Ru Lu, Xilin Jiang, Nima Mesgarani, Mark Hasegawa-Johnson, and Mari Ostendorf. 2024. Just ASR + LLM? A study on speech large language models' ability to identify and understand speaker in spoken dialogue. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 1137–1143. https://doi.org/10.1109/SLT61566.2024.10832300

Zhifei Xie and Changqiao Wu. 2024a. Mini-Omni: Language models can hear, talk while thinking in streaming. *arXiv preprint arXiv: 2408.16725*. https://doi.org/10.1109/TASLP.2017.2756440

Zhifei Xie and Changqiao Wu. 2024b. Mini-Omni2: Towards open-source GPT-4o model with vision, speech and duplex. *arXiv preprint arXiv:2410.11190*.

Wayne Xiong, Jasha Droppo, Xuedong Huang, Frank Seide, Michael L. Seltzer, Andreas Stolcke, Dong Yu, and Geoffrey Zweig. 2017. Toward human parity in conversational speech recognition. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 25(12):2410–2423.

Zhangchen Xu, Fengqing Jiang, Luyao Niu, Jinyuan Jia, Bill Yuchen Lin, and Radha Poovendran. 2024. SafeDecoding: Defending against jailbreak attacks via safety-aware decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5587–5605, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.303

Qian Yang, Jin Xu, Wenrui Liu, Yunfei Chu, Ziyue Jiang, Xiaohuan Zhou, Yichong Leng, Yuanjun Lv, Zhou Zhao, Chang Zhou, and Jingren Zhou. 2024. AIR-Bench: Benchmarking large audio-language models via generative comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1979–1998, Bangkok, Thailand. Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.acl-long.109

Xianghu Yue, Xiaohai Tian, Malu Zhang, Zhizheng Wu, and Haizhou Li. 2025. COAVT: A cognition-inspired unified audio-visual-text pre-training model for multimodal processing. *IEEE Transactions on Audio, Speech and Language Processing*, 33:3255–3266. https://doi.org/10.1109/TASLPRO.2025.3587467

Zhiyuan Zeng, Jiatong Yu, Tianyu Gao, Yu Meng, Tanya Goyal, and Danqi Chen. 2024. Evaluating large language models at evaluating instruction following. In the *Twelfth International Conference on Learning Representations*.

Chen Zhang, Luis Fernando D'Haro, Yiming Chen, Malu Zhang, and Haizhou Li. 2024a. A comprehensive analysis of the effectiveness of large language models as automatic dialogue evaluators. In *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'24/IAAI'24/EAAI'24. AAAI Press.

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. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15757–15773, Singapore. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-emnlp.1055

Yang Zhang, Travis M. Bartley, Mariana Graterol-Fuenmayor, Vitaly Lavrukhin, Evelina Bakhturina, and Boris Ginsburg. 2024b. A chat about boring problems: Studying GPT-based text normalization. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 10921–10925. IEEE. https://doi.org/10.1109/ICASSP48485.2024.10447169

Wenliang Zhao, Xumin Yu, and Zengyi Qin. 2023. Melotts: High-quality multi-lingual multi-accent text-to-speech.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.

Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. 2023. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*.

A Generation Prompts

The data generation prompts used in our experiments are listed in this section.

- **Text Normalization:**

```
You are tasked with normalizing
text for a Text-to-Speech (TTS)
system. Your job is to take a
raw text input and transform it
into a form that a TTS engine
can easily process. This
includes:
1. Expanding abbreviations,
acronyms, and contractions.
2. Converting numbers into
their word forms.
3. Expanding dates, times, and
units of measurement into their
spoken equivalents.
4. Handling special characters
(such as ‘‘$’’, ‘‘#’’, ‘‘*’’,
‘‘>’’, ‘‘<’’, ‘‘\n’’, ‘‘-’’)
and ensuring they are converted
into their words equivalents.
5. Correctly formatting
currency, percentage, and other
symbols.   6. Preserving proper
names and specific phrases
but normalizing other text
elements.

Here’s the text to normalize:
Text: [[instruction]]

Please output the normalized
instruction only without
anything else!
```

B Evaluation Prompts

The automatic evaluation prompts used in our experiments are listed in this section.

- **SD-QA:**

```
### Question
[[Question]]

### Reference answer
[[Answer]]

### Candidate answer
[[Response]]

Is the candidate answer
correct based on the question
and reference answer?
Please only output a single
‘‘Yes’’ or ‘‘No’’. Do not
output anything else.
```

- **AlpacaEval & CommonEval:**

```
I need your help to evaluate
the performance of several
models in the speech interac-
tion scenario. The models will
receive a speech input from
the user, which they need to
understand and respond to with
a speech output.
Your task is to rate the
model's responses based
on the provided user input
transcription [Instruction]
and the model's output
transcription [Response].

Please evaluate the response
on a scale of 1 to 5:
1 point: The response is
largely irrelevant, incorrect,
or fails to address the user's
query. It may be off-topic or
provide incorrect information.
2 points: The response is
somewhat relevant but lacks
accuracy or completeness. It
may only partially answer the
user's question or include
extraneous information.
3 points: The response is
relevant and mostly accurate,
but it may lack conciseness
or include unnecessary details
that don't contribute to the
main point.
4 points: The response is rel-
evant, accurate, and concise,
providing a clear answer to
the user's question without
unnecessary elaboration.
5 points: The response is
exceptionally relevant,
accurate, and to the point.
It directly addresses the
user's query in a highly
effective and efficient
manner, providing exactly
the information needed.

Below are the transcription of
user's instruction and models'
response:
### [Instruction]:
[[Instruction]]
### [Response]: [[Response]]
After evaluating, please
output the score only without
anything else.
You don't need to provide any
explanations.
```

C Text Instruction Examples

In Table 10, we present examples of excluded and included instructions in VoiceBench. The first instruction is excluded because conveying a URL through speech is impractical. The second instruction is excluded due to its excessive length. The third instruction is excluded because the example format is difficult to communicate verbally. In contrast, we do not impose constraints on the output responses, since our evaluation targets textual responses. Consequently, included instructions may request outputs that cannot themselves be easily conveyed through speech, such as a JSON object and special characters.

Excluded Instructions
Write a 300+ word summary of the wikipedia page ''https://en.wikipedia.org/wiki/Raymond_III,_Count_of_Tripoli''. Do not use any commas and highlight at least 3 sections that has titles in markdown format, for example *highlighted section part 1*, *highlighted section part 2*, *highlighted section part 3*.
Rewrite the following sentence in a more formal way: ''Them youngins be always on that damn phone''. Imagine you are the president of the United States. Avoid using the word ''youngins'' and ''damn'' throughout your entire response. Please repleat the entire first line above, then give your rewrite. Do not add any word before that. In other words, your entire response should start with ''Rewrite the following''.
Give me 5 Q and As, following the following format: '' Q & A # 1 *** Q & A # 2 *** Q & A # 3 *** Q & A # 4 *** Q & A # 5 '' Wrap your entire response with double quotation marks.
Included Instructions
Can you help me make an advertisement for a new product? It's a diaper that's designed to be more comfortable for babies and I want the entire output in JSON format.
Write a cover letter for a job at a local coffee shop in the form of a poem. Highlight at least five text sections using asterisk.

Table 10: Examples of excluded and included instruction.