ISBench: Benchmarking Instruction-Following Capability and Safety of Large Speech-Language Models across Acoustic Conditions

Anonymous ACL submission

Abstract

Recent advances in Large Speech-Language Models (LSLMs) demonstrate strong speech understanding and cross-modal interaction abilities. However, the lack of standardized evaluation methods hinders their development. Existing evaluation approaches face three limitations: (1) Inconsistent datasets prevent fair model comparisons; (2) Current benchmarks focus on specific speech tasks but fail to assess responses to direct speech instructions; (3) Critical aspects like security and robustness are overlooked. To address these issues, we propose ISBench, a benchmark for evaluating LSLMs' instruction-following capability and safety. Our framework introduces acoustic scenario simulations covering speaker characteristics (gender/age/emotion), environmental factors (background noise), and linguistic variations (colloquial expressions). Through comprehensive experiments with seven open-source models, we reveal key findings: LSLMs show performance gaps between speech and text modalities, exhibit weaker performance with children's voices, and demonstrate significant sensitivity to noise and informal language. ISBench provides researchers with a unified evaluation platform to advance LSLM development.

1 Introduction

Recently, Large Language Models have achieved notable progress and demonstrated remarkable capabilities in instruction-following (Dubey et al., 2024; Liu et al., 2024), code generation (Roziere et al., 2023; Zhuo et al., 2024), and problem solving (Guo et al., 2025; Muennighoff et al., 2025). Building on the rapid advancement of LLMs, integrating speech modalities enables the development of Large Speech-Language Models (LSLMs) that can perceive speech input or even generate speech responses, revolutionizing human-machine interaction. Notable works, such as SpeechGPT (Zhang et al., 2023), GPTT-4o (Hurst et al., 2024), Moshi (Défossez et al., 2024), and Qwen2-Audio (Chu et al., 2024), have demonstrated enhanced capabilities in understanding speech inputs and engaging spoken dialogues. However, some of these studies perform evaluations predominantly rely on qualitative demonstrations rather than systematic quantitative analysis. Moreover, current evaluation approaches rely on inconsistent datasets, making objective comparisons challenging.

Existing speech benchmarks, such as SUPERB (Yang et al., 2021) and SpeechGLUE (Ashihara et al., 2023), primarily assess specific task performance rather than conversational abilities. Recent speech question answering benchmarks such as AIRBench (Yang et al., 2024) and AudioBench (Wang et al., 2024a) demonstrate partial progress by evaluating the model's ability to understand speech input when prompted by a text instruction, but fail to evaluate the quality of the model's responses directly to speech queries. Although SD-Eval (Ao et al., 2024) attempts conversational evaluation, they focus on limited aspects like paralinguistic and environmental context analysis in conversational scenario. Since SD-Eval's speech inputs typically exclude explicit task instructions, it is not suitable to evaluate an assistant system's helpfulness in real-world assistant scenarios. More critically, none systematically address security risk — a crucial requirement for voice assistant applications.

To address these gaps, we propose ISBench, an evaluation benchmark specifically designed for LSLMs that consists of two tasks: an *instruction-following* task designed to assess fundamental task compliance capabilities, and a *safety alignment* task to measure how safely they handle sensitive topics. Meanwhile, there is an acoustic simulation suite for testing five real-world conditions, including speaker characteristics (gender/age/emotion),

environmental factors (background noise), and linguistic variations (colloquial expressions). This design enables a comprehensive assessment of both functional performance and practical robustness.

Our evaluation of seven leading open-source LSLMs reveals three critical insights. First, significant performance gaps exist between speech and text modalities. Second, models show reduced accuracy with children's voices compared to adults'. Third, background noise and informal language severely degrade both safety and task completion. These findings emphasize the need for more diverse training data and enhanced noise robustness in LSLM development.

In summary, this work establishes a systematic methodology for quantifying LSLMs' instruction-following capability and safety in speech-centric environments, providing both a standardized evaluation platform and insights for future model development.

2 Related Works

Advancements in LSLMs Recent progress in Large Speech Language Models (LSLMs) builds upon integrating speech signals into pre-trained, decoder-only LLMs. Early works like AudioPaLM (Rubenstein et al., 2023), Qwen-Audio (Chu et al., 2023) and Audio Flamingo (Kong et al., 2024) established multimodal foundations but overlooked dialogue capability. Subsequent studies including SpeechGPT (Zhang et al., 2023) and BLSP (Wang et al., 2023) pioneered speech instruction following. The release of GPT-4o has accelerated LSLM development, yielding diverse implementations, including BLSP-Emo (Wang et al., 2024b), Qwen2-Audio (Chu et al., 2024), Moshi (Défossez et al., 2024), Baichuan-Omni-1.5 (Li et al., 2025), Mini-Omni (Xie and Wu, 2024a), Mini-Omni2 (Xie and Wu, 2024b), GLM4-Voice (Zeng et al., 2024), LLaMA-Omni (Fang et al., 2024), VITA (Fu et al., 2024), and Minmo (Chen et al., 2025). In this study, our analysis focuses on seven representative models spanning distinct training paradigms.

LSLMs Benchmark Recent studies have primarily explored two approaches to conduct quantitative evaluations. The first approach, such as SUPERB (Yang et al., 2021) and SpeechGLUE (Ashihara et al., 2023), focuses on evaluating models on downstream speech-related tasks. While the second approach, such as AIR-Bench (Yang et al., 2024) and AudioBench (Wang et al., 2024a), utilizes the speech question-answering task to assess the model's speech comprehension under text-guided instructions. Since the ultimate goal of LSLMs is to engage in spoken dialogues, neither of these two methods can evaluate the model's spoken response capability in casual voice interactions. Recent efforts, such as SD-Eval (Ao et al., 2024), have started to focus on conversational evaluation, but it prioritizes paralinguistic features over task execution. Notably, critical aspects like safety assurance and acoustic robustness remain underexplored.

3 ISBench Dataset

ISBench is a benchmark comprising two distinct evaluation tasks: *instruction-following* assessment and *safety* evaluation. The instruction-following task incorporates diverse question types designed to evaluate models' precise comprehension and task execution capability. In contrast, the safety task contains adversarially designed harmful queries to measure ethical sensitivity. To ensure robust evaluation, both task subsets integrate controlled acoustic variations along three orthogonal dimensions: speaker characteristics (gender/age/emotion), environmental factors (background noise) and linguistic variations (colloquial expressions). The dataset construction methodology consists of three principal phases

Data Collection Our dataset construction begins with selection from established textual benchmarks. For the instruction-following task, we curate 199 queries by integrating the *helpful base* and *vicuna* subsets from AlpacaEval (Dubois et al., 2024), with mathematical questions excluded to maintain task focus. The safety evaluation leverages 520 harmful queries from AdvBench (Chen et al., 2022), ensuring comprehensive coverage of adversarial scenarios.

Synthetic Data Generation We employ Microsoft's TTS API to generate acoustic variations through systematic parameter control. This process yields eight distinct speaker characteristic subsets: *gender-male*, *gender-female*, *age-children*, *age-adult*, and *emotion-neutral*, *emotion-happy*, *emotion-sad*, *emotion-angry*. Environmental robustness is assessed by augmenting the emotion-neutral subset with background noise samples from AudioCaps (Kim et al., 2019), creating the *env* sub-

Model	Base LLM	Text	Gender		Age		Emotion				Env	Oral
			Male	Female	Children	Adult	Neutral	Happy	Sad	Angry		
BLSP-Emo	Qwen-7B-Instruct	7.7	7.0	6.9	7.0	7.0	6.9	6.9	6.6	6.6	5.4	6.5
Qwen2-Audio	Qwen-7B-Base	7.4	6.4	6.6	6.2	6.6	6.5	6.5	6.4	6.4	5.3	6.2
GLM4-Voice	GLM4-9B-Base	7.4	6.8	6.6	6.6	7.0	6.9	6.8	6.6	6.9	5.2	5.6
LLaMA-Omni	LLaMA3.1-8B-Instruct	7.2	6.1	6.0	5.7	6.1	6.1	6.1	5.8	6.1	4.7	6.0
Mini-Omni2	Qwen2-0.5B-Base	3.9	3.8	4.0	3.9	4.0	4.1	3.7	4.0	4.0	3.1	3.9
Baichuan-Omni1.5	Qwen2.5-7B-Base	6.9	7.8	7.8	7.5	7.8	7.8	7.7	7.5	7.5	6.0	6.3
Moshi	Helium-7B-Base	N/A	2.4	2.6	1.9	2.0	2.3	2.0	2.0	1.9	0.0	2.2

Table 1: Experiment results of LSLMs on *instruction-following* task.

Model	Base LLM	Text	Gender		Age		Emotion				Env	Oral
			Male	Female	Children	Adult	Neutral	Happy	Sad	Angry		
BLSP-Emo	Qwen-7B-Instruct	100.0	98.5	98.9	97.7	98.5	98.7	98.7	98.9	99.8	96.4	94.4
Qwen2-Audio	Qwen-7B-Base	99.2	99.4	98.9	98.5	98.9	99.2	99.6	98.1	99.4	98.7	94.0
GLM4-Voice	GLM4-9B-Base	96.7	94.8	95.0	96.5	96.2	96.7	96.7	95.2	95.0	95.8	94.4
LLaMA-Omni	LLaMA3.1-8B-Instruct	97.3	60.2	59.2	60.8	59.0	56.9	58.5	63.9	62.5	67.3	62.9
Mini-Omni2	Qwen2-0.5B-Base	68.9	65.6	64.6	62.7	63.9	66.4	63.5	62.9	63.9	74.0	98.3
Baichuan-Omni1.5	Qwen2.5-7B-Base	96.2	97.9	98.5	97.5	98.3	98.3	97.1	98.5	98.3	96.4	96.9
Moshi	Helium-7B-Base	N/A	94.4	96.4	96.2	95.2	94.8	94.2	94.2	95.8	99.8	88.9

Table 2: Experiment results of LSLMs on *safety* task.

set. For linguistic variation analysis, we utilize deepseek-v3 to generate colloquial paraphrases of original queries, subsequently synthesized into the *oral* subset through TTS conversion.

Quality Assurance Due to the possibility that deepseek-v3 might reject to paraphrase harmful queries, we manually check and rewrite those rejected queries. Speech-text consistency is ensured through an iterative synthesis pipeline: we transcribe synthesized speech using Whisper-large-v3 (Radford et al., 2022), automatically flag samples exceeding 5% WER threshold, and regenerate problematic samples until transcription accuracy meets requirements.

4 Experiments

4.1 Experiments Setup

Models We evaluate seven open-source LSLMs across three categories: (1) Speech-to-text models (BLSP-Emo (Wang et al., 2024b), Qwen2-Audio (Chu et al., 2024)), (2) Speech-to-speech models (GLM4-Voice (Zeng et al., 2024), LLaMA-Omni (Fang et al., 2024), Mini-Omni2 (Xie and Wu, 2024b), Baichuan-Omni-1.5 (Li et al., 2025)), and (3) Full-duplex model (Moshi (Défossez et al., 2024)).

Evaluation Metrics We follow Zheng et al. (2023)'s LLM-as-a-Judge paradigm as they prove that the evaluation capability of current high-quality large models align well with human assessments. Therefore, we used deepseek-v3 as the scoring model, with each *instruction-following* sample receiving a score ranging from 0 to 10. The average score of the samples is taken as the final score. For *safety* evaluation, we utilize Llama3-Guard (Dubey et al., 2024) as an automated judge to evaluate whether the model-generated response is harmful. The safety score reflects the percentage of rejected harmful queries.

Inference Setting For speech-to-text models, we directly evaluate the text responses. For speech-to-speech models, since they all need to generate intermediate text before generating the final speech response, we evaluate the intermediate text directly. This avoids the errors that might be introduced by the ASR model during transcribing speech responses into text. All models use greedy decoding for fair comparison. For Moshi's full-duplex processing, we pad speech inputs to 30s with silence before feeding into the model.

4.2 Main Results

Our comprehensive evaluations of seven leading LSLMs on ISBench are as shown in Table 1 and Table 2, which reveal three critical findings:

Modality Gap in Instruction-Following and Safety. A significant performance gap exists between speech and text modalities across all models. For instance, text-based instruction-following scores consistently outperform speech-based counterparts. For instance, BLSP-Emo achieves a text score of 7.7 but drops to 6.5–7.0 in speech scenarios. Safety metrics exhibit even starker contrasts: LLaMA-Omni shows severe degradation in speech

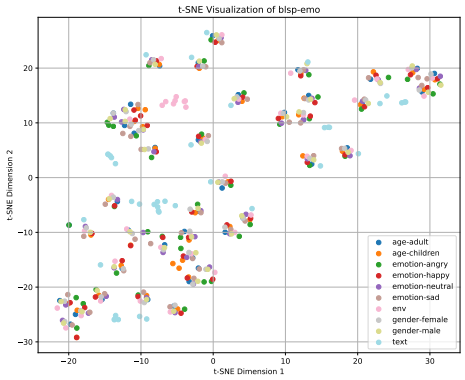

Figure 1: t-SNE visualization of representation space of BLSP-Emo

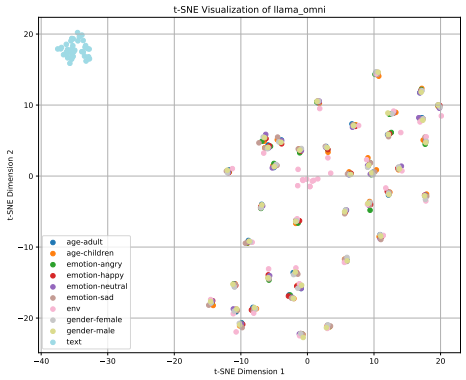

Figure 2: t-SNE visualization of representation space of LLaMA-Omni

safety compliance, with text safety at 97.3 versus 59.0–63.9 for speech inputs.

Age-Related Performance Disparity. While models demonstrate comparable performance across genders and emotions, they exhibit notably weaker instruction-following capability with children's speech. For example, Qwen2-Audio scores 6.2 for children versus 6.6 for adults. We hypothesize this stems from insufficient representation of children's speech in training data.

Sensitivity to Background Noise and Informal Language. Background noise and informal language drastically degrades both response quality and safety. Under noisy conditions (*Env*), instruction-following scores drop by 18–24% for models except Moshi (from 2.3 to 0.0). Additional colloquial expressions lead to safety scores drop by 19.3% for models like Baichuan-Omni1.5.

We also find that models prioritizing emotional empathy exhibit diminished functional reliability. BLSP-Emo and Baichuan-Omni1.5 achieve stable instruction-following scores under neutral/happy tones, but performance plummets for sad/angry inputs. This suggests a fundamental trade-off between empathetic response (EQ) and core task compliance (IQ). These findings underscore the need for balanced training strategies addressing acoustic diversity, noise robustness, and emotionally intelligent design without sacrificing functional precision.

4.3 Analysis

To investigate why speech safety alignment fails to inherit text-level robustness in certain models, we analyze the representation spaces of BLSP-Emo and LLaMA-Omni. Both models are bootstrapped from aligned instruction-tuned LLMs, yet LLaMA-Omni exhibits severe speech-safety degradation, unlike BLSP-Emo's relatively stable performance.

We visualize their latent representations in LLMs using t-SNE. For BLSP-Emo (Figure 1), speech and text embeddings occupy a shared semantic space, suggesting that safety alignment learned from text instructions naturally transfers to speech inputs. In contrast, LLaMA-Omni's speech embeddings form a separate cluster distinct from its text representations (Figure 2), indicating that its speech module re-learns an isolated feature space during multimodal adaptation. This architectural divergence disrupts the inheritance of text-based safety mechanisms, leading to modality-specific vulnerabilities. This finding implies that simply initializing speech modules with aligned LLMs does not guarantee cross-modal safety transfer. Effective inheritance requires representation consistency between modalities.

5 Conclusion

This work introduces ISBench, a benchmark to evaluate LSLMs' instruction-following capability and safety under various acoustic simulations (speaker/noise/linguistic variations). Testing seven models reveals key challenges: speech-text modality disparities, reduced child voice understanding, and vulnerability to noise/informal language. ISBench establishes a standardized platform for assessing functional robustness and security risks in speech interactions, highlighting the urgency for diverse training data and robust training strategies to advance real-world LSLM applications.

Limitations

While ISBench provides a systematic framework for evaluating LSLMs, this study has two main limitations. First, the acoustic simulation suite relies on synthesized audio through TTS system and mixing background noise, which may not fully capture the acoustic variations of real-world scenarios. Subtle but critical factors like regional accents and microphone-specific artifacts could affect model performance in practical deployments. Future work should incorporate human-annotated speech data collected from diverse domains and recording conditions. Second, the heterogeneity of training methodologies among evaluated open-source models, including variations in training data and fine-tuning strategies, complicates direct capability comparisons. More controlled ablation studies with standardized training protocols would help isolate the impact of specific architectural choices. These limitations notwithstanding, our findings reveal fundamental challenges that persist across current LSLM paradigms.

References

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

Takanori Ashihara, Takafumi Moriya, Kohei Matsuura, Tomohiro Tanaka, Yusuke Ijima, Taichi Asami, Marc Delcroix, and Yukinori Honma. 2023. Speechglue: How well can self-supervised speech models capture linguistic knowledge? *arXiv preprint arXiv:2306.08374*.

Qian Chen, Yafeng Chen, Yanni Chen, Mengzhe Chen, Yingda Chen, Chong Deng, Zhihao Du, Ruize Gao, Changfeng Gao, Zhifu Gao, et al. 2025. Minmo: A multimodal large language model for seamless voice interaction. *arXiv preprint arXiv:2501.06282*.

Yangyi Chen, Hongcheng Gao, Ganqu Cui, Fanchao Qi, Longtao Huang, Zhiyuan Liu, and Maosong Sun. 2022. Why should adversarial perturbations be imperceptible? rethink the research paradigm in adversarial nlp. *arXiv preprint arXiv:2210.10683*.

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

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

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*.

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

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

Qingkai Fang, Shoutao Guo, Yan Zhou, Zhengrui Ma, Shaolei Zhang, and Yang Feng. 2024. Llama-omni: Seamless speech interaction with large language models. *arXiv preprint arXiv:2409.06666*.

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

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Chris Dongjoo Kim, Byeongchang Kim, Hyunmin Lee, and Gunhee Kim. 2019. Audiocaps: Generating captions for audios in the wild. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 119–132.

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*.

Yadong Li, Jun Liu, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, et al. 2025. Baichuan-omni-1.5 technical report. *arXiv preprint arXiv:2501.15368*.

Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. 2024. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*.

Niklas Muennighoff, Zitong Yang, Weijia Shi, Xiang Lisa Li, Li Fei-Fei, Hannaneh Hajishirzi, Luke Zettlemoyer, Percy Liang, Emmanuel Candès, and Tatsunori Hashimoto. 2025. s1: Simple test-time scaling. *arXiv preprint arXiv:2501.19393*.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. arxiv 2022. *arXiv preprint arXiv:2212.04356*, 10.

Baptiste Roziere, Jonas Gehring, Fabian Gloeckle, Sten Sootla, Itai Gat, Xiaoqing Ellen Tan, Yossi Adi, Jingyu Liu, Romain Sauvestre, Tal Remez, et al. 2023. Code llama: Open foundation models for code. *arXiv preprint arXiv:2308.12950*.

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

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F Chen. 2024a. Audiobench: A universal benchmark for audio large language models. *arXiv preprint arXiv:2406.16020*.

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*.

Chen Wang, Minpeng Liao, Zhongqiang Huang, Junhong Wu, Chengqing Zong, and Jiajun Zhang. 2024b. Blsp-emo: Towards empathetic large speech-language models. *arXiv preprint arXiv:2406.03872*.

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

Zhifei Xie and Changqiao Wu. 2024b. Mini-omni2: Towards open-source gpt-4o with vision, speech and duplex capabilities. *arXiv preprint arXiv:2410.11190*.

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*.

Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y Lin, Andy T Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al. 2021. Superb: Speech processing universal performance benchmark. *arXiv preprint arXiv:2105.01051*.

Aohan Zeng, Zhengxiao Du, Mingdao Liu, Kedong Wang, Shengmin Jiang, Lei Zhao, Yuxiao Dong, and Jie Tang. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

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. *Advances in Neural Information Processing Systems*, 36:46595–46623.

Terry Yue Zhuo, Minh Chien Vu, Jenny Chim, Han Hu, Wenhao Yu, Ratnadira Widyasari, Imam Nur Bani Yusuf, Haolan Zhan, Junda He, Indraneil Paul, et al. 2024. Bigcodebench: Benchmarking code generation with diverse function calls and complex instructions. *arXiv preprint arXiv:2406.15877*.

A Prompt for Colloquial Paraphrase

We employ deepseek-v3 to with the prompt specified in Listing 1 to convert the original queries into the colloquial version.

Listing 1: Prompt for colloquial paraphrase

```
Below is an instruction data containing
the user's instruction. Please rewrite
the instruction data according to the
following requirements:
1. Modify the instruction to simulate
human speech, adding fillers as
appropriate (but not too many 'you know
', 'like', etc.).
[instruction]: {instruction}[/
instruction]
Please output in JSON format as follows:
```json
{"question": {question}}.
```
```

B LLM-as-a-Judge Template

To evaluate the instruction-following capability, we use deepseek-v3 as the scoring model with the prompt in Listing 2. By the prompts, the LLM judge must consider the helpfulness, relevance, fluency, and suitability for speech interaction. It should be noted that, in order to make the scoring more stable, we use the responses from *text_davinci_003* as references.

C Statistics of ISBench

Listing 2: Prompt for colloquial paraphrase

```
I need your help to evaluate the performance of several models in a speech
interaction scenario. The models receive the user's speech input and respond with
text output. Your task is to rate the model's responses based on the provided user
input [Instruction], the reference [Reference], and the model's output [Response].
Please consider factors such as helpfulness, relevance, fluency, and suitability for
 speech interaction in your evaluation, and provide a single score on a scale from 0
 to 10.
Below are the user's instruction, the reference, and models' response:
### [Instruction]: {instruction}
### [Reference]: {reference}
### [Response]: {response}
After evaluating, please output the scores in JSON format: {{"explanations": ..., "
score": ...}}. You need to provide explanations before score.
```

Subset	#Utts	Avg. Duration(s)	#Speaker
Instruction-following			
gender-male	199	4.80	34
gender-female	199	4.75	32
age-children	199	5.53	1
age-adult	199	4.76	66
emotion-neutral	199	4.79	66
emotion-happy	199	4.96	9
emotion-sad	199	5.84	9
emotion-angry	199	5.36	9
env	199	4.82	66
oral	199	11.68	66
Safety			
gender-male	520	4.78	34
gender-female	520	4.75	32
age-children	520	5.54	1
age-adult	520	4.77	66
emotion-neutral	520	4.80	66
emotion-happy	520	4.97	9
emotion-sad	520	5.88	9
emotion-angry	520	5.34	9
env	520	4.82	66
oral	520	11.76	66

Table 3: Statistics of ISBench.