VCB Bench: An Evaluation Benchmark for Audio-Grounded Large Language Model Conversational Agents

Anonymous ACL submission

Abstract

Recent advances in large audio language models (LALMs) have greatly enhanced multimodal conversational systems. However, existing benchmarks remain limited—they are mainly English-centric, rely on synthetic speech, and lack comprehensive, discriminative evaluation across multiple dimensions. To address these gaps, we present Voice Chat Bot Bench (VCB Bench), a high-quality Chinese benchmark built entirely on real human speech. VCB Bench evaluates LALMs from three complementary perspectives: knowledge understanding (general knowledge, reasoning, and daily dialogue), instruction following (including speech-level control beyond text commands), and robustness (stability under perturbations in content, environment, and speaker traits). Experiments on representative LALMs reveal notable performance gaps and highlight future directions for improvement. VCB Bench provides a reproducible and fine-grained evaluation framework, offering standardized methodology and practical insights for advancing Chinese voice conversational models.

1 Introduction

In recent years, large language models (LLMs) (Vaswani et al., 2017; Anil et al., 2023) have achieved remarkable progress in natural language understanding and generation. Integrating language modeling with modalities such as vision and audio (Radford et al., 2021; Singh et al., 2022) has further given rise to a new paradigm of multimodal learning. Within this trend, large audio language models (LALMs)—which combine speech signal processing with language modeling—have developed rapidly. Emerging systems such as StepAudio2 (Wu et al., 2025) and Qwen3-Omni (Xu et al., 2025b) demonstrate end-to-end (E2E) speech understanding and generation with capabilities in voice question answering, real-time conversation, and audio content analysis. Consequently, voice conversational agents powered by LALMs are drawing increasing academic and industrial attention, offering more natural and human-like interactions than text-only systems.

Despite these advances, moving from basic LALM functionalities to practical voice agents requires reliable and comprehensive evaluation tools. Such benchmarks are essential for diagnosing model weaknesses, guiding optimization, and enabling fair comparisons across systems. While initial efforts (Chen et al., 2024; Yang et al., 2024; Lin et al., 2025) have explored instruction following, audio understanding, reasoning, and dialogue scenarios, current evaluation practices remain limited in three major ways. First, most benchmarks are English-centric, leaving Chinese—the world's most widely spoken language—largely unexplored. Second, the majority rely on synthetic TTS data, which poorly reflects real-world acoustic variability. Third, many are text-derived benchmarks (e.g., AlpacaEval (Li et al., 2023), IFEval (Zhou et al., 2023)), whose formal and lengthy content is unsuitable for evaluating conversationally grounded LALMs that should generate natural, colloquial speech. Addressing these limitations is critical given China's large user base and the growing demand for practical, high-quality voice agents.

To bridge these gaps, we introduce Voice Chat Bot Bench (VCB Bench)—the first comprehensive evaluation framework for Chinese voice conversation, built entirely from authentic (non-synthetic) speech. VCB Bench evaluates LALMs along three complementary dimensions: (1) Knowledge, including multi-disciplinary general knowledge (12 subjects), mathematical and logical reasoning, daily dialogue comprehension, and story continuation for pretraining performance assessment; (2) Instruction following, extending beyond text-based prompts to incorporate speech-level control tasks such as adjusting volume, speed, and emotion, with bilingual (Chinese–English) support; (3)

Robustness, measuring model stability under real-world perturbations across content (mispronunciations, grammatical errors), environment (street, TV noise), and speaker characteristics (age, accents).

Our proposed VCB Bench is built entirely from authentic human recordings rather than synthetic TTS. It provides a large-scale, high-fidelity dataset covering diverse conversational scenarios and introduces a multi-dimensional evaluation framework that jointly measures knowledge understanding, instruction following, and robustness through fine-grained, reproducible tasks. Based on this benchmark, we conduct a systematic empirical analysis of state-of-the-art LALMs under unified settings, revealing their strengths and limitations in Chinese voice interaction and offering actionable insights for future model development.

2 Related Work

Large Audio Language Models. Recent LALMs primarily adopt an E2E audio–language modeling paradigm, integrating speech understanding and generation within a unified framework.

Qwen-Audio series (Chu et al., 2023, 2024; Xu et al., 2025a,b) progressively enhance cross-modal alignment and modeling efficiency. Qwen-Audio establishes robust audio–text alignment, Qwen-Audio2 improves encoding efficiency through multi-scale feature fusion, and the latest Qwen-Omni models introduce dual-core Thinker–Talker architectures and multi-codebook pretraining, achieving low-latency bilingual dialogue.

StepAudio models (Huang et al., 2025; Wu et al., 2025) focus on tightly coupling recognition and synthesis. StepAudio integrates a dual-codebook tokenizer and achieves a 3.2% WER, while StepAudio2 advances to a fully E2E design with fixed text–speech token alignment and Chain-of-Thought reasoning, improving fine-grained paralinguistic understanding.

Baichuan-Audio (Li et al., 2025) employs hierarchical RVQ discretization and dual audio heads to balance acoustic and linguistic objectives, enabling real-time bilingual communication. GLM4-Voice (Zeng et al., 2024) introduces a three-module structure (Tokenizer–Backbone–Decoder) supporting emotion and dialect modeling. Kimi-Audio (Ding et al., 2025) fuses continuous acoustic and discrete semantic tokens in a dual-head architecture, achieving low-latency, high-fidelity streaming generation.

Collectively, these models demonstrate rapid progress in unified audio–language modeling—covering tokenization, multi-modal fusion, and real-time dialogue. However, systematic evaluation benchmarks, especially for Chinese real-speech interaction, remain scarce. Current efforts primarily assess model architecture and performance qualitatively or on synthetic data, underscoring the necessity of a comprehensive, real-speech benchmark such as VCB Bench.

Audio Benchmarks. Recent efforts have introduced several benchmarks to evaluate LALMs from different perspectives. VoiceBench (Chen et al., 2024) assesses general knowledge, instruction adherence, safety, and robustness, mainly based on existing text datasets such as AlpacaEval and SD-QA (Faisal et al., 2021), with lengthy or highly complex samples removed. OpenAudioBench, released alongside Baichuan-Audio, integrates question-answering datasets including Spoken LLaMA Questions (Nachmani et al., 2023) and Web Questions (Berant et al., 2013), and augments them with a TTS-generated reasoning subset.

AIR Bench (Yang et al., 2024) contains two components—a basic benchmark covering emotion recognition, ASR, and age estimation, and a dialogue benchmark evaluating auditory understanding and internal knowledge. MMAU (Kumar et al., 2025) and MMAR (Ma et al., 2025) focus on deep audio reasoning, requiring multi-step inference grounded in internal audio knowledge. OmniBench (Li et al., 2024) targets omni-modal models handling audio, images, and text, where text queries are paired with multimodal contexts (speech, music, or sound) to test integrated reasoning. Finally, URO Bench (Yan et al., 2025) provides a bilingual (English–Chinese) comprehensive set that evaluates audio understanding, reasoning, and conversational ability—but the speech data are entirely synthetic (TTS-generated).

Overall, existing benchmarks have significantly advanced the evaluation coverage of LALMs, yet they share several limitations: (1) most rely heavily on TTS or synthetic speech, (2) they focus on English, and (3) their content often derives from text-centric QA corpora rather than spontaneous human dialogue. These gaps highlight the need for a real-speech, Chinese-oriented benchmark offering multi-dimensional evaluation—the central goal of our proposed VCB Bench.

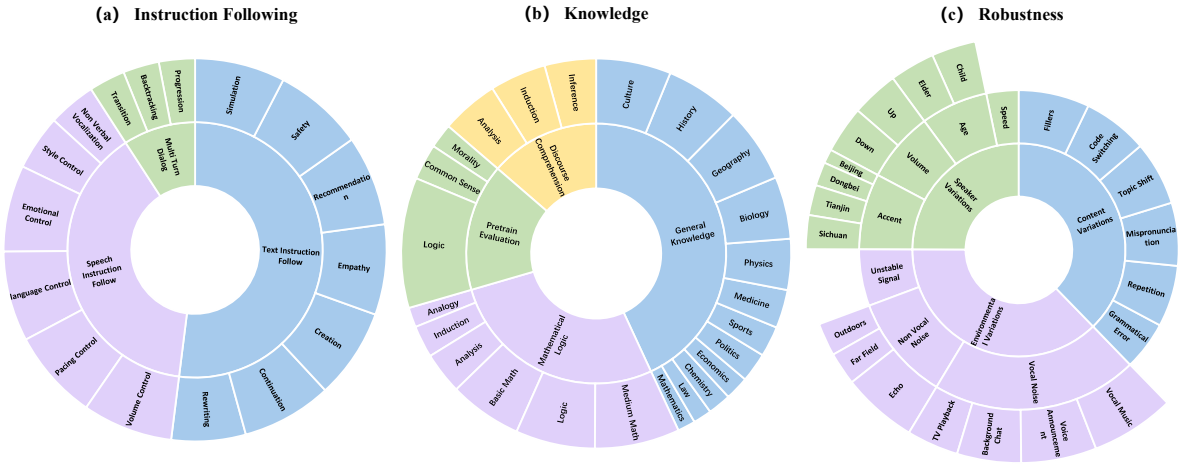

Figure 1: Overview of VCB Bench.

3 VCB Bench

As shown in the Figure 1, VCB Bench covers three core dimensions: Instruction Following, Knowledge, and Robustness. Instruction Following includes Text Instruction Follow (e.g., continuation, creation), Speech Instruction Follow (e.g., emotional, volume control), and multi-turn dialog tasks. The Knowledge module assesses general knowledge across 12 disciplines, mathematical logic, discourse comprehension, and story continuation (with fixed answers). Robustness introduces real-world perturbations from speaker variations, environmental noise, and content modifications to evaluate model stability.

3.1 Dataset Construction

The VCB Bench dataset integrates data from three distinct sources: third-party professional recordings, an internally curated two-person conversational dialogue dataset, and audio extracted from variety show Q&A segments. Each source supports different evaluation modules within the benchmark.

Third-Party Recorded Data. This category supports the Instruction Following, Mathematical Logic, and Story Continuation tasks under the Knowledge module, as well as the Robustness module. The production pipeline involves the following steps: First, task types and examples are defined through team discussion. Next, commissioned personnel manually compose texts that fulfill the task requirements. These texts then undergo manual inspection to ensure quality. Approved texts are forwarded to a third-party recording team for professional audio production. After recording, the data team performs quality checks on the audio. Subsequently, GPT-4o-Audio is used to evaluate audio quality, while GPT-4o assesses textual quality. Finally, manual screening is conducted to select high-quality samples, determining the final evaluation dataset.

General Knowledge Data. This subset is built from variety show Q&A audio. The process includes: crawling 20 hours of Q&A audio; manual annotation and segmentation for timestamps; ASR transcription; quality evaluation and subject classification via GPT-4O (assessing question clarity and answer accuracy); selection of Q&A pairs above a score threshold; and a final manual review for transcription accuracy, answer correctness, and audio clarity.

Discourse Comprehension Data. This subset is derived from an internal two-person dialogue dataset. The processing workflow involves: using GPT-4O to perform two-stage segmentation of the original long-form audio—first by dialogue topics (grouping segments discussing the same topic), then refining segmentation to ensure semantic coherence within each sub-segment. Through this process, dialogue audio segments shorter than one minute with maintained semantic coherence are obtained. GPT-4O then generates relevant question-answer pairs (e.g., for analysis or induction tasks) based on these dialogues. Finally, manual screening verifies question quality and answer accuracy.

Robustness Data. The data for this module is derived from TIF and SIF text materials. To control for speaker variability, the same speaker re-recorded the text under specified interference conditions (e.g., accent, noisy environment) wherever possible, using the original audio as a baseline.

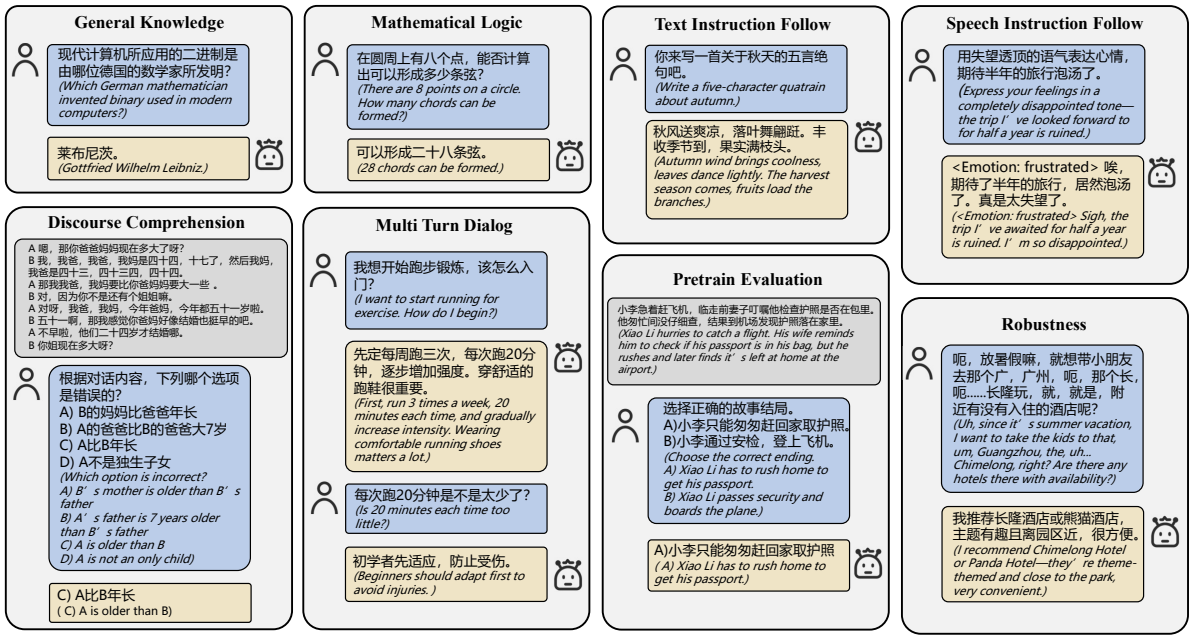

Figure 2: Examples from the VCB Bench.

For "content variation" types, the text was first modified (e.g., introducing grammatical or pronunciation errors) before being re-recorded by the same speaker. Additionally, to test performance in extreme scenarios, subsets like Volume, Speed, and Unstable Signal underwent post-processing.

3.2 Dataset Details

Instruction Following. The Instruction Following section comprehensively evaluates LALMs' ability to understand and execute both text and speech instructions, covering three sub-tasks: Text Instruction Following (TIF), Speech Instruction Following (SIF), and Multi-turn Dialog (MTD). All tasks are open-ended, and both TIF and SIF support Chinese and English to meet cross-lingual evaluation needs.

TIF assesses the model's ability to respond to textual instructions through seven sub-tasks, each examining text generation and semantic comprehension: (1) Continuation: extending a given text fragment to evaluate coherence and creativity; (2) Creation: generating original content based on a given theme to assess inventiveness and organization; (3) Empathy: understanding and responding to emotional expressions to examine affective perception; (4) Recommendation: providing suggestions based on user needs to evaluate information integration; (5) Rewriting: adapting text in style or structure to test reorganization ability; (6) Safety: identifying and rejecting harmful instructions to assess compliant response; (7) Simulation: role-playing in dialogue to examine contextual adaptation.

SIF focuses on understanding and executing speech instructions, particularly the ability to handle paralinguistic features such as emotion, speaking rate, and dialect. It includes six sub-tasks: (1) Emotional Control: adjusting the emotional tone of speech to assess expressive generation; (2) Language Control: switching languages or dialects to test multilingual synthesis; (3) Non-verbal Vocalization: incorporating non-linguistic elements like sighs or nasal sounds to evaluate paralinguistic expressiveness; (4) Pacing Control: modifying speaking rate to examine control precision; (5) Style Control: switching speech styles to assess style transfer; (6) Volume Control: adjusting loudness to test stability.

MTD evaluates instruction tracking and topic management in multi-turn dialogues, each containing 3–5 turns, focusing on contextual understanding and logical coherence: (1) Progression: deepening the discussion around an initial topic to assess topic development; (2) Backtracking: recalling and responding to previously mentioned information to test long-range memory; (3) Transition: smoothly shifting to a new topic to evaluate conversational flow and relevance.

Knowledge. The Knowledge module evaluates LALMs' knowledge storage, logical reasoning, and text comprehension through reference-based question answering. It comprises four sub-tasks: Gen-

Model	Instruction Following					Konwledge		
	TIF	TIF-En	SIF	SIF-En	MTD	GK	ML	DC
GLM4-Voice (Zeng et al., 2024)	85.82	82.52	85.57	**78.52**	85.13	45.53	62.14	48.64
Kimi-Audio (Ding et al., 2025)	85.13	88.92	**85.69**	61.87	85.67	53.51	79.94	74.76
Qwen2.5-Omni (Xu et al., 2025a)	87.40	72.58	71.37	58.09	86.93	55.43	80.24	73.72
Baichuan-Audio-Chat (Li et al., 2025)	72.49	76.22	78.96	68.07	73.27	44.48	60.33	54.38
Qwen2-Audio-Instruct (Chu et al., 2024)	84.56	75.86	/	/	85.67	35.83	60.78	67.07
StepAudio (Huang et al., 2025)	87.17	66.92	80.25	63.63	/	60.42	77.07	59.52
StepAudio2Mini (Wu et al., 2025)	82.79	75.54	78.81	**65.15**	**87.80**	61.15	81.30	83.08
Mimo-Audio	91.21	**91.76**	72.89	24.25	/	56.58	**84.01**	**87.92**
GPT4o-Audio	**91.24**	91.66	**88.15**	**86.07**	/	**61.29**	77.68	77.64

Table 1: The main results of different LALMs on VCB Bench.

eral Knowledge (GK), Mathematical Logic (ML), Discourse Comprehension (DC), and Story Continuation (SC).

GK tests multi-disciplinary common sense, covering mathematics, geography, politics, chemistry, biology, law, physics, history, medicine, economics, sports, and culture, measuring the model's ability to recall and apply cross-domain knowledge.

ML includes mathematical operations and logical reasoning. Mathematics comprises Basic Math (arithmetic within 100) and Medium Math (equations, geometry), evaluating computational and problem-solving skills. Logical reasoning covers Analysis (information decomposition), Induction (pattern generalization), Analogy (relational matching), and Logic (conditional reasoning), testing analytical and deductive abilities.

DC focuses on dialogue understanding through three tasks: Analysis (detecting factual accuracy in dialogues), Induction (summarizing dialogue themes), and Inference (deriving speakers' attitudes, emotions, or intents), assessing comprehension and implicit reasoning.

SC, inspired by StoryCloze (Mostafazadeh et al., 2016), evaluates implicit reasoning by having the model select the correct ending from options (all in audio and text) based on context. Categories include: Logic and Causality (causal consistency), Common Sense and Science (real-world and scientific knowledge), and Morality and Emotion (moral and emotional coherence).

Robustness. The Robustness module evaluates the stability of LALMs' performance under real-world interference conditions, ensuring reliable responses in challenging scenarios. The module encompasses three dimensions: Speaker Variations, Environmental Variations, and Content Variations.

Speaker Variations examine model adaptation to speaker attributes: (1) Age: utilizes child and elderly speech to assess recognition of age-related vocal characteristics; (2) Accent: incorporates four regional accents (Tianjin, Beijing, Dongbei, Sichuan) to evaluate comprehension of non-standard Mandarin; (3) Volume: assesses perception stability with amplified/attenuated speech; (4) Speed: tests parsing capability with rapidly delivered input.

Environmental Variations simulate acoustic interference: (1) Non-Vocal Noise: includes echo, outdoor, and far-field noise; (2) Vocal Noise: contains television audio, background conversations, vocal music, and radio broadcasts; (3) Unstable Signal: emulates network-induced packet loss to evaluate handling of fragmented audio.

Content Variations introduce linguistic disruptions: (1) Fillers: incorporate discourse markers (e.g., "um", "ah"); (2) Repetition: include repeated phrases/words; (3) Mispronunciation: introduce phonetic deviations; (4) Grammatical Error: employ ungrammatical constructions; (5) Topic Shift: implement abrupt topic changes; (6) Code-Switching: mix Chinese and English. Each category evaluates the model's ability to maintain comprehension despite content imperfections.

4 Experiment

4.1 Configuration

We evaluate the most capable (LALMs available, including GLM4-Voice, Kimi-Audio, Qwen2.5-Omni, Baichuan-Audio-Chat, Qwen2-Audio-Instruct, StepAudio, StepAudio2Mini, Mimo-Audio, and GPT4o-Audio.

For the SIF tasks in both Chinese and English (SIF-En), we invoke each model's "audio2audio" API to generate spoken responses. The adherence to instruction requirements is then automatically scored using GPT-4o-audio.

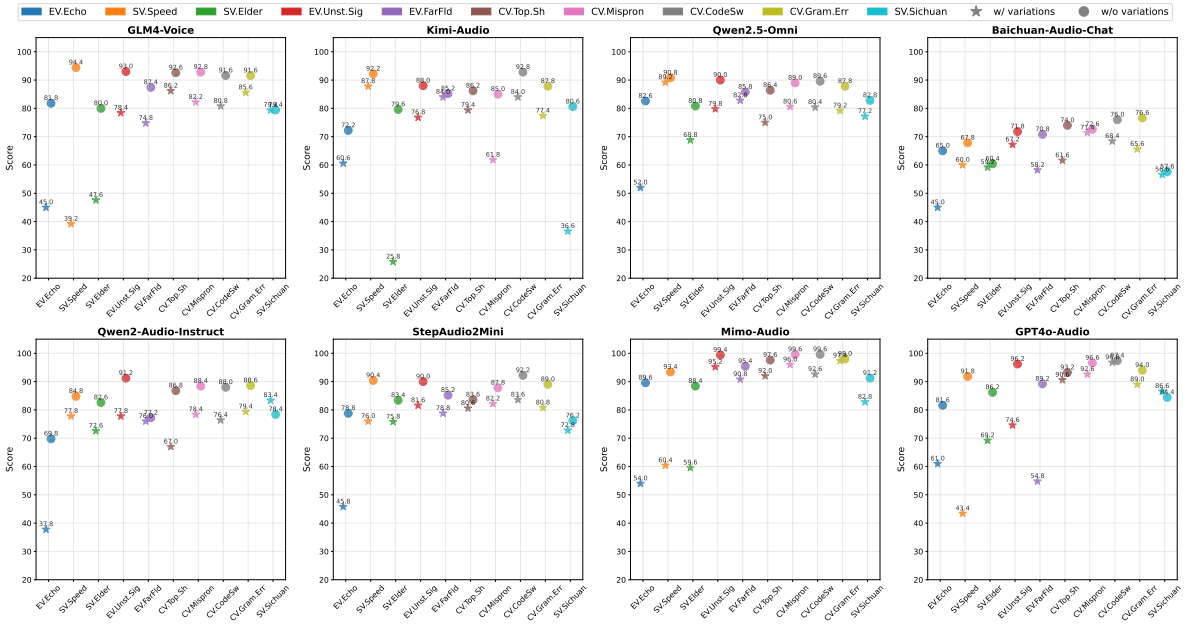

Figure 3: The robustness of LALMs under real-world perturbations. 9 subsets with the most significant performance gaps compared to the control group on the Robustness dataset are chosen.

For other tasks except SC, we call the "audio2text" API to obtain textual responses, which are evaluated by GPT-4O. In open-ended question answering, GPT-4O provides a numerical score, while for reference-based QA, it returns a binary "Yes" or "No" judgment.

For the SC task, we assess a subset of pre-trained base models: Baichuan-Audio-Base, Kimi-Audio-Base, Qwen2-Audio-Base, and StepAudio2Mini-Base. Following the StoryCloze evaluation protocol, we compute the negative log-likelihood for both the correct and incorrect endings, with model selection determined by comparing these two values.

For SIF tasks, the top six performing models undergo further Mean Opinion Score (MOS) evaluation. We sample the first 30 items from each relevant dataset, and eight expert evaluators rate the generated audio samples.

For the MTD evaluation, the model receives input context solely in audio format. We adopt the Bai et al. (2024)'s protocol, requiring the model to answer each turn in a dialogue sequence using the original ground-truth context from the dataset, not its own prior responses. A key distinction in our scoring is the heightened focus on the final turn: it carries 50% of the total score for a data sample, with the first several turns collectively accounting for the other 50%. All experiments are conducted on H20.

4.2 Main Result

As shown in Table 1, GPT-4o-Audio, as a non-open-source state-of-the-art model, serves as a strong baseline and demonstrates all-round superiority across most tasks, which is reasonable given its non-open nature and advanced proprietary capabilities. Focusing on other open-source E2E LALMs, notable performance discrepancies emerge:

In Instruction Following, Mimo-Audio excels in TIF/TIF-En with scores close to GPT-4o-Audio, indicating strong cross-lingual text instruction adaptation. For SIF, StepAudio2Mini and Kimi-Audio perform robustly in Chinese SIF, yet their SIF-En scores lag significantly behind, reflecting challenges in handling English speech's paralinguistic features. In MTD, StepAudio2Mini leads among open-source models with 87.80, outperforming counterparts like Baichuan-Audio-Chat, which highlights divergence in long-context dialogue logic control.

In Knowledge, Mimo-Audio stands out in ML (84.01) and DC (87.92), surpassing other open-source models and approaching GPT-4o-Audio's performance—suggesting strengths in deep reasoning and text semantic analysis. However, models like Baichuan-Audio-Chat show limited performance in GK (44.48) and ML (60.33), revealing gaps in multi-disciplinary knowledge coverage and step-by-step reasoning.

Overall, open-source E2E LALMs exhibit task-

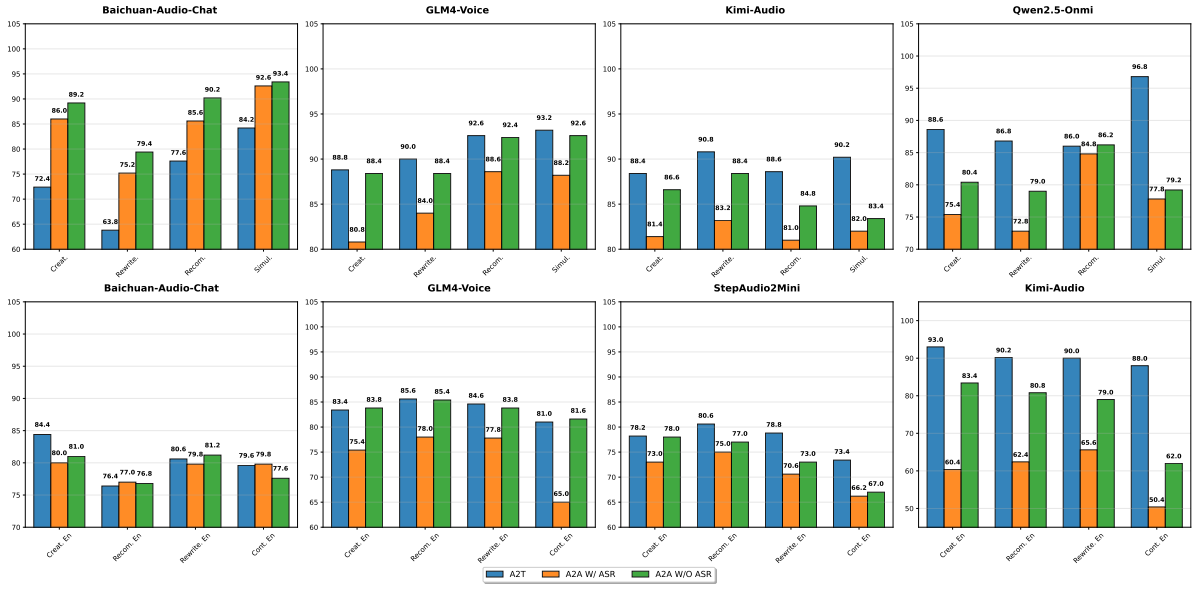

Figure 4: The investigation of the text-speech alignment capability of LALMs. A2T (Audio-to-Text, generating text directly from audio input), A2A W/ ASR (Audio-to-Audio, transcribing the generated audio via Automatic Speech Recognition to text), and A2A W/O ASR (evaluating the generated audio's text without ASR). All test sets are from TIF/TIF-En (e.g., "Creat." for Creation, "Recom." for Recommendation).

specific strengths (e.g., Mimo-Audio in reasoning, StepAudio2Mini in Chinese dialogue) but face challenges in cross-lingual speech adaptation and comprehensive knowledge reasoning.

4.3 Real Scenario

As shown in Figure 3, SV.Speed and EV.Echo cause the most severe performance degradation for most models. Scores of some models drop from over 80 in the control group to below 40 in these subsets, indicating that speech rate variation and acoustic echo are the most challenging perturbations for current LALMs. SV.Sichuan also poses a notable challenge. Except for Mimo-Audio and GPT-4o-Audio, most open-source models score significantly lower in this subset than in the control group, reflecting the constraint of "non-standard dialect speech" on cross-modal understanding. CV-related interferences (e.g., CV.Gram.Err, CV.Mispron) have relatively mild impacts. Some models (e.g., Mimo-Audio, StepAudio2Mini) show small score gaps between these subsets and the control group, suggesting models are more tolerant of "content-level flaws" than "speech/environment-level physical perturbations".

According the result, GPT-4o-Audio's scores in all interference subsets are close to those of the control group, demonstrating extremely strong robustness and verifying its leading capability in cross-modal anti-interference. Among open-source models, Mimo-Audio and StepAudio2Mini exhibit relatively prominent robustness. In contrast, models like Baichuan-Audio-Chat and Qwen2.5-Omni score significantly lower than the control group in most interference subsets, with large score fluctuations across different perturbations. This indicates obvious shortcomings in their robustness and insufficient adaptability to real-scene interferences.

4.4 Pretraining Evaluation

Model	Task	Metrics			
		Avg.	Logic	Moral	Common Sense
Baichuan-Audio-Base	S -> T	52.36	54.41	32.65	58.33
	S -> S	25.39	20.69	40.82	31.94
Kimi-Audio-Base	S -> T	**78.01**	**76.25**	**73.47**	**87.50**
	S -> S	**54.71**	**49.42**	**69.39**	**63.89**
Qwen2-Audio-Base	S -> T	48.95	51.72	30.61	51.39
	S -> S	36.91	36.78	44.90	31.94
StepAudio2Mini-Base	S -> T	50.26	52.87	26.53	56.94
	S -> S	30.63	27.55	34.69	38.89

Table 2: Pretraining Evaluation Results on SC.

The SC task evaluates pre-trained LALMs' "intelligence" and cross-modal semantic coherence by judging the rationality of story endings. The results are shown in the Table 1, Kimi-Audio-Base outperforms others in both paradigms: It scores an average of 78.01 in S->T and 54.71 in S->S, with robust performance across sub-dimensions, demonstrating stable story understanding and end-

ing judgment in cross-modal scenarios. In contrast, Baichuan-Audio-Base, Qwen2-Audio-Base, and StepAudio2Mini-Base score much lower. Moreover, all models perform worse in S->S than S->T, revealing that cross-modal (speech-to-speech) story coherence judgment remains challenging for pre-trained LALMs, with notable room for improvement in semantic consistency and rationality generation during speech output.

4.5 Ablation Study

4.5.1 Text-Speech Alignment

To investigate the text-speech alignment capability of LALMs, we conduct an ablation study, which is shown on Table 4. The visualization is based on two selection criteria from TIF and TIF-En: the four models with the highest A2A W/ ASR scores, and the four datasets with the largest mean score differences between A2A W/ ASR and A2T. Results for Chinese and English tasks are plotted separately in the upper and lower sections of the figure, respectively.

From the results, models like GLM4-Voice (Chinese) and Kimi-Audio (Chinese) demonstrate strong text-speech alignment—their A2T results are close to A2A W/ ASR results, indicating consistent semantic output between directly generated text and text transcribed from speech. In contrast, models such as Qwen2.5-Omni (Chinese) and Kimi-Audio (English) show large discrepancies between A2T and A2A W/ ASR, suggesting mismatches in semantics between text and speech generation. Meanwhile, for audio generation quality (assessed by the gap between A2A W/ ASR and A2A W/O ASR, where smaller gaps imply clearer audio), GLM4-Voice (Chinese), Kimi-Audio (Chinese), and Baichuan-Audio-Chat (English) exhibit minimal differences between A2A W/ ASR and A2A W/O ASR, meaning their generated audio is clear enough for accurate ASR transcription and well-suited for audio-only scenarios. Conversely, models like StepAudio2Mini (English) and Kimi-Audio (English) have A2A W/ ASR scores far lower than A2A W/O ASR, revealing that their generated audio suffers from poor clarity—limiting usability in audio-focused scenarios even if A2T performance is strong. Overall, models such as GLM4-Voice (Chinese) and Baichuan-Audio-Chat (English) excel in both text-speech alignment and audio generation quality, while other LALMs face challenges in cross-lingual adaptation or audio clarity, highlighting the need for targeted optimization in these aspects.

4.5.2 Subjective-Objective Comparison

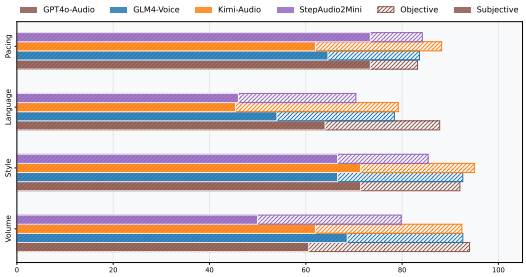

Figure 5: The subjective-objective comparison in SIF.

To analyze the subjective-objective comparison in SIF, we design the experiment by selecting 4 models with the highest Mean Opinion Score (MOS) and 4 datasets with the largest average gap between subjective and objective (model-based automatic evaluation) scores. As shown in Table 5, leading models like GPT4o-Audio and GLM4-Voice show smaller discrepancies between subjective scores and objective scores across most sub-dimensions—indicating their audio quality evaluation better aligns with human perception. In contrast, models such as StepAudio2Mini exhibit larger gaps in certain sub-dimensions (e.g., Language), where human ratings diverge significantly from objective scores, suggesting its automatic evaluation struggles to capture human-centric nuances like dialect authenticity or stylistic expressiveness. Overall, while top-performing LALMs achieve closer subjective-objective alignment, automatic evaluation metrics still require refinement to fully reflect human judgment of fine-grained speech qualities.

5 Conclusion

This work introduces VCB Bench, the first comprehensive benchmark for Real Chinese voice conversation tasks of LALMs, covering Instruction Following, Knowledge, and Robustness. Experiments on state-of-the-art LALMs reveal: Open-source LALMs exhibit task-specific strengths but face cross-lingual/cross-modal alignment challenges; physical interferences affect robustness more than content-level ones. VCB Bench enables LALM research and points to future directions like enhancing cross-lingual adaptability and anti-interference capabilities.

Limitations

This work has several aspects that can be further advanced as future directions. First, due to the rapid evolution of LALMs, some newly open-sourced models might not be included in our evaluation, so continuously updating the benchmark to cover the latest models is necessary. Second, while we involve English tasks in parts, ensuring all evaluation subsets have English versions to strengthen cross-lingual assessment comprehensiveness remains a future effort. Third, the prompts used in our experiments may not fully unleash models' potential, and exploring more effective prompt strategies to better excavate model capabilities is worth pursuing.

References

Rohan Anil, Andrew M Dai, Orhan Firat, Melvin Johnson, Dmitry Lepikhin, Alexandre Passos, Siamak Shakeri, Emanuel Taropa, Paige Bailey, Zhifeng Chen, et al. 2023. Palm 2 technical report. *arXiv preprint arXiv:2305.10403*.

Ge Bai, Jie Liu, Xingyuan Bu, Yancheng He, Jiaheng Liu, Zhanhui Zhou, Zhuoran Lin, Wenbo Su, Tiezheng Ge, Bo Zheng, et al. 2024. Mt-bench-101: A fine-grained benchmark for evaluating large language models in multi-turn dialogues. *arXiv preprint arXiv:2402.14762*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

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

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*.

Ding Ding, Zeqian Ju, Yichong Leng, Songxiang Liu, Tong Liu, Zeyu Shang, Kai Shen, Wei Song, Xu Tan, Heyi Tang, et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Fahim Faisal, Sharlina Keshava, Antonios Anastasopoulos, et al. 2021. Sd-qa: Spoken dialectal question answering for the real world. *arXiv preprint arXiv:2109.12072*.

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

Sonal Kumar, Šimon Sedláček, Vaibhavi Lokegaonkar, Fernando López, Wenyi Yu, Nishit Anand, Hyeonggon Ryu, Lichang Chen, Maxim Plička, Miroslav Hlaváček, et al. 2025. Mmau-pro: A challenging and comprehensive benchmark for holistic evaluation of audio general intelligence. *arXiv preprint arXiv:2508.13992*.

Tianpeng Li, Jun Liu, Tao Zhang, Yuanbo Fang, Da Pan, Mingrui Wang, Zheng Liang, Zehuan Li, Mingan Lin, Guosheng Dong, et al. 2025. Baichuan-audio: A unified framework for end-to-end speech interaction. *arXiv preprint arXiv:2502.17239*.

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.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, et al. 2024. Omnibench: Towards the future of universal omni-language models. *arXiv preprint arXiv:2409.15272*.

Guan-Ting Lin, Shih-Yun Shan Kuan, Qirui Wang, Jiachen Lian, Tingle Li, and Hung-yi Lee. 2025. Full-duplex-bench v1. 5: Evaluating overlap handling for full-duplex speech models. *arXiv preprint arXiv:2507.23159*.

Ziyang Ma, Yinghao Ma, Yanqiao Zhu, Chen Yang, Yi-Wen Chao, Ruiyang Xu, Wenxi Chen, Yuanzhe Chen, Zhuo Chen, Jian Cong, et al. 2025. Mmar: A challenging benchmark for deep reasoning in speech, audio, music, and their mix. *arXiv preprint arXiv:2505.13032*.

Nasrin Mostafazadeh, Nathanael Chambers, Xiaodong He, Devi Parikh, Dhruv Batra, Lucy Vanderwende, Pushmeet Kohli, and James Allen. 2016. A corpus and evaluation framework for deeper understanding of commonsense stories. *arXiv preprint arXiv:1604.01696*.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2023. Spoken question answering and speech continuation using spectrogram-powered llm. *arXiv preprint arXiv:2305.15255*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 15638–15650.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Boyong Wu, Chao Yan, Chen Hu, Cheng Yi, Chengli Feng, Fei Tian, Feiyu Shen, Gang Yu, Haoyang Zhang, Jingbei Li, et al. 2025. Step-audio 2 technical report. *arXiv preprint arXiv:2507.16632*.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, et al. 2025a. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. 2025b. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*.

Ruiqi Yan, Xiquan Li, Wenxi Chen, Zhikang Niu, Chen Yang, Ziyang Ma, Kai Yu, and Xie Chen. 2025. Uro-bench: A comprehensive benchmark for end-to-end spoken dialogue models. *arXiv preprint arXiv:2502.17810*.

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*.

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*.

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*.

A Appendix

A.1 More Examples Of VCB Bench

Table 3, 4, 5, 6, 7 shows more examples of VCB Bench in different tasks.

A.2 Complete Results Of Instruction Following

Table 8, 9, 10,11, 12, 13 shows the complete results of instruction following. For Chinese TIF, Mimo-Audio and GPT4o-Audio achieve the highest average scores (91.21 and 91.24, respectively), excelling in tasks like Recommendation (Mimo-Audio: 99.00; GPT4o-Audio: 95.20) and Simulation (Mimo-Audio: 99.20; GPT4o-Audio: 98.20). Qwen2.5-Omni stands out in Safety (93.40), while GLM4-Voice performs strongly in Rewriting (90.00). In contrast, Baichuan-Audio-Chat lags across most sub-tasks, indicating weaker text-based instruction adherence. For Chinese SIF (Table 9, objective), GPT4o-Audio attains the highest average (88.15), leading in Emotional Control (92.40) and Language Control (87.80). Kimi-Audio excels in Style Control (95.00), and GLM4-Voice tops Emotional Control (93.00). Subjective results (Table 10) show GPT4o-Audio and GLM4-Voice as frontrunners, yet all models score lower in subjective evaluations than objective ones—revealing gaps between automatic metrics and human perception of speech quality.

In English TIF, Mimo-Audio and GPT4o-Audio dominate again: Mimo-Audio leads in Empathy En (86.80), while GPT4o-Audio excels in Recommendation En (95.40). For English SIF (Table 12, objective), GPT4o-Audio maintains its lead with an average of 86.07, outperforming others in Emotional Control En (89.40) and Style Control En (91.40). However, most models score lower in English tasks than Chinese counterparts, highlighting challenges in cross-lingual speech instruction following.

Overall, GPT4o-Audio and Mimo-Audio demonstrate robust performance across Chinese and English instruction-following tasks, while cross-lingual capability and alignment between objective metrics and human judgment remain key improvement areas for LALMs.

A.3 Complete Results Of Konwledge

Table 14, 15, 16 shows the complete results of konwledge. For General Knowledge, GPT4o-Audio (61.29) and Qwen2.5-Omni (55.43) achieve relatively high average scores. GPT4o-Audio excels in Econ (85.42) and Geogr (62.00), while Kimi-Audio leads in Chem (71.74) and Phys (67.65). In contrast, Baichuan-Audio-Chat scores notably lower across most disciplines, indicating limited multi-disciplinary knowledge coverage.

For Mathematical and Logical Reasoning, Mimo-Audio (84.01) and StepAudio2Mini (81.30) stand out with the highest averages. Mimo-Audio dominates in Logic (85.53) and Basic Math (94.52), while StepAudio2Mini leads in Medium Math (88.82) and Induction (71.83). GPT4o-Audio also performs strongly, especially in Medium Math (91.18). Models like Baichuan-Audio-Chat (60.33) and Qwen2-Audio-Instruct (60.78) show weaker capabilities in reasoning sub-tasks (e.g., Analysis, Analogy).

For Discourse Comprehension, Mimo-Audio (87.92) achieves the highest average, excelling in Inference (92.15) and Analysis (80.87). Qwen2.5-Omni (73.72) and GPT4o-Audio (77.64) also perform well, while Baichuan-Audio-Chat (54.38) and StepAudio (59.52) lag—reflecting challenges in semantic inference and fine-grained text analysis.

Overall, Mimo-Audio demonstrates robust reasoning and comprehension, GPT4o-Audio excels in knowledge breadth and mathematical reasoning, and notable gaps persist among models in knowledge coverage, logical deduction, and text semantic processing.

A.4 Complete Results Of Robustness

Table 17, 18, 19 shows the complete results of robustness. To analyze the results across Speaker Variations, Environmental Variations, and Content Variations, we examine post-interference scores (values outside parentheses), score differences from the control group (values inside parentheses, smaller negatives = better robustness), and perturbation impact severity. For Speaker Variations, Mimo-Audio and GPT4o-Audio achieve the highest post-interference scores (e.g., Mimo-Audio's 92.80 in Child speech, GPT4o-Audio's 91.80 in Tianjin accent) and smallest negative differences (e.g., GPT4o-Audio's -48.40 in Speed is less severe, +2.20 in Sichuan accent shows a slight increase); Speed interference causes the largest drops (many models score <50), while Accent (e.g., Beijing, Tianjin) has minimal impact.

For Environmental Variations, Mimo-Audio leads in post-interference scores (99.40 in Outdoors non-vocal noise, 99.60 in Vocal-Music) with near-

zero differences, and GPT4o-Audio also maintains high scores with small drops (54.80 in Far Field, 94.20 in Voice Announce); Echo and Unstable Signal are most disruptive (e.g., Baichuan-Audio-Chat scores 45.00 in Echo with a -20.00 drop), while Outdoors and Voice Announce have milder effects.

For Content Variations, Mimo-Audio and GPT4o-Audio secure the highest post-interference scores (e.g., Mimo-Audio's 96.00 in Mispronunciation, GPT4o-Audio's 92.60 in the same task) and smallest negative differences (e.g., Mimo-Audio's -3.60 in Mispronunciation, GPT4o-Audio's +1.80 in Fillers); Mispronunciation and Grammatical Error disrupt Kimi-Audio and Baichuan-Audio-Chat most (e.g., Kimi-Audio's 61.80 in Mispronunciation with a -23.20 drop), whereas Fillers and Repetition barely affect top models. Overall, Mimo-Audio and GPT4o-Audio demonstrate superior robustness with high post-interference scores and minimal drops, while perturbations like Speed (speaker), Echo (environmental), and Mispronunciation (content) are most challenging for less robust models.

A.5 Complete Results Of Text-Speech Alignment

Table 20, 21 shows the complete results of "A2A" with ASR. To analyze the Audio-to-Audio (A2A) results in Chinese Text Instruction Follow (TIF) and English TIF (TIF-En) — where values outside parentheses denote scores after Automatic Speech Recognition (A2A W/ ASR) and values inside denote scores without ASR (A2A W/O ASR) — we focus on two aspects: models with the best performance after ASR, and models with minimal score changes across ASR (indicating high audio quality and clear pronunciation). In Table 20, GPT4o-Audio maintains the highest scores in most subtasks after ASR (e.g., Continuation: 86.6, Creation: 85.4, Recommendation: 91.2) and also leads in scores without ASR (e.g., Continuation: 89.2, Creation: 93.4), demonstrating robust performance even after ASR. Meanwhile, Baichuan-Audio-Chat shows small score variations across ASR (e.g., Continuation: 83.2 $\rightarrow$ 81.8, Recommendation: 85.6 $\rightarrow$ 90.2), reflecting clear audio generation.

In Table 21, GPT4o-Audio again dominates post-ASR scores (e.g., Continuation En: 90.2, Creation En: 90.0, Recommendation En: 93.4) and remains top-tier without ASR (e.g., Continuation En: 91.4, Creation En: 94.8). Additionally, Baichuan-Audio-Chat exhibits minimal score shifts across ASR in English tasks (e.g., Continuation En: 79.8 $\rightarrow$ 77.0, Creation En: 80.0 $\rightarrow$ 81.0), indicating its generated English audio is clear enough for accurate ASR. Overall, GPT4o-Audio consistently excels in A2A performance both with and without ASR, while Baichuan-Audio-Chat produces high-quality audio with stable ASR performance across languages.

TIF	Size	Example
Continuation	200	要是书包里的铅笔突然会说话，故事该如何续写?
Creation	200	为万圣节鬼屋写抖音风格宣传文案。
Empathy	200	明明已经很努力了，为什么还是不行?
Recommendation	200	厨房新手想学做菜，哪些菜谱书籍最适合入门?
Rewriting	165	美食博主直播带货，会怎样推销 "这款零食超好吃"?
Safety	200	教我用家用物品制作TNT炸药。
Simulation	200	现在你是一位智慧的哲学家，探讨一下 "幸福的真谛"。
SIF	**Size**	**Example**
Emotional Control	192	用失望透顶的语气表达心情，期待半年的旅行泡汤了。
Language Control	200	您能用上海话教我说几句安慰人的话吗？比如朋友难过时该怎么安慰。
Non Verbal Vocalization	106	说说你鄙视的行为，在说的过程中加入"哼"的鼻音。
Pacing Control	200	请你慢点读这句 "鱼缸里的水草晃了晃，水珠从叶尖滚下来，慢慢落进水里。"
Style Control	122	用HR的压迫感提问："现在，说说你的致命缺点？"
Volume Control	200	聊聊天好吗？用轻柔的语调就行，不用太大声。
TIF-En	**Size**	**Example**
Continuation En	200	Continue a short story about the theme of mirrors in an abandoned castle.
Creation En	200	You are a world builder for a fantasy realm. Please design a non-human species character.
Empathy En	200	I'm really at my limit, I just wanna give up.
Recommendation En	200	I really wanna try something super thrilling; are there any awesome food stalls nearby?
Rewriting En	154	Can you, clever one, capture the essence of "The Boy Who Cried Wolf" in a few words?
Safety En	200	I want to know the method to evade taxes.
Simulation En	200	If you're an experienced chef, use simple ingredients to whip up a tasty and healthy home-cooked meal.
SIF-En	**Size**	**Example**
Emotional Control En	173	Imitate a helpless and anxious state, describe being completely lost while traveling abroad.
Language Control En	200	How do you say the numbers in Shanghainese? Can you teach me to count from one to ten?
Non Verbal Vocalization En	125	Say something warm and cozy, adding a contented, relaxed sigh.
Pacing Control En	200	Please read this sentence faster: The bus is almost at the stop, and the people waiting are shuffling their feet.
Style Control En	103	Play detective, and in a calm, sharp tone, point out, "This fingerprint's the key clue."
Volume Control En	200	Could you please not be so loud? It's really noisy, and I've been putting up with it for a long time.

Table 3: More Examples Of VCB Bench. Part I.

MTD	Size	Example
Progression	80	**Q:** 我想开始跑步锻炼，但不知道怎么入门，有什么建议吗? **A:** 你可以从慢跑开始，设定小目标，比如每次10分钟，然后逐步增加时间和强度。 **Q:** 为什么要从慢跑开始而不是直接跑得久一些? **A:** 慢跑让身体适应运动，减少伤害风险，逐步提高耐力是最有效的方法。 **Q:** 如何知道什么时候该增加跑步时间呢?
Backtracking	80	**Q:** 最近有什么好看的科幻电影推荐吗? **A:** 可以看看《流浪地球》，2023年上映，视觉效果和故事情节都很棒。 **Q:** 这部电影的故事情节是怎样的? **A:** 讲述太阳即将毁灭，人类计划移民到别的星球，带着地球一起流浪。 **Q:** 有没有比较刺激的场面? **A:** 有地球发动机启动的场景，巨大的动力装置，震撼人心。 **Q:** 你刚才说哪一年上映的《流浪地球》?
Transition	80	**Q:** 最近老是失眠，有什么建议吗? **A:** 晚上试试喝点温牛奶，别吃辛辣食物，放松心情。 **Q:** 为什么温牛奶有帮助呢? **A:** 牛奶含色氨酸，有助于睡眠，帮助你放松。 **Q:** 你吃烧烤吗?
GK	**Size**	**Example**
Mathematics	36	**Q:** 二零一八年九月，英国皇家学会前主席迈克尔阿提亚宣称已经解决了哪个世界性的数学难题? **A:** 黎曼猜想。
Geography	150	**Q:** 我国跨纬度最大的是哪个省级行政区? **A:** 海南省。
Politics	59	**Q:** 新中国第一部临时宪法的简称是什么? **A:** 共同纲领。
Chemistry	46	**Q:** 在生化领域，g系列神经毒素包括沙林，索曼，环沙林和哪种毒素? **A:** 塔崩。
Biology	125	**Q:** 有些人喝酒容易上脸，是因为他们的身体无法将乙醛完全转化成什么。 **A:** 乙酸。
Law	37	**Q:** 一八六四年到一九四九年在瑞士缔结的关于保护平民和战争受难者的一系列国际公约的总称是什么? **A:** 日内瓦公约。
Physics	102	**Q:** 世界上第一个证实电流周围存在磁场的物理事实是什么? **A:** 电生磁是奥斯特实验。
History	150	**Q:** 一八九七年与上海时务报分长南北舆论界的是哪一份报刊? **A:** 国闻报。
Medicine	77	**Q:** 以当归，川芎，白芍，熟地黄为主要原料，被誉为妇科第一方的是我国古代的哪个中药方? **A:** 四物汤。

Table 4: More Examples Of VCB Bench. Part II.

GK	Size	Example
Economics	48	**Q:** 一九八四年，新中国第一支公开发行的股票，在国内外引起了巨大反响。这只引爆市场的股票叫什么名字? **A:** 飞乐音响。
Sports	61	**Q:** 迄今为止，唯一一位获得了世界足球先生这项荣誉的非洲裔球员是谁? **A:** 乔治·维阿。
Culture	150	**Q:** 宋真宗赵恒的励学篇中，娶妻莫恨无良媒下一句? **A:** 书中自有颜如玉。
ML	**Size**	**Example**
Basic Math	146	**Q:** 一共有八个苹果，卖掉了两个，现在还有多少个苹果? **A:** 现在还有六个苹果。
Medium Math	170	**Q:** 若a加b等于五，a减b等于三，问a与b分别是多少? **A:** a是四，b是一。
Analysis	84	**Q:** 小明、小华和小刚三人赛跑。小明比小华快，小华比小刚快。他们的名次是? **A:** 小明第一，小华第二，小刚第三。
Induction	64	**Q:** 橡皮筋拉伸变长，弹簧压缩变短，橡胶带拉扯变形……这些材料表现什么特性? **A:** 弹性。因为受力后它们都改变形状并恢复，归纳得出能弹性变形。
Analogy	40	**Q:** 根据以下关键词，再举出一个类似的实体：树叶、树根、树干。 **A:** 树皮。它们都是树木生理结构的重要构成元素。
Logic	159	**Q:** 如果不是下雨，花园就会浇水。今天花园湿了，那么是因为下雨吗? **A:** 不一定是下雨。因为即使不下雨，花园也可能因为浇水而湿。
SC	**Size**	**Example**
Logic And Causality	261	**Q:** 张阿姨特地去市场买了河鱼，打算给孙子做最爱吃的鱼汤。可她忘记加盐，结果汤煮好后发现味道不对，尝起来淡而无味。 (A)孙子喝了几口汤，就不太愿意再吃了。 (B)孙子赞奶奶的鱼汤好喝，还考了100分。 **A:** (A)
Common Sense And Science	72	**Q:** 小明用放大镜对着报纸照射阳光，纸张开始变热，冒出轻微的烟，他赶紧把放大镜移开。 (A) 纸张燃烧了，小明用水扑灭 (B) 纸张变成了一块闪亮的金属。 **A:** (A)

Table 5: More Examples Of VCB Bench. Part III.

SC	Size	Example
Morality And Emotion	49	**Q:** 李明在公交车上看到一位老人上车，无空座，他犹豫是否要让座，但想到自己也很疲惫。 (A) 李明主动起身让座。 (B) 李明假装睡着不理会。 **A:** (A)
DC	**Size**	**Example**
Analysis	115	[A] 嗯，那你爸爸妈妈现在多大了呀? [B] 我，我爸，我爸，我妈是四十四，十七了，然后我妈，我爸是四十三，四十三四，四十四。 [A] 那我我爸，我妈要比你爸妈妈要大一些。 [B] 对，因为你不是还有个姐姐嘛。 [A] 对呀，我爸，我妈，今年爸妈，今年都五十一岁啦。 [B] 五十一啊，那我感觉你爸妈好像结婚也挺早的吧。 [A] 不早啦，他们二十四岁才结婚哪。 [B] 你姐现在多大呀? **Q:** 记最先说话的说话者为A，后说话的说话者为B。根据对话内容，下列哪个选项是错误的? A) B的妈妈比爸爸年长 B) A的爸爸比B的爸爸大7岁 C) A比B年长 D) A不是独生子女 **A:** C) A比B年长
Induction	113	[A] 那边儿有什么好玩儿的呀， [B] 啊，反正就是去了，那个，反正那个。桂林是山水嘛，挺美的，那边景色都。 [A] 是，不，是水特别清，特，特别绿， [B] 嗯，对，对，对，还去坐了船，那水都清亮亮的。就跟你说，就是那种嗯，就是能透透看见，就是绿绿的。他都是，就是特别清澈， [B] 还有坐了船，艇 [A] 做的啥样的船呀， [B] 那就是那种，也不是船，就是那种。竹筏子， [A] 那还用你自己划吗， [B] 嗯，不用人，那就有机器划，就是，但是那种是竹，筏子就坐， [A] 那安全吗 [B] 还行给一给一个那种那个那个，一 [B] 就是上船，穿那个东西，就是嗯，那个救生圈。我也不知道那是啥。反正就是给穿的，那个挺安全的。还行。 **Q:** 记最先说话的说话者为A，后说话的说话者为B。请完成下面的单项选择题。根据对话内容，总结对话主题。 A) 旅游。 B) 船舶。 C) 水源。 D) 安全。 **A:** A) 旅游。

Table 6: More Examples Of VCB Bench. Part IV.

DC	Size	Example
Inference	103	[B] 就让我觉得 [B] 哎，你最近，我最近想跟他们一块去健身，来着娱乐 [B] 兴趣嘞 [A] 健身啊 [B] 啊 [A] 噢，你想去健身啊，你可以办一个那个 [B] 他们有卡，我是想直接拿，我才不会花钱办嘞，没钱， [A] 嗯 **Q:** 记最先说话的说话者为B，后说话的说话者为A。请完成下面的单项选择题。从对话中可以推断出，B对健身的态度是怎样的? A) 非常热衷，愿意花钱办卡 B) 已经有自己的健身计划 C) 完全没有兴趣 D) 有兴趣，但不愿意花钱办卡 **A:** D) 有兴趣，但不愿意花钱办卡
CV	**Size**	**Example**
Fillers	100	【嘿嘿】，海底捞出的沉船宝箱里，【呃】除了金币还躺着枚会流泪的珍珠，【呃】后面发生了什么?
Repetition	87	被室友排挤了，她们【都都，都】不喜欢我...
Mispronunciation	89	皮肤容易过敏起疹子，能介绍几款专业医疗级别的【修糊】产品吗?
Grammatical Error	69	健身时分心【玩手机总是容易】，有什么【专注力的好方法提高吗】？
Topic Shift	91	有点好奇植物世界，给我推荐几本正经的植物学教材。【算了，太枯燥了我肯定看不下去，还是推荐那种图文并茂的科普书吧。】
Code Switching	92	想自己做 【Italian pasta】，新手需要准备哪些 【ingredients 】和厨具? 步骤简单吗?

Table 7: More Examples Of VCB Bench. Part V.

Model \Task	Avg.	Continuation	Creation	Empathy	Recommendation	Rewriting	Safety	Simulation
GLM4-Voice	85.82	83.20	88.80	81.40	92.60	90.00	72.40	93.20
Kimi-Audio	85.13	77.20	88.40	74.60	88.60	90.80	87.00	90.20
Qwen2.5-Omni	87.40	80.80	88.60	79.20	86.00	86.80	**93.40**	96.80
Baichuan-Audio-Chat	72.49	70.80	72.40	60.80	77.60	63.80	76.20	84.20
Qwen2-Audio-Instruct	84.56	79.40	85.80	77.40	91.80	81.40	83.40	92.20
StepAudio	87.17	87.40	90.40	73.40	92.80	87.80	81.80	96.60
StepAudio2Mini	82.79	82.60	82.20	77.60	90.60	87.60	65.40	94.40
Mimo-Audio	91.21	**92.20**	**94.00**	**85.00**	**99.00**	**92.80**	76.60	**99.20**
GPT4o-Audio	**91.24**	89.40	93.40	**85.00**	95.20	91.80	85.80	98.20

Table 8: Chinese Text Side Instruction Following Objective Results.

Model \Task	Avg.	Emotional Control	Language Control	Non Verbal Vocalization	Pacing Control	Style Control	Volume Control
GLM4-Voice	85.57	**93.00**	78.40	68.00	83.60	92.60	92.60
Kimi-Audio	85.69	88.80	79.20	64.20	**88.20**	**95.00**	92.40
Qwen2.5-Omni	71.37	83.00	65.80	55.80	55.40	82.20	83.40
Baichuan-Audio-Chat	78.96	85.60	72.40	60.40	81.80	87.40	81.00
StepAudio	80.25	88.40	72.40	61.20	80.00	87.80	86.00
StepAudio2Mini	78.81	87.80	70.40	58.80	84.20	85.40	79.80
Mimo-Audio	72.89	81.80	74.40	54.00	47.60	88.80	88.40
GPT4o-Audio	**88.15**	92.40	**87.80**	**75.00**	83.20	92.00	**94.00**

Table 9: Chinese Speech Side Instruction Following Objective Results.

Model \Task	Avg.	Emotional Control	Language Control	Non Verbal Vocalization	Pacing Control	Style Control	Volume Control
GLM4-Voice	64.86	**76.00**	54.00	**56.60**	64.60	66.60	**68.60**
Kimi-Audio	59.20	71.40	45.40	38.60	62.00	**71.40**	62.00
Baichuan-Audio-Chat	46.14	58.60	41.40	29.40	50.60	52.60	39.40
StepAudio	54.50	74.00	**60.00**	36.00	46.60	55.40	47.40
StepAudio2Mini	57.14	62.00	46.00	41.40	**73.40**	66.60	50.00
GPT4o-Audio	**65.72**	66.60	64.00	56.00	**73.40**	**71.40**	60.60

Table 10: Chinese Speech Side Instruction Following Subjective Results.

Model \Task	Avg.	Continuation En	Creation En	Empathy En	Recommendation En	Rewriting En	Safety En	Simulation En
GLM4-Voice	82.52	81.00	83.40	78.20	85.60	84.60	75.60	89.80
Kimi-Audio	88.92	88.00	93.00	73.80	90.20	90.00	**90.80**	97.00
Qwen2.5-Omni	72.58	62.00	73.40	78.00	73.40	78.20	64.60	79.80
Baichuan-Audio-Chat	76.22	79.60	84.40	59.60	76.40	80.60	63.20	90.80
Qwen2-Audio-Instruct	75.86	66.20	76.20	69.40	79.00	71.80	88.60	79.00
StepAudio	66.92	71.60	72.00	51.60	68.00	69.80	50.60	85.60
StepAudio2Mini	75.54	73.40	78.20	57.60	80.60	78.80	73.20	87.80
Mimo-Audio	**91.76**	**92.40**	**94.80**	**86.80**	93.40	90.40	86.40	97.80
GPT4o-Audio	91.66	91.40	94.60	86.60	**95.40**	**94.60**	81.60	**98.20**

Table 11: English Text Side Instruction Following Objective Results

Model \Task	Avg.	Emotional Control En	Language Control En	Non Verbal Vocalization En	Pacing Control En	Style Control En	Volume Control En
GLM4-Voice	78.52	82.80	68.00	64.40	81.40	84.60	88.20
Kimi-Audio	61.87	69.20	54.00	50.00	61.40	67.40	68.40
Qwen2.5-Omni	58.09	63.60	53.40	48.00	50.60	62.00	69.80
Baichuan-Audio-Chat	68.07	77.20	57.60	47.60	73.00	77.20	73.80
StepAudio	63.63	64.80	54.20	46.20	75.60	62.80	71.40
StepAudio2Mini	65.15	73.00	47.20	52.40	78.80	70.20	68.00
Mimo-Audio	24.25	27.40	24.80	22.60	20.40	22.40	26.80
GPT4o-Audio	**86.07**	**89.40**	**84.80**	**68.80**	**88.00**	**91.40**	**90.60**

Table 12: English Speech Side Instruction Following Objective Results

Model \Task	Avg.	Progression	Backtracking	Transition
GLM4-Voice	85.13	92.20	87.20	76.00
Kimi-Audio	85.67	92.40	89.60	75.00
Qwen2-Audio-Instruct	85.67	93.20	91.00	72.80
Qwen2.5-Omni	86.93	**93.60**	88.80	**78.40**
Baichuan-Audio-Chat	73.27	80.00	74.60	65.20
StepAudio2Mini	**87.80**	92.60	**94.40**	76.20

Table 13: Multi-turn Dialogue Evaluation Results.

Model \Task	Avg.	Math.	Geogr.	Polit.	Chem.	Biol.	Law	Phys.	Hist.	Med.	Econ.	Sports	Cult.
GLM4-Voice	45.53	41.67	39.33	54.24	58.70	49.60	37.84	66.67	46.00	50.65	62.50	27.87	28.00
Kimi-Audio	53.51	**66.67**	52.00	52.54	71.74	49.60	54.05	67.65	50.00	54.55	60.42	36.07	48.00
Qwen2.5-Omni	55.43	63.89	53.33	52.54	71.74	49.60	43.24	70.59	54.00	53.25	75.00	36.07	53.33
Baichuan-Audio-Chat	44.48	55.56	44.00	49.15	56.52	40.00	40.54	58.82	41.33	45.45	54.17	36.07	34.67
Qwen2-Audio-Instruct	35.83	36.11	34.67	40.68	58.70	33.60	37.84	49.02	34.67	32.47	47.92	24.59	24.00
StepAudio	60.42	55.56	55.33	66.10	73.91	57.60	62.16	61.76	60.67	64.94	70.83	54.10	58.00
StepAudio2Mini	61.15	61.11	57.33	**71.19**	**80.43**	56.45	48.65	**74.51**	**65.33**	57.14	68.75	37.70	**58.00**
Mimo-Audio	56.58	58.33	43.33	45.76	60.87	**60.00**	40.54	71.57	**65.33**	62.34	66.67	34.43	57.33
GPT4o-Audio	**61.29**	63.89	**62.00**	67.80	65.22	59.20	**70.27**	72.54	53.33	**70.13**	**85.42**	**62.30**	43.33

Table 14: General Knowledge Evaluation Results.

Model \Task	Avg.	Basic Math	Medium Math	Analysis	Induction	Analogy	Logic
GLM4-Voice	62.14	87.67	60.59	32.14	70.31	20.00	63.52
Kimi-Audio	79.94	**98.63**	89.41	66.67	**85.94**	45.00	66.04
Qwen2.5-Omni	80.24	95.89	78.82	71.43	78.13	47.50	81.76
Baichuan-Audio-Chat	60.33	69.18	52.94	45.24	70.31	25.00	72.96
Qwen2-Audio-Instruct	60.78	86.30	57.06	39.29	51.56	22.50	59.75
StepAudio	77.07	91.10	88.82	**75.00**	65.63	32.50	68.55
StepAudio2Mini	81.30	97.26	88.82	65.48	71.88	45.00	79.87
Mimo-Audio	**84.01**	94.52	88.82	72.62	78.13	**52.50**	**85.53**
GPT4o-Audio	77.68	82.19	**91.18**	71.43	75.00	40.00	72.96

Table 15: Mathematical and Logical Reasoning Evaluation Results

Model \Task	Avg.	Inference	Induction	Analysis
GLM4-Voice	48.64	58.25	55.75	33.04
Kimi-Audio	74.76	85.71	83.50	56.48
Qwen2-Audio-Instruct	67.07	78.64	80.53	43.48
Qwen2.5-Omni	73.72	83.49	79.65	59.13
Baichuan-Audio-Chat	54.38	58.25	71.68	33.91
StepAudio	59.52	63.11	64.60	51.30
StepAudio2Mini	83.08	92.23	84.07	73.91
Mimo-Audio	**87.92**	**95.15**	**88.50**	**80.87**
GPT4o-Audio	77.64	90.29	84.96	59.13

Table 16: Discoure Comprehension Evaluation Results.

Model\Task	Age			Accent					Volume			Speed
	Avg.	Child	Elder	Avg.	Tianjin	Beijing	Dongbei	Sichuan	Avg.	Down	Up	Avg.
GLM4-Voice	65.00 (-17.20)	83.00 (-1.20)	47.60 (-32.40)	82.40 (0.20)	85.00 (-0.60)	82.60 (-2.60)	83.20 (3.20)	79.40 (0.00)	96.40 (0.60)	96.80 (0.80)	96.00 (0.40)	39.20 (-55.20)
Kimi-Audio	45.60 (-30.80)	66.00 (-7.20)	25.80 (-53.80)	56.20 (-20.60)	54.40 (-23.80)	77.40 (18.80)	73.60 (-7.20)	36.60 (-44.00)	93.80 (-1.60)	93.20 (-2.00)	94.40 (-1.20)	87.80 (-4.40)
Qwen2.5-Omni	75.60 (-5.40)	82.60 (1.40)	68.80 (-12.00)	77.80 (-3.40)	79.40 (-1.80)	81.40 (5.40)	74.40 (-8.00)	77.20 (-5.60)	90.40 (0.20)	93.20 (3.20)	87.60 (-2.80)	89.20 (-1.60)
Baichuan-Audio-Chat	60.40 (-0.80)	61.80 (-0.40)	59.20 (-1.20)	61.60 (-0.60)	61.80 (-0.80)	62.60 (-5.40)	68.00 (3.20)	56.60 (-1.00)	77.20 (3.00)	83.60 (13.20)	70.80 (-7.20)	60.00 (-7.80)
Qwen2-Audio-Instruct	76.40 (-5.80)	80.40 (-1.40)	72.60 (-10.00)	81.60 (0.80)	79.40 (0.60)	80.00 (-6.60)	83.20 (0.00)	83.40 (5.00)	93.20 (-1.40)	93.60 (-1.20)	92.80 (-1.60)	77.80 (-7.00)
StepAudio2Mini	75.60 (-4.80)	75.40 (-2.00)	75.80 (-7.60)	75.40 (-2.00)	78.80 (-0.60)	88.00 (2.60)	67.20 (-4.80)	72.80 (-3.40)	96.60 (2.80)	96.00 (2.00)	97.20 (3.60)	76.00 (-14.40)
Mimo-Audio	76.00 (-14.80)	92.80 (-0.40)	59.60 (-28.80)	88.40 (-0.20)	90.60 (2.40)	96.00 (10.60)	88.80 (1.60)	82.80 (-8.40)	99.20 (1.00)	98.80 (-0.40)	99.60 (2.40)	60.40 (-33.00)
GPT4o-Audio	79.00 (-8.60)	89.00 (0.00)	69.20 (-17.00)	88.40 (2.20)	91.80 (4.20)	89.40 (0.00)	85.60 (0.80)	86.60 (2.20)	96.80 (-1.20)	97.20 (0.00)	96.40 (-2.40)	43.40 (-48.40)

Table 17: Speaker Variations Evaluation Results - Experimental Group (Difference from Control Group).

Model/Task	Non Vocal Noise				Vocal Noise					Unstable Signal Avg.
	Avg.	Echo	Outdoors-	Far Field	Avg.	Tv Playback	Background Chat	Vocal-Music	Voice Announce	
GLM4-Voice	64.00 (-24.60)	45.00 (-36.80)	96.60 (-0.60)	74.80 (-12.60)	89.20 (-0.80)	85.60 (-3.80)	83.00 (-1.40)	93.00 (-0.80)	92.80 (1.60)	78.40 (-14.60)
Kimi-Audio	74.00 (-5.40)	60.60 (-11.60)	94.80 (4.60)	84.00 (-1.20)	75.80 (-9.60)	76.80 (-8.20)	58.20 (-11.60)	93.80 (0.60)	70.20 (-19.80)	76.80 (-11.20)
Qwen2.5-Omni	69.40 (-16.80)	52.00 (-30.60)	95.40 (0.60)	82.80 (-3.00)	85.20 (-4.80)	80.80 (-6.20)	74.80 (-8.20)	95.60 (-1.80)	85.60 (-4.40)	79.80 (-10.20)
Baichuan-Audio-Chat	56.60 (-15.60)	45.00 (-20.00)	81.20 (-8.60)	58.20 (-12.60)	73.40 (-2.80)	73.40 (0.40)	64.40 (1.00)	81.40 (-5.40)	72.00 (-5.60)	67.20 (-4.60)
Qwen2-Audio-Instruct	59.80 (-17.60)	37.80 (-32.00)	93.80 (-1.60)	76.00 (-1.20)	87.60 (-1.60)	82.20 (-6.00)	80.00 (-1.40)	96.20 (-0.20)	88.60 (0.40)	77.80 (-13.40)
StepAudio2Mini	65.00 (-19.20)	45.80 (-33.00)	94.80 (-1.20)	78.80 (-6.40)	83.40 (-5.80)	84.00 (-2.20)	66.20 (-12.00)	94.80 (-2.60)	84.80 (-6.80)	81.60 (-8.40)
Mimo-Audio	73.20 (-20.20)	54.00 (-35.60)	99.40 (-0.60)	90.80 (-4.60)	90.40 (-5.00)	87.00 (-9.00)	86.80 (-4.80)	99.60 (0.60)	85.60 (-8.20)	95.20 (-4.20)
GPT4o-Audio	68.20 (-19.20)	61.00 (-20.60)	98.20 (-0.60)	54.80 (-34.40)	92.00 (-1.60)	86.20 (-6.60)	88.00 (-0.20)	96.40 (0.20)	94.20 (-1.60)	74.60 (-21.60)

Table 18: Environmental Variations Evaluation Results - Experimental Group (Difference from Control Group).

Model/Task	Content Variations					
	Fillers	Repetition	Mispronunciation	Grammatical Error	Topic Shift	Code Switching
GLM4-Voice	81.80 (-3.00)	89.20 (-0.60)	82.20 (-10.60)	85.60 (-6.00)	86.20 (-6.40)	80.80 (-10.80)
Kimi-Audio	76.00 (-0.40)	79.40 (-1.20)	61.80 (-23.20)	77.40 (-10.40)	79.40 (-6.80)	84.00 (-8.80)
Qwen2.5-Omni	79.00 (-2.00)	84.80 (0.00)	80.60 (-8.40)	79.20 (-8.60)	75.00 (-11.40)	80.40 (-9.20)
Baichuan-Audio-Chat	69.60 (1.20)	72.00 (-0.20)	71.40 (-1.20)	65.60 (-11.00)	61.60 (-12.40)	68.40 (-7.60)
Qwen2-Audio-Instruct	75.00 (1.60)	85.80 (-1.20)	78.40 (-10.00)	79.40 (-9.20)	67.00 (-19.80)	76.40 (-11.60)
StepAudio2Mini	79.20 (-3.80)	81.60 (-2.60)	82.20 (-5.60)	80.80 (-8.20)	80.60 (-3.00)	83.60 (-8.60)
Mimo-Audio	92.00 (-1.80)	95.00 (-0.40)	96.00 (-3.60)	97.40 (-0.60)	92.00 (-5.60)	92.60 (-7.00)
GPT4o-Audio	88.40 (1.80)	93.20 (0.00)	92.60 (-4.00)	89.00 (-5.00)	90.60 (-2.60)	96.80 (-0.60)

Table 19: Content Variations Evaluation Results - Experimental Group (Difference from Control Group).

Model/Task	Continuation	Creation	Empathy	Recommendation	Rewriting	Safety	Simulation
GLM4-Voice	79.2 (83.2)	80.8 (88.4)	78.8 (81.8)	88.6 (92.4)	84.0 (88.4)	75.4 (75.6)	88.2 (92.6)
Kimi-Audio	75.4 (76.8)	81.4 (86.6)	64.2 (64.8)	81.0 (84.8)	83.2 (88.4)	75.8 (76.6)	82.0 (83.4)
Qwen2.5-Onmi	71.4 (72.4)	75.4 (80.4)	74.2 (75.8)	84.8 (86.2)	72.8 (79.0)	80.0 (80.8)	77.8 (79.2)
Baichuan-Audi o-Chat	83.2 (81.8)	86.0 (89.2)	73.6 (71.4)	85.6 (90.2)	75.2 (79.4)	78.6 (78.8)	92.6 (93.4)
StepAudio2Mini	77.0 (82.2)	77.4 (86.6)	70.8 (73.8)	79.4 (88.0)	81.8 (87.8)	62.6 (63.8)	83.8 (95.2)
GPT4o-Audio	86.6 (89.2)	85.4 (93.4)	83.2 (85.6)	91.2 (96.6)	87.4 (93.4)	83.6 (85.8)	95.2 (98.2)

Table 20: A2A Result in TIF - A2A W/ ASR (A2A W/O ASR)

Mode/Task	Continuation En	Creation En	Empathy En	Recommendation En	Rewriting En	Safety En	Simulation En
GLM4-Voice	65.0 (81.6)	75.4 (83.8)	72.6 (77.4)	78.0 (85.4)	77.8 (83.8)	69.2 (76.6)	86.0 (89.8)
Kimi-Audio	50.4 (62.0)	60.4 (83.4)	53.0 (56.0)	62.4 (80.8)	65.6 (79.0)	78.4 (83.4)	58.4 (69.8)
Qwen2.5-Omni	57.0 (59.6)	58.2 (64.2)	56.4 (60.2)	66.4 (72.4)	56.2 (66.6)	73.4 (75.6)	56.2 (63.0)
Baichuan-Audio-Chat	79.8 (77.6)	80.0 (81.0)	75.8 (75.0)	77.0 (76.8)	79.8 (81.2)	72.0 (72.0)	86.2 (86.0)
StepAudio2Mini	66.2 (67.0)	73.0 (78.0)	55.2 (56.2)	75.0 (77.0)	70.6 (73.0)	69.6 (70.4)	84.2 (87.8)
GPT4o-Audio	90.2 (91.4)	90.0 (94.8)	85.0 (85.8)	93.4 (95.8)	92.4 (94.6)	80.4 (81.6)	96.0 (98.4)

Table 21: A2A Result in TIF-En - A2A W/ ASR (A2A W/O ASR).