MSU-Bench: Towards Understanding the Conversational Multi-talker Scenarios

Shuai Wang[1*], **Zhaokai Sun**[2*], **Zhennan Lin**[2], **Chenyou Wang**[2], **Zhou Pan**[3], **Lei Xie**[2]

[1]Nanjing University,
[2]Audio, Speech and Language Processing Lab (ASLP@NPU)
[3] Li Auto Inc.
shuaiwang@nju.edu.cn, zxsun@mail.nwpu.edu.cn, lxie@nwpu.edu.cn

Abstract

Spoken Language Understanding (SLU) has progressed from traditional single-task methods to large audio language model (LALM) solutions. Yet, most existing speech benchmarks focus on single-speaker or isolated tasks, overlooking the challenges posed by multi-speaker conversations that are common in real-world scenarios. We introduce **MSU-Bench**, a comprehensive benchmark for evaluating multi-speaker conversational understanding with a speaker-centric design. Our hierarchical framework covers four progressive tiers: single-speaker static attribute understanding, single-speaker dynamic attribute understanding, multi-speaker background understanding, and multi-speaker interaction understanding. This structure ensures all tasks are grounded in speaker-centric contexts, from basic perception to complex reasoning across multiple speakers. By evaluating state-of-the-art models on MSU-Bench, we demonstrate that as task complexity increases across the benchmark's tiers, all models exhibit a significant performance decline. We also observe a persistent capability gap between open-source models and closed-source commercial ones, particularly in multi-speaker interaction reasoning. These findings validate the effectiveness of MSU-Bench for assessing and advancing conversational understanding in realistic multi-speaker environments. Demos can be found in the supplementary material.

Spoken Language Understanding (SLU) constitutes a fundamental task in artificial intelligence, enabling machines to interpret human speech beyond mere transcription. Recent advances in SLU research have transitioned from traditional single-task approaches, such as automatic speech recognition (ASR), automatic speaker verification (ASV), and spoken sentiment analysis (SSA), toward Large Audio Language Models (LALMs) (Peng et al. 2024; Su et al. 2025). Building upon established Large Language Model (LLM) paradigms, sophisticated LALMs such as LTU-AS (Gong et al. 2023), Salmonn (Tang et al. 2024a), Qwen-Audio (Chu et al. 2023, 2024a) and OSUM (Geng et al. 2025) have emerged, demonstrating exceptional general speech understanding capabilities.

However, real-world conversations inherently involve multiple speakers and present fundamentally different challenges than single-speaker scenarios. Human dialogues are inherently collaborative and social, involving complex interactions among multiple participants where speakers frequently interrupt one another, reference previous statements, and dynamically shift conversational roles. While researchers have developed sophisticated techniques for individual aspects of multi-speaker processing, including speaker diarization, speech separation, and target speaker extraction, these methods predominantly focus on isolated technical problems without capturing the holistic complexity of conversational dynamics. Critical speaker-centric phenomena such as social role analysis, power dynamics, and interactional patterns remain largely unexplored.

Existing speech benchmarks (Chen et al. 2024; Yang et al. 2024; Ao et al. 2024; Sakshi et al. 2025; Wang et al. 2025b,a) typically aggregate speaker-related tasks with general speech or dialogue tasks, rarely isolating the unique challenges inherent in speaker-centric understanding within authentic conversational contexts. Consequently, essential multi-speaker dynamics, including dominance detection, turn-taking analysis, and social role identification, remain underexplored. This gap is particularly significant given that most real-world applications require understanding not just individual speakers but the complex interplay between multiple participants in dynamic conversational settings.

To bridge this critical gap, we introduce **MSU-Bench**, the first comprehensive benchmark specifically designed to define and evaluate multi-speaker understanding in authentic interactive scenarios. MSU-Bench employs a hierarchical framework that decomposes speaker-centric understanding into four progressive tiers of complexity: single-speaker static attribute understanding (e.g., speaker counting, demographic profiling), single-speaker dynamic attribute understanding (e.g., emotion state tracking, voice quality evolution), multi-speaker background understanding (e.g., venue/event inference, role identification), and multi-speaker interaction understanding (e.g., dominance detection, interruption pattern analysis). This systematic decomposition enables targeted evaluation of the social and interactive dimensions of conversational understanding while maintaining clear progression from basic perceptual tasks to complex reasoning scenarios.

Our work makes the following key contributions:

- We present the first benchmark that is dedicated to conversational speaker-centric understanding. It uses a clear

[*]These authors contributed equally.

Characteristics	Speech Understanding Benchmarks						
	VoiceBench	MMSU	MMAU	AudioBench	AIR-Bench	SD-Eval	MSU-Bench
Speaker-Oriented	×	×	×	×	×	×	✓
Audio Source	TTS+RPC	RPC	RPC	RPC	RPC	TTS+RPC	RPC
Conversation Type	Mono.	Dial.	Dial.	Dial.	Dial.	Dial.	Dial.
Multi-speaker	×	×	×	×	×	×	✓
Speaker-related Task	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparative analysis of construction characteristics across seven speech understanding benchmarks. ✓ indicates support, × indicates no support. TTS: Text-to-Speech, RPC: Real Person Recordings, Mono.: Monologue, Dial.: Dialogue.

hierarchical structure from basic perception to advanced reasoning.

- We propose a comprehensive "5M" design. This covers multi-tier, multi-speaker, multi-lingual, multi-scenario, and multi-task aspects for thorough evaluation.
- We conduct extensive empirical analysis of both open-source and closed-source models. Our results show persistent performance gaps, especially in multi-speaker interaction reasoning.
- We provide a detailed task construction pipeline and open-source codebase. This supports reproducibility and promotes future research on speaker-centric understanding in conversations.

Related Works

Speaker Modeling and Analysis

Speaker information constitutes a crucial dimension of the acoustic signal (Wang et al. 2024). Conventionally, the scope of speaker modeling has been narrowly centered on identity for tasks like recognition and verification. However, a more holistic perspective, often termed speaker understanding, extends to a rich set of paralinguistic traits, including a speaker's accent, age, and emotional state. While early research addressed these characteristics through separate, task-specific systems, a recent paradigm shift has led to the development of benchmarks that evaluate speaker attributes more comprehensively. For instance, the VoxProfile (Feng et al. 2025) benchmark was introduced to analyze speaker profiles across various facets. A limitation of such approaches is their inherent focus on single-utterance, single-speaker scenarios, which precludes the analysis of more complex, interactive contexts that are essential for understanding real-world conversational dynamics.

Speech Understanding Models

Speech understanding encompasses the machine interpretation of semantic content and emotional nuances in spoken language, extending beyond simple transcription to include intent recognition, sentiment analysis, and dialogue act identification. Recent advances in large language models have significantly enhanced this capability by integrating audio modalities with LLMs (Huang et al. 2024; Zhang et al. 2023; Ghosh et al. 2025; Goel et al. 2025; Chu et al. 2024a; Xu et al. 2025). Current approaches fall into two categories: cascade and end-to-end methods. Cascade approaches utilize automatic speech recognition followed by natural language processing, as exemplified by AudioGPT (Huang et al. 2024) combining Whisper with LLMs. While modular and industrially mature, this method suffers from error propagation and acoustic information loss. End-to-end approaches directly map speech signals to semantic representations, demonstrated by models such as SpeechGPT (Zhang et al. 2023), Salmonn (Tang et al. 2024b; Yu et al. 2025), Glm-4-voice (Zeng et al. 2024), GPT 4o-Audio, Gemini (Team et al. 2023), Kimi-Audio (Ding et al. 2025), Step-audio (Wu et al. 2025; Huang et al. 2025) and the Qwen-Audio series (Chu et al. 2023, 2024b). These models exhibit greater robustness and universal audio understanding capabilities. Recent multimodal extensions like Gemini and Qwen2.5-Omni further integrate audio and visual information.

Speech Understanding Benchmarks

Recent advances in LALMs have catalyzed the emergence of diverse benchmarks for speech understanding. To systematically assess the current landscape, we compare several representative benchmarks with our proposed MSU-Bench in Table 1. Early efforts such as AudioBench (Wang et al. 2025a) mainly target foundational capabilities, including automatic speech recognition (ASR) and audio classification. More recent benchmarks, such as MMAU (Sakshi et al. 2025) and MMSU (Wang et al. 2025b), extend the scope to encompass audio-based question answering and the assessment of paralinguistic features. Nevertheless, as summarized in Table 1, none of the existing benchmarks are explicitly speaker-centric or designed to evaluate multi-speaker interactions. While many adopt dialogue recordings (Dial.) and include speaker-aware tasks, they fail to address the critical challenge of modeling the relationships and interaction logic among different speakers, which is essential for a comprehensive conversational understanding. To bridge this gap, we introduce **MSU-Bench**, the first benchmark specifically dedicated to the rigorous evaluation of multi-speaker understanding in realistic conversational scenarios.

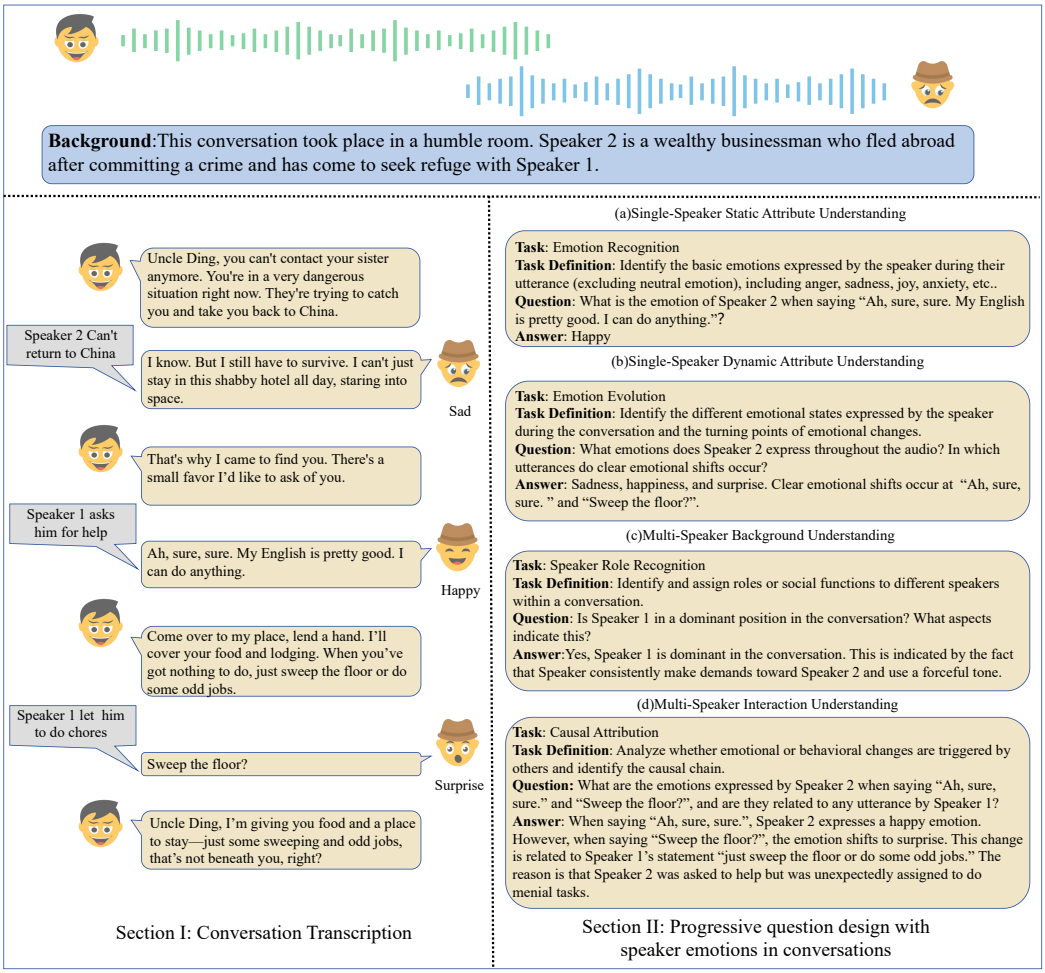

Figure 1: **Example of MSU-Bench QA.** The conversation scenario involves two speakers with different emotional states and social dynamics, demonstrating the progression from basic emotion recognition to complex causal reasoning.

MSU-Bench: Hierarchical Design for Multi-Speaker Understanding

Hierarchical Task Framework

We propose a four-tier hierarchical framework for multi-speaker understanding tasks, organized by increasing complexity to systematically evaluate model capabilities across different tiers of conversational understanding, the definitions and scope of each tier are demonstrated in Table 2.

Framework Overview. Our framework progresses from basic speaker-level perception to complex multi-party interaction reasoning as Figure 1. The progression follows a natural cognitive hierarchy: Tier 1 establishes foundational recognition capabilities for static speaker attributes, Tier 2 extends to temporal dynamics analysis within individual speakers, Tier 3 advances to contextual inference and background understanding across multiple speakers, and Tier 4 culminates in comprehensive multi-speaker interaction understanding. Models can be assessed at each tier independently, enabling precise identification of strengths and limitations across the full spectrum of multi-speaker understanding tasks.

Tier 1: Single-Speaker Static Attribute Understanding. This tier focuses on identifying and characterizing individual speakers' static attributes. The primary objectives include speaker differentiation, demographic profiling (gender, age, accent), and paralinguistic analysis (voice quality, emotional tone). These capabilities establish the perceptual foundation necessary for higher-level reasoning tasks.

Tier 2: Single-Speaker Dynamic Attribute Understanding. Building upon static attribute recognition, this tier addresses temporal dynamics within individual speakers. Key capabilities include tracking emotional evolution, detecting voice quality changes, and identifying opinion shifts throughout conversations. Unlike Tier 1, this level requires models to reason about causality and context–understanding not just *what* changes occur, but *why* they occur. This tier evaluates models' ability to capture speaker-internal dynamics and infer cultural identity through language patterns and expression preferences.

Tier 3: Multi-Speaker Background Understanding. This tier focuses on contextual inference beyond immediate interaction dynamics. Tasks include inferring conversa-

	Capability	Description	Representative Tasks
	Tier 1: Single-Speaker Static Attribute Understanding		
Single-Speaker → Multi-Speaker	**Speaker Recognition (SR)**	Identify and track speakers in multi-speaker environments, focusing on speech content, timing, alternation, frequency, and interaction structure.	Speaker Recognition Speaker Counting Silence/Overlap Detection
	Speaker Attribute Comprehension (SAC)	Determine static attributes such as gender, age, accent, and language background of the speaker.	Accent/Dialect Recognition Language Recognition Gender Recognition Age Recognition
	Speaker Paralinguistic Analysis (SPA)	Analyze vocal characteristics such as timbre, fluency, and emotional tone.	Voice Quality Analysis Speech Flow Analysis Emotion Recognition
	Tier 2: Single-Speaker Dynamic Attribute Understanding		
Static → Dynamic	**Speaker Dynamic Analysis (SDA)**	Detect and interpret dynamic changes in emotion, voice quality, and perspective over the course of a conversation.	Emotion Evolution Voice Quality Evolution Opinion Change Recognition
	Speaker Cultural Identity Integration (SCII)	Infer cultural background, geographical affiliation, age group, and cognitive style by analyzing language, accent, and expression preferences.	Language/Accent Cultural Reasoning Expression Preference Recognition Geographical Location Estimation
	Tier 3: Multi-Speaker Background Understanding		
Perception → Reasoning	**Multi-Speaker Scene Inference (MSSI)**	Infer conversational venue and predict conversational outcomes from topics and language styles.	Dialogue Background Reasoning Dialogue Result Reasoning
	Multi-Speaker Relationship Inference (MSRI)	Understand speaker relationships and infer social roles in multi-speaker conversations.	Speaker Role Recognition Social Role Recognition
	Tier 4: Multi-Speaker Interaction Understanding		
	Multi-Speaker Transcription (MST)	Identify and distinguish multiple speakers in conversations, ensuring accurate restoration of semantic content.	Dialogue Transcription
	Multi-Speaker Interaction Analysis (MSIA)	Understand interpersonal dynamics in multi-speaker interactions through paralinguistic and social cues.	Paralinguistic Interaction Analysis Social Interaction Analysis
	Multi-Speaker Contextual Reasoning (MSCR)	Analyze emotional shifts, intention changes, and interaction logic among speakers, enabling semantic-based cross-speaker reasoning.	Causal Attribution Motivation Reasoning

Table 2: **Hierarchical Multi-Speaker Understanding Tasks.** The framework systematically progresses from single-speaker static perception to multi-speaker dynamic reasoning, encompassing 10 core capabilities and 25 representative tasks, detailed description and examples of individual tasks can be found in the appendix.

tional venues, predicting dialogue outcomes, and determining speaker roles and social relationships. Models must analyze contextual cues, topic patterns, and language styles to understand the broader situational context. This tier evaluates the ability to reason about environmental factors and social structures that influence multi-speaker conversations.

Tier 4: Multi-Speaker Interaction Understanding. The highest tier extends analysis to inter-speaker dynamics and conversational interactions. Tasks include multi-speaker transcription with accurate attribution, paralinguistic and social interaction analysis, and cross-speaker reasoning about emotional shifts and motivations. This tier requires models to simultaneously track multiple speakers while reason-

Models	Tier 1				Tier 2			Tier 3			Tier 4				Avg
	SR	SAC	SPA	Avg	SDA	SCII	Avg	MSSI	MSRI	Avg	MST	MSIA	MSCR	Avg	
Kimi-Audio	0.39	0.53	0.38	0.44	0.21	0.29	0.25	0.38	0.40	0.39	0.35	0.23	0.24	0.25	0.35
Qwen2.5-Omni	0.48	0.48	0.37	0.45	0.26	0.34	0.29	0.33	0.44	0.36	0.29	0.34	0.26	0.30	0.37
GPT-4o-Audio	0.52	0.65	**0.55**	0.58	0.38	0.52	0.44	0.70	0.51	0.64	0.37	0.49	0.36	0.43	0.52
Gemini-2.5-Flash	0.49	**0.70**	0.51	0.58	0.41	0.57	0.48	0.76	0.66	0.73	0.38	0.51	0.39	0.45	0.55
Gemini-2.5-Pro	**0.55**	**0.70**	0.54	**0.61**	**0.46**	**0.61**	**0.53**	**0.80**	**0.67**	**0.76**	**0.44**	**0.56**	**0.47**	**0.51**	**0.59**

Table 3: Performance comparison of different models across four tiers and various capabilities. Bold values indicate the best performance in each group. Definitions of all capability abbreviations are provided in Table 2. Note that the Avg values are computed on all involved cases, instead of simply averaging individual capability values.

ing about their mutual influence, conversational control, and collaborative behaviors. Success at this level demonstrates comprehensive understanding of dynamic multi-party conversational structures.

The four-tier architecture ensures systematic progression from perceptual to reasoning tasks, from static to dynamic understanding, and from single-speaker to multi-speaker analysis. This design principle allows for granular evaluation of model capabilities while maintaining clear relationships between different complexity levels.

QA Construction Pipeline

To construct the four-tier benchmark with diverse speaker-centric audio-text tasks, we build a rigorous QA generation pipeline that automatically produces high-quality question–answer pairs from multi-speaker dialogues spanning various real-world scenarios and acoustic conditions. For each core ability, we design dedicated prompts to guide template construction and question formulation, ensuring that the resulting QA samples are tightly aligned with task-specific objectives. All selected audio segments span 60–120 seconds and include at least two speakers, thereby guaranteeing meaningful multi-speaker interaction and benchmark reliability. The overall QA pipeline(see Figure 2), will be further detailed in the subsequent sections.

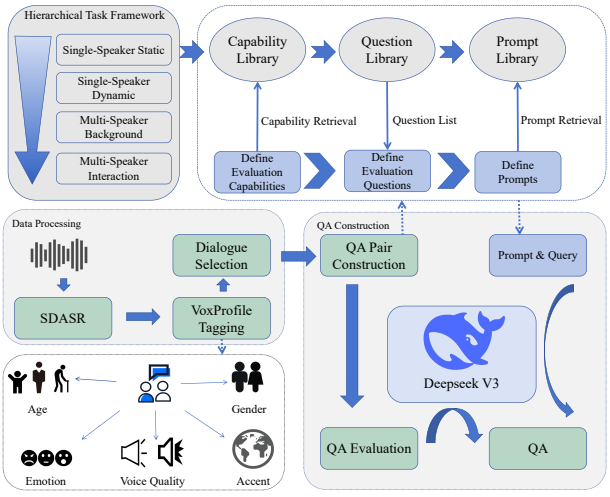

Figure 2: The QA Construction Pipeline

Data Selection and Preparation. To support the construction of multi-lingual and multi-scenario conversational datasets, we curate dialogue data from diverse real-world sources. For both near-field and far-field multi-speaker interactions, the full QA pipeline is applied to generate high-quality QA pairs. For film and television content, a denoising module was applied prior to subsequent processing. All QA data generated for downstream evaluation go through an additional filtering stage, guided by large language model (LLM) assessments, to ensure its quality and relevance.

Specifically, the Chinese near-field data is sourced from *MDT-AA007*, while the far-field data comes from *alimeeting (Yu et al. 2022a,b)*, a corpus of multi-speaker, meeting-style conversations recorded with distant microphones in acoustically complex environments. For English, near-field data is collected from *MDT-AD015*, which contains telephone-based conversational speech. The far-field data is drawn from *CHiME6 (Watanabe et al. 2020)*, comprising real home-based multi-party conversations captured with distant microphones, featuring challenges such as background noise and overlapping speech.

Pipeline Design. Considering the complexity and heterogeneity of the data, the following sequential modules[1] are employed, as demonstrated in Figure 2:

1. **Transcription with Speaker Diarization:** Multi-speaker dialogues are first transcribed using an SDASR (Speaker-Diarized ASR) system, which aligns speech content with speaker identities over time.
2. **Speaker Attribute Tagging:** The transcriptions are then processed by the *VoxProfile* module, which annotates each speaker with metadata such as gender, role (e.g., host or guest), and secondary language usage.
3. **Dialogue Segment Selection:** Utilizing Deepseek v3, dialogue segments are selected based on annotated speaker information and task-specific configuration, ensuring both relevance and diversity in the selected content.
4. **QA Pair Construction:** QA pairs are generated from the selected segments using Deepseek v3, following predefined strategies tailored to the model's capabilities and the conversation context.

[1]The QA construction pipeline will be released alongside the benchmark

5. **Automated QA Evaluation:** The resulting QA pairs undergo automated evaluation via Deepseek v3's built-in assessment mechanism, which measures QA quality against the original dialogue context and annotations.

This structured and modular pipeline enables consistent, scalable, and high-quality QA data generation across complex, language-diverse, and acoustically challenging multi-speaker conversational scenarios.

Experiments and Results

Dataset and Evaluation Protocols

Dataset MSU-Bench leverages six open-source datasets to cover diverse conversational scenarios [2]: MDT-AA007 (Chinese telephone) and MDT-AD015 (English telephone) for near-field dialogue, AliMeeting (Yu et al. 2022a,b) (Chinese meeting) and CHiME6 (Barker et al. 2018; Watanabe et al. 2020) (English home dialogue) for far-field settings, and Chinese MovieClips and English MovieClips for challenging acoustic conditions (film audio post-processed to remove background music). For each task, sessions are randomly sampled from all datasets to ensure linguistic and acoustic diversity. Ten question-answer pairs per session are manually verified and constructed using varied templates, balancing phrasing diversity with consistent reasoning requirements. This ensures that performance differences are mainly attributable to audio complexity rather than variations in question form. We find that multiple-choice questions can unintentionally provide LALMs with additional cues, potentially inflating performance. Instead, we use open-ended questions with answer formats and constraints in the inference prompts.

Evaluation Protocols The evaluation uses a dedicated scoring prompt to assess LALM outputs along 3 dimensions: relevance, accuracy, and causal soundness. Relevance ensures that responses are tightly aligned with questions, filtering out hallucinated or off-topic content. Accuracy measures the factual correctness of the response against the ground truth. Causal soundness evaluates the logical consistency of cause-effect reasoning in inference tasks; for non-causal questions, this component receives full marks by default.

Speech Understanding Models This work evaluates five representative models[3] on MSU-Bench to assess multi-speaker understanding. Specifically, Gemini-2.5-Pro (Team et al. 2023), Gemini-2.5-Flash (Team et al. 2023), and GPT-4o-Audio are closed-source commercial systems, while Kimi-Audio (Ding et al. 2025) and Qwen2.5-Omni (Xu et al. 2025) are open-source models. This diverse selection enables a systematic comparison between open-source and commercial solutions across all benchmark tiers and tasks.

[2]Detailed statistics such as audio sources, duration distributions can be found in the appendix

[3]Due to input length restrictions (e.g., a 30-second audio limit), certain models could not be evaluated on multi-speaker scenarios. We also noticed the latest Step-Audio2 (Wu et al. 2025) and Audio Flamingo3 (Goel et al. 2025); however, as of submission, their model weights and inference code were not publicly available.

Evaluation Results and Analysis

We present comprehensive evaluation results of both state-of-the-art open-source and commercial models on our benchmark (see Table 3). The test set is carefully balanced, with samples drawn uniformly from six diverse data sources. For each benchmark tier and capability, we report the average results of all tasks related to that capability, with the task-capability mappings detailed in Table 2. Moreover, the comparison of different systems across all 25 tasks is illustrated in Figure 3, providing a highly intuitive performance comparison across models.

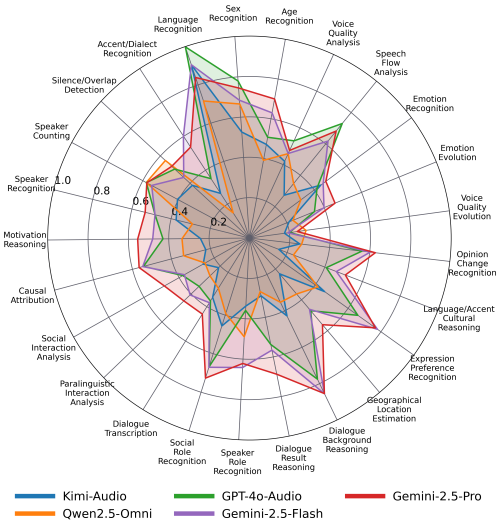

Figure 3: Overall Performance Comparison

Commercial vs. Open-source Performance Gap. Our evaluation reveals a significant and consistent performance gap between commercial and open-source models. The Gemini series shows superior performance across all tiers, with Gemini-2.5-Pro achieving the best results in 8 out of 9 capabilities. GPT-4o-Audio shows strong performance in paralinguistic analysis (0.55 in SPA) but weaker performance in cultural identity integration (0.52 in SCII). Among open-source models, Kimi-Audio shows relatively strong performance in speaker recognition (0.39 in SR) but struggles significantly with multi-speaker interaction analysis (0.23 in MSIA), indicating limitations in handling complex multi-party dynamics.This suggests that the involved commercial models have significantly better capabilities in handling complex multi-speaker reasoning tasks.

Static vs. Dynamic Attribute Understanding Our analysis reveals fundamental differences in how models handle static versus dynamic speaker attributes. Static attributes (e.g., age, gender) remain constant in dialogue, while dynamic ones (e.g., emotion, prosody) vary with context. To bridge the training–inference label gap in audio models, we provide explicit labels during inference. Evaluation shows large performance gaps across models: Gemini (Team et al. 2023) excels in age, gender, and accent recognition, while GPT-4o is accurate on gender but weaker on age and accent.

Most models struggle with dynamic attributes like timbre and emotion, revealing limitations in current LALMs' paralinguistic understanding.

Acoustic vs. Semantic Processing Patterns Our results reveal a clear preference for semantic over acoustic processing across all models. Tasks that primarily rely on semantic content (e.g., SAC, MSSI) achieve significantly higher performance than those requiring acoustic analysis (e.g., SPA, SDA). This pattern is consistent across all tiers and model families, suggesting a fundamental limitation in current LALM architectures. This pattern suggests that current LALMs can not perfectly leverage fine-grained acoustic features, even when these features are explicitly relevant to the task. This limitation has significant implications for real-world applications where acoustic cues are crucial for understanding speaker intent and emotional state.

Cross-Lingual Performance Analysis MSU-Bench enables natural cross-lingual evaluation, revealing interesting patterns in model performance across English and Chinese. We present the average results of different tier tasks for each model through a heatmap visualization, as shown in Figure 4. The results demonstrate that different models exhibit largely consistent trends across both English and Chinese languages, indicating that these models do not exhibit over-optimization for any specific language. This also demonstrates that the difficulty levels of our selected English and Chinese data sources are comparable.

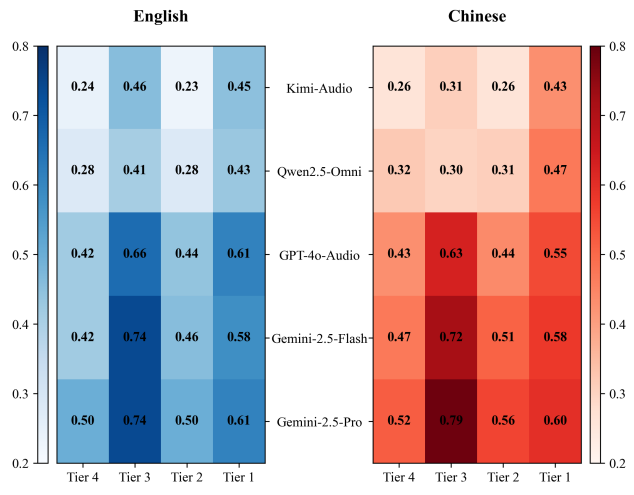

Figure 4: Performance comparison of different models on tasks of different languages

Paralinguistic context v.s. Semantic context Our analysis of Tier 3 (Multi-Speaker Background Understanding) versus Tier 4 (Multi-Speaker Interaction Understanding) reveals important distinctions in how models process different types of contextual information. Tier 3 tasks focus on background understanding and rely heavily on semantic content to infer dialogue scenes and speaker relationships. Notably, even with speaker attribution errors, models with transcription capability can still identify the dialogue setting and social relations based on topical cues. In contrast, Tier 4 tasks require deeper analysis of emotional and semantic exchanges, serving as a more sophisticated analysis of speaker dynamics. When emotional shifts are explicitly marked in QA settings, models leveraging transcription and semantic reasoning can infer causes accurately. However, when inferring emotions solely from dialogue segments, models often accumulate errors due to the difficulty of tracking emotional dynamics across multiple speakers.

Error Analysis

For a systematic comparison of error causes, we select Gemini-2.5-Pro and Kimi-Audio as two representatives of commercial and open-source models, aggregating results across all 25 sub-tasks and follow MMSU's taxonomy to classify errors into several key categories.We randomly sampled 200 QA pairs from error cases across different tiers for both Gemini-2.5-Pro and Kimi-Audio and analyzed the error causes, as summarized in Table 4.

Error Type	Gemini-2.5-Pro (%)	Kimi-Audio (%)
Rejection of Answer	–	0.50
Answer Extraction Errors	4.04	48.24
Perceptual Errors	56.06	37.68
Reasoning Errors	37.37	13.57
Lack of Knowledge	2.53	–

Table 4: Error distribution analysis

Our evaluation reveals that Gemini-2.5-Pro provides comprehensive responses with strong instruction-following while Kimi-Audio frequently delivers partial answers. Both models exhibit significant perception errors, consistent with MMSU findings. Additional analysis of Tier 4 error distribution shows perception errors dominate at 68.09%, highlighting the necessity of strong multi-speaker perception for effective interaction comprehension[4].

Conclusion

We present MSU-Bench, a comprehensive four-tier benchmark for multi-speaker conversational understanding that systematically evaluates speaker-centric capabilities from basic perception to complex reasoning. Our hierarchical framework covers single-speaker static/dynamic attribute understanding and multi-speaker background/interaction understanding, with all tasks grounded in authentic conversational contexts. Through extensive evaluation of state-of-the-art models, we demonstrate significant performance degradation as task complexity increases across tiers, revealing persistent gaps between open-source and commercial solutions, particularly in multi-speaker interaction reasoning. Our analysis reveals critical limitations in current LALMs' ability to handle fine-grained acoustic cues and complex multi-speaker dynamics, highlighting the need for more efforts to the complex conversational understanding. MSU-Bench provides a standardized evaluation platform to facilitate future research in multi-speaker speech under-

[4]Tier-wise error distribution can be found in the appendix

standing and guide development of more robust conversational AI systems.

Limitations: Despite efforts to diversify scenarios and models, access limitations may constrain comprehensiveness. We hope this dataset helps identify audio-language model performance on speaker-centric tasks.

References

Ao, J.; Wang, Y.; Tian, X.; Chen, D.; Zhang, J.; Lu, L.; Wang, Y.; Li, H.; and Wu, Z. 2024. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Advances in Neural Information Processing Systems*, 37: 56898–56918.

Barker, J.; Watanabe, S.; Vincent, E.; and Trmal, J. 2018. The Fifth 'CHiME' Speech Separation and Recognition Challenge: Dataset, Task and Baselines. In *Proc. Interspeech 2018*, 1561–1565.

Chen, Y.; Yue, X.; Zhang, C.; Gao, X.; Tan, R. T.; and Li, H. 2024. Voicebench: Benchmarking llm-based voice assistants. *arXiv preprint arXiv:2410.17196*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024a. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; et al. 2024b. Qwen2-audio technical report. *arXiv preprint arXiv:2407.10759*.

Chu, Y.; Xu, J.; Zhou, X.; Yang, Q.; Zhang, S.; Yan, Z.; Zhou, C.; and Zhou, J. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; et al. 2025. Kimi-audio technical report. *arXiv preprint arXiv:2504.18425*.

Feng, T.; Lee, J.; Xu, A.; Lee, Y.; Lertpetchpun, T.; Shi, X.; Wang, H.; Thebaud, T.; Moro-Velazquez, L.; Byrd, D.; et al. 2025. Vox-Profile: A Speech Foundation Model Benchmark for Characterizing Diverse Speaker and Speech Traits. *arXiv preprint arXiv:2505.14648*.

Geng, X.; Wei, K.; Shao, Q.; Liu, S.; Lin, Z.; Zhao, Z.; Li, G.; Tian, W.; Chen, P.; Li, Y.; et al. 2025. OSUM: Advancing open speech understanding models with limited resources in academia. *arXiv preprint arXiv:2501.13306*.

Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025. Audio flamingo 2: An audio-language model with long-audio understanding and expert reasoning abilities. *arXiv preprint arXiv:2503.03983*.

Goel, A.; Ghosh, S.; Kim, J.; Kumar, S.; Kong, Z.; Lee, S.-g.; Yang, C.-H. H.; Duraiswami, R.; Manocha, D.; Valle, R.; et al. 2025. Audio Flamingo 3: Advancing Audio Intelligence with Fully Open Large Audio Language Models. *arXiv preprint arXiv:2507.08128*.

Gong, Y.; Liu, A. H.; Luo, H.; Karlinsky, L.; and Glass, J. 2023. Joint audio and speech understanding. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8. IEEE.

Huang, A.; Wu, B.; Wang, B.; Yan, C.; Hu, C.; Feng, C.; Tian, F.; Shen, F.; Li, J.; Chen, M.; et al. 2025. Step-audio: Unified understanding and generation in intelligent speech interaction. *arXiv preprint arXiv:2502.11946*.

Huang, R.; Li, M.; Yang, D.; Shi, J.; Chang, X.; Ye, Z.; Wu, Y.; Hong, Z.; Huang, J.; Liu, J.; et al. 2024. Audiogpt: Understanding and generating speech, music, sound, and talking head. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23802–23804.

Peng, J.; Wang, Y.; Fang, Y.; Xi, Y.; Li, X.; Zhang, X.; and Yu, K. 2024. A survey on speech large language models. *arXiv preprint arXiv:2410.18908*.

Sakshi, S.; Tyagi, U.; Kumar, S.; Seth, A.; Selvakumar, R.; Nieto, O.; Duraiswami, R.; Ghosh, S.; and Manocha, D. 2025. MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark. In *The Thirteenth International Conference on Learning Representations*.

Su, Y.; Bai, J.; Xu, Q.; Xu, K.; and Dou, Y. 2025. Audio-Language Models for Audio-Centric Tasks: A survey. *arXiv preprint arXiv:2501.15177*.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; MA, Z.; and Zhang, C. 2024a. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations*.

Tang, C.; Yu, W.; Sun, G.; Chen, X.; Tan, T.; Li, W.; Lu, L.; Ma, Z.; and Zhang, C. 2024b. SALMONN: Towards Generic Hearing Abilities for Large Language Models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Team, G.; Anil, R.; Borgeaud, S.; Alayrac, J.-B.; Yu, J.; Soricut, R.; Schalkwyk, J.; Dai, A. M.; Hauth, A.; Millican, K.; et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; and Chen, N. 2025a. AudioBench: A Universal Benchmark for Audio Large Language Models. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics*, 4297–4316.

Wang, D.; Wu, J.; Li, J.; Yang, D.; Chen, X.; Zhang, T.; and Meng, H. 2025b. MMSU: A Massive Multi-task Spoken Language Understanding and Reasoning Benchmark. *arXiv preprint arXiv:2506.04779*.

Wang, S.; Chen, Z.; Lee, K. A.; Qian, Y.; and Li, H. 2024. Overview of speaker modeling and its applications: From the lens of deep speaker representation learning. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

Watanabe, S.; Mandel, M.; Barker, J.; Vincent, E.; Arora, A.; Chang, X.; Khudanpur, S.; Manohar, V.; Povey, D.; Raj, D.; et al. 2020. CHiME-6 Challenge: Tackling multispeaker speech recognition for unsegmented recordings. In *CHiME 2020-6th International Workshop on Speech Processing in Everyday Environments*.

Wu, B.; Yan, C.; Hu, C.; Yi, C.; Feng, C.; Tian, F.; Shen, F.; Yu, G.; Zhang, H.; Li, J.; et al. 2025. Step-Audio 2 Technical Report. *arXiv preprint arXiv:2507.16632*.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; et al. 2025. Qwen2. 5-omni technical report. *arXiv preprint arXiv:2503.20215*.

Yang, Q.; Xu, J.; Liu, W.; Chu, Y.; Jiang, Z.; Zhou, X.; Leng, Y.; Lv, Y.; Zhao, Z.; Zhou, C.; et al. 2024. AIR-Bench: Benchmarking Large Audio-Language Models via Generative Comprehension. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 1979–1998.

Yu, F.; Zhang, S.; Fu, Y.; Xie, L.; Zheng, S.; Du, Z.; Huang, W.; Guo, P.; Yan, Z.; Ma, B.; Xu, X.; and Bu, H. 2022a. M2MeT: The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Challenge. In *Proc. ICASSP*. IEEE.

Yu, F.; Zhang, S.; Guo, P.; Fu, Y.; Du, Z.; Zheng, S.; Huang, W.; Xie, L.; Tan, Z.-H.; Wang, D.; Qian, Y.; Lee, K. A.; Yan, Z.; Ma, B.; Xu, X.; and Bu, H. 2022b. Summary On The ICASSP 2022 Multi-Channel Multi-Party Meeting Transcription Grand Challenge. In *Proc. ICASSP*. IEEE.

Yu, W.; Wang, S.; Yang, X.; Chen, X.; Tian, X.; Zhang, J.; Sun, G.; Lu, L.; Wang, Y.; and Zhang, C. 2025. SALMONN-omni: A Standalone Speech LLM without Codec Injection for Full-duplex Conversation. *CoRR*, abs/2505.17060.

Zeng, A.; Du, Z.; Liu, M.; Wang, K.; Jiang, S.; Zhao, L.; Dong, Y.; and Tang, J. 2024. Glm-4-voice: Towards intelligent and human-like end-to-end spoken chatbot. *arXiv preprint arXiv:2412.02612*.

Zhang, D.; Li, S.; Zhang, X.; Zhan, J.; Wang, P.; Zhou, Y.; and Qiu, X. 2023. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.