MULTIVOX: A Benchmark for Evaluating Voice Assistants for Multimodal Interactions

Ramaneswaran Selvakumar*, Ashish Seth*, Nishit Anand,
Utkarsh Tyagi, Sonal Kumar, Sreyan Ghosh, Dinesh Manocha
University of Maryland, College Park
{ramans, aseth125, dmanocha}@umd.edu

Abstract

The rapid progress of Large Language Models (LLMs) has empowered *omni* models to act as voice assistants capable of understanding spoken dialogues. These models can process multimodal inputs beyond text, such as speech and visual data, enabling more context-aware interactions. However, current benchmarks fall short in comprehensively evaluating how well these models generate context-aware responses, particularly when it comes to implicitly understanding fine-grained speech characteristics, such as pitch, emotion, timbre, and volume or the environmental acoustic context such as background sounds. Additionally, they inadequately assess the ability of models to align paralinguistic cues with complementary visual signals to inform their responses. To address these gaps, we introduce MULTIVOX, the first omni voice assistant benchmark designed to evaluate the ability of voice assistants to integrate spoken and visual cues including paralinguistic speech features for truly multimodal understanding. Specifically, MULTIVOX includes 1000 human-annotated and recorded speech dialogues that encompass diverse paralinguistic features and a range of visual cues such as images and videos. Our evaluation on 10 state-of-the-art models reveals that, although humans excel at these tasks, current open-source models consistently struggle to produce contextually grounded responses.[1]

1 Introduction

With recent advancements in Multimodal Large Language Models (MLLMs) (Xu et al., 2025; Microsoft et al., 2025), there is a growing interest in developing models that can understand and generate information across multiple modalities, such as images, video, and audio-simultaneously. This evolution is paving the way for the development of Omni Language Models (OLMs), which are crucial for building efficient and versatile Artificial General Intelligence (AGI) (Bubeck et al., 2023; Morris et al., 2024). While OLMs provide a wide range of applications (Xu et al., 2025), one of their primary use cases is developing *omni-modal voice assistants (OVA)* (Huang et al., 2024). Unlike traditional speech voice assistants that rely solely on speech instruction, OVAs powered by OLMs such as GPT-4o (OpenAI, 2024) and Qwen2.5 Omni (Xu et al., 2025), can understand speech dialogues and reason over multimodal inputs, including images and videos.

Advancing the application of OLMs in voice assistants poses challenges not only in model development but also in constructing effective evaluation benchmarks. While existing OLM benchmarks like OmniBench (Li et al., 2024) incorpo-

Figure 1: Comparison of existing benchmark with MULTIVOX. Existing benchmark for omni-modal voice assistant evaluation are derived from vision-centric text VQA benchmarks. In MULTIVOX, models need leverage not only visual cues but also non-verbal speech signals.

[1] https://github.com/ramaneswaran/multivox

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 28481–28493
November 4-9, 2025

rate multimodal inputs such as images and video, they lack spoken dialogues—an essential modality for assessing the conversational and auditory capabilities required of voice assistants. On the other hand, current voice assistant benchmarks such as VoXDialogue (Cheng et al., 2025), SD-Eval (Ao et al., 2025), and S2S-Arena (Jiang et al., 2025) focus primarily on evaluating a model's ability to generate contextually appropriate responses based on auditory cues like content, emotion, or speaker demographics embedded in a speech instruction. However, these benchmarks fall short of capturing the full multimodal reasoning abilities expected of OVAs, particularly in integrating visual cues alongside speech instructions.

To address this gap, we introduce MULTIVOX, a novel benchmark designed to evaluate an OLMs ability to incorporate multimodal cues to provide accurate and contextual responses. MULTIVOX includes *1000* questions consisting of *human spoken questions paired with either a video or an image*. Unlike existing benchmarks which primarily test visual grounding and use speech to deliver the content of a straightforward instruction, MULTIVOX consists of questions which require a model to combine visions skills such as object recognition, scene understanding, scene text understanding with speech skills such as acoustic scene understanding, paralanguage understanding and speaker profiling (See fig. 1). The spoken questions in MULTIVOX are recorded by professional voice actors to cover a diverse range of paralinguistic and emotional features that are not possible with current text-to-speech systems. A key problem in benchmarks that evaluate multi-modal reasoning capability is that models can take shortcuts by exploiting priors from other modalities, to mitigate this we introduce confounding samples in MULTIVOX. Specifically, each question in our benchmark has another associated question which has the same textual and visual content but their speech property is flipped such that expected answers should be different. Our key contributions are:

1. We present MULTIVOX, the first benchmark designed to evaluate omni-modal language models (OLMs) using human-spoken queries paired with visual inputs. The 1000 examples require models to jointly ground visual and paralinguistic speech cues to produce accurate, context-aware responses.

2. We evaluate 10 omni-modal models on MULTIVOX and find that, while humans excel with ease, current OLMs consistently struggle, particularly with grounding speech signals, revealing a critical bottleneck in their capabilities.

3. We perform extensive qualitative and quantitative analysis on model's responses and uncover key insights: Models exhibit strong visual grounding but rely heavily on textual cues for speech-related tasks; they often ignore non-verbal audio signals like tone or background sounds.

2 Related Work

Omni Voice Assistants The recent development of Omni Language Models (OLMs) has enabled development of omni-modal voice assistants that can simultaneously infer across both visual and speech inputs. Recent iterations of previously mentioned voice assistants now support visual inputs in form of either images like in the case of Mini-Omni2 (Xie and Wu, 2024) and MoshiVis (Royer et al., 2025) or video inputs such as Qwen2.5-Omni (Xu et al., 2025). While these models demonstrate impressive instruction-following capabilities as voice assistants, *when extensively evaluated on* MULTIVOX*, we find that they often overlook crucial paralinguistic cues—such as tone, emotion, and pitch—in speech input, which are essential for generating context-aware responses.*

Benchmarks For Voice Assistants While there are works such as OmniBench (Li et al., 2024) and OmniXR (Chen et al., 2024a) that evaluate OLMs, there have been few efforts to standardize the evaluation of omni voice assistants. Recent work such as Lyra (Zhong et al., 2024) repurpose existing visual question answering (VQA) benchmarks by converting the textual questions to speech. However, such approach overlooks crucial non-verbal information typically present in spoken conversations. While some progress has been made in evaluating speech VAs, many existing benchmarks still fall short. VoiceBench (Chen et al., 2024b) assesses capabilities like world knowledge and instruction following by converting textual benchmarks like MMLU (Hendrycks et al., 2021) and AlpacaEval (Dubois et al., 2024) to speech, but overlooks the non-verbal speech information. SD-Eval (Ao et al., 2025) is a pioneering work in evaluating paralinguistic features but is limited to only four

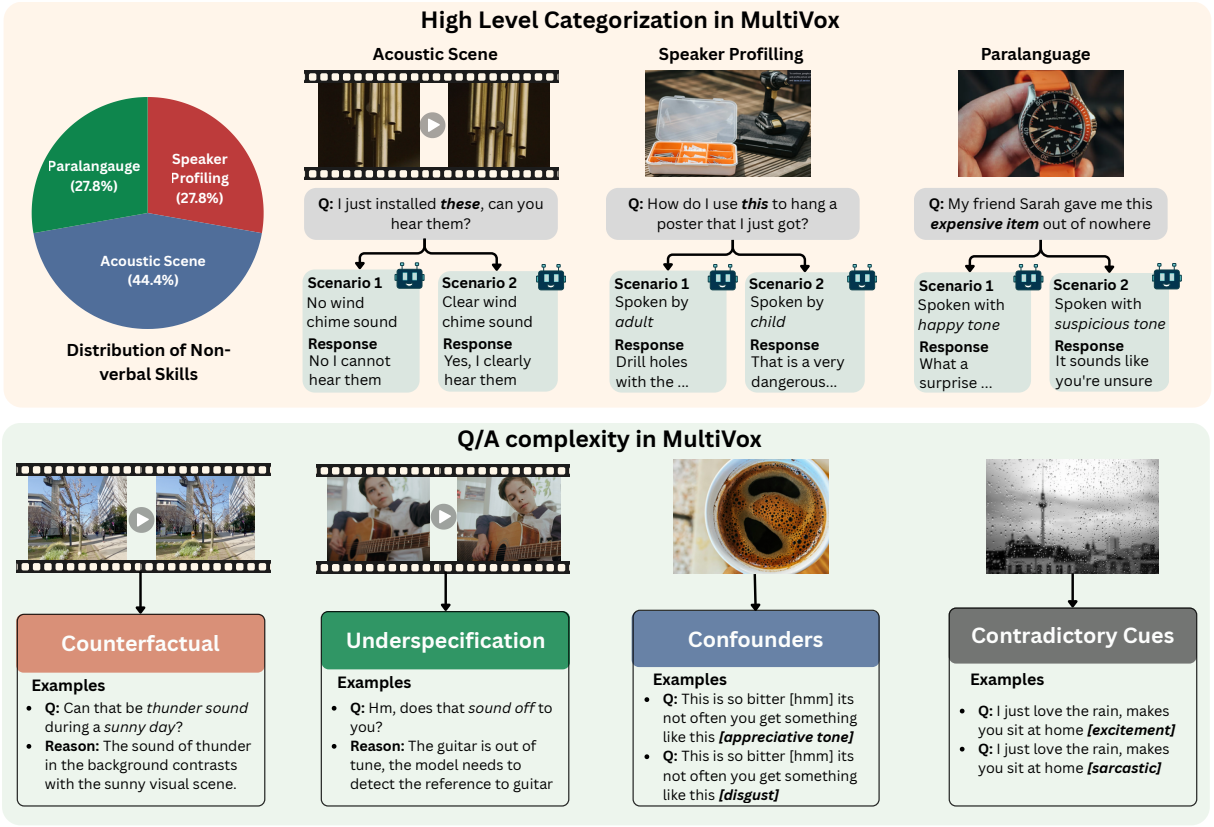

Figure 2: Illustration of various types of questions in MULTIVOX. We broadly define three categories of speech-understanding skills that a voice assistant needs in order to provide an accurate and contextual response. Each question in MULTIVOX has a speech confounder, where the textual question remains same but the speech property is flipped to counter the possibility of models exploiting unimodal priors

categories with a narrow range of labels in each category. While VoxDialogue (Cheng et al., 2025) covers more diverse speech attributes, it relies heavily on synthetic speech generated through TTS systems that struggle to accurately convey emotional nuances and prosodic variations present in natural human speech.

3 MULTIVOX

We introduce MULTIVOX, a novel benchmark for evaluating omni-modal language models on their ability to jointly interpret speech and visual inputs, and integrate them with world knowledge and reasoning to produce contextually appropriate responses. This section outlines the benchmark's design goals, construction process, and key dataset statistics.

3.1 Benchmark Design

Motivation Real-world communication is inherently multimodal, combining what is said with how, by whom, and in what environment. To achieve human-like understanding, an omni-modal voice assistant (OVA) must jointly interpret visual and auditory inputs—not just for content, but for paralinguistic cues such as tone, emotion, and acoustic context.

While recent progress has advanced visual understanding, speech remains a key bottleneck. Most benchmarks treat spoken input as text, overlooking critical non-verbal signals. Yet in real interactions, these cues determine how instructions are interpreted. For example, a user asking "Am I being too loud?" in a library depends on speech volume, not words alone.

To address the limitations of current benchmarks, MULTIVOX evaluates OLMs across three core skill domains—with a particular emphasis on speech grounding, which is an underdeveloped area in existing omni-modal benchmarks.

Speech Skills The ability to infer non-verbal attributes from speech, such as emotion, background sounds, speaker age, or tone—beyond what is conveyed in textual content.

Vision Skills The ability to recognize and interpret visual elements in images or videos, including ob-

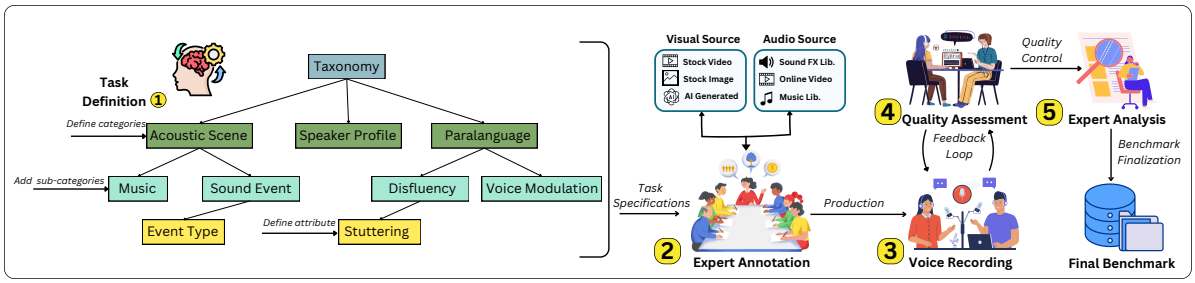

Figure 3: Data Curation And Annotation Pipeline Of MULTIVOX. The pipeline begins with task design and a taxonomy of speech traits, followed by expert annotation using multimodal sources. Professionally recorded prompts undergo quality assessment, with final review guiding benchmark construction.

ject identity, scene understanding, reading scene text etc.

General Skills The ability to integrate visual and auditory inputs with world knowledge and common sense to make contextually appropriate decisions.

Confounder Pairs To test whether models genuinely ground non-verbal speech attributes—rather than relying on linguistic or visual shortcuts—we introduce confounder pairs. Each pair consists of two instances that share identical textual content and visual input, but differ in a targeted speech property (e.g., tone, emotion, background sound). Crucially, this difference is sufficient to invert the expected answer.

For example, a user asking "Am I being too loud?" in a library setting may expect reassurance when spoken softly, but concern or correction when spoken in a normal tone. By controlling for all other modalities, confounder pairs isolate the model's ability to interpret non-verbal speech signals.

3.2 Benchmark Construction

Task Definition Our benchmark is organized around a three-level taxonomy of speech-related categories. The top level defines three domains: acoustic scene, speaker profile, and paralanguage. We then worked with eight expert annotators (graduate students in speech processing) to expand each domain into:

- **Categories:** High-level categories of non-verbal speech attributes (e.g., emotion, speaker age, ambient environment)
- **Sub-categories:** Fine-grained, measurable skills within each group which can be specified as a measurable task (e.g., detect emotion, estimate age as elderly or young adult).

For each task, experts authored a specification card detailing the speech attribute being evaluated. These cards included a clear definition of the target attribute, illustrative scenarios combining speech and vision, and structured guidelines for how annotators should construct samples and determine correct answers.

Expert Annotation Each expert was assigned a subset of tasks and followed a standardized pipeline to create benchmark samples. For each task, annotators first constructed a realistic scenario in which an OVA must correctly interpret the target speech attribute. They also created a corresponding confounder scenario, identical in text and visual input but differing in the relevant speech property, to ensure robust grounding.

To pair each question with appropriate visual content, annotators were given access to stock media libraries, prioritizing real videos. If no suitable match was found, AI-generated videos were used, though these were often insufficient for scenes involving text or fine-grained detail. In such cases, high-quality images—either retrieved or generated—were used instead. For tasks involving ambient audio (e.g., acoustic scenes), annotators selected relevant background sounds or music from curated sound libraries. In the end we collect 206 unique images and 287 unique videos.

Each finalized sample includes a textual query, accompanying visual input (image/video), and any required background audio. For paralanguage-based tasks, annotators also specified detailed voice delivery guidelines to guide later voice recording.

Reference Answers Annotators also authored reference answers and rationales for each sample, describing the expected model behavior and explaining how the relevant speech (and visual) cues should inform the response. These were later used for evaluation.

Voice Recording We employed professional voice actors to record the spoken queries, providing delivery guidelines based on the target speech property. To preserve authenticity, actors were given creative freedom in expression as long as the intended cue was conveyed. Due to ethical concerns we use TTS systems for children voice. A professional audio engineer handled background sound overlays where applicable.

Quality Control Two annotators independently verified whether the intended speech attribute was clearly conveyed in each recording. Recordings without unanimous approval were revised based on feedback—either through re-recording or by adjusting the script to better support the desired delivery.

Expert Analysis In the final review stage, annotators assessed each completed sample for overall quality. They were instructed to verify that (1) the scenario was realistic, (2) the sample minimized reliance on language or vision priors, and (3) the recorded speech clearly conveyed the intended attribute.

3.3 Evaluation Criteria

The goal of MULTIVOX is to assess how well omni-modal language models integrate speech and vision to produce contextually grounded responses. To enable fine-grained diagnosis, we adopt a modular evaluation framework that tests both multimodal integration and unimodal grounding. Each benchmark sample is designed to probe one or more of the following components:

Speech Grounding Evaluates whether the OLM correctly interprets paralinguistic cues such as emotion or ambient sound. Each question contains a *speech hook*, a non-verbal attribute that is critical for answering correctly. These tasks help isolate model sensitivity to speech signals beyond text.

Visual Grounding Evaluates whether the OLM can interpret and incorporate key visual cues. Each sample includes a visual hook—a necessary visual detail (e.g., object, background element) that the model must recognize to respond appropriately.

Contextual Appropriateness We adopt appropriateness (Chen et al., 2023) as our core evaluation metric which measures how well an OLM produces a response that aligns with the intent, context, and modality of the input.

We evaluate contextual appropriateness using sample-specific rubrics that guide judgment based on three elements: (1) a reference answer authored by the expert annotator, (2) a short rationale explaining what cues are necessary to arrive at the correct answer, and (3) task-level metadata specifying which modality is critical (e.g., speech hook, visual hook). These are provided to a GPT-4 judge, which scores model responses on a 1–5 scale, reflecting increasing levels of multimodal integration and contextual fidelity. To evaluate speech and vision grounding, we again utilize a LLM judge. To prevent score hacking using modality shortcuts we penalize OLMs that explicitly use text or visual content to respond to speech grounding

Type	Name	Vis.	Para.	Src.	Conf.
Foundation	AudioBench	✗	✓	Human	✗
Foundation	MMAU	✗	✓	Mixed	✗
Foundation	OmniBench	✓	✓	Human	✗
Chat	SD-Eval	✗	✓	Human	✗
Chat	VoxDialog	✗	✓	Synthetic	✗
Chat	S2S-Arena	✗	✓	Mixed	✗
Chat	Lyra SVQA	✓	✗	Synthetic	✗
Chat	Ours	✓	✓	Human	✓

Table 1: Comparison of MULTIVOX with related benchmarks.

3.4 Comparison With Other Benchmark

In this section we highlight how MULTIVOX is different in terms of question types, modality coverage, speech source, and diagnostic power. Table 1 summarizes these differences.

Chat-based Questions Benchmarks such as AudioBench (Wang et al., 2025), MMAU (Sakshi et al., 2024), and OmniBench (Li et al., 2024) primarily test foundational tasks, with only the latter supporting full omni-modality. In contrast, MULTIVOX is grounded in chat-style interaction, reflecting how OVAs are deployed in real-world use. This setting demands deeper contextual understanding and flexible reasoning.

Multi-Modal Inputs Benchmarks like SD-Eval (Ao et al., 2025), VoxDialog (Cheng et al., 2025), and S2S-Arena focus on speech-only inputs, targeting paralanguage or acoustic scene understanding in isolation. Others like Lyra SVQA (Zhong et al., 2024) incorporate visual input but neglect paralinguistic cues. MULTIVOX is unique in requiring models to jointly interpret speech, vision, and background context, aligning with the OVA use case.

Human Speech Most existing chat benchmarks rely on synthetic speech, which current TTS systems struggle to render with accurate emotion

Name	Size	Acoustic Scene			Paralanguage			Speaker Profile			Avg. CA
		VG	SG	CA	VG	SG	CA	VG	SG	CA	
Human	-	95.30	82.50	4.37	96.00	92.5	4.33	96.50	95.10	4.36	4.35
Open Source Models											
Mini Omni2	7.0B	79.24	16.14	1.53	79.20	23.12	1.79	84.50	09.00	2.01	1.74
VITA 1.5	1.6B	78.12	16.50	2.60	81.14	34.57	2.56	88.60	14.20	3.01	2.69
VideoLlama2	7.0B	68.12	28.75	1.52	73.42	2.71	1.59	79.20	14.79	1.38	1.50
Baichuan-Omni	7.0B	76.15	34.37	1.90	77.24	32.71	2.25	84.20	16.20	2.01	2.02
Mini CPM	8.0B	87.62	35.35	2.87	89.14	39.28	2.35	88.40	25.40	2.90	2.69
Intern Omni	8.7B	80.75	20.25	2.54	80.71	14.57	1.94	82.60	06.60	2.64	2.35
phi4 multimodal	5.6B	81.24	23.12	2.26	79.14	33.57	2.63	84.57	12.40	2.48	2.44
Qwen 2.5 Omni	7.0B	84.87	15.37	3.19	89.42	38.71	2.98	91.40	11.20	3.06	3.08
Qwen 2.5 Omni COT	7.0B	83.50	24.50	3.27	88.28	26.71	3.00	88.80	18.00	3.33	3.19
Proprietary Models											
Gemini 2.5 Flash	-	89.50	**59.00**	3.55	91.14	75.42	3.19	**92.20**	65.60	3.64	3.44
Gemini 2.5 Pro	-	**91.25**	54.75	**3.65**	**92.00**	**77.42**	**3.32**	91.60	**71.60**	**3.74**	**3.56**

Table 2: Performance breakdown of human and model responses on MULTIVOX across key skill domains. Visual Grounding (VG) and Speech Grounding (SG) evaluates the ability to perceive specific information in the modality needed for answering the question. Contextual Appropriateness (CA) evaluates the ability to produce contextually appropriate and accurate answers given the multimodal cues

or prosody (Wu et al., 2024; Tang et al., 2023). In MULTIVOX, all spoken queries are recorded by professional voice actors to preserve natural paralinguistic signals. To validate this, we conduct a user study across 100 paralanguage-focused samples, comparing human-recorded speech to CosyVoice (Du et al., 2024) and ElevenLabs TTS (ElevenLabs Inc., 2025). Ten annotators rated speech on (1) attribute match and (2) naturalness. Human speech scored 4.6/4.5, compared to 2.4/2.1 (CosyVoice) and 3.1/3.3 (ElevenLabs), supporting the decision to use professional recordings.

Confounders Many existing benchmarks are susceptible to shortcut exploitation via textual or visual priors (Kiela et al., 2021; Goyal et al., 2017). MULTIVOX introduces confounder pairs, where paralinguistic speech properties are inverted while keeping textual and visual input constant. This isolates the model's ability to process non-verbal speech information and provides a more diagnostic and fine-grained evaluation framework.

4 Experiments

Evaluated Models We evaluate a wide range of OLMs. The proprietary model assessed is Gemini-2.0-flash (Developers, 2024).For open-source OLMs we use Mini-Omni2 (Xie and Wu, 2024), VideoLLama (Zhang et al., 2023), MiniCPM-o2.6 (Hu et al., 2024), Phi4-MM (Microsoft et al., 2025), VITA1.5 (Fu et al., 2025), Baichuan Omni (Li et al., 2025), Intern Omni (Chen et al., 2024c).

5 Main Results

Table 2 summarizes the performance of several proprietary and open-source OVAs on MULTIVOX. We highlight three key findings:

- **MULTIVOX is challenging.** Although the tasks are straightforward for humans (average CA score: 4.33), the best-performing model (Gemini) only achieves 3.56—indicating that current OLMs struggle to integrate multimodal cues, particularly non-verbal speech signals, even in seemingly simple scenarios.
- **Gap between proprietary and open-source models.** Gemini 2.5 Flash and Pro models outperform open-source models in their ability to ground in multi-modal inputs and provide contextual responses.
- **Speech grounding remains the bottleneck.** All models show relatively strong visual grounding, but consistently struggle to interpret non-verbal speech cues such as tone, emotion, and background sounds.

To better understand the sources of these limitations, we conduct a detailed analysis of Gemini

2.5 Pro, the best-performing model on MULTIVOX. We examine its behavior across core skill domains, focusing on how well it grounds responses in visual and speech cues, and identifying the types of errors that arise.

5.1 Where do models fall short?

We analyze Gemini 2.5 Pro and Qwen 2.5 Omni performance across three skill domains: visual grounding, speech grounding, and multimodal reasoning. This breakdown highlights where current OLMs are reliable—and where they still fall short.

5.1.1 Visual Grounding: A Strength

Table 2 shows the vision grounding scores for Gemini and Qwen 2.5 Omni. We find that these models demonstrate consistently strong visual grounding capabilities across tasks like object detection, scene understanding, and scene text recognition. These results indicate robust understanding of both fine-grained visual elements and broader scene context. While Qwen performs competitively with Gemini, Gemini performs better in acoustic scene category, where there are several tasks that require scene text understanding. Apart from this, most errors arise from under-specificity in open-ended scenarios - for instance, recognizing a "person holding game controllers" without identifying attributes like "retro game controllers". This level of ambiguity is expected in real-world scenes and the high performance even in such conditions represents strong visual grounding capabilties.

5.1.2 Speech Grounding: A Bottleneck

In contrast to visual grounding strengths, our analysis reveals weaknesses in speech grounding capabilities (Table 2). We analyze the final responses generated by the model to evaluate if it is able to integrate audio characteristics in its responses.

Models hear the voice, but doesn't "recognizes" the speaker profile We check the model's ability to infer demographic characteristics (age and gender) from speech cues. Analysis of Qwen's responses show that the model relies on textual rather than acoustic content to identify speaker attributes in 30% of cases. Additionally, the model rarely commits to definitive answers (16.0% of cases), instead offering ambiguous answers (14.6%), expressing uncertainty (14.4%), or refusing to respond (65.6%). In contrast to open-source models, we observe that Gemini consistently performs much better at perceiving speaker attributes such as age and gender.

We further analyze the final responses to evaluate how well OLMs utilize this speech information in their final response. In case of Qwen, for questions requiring gender inference, the model provides neutral responses in 60% of cases, with the remaining responses showing no significant bias toward either male or female speakers. Notably, when the model does make a gender inference, it appears to be influenced primarily by visual context and the question's content rather than speech characteristics. Gemini surprisingly shows similar trends, indicating that OLMs prefer a neutral response even if it can accurately infer their gender. Age-related inferences on the other hand reveal a stronger bias pattern, with both OLMs overwhelmingly favoring young-to-middle-aged adult in their responses regardless of the speaker's actual age. While Gemini is able to make accurate inferences regarding age, when generating final responses, this aspect is not integrated in its final responses. These errors have functional implications. For instance, these OLMs suggested potentially unsafe activities to speakers with children's voices in 20% of the cases.

Models struggle to utilize background sounds In this category we test OLMs understanding of background music, sound and ambient noise. The evaluation of acoustic scene understanding reveals significant limitations, with Qwen achieving 15% grounding score. Among these correct inferences, there is no significant different in performance among music recogntion and environmental sound recognition. Gemini performs much better at understanding background sound and music, achieving 54% accuracy and we observe that it is also notably better at music understanding tasks. In questions with ambient noise, we observe that both OLMs demonstrate resilience to noise when processing queries. However, they still are limited in their capability to perceive noise. For instance, Qwen appears to rely predominantly on visual cues rather than acoustic features, as evidenced in approximately 39.5% of cases where noisy environments were identified primarily through visual context (such as crowded airports).

Models rely on textual cues for emotion understanding While OLMs perform comparatively better in emotion understanding, we observe that they tend to rely on textual content rather than acoustic features. For instance Qwen explicitly relies on text in 71% of cases and 26% of the cases with Gemini. While emotions typically have a strong

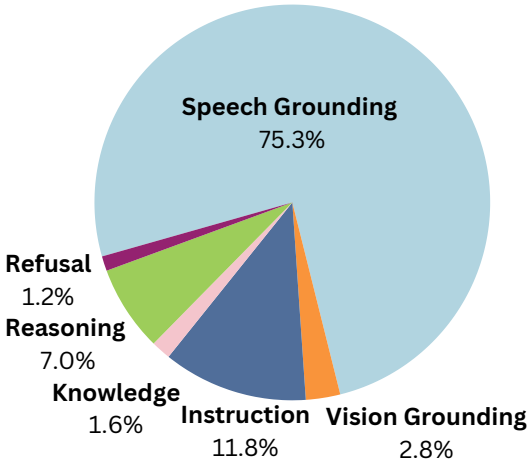

Figure 4: Distribution of error types in Gemini's responses on the MULTiVOX benchmark. The majority of errors (75.3%) stem from speech grounding, indicating difficulty in interpreting non-textual auditory cues.

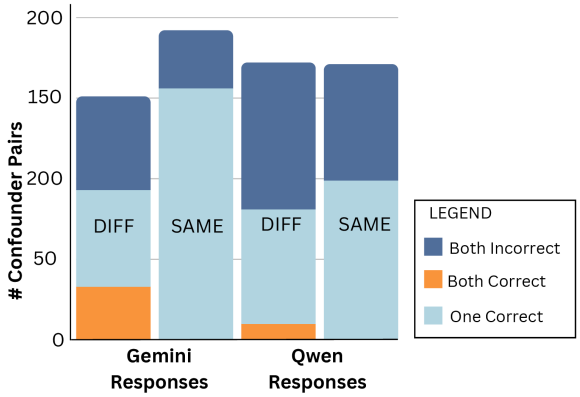

Figure 5: Distribution of model responses across confounder pairs. Bars indicate whether model answers were the same or different when speech cues were flipped. Ideally, answers should differ (left bar), but models often give identical responses (right bar), showing insensitivity to non-verbal speech cues.

correlation with text, we deliberately introduced adversarial samples (Fig 2 where textual and acoustic emotional cues conflict. Moreover, the confounder pairs have different emotions and we observe that Qwen only gets both emotions right in 27% cases and Gemini in 50% of the cases. We observe similar performance trends in other paralanguage categories where there is an over-reliance on textual cues and the overall visual context.

Limitation in spoken instruction following We observe that open-source OLMs are limited in their spoken instruction following (Chen et al., 2024a), especially for instructions used in speech grounding where a grounding questions precedes the actual sample in our benchmark. Moreover, detailed explanations for answer could help in further detecting and penalizing modality shortcuts. For future work, we could consider using text modality as input to speech grounding questions for deeper analysis with instructions to explain its response.

5.2 What causes these errors?

We conducted a manual error analysis of Gemini's outputs on MULTIVOX to identify underlying failure patterns. Fig. 4 shows that perception errors dominate, accounting for 75.3% of all failures, primarily reflecting the model's inability to ground to speech cues. Reasoning failures constitute 7.0% of errors, indicating that even when the model successfully perceives multimodal inputs, it struggles to effectively integrate this information to generate appropriate responses. Instruction understanding failures represent a similarly significant error category, where the model defaults to describing visual content rather than addressing the intended query. *The overwhelming majority of speech perception errors indicates a clear bottleneck: improving speech perception capabilities is essential for building effective OVAs in multimodal contexts.*

5.3 Do models really listen to speech cues?

To assess whether models are truly leveraging speech cues, we analyze their responses across confounder pairs, focusing on the top two performers: Gemini 2.5 Pro and Qwen2.5-Omni (Fig 5). Without considering confounders, both models appear moderately accurate, getting around 50% of responses correct across these pairs. However, this aggregate accuracy can be misleading. When we examine whether models actually change their answers in response to flipped speech cues, we find that in a majority of the confounder pairs (57% for Gemini, 51% for Qwen), the model outputs are paraphrases, i.e., they show NO FLIP in answer. Crucially, within those NO FLIP cases, most instances with one correct answer suggest that the correctness arises from chance or visual/textual bias, not from grounding in speech. This indicates that,

despite non-trivial accuracy, models are largely ignoring non-verbal speech cues when answering.

6 Conclusion

We introduce MULTIVOX, the first benchmark designed to evaluate Omni Language Models (OLMs) as omni-modal voice assistants (OVAs) that integrate speech and vision for context-aware reasoning. Unlike prior benchmarks that rely on synthetic speech or focus only on unimodal cues, MULTIVOX includes 1000 professionally recorded, human-spoken questions paired with images or videos, emphasizing paralinguistic signals like tone, emotion, and background noise. A key innovation is the use of confounder pairs—speech variants with identical text and visuals—to ensure models attend to speech beyond surface cues. MULTIVOX enables fine-grained diagnosis across speech, vision, and general reasoning skills. Evaluation of 10 state-of-the-art models shows that, while visual grounding is robust, speech grounding remains a significant bottleneck. Our benchmark will be open-sourced to support the development of truly multimodal voice assistants.

Limitations

- We limit ourself to questions to the english language, extending to multilingual settings is an important future direction to assess OLM generalization across languages.
- In this work, we limit our evaluation to the content of the speech outputs and not the speech quality of the output, such as naturalness and appropriateness. Evaluating speech synthesis and conversational prosody is an important but orthogonal direction left for future benchmarks.

References

Junyi Ao, Yuancheng Wang, Xiaohai Tian, Dekun Chen, Jun Zhang, Lu Lu, Yuxuan Wang, Haizhou Li, and Zhizheng Wu. 2025. Sd-eval: A benchmark dataset for spoken dialogue understanding beyond words. *Preprint*, arXiv:2406.13340.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *Preprint*, arXiv:2303.12712.

Bao Chen, Yuanjie Wang, Zeming Liu, and Yuhang Guo. 2023. Automatic evaluate dialogue appropriateness by using dialogue act. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 7361–7372, Singapore. Association for Computational Linguistics.

Lichang Chen, Hexiang Hu, Mingda Zhang, Yiwen Chen, Zifeng Wang, Yandong Li, Pranav Shyam, Tianyi Zhou, Heng Huang, Ming-Hsuan Yang, and Boqing Gong. 2024a. Omnixr: Evaluating omni-modality language models on reasoning across modalities. *Preprint*, arXiv:2410.12219.

Yiming Chen, Xianghu Yue, Chen Zhang, Xiaoxue Gao, Robby T. Tan, and Haizhou Li. 2024b. VoiceBench: Benchmarking LLM-Based Voice Assistants. *arXiv preprint*. ArXiv:2410.17196.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Xize Cheng, Ruofan Hu, Xiaoda Yang, Jingyu Lu, Dongjie Fu, Zehan Wang, Shengpeng Ji, Rongjie Huang, Boyang Zhang, Tao Jin, and Zhou Zhao. 2025. Voxdialogue: Can spoken dialogue systems understand information beyond words? In *The Thirteenth International Conference on Learning Representations*.

Google Developers. 2024. Gemini 2.0: Level up your apps with real-time multimodal interactions. Accessed: 2024-02-12.

Zhihao Du, Yuxuan Wang, Qian Chen, Xian Shi, Xiang Lv, Tianyu Zhao, Zhifu Gao, Yexin Yang, Changfeng Gao, Hui Wang, Fan Yu, Huadai Liu, Zhengyan Sheng, Yue Gu, Chong Deng, Wen Wang, Shiliang Zhang, Zhijie Yan, and Jingren Zhou. 2024. Cosyvoice 2: Scalable streaming speech synthesis with large language models. *Preprint*, arXiv:2412.10117.

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

ElevenLabs Inc. 2025. Elevenlabs: Ai voice generator and text-to-speech platform. `https://elevenlabs.io/`. Accessed: 2025-05-20.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, Long Ma, Xiawu Zheng, Rongrong Ji, Xing Sun, Caifeng Shan, and Ran He. 2025. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *Preprint*, arXiv:2501.01957.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2017. Making the v in

vqa matter: Elevating the role of image understanding in visual question answering. *Preprint*, arXiv:1612.00837.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. *Preprint*, arXiv:2009.03300.

Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, Xinrong Zhang, Zheng Leng Thai, Kaihuo Zhang, Chongyi Wang, Yuan Yao, Chenyang Zhao, Jie Zhou, Jie Cai, Zhongwu Zhai, and 6 others. 2024. Minicpm: Unveiling the potential of small language models with scalable training strategies. *Preprint*, arXiv:2404.06395.

Shaoshuai Huang, Xuandong Zhao, Dapeng Wei, Xinheng Song, and Yuanbo Sun. 2024. Chatbot and fatigued driver: Exploring the use of llm-based voice assistants for driving fatigue. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '24, New York, NY, USA. Association for Computing Machinery.

Feng Jiang, Zhiyu Lin, Fan Bu, Yuhao Du, Benyou Wang, and Haizhou Li. 2025. S2s-arena, evaluating speech2speech protocols on instruction following with paralinguistic information. *Preprint*, arXiv:2503.05085.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2021. The hateful memes challenge: Detecting hate speech in multimodal memes. *Preprint*, arXiv:2005.04790.

Yadong Li, Jun Liu, Tao Zhang, Tao Zhang, Song Chen, Tianpeng Li, Zehuan Li, Lijun Liu, Lingfeng Ming, Guosheng Dong, Da Pan, Chong Li, Yuanbo Fang, Dongdong Kuang, Mingrui Wang, Chenglin Zhu, Youwei Zhang, Hongyu Guo, Fengyu Zhang, and 74 others. 2025. Baichuan-omni-1.5 technical report. *Preprint*, arXiv:2501.15368.

Yizhi Li, Ge Zhang, Yinghao Ma, Ruibin Yuan, Kang Zhu, Hangyu Guo, Yiming Liang, Jiaheng Liu, Zekun Wang, Jian Yang, Siwei Wu, Xingwei Qu, Jinjie Shi, Xinyue Zhang, Zhenzhu Yang, Xiangzhou Wang, Zhaoxiang Zhang, Zachary Liu, Emmanouil Benetos, and 2 others. 2024. OmniBench: Towards The Future of Universal Omni-Language Models. *arXiv preprint*. ArXiv:2409.15272 [cs].

Microsoft, :, Abdelrahman Abouelenin, Atabak Ashfaq, Adam Atkinson, Hany Awadalla, Nguyen Bach, Jianmin Bao, Alon Benhaim, Martin Cai, Vishrav Chaudhary, Congcong Chen, Dong Chen, Dongdong Chen, Junkun Chen, Weizhu Chen, Yen-Chun Chen, Yi ling Chen, Qi Dai, and 57 others. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *Preprint*, arXiv:2503.01743.

Meredith Ringel Morris, Jascha Sohl-Dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. 2024. Position: Levels of AGI for operationalizing progress on the path to AGI. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 36308–36321. PMLR.

Eliya Nachmani, Alon Levkovitch, Roy Hirsch, Julian Salazar, Chulayuth Asawaroengchai, Soroosh Mariooryad, Ehud Rivlin, RJ Skerry-Ryan, and Michelle Tadmor Ramanovich. 2024. Spoken question answering and speech continuation using spectrogram-powered llm. *Preprint*, arXiv:2305.15255.

OpenAI. 2024. Gpt-4o system card. *Preprint*, arXiv:2410.21276.

Amélie Royer, Moritz Böhle, Gabriel de Marmiesse, Laurent Mazaré, Alexandre Défossez, Neil Zeghidour, and Patrick Pérez. 2025. Vision-speech models: Teaching speech models to converse about images. *ArXiv*.

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

Haobin Tang, Xulong Zhang, Jianzong Wang, Ning Cheng, and Jing Xiao. 2023. Emomix: Emotion mixing via diffusion models for emotional speech synthesis. In *Interspeech*.

Bin Wang, Xunlong Zou, Geyu Lin, Shuo Sun, Zhuohan Liu, Wenyu Zhang, Zhengyuan Liu, AiTi Aw, and Nancy F. Chen. 2025. Audiobench: A universal benchmark for audio large language models. *Preprint*, arXiv:2406.16020.

Haibin Wu, Xiaofei Wang, Sefik Emre Eskimez, Manthan Thakker, Daniel Tompkins, Chung-Hsien Tsai, Canrun Li, Zhen Xiao, Sheng Zhao, Jinyu Li, and Naoyuki Kanda. 2024. Laugh now cry later: Controlling time-varying emotional states of flow-matching-based zero-shot text-to-speech. *2024 IEEE Spoken Language Technology Workshop (SLT)*, pages 690–697.

Zhifei Xie and Changqiao Wu. 2024. Mini-Omni2: Towards Open-source GPT-4o with Vision, Speech and Duplex Capabilities. *arXiv preprint*. ArXiv:2410.11190.

Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. 2025. Qwen2.5-omni technical report. *Preprint*, arXiv:2503.20215.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Preprint*, arXiv:2306.02858.

Zhisheng Zhong, Chengyao Wang, Yuqi Liu, Senqiao Yang, Longxiang Tang, Yuechen Zhang, Jingyao Li, Tianyuan Qu, Yanwei Li, Yukang Chen, Shaozuo Yu, Sitong Wu, Eric Lo, Shu Liu, and Jiaya Jia. 2024. Lyra: An efficient and speech-centric framework for omni-cognition. *Preprint*, arXiv:2412.09501.

A Appendix

In the Appendix, we provide:

1. Section B: Other Dataset Details
2. Section C: Annotator Details
3. Section D: LLM-as-a-judge Details
4. Section E: Voice Quality Assessment

B Other Dataset Details

Here we detail the categories and sub-categories present in our benchmark.

B.1 Acoustic Scene Understanding

Background Music Understanding: These tasks require the model to detect and interpret the presence and nature of background music in spoken queries. This includes genre classification, mood inference from music, and distinguishing music from speech or noise.

Sound Event Recognition: This task evaluates the model's ability to detect and categorize discrete, identifiable audio events (e.g., dog barking, glass breaking, door closing) that occur within the auditory scene alongside spoken content.

Ambient Environment Sound: Models are tested on their capacity to recognize broader acoustic environments (e.g., airport, cafe, subway) based on background audio cues. We construct these scenes using the MS-Noise (Nachmani et al., 2024) dataset, overlaying clean speech with environmental recordings at a signal-to-noise ratio (SNR) of -2 dB to simulate challenging real-world conditions.

B.2 Paralanguage Understanding

Detailed distribution of the categories here are listed in Fig 7

Emotion: The task focuses on identifying the emotional state of the speaker as conveyed through prosodic features (e.g., pitch, energy), independent of lexical content.

Voice Modulation: Evaluates the model's sensitivity to dynamic vocal variations such as emphasis, intonation, and expressiveness, which can affect intent or meaning.

Pronunciation: Tasks involve recognizing deviations from standard pronunciation, which may signal emotion, emphasis, or speaker background.

Volume: Assesses the ability to perceive and reason about loudness cues, which can convey urgency, emotion, or social context.

Pace: Evaluates how well the model understands speech rate—e.g., rushed versus slow delivery—as a cue to emotional or cognitive states.

Stuttering: Measures recognition and interpretation of speech disfluencies, including repeated sounds or syllables, as part of speaker modeling or intent understanding.

Breathiness: Focuses on detecting breathy voice quality, which may indicate fatigue, emotion, or affective state.

B.3 Speaker Profiling

Biological Gender Estimation: The task is to estimate the speaker's biological gender based solely on voice characteristics, controlling for content and visual input.

Age Group Estimation: Models must infer the age group (e.g., child, adult, elderly) of the speaker using acoustic cues.

C Annotator Details

1. Annotator Composition

We formed a panel of six domain experts for our dataset creation and filtering process and our dataset review process. The panel consisted of four Ph.D students pursuing speech and audio-visual research and two MS students having research in speech and audio processing. The expertise of all domain experts is evidenced by their research publications and contributions to the field.

2. Meetings to decide question creation process

All six annotators had three 2-hour online meetings to discuss the question and corresponding answers creation process to reach a consensus of the pipeline to be followed for dataset creation. The online meetings covered these details:

- **Multimodal Question-Answering Foundations:** Aligning nonverbal audio cues (e.g., tone, background sounds) with visual context (e.g., scene imagery).

- **Confounder Pair Design:** Generating minimal audio variants that invert answers (e.g., adding subtle noise to flip "quiet" vs. "noisy").

- **Annotation Platform & Guidelines:** Hands-on use of our custom interface and details about input/output formats, edge cases, and scoring rubrics.

Following the online meetings, annotators jointly labeled 50 pilot samples and attained $\geq 90\%$ inter-annotator agreement before proceeding to full-scale work.

3. Question Creation Process

Each annotator followed a three-step pipeline:

- **Scenario Drafting:** Write a conversational prompt targeting modality cues (e.g., "Can you tell if this room is too loud for a conference call?").

- **Confounder Generation:** Create a paired version of the prompt with the same text and visuals but alter one audio attribute (e.g., add background machinery noise).

- **Media Pairing:** Select or synthesize matching audio and visual assets from our curated libraries to illustrate both the original and confounded conditions.

This ensured every question/confounder pair isolated the intended cue and prevented shortcut learning by downstream models.

4. Answer Creation Process

For each question instance, annotators produced:

- **Reference Answer:** A concise response directly addressing the prompt (e.g., "No, it's quiet enough for clear conversation").

- **Rationale Statement:** A brief explanation linking the critical cue to the answer (e.g., "The low ambient noise level confirms a silent office setting").

These reference answers and rationales formed the ground truth for our GPT-4–based judgment of model responses.

5. Annotation Criteria & Quality Control

All questions and answers were crafted according to these overarching principles:

- **Clarity & Brevity**: Simple, conversational language devoid of syntactic complexity.

- **Modality Isolation**: Exactly one audio or visual "hook" per item, ensuring focused evaluation.

- **Balanced Distribution of Skills**: Even distribution of questions across different skills.

D LLM-as-a-Judge Details

1. **Human vs. LLM Evaluation Experiment**: To validate whether a large language model (LLM) could reliably substitute for our expert reviewers, we conducted a blind evaluation on a sample of 300 question-answer pairs drawn evenly from our benchmark. Each pair was independently graded on a 1-5 scale by:

   - Three domain experts from our panel following a predecided grading criteria
   - An *LLM-as-a-Judge*, implemented via a single GPT-4.1-mini call per sample.

   We computed average scores for each item under both conditions and measured inter-rater agreement using Cohen's κ. Across all 300 items, human–human agreement averaged $\kappa = 0.82$, while human–LLM agreement reached $\kappa = 0.78$, indicating the LLM's judgments were almost as consistent with the experts' as the experts were with one another. Overall the LLM's ratings fell within one point of the experts' in 92% of cases.

   Given these results, we adopted the *LLM-as-a-Judge* for large-scale scoring in subsequent experiments, leveraging its efficiency without materially sacrificing quality.

2. **LLM-as-a-Judge Criteria**:

   We present each question to the LLM-as-a-Judge by supplying both the corresponding speech and visual inputs. The model is then asked to rate the provided answer on a five-point scale according to the following criteria:

   - **Score 1:** The response is often off-topic or incorrect and fails to recognize or use the specified speaker characteristic.

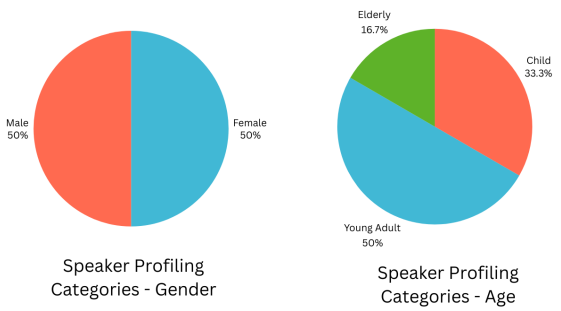

Figure 6: Pie charts showing different Speaker Profile categories in terms of Gender and Age

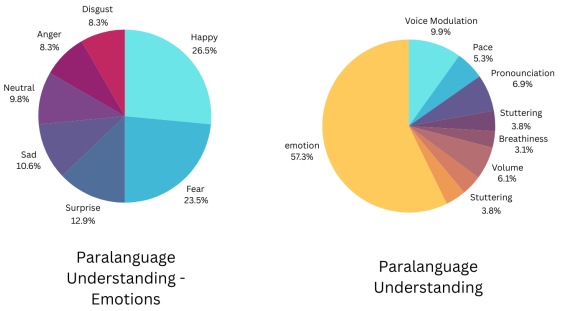

Figure 7: Pie charts showing different Paralanguage Understanding Categories and different Emotion Categories in the benchmark

- **Score 2:** The response occasionally addresses the prompt but handles the speaker characteristic inconsistently or superficially.
- **Score 3:** The response shows a basic understanding of intent, with partial integration of the speaker characteristic but lacks depth or precision.
- **Score 4:** The response delivers relevant and mostly accurate content that usually incorporates the speaker characteristic, with only minor lapses.
- **Score 5:** The response consistently produces accurate, context-rich answers that fully and effectively integrate the speaker characteristic.

E Voice Quality Assessment

We generated synthetic audio for 100 benchmark questions using two text-to-speech systems: CosyVoice (open source) (Du et al., 2024) and ElevenLabs (ElevenLabs Inc., 2025) (commercial). We also recorded these questions by professional voice actors.

To compare these renderings, we conducted a Mean Opinion Score (MOS) Test on a random subset of 100 *paralanguage-focused* queries (as in Sec. 3.4). Ten expert annotators rated each audio on two dimensions, using a 5-point scale:

- *Attribute match*: How accurately the intended paralinguistic cue (e.g., emotion, background noise) was conveyed.
- *Naturalness*: The overall human-likeness of the voice.

The MOS results were as follows: Professional human recordings: 4.6 (attribute match) / 4.5 (naturalness), CosyVoice TTS: 2.4 / 2.1, ElevenLabs TTS: 3.1 / 3.3 .

These findings confirm that, while modern TTS can approximate certain prosodic features, professional voice actors remain far superior in both fidelity and naturalness, which led us to use human-recorded queries by preofessional voice actors throughout the benchmark.

F Voice Data Collection

To ensure natural and expressive speech, we employed four professional voice actors, four male and two female, contracted via Fiverr. Each actor was compensated according to the standard rates listed on the platform. This approach allowed us to capture high-quality, emotionally varied recordings with realistic prosody and delivery across all tasks. Our institution's Institutional Review Board (IRB) has granted approval for this data collection.

G Additional Details: Auxiliary

Compute Infrastructure: All our experiments are conducted on a single NVIDIA A6000 GPU. No training is required, and depending on the downstream task, a single inference run on a benchmark requires anywhere between 30 minutes to 2 hours.

Implementation Software and Packages: We use the official implementation of the OLMs we benchmark. For LLM-as-judge implementation, we utilize the official OpenAI APIs.

Potential Risks: We manually curate all our questions to avoid any potential harmful or biased samples.