MMAU-Pro: A Challenging and Comprehensive Benchmark for Holistic Evaluation of Audio General Intelligence

Sonal Kumar[1*], Šimon Sedláček[2*], Vaibhavi Lokegaonkar[1*], Fernando López[3,4*], Wenyi Yu [5], Nishit Anand[1], Hyeonggon Ryu[6], Lichang Chen[1], Maxim Plička[2], Miroslav Hlaváček[7], William Fineas Ellingwood[8], Sathvik Udupa[2], Siyuan Hou[5], Allison Ferner[9], Sara Barahona[3], Cecilia Bolaños[10], Satish Rahi[11], Laura Herrera-Alarcón[3], Satvik Dixit [13], Siddhi Patil[1], Soham Deshmukh[12], Lasha Koroshinadze [1], Yao Liu[14], Leibny Paola Garcia Perera[15], Eleni Zanou[16], Themos Stafylakis[16], Joon Son Chung[6], David Harwath[17], Chao Zhang[5,18], Dinesh Manocha[1], Alicia Lozano-Diez[3], Santosh Kesiraju[2 †], Sreyan Ghosh[1†], Ramani Duraiswami[1†]

[1]University of Maryland, College Park, USA
[2] Brno University of Technology, Czech Republic
[3] Universidad Autónoma de Madrid
[4] Telefónica
[5] Tsinghua University
[6] KAIST, Daejeon
[7] Phonexia
[8] Middlebury College, USA
[9] Tufts University
[10] Universidad de Buenos Aires
[11] Indian Institute of Technology, Bombay
[12] Microsoft
[13] Carnegie Mellon University, USA
[14] Universiti Sains Malaysia
[15] Johns Hopkins University, USA
[16] Athens University of Economics and Business
[17] University of Texas, Austin, USA
[18] Shanghai Artificial Intelligence Laboratory
{sonalkum, sreyang}@umd.edu

Abstract

Audio comprehension-including speech, non-speech sounds, and music-is essential for achieving human-level intelligence. Consequently, AI agents must demonstrate holistic audio understanding to qualify as generally intelligent. However, evaluating auditory intelligence comprehensively remains challenging. To address this gap, we introduce MMAU-Pro, the most comprehensive and rigorously curated benchmark for assessing audio intelligence in AI systems. MMAU-Pro contains 5,305 instances, where each instance has one or more audios paired with human expert-generated question-answer pairs, spanning speech, sound, music, and their combinations. Unlike existing benchmarks, MMAU-Pro evaluates auditory intelligence across 49 unique skills and multiple complex dimensions, including long-form audio comprehension, spatial audio reasoning, multi-audio understanding, among others. All questions are meticulously designed to require deliberate multi-hop reasoning, including both multiple-choice and open-ended response formats. Importantly, audio data is sourced directly "from the wild" rather than from existing datasets with known distributions. We evaluate 22 leading open-source and proprietary multimodal AI models, revealing significant limitations: even state-of-the-art models such as Gemini 2.5 Flash and Audio Flamingo 3 achieve only 59.2% and 51.7% accuracy, respectively, approaching random performance in multiple categories. Our extensive analysis highlights specific shortcomings and provides novel insights, offering actionable perspectives for the community to enhance future AI systems' progression toward audio general intelligence.

Datasets — https://sonalkum.github.io/mmau-pro

Introduction

Comprehensive audio understanding-from spoken language to environmental sounds and music-is fundamental to human general intelligence. Correspondingly, AI systems must possess comparable capabilities for effective real-world interaction (Sakshi et al. 2025). Recent advancements in multimodal large language models (MLLMs) have led to the emergence of Large Audio-Language Models (LALMs), demonstrating notable audio comprehension skills (Ghosh

[*]Core Contributors
[†]Core advisors

Category	Example
Spatial QA	**Q:** Who's order does the waiter take first? **Opt:** (A) The person to the left of the mic holder (B) The person to the right... (C) The person in front... **A:** (A) The person to the left of the mic holder
Open-ended QA	**Q:** What is hyper-foreignism with respect to pronunciation according to the clip? **A:** When a speaker changes the way they say a word to sound more like that of the stereotype they hold for a foreign language
Voice QA	**Q:** Answer the question in the audio. **Opt:** (A) It sounds like this decision carries a lot of weight... (B) Perhaps the best approach is to systematically list... (C) Your tranquil state suggests... (D) Making major decisions during emotional... **A:** Your tranquil state...
Long Audio	**Q:** What did the winning team get at the end of the game? **Opt:** (A) They choose the next vacation destination (B) The other team has to get rid of their rooster (C) They win the other team's apartment (D) Tie... **A:** They win the other team's apartment
Multi-Audio	**Q:** What effect needs to be applied to the first recording to achieve sound of the second recording? **Opt:** (A) echo (B) distortion (C) phaser (D) reverb **A:** (D) reverb
Sound	**Q:** What trend can be observed in the weight of the cloths thrown in the audio? **Opt:** (A) Increasing (B) Decreasing (C) Constant (D) None **A:** (A) Increasing
Voice STEM QA	**Q:** Choose the correct option that answers the question in the audio... **Opt:** (A) 25% (B) 35% (C) 55% (D) 75% **A:** (B) 35%
Speech	**Q:** In the audio with roughly the same phrase being repeated, explain how the different tone effects the meaning... **Opt:** (A) Bored $\rightarrow$ enthusiastic $\rightarrow$ questioning $\rightarrow$ angry (B) Cheerful $\rightarrow$ playful $\rightarrow$ serious $\rightarrow$ stern (C) Sarcastic $\rightarrow$ sincere $\rightarrow$ doubtful $\rightarrow$ irritated (D) Genuine $\rightarrow$ sarcastic $\rightarrow$ questioning $\rightarrow$ frustrated. **A:** (D)
Instruction Following	**Instruction:** Explain what's happening... contain a title wrapped in double angular brackets... **Correct A:** <<Harmonica's Ascending and Descending Scale>> This audio clip features a harmonica playing the C major scale... **Wrong A:** This is an audio clip of a person playing the C major scale...
Music	**Q:** Can you guess the singer in this song? **Opt:** (A) Jeff Beck (B) Tenacious B (C) Jimmy Hendricks (D) Jack Black **A:** (D) Jack Black
Multicultural Music	**Q:** What *raag* is this *bandish* composed in? **Opt:** (A) Bhimpalasi (B) Kannada (C) Durga (D) Malkaunse (E) Bhairavi (F) Yaman Kalyan **A:** (C) Durga
Speech-Sound-Music	**Q:** Which of the following songs is made according to the speaker? **Opt:** (A) Not Like Us (B) Star Boy (C) All the stars (D) HUMBLE **A:** (D) HUMBLE

Table 1: Overview of the MMAU-Pro benchmark. MMAU-Pro provides comprehensive coverage across all three core audio domains-speech, sound, and music-and extends evaluation to their mixtures. It further includes multi-audio reasoning, long-form audio (up to 10 minutes), voice-chat QA, spatial audio understanding, open-ended QA, and multimodal instruction following, offering a broad and realistic assessment of audio intelligence.

et al. 2024, 2025b; Goel et al. 2025; Gong et al. 2024; Deshmukh et al. 2023; KimiTeam et al. 2025; Xie et al. 2025a; Xu et al. 2025; Chu et al. 2024). Despite numerous benchmarks assessing progress toward Artificial General Intelligence (AGI) through text, audio intelligence evaluation remains notably underserved. Given audio's inherent diversity and complexity, we contend that progress toward AGI is incomplete without strong audio intelligence capabilities-and that their rigorous evaluation remains an open challenge.

Recently, several benchmarks have emerged to evaluate LALMs. MMAU (Sakshi et al. 2025), a pioneering comprehensive benchmark, comprises 10,000 carefully selected audio clips across speech, sounds, and music, with single-turn, single-audio questions requiring knowledge and reasoning. Following MMAU, MMAR (Ma et al. 2025) introduced more challenging queries, while MMSU (Wang et al. 2025b) expanded spoken language understanding assessments. Domain-specific benchmarks like Speech-IFEval (Lu, Kuan, and Lee 2025) focus on instruction-following and CMM (Leng et al. 2024) focuses on hallucinations. Nevertheless, existing benchmarks inadequately represent the complexity of realistic auditory scenarios - such as multiple and overlapping audios, long-duration inputs, open-ended answers, and culturally varied content-which demand deeper comprehension and multi-hop reasoning beyond basic recognition.

Our Contributions. To this end, we present **MMAU-Pro**, a novel benchmark consisting of 5,305 expert-annotated instances designed to evaluate 49 distinct auditory intelligence skills spanning speech, environmental sounds, and music. MMAU-Pro presents challenges overlooked by prior benchmarks, including long-form audio understanding (up

to 10 minutes), reasoning across multiple clips, spatial audio perception, multicultural music interpretation, instruction-following abilities, etc. All questions are crafted to require deliberate multi-hop reasoning and include a balanced mix of multiple-choice and open-ended formats. To address the shortcomings of existing evaluation methodologies, we further propose a retrieval-based evaluation framework that enables more robust and reliable assessment. By emphasizing realistic and demanding auditory tasks, MMAU-Pro provides a comprehensive testbed to accelerate the development of auditory intelligence in multimodal AI systems. To summarize, our main contributions are:

- We introduce **MMAU-Pro**, the most comprehensive benchmark to date for evaluating auditory intelligence. It comprises 5,305 expert-annotated question–answer pairs spanning 49 distinct skills across speech, environmental sounds, music, and their mixtures. MMAU-Pro introduces novel challenges, including spatial audio reasoning, multi-clip audio reasoning, voice–chat comprehension, and tasks requiring prosodic, world-knowledge, and STEM-based reasoning. All audio samples are drawn from the wild, with durations up to ten minutes, significantly surpassing the short clips typical of prior benchmarks where current models are near-saturated.
- We benchmark over 15 open-source and proprietary multimodal LLMs on MMAU-Pro, finding that even the strongest models face substantial challenges. Gemini 2.5 Flash achieves only 59.2% accuracy; the best-performing fully open-source model, Audio Flamingo 3, reaches 51.7%; and the strongest open-weights omni model, Qwen2.5-Omni-7B-Instruct, achieves just 52.2%.
- We provide an in-depth analysis of model responses, uncovering key failure modes in auditory perception and reasoning. These include shallow audio grounding, degradation in text-only and STEM reasoning, poor performance in multi-audio and spatial reasoning, and limited understanding of multicultural music.

Related Work

Large Audio Language Models

Recent advances in multimodal modeling have led to (L)ALMs-models that pair audio perception with (L)LMs to tackle complex audio tasks. Early systems such as Whisper (Li et al. 2024a; Peng et al. 2023) and CLAP (Wu et al. 2023; Elizalde et al. 2023; Elizalde, Deshmukh, and Wang 2024) focused on foundational tasks like transcription, captioning, and retrieval, but struggled with reasoning-centric challenges. More recent models-GAMA (Ghosh et al. 2024), Audio Flamingo (Ghosh et al. 2025b; Goel et al. 2025), Mellow (Deshmukh et al. 2025), Phi-4MM (Abouelenin et al. 2025) Qwen2-Audio (Chu et al. 2024), and Audio-PALM (Rubenstein et al. 2023) proposed improved architectures and large-scale training, targeting deeper understanding. These efforts have culminated in large audio reasoning models (LARMs), including Audio-Reasoner (Xie et al. 2025a), SoundMind (Diao et al. 2025), R1-AQA (Li et al. 2025b), and Audio-CoT (Xie et al. 2025b), which explicitly model step-by-step reasoning. In parallel, general-purpose Omni-Language Models (OLMs) such as Qwen2.5-Omni (Xu et al. 2025), Baichuan-Omni (Li et al. 2024b, 2025c), and Ming-Omni (Gong et al. 2025) - though not tailored for audio-have demonstrated surprising proficiency on audio tasks. While progress is promising, robust benchmarks remain essential to evaluate audio intelligence.

Audio Benchmarks

Existing benchmarks provide strong foundations but fall short in evaluating holistic audio intelligence. MMAU (Sakshi et al. 2025) introduced 10,000 QA pairs across 27 skills for speech, sounds, and music, but used existing datasets and short, single-source clips-achieving only 52–60% accuracy. MMAR (Ma et al. 2025) added 1,000 real-world QA triplets with hierarchical reasoning layers and rationales, yet remained limited in scale and scope. AudioBench (Wang et al. 2025a) unifies 26 datasets across 8 tasks, while MuChoMusic (Weck et al. 2024) probes 1,100 MCQs on culturally diverse music, exposing models' over-reliance on text. MMSU (Wang et al. 2025b) tests 5,000 spoken-language QA pairs across 47 speech skills. Beyond Single-Audio (Chen et al. 2024b) evaluates multi-audio reasoning across 20 datasets, showing that most ALLMs struggle when reasoning over more than one audio stream. Dynamic-SUPERB Phase-2 (Huang et al. 2024) expands to 180 tasks.

While some recent models, such as Mellow (Deshmukh et al. 2025) and BAT (Zheng et al. 2024), begin to address multi- and spatial-audio tasks, benchmark evaluations remain shallow and fragmented. Moreover, no existing benchmark systematically tests instruction following or jointly evaluates long-form (up to 10 minutes), multi-audio, spatial, open-ended, and multicultural scenarios. Addressing these gaps, MMAU-Pro offers the most comprehensive benchmark to date, targeting underexplored dimensions critical to advancing real-world audio general intelligence.

The MMAU-Pro Benchmark

Overview

MMAU-Pro is designed to holistically evaluate audio intelligence in AI systems. It comprises 5,305 expert-annotated question–answer pairs covering 49 distinct skills. Table 2 summarizes the core statistics. Questions require multi-step, multi-hop reasoning and were authored and validated by domain experts to ensure high quality. To avoid data leakage, all audio (except our spatial subset) is sourced from in-the-wild recordings. For spatial audio reasoning, we reuse high-fidelity multi-channel recordings from the EasyCom dataset (Donley et al. 2021).

While prior benchmarks such as MMAU and MMAR primarily evaluate models using multiple-choice questions, MMAU-Pro extends evaluation to include open-ended responses and MCQs with up to 10 options, thereby substantially reducing the likelihood of models succeeding by random guessing. It also categorizes audio clips by duration: short ($\leq$30s), medium (30s–3min), long (3–8min), and

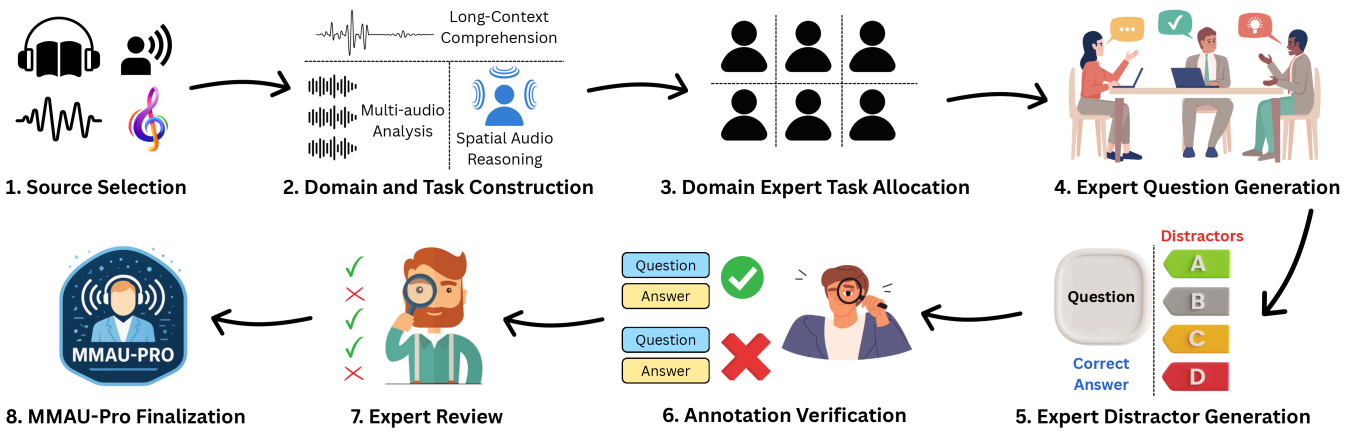

Figure 1: Overview of dataset-construction pipeline for MMAU-Pro.

ultra-long (8–10min), enabling characterization and analysis across varying temporal contexts.

Statistics	Number
Total Questions	5,305
Domains	11
Speech Questions	891
Sound Question	1654
Music Questions	1618
Sound-Speech Mix	88
Music-Speech Mix	46
Sound-Music Mix	50
Sound-Speech-Music Mix	7
Spatial Understanding Questions	325
Voice STEM Questions	94
Voice Prosodic Questions	96
Voice World Knowledge Questions	100
Instruction Following Questions	87
Multi-Audio QA (sound:speech:music)	247:111:72
Multiple Choice Questions	4593
Open-ended Questions	625
Average Audio Length	123.78 s
Durations (short:med:long:ultra-long)	2589:1897:1307:348

Table 2: Core statistics of the MMAU-Pro Benchmark.

Data Curation, Annotation and Validation

Inspired by prior benchmarks in this space, we design a specialized multi-stage pipeline with more human involvement in the process to construct high-quality data for MMAU-Pro.

1. **Domain & Task Design:** We define a diverse set of reasoning tasks across speech, sound, music, and their mixtures, including long-context comprehension, spatial reasoning, multi-audio analysis, and multicultural music understanding.
2. **Task Allocation:** Domain experts (authors) are assigned tasks based on specialization, guided by detailed instructions to ensure comprehensive domain coverage.
3. **QA Generation:** The experts then manually collect audio and craft QA pairs. The QA pairs are created with an emphasis on *multi-hop reasoning* and *real-world use cases*. We include both MCQs and open-ended questions in the benchmark, with explanatory answers authored for the latter. Annotators determine the appropriate format based on factors such as susceptibility to elimination via language cues and the added value of open-endedness for deeper evaluation.
4. **Distractor Construction:** For MCQs, experts create challenging distractors that discourage superficial pattern matching. Distractors are crafted to avoid trivial elimination and encourage careful audio grounding. Unlike other benchmarks, which resort to LLMs for the generation of distractors, experts carefully create distractors for each question to pose a higher level of challenge to the models being evaluated.
5. **Annotation Verification:** A second expert independently verifies each QA instance for accuracy, clarity, and reasoning validity. Discrepancies are resolved iteratively, followed by grammar and style checks using both experts and LLMs.
6. **Expert Review:** Final review ensures cultural sensitivity, task appropriateness, and explanatory quality, particularly for open-ended responses.
7. **Benchmark Finalization:** The finalized dataset balances domain, task type, and audio length to ensure diverse and representative coverage (Table 2).

Over 25 individuals were involved in this process of data collection, QA categorization and design, validation, curation, and evaluation.

Comparison and Task Coverage

Table 3 compares MMAU-Pro with existing popular benchmarks across core and novel evaluation dimensions. As shown, MMAU-Pro introduces several key advancements, including multi-audio reasoning, spatial audio understanding, and STEM-based evaluation. In the following subsections, we describe these core innovations in detail.

Capability	MMAU-Pro	MMAU	MMAR	AIR-Bench	AudioBench	MMSU	DynSuperb-1	DynSuperb-2
Long Audio Understanding	✓	×	×	×	✓	×	×	✓
Multi-Audio Understanding	✓	×	✓	×	×	×	×	×
Spatial Audio Understanding	✓	×	✓	×	×	×	×	✓
Open-Ended QA	✓	✓	✓	✓	✓	✓	×	✓
Multi-Step Reasoning	✓	✓	✓	×	×	✓	×	×
Multicultural Music	✓	×	✓	×	×	×	×	×
Instruction Following	✓	×	×	✓	✓	×	✓	✓
In-the-wild Audios	✓	×	✓	×	×	×	×	✓
Voice Chat	✓	×	×	×	×	×	×	×
STEM Reasoning	✓	✓	✓	×	✓	×	×	×
Fully Human-Annotated	✓	✓	✓	×	×	✓	✓	✓

Table 3: Comparison of MMAU-Pro with existing audio understanding and reasoning benchmarks across various statistics.

Long-Audio Understanding Previous benchmarks such as MMAU (avg. 10.1 sec), MMAR (avg. 19.4 sec), and MMSU (7.01 sec) primarily focus on short audio clips, limiting evaluation to brief segments. Audio Flamingo 2 introduced long-audio perception with LALMs, motivated by real-world applications such as video, podcast, and movie analysis. This was followed by models like Qwen2.5-Omni and Audio Flamingo 3. MMAU-Pro builds on recent efforts such as LongAudioBench (Ghosh et al. 2025b) and BLAB (Ahia et al. 2025) by incorporating long-form audio inputs, categorized into four duration bins: short ($\leq$30s), medium (30s–3min), long (3–8min), and ultra-long (8–10min), comprising 2,589; 1,897; 1,307; and 348 QA instances respectively. Long-form comprehension poses unique challenges-such as locating sparse events ("needle in a haystack") and understanding narrative or temporal structure, which is explicitly tested in MMAU-Pro through specialized QA designs (more fine-grained stats and details in Appendix B.4)

Multi-Audio Understanding While multi-image understanding has been extensively studied (Jiang et al. 2024; Zhao et al. 2024; Li et al. 2025a), multi-audio understanding remains largely underexplored. Although many real-world use cases require understanding and reasoning over multiple audio inputs, most frontier MLLMs with audio perception capabilities do not natively support multi-audio processing. Chen et al. (2024c) make an initial attempt, and Audio Flamingo 3 supports multi-audio multi-turn dialogue, but lacks explicit multi-audio analysis support. MMAU-Pro addresses this gap by extending beyond single-audio QA. It includes 430 and 26 QA instances with two and three audios, respectively, each requiring understanding all individual audios for answering the QA correctly.

Multicultural Music Understanding

Most existing benchmarks evaluating music understanding focus predominantly on Western music, overlooking the rich diversity of global musical traditions. MMAU-Pro expands this scope by incorporating music from eight culturally distinct regions: African (21), Chinese (496), European (54), Indian (112), Latin American (11), Middle Eastern (7), Western (901), and Other Asian cultures (16). In Appendix B, we show that models trained primarily on Western music struggle with non-Western musical reasoning, highlighting the need to diversify training datasets for more inclusive music understanding.

Spatial Audio Understanding Understanding properties such as directionality, reverberation, and acoustic environment is a critical component of spatial awareness in auditory intelligence. Unlike visual spatial reasoning, spatial cues in audio often require multi-channel input. MMAU-Pro includes 325 expertly curated QA pairs paired with binaural recordings, designed to assess models' ability to perceive spatial relationships, such as sound direction and room characteristics, requiring fine-grained spatial awareness.

Voice QA As AI agents become more capable and widely adopted, voice-to-voice interaction is poised to become the default interface (Seaborn et al. 2021). However, enabling faithful voice-based interaction requires more than just spoken language understanding. It demands robust paralinguistic comprehension, including age, emotion, demographic cues, and urgency (Pias et al. 2024). Moreover, models must process spoken content that extends beyond natural language-such as mathematical expressions and STEM-related queries.

To evaluate these capabilities, MMAU-Pro introduces questions that assess paralinguistic understanding, including age, emotion, and urgency (see Appendix B.5). Additionally, we convert STEM questions into spoken form using GPT-4o TTS to test voice-based comprehension of mathematical expressions. These tasks also allow us to probe the model's STEM reasoning abilities, a known challenge for MLLMs. (see Section for analysis).

Instruction Following Enabling foundation models to follow human instructions is essential for building controllable and reliable AI assistants (Ouyang et al. 2022; Chen et al. 2024a; Zhou et al. 2023a). However, evaluating instruction-following remains challenging due to the open-ended nature of many prompts (e.g., "Write a short poem based on the sound you hear") (Chen et al. 2024a; Wang et al. 2024). To enable objective evaluation, we adopt the constraint-based approach of Zhou et al. (2023b), framing instruction-following as a verifiable subtask within MMAU-Pro. Our design is further inspired by IFEval-Audio (Gao et al. 2025), which introduced structured spoken instruction evaluation for audio-language models.

We construct a dedicated subset with 87 constraint in-

Models	Sound	Music	Speech	Sound Music	Speech Music	Speech Sound	Sound-Mus. Speech	Spatial	Voice	Multi Audio	Open ended	IF	Avg.
Random Choice	28.3	26.1	29.4	24.2	25.2	30.5	14.8	21.2	29.3	25.2	-	-	23.4
Human	78.2	70.5	82.3	79.3	78.5	82.4	85.7	88.2	78.4	79.8	77.3	100	77.9
Large Audio Language Models													
SALMONN 7B	32.2	44.9	38.3	22.0	34.8	28.4	28.6	26.5	36.5	11.4	31.2	33.9	34.5
SALMONN 13B	43.6	47.2	37.3	28.0	47.8	38.4	42.8	30.8	53.2	17.4	33.6	38.5	39.6
GAMA	45.4	41.2	29.8	24.0	27.9	27.3	14.8	12.0	28.4	20.2	24.2	31.7	33.2
DeSTA2	31.0	43.3	46.5	32.6	47.8	39.7	42.8	32.6	54.8	13.2	25.4	41.5	36.7
DeSTA2.5-Audio	35.7	48.2	49.9	22.0	36.9	35.2	28.6	28.0	51.0	19.8	36.4	46.5	40.6
BAT	28.9	22.7	25.9	30.0	23.9	25.0	14.8	23.7	24.5	20.2	24.6	31.8	24.8
Audio Flamingo 2	39.5	55.7	43.0	36.0	34.8	29.5	14.8	**44.1**	37.2	15.5	43.2	29.6	42.6
Phi4-MM	25.7	47.8	47.6	30.0	39.1	30.1	28.6	39.7	42.7	11.4	42.5	65.4	38.7
Kimi-Audio	46.0	57.6	52.2	**46.0**	**54.3**	48.9	42.8	43.7	50.6	17.2	34.5	42.3	46.6
Audio Flamingo 3	**55.9**	61.7	58.8	40.0	41.3	47.7	**57.1**	26.8	**58.6**	26.0	**44.2**	33.3	51.7
Gemma-3n-E2B-it	40.1	44.1	41.3	26.0	33.2	30.6	28.6	12.0	51.4	11.4	23.2	29.6	35.4
Gemma-3n-E4B-it	42.4	46.4	44.9	38.0	45.6	31.8	**57.1**	21.8	58.3	19.6	28.5	36.4	39.7
GPT4o-mini-Audio	40.2	59.7	66.1	35.3	42.2	55.9	42.8	12.0	52.7	22.4	41.6	79.7	48.3
GPT4o-Audio	44.7	**63.1**	**68.2**	40.4	43.5	**62.5**	**57.1**	21.4	57.5	**32.6**	43.2	**82.5**	**52.5**
Large Audio Reasoning Models													
R1-AQA	**47.9**	31.9	33.7	**32.0**	**36.9**	20.4	28.5	23.6	32.7	11.4	38.5	**44.2**	34.1
Audio-Reasoner	34.2	**50.1**	**44.0**	26.0	**36.9**	**43.2**	**28.6**	20.3	**43.4**	**22.6**	**38.6**	43.4	**39.5**
Mellow	27.6	32.9	27.9	24.0	34.8	27.3	14.3	**23.7**	28.3	20.8	21.4	23.5	27.5
Omni Models													
Ming-Lite-Omni-1.5	47.9	56.2	49.1	30.0	39.1	45.4	**42.8**	31.7	44.5	**37.4**	42.7	48.2	47.4
Baichuan-Omni-1.5	34.6	32.5	36.5	30.0	19.5	30.7	28.5	21.2	40.0	28.8	39.7	47.2	33.9
Qwen2.5-Omni-3B	38.5	60.3	53.9	40.0	45.6	46.6	**42.8**	28.9	46.5	11.4	47.6	58.4	46.1
Qwen2.5-Omni-7B	47.6	61.5	57.4	40.0	53.2	60.2	28.5	**41.2**	60.0	24.3	52.3	61.3	52.2
Gemini-2.0 Flash	48.4	56.9	69.5	39.6	57.6	55.9	**42.8**	34.6	68.6	26.5	66.8	94.2	55.7
Gemini-2.5 Flash	**51.9**	**64.9**	**73.4**	**42.8**	**58.7**	**61.3**	**42.8**	36.3	**71.7**	21.2	**67.5**	**95.1**	**59.2**
Cascaded Systems													
Caption + GPT4o	**38.6**	40.6	**38.4**	**21.6**	**38.2**	**25.5**	**28.6**	**9.5**	**38.6**	**24.7**	**27.6**	**88.2**	**35.3**
Captions + Qwen235B	36.4	**41.3**	36.1	18.6	37.4	24.5	14.3	5.8	35.6	22.5	25.6	85.5	33.7

Table 4: Accuracy of evaluated models on MMAU-Pro across single-modality tasks (Sound, Music, Speech), mixed-modality tasks (Sound–Music, Speech–Music, Speech–Sound, Sound–Music–Speech), and specialized tasks (Spatial, Voice-chat, Multi-Audio reasoning, Open-ended QA, Instruction-Following), along with overall weighted averages. **Bold** values highlight the highest value and underlined values highlight the second-highest value in each category for each type of model.

stances drawn from 28 instruction types, grouped into six categories (e.g., *Length Constraints*, *Keyword Usage*, *Format*). Each instruction is paired with one of seven open-ended prompt templates (e.g., "Describe the audio") and instantiated with variations to test robustness across prompt styles. Final inputs are synthesized using `ChatterboxTTS` (Resemble AI 2025), combining spoken instructions with audio segments from the `MMAU` dataset (e.g., speech, music, ambient sounds). We provide deterministic regex-based evaluation scripts for each constraint, enabling scalable, reproducible scoring in realistic conditions.

Experimental Setup

LALMs. We evaluate a wide range of Large Audio-Language Models on the MMAU-Pro benchmark to assess their capabilities in long and short form reasoning, spatial understanding, multicultural music interpretation, and multi-audio comparisons.

Cascaded Systems. To evaluate the robustness of our benchmark MMAU-Pro, we also conduct assessments on cascaded systems. In this approach, we first obtain captions for sound and music, and transcripts for speech-based questions. Subsequently, we combine these captions and questions and pass them to text-only open and closed-source Large Language Models (LLMs). These LLMs include GPT-4o (OpenAI et al. 2024), one closed-source, state-of-the-art LLM, and Qwen3-235B-A22B-Instruct (Yang et al. 2025), an open-source, instruction-tuned model. For obtaining the captions of sound and music audios, we resort to Audio Flamingo 3, and for obtaining the speech transcripts, we use Whisper-Large-v3 (Radford et al. 2023).

Evaluation Strategy. To evaluate MCQs, we compute the embedding of each answer choice using a pretrained transformer model, i.e, NV-Embed-v2 (Lee et al. 2024; Moreira et al. 2024) in our case, and compare it to the model's output embedding for the question context. Rather than computing the question embedding directly, the model generates an output vector representing its predicted response. This output embedding is compared against the embeddings of all available answer choices using cosine similarity. The choice

with the highest similarity is selected as the predicted answer. The evaluation is then conducted by comparing this prediction to the ground truth label. This embedding-based selection strategy allows for semantically meaningful predictions even when explicit answer tokens are not generated, and avoids reliance on string-based pattern matching. For evaluating open-ended responses, we employ Qwen2.5-7B-Instruct as a judge and provide it with the ground truth answer and the prediction. We evaluate each model's response on 5 fronts - (i) Correctness: How factually accurate is the response compared to the reference? (ii) Relevance: How well does the response address the specific question asked? (iii) Completeness: Does the response cover all important aspects mentioned in the reference? and (iv) Clarity: How clear and well-structured is the response? and we also ask the LLM to assign an overall assessment score, which we report in Table 4. For evaluating open-ended evaluations, we first obtain scores on a scale of 1 to 5. Then, we convert these scores into percentage values to ensure that all reported scores remain on the same scale. We also evaluate LLM as a judge vs Human annotation score, and find a high correlation value to validate the strength of our LLM-as-a-judge framework. We show these correlations on the MMAU-test-mini and MMAR dataset in Appendix D. For evaluation of multi-audio, for models that do not support multi-audio analysis, we concatenate the audios with a silence of 2 seconds and feed it to the model, and mention it in the prompt, whereas for the models that support multiple audios, we feed them sequentially.

Results and Discussion

Table 4 presents a comprehensive breakdown of model performance across twelve main modalities. Several clear patterns emerge. First, on the core single-modality tasks (Sound, Music, Speech), most end-to-end LALMs achieve only moderate accuracy (30–60%), with smaller models such as SALMONN-7B and Phi4-MM-Instruct often below 50%. Even the strongest open-source models Qwen2.5-Omni-7B rarely exceed 65% on Music or Speech, indicating that foundational audio understanding remains challenging.

Performance degrades further as tasks grow more complex. On mixed modalities (Sound–Music, Speech–Music, Speech–Sound, and three-way mixtures), accuracies typically fall into the 20–50% range. This suggests that models cannot yet reliably fuse information across multiple audio streams. A similar drop can be seen in the spatial audio understanding performance, where even the top models rarely surpass 40%.

Voice-chat reasoning, which tests conversational and world-knowledge, and STEM knowledge inference, also exposes weaknesses, with most models scoring between 25% and 60%. Notably, Qwen2.5-Omni-7B and Gemini-2.5 Flash perform decently well on these tasks, scoring 60% and 71.7% respectively, but smaller or less instruction-tuned models often languish below 50%.

Multi-audio reasoning and open-ended question answering remain the most challenging tasks: no model surpasses 30% accuracy on the "Multi-Audio" subset, and open-ended QA tops out at only 45% even for the largest models. For multiple-choice questions with four options, performance may be inflated because models can rely on elimination strategies or benefit from the higher probability of guessing correctly (25%). To further stress-test this, we expand certain questions to include 10 options and observe a substantial drop in accuracy as the number of options increases. Models like Mellow, Qwen2.5-Omni, and AF3, which support multi-audio, still do not exceed 30%. GPT4o-Audio achieves slightly higher than 30%. Finally, instruction-following ("IF") is dominated by large closed-source models.

In summary, while most closed-source LALMs can handle many single-domain input tasks well, they still struggle with nuances in audio like temporal understanding, prosodic, and emotional reasoning. In addition, they struggle much more with multi-audio analysis, spatial reasoning, free-form answering, and instruction following. These areas represent clear directions for future model and benchmark development.

Do MLLMs Retain Skills Acquired from Text Pre-training?

Performance on STEM QA. To examine whether models can effectively link audio understanding with text-based knowledge and reasoning skills, we compare the STEM-focused QA performance of AF3 (in "Think" mode) with its base LLM, Qwen2.5-7B-Instruct. In this setup, Qwen2.5-7B-Instruct is evaluated on the original text-only STEM questions from the source dataset, while AF3+Think is evaluated on the corresponding audio-based MMAU-Pro Voice STEM subset. Qwen2.5-7B-Instruct achieves 36.17% accuracy, whereas AF3+Think reaches only 31.91%. We identify two possible causes: (i) AF3 may lose part of its text-based math reasoning ability during audio fine-tuning—a gap that could potentially be mitigated with high-quality instruction tuning data; or (ii) an *auditory perception gap*, where the model correctly interprets the audio and retains the necessary reasoning skills, but fails to connect perception with knowledge. Similar issues have been observed in LVLMs (Ghosh et al. 2025a), where models demonstrate sufficient reasoning ability in text form but struggle to bridge perception with understanding, a phenomenon described as the *visual perception gap*.

Instruction Following. We compare the performance of AF3 and Qwen-2.5-Omni-7B on the instruction-following subset of MMAU-Pro. For the *Change Cases* task, AF3 attains 35.4% accuracy, whereas Qwen2.5-Omni-7B reaches 75.2%. On *Detectable Format*, AF3 fails to produce many correct responses (8.6%), while Qwen2.5-Omni-7B correctly formats 40.3% of cases. In *Length Constraints*, AF3 scores 30.7% compared to 68.5% for Qwen2.5-Omni-7B. AF3's only relative strength appears on *Detectable Content*, where it achieves 67.8% accuracy versus Qwen2.5-Omni-7B's 60.4%. *The Keywords* task again highlights the gap-AF3 manages just 5.9% while Qwen2.5-Omni-7B succeeds on 65.1%. Finally, for *Multi-Part Response*, AF3 records 55.6% accuracy compared with 60.8% for Qwen2.5-Omni-7B. This consistent advantage for Qwen2.5-Omni-7B on five of six subtasks underscores the crucial role of extensive text-

only pretraining and instruction-tuning: without a robust textual instruction corpus, AF3's performance on language-centric directives remains significantly weaker despite its strong audio understanding and reasoning ability.

Conclusion, Limitations and Future Work

In this paper, we introduced MMAU-Pro, a comprehensive benchmark designed to holistically evaluate general audio intelligence in multimodal language models. MMAU-Pro advances prior efforts by incorporating 5,305 expert-annotated QA pairs spanning 49 diverse skills across speech, sounds, music, and their combinations. The benchmark introduces several key innovations, including long-audio understanding, multi-audio reasoning, spatial audio comprehension, multicultural music understanding, instruction following, etc. These tasks mirror real-world challenges and require advanced perception, contextual understanding, and complex reasoning. Our evaluation across open and proprietary LALMs demonstrates that even the strongest models struggle across several categories. Of course, humans are able to do amazing feats with the sound they hear, and to truly benchmark an audio model's ability to do all of these will always be a work in progress, and we do not claim to explore all dimensions of audio processing/reasoning ability in the benchmark.

As part of future work, we plan to: (i) further expand the scale of MMAU-Pro to include more languages and low-resource acoustic environments; (ii) introduce dynamic and interactive audio tasks, such as real-time reasoning over streaming audio; (iii) refine instruction-following evaluation with free-form generation and adversarial constraints; and (iv) develop better metrics for evaluating paralinguistic understanding and culturally-grounded reasoning. We hope MMAU-Pro serves as a stepping stone toward developing more capable and general-purpose audio-language models.

References

Abouelenin, A.; Ashfaq, A.; Atkinson, A.; Awadalla, H.; Bach, N.; Bao, J.; Benhaim, A.; Cai, M.; Chaudhary, V.; Chen, C.; et al. 2025. Phi-4-mini technical report: Compact yet powerful multimodal language models via mixture-of-loras. *arXiv preprint arXiv:2503.01743*.

Ahia, O.; Bartelds, M.; Ahuja, K.; Gonen, H.; Hofmann, V.; Arora, S.; Li, S. S.; Puttagunta, V.; Adeyemi, M.; Buchireddy, C.; Walls, B.; Bennett, N.; Watanabe, S.; Smith, N. A.; Tsvetkov, Y.; and Kumar, S. 2025. BLAB: Brutally Long Audio Bench. arXiv:2505.03054.

Chen, L.; Zhu, C.; Chen, J.; Soselia, D.; Zhou, T.; Goldstein, T.; Huang, H.; Shoeybi, M.; and Catanzaro, B. 2024a. ODIN: Disentangled Reward Mitigates Hacking in RLHF. In *ICML*.

Chen, Y.; Yue, X.; Gao, X.; Zhang, C.; D'Haro, L. F.; Tan, R. T.; and Li, H. 2024b. Beyond single-audio: Advancing multi-audio processing in audio large language models. *arXiv preprint arXiv:2409.18680*.

Chen, Y.; Yue, X.; Gao, X.; Zhang, C.; D'Haro, L. F.; Tan, R. T.; and Li, H. 2024c. Beyond Single-Audio: Advancing Multi-Audio Processing in Audio Large Language Models. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Findings of the Association for Computational Linguistics: EMNLP 2024*, 10917–10930. Miami, Florida, USA: Association for Computational Linguistics.

Chu, Y.; Xu, J.; Yang, Q.; Wei, H.; Wei, X.; Guo, Z.; Leng, Y.; Lv, Y.; He, J.; Lin, J.; Zhou, C.; and Zhou, J. 2024. Qwen2-Audio Technical Report. arXiv:2407.10759.

Deshmukh, S.; Dixit, S.; Singh, R.; and Raj, B. 2025. Mellow: a small audio language model for reasoning. arXiv:2503.08540.

Deshmukh, S.; Elizalde, B.; Singh, R.; and Wang, H. 2023. Pengi: An Audio Language Model for Audio Tasks. In Oh, A.; Naumann, T.; Globerson, A.; Saenko, K.; Hardt, M.; and Levine, S., eds., *Advances in Neural Information Processing Systems*, volume 36, 18090–18108. Curran Associates, Inc.

Diao, X.; Zhang, C.; Kong, K.; Wu, W.; Ma, C.; Ouyang, Z.; Qing, P.; Vosoughi, S.; and Gui, J. 2025. SoundMind: RL-Incentivized Logic Reasoning for Audio-Language Models. arXiv:2506.12935.

Donley, J.; Tourbabin, V.; Lee, J.-S.; Broyles, M.; Jiang, H.; Shen, J.; Pantic, M.; Ithapu, V. K.; and Mehra, R. 2021. Easycom: An augmented reality dataset to support algorithms for easy communication in noisy environments. *arXiv preprint arXiv:2107.04174*.

Elizalde, B.; Deshmukh, S.; Ismail, M. A.; and Wang, H. 2023. CLAP Learning Audio Concepts from Natural Language Supervision. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Elizalde, B.; Deshmukh, S.; and Wang, H. 2024. Natural Language Supervision For General-Purpose Audio Representations. In *ICASSP 2024 - 2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 336–340.

Gao, Y.; Wang, B.; Wei, C.; Sun, S.; and Aw, A. 2025. IFEval-Audio: Benchmarking Instruction-Following Capability in Audio-based Large Language Models. arXiv:2505.16774.

Ghosh, S.; Evuru, C. K. R.; Kumar, S.; Tyagi, U.; Nieto, O.; Jin, Z.; and Manocha, D. 2025a. Visual Description Grounding Reduces Hallucinations and Boosts Reasoning in LVLMs. In *The Thirteenth International Conference on Learning Representations*.

Ghosh, S.; Kong, Z.; Kumar, S.; Sakshi, S.; Kim, J.; Ping, W.; Valle, R.; Manocha, D.; and Catanzaro, B. 2025b. Audio Flamingo 2: An Audio-Language Model with Long-Audio Understanding and Expert Reasoning Abilities. arXiv:2503.03983.

Ghosh, S.; Kumar, S.; Seth, A.; Evuru, C. K. R.; Tyagi, U.; Sakshi, S.; Nieto, O.; Duraiswami, R.; and Manocha, D. 2024. GAMA: A Large Audio-Language Model with Advanced Audio Understanding and Complex Reasoning Abilities. In Al-Onaizan, Y.; Bansal, M.; and Chen, Y.-N., eds., *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 6288–6313. Miami, Florida, USA: Association for Computational Linguistics.

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

Gong, B.; Zou, C.; Zheng, C.; Zhou, C.; Yan, C.; Jin, C.; et al. 2025. Ming-Omni: A Unified Multimodal Model for Perception and Generation. arXiv:2506.09344.

Gong, Y.; Luo, H.; Liu, A. H.; Karlinsky, L.; and Glass, J. R. 2024. Listen, Think, and Understand. In *The Twelfth International Conference on Learning Representations*.

Huang, C.-y.; Chen, W.-C.; Yang, S.-w.; Liu, A. T.; Li, C.-A.; Lin, Y.-X.; Tseng, W.-C.; Diwan, A.; Shih, Y.-J.; Shi, J.; et al. 2024. Dynamic-superb phase-2: A collaboratively expanding benchmark for measuring the capabilities of spoken language models with 180 tasks. *arXiv preprint arXiv:2411.05361*.

Jiang, D.; He, X.; Zeng, H.; Wei, C.; Ku, M.; Liu, Q.; and Chen, W. 2024. Mantis: Interleaved Multi-Image Instruction Tuning. *Transactions on Machine Learning Research*.

KimiTeam; Ding, D.; Ju, Z.; Leng, Y.; Liu, S.; Liu, T.; Shang, Z.; Shen, K.; Song, W.; Tan, X.; Tang, H.; Wang, Z.; Wei, C.; Xin, Y.; Xu, X.; Yu, J.; Zhang, Y.; Zhou, X.; Charles, Y.; Chen, J.; Chen, Y.; Du, Y.; He, W.; Hu, Z.; Lai, G.; Li, Q.; Liu, Y.; Sun, W.; Wang, J.; Wang, Y.; Wu, Y.; Wu, Y.; Yang, D.; Yang, H.; Yang, Y.; Yang, Z.; Yin, A.; Yuan, R.; Zhang, Y.; and Zhou, Z. 2025. Kimi-Audio Technical Report. arXiv:2504.18425.

Lee, C.; Roy, R.; Xu, M.; Raiman, J.; Shoeybi, M.; Catanzaro, B.; and Ping, W. 2024. NV-Embed: Improved Techniques for Training LLMs as Generalist Embedding Models. *arXiv preprint arXiv:2405.17428*.

Leng, S.; Xing, Y.; Cheng, Z.; Zhou, Y.; Zhang, H.; Li, X.; Zhao, D.; Lu, S.; Miao, C.; and Bing, L. 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. *arXiv preprint arXiv:2410.12787*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; and Li, C. 2025a. LLaVA-OneVision: Easy Visual Task Transfer. *Transactions on Machine Learning Research*.

Li, G.; Liu, J.; Dinkel, H.; Niu, Y.; Zhang, J.; and Luan, J. 2025b. Reinforcement Learning Outperforms Supervised Fine-Tuning: A Case Study on Audio Question Answering. arXiv:2503.11197.

Li, M.; Do, C.-T.; Keizer, S.; Farag, Y.; Stoyanchev, S.; and Doddipatla, R. 2024a. WHISMA: A Speech-LLM to Perform Zero-Shot Spoken Language Understanding. In *2024 IEEE Spoken Language Technology Workshop (SLT)*, 1115–1122.

Li, Y.; Liu, J.; Zhang, T.; Zhang, T.; Chen, S.; Li, T.; Li, Z.; Liu, L.; Ming, L.; Dong, G.; Pan, D.; Li, C.; Fang, Y.; Kuang, D.; Wang, M.; Zhu, C.; Zhang, Y.; Guo, H.; Zhang, F.; Wang, Y.; Ding, B.; Song, W.; Li, X.; Huo, Y.; Liang, Z.; Zhang, S.; Wu, X.; Zhao, S.; Xiong, L.; Wu, Y.; Ye, J.; Lu, W.; Li, B.; Zhang, Y.; Zhou, Y.; Chen, X.; Su, L.; Zhang, H.; Chen, F.; Dong, X.; Nie, N.; Wu, Z.; Xiao, B.; Li, T.; Dang, S.; Zhang, P.; Sun, Y.; Wu, J.; Yang, J.; Lin, X.; Ma, Z.; Wu, K.; li, J.; Yang, A.; Liu, H.; Zhang, J.; Chen, X.; Ai, G.; Zhang, W.; Chen, Y.; Huang, X.; Li, K.; Luo, W.; Duan, Y.; Zhu, L.; Xiao, R.; Su, Z.; Pu, J.; Wang, D.; Jia, X.; Zhang, T.; Ai, M.; Wang, M.; Qiao, Y.; Zhang, L.; Shen, Y.; Yang, F.; Zhen, M.; Zhou, Y.; Chen, M.; Li, F.; Zhu, C.; Lu, K.; Zhao, Y.; Liang, H.; Li, Y.; Qin, Y.; Sun, L.; Xu, J.; Sun, H.; Lin, M.; Zhou, Z.; and Chen, W. 2025c. Baichuan-Omni-1.5 Technical Report. arXiv:2501.15368.

Li, Y.; Sun, H.; Lin, M.; Li, T.; Dong, G.; Zhang, T.; Ding, B.; Song, W.; Cheng, Z.; Huo, Y.; Chen, S.; Li, X.; Pan, D.; Zhang, S.; Wu, X.; Liang, Z.; Liu, J.; Zhang, T.; Lu, K.; Zhao, Y.; Shen, Y.; Yang, F.; Yu, K.; Lin, T.; Xu, J.; Zhou, Z.; and Chen, W. 2024b. Baichuan-Omni Technical Report. arXiv:2410.08565.

Lu, K.-H.; Kuan, C.-Y.; and Lee, H.-y. 2025. Speech-ifeval: Evaluating instruction-following and quantifying catastrophic forgetting in speech-aware language models. *arXiv preprint arXiv:2505.19037*.

Ma, Z.; Ma, Y.; Zhu, Y.; Yang, C.; Chao, Y.-W.; Xu, R.; Chen, W.; Chen, Y.; Chen, Z.; Cong, J.; Li, K.; Li, K.; Li, S.; Li, X.; Li, X.; Lian, Z.; Liang, Y.; Liu, M.; Niu, Z.; Wang, T.; Wang, Y.; Wang, Y.; Wu, Y.; Yang, G.; Yu, J.; Yuan, R.; Zheng, Z.; Zhou, Z.; Zhu, H.; Xue, W.; Benetos, E.; Yu, K.; Chng, E.-S.; and Chen, X. 2025. MMAR: A Challenging Benchmark for Deep Reasoning in Speech, Audio, Music, and Their Mix. arXiv:2505.13032.

Moreira, G. d. S. P.; Osmulski, R.; Xu, M.; Ak, R.; Schifferer, B.; and Oldridge, E. 2024. NV-Retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*.

OpenAI; :; Hurst, A.; et al. 2024. GPT-4o System Card. arXiv:2410.21276.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C. L.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*, volume 35.

Peng, Y.; Tian, J.; Yan, B.; Berrebbi, D.; Chang, X.; Li, X.; Shi, J.; Arora, S.; Chen, W.; Sharma, R.; Zhang, W.; Sudo, Y.; Shakeel, M.; Jung, J.-W.; Maiti, S.; and Watanabe, S. 2023. Reproducing Whisper-Style Training Using An Open-Source Toolkit And Publicly Available Data. In *2023 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 1–8.

Pias, S. B. H.; Huang, R.; Williamson, D. S.; Kim, M.; and Kapadia, A. 2024. The Impact of Perceived Tone, Age, and Gender on Voice Assistant Persuasiveness in the Context of Product Recommendations. In *ACM Conversational User Interfaces 2024*, CUI '24, 1–15. ACM.

Radford, A.; Kim, J. W.; Xu, T.; Brockman, G.; McLeavey, C.; and Sutskever, I. 2023. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, 28492–28518. PMLR.

Resemble AI. 2025. Chatterbox-TTS. https://github.com/resemble-ai/chatterbox. GitHub repository.

Rubenstein, P. K.; Asawaroengchai, C.; Nguyen, D. D.; Bapna, A.; Borsos, Z.; Quitry, F. d. C.; Chen, P.; Badawy, D. E.; Han, W.; Kharitonov, E.; et al. 2023. Audiopalm: A large language model that can speak and listen. *arXiv preprint arXiv:2306.12925*.

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*.

Seaborn, K.; Miyake, N. P.; Pennefather, P.; and Otake-Matsuura, M. 2021. Voice in Human–Agent Interaction: A Survey. *ACM Comput. Surv.*, 54(4).

Wang, B.; Zou, X.; Lin, G.; Sun, S.; Liu, Z.; Zhang, W.; Liu, Z.; Aw, A.; and Chen, N. F. 2025a. AudioBench: A Universal Benchmark for Audio Large Language Models. In Chiruzzo, L.; Ritter, A.; and Wang, L., eds., *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, 4297–4316. Albuquerque, New Mexico: Association for Computational Linguistics. ISBN 979-8-89176-189-6.

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, P.; Li, L.; Chen, L.; Cai, Z.; Zhu, D.; Lin, B.; Cao, Y.; Kong, L.; Liu, Q.; Liu, T.; and Sui, Z. 2024. Large Language Models are not Fair Evaluators. In Ku, L.-W.; Martins, A.; and Srikumar, V., eds., *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 9440–9450. Bangkok, Thailand: Association for Computational Linguistics.

Weck, B.; Manco, I.; Benetos, E.; Quinton, E.; Fazekas, G.; and Bogdanov, D. 2024. MuChoMusic: Evaluating Music Understanding in Multimodal Audio-Language Models. In *Proceedings of the 25th International Society for Music Information Retrieval Conference*, 825–833. ISMIR.

Wu, Y.; Chen, K.; Zhang, T.; Hui, Y.; Berg-Kirkpatrick, T.; and Dubnov, S. 2023. Large-Scale Contrastive Language-Audio Pretraining with Feature Fusion and Keyword-to-Caption Augmentation. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1–5.

Xie, Z.; Lin, M.; Liu, Z.; Wu, P.; Yan, S.; and Miao, C. 2025a. Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models. arXiv:2503.02318.

Xie, Z.; Lin, M.; Liu, Z.; Wu, P.; Yan, S.; and Miao, C. 2025b. Audio-Reasoner: Improving Reasoning Capability in Large Audio Language Models. arXiv:2503.02318.

Xu, J.; Guo, Z.; He, J.; Hu, H.; He, T.; Bai, S.; Chen, K.; Wang, J.; Fan, Y.; Dang, K.; Zhang, B.; Wang, X.; Chu, Y.; and Lin, J. 2025. Qwen2.5-Omni Technical Report. arXiv:2503.20215.

Yang, A.; Li, A.; Yang, B.; Zhang, B.; Hui, B.; Zheng, B.; Yu, B.; Gao, C.; Huang, C.; Lv, C.; et al. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Zhao, H.; Cai, Z.; Si, S.; Ma, X.; An, K.; Chen, L.; Liu, Z.; Wang, S.; Han, W.; and Chang, B. 2024. MMICL: Empowering Vision-language Model with Multi-Modal In-Context Learning. In *The Twelfth International Conference on Learning Representations*.

Zheng, Z.; Peng, P.; Ma, Z.; Chen, X.; Choi, E.; and Harwath, D. 2024. BAT: Learning to Reason about Spatial Sounds with Large Language Models. In Salakhutdinov, R.; Kolter, Z.; Heller, K.; Weller, A.; Oliver, N.; Scarlett, J.; and Berkenkamp, F., eds., *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, 61454–61469. PMLR.

Zhou, C.; Liu, P.; Xu, P.; Iyer, S.; Sun, J.; Mao, Y.; Ma, X.; Efrat, A.; Yu, P.; Yu, L.; et al. 2023a. Lima: Less is more for alignment. *Advances in Neural Information Processing Systems*, 36: 55006–55021.

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023b. Instruction-Following Evaluation for Large Language Models. arXiv:2311.07911.