MMAU: A Massive Multi-Task Audio Understanding and Reasoning Benchmark

S Sakshi[♠*] Utkarsh Tyagi[♠*] Sonal Kumar[♠*] Ashish Seth[♠*] Ramaneswaran Selvakumar[♠*]
Oriol Nieto[♣] Ramani Duraiswami[♠†] Sreyan Ghosh[♠*†] Dinesh Manocha[♠†]

[♠]University of Maryland, College Park, USA [♣]Adobe, USA [*] Equal Contribution [†] Equal Advising

Evaluating Advanced Audio Understanding and Reasoning

- Current Large Audio-Language Models (LALMs) are mainly evaluated on fundamental tasks such as ASR and audio scene classification, which assess basic audio perceptions but lack complex reasoning that characterizes more sophisticated forms of intelligence.
- We present **MMAU**, the first benchmark for evaluating advanced audio perception and reasoning in LALMs. Featuring ***10K expert-annotated instances*** across ***speech, sounds, and music, MMAU tests 27 distinct skills***, requiring advanced audio understanding and domain knowledge.

MMAU Vs. Prior Audio Benchmarks

Benchmark	Size	Domain			Tasks		Expert Comments	Difficulty Level
		Speech	Sound	Music	Info Extraction	Reasoning		
CompA	600	×	✓	×	0 ×	0.6k ✓	Requires only sound event sequence understanding. Limited in reasoning depth and knowledge scope.	2.0
CompA-R	1.5k	×	✓	×	0 ×	1.5k ✓	Restricted to sounds and only contextual event understanding. Limited in knowledge scope.	3.0
MuChin	1k	×	×	×	0 ×	0 ×	Restricted to music with minimal reasoning depth. Limited in knowledge scope.	2.5
MusicBench	0.4k	×	×	✓	0 ×	0 ×	Restricted to music with minimal reasoning depth. Limited in knowledge scope.	2.5
MuChoMusic	1.2k	×	×	✓	0.7k ✓	0.4k ✓	Restricted to music with open-ended answers. Limited in knowledge scope.	3.5
OpenASQA	8.8k	✓	✓	×	8.8k ✓	0 ×	Requires limited acoustic scene understanding. Does not require external or expert knowledge.	3.0
AudioBench	100k+	✓	✓	✓	5k ✓	0 ×	Basic acoustic information retrieval with minimal reasoning depth and complexity. Does not require external or expert knowledge.	3.5
AIR-Bench	19k	✓	✓	✓	1.2k ✓	0.8k ✓	Basic acoustic information retrieval with minimal reasoning depth and complexity. Does not require external or expert knowledge.	2.5
MMAU *(ours)*	10K	✓	✓	✓	4.5k ✓	5.2k ✓	Requires fine-grained audio understanding with expert-level, multi-step reasoning and specialized knowledge across a broad range of topics.	4.5

Table 1. Comparison of MMAU with existing audio understanding and reasoning benchmarks across various statistics. MMAU covers all three domains—speech, sound, and music—while having the highest number of information extraction and complex reasoning tasks.

Skill Distribution in MMAU

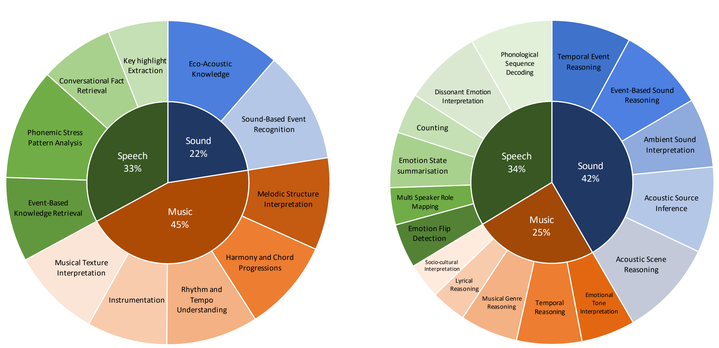

Figure 1. **(Left)** Distribution of skills required for *information extraction questions* and **(Right)** for *reasoning questions* in the MMAU benchmark across the multiple domains. Each question in MMAU demands the model to apply one or more of these skills.

Examples of Question-Answer pairs in MMAU

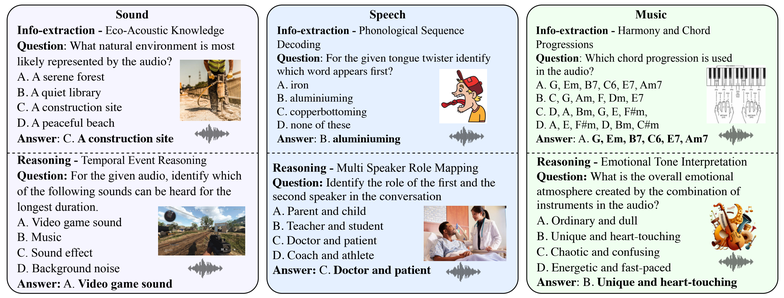

Figure 2. Examples from the MMAU benchmark illustrating the diverse range of reasoning and information extraction tasks across the domains of sound, speech, and music.

- Each task in MMAU involves rich, context-specific audio paired with human-annotated QA pairs that require expert-level knowledge and reasoning abilities.
- The benchmark covers a wide range of challenges, illustrating the breadth and depth of MMAU's evaluation scope.

MMAU Benchmark Construction Pipeline

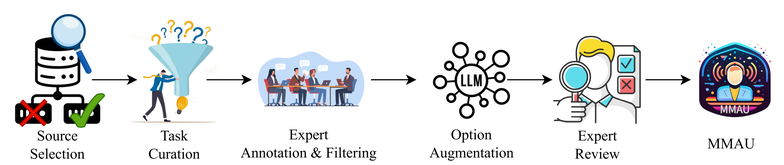

Figure 3. MMAU Benchmark Construction Pipeline.

Performance comparison of various LALM and LLM on MMAU

Models	Size	{So, Mu, Sp}			Sound		Music		Speech		Avg	
					Test-mini	Test	Test-mini	Test	Test-mini	Test	Test-mini	Test
Random Guess	-		-		26.72	25.73	24.55	26.53	26.72	25.50	26.00	25.92
Most Frequent Choice	-		-		27.02	25.73	20.35	23.73	29.12	30.33	25.50	26.50
Human (test-mini)	-		-		86.31	-	78.22	-	82.17	-	82.23	-
Large Audio Language Models (LALMs)												
Pengi	323M	✓	✓	×	06.10	08.00	02.90	03.05	01.20	01.50	03.40	04.18
Audio Flamingo Chat	2.2B	✓	✓	×	23.42	28.26	15.26	18.20	11.41	10.16	16.69	18.87
LTU	7B	✓	✓	×	22.52	25.86	09.69	12.83	17.71	16.37	16.89	18.51
LTU AS	7B	✓	✓	✓	23.35	24.96	9.10	10.46	20.60	21.30	17.68	18.90
MusiLingo	7B	×	✓	×	23.12	27.76	03.96	06.00	05.88	06.42	10.98	13.39
MuLLaMa	7B	×	✓	×	40.84	44.80	32.63	30.63	22.22	16.56	31.90	30.66
M2UGen	7B	×	✓	×	03.60	03.69	32.93	30.40	06.36	04.53	14.28	12.87
GAMA	7B	✓	✓	×	41.44	45.40	32.33	30.83	18.91	19.21	30.90	31.81
GAMA-IT	7B	✓	✓	×	43.24	43.23	28.44	28.00	18.91	15.84	30.20	29.02
Qwen-Audio-Chat	8.4B	✓	×	×	55.25	56.73	44.00	40.90	30.03	27.95	43.10	41.86
Qwen2-Audio	8.4B	✓	✓	✓	07.50	08.20	05.14	06.16	03.10	04.24	05.24	06.20
Qwen2-Audio-Instruct	8.4B	✓	✓	✓	54.95	45.90	50.98	53.26	42.04	45.90	49.20	52.50
SALMONN	13B	✓	✓	✓	41.00	40.30	34.80	33.76	25.50	24.24	33.70	32.77
Audio Flamingo 2		✓	✓	×	**61.56**	**65.12**	**73.95**	**72.94**	38.91	30.93	55.50	58.99
Gemini Pro v1.5	-		-		56.75	54.46	49.40	48.56	**58.55**	55.90	54.90	52.97
Gemini 2.0 Flash	-		-		56.46	61.73	58.68	56.53	51.65	**61.53**	**55.60**	**59.93**
GPT4o-Audio	-		-		34.23	57.33	21.26	41.03	42.64	57.83	32.70	52.07
Large Language Models (LLMs)												
GPT4o + weak cap.	-		-		39.33	35.80	39.52	41.9	58.25	68.27	45.70	48.65
GPT4o + strong cap.	-		-		**57.35**	**55.83**	49.70	**51.73**	**64.86**	**68.66**	**57.30**	**58.74**
Llama-3-Ins. + weak cap.	8B		-		34.23	33.73	38.02	42.36	54.05	61.54	42.10	45.87
Llama-3-Ins. + strong cap.	8B		-		50.75	49.10	**50.29**	48.93	55.25	62.70	52.10	53.57

Table 2. Performance comparison of various LALMs and LLMs on the MMAU. The best performing models in each category are highlighted in **bold**, and the second-best scores are underlined.

Deep Dive into Skill-Specific LALMs Performance

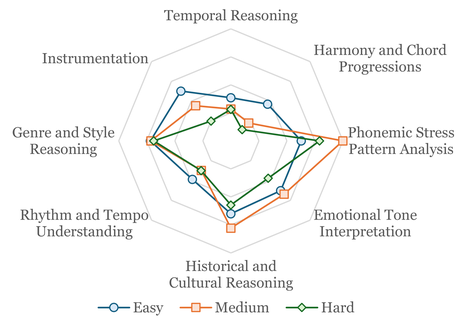

Figure 4. Accuracy for Gemini 2.0 Flash across easy, medium, and hard questions, categorized by skills. LALMs excel in some skills across all difficulty levels (e.g., Phonemic Stress Pattern Analysis) but struggle with others (e.g., Temporal Reasoning) regardless of difficulty.

Are LALMs Really Listening?

To assess LALMs' attention to audio, we replace the original audio in MMAU's test set with Gaussian noise and compare performance (Fig. 5). MuLLaMa and SALMONN show little change, indicating limited audio reliance, while others drop significantly, suggesting greater dependence on audio inputs.

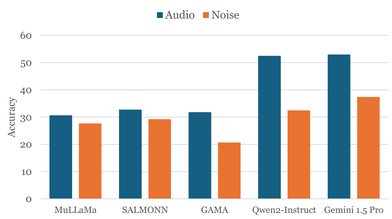

Figure 5. Performance on replacing audio with Gaussian noise.

Where are the LALMs Falling Short?

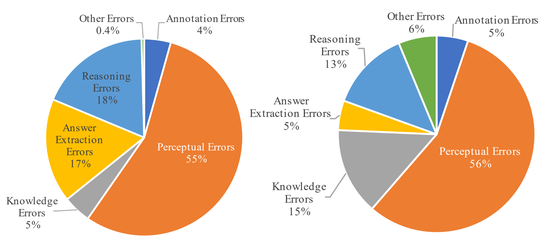

Figure 6. Distribution of error types across 500 instances for Qwen2-Audio-Instruct **(Left)** and Gemini 2.0 Flash **(Right)**. The dominant error type is **Perceptual Errors**.

Future Work

- **MMAU-pro is almost here!** Larger skill-set, multi-hop reasoning, multiple audios, variable length (and long up to 10 minutes!), speech + sounds, and much more!
- **Improving certain aspects from community feedback:** 1) Reduce language priors in the question, 2) have more options and specifically a "none of the above option", 3) Include open-ended QAs beyond just MCQ and 4) Work on improving the evaluation algorithm.

Figure 7. MMAU Project Page.

Figure 8. Reasoning training data.