MAEB: Massive Audio Embedding Benchmark

Adnan El Assadi, Isaac Chung, Chenghao Xiao, Roman Solomatin, Animesh Jha, Rahul Chand, Silky Singh, Kaitlyn Wang, Ali Sartaz Khan, Marc Moussa Nasser, Sufen Fong, Pengfei He, Alan Xiao, Ayush Sunil Munot, Aditya Shrivastava, Artem Gazizov, Niklas Muennighoff, Kenneth Enevoldsen

MAEB is the largest audio embedding benchmark to date covering 7 task types, 30 tasks, and 100+ languages.

NEURAL INFORMATION PROCESSING SYSTEMS

Benchmark Overview

Task type	Example tasks	Metric	# Tasks
Classification	CREMA-D, MInDS-14, SIB-FLEURS	Accuracy	11
Retrieval	Clotho, FLEURS, JamAlt artist	Hit@5	9
Pair Classification	VoxPopuli accent, NMSQA	AP	3
Clustering	Vehicle sounds, CREMA-D	V-measure	3
Zero-shot Classification	RAVDESS, SpeechCommands	Accuracy	2
Multi-label Classification	FSD2019Kaggle	LRAP	1
Reranking	GTZAN genre	MAP	1
MAEB			**30**
MAEB+ (extended)			**98**

MAEB uses 30 of the 98 MAEB+ tasks with the **same model ranking** (ρ = 0.98).

A100 GPU hours	MAEB	MAEB+	Speedup
YAMNet	2.0	6.0	3.0×
larger_clap_general	11.5	32.2	2.8×
CLAP-htsat-fused	13.0	35.4	2.7×
wav2vec2-xls-r-2b	26.9	45.6	1.7×

Acoustic vs. Linguistic

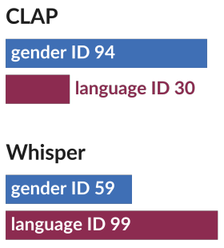
Top 5 acoustic	linguistic rank
1 larger_clap_general	#29
2 larger_clap_music_speech	#32
3 Qwen2-Audio-7B	#12
4 clap-htsat-unfused	#39
5 msclap-2023	#38

Top 5 linguistic	acoustic rank
1 BidirLM-Omni-2.5B	#18
2 jina-v5-omni-small	#15
3 LCO-Omni-7B	#12
3 jina-v5-omni-nano	#19
3 LCO-Omni-3B	#20

Rankings on **acoustic** and **linguistic** tasks are unrelated: **ρ = 0.04.**

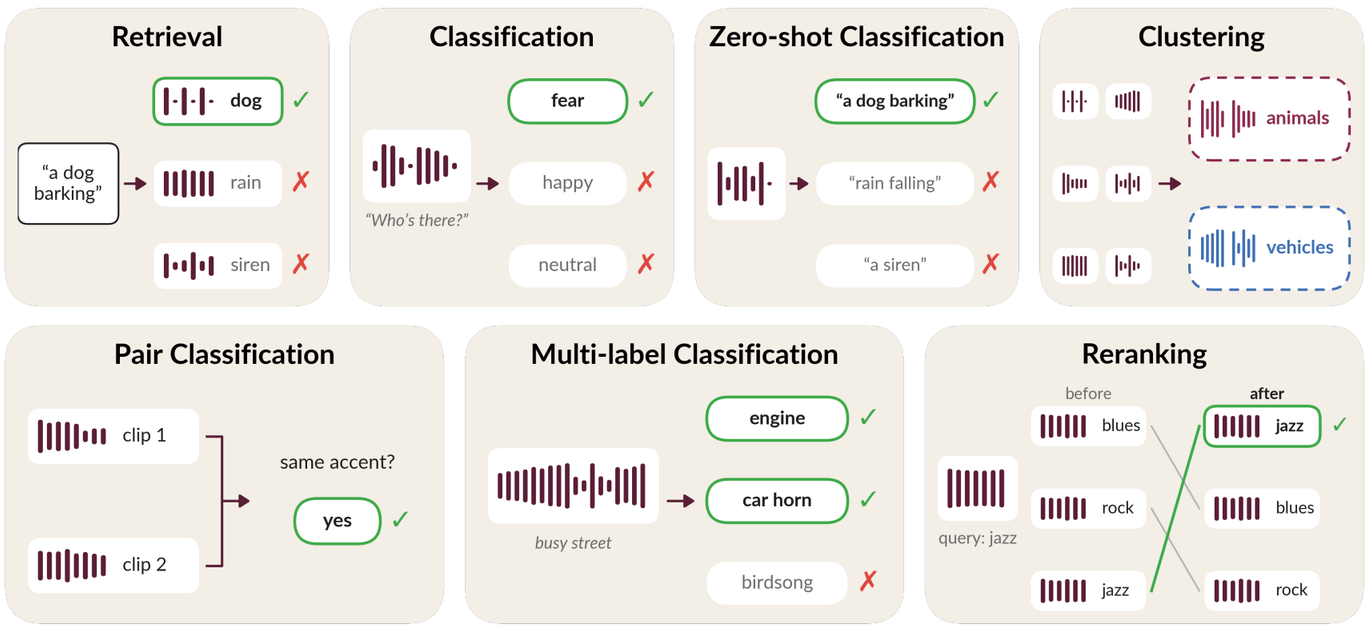
No model is top 5 on both.

Same in VoxPopuli:



Massive Audio Embedding Benchmark

100+ languages · 30 tasks



Main Result

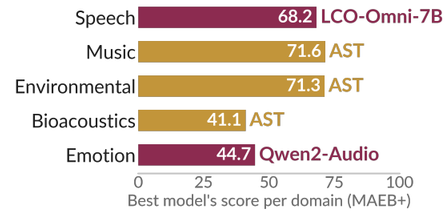
	Borda rank		Score by task type (higher is better)								
Model	Rank	95% CI	Avg	Classif.	Multi-label	Pair	Rerank	Cluster	Audio→ Audio	Text→ Audio	Zero-shot
LCO-Embedding-Omni-7B	1	1–2	53.5	56.2	75.7	67.3	78.7	1.7	75.0	52.8	64.5
jina-emb-v5-omni-small	2	2–4	50.4	52.0	69.3	65.4	84.4	6.1	83.9	45.5	62.2
Qwen2-Audio-7B	3	1–6	34.5	58.3	32.6	56.9	80.8	12.7	37.6	1.3	12.4
BidirLM-Omni-2.5B	4	2–5	52.4	59.8	60.9	66.7	74.8	6.2	66.9	45.0	65.8
larger_clap_general	5	3–7	35.0	47.9	80.6	51.9	66.8	6.6	95.6	9.3	14.9
larger_clap_music_speech	6	4–9	34.7	47.4	80.6	52.1	65.6	7.7	96.9	8.8	13.2
LCO-Embedding-Omni-3B	8	5–8	52.3	54.5	74.1	66.7	75.4	1.3	67.1	53.0	62.2
whisper-medium	9	9–12	48.8	54.8	56.8	53.9	67.6	5.0	72.1	–	–
clap-htsat-fused	12	10–13	33.5	41.2	81.1	52.0	61.3	22.7	87.7	8.8	13.2
MuQ-MuLan-large	16	15–19	28.2	38.8	33.1	51.9	85.4	4.3	96.3	1.2	12.6
AST (AudioSet)	19	18–20	46.1	45.7	72.7	51.2	77.6	6.9	94.6	–	–

Highlighted: best per column. 11 of 58 models; – = no text encoder.

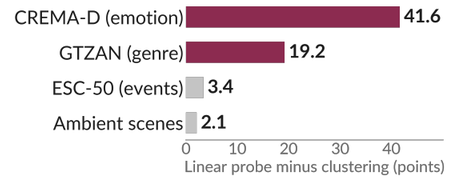
Audio LLM · CLAP · Speech · Encoder

- **Embedding-trained audio LLMs lead** (LCO-Omni-7B first).
- Generative audio LLMs fail at text→audio retrieval (Qwen2-Audio: 1.3).
- CLAP leads audio→audio retrieval but is weak on speech.
- **No model is best everywhere:** 9 column bests, 6 models.
- **Clustering is unsolved:** the best of 58 models scores 22.7.

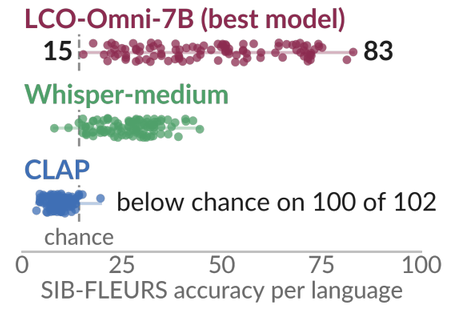
Analysis



- No model wins every domain: AST leads 3, audio LLMs lead 2.



- Linear probes find emotion and genre; clustering does not (gaps of 42 and 19).



- Best model: 15 (Burmese) to 83 (Indonesian) across 102 languages.

MTEB · `$ pip install mteb`

arXiv Paper · Code · Leaderboard