CMI-Bench (ISMIR 2025): open audio-LLMs perform poorly on standard MIR tasks. Is it still true?

Zero-shot audio-LLMs are **now usable for key detection**, but still behind ASR on lyrics.

Key detection results

Lyrics results

Tomáš Sourada **Jan Hajič jr.**
sourada@ufal.mff.cuni.cz Institute of Formal and Applied Linguistics, Charles University, Prague

What CMI-Bench found

	Key detection	Lyrics transcription	
Audio LLM	**GS-K GES ↑**	**DS WER ↓**	**DS CER ↓**
Qwen2-Audio	8.28	793.0	818.6
Qwen-Audio	6.51	115.7	96.2
SALMONN-Audio	7.70	816.1	760.00
MusiLingo	**9.50**	2019	2311
LTU	7.61	235.5	210.8
LTU-AS	1.42	191.7	185.5
MU-LLaMA	7.56	191.9	168.3
Audio-Flamingo	8.21	275.7	262.6
GAMA	7.69	225.4	201.3
GAMA-IT	7.70	152.6	165.2
Pengi	0.00	343.2	368.0
SOTA (non-LLM)	*74.3*	*12.99*	–

From Table 3 of CMI-Bench (Ma et al., ISMIR 2025). GS-K: GiantSteps Key, weighted score ×100. DS: DSing, in %.

Key detection: best 9.50 (MusiLingo) vs. supervised 74.3
Lyrics: WER > 100 % for all models, "none usable"

What we did

No training, no fine-tuning: zero-shot prompting of open-weight audio-LLMs, text answer parsed.

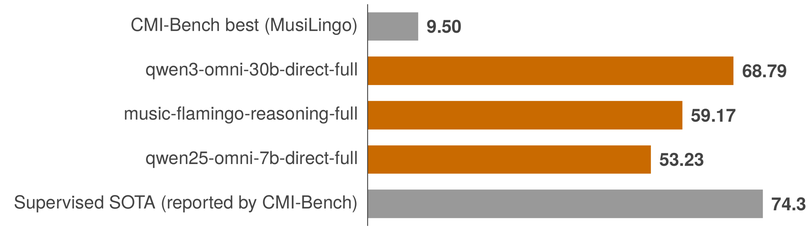
Key detection: 10 audio-LLMs

Qwen2.5-Omni 3B, 7B · Qwen3-Omni-30B-A3B Instruct, Thinking · Gemma 4 E2B, E4B, 12B · Audio Flamingo 3 · Music Flamingo, Music Flamingo think

Lyrics: 5 audio-LLMs, 2 ASR models

Qwen2.5-Omni-3B, -7B · Gemma 4 E2B, E4B, 12B · ASR: Whisper large-v2, Qwen3-ASR-1.7B

Key detection: official MIREX results



GiantSteps Key (604 excerpts), weighted score (asym.) ×100. Orange: our three submitted systems. Grey: CMI-Bench Table 3.

Same test set, same metric as CMI-Bench.

Prompt and parser adopted from CMI-Bench *(may differ slightly)*

How

- Answer parsed into 1 of 24 major/minor keys
- 3 prompting setups: direct (CMI-Bench prompt), reasoning, two-turn
- Input: middle or first 30 s, vote over 30 s clips, or the full song
- **New audio-LLMs hear the whole 2-min excerpt** *(CMI-Bench: 30 s clips)*

Development findings *(GiantSteps MTG Key)*

- Whole song better than any 30 s clip (Qwen Omni, Music Flamingo)
- Thinking hurt
- No prompting setup best for all models

We also tried zero-shot audio-LLMs on lyrics transcription

System	WER ↓	CER ↓
Whisper	33.3	21.8
Qwen-ASR	**28.9**	**16.9**
Whisper+Qwen-ASR+Qwen3-8b	29.7	18.1
QwenOmni2.5-7b-nohint	42.5	28.9
Gemma-e4b-nohint	52.5	31.3

Development set multiling500: 500 DALI songs (EN, FR, DE, ES), 31.6 h; in %. Audio-LLMs without ASR hints, on clips of at most 30 s.

- Audio-LLMs: WER **safely below 100 %**, but behind much smaller ASR models
- CMI-Bench used DSing: English solo singing, simpler than our multilingual polyphonic data. *Not directly comparable.*
- Submitted: Qwen3-8B merges Whisper + Qwen3-ASR transcripts clip by clip, then inserts section breaks.
- The winner (Frangiadaki et al., MIREX 2026): similar idea, 3 transcripts (Whisper, Parakeet, Music Flamingo), a much larger text-only LLM.

Open question

Why so different from CMI-Bench?

- Different test data? *(lyrics only)*
- Different prompting?
- Newer models really better?
- … or our test sets in their training data? *Unpublished: we cannot tell.*

Co-funded by the European Union

MINISTRY OF EDUCATION, YOUTH AND SPORTS

Supported by project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union, and partially by the SVV project number 260 821. The work described herein has been using services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).