

UFAL-PMCG SYSTEMS AT THE MIREX 2026 AUDIO KEY DETECTION CHALLENGE: AUDIO-LLMS ARE NO LONGER OFF-KEY

Tomáš Sourada Jan Hajič jr.
Institute of Formal and Applied Linguistics
Charles University
Prague, Czech Republic
{sourada,hajicj}@ufal.mff.cuni.cz

ABSTRACT

Our three systems for the MIREX 2026 Audio Key Detection task each ask a pretrained audio large language model (audio LLM) for the key of a recording and parse its text answer into one of the 24 major and minor keys; nothing is trained or fine-tuned. We compared eleven open-weight checkpoints, three prompting setups and four ways of selecting the input audio clip on 1157 tracks of GiantSteps MTG Key. The best system, Qwen3-Omni 30B with a direct prompt, reaches a MIREX weighted score of 0.680, against 0.104 for a random guess and 0.180 for the best constant answer. Giving the model the whole song instead of a 30-second clip helped the Qwen Omni models and Music Flamingo, thinking variants scored far below their non-thinking counterparts, and no prompting setup was best for all models. We submit the best configuration of Qwen3-Omni 30B, Music Flamingo and Qwen2.5-Omni 7B.

1. INTRODUCTION

This extended abstract describes the systems submitted by the Prague Music Computing Group (PMCG) at the Institute of Formal and Applied Linguistics (ÚFAL), Charles University, to the Audio Key Detection task of MIREX 2026.¹

Our motivation comes from CMI-Bench [1], which reported very poor key detection for open audio LLMs: the best of its eleven models, MusiLingo, reached a weighted score of 0.095 on GiantSteps Key, against 0.743 for a supervised system. To see whether this still holds for recent models, we trained nothing: we ask pretrained audio LLMs for the key in a zero-shot fashion and compare eleven checkpoints under different prompts and audio inputs.

On our development set, the best system reaches a weighted score of 0.680, or 0.649 in the scoring variant

¹https://music-ir.org/mirex/wiki/2026:Audio_Key_Detection



© T. Sourada and J. Hajič jr.. Licensed under a Creative Commons Attribution 4.0 International License (CC BY 4.0). **Attribution:** T. Sourada and J. Hajič jr., “UFAL-PMCG Systems at the MIREX 2026 Audio Key Detection Challenge: Audio-LLMs Are No Longer Off-Key”, extended abstract in *MIREX 2026 Audio Key Detection Task*, 2026.

of CMI-Bench. The numbers are not directly comparable: our development set is not GiantSteps Key, and CMI-Bench scores 30-second clips while our best systems hear the whole song. We cannot say how much of the gap comes from newer models and how much from the evaluation setup, but recent audio LLMs are usable for key detection.

2. TASK AND EVALUATION

The task is to identify the key of a recording, one of the 12 major and 12 minor keys. A system is called once per recording: it gets a single-channel 16-bit WAV file at 44.1 kHz and writes one line with the tonic and the mode separated by a tab. The test set is GiantSteps Key [2], 604 two-minute excerpts of electronic dance music; we checked that its annotations spell no tonic with a sharp. It must not be used for training.

A prediction scores 1 for the correct key, 0.5 for a key a perfect fifth away (same mode), 0.3 for the relative key, 0.2 for the parallel key, and 0 otherwise; the weighted score of a system is the mean over all recordings. The task description does not say whether the fifth counts in both directions. *mir_eval* [3] gives the 0.5 by default only to a prediction a fifth above the reference key. According to the documentation of its development version,² MIREX has counted the fifth in both directions since 2017; the latest release (0.8.2) does not implement this yet. We therefore assume both directions, implement the descending fifth ourselves, and use this *MIREX score* as our main development metric. We also report the *CMI score*, the variant with ascending fifths only that CMI-Bench uses. An answer in which no key is found scores 0 and is counted.

3. DATA

We do not train or fine-tune anything, so we have no training data to report. The audio LLMs we use were pretrained on large amounts of data, but for none of them is the list of training data published.

For development we use GiantSteps MTG Key³ which the task description suggests for training. The other suggested dataset, Isophonics, ships only annotations and no

²https://github.com/mir-evaluation/mir_eval/blob/main/mir_eval/key.py#L135-L137

³<https://github.com/GiantSteps/giantsteps-mtg-key-dataset>

audio. Of the 1486 annotated two-minute clips we keep those with annotator confidence 2 and exactly one major or minor key: 1157 tracks (782 minor, 375 major; the most frequent key is C minor with 112). We downloaded the MP3 previews from the JKU backup mirror (the Beatport URL is dead), verified their MD5 checksums, converted them with ffmpeg to the MIREX input format, and respelled the gold keys with flats (D# minor becomes Eb minor) as in GiantSteps Key. Our selection shares no track id and no audio MD5 checksum with GiantSteps Key; a different mix or a re-recording of a test song would not be detected this way. We drew three random subsets: *small* (25 songs) for preliminary experiments, *large* (250) for decisions about hyperparameters, and *full* (1157) for a few selected systems.

4. SYSTEMS

All systems share one model-agnostic pipeline; the text answer of the model is parsed into one of the 24 keys.

Pipeline. The audio is decoded with PyAV to 16 kHz mono float32 and sent in a single user message, with the text before the audio. We add no system message; the chat templates of the Flamingo models and of Qwen2.5-Omni add their own default one. The models run with bfloat16 weights and greedy decoding in Hugging Face transformers 5.17.0 and torch 2.11.0 with CUDA 12.8. Of the Qwen Omni models we load only the text-generating “thinker”, not the speech-generating “talker”. In the MIREX call, an unparsable answer is replaced by a fixed fallback key, C minor, the most frequent key of the development data.

Prompting setups. *Direct* uses the prompt of CMI-Bench [1] word for word (“Return only the name of the predicted key”) and an exact port of its parser. *Reasoning* asks for a free answer whose last line is `Answer: $KEY`. *Two-turn* asks for a free answer and then, in a second turn, for `Answer: $KEY`. The last two are parsed with a port of the MusICA-MetaBench parser [4]: the last key mentioned, sharps mapped to flats, upper-case tonic required. The token limits are 32 (direct), 1024 (reasoning) and 1024 + 32 (two-turn), with 2048 more when the model thinks. Gemma 4 12B got 512 tokens for the direct answer, because 11 of its 25 direct answers on small were cut off at 32. If a model outputs the key in the sharpened variant, the postprocessing respells it to flat.

Audio clip. Qwen2-Audio and the Gemma 4 models accept at most 30 s of audio; the Qwen Omni and the Flamingo models also accept whole songs (two-minute clips in our data). We compare the *middle* 30 s, the *first* 30 s, a plurality *vote* over consecutive 30 s clips (unparsable clips do not vote, a tie goes to the clip nearest the centre of the song), and the *full* song.

Audio LLMs. We ran eleven open-weight checkpoints: Qwen2-Audio-7B-Instruct; Qwen2.5-Omni 3B and 7B; Qwen3-Omni-30B-A3B Instruct and Thinking; Gemma 4 E2B, E4B and 12B; Audio Flamingo 3; and Music Flamingo, basic and think. Thinking is a switch of the chat template for Gemma 4 and Qwen3-Omni Thinking; Music Flamingo think is a separate checkpoint that always

thinks.

Baselines. Two baselines need no model: a random guess (0.104, computed analytically), and the fallback key (0.180 on full).

5. DEVELOPMENT EXPERIMENTS

On small with the middle clip, all Gemma 4 models scored below or very close to the random guess in all three prompting setups, so we excluded them from further experiments. The following observations are from large.

For Qwen2-Audio, direct is worse than reasoning, which is worse than two-turn; its best combination is two-turn with the first clip, and it often spells the tonic with a sharp. For both Qwen2.5-Omni models the full song makes the difference. With it, reasoning is slightly better than the other setups for the 3B model and direct is the best setup for the 7B model; their best systems are almost on par (0.5692 for 7B direct, 0.5604 for 3B reasoning).

Audio Flamingo 3 stays below 0.34 in every configuration, with small differences between the setups; its best one is direct with the first clip. For Music Flamingo the full song again makes the difference: it reaches 0.60 with reasoning and 0.58 with two-turn, slightly more than the Qwen2.5-Omni models. It still returns sharps, against the instruction. Its think checkpoint stays below 0.3, so thinking hurts rather than helps.

Qwen3-Omni Instruct is the best of all models in the direct and the two-turn setup, while its reasoning setup is clearly worse; the Thinking checkpoint scores below 0.22. The eleven best systems on large all hear the full song.

6. RESULTS AND DISCUSSION

Table 1 reports the selected systems on full. Qwen3-Omni 30B is the clear winner: with the direct setup it scores 0.6799, far above chance. Music Flamingo follows, then Qwen2.5-Omni 7B and 3B, and Qwen2-Audio comes last.

For the Qwen Omni models and Music Flamingo, giving the model the whole song is much better for key detection than any setup with 30-second clips. For the prompting setup there is no such preference: the winning model, Qwen3-Omni 30B, fails in the reasoning setup, which is the best one for Music Flamingo.

Ensemble analysis. We also built pseudo-systems from the development answers of several systems. An oracle that picks, for each song, the best answer among the best systems of the top three models reaches 0.844 on full, suggesting that the models make errors on different songs. A plurality vote of the best systems of the top three models reaches 0.691, against 0.680 for the best single system, which is not much help. This vote is in-sample: its members were selected and ordered on the scored set. We did not implement it as a runnable system and did not submit it.

Model	Setup	Clip	Score	CMI	Acc.	Fifth	Rel.	Par.
Random guess (exp. values)			0.1042	0.0833	0.0417	96.4	48.2	48.2
Constant key C minor			0.1804	0.1523	0.0968	168	27	23
Oracle (top3)			0.8438	0.8192	0.7753	122	31	45
Vote (top3)			0.6913	0.6653	0.5990	123	77	111
Qwen3-Omni-30B-A3B-Instruct	direct	full	0.6799	0.6487	0.5799	136	78	121
Qwen3-Omni-30B-A3B-Instruct	two-turn	full	0.6445	0.6104	0.5411	152	73	109
music-flamingo-2601-hf	reasoning	full	0.5911	0.5751	0.4892	101	86	208
music-flamingo-2601-hf	two-turn	full	0.5870	0.5719	0.4823	96	78	249
Qwen2.5-Omni-7B	direct	full	0.5678	0.5346	0.4615	171	69	84
music-flamingo-2601-hf	direct	full	0.5591	0.5444	0.4503	95	94	251
Qwen2.5-Omni-3B	two-turn	full	0.5393	0.5134	0.4443	149	55	95
Qwen2.5-Omni-3B	direct	full	0.5391	0.5114	0.4443	152	49	95
Qwen2.5-Omni-3B	reasoning	full	0.5366	0.5115	0.4443	143	59	88
Qwen2.5-Omni-7B	reasoning	full	0.5275	0.5003	0.4270	152	67	101
Qwen2.5-Omni-7B	two-turn	full	0.5005	0.4716	0.4036	152	69	77
Qwen2-Audio-7B-Instruct	two-turn	first	0.4903	0.4462	0.3665	196	83	102
Qwen2-Audio-7B-Instruct	two-turn	vote	0.4901	0.4382	0.3544	223	75	115
Qwen2-Audio-7B-Instruct	two-turn	middle	0.4541	0.4018	0.3181	230	74	101
Qwen3-Omni-30B-A3B-Instruct	reasoning	full	0.3882	0.3570	0.2844	171	52	95

Table 1. Results on full (1157 tracks), sorted by score. Score: MIREX score; CMI: CMI score; Acc.: share of recordings with exactly the reference key; Fifth, Rel., Par.: number of recordings whose predicted key is a perfect fifth away, the relative or the parallel key. The most frequent key, the best-scoring constant key and the fallback key are all C minor here. Thinking is off in all runs. We submit the best row of Qwen3-Omni-30B-A3B-Instruct, music-flamingo-2601-hf and Qwen2.5-Omni-7B. Oracle and Vote are pseudo-systems, both combining the answers of the best-performing systems of the top three models (that is, Qwen3-Omni-30B with direct prompt, Music Flamingo with reasoning prompt, and Qwen2.5-Omni-7B with direct prompt).

System	Params	Inf.	Mem.
qwen3-omni-30b-direct-full	35.3 B	1.07	64.6
music-flamingo-reasoning-full	8.3 B	1.11	16.5
qwen25-omni-7b-direct-full	10.7 B	1.01	18.7

Table 2. Resources on full. Params: whole checkpoint (the loaded thinkers of the Qwen Omni models have 31.7 B and 8.9 B). Inf.: seconds per recording, model loading included. Mem.: peak GPU memory in GB, on two NVIDIA A100 40 GB (summed), an A40 and an A30. Nothing was trained.

7. SUBMITTED SYSTEMS

We submit three systems, one per audio LLM: the best configuration on full of each of the three best models. They are submitted with the prefix `UFAL-PMCG-`, which we omit here. All three give the whole recording to the model, write the tonic with flats only, and write C minor when no key is found in the answer. Table 2 lists their resources. **qwen3-omni-30b-direct-full**: Qwen3-Omni-30B-A3B-Instruct is asked to return only the name of the key (at most 32 new tokens); the answer is parsed with the port of the CMI-Bench parser. It scores 0.680 on full. **music-flamingo-reasoning-full**: Music Flamingo (basic) is asked to reason about the key and to end with `Answer: $KEY` (at most 1024 new tokens); the last key named is taken. It scores 0.591. **qwen25-omni-7b-direct-full**: Qwen2.5-Omni-7B with the direct setup, as above. It scores 0.568.

8. LIMITATIONS AND CONCLUSION

The general-purpose pretrained audio LLMs we use may have had the evaluation data in their training data, which would be a data leak. Their training data are not published, so we cannot confirm or rule this out.

Without any training, Qwen3-Omni 30B with a direct prompt scored 0.680 on 1157 tracks of GiantSteps MTG Key, followed by Music Flamingo (0.591) and Qwen2.5-Omni 7B (0.568). For these models the whole song was a better input than any 30-second clip, thinking hurt, and the best prompting setup depended on the model. A like-for-like comparison with CMI-Bench needs the scores on GiantSteps Key itself.

9. ACKNOWLEDGMENTS

This work was supported by the 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).

10. AI USAGE STATEMENT

We used Claude Code (Anthropic) to write the code of our experiments and to help draft, edit, and shorten this paper.

The authors checked all content, results, and references.

11. REFERENCES

- [1] Y. Ma, S. Li, J. Yu, E. Benetos, and A. Maezawa, “CMI-Bench: A comprehensive benchmark for evaluating music instruction following,” in *Proc. of the 26th Int. Society for Music Information Retrieval Conf.*, 2025, pp. 416–425.
- [2] P. Knees, Á. Faraldo, P. Herrera, R. Vogl, S. Böck, F. Hörschläger, and M. L. Goff, “Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections,” in *Proc. of the 16th Int. Society for Music Information Retrieval Conf.*, 2015, pp. 364–370.
- [3] C. Raffel, B. McFee, E. J. Humphrey, J. Salamon, O. Nieto, D. Liang, and D. P. W. Ellis, “mir_eval: A transparent implementation of common MIR metrics,” in *Proc. of the 15th Int. Society for Music Information Retrieval Conf.*, 2014, pp. 367–372.
- [4] T. Sourada, K. Vendrame, and J. Hajič jr., “Music I Care About: Automated Multimodal Benchmarking of LLM Music Perception Skills on (Almost) Any Music,” 2026, arXiv:2607.06015.