MIREX 2026 Overall Result Poster

ISMIR 2026 Late-Breaking/Demo · Abu Dhabi

Future MIREX Team

All 2026 results on the MIREX wiki

Lyrics Transcription <TC: Alexandre D'Hooge>

Dataset: Jam-ALT (multilingual)

Team	CER	WER	WER′	F (P)	F (B)	F (L)	F (S)
Frangiadaki et al.[1]	**0.0881**	**0.1470**	**0.1913**	**0.5134**	**0.4932**	0.8026	**0.7587**
UFAL-PMCG Whisper	0.1991	0.2860	0.3317	0.4295	NA	0.7442	0.0000
UFAL-PMCG Qwen-ASR	0.1598	0.2593	0.3190	0.2584	NA	0.5438	NA
UFAL-PMCG Whisper+Qwen	0.1714	0.2622	0.3162	0.3210	NA	0.6764	0.2661
UFAL-PMCG Gemma	0.1785	0.2806	0.3301	0.1447	0.0943	**0.8118**	0.2431

Note: CER, WER and WER′: lower is better. F: formatting F-scores for punctuation (P), parentheses (B), line breaks (L) and section breaks (S). [1] Spanish songs transcribed by the authors (out of memory on a V100-32GB). MUSDB-ALT and JamendoLyrics++ results are on the wiki.

Audio Instrument Recognition <TC: Wenye Ma>

Dataset: Hidden set (316 MedleyDB excerpts)

System	Macro-F1	Micro-F1	mAP	Labels / clip
Baseline: OpenMIC-RF	**0.497**	**0.607**	0.632	3.16
SSR	0.488	0.451	**0.688**	11.18

Note: Ranked by macro-F1 over the 12 classes with at least 20 positive excerpts (3.43 true labels per clip). On the public OpenMIC-2018 test set, SSR leads all three metrics (macro-F1 0.801 vs 0.707).

Standardized Evaluation Suites <TC: Junyan Jiang, Akira Maezawa>

Audio Chord Estimation

Group	Billboard 2013 (161)			JSIR (267)			Yamaha Balanced (228)			Yamaha JPop (249)		
	Root	MajMin	7ths	Root	MajMin	7ths	Root	MajMin	7ths	Root	MajMin	7ths
ChordyV-ACE	54.88	37.58	25.81	25.12	13.28	5.89	62.19	41.06	28.98	58.38	38.33	26.88
MuQ-BTC	78.02	74.65	62.43	54.03	47.92	35.43	83.83	81.65	66.61	**83.12**	80.95	64.27
SLASH	**82.02**	★ **80.37** prev. 79.15	62.92	47.63	41.81	25.98	83.10	81.54	63.66	81.17	78.97	59.25
Baseline: Chordino	71.06	67.18	48.88	38.39	32.88	17.09	77.57	74.64	56.38	74.48	71.98	52.39
Baseline: ChordFormer	81.11	80.01	**69.02**	**55.89**	**52.10**	**39.61**	**84.11**	**83.27**	**70.43**	**83.12**	**81.94**	**67.57**

Note: +Bass and segmentation scores and the additional test sets (Billboard 2012, RWC Popular, JAAH) are on the wiki.

Audio Beat Tracking

Group	GTZAN (999)		SMC (217)		osu2beat2025 (703)		Yamaha Bal. (239)		Yamaha JPop (250)		JSIR (260)	
	F1	CMLt	F1	CMLt	F1	CMLt	F1	CMLt	F1	CMLt	F1	CMLt
CASM (default)	88.75	**79.68**	**57.95**	**43.30**	**87.91**	**69.31**	90.33	**85.19**	**94.43**	**88.46**	92.48	78.08
infgon	42.02	24.62	25.23	11.07	39.21	6.85	36.55	17.23	38.54	18.24	37.75	5.54
Baseline: CD1	81.19	66.94	33.66	13.12	71.75	42.99	76.62	59.60	77.61	58.90	65.44	27.86
Baseline: BeatThis	**89.02**	79.64	71.81*	69.89*	87.53	67.77	**90.59**	84.52	94.00	86.64	**94.70**	**83.50**

Note: CASM: the default of its 12 variants. The other variants and metrics are on the wiki.

Audio Downbeat Estimation

Group	GTZAN (993)	osu2beat2025 (703)	Yamaha Balanced (239)	Yamaha JPop (250)	JSIR (260)
CASM (default)	77.24	81.99	85.57	**91.78**	85.52
infgon	13.95	11.15	11.40	12.36	10.60
Baseline: CD4	46.36	44.01	60.45	63.10	27.82
Baseline: BeatThis	**78.12**	**83.22**	**86.09**	91.46	**87.27**

Note: F-measure, ±70 ms. Precision, recall and CASM's other variants are on the wiki.

Audio Key Detection

Group	GiantSteps (604)		GTZAN (837)		FMAKv2 (5,489)		JSIR (226)		Yamaha Bal. (185)		Yamaha JPop (174)	
	Score	Exact	Score	Exact	Score	Exact	Score	Exact	Score	Exact	Score	Exact
Chord-LLM	74.40	68.05	**73.76**	**67.03**	77.60*	71.51*	76.90	68.58	91.46	88.11	**91.49**	**87.36**
ChordyV-Key	39.74	28.81	48.64	36.44	53.40	41.12	53.81	39.82	76.16	64.86	63.62	52.87
Hurliman_K10	74.57	67.88	72.03	64.64	72.43	64.40	34.78	21.68	89.30	85.95	85.80	79.89
MusicFM-MLP	★ **77.62** prev. 75.81	**72.19**	73.42	66.91	**75.63**	**68.46**	**79.82**	**71.68**	**91.68**	**89.19**	87.99	83.33
RESON	42.30	29.30	48.48	34.29	56.66	43.58	31.81	18.14	62.70	47.03	55.69	37.93
UFAL-PMCG music-flamingo-reasoning	60.17	52.15	54.15	44.68	69.52	61.14	70.04	59.29	90.59	87.03	85.86	79.89
UFAL-PMCG qwen25-omni-7b-direct	56.54	45.86	55.94	45.04	66.00	56.33	62.74	51.33	79.51	72.97	77.53	70.69
UFAL-PMCG qwen3-omni-30b-direct	70.94	62.91	67.60	58.90	74.86	67.41	63.54	53.54	90.81	88.11	87.47	82.18
Baseline: CN1	50.53	39.74	47.08	36.68	63.02	52.94	53.05	42.92	83.03	77.84	75.63	67.82

Note: Score: 1 correct, 0.5 a fifth above or below, 0.3 relative, 0.2 parallel. Exact: % correct keys. The asymmetric score and error categories are on the wiki.

Music Structure Analysis

Dataset: Harmonix Set (test split, 200 songs)

Group	ACC	HR.5F	HR3F
SCALE	0.799	0.728	0.828
songmamba	0.776	**0.739**	★ **0.840** prev. 0.762
Baseline: SongFormer	**0.802**	0.704	0.798
Baseline: MusicFM (2025, reported)	0.705	0.644	0.710

Lyrics-to-Audio Alignment

Dataset: Jamendo (20 songs)

Group	AAE (s) ↓	MedAE (s) ↓	PCS ↑	PCO ↑
Repetition-Aware Phrase Priors	0.241	**0.040**	0.635	0.944
Baseline: NUS (AutoLyrixAlign)	**0.216**	0.046	**0.751**	**0.946**

Symbolic Music Generation <TC: Ziyu Wang, Xinyue Li, Jingwei Zhao>

Dataset: Subjective evaluation (12 listeners, 38 ratings per model)

Team	Coherency	Structure	Creativity	Musicality
DAMN	**3.68 ± 0.16[a]**	**3.39 ± 0.18[a]**	3.18 ± 0.16[a]	**3.47 ± 0.14[a]**
PhraseSplitGen	2.95 ± 0.18[b]	2.74 ± 0.15[b]	2.89 ± 0.15[a]	2.61 ± 0.14[b]
MuseCoco (BL-1)	2.79 ± 0.17[b]	2.76 ± 0.15[b]	3.42 ± 0.12[a]	2.68 ± 0.16[b]
Anticipatory Music Transformer (BL-2)	**3.47 ± 0.20[a]**	**3.34 ± 0.16[a]**	3.34 ± 0.16[a]	**3.61 ± 0.17[a]**

Note: Mean ± SEM. Different letters within a column mark significant differences (Wilcoxon signed-rank test, Holm–Bonferroni, $p < 0.05$).

Audio-to-Score Transcription <TC: Alexandre D'Hooge>

Dataset: QuartetsMIREX 2026 (373 excerpts of 8 unseen string quartets)

Submission	WER ↓	CER ↓	LER ↓	MV2H ↑
Staves-Informed track				
Lu et al.	0.3177	0.3481	0.7456	0.8179
Blind track				
Cummins et al.	**0.3202**	**0.3150**	0.7679	0.5981
Hidalgo-Centeno et al.	0.3385	0.3258	**0.7460**	**0.6053**

Note: Quartets by Schubert, Brahms and Mendelssohn, composers absent from the Quartets training data. On the original Quartets test set, the blind systems reach a WER of 0.029–0.035.

Rhythm Game Chart Generation <TC: Ziyun Liu, Carolina Carusi>

Dataset: Expert ratings, 5 songs (3 of 5 mappers per song)

Model	Musical repr.		Creativity		Gameplay		Combined		Human-likeness	
	Within	Across	Within	Across	Within	Across	Within	Across	Within	Across
DOSU	−1.8850	−1.8592	−1.5620	−1.3766	−0.4957	−0.6081	−1.9416	−1.4305	−1.6278	−1.5213
FumenLab	**1.4122**	**1.4946**	**1.4761**	**1.5760**	0.6217	**0.8282**	**1.2576**	**1.5429**	0.8437	0.8177
HiTaiko	1.0530	0.7580	−0.2570	−0.4161	**1.1337**	0.7907	1.1165	0.3365	**2.4409**	**1.9699**
MALer	−0.5802	−0.3934	0.3429	0.2167	−1.2597	−1.0107	−0.4325	−0.4489	−1.6567	−1.2663

Note: Sum of the 3 judges' normalized scores (z-score per judge, within one song or across songs), averaged over the 5 songs.

Dataset: Player ratings, Himitsu feat. KAFU (11 players)

Mapper	Musical repr.	Creativity	Gameplay	Overall	Human-likeness
DOSU	4.73	4.91	4.45	4.70	3.09
FumenLab	5.91	4.55	5.18	5.21	**5.27**
HiTaiko	4.55	3.82	4.00	4.12	3.27
MALer	**6.00**	**5.45**	**5.27**	**5.58**	4.18
Amasugi (human mapper)	8.36	7.36	8.64	8.12	8.82

Note: Mean of 1–10 ratings; players were not told that one chart was human-made. Bold: best AI chart; shaded: human reference. Per-song and per-difficulty results are on the wiki.

Music Evaluation via CMI-RewardBench <TC: Yinghao Ma, Yonghyun Kim, Haiwen Xia, Junwon Lee>

Dataset: Agreement with human preference (%)

System	Core (249)	Vocal ext.	Instrumental ext.
CloudEar-overall	**73.9**	60.9	72.6
SongEval (Average)	72.7	62.0	72.0
first_submission	72.3	**63.3**	**75.4**
CloudEar (Musicality)	69.9	60.6	70.7
CloudEar (Alignment)	69.5	61.6	69.5
SongBench (Musicality)	69.1	62.0	68.8
SongBench (Average)	69.1	59.6	72.6
Baseline: CMI-RM	65.9	57.2	69.2
aimusic	65.5	55.9	70.7
TuneJury	64.3	53.9	66.0
Random guess (always B)	59.0	55.2	56.1
MuseCritic	55.4	59.9	46.7
SYSTASYS Resource-Lite V2	39.8	43.4	40.2
SYSTASYS, flipped	60.2	56.6	59.8

Note: Pairwise preference accuracy, ranked by the core set.

Music Performance Difficulty Prediction <TC: Huan Zhang, Pedro Ramoneda>

Dataset: Held-out test set (761 pieces)

Submission	Kendall's τ-c ↑	Time / song
CrescendAI	**0.6654**	24.96 s
KaraOrchee	0.5864	**3.70 s**
VG1	0.5582	23.32 s

MIREX 2026 Overview

ISMIR 2026 Late-Breaking/Demo · Abu Dhabi

Future MIREX Team

All 2026 results on the MIREX wiki

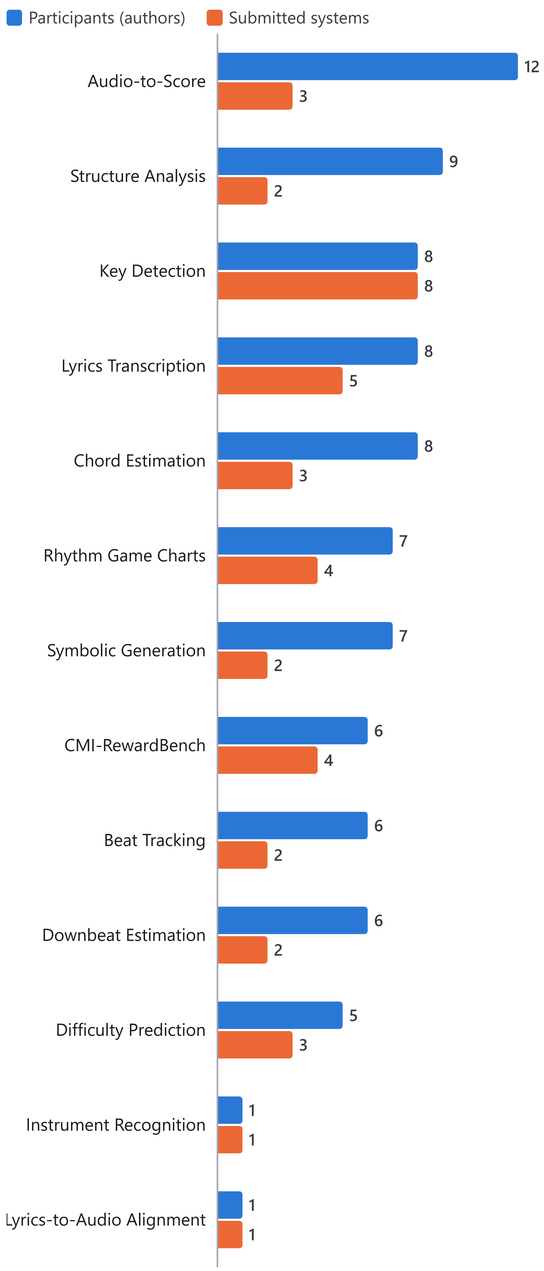
MIREX 2026 at a glance

13 tasks

40 submitted systems

63 authors

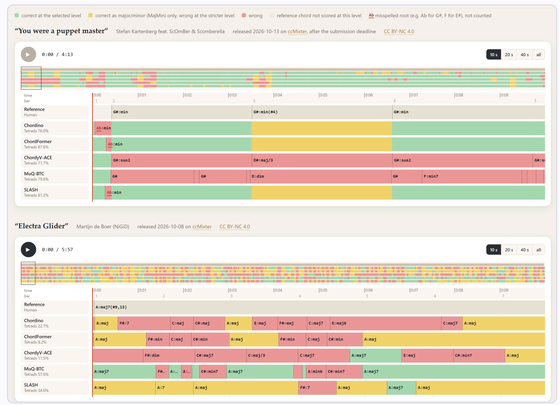
Participants and systems per task



Baselines are not counted. Task descriptions on this page are quoted from the 2026 task pages of the MIREX wiki.

Audio Chord Estimation

This task requires participants to extract or transcribe a sequence of chords from an audio music recording. This includes the full characterisation of chords – root, quality, and bass note – as well as their chronological order, including specific onset times and durations.



Interactive demo
Two Creative Commons songs released after the submission deadline, with every system's chords.
futuremirex.com/demos/audio-chord-estimation-2026/

Audio Key Detection

Audio Key Detection aims to identify the musical key (e.g., C major, A minor) of an audio recording. This involves determining both the tonic (root pitch) and the mode (major or minor) from the audio signal.

Audio Beat Tracking

The aim of the automatic beat tracking task is to track each beat location in a collection of sound files. The algorithms are evaluated in terms of their accuracy in predicting beat locations annotated by a group of listeners.

Audio Downbeat Estimation

The aim of the automatic downbeat estimation task is to identify the locations of downbeats in a collection of sound files. While this is similar to the Audio Beat Tracking task, here the aim is to find the first beat of each bar (measure) rather than all beat times.

Music Audio Generation

The MIREX 2026 **Cover Song Generation** task asks an AI music system to remake a full song according to a style or arrangement prompt. The output can change the style, instruments, arrangement, timbre, or tempo, but it should still be recognizable as a version of the source song.

Music Structure Analysis

The aim of the MIREX Music Structure Analysis task is to identify and label key structural sections in musical audio. For MIREX 2026, participants are required to segment musical audio and classify each segment into one of seven functional categories: 'intro', 'verse', 'chorus', 'bridge', 'inst' (instrumental), 'outro', or 'other'.

Lyrics-to-Audio Alignment

The task of automatic lyrics-to-audio alignment has as an end goal the synchronization between an audio recording of singing and its corresponding written lyrics. The algorithm receives mixed singing audio (singing voice + musical accompaniment) and its lyrics at word level, and outputs the onset and offset timestamps (seconds) of each word.

```
Lyrics:  no more carefree ...
Output:  0.123  0.798  no
         0.798  1.123  more
         1.345  2.176  carefree
```

Lyrics Transcription

The task of Lyrics Transcription aims to identify the words from sung utterances, in the same way as in automatic speech recognition. For this year's edition, the input is always a polyphonic mix (singing voice + musical accompaniment).

Audio Instrument Recognition

The task is clip-level multi-label instrument recognition. Given a music audio excerpt, a submitted system should predict which instruments are present in the excerpt.

Audio-to-Score Transcription

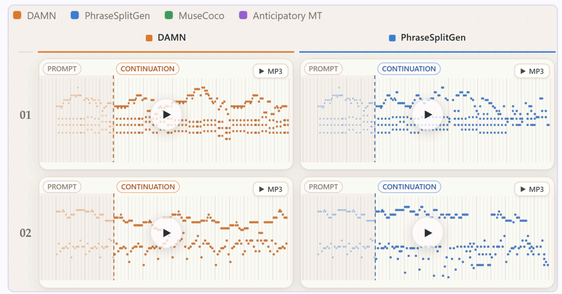
Participating systems receive an audio recording of a polyphonic music piece and must output a digital symbolic score that can be read by a musician or notation program: a kern score encoding pitches, durations, accidentals, staves, meter, clef, key signature, and bar structure. Two tracks: **Staves-Informed A2S** (staves metadata is provided) and **Blind A2S** (only the audio recording is provided).

Music Performance Difficulty Prediction

Participants submit systems that take a solo piano audio recording as input and output a predicted difficulty score. This shared task establishes the first community-wide benchmark for piano difficulty estimation from audio.

Symbolic Music Generation

For this year's challenge, the selected task is **Piano Music Continuation**. Given a 4-measure piano prompt (plus an optional pickup measure), the goal is to generate a 12-measure continuation that is musically coherent with the prompt, forming a complete 16-measure piece.



Interactive demo
8 prompts × 4 systems, playable in the browser (excerpt shown).
futuremirex.com/demos/symbolic-music-generation-2026/

Rhythm Game Chart Generation

A rhythm game chart consists of a sequence of timed note events synchronized with an audio track, defining what actions players should perform and when. This year's task is generating **osu!taiko** beatmaps: osu!taiko is a game mode in osu! inspired by Taiko no Tatsujin, which features a horizontal conveyor belt of drum notes moving towards a judgment circle.



Hit circles | Drumroll | Spinner

Images: MIREX wiki task page.

Music Evaluation via CMI-RewardBench

Participants are invited to submit evaluation scripts or models that output preference scores given a prompt and a pair of generated music tracks. The benchmark evaluates the systems on their ability to mirror human judgment.

```
Prompt:  "ambient serene instrumental"
Input:   Candidate A, Candidate B (generated tracks)
Output:  {"battle_uuid": "toy_example_001",
          "preferred_candidate": "A"}
```