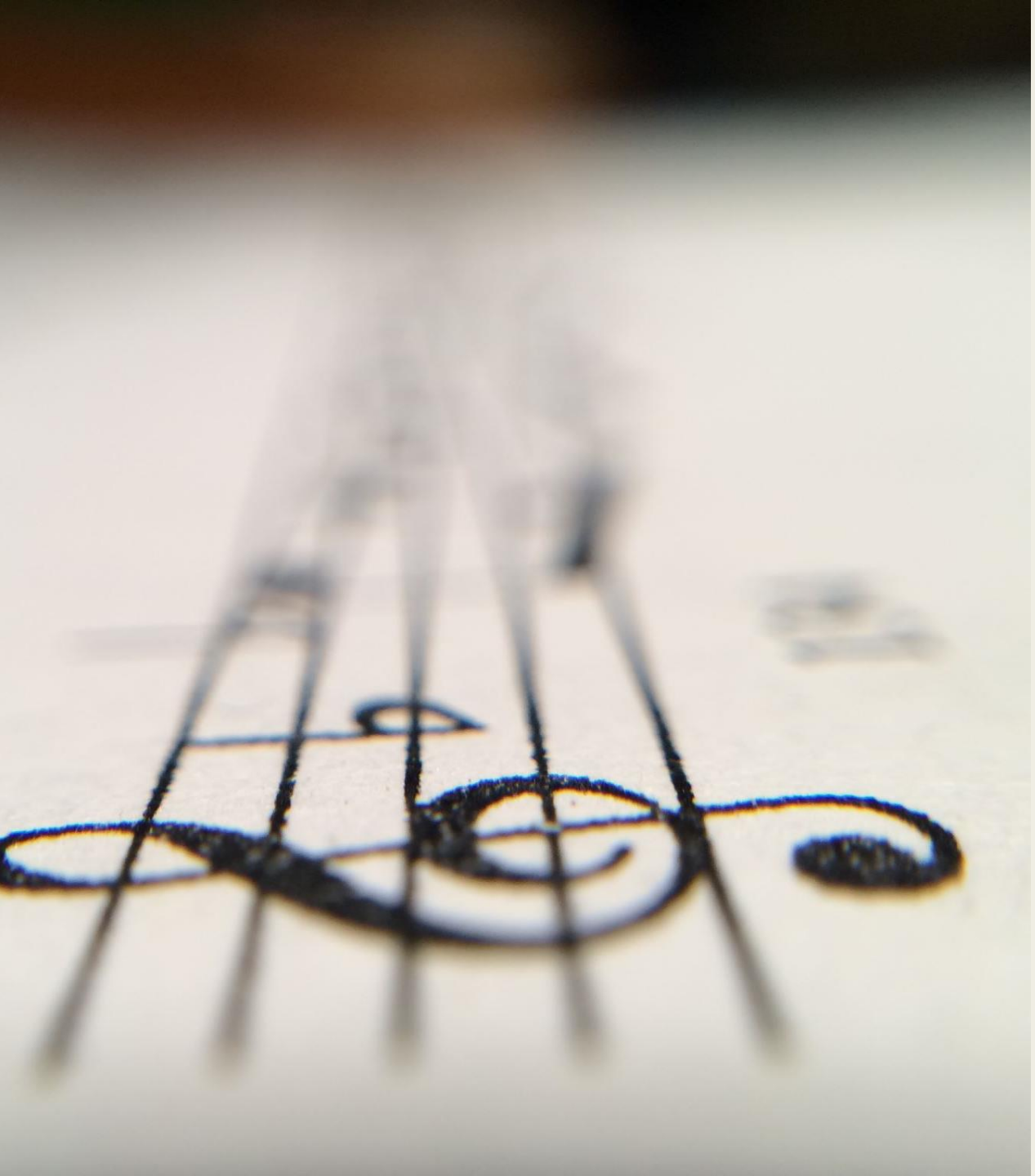


ChatMusician: Understanding and Generating Music Intrinsically with LLM

Ruibin Yuan, Hanfeng Lin, Yi Wang, Zeyue Tian, Shangda Wu, Tianhao Shen, Ge Zhang, Yuhang Wu, Cong Liu, Ziya Zhou, Ziyang Ma, Liumeng Xue, Ziyu Wang, Qin Liu, Tianyu Zheng, Yizhi Li, Yinghao Ma, Yiming Liang, Xiaowei Chi, Ruibo Liu, Zili Wang, Pengfei Li, Jingcheng Wu, Chenghua Lin, Qifeng Liu, Tao Jiang, Wenhao Huang, Wenhu Chen, Emmanouil Benetos, Jie Fu, Gus Xia, Roger Dannenberg, Wei Xue, Shiyin Kang, Yike Guo

Zofia Łągiewka, Jacek Zalewski

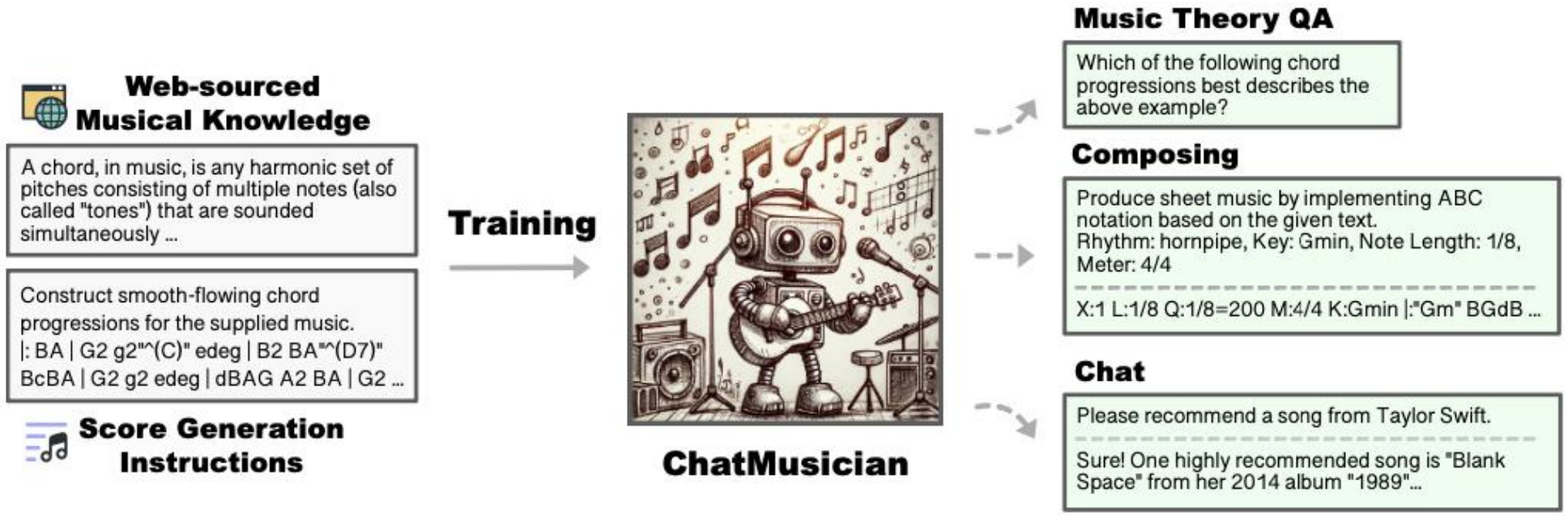


Figure 1: ChatMusician learns from web-sourced musical knowledge and handcrafted music score generation instructions, unifies music generation and music understanding, and can chat, compose, and answer college-level music theory questions.

Why adapting LLMs for music is so challenging

- ChatMusician integrates music understanding and generation into LLMs.
- AI in the arts, particularly music, faces unique challenges due to music's structured nature.
- GPT-4 and other models perform poorly on complex musical tasks, motivating the creation of a specialized model.

Music representation

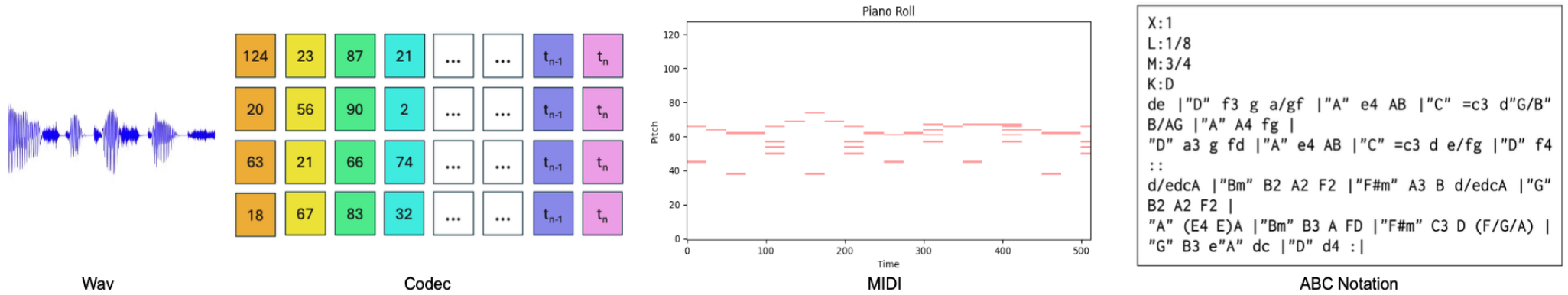


Figure 2: Commonly used music representations, including Wav, Codec, MIDI (visualized as piano roll), and ABC notation. From left to right, the compression rate gets higher.

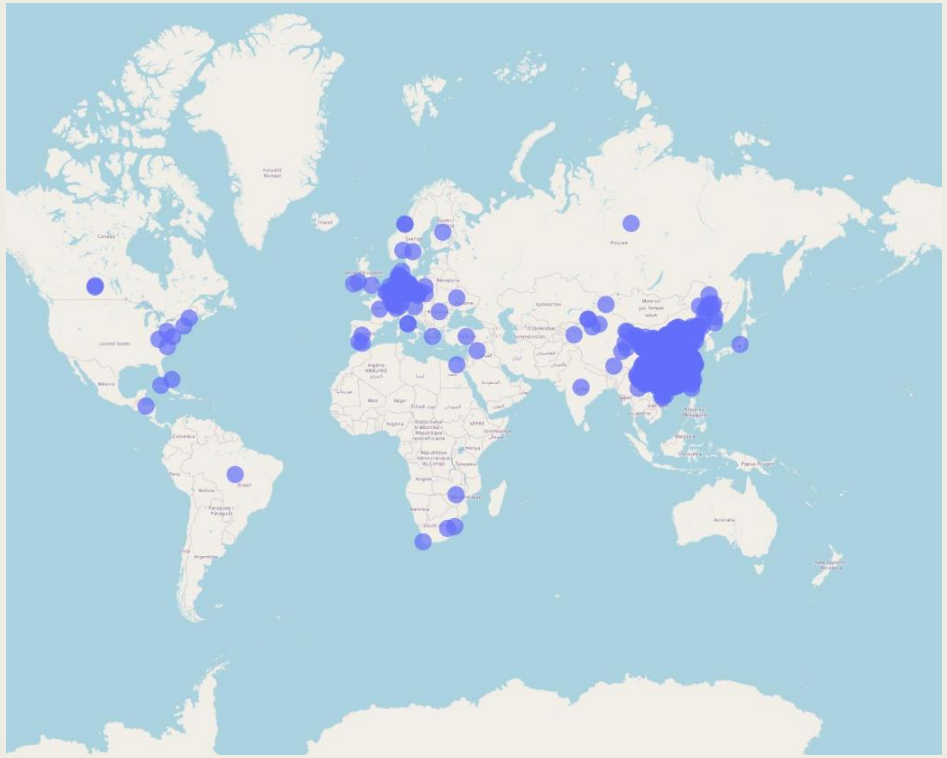
Dataset Composition

Datasets	Sourced from	Tokens	# Samples	Category	Format
Pile (Gao et al., 2020)	public dataset	0.83B	18K	general	article
Falcon-RefinedWeb (Penedo et al., 2023)	public dataset	0.80B	101K	general	article
Wikipedia (Wikipedia contributors, 2023)	public dataset	0.39B	588K	general	article
OpenChat (Wang et al., 2023a)	public dataset	62.44M	43K	general	chat
LinkSoul (LinkSoul-AI, 2023)	public dataset	0.6B	1.5M	general	chat
GPT4-Alpaca (Peng et al., 2023)	public dataset	9.77M	49K	general	chat
Dolly (Conover et al., 2023)	public dataset	3.12M	14K	general	chat
Irishman (Wu and Sun, 2023)	public dataset + Human-written Instructions	0.23B	868K	music score	chat
KernScores (CCARH at Stanford University, 2023)	public dataset + Human-written Instructions	2.76M	10K	music score	chat
Bach (Wu et al., 2023)	public dataset + Human-written Instructions	0.44M	349	music score	chat
synthetic music chat☆	public dataset + Human-written Instructions	0.54B	50K	music score	chat
music knowledge★	Generated w/ GPT-4	0.22B	255K	music verbal	chat
music summary★	Generated w/ GPT-4	0.21B	500K	music verbal	chat
GSM8k (Cobbe et al., 2021)	public dataset	1.68M	7K	math	chat
math (Kenney, 2023)	public dataset	7.03M	37K	math	chat
MathInstruct (Yue et al., 2023)	public dataset	55.50M	188K	math	chat
Camel-Math (Li et al., 2023)	public dataset	27.76M	50K	math	chat
arxiv-math-instruct-50k (Kenney, 2023)	public dataset	9.06M	50K	math	chat
Camel-Code (Li et al., 2023)	public dataset	0.13B	366K	code	chat
OpenCoder (Wang et al., 2023a)	public dataset	36.99M	28K	code	chat
Total		4.16B	5.17M		

Music Score Corpora

- Chord Conditioned Music Generation
- Musical Form Conditioned Music Generation
- Alphabetic Musical Form and Motif Conditioned
- Music Generation Terminology Musical Form and Motif conditioned Music Generation
- Melody Harmonization
- Bach's Style Music Generation

❑ Motif Extraction
❑ Musical Form Extraction



MusicTheoryBench

Examples of questions from:

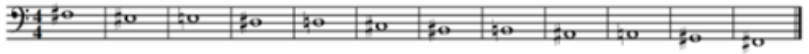
(a) music knowledge

(b) music reasoning

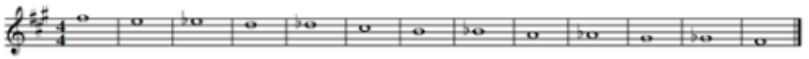
Question: "Which of the following statements about triads is correct?",
Options:
A: "A triad can only be composed of three notes.",
B: "All triads are consonant chords.",
C: "When a triad is inverted, its properties and consonance remain consistent with its original position.",
D: "A triad is defined as a chord formed by stacking three notes in a third relationship."},
Answer: D

(a)

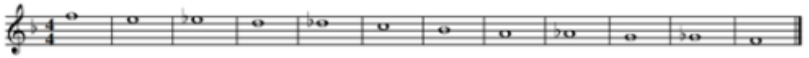
Question: "Which of the following is a descending natural minor scale with E as the leading tone?",
Options:
A: "L:1/4\nM:4/4\nK:C\n ^F,4 | ^E,4 | =E,4 | ^D,4 | =D,4 | ^C,4 | ^B,,4 | =B,,4 | ^A,,4 | =A,,4 | ^G,,4 | ^F,,4 |] %12",



B: "L:1/4\nM:4/4\nK:A\n f4 | e4 | _e4 | d4 | _d4 | c4 | B4 | _B4 | A4 | _A4 | G4 | _G4 | F4 |] %13",



C: "L:1/4\nM:4/4\nK:F\n f4 | e4 | _e4 | d4 | _d4 | c4 | B4 | A4 | _A4 | G4 | _G4 | F4 |] %12",



D: "None of the first three options are correct"},
Answer: D

(b)

Training Settings and Data Settings

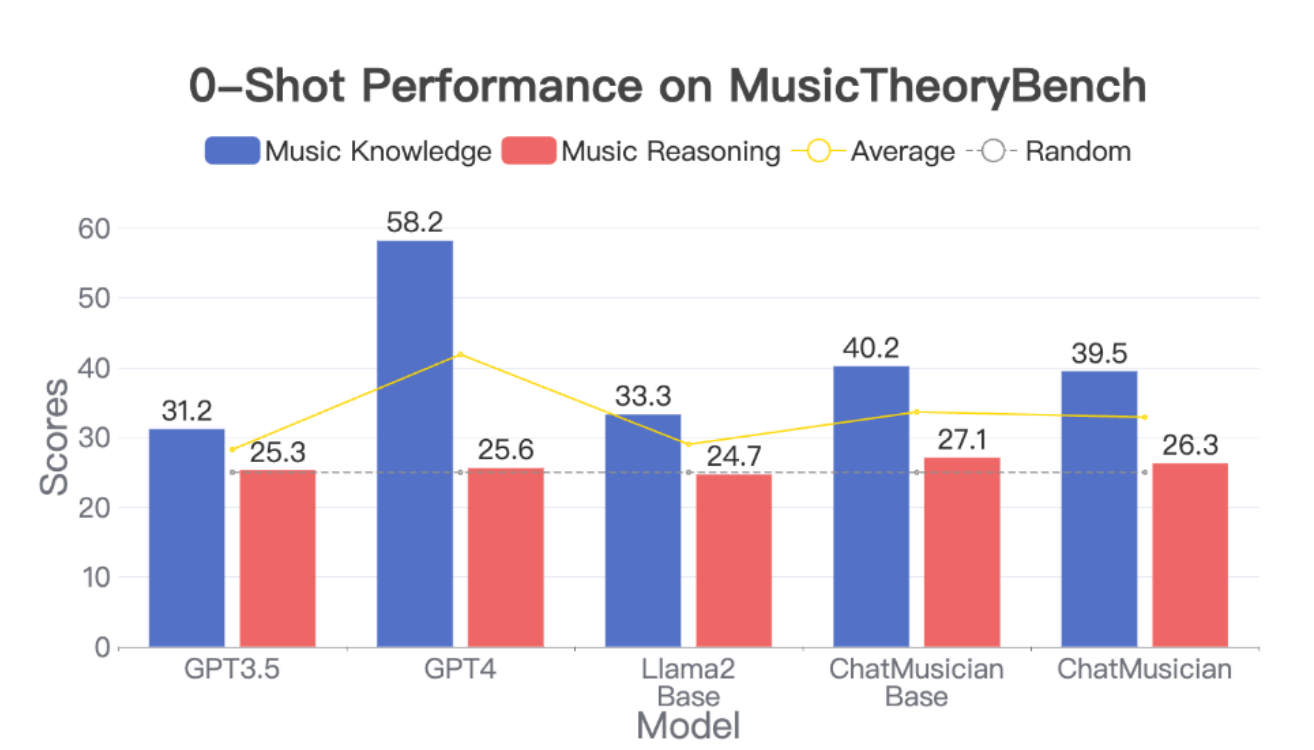
- LLaMA2-7B-Base weights
- LoRA adapters
- maximum sequence length was 2048

- one epoch training
- 2:1 ratio between music scores and music knowledge&music summary data
- 78K samples from the training set and trained for 10 epoch
- 1.1M samples and trained for 2 epochs

Evaluation and Baseline Systems

- Baseline Systems (GPT-3.5, GPT-4, and LLaMA-2)
- Evaluation of General Language Abilities (MMLU)
- Evaluation of Music Understanding Abilities (MusicTheoryBench, average accuracy)
- Evaluation of Music Generation Abilities. (phrase-level repetition metric and a parsing success rate metric, human preference)

Music Understanding



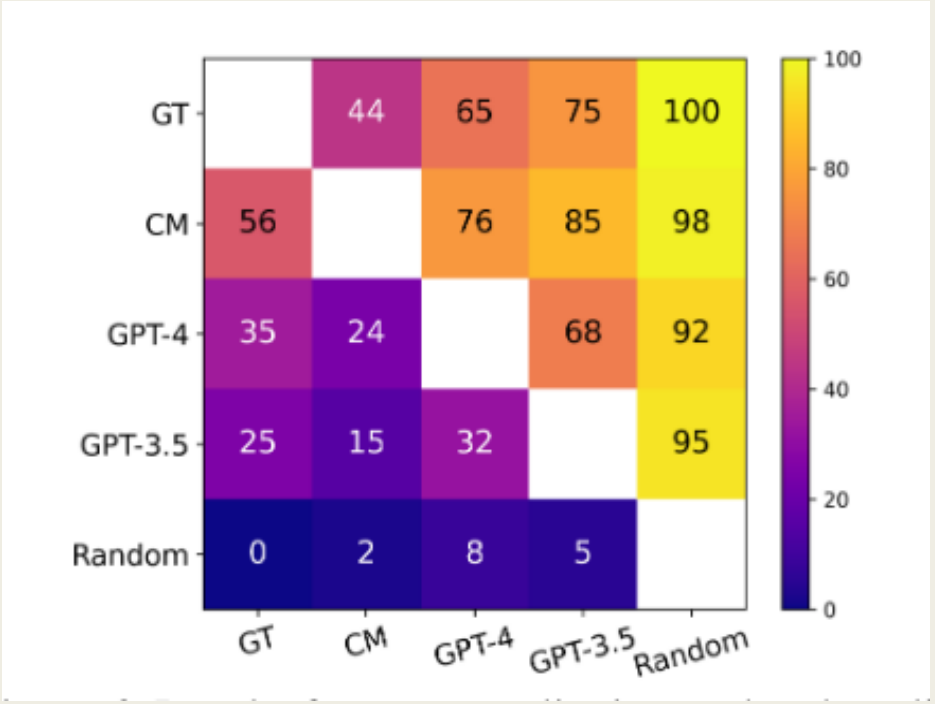
Method	Mus. Knowledge	Mus. Reasoning
GPT4-0-shot	58.2	25.6
+5-shot ICL	64.1	38.0
GPT4-RolePlay	68.3	36.6
+5-shot ICL	68.8	**39.5**
GPT4-CoT	68.4	36.7
+5-shot ICL	**69.9**	34.9

Compression Ratio of ABC Notation

Format	Tokenizer	Tok./Song	Tok./Sec.
ABC	LLaMA Tokenizer	288.21	5.16
MIDI	REMI(Huang and Yang, 2020b)	753.41	12.84
MIDI	MIDI-like(Oore et al., 2018)	728.60	12.42
WAV	EnCodec(Défossez et al., 2022)	12577.46	200.00

Musicality

- Learning Music Repetitions
- Human Evaluation.
- Qualitative Study



System	Repetition Det. Rate(%)
ChatMusician	**76.0**
GPT-4	70.2
GPT-3.5	32.2

Example

```
X:1
L:1/8
M:2/4
K:F
F/G/ |:"F" BA"C7" GG |"F" FA"C7" G2 |"F" F>G"C7" AB |
"Am" cA"C7" GF/G/ |"F" BA"C7" GG |"F" FA"C7" G2 |"F" F>G"Bb" Bd |
1"C7" cE"F" FF/G/ :|2"C7" cE"F" F z |:"F" f3 (c/d/)(d/e/) |
"Gm" (e/f/)(f/g/) g>ec |"C7" e/d/ d/c/c/B/ B/A/A/G/ |"F" GA/B/ c/d/e/f/ | f3 (c/d/)(d/e/) |
"Gm" (e/f/)(f/g/) g>ec |"C7" e/d/ d/c/c/B/ B/A/A/G/ |"F" FA/c/ f z :|
```

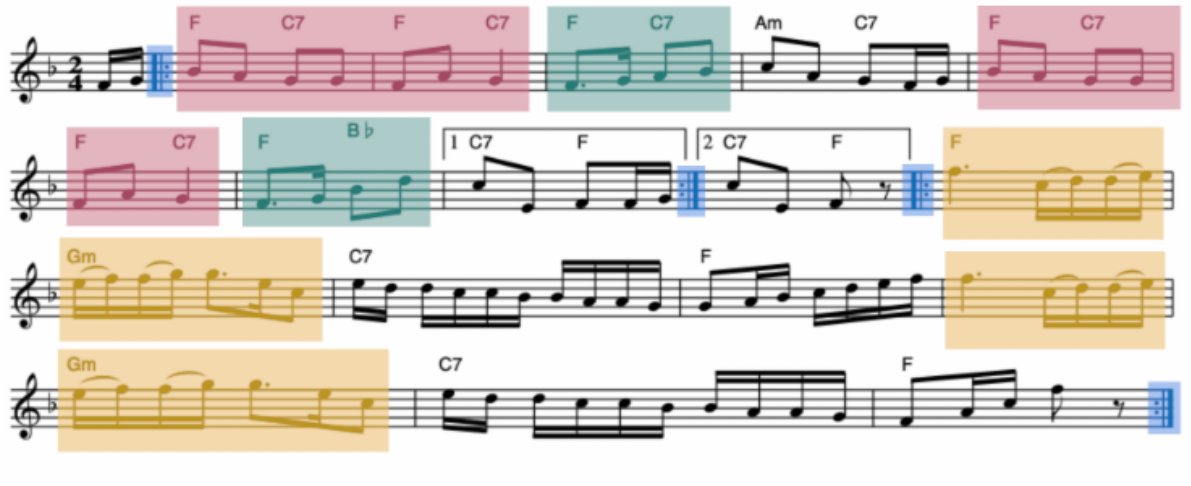
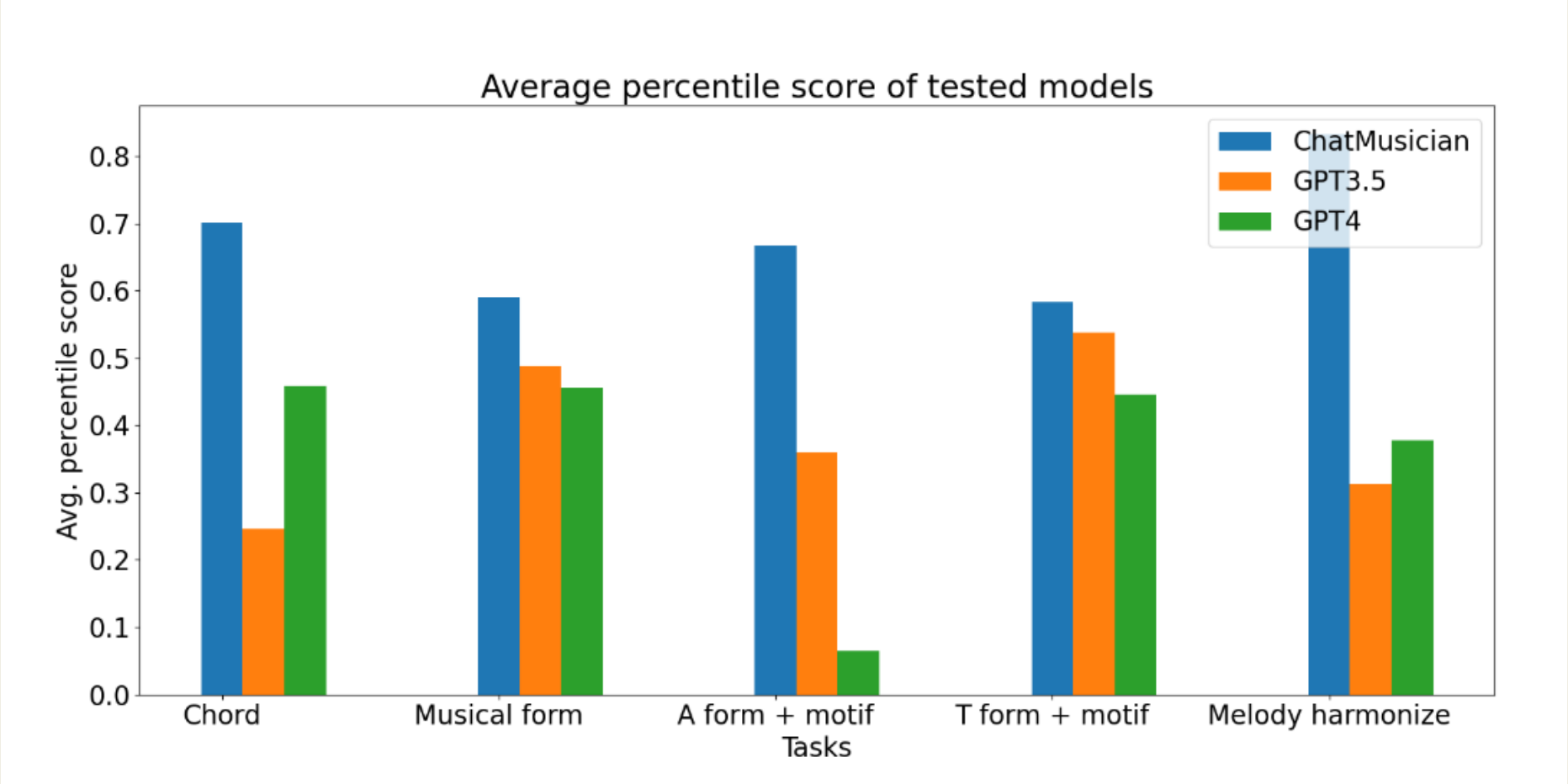


Figure 7: ABC notation and corresponding staff notation of a generated music. Repetition symbols are marked blue in both notations and demonstrate a clear phrase-level repetition. Red and yellow rectangles mark clear motif-level repetition in both sections. Green rectangles mark variation notes following the motif of the first section.

Task-wise metric



Language Ability

System	MMLU Score(%)
ChatMusician-Base	**48.50**
ChatMusician	46.80
LLaMA2-7B-Base	46.79

Table 7: MMLU score of ChatMusicians and LLaMA2-7B-Base.

Demo