TAPA-ICL: Taxonomy-Aware Prompt Augmentation for In-Context Learning in Music Understanding

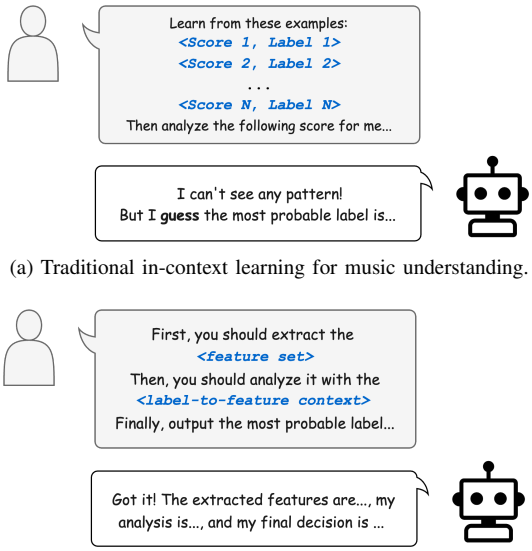
Jiahao Zhao*, Yunjia Li† and Kazuyoshi Yoshii‡
* Graduate School of Informatics, Kyoto University, Kyoto, Japan
† School of Computer Science and Technology, Fudan University, Shanghai, China
‡ Graduate School of Engineering, Kyoto University, Kyoto, Japan
E-mail: zhao.jiahao.56h@st.kyoto-u.ac.jp, 24210240033@m.fudan.edu.cn, yoshii.kazuyoshi.3r@kyoto-u.ac.jp

***Abstract*—This paper presents TAPA-ICL, a novel in-context learning (ICL) framework for few/zero-shot symbolic music understanding task that generates human-readable analysis. Conventional ICL approaches mainly rely on the ability of large language models (LLMs) to infer patterns and tasks from contextual examples. However, while LLMs have shown basic capability of understanding symbolic music, how to enhance such capability via additional context still remains unexplored. To tackle this issue, we focus on the two major challenges as follows: (1) *Sparsity of Score Input* through taxonomy-aware prompt augmentation (TAPA) that distills label-to-feature profiles, reducing context length by over 10 times; and (2) *Complexity of Musical Semantics Reasoning* via structured chain-of-thought (CoT) prompts enforcing feature extraction, context-aware analysis, and decision-making. By combining TAPA strategy and the CoT reasoning prompts, our method enables effective few/zero-shot adaptation across emotion recognition, composer identification, and genre classification tasks. Experimental results show that our TAPA-ICL method significantly outperforms conventional few-shot ICL baselines (including those based on ultra-large and mixture of experts (MoE) models) on each downstream task and achieves slightly weaker performance to existing many-shot approaches.**

I. Introduction

Symbolic music understanding refers to the computational analysis of music represented in symbolic formats (e.g., sheet music, MIDI, or ABC notations) [1], [2]. The mainstream paradigm in current research involves pre-training large-scale music-language models for general music modeling, then fine-tuning them on downstream tasks [3]. However, such approaches are limited to producing music embeddings and fixed labels, while developing user-facing systems requires human-readable analytical content. In this work, we propose to leverage large language models (LLMs) to bridge this gap by generating interpretable, natural-language descriptions of symbolic musics.

Extensive validation across tasks such as music generation [4] and music knowledge QA [5], [6] demonstrates that LLMs has acquired systematic music-theoretical knowledge and basic capabilities in understanding and generating symbolic music via pre-training. While few-shot in-context learning (ICL) has proven to be able to achieve training-free adaption through limited samples on many text-based tasks [7] (e.g., named entity recognition (NER) [8] or sentiment analysis [9]), it remains limited in adaption of symbolic music understanding tasks. To tackle this issue, we identify two core challenges of few-shot ICL in music understanding and develop targeted solutions:



(a) Traditional in-context learning for music understanding.

(b) Our proposed TAPA-ICL for music understanding.

Fig. 1: Conceptual comparison between traditional ICL and proposed TAPA-ICL, blue text refers to the inserted context.

Challenge 1: Sparsity of Score Input. Few-shot in-context learning (ICL) for symbolic music analysis struggles with extreme input sequence lengths. Even using compact ABC notation, the length of individual samples could reach several thousand tokens. This leads to two major bottlenecks: information degradation from long-range attention sparsity, and computational inefficiency from quadratic memory growth. To overcome this, we redesign the contextual demonstration paradigm. Rather than presenting raw score-to-label examples, we instead employ distilled label-to-feature profiles, achieving a great reduction in context length (10 times shorter). These high-efficiency demonstrations are automatically generated by LLMs through two novel strategies: (1) a self-heuristic zero-shot approach that generate label profiles from the LLM's inherent knowledge, and (2) a few-shot method that distills label profiles from several provided ground truth annotations.

Challenge 2:Complexity of Musical Semantics Reason-

979-8-3315-7206-8/25/$31.00

ing. Explicit and learnable patterns can hardly be found in music understanding tasks with limited amount of sample-to-label demonstrations. This is primarily due to music's inherent hierarchical organization and the implicit multi-stage reasoning process required for musical analysis, which makes it even more difficult to tackle the task through single LLM inference. Current solutions include task decomposition strategies like multi-agent systems with iterative inference [4]. To address this, we augment few-shot ICL with a structured multi-stage Chain-of-Thought (CoT) prompt: First, the model is instructed to extract 12 predefined musical features (spanning harmony, rhythm, melody, and structure) from ABC notation inputs; second, it establishes feature-label mappings using the provided label-to-feature context; finally, it synthesizes these reasoning steps to produce the final prediction. This approach effectively collapses traditional multi-task, multi-stage processing into a single inference.

Based on the solutions above, in this work we propose TAPA-ICL, a novel in-context learning method for few/zero-shot symbolic music understanding with LLMs. TAPA-ICL enables the LLMs to perform better-organized reasoning than traditional few-shot ICL methods, we show their conceptual comparison in Fig1. The main contributions of this work can be summarized as follows:

- We present the first attempt of utilizing few/zero-shot ICL strategy with LLMs for general symbolic music understanding tasks. We introduce a novel taxonomy-aware prompt augmentation (TAPA) strategy to obtain label-to-feature context for effective downstream task adaption, which requires no manual prompt engineering under both few-shot and zero-shot scenarios.
- We develop a multi-stage CoT prompt injection mechanism specifically designed for symbolic music characteristics. Our approach explicitly provides the model with structured reasoning templates through demonstration examples, enforcing a three-stage reasoning process of feature extraction, context-aware analysis, and decision making. This achieves an efficient and interpretable mapping from symbols to semantics within a single inference of LLMs.
- We implement and evaluate our method across multiple downstream music understanding tasks, including emotion recognition, composer identification, and genre classification. Our model outperforms all few-shot ICL baselines, including implementations based on ultra-large or Mixture-of-Expert (MoE) models. It demonstrates training-free generalization capability on music understanding and achieves performance comparable to some many-shot baselines on certain tasks.

II. Related Studies

A. LLMs for Symbolic Music

Symbolic music can be regarded as the "language of music," represented by discrete token sequences [10], [11], which enables processing similar to natural language. Inspired by the powerful BERT encoder pre-training architecture, pre-trained music language models (MLMs) such as MidiBERT and MusicBERT [3], [12] have been developed. Although MLMs have demonstrated promising performance in both token-level and sequence-level downstream understanding tasks, their practical applications face limitations due to strict input/output format requirements, human-unreadable characteristics, and suboptimal task adaptation capabilities when fine-tuning data is limited.

To address these challenges, the potential of LLMs for symbolic music generation and understanding tasks is being progressively explored. Zhou et al. [13] investigated the fundamental understanding and generation capabilities of LLMs on text-formatted ABC notation scores across multiple base models, while enhancing these capabilities through prompt engineering. Under the training-free paradigm, Deng et al. proposed ComposerX [4], which decomposes music generation into three stages (plan, compose, and review) with defined subtasks, achieving highly controllable symbolic music generation through multi-agent collaboration. For musical knowledge understanding and question-answering tasks, Yuan et al.'s ChatMusician [5] demonstrates promising music knowledge comprehension and reasoning abilities using supervised fine-tuning (SFT) and few-shot ICL. However, the few/zero-shot understanding potential of LLMs for high-level semantic tasks such as genre classification and emotion recognition remains underexplored.

III. Proposed Method

The overall workflow of our proposed method is illustrated in figure 2. The overall input consists of three components: TAPA context consisting of label-to-feature profiles, structured CoT prompts, and ABC score input. The LLM base model performs single inference based on this input, outputting the extracted feature set, reasoning process, and final predicted labels. In the following subsections, we will introduce the two kinds of employed TAPA strategies and the structured CoT prompts in detail.

A. Taxonomy-Aware Prompt Augmentation

In this subsection we introduce our proposed taxonomy-aware prompt augmentation (TAPA) strategy. Instead of using sample-label pairs as input context, TAPA constructs label-feature profiles as more concise learnable contexts. Given the target label set and a handcrafted feature set (12 features derived from harmony, rhythm, melody, and structure), TAPA context is to build label-feature profiles that describe the mapping relationships between each label and each feature.

These label-feature profiles consist of descriptions in the form of (Label Name: Feature Name - Corresponding Feature Attribute). Considering the instability of LLMs in numerical computation during long-range reasoning and their strong capability for high-level semantic abstraction, we use qualitative descriptions (e.g., high/low/positive) when annotating feature attributes. In practice, we adopt two distinct TAPA strategies, corresponding to zero-shot and few-shot scenarios:

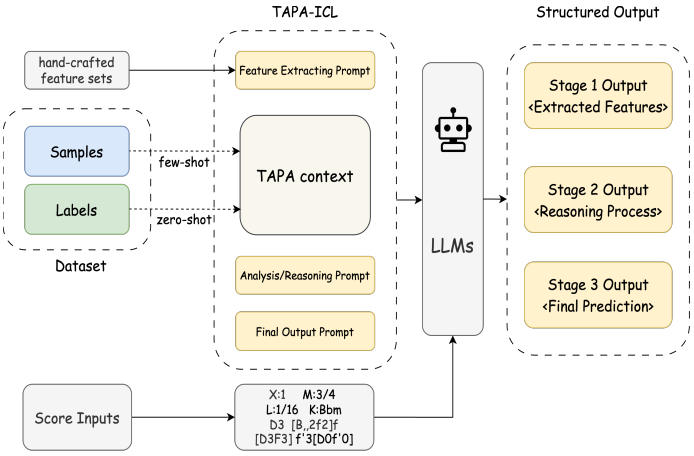


Fig. 2: The overall structure and workflow of our proposed TAPA-ICL method. The input consists of hand-crafted CoT prompts, TAPA context obtained from ground truth samples or labels and the ABC score input.

Zero-shot TAPA: In this case, we instruct the LLM to summarize the characteristics of each label based on its background knowledge and generate corresponding feature descriptions according to the provided feature set. The input to the model consists of the label set and the feature set to be analyzed.

Few-shot TAPA: Here, we provide the LLM with several samples of the same label as demonstrations and it is instructed to summarize the common characteristics of the input samples as the corresponding label profile, based on the feature set.

It is worth noting that the few-shot TAPA also basically meets the criteria of zero-shot methods. Although in this implementation the few-shot TAPA context is derived from ground truth samples, it only summarizes the features associated with the labels, and this content cannot be reconstructed into the original ground truth samples. Nevertheless, for the sake of thorough rigor, we still consider this approach to be few-shot.

B. Structured Chain-of-Thought Reasoning

We abstract and formalize the feature-to-label reasoning process as three distinct tasks, employing structured prompts to guide the model through this three-stage reasoning.

The first stage involves extracting assigned feature groups from the input $\mathbf{x}$, which can be abstracted by the following formula:

$$\phi(\mathbf{x}) = \big(f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_m(\mathbf{x})\big)^\top \in \mathbb{R}^m \tag{1}$$

where f_n denotes the n-th feature extraction function in the feature group, m represents the feature dimension (set to 12 in our implementation), and $\phi(\mathbf{x})$ is the feature mapping from input $\mathbf{x}$ to the feature space.

The second stage estimates the feature-to-label probability mapping by comparing the extracted features $\phi(\mathbf{x})$ with the provided TAPA context, abstracted by:

TABLE I: Content of the structured CoT prompt we used, the feature set is omitted due to insufficient space.

Structured CoT Prompt

```
Analyze the provided ABC music notation
and classify it through strict
multi-stage reasoning. Proceed step by
step without skipping stages.

=== Stage 1: Core Feature Extraction ===
Extract these 12 objective features from
the ABC notation (Output the feature name
and result without inference process):
<Detailed features omitted for space>
Output: <Extracted_Features>

=== Stage 2: Analysis and Reasoning ===
Compare each extracted feature against
the provided label profile database.
For each feature, identify which genre
characteristics it matches and how
strongly (output format: 'feature name:
label name, strong/weak/moderate').
Output the only the results of matched
feature and characteristic concisely in
plain text.
Output: <Feature_Matching_Analysis>

=== Stage 3: Label Selection ===
Based on the reasoning, output only the
single most probable label name from the
label profile without any formatting,
punctuation or additional.
Output: <Selected_Label>
```

$$\Gamma(\phi(\mathbf{x})) = \big(p(y_1 \mid \phi(\mathbf{x})),\ p(y_2 \mid \phi(\mathbf{x})),\ \ldots,\ p(y_k \mid \phi(\mathbf{x}))\big)^\top \tag{2}$$

where y_n refers to the n-th label in the label space, k is the label dimension, and Γ represents the probability mapping function.

Finally, based on the probability mapping $\Gamma(\phi(\mathbf{x}))$ obtained in the second stage, the third stage aims to identify the predicted label y with maximum probability:

$$\hat{y} = \arg\max_{y \in \mathcal{Y}} p(y \mid \phi(\mathbf{x})) = \arg\max_{y \in \mathcal{Y}} \Gamma_y(\phi(\mathbf{x})) \tag{3}$$

where $\mathcal{Y}$ denotes the complete label space and $\hat{y}$ is the predicted label with highest probability.

Based on these three sub-tasks, we manually designed a structure CoT prompt to guide the reasoning process. The exact content of the prompt is shown in Table.I.

IV. Experiments

A. Research Questions

We designed our experiments and ablation studies to answer the two core research questions:

TABLE II: Performance Comparison Across Baseline Methods (Accuracy)

	Proposed Methods		Few-Shot In-Context Learning				
Task	**TAPA-ICL (zero-shot)**	**TAPA-ICL (few-shot)**	**DS-Reasoner**	**DS-Chat**	**Gemini-2.5-flash**	**Gemini-2.5-pro**	**Qwen3-32B**
Emotion	**59.09%**	52.27%	25.00%	34.09%	36.36%	36.36%	18.18%
Composer	47.73%	**53.41%**	26.67%	20.00%	18.89%	22.22%	26.67%
Genre	**61.82%**	56.36%	30.91%	38.18%	32.73%	29.09%	41.81%

TABLE III: Performance Comparison Across Many-Shot Methods (Accuracy)

	Proposed Methods		Pre-trained MLMs			MER Model	
Task	**TAPA-ICL (zero-shot)**	**TAPA-ICL (few-shot)**	**MidiBERT-Piano**	**Music-BERT**	**RoAR**	**MM_MER**	**SCMA**
Emotion	59.09%	52.27%	70.64%	71.06%	76.15%	69.20%	71.40%
Composer	47.73%	53.41%	78.57%	86.05%	80.95%	/	/

- Do LLMs possess inherent few/zero-shot music understanding capabilities? If not, to what extent can our method enhance such abilities?
- Does our approach demonstrate sufficient generalization and robustness? Can it consistently improve performance across different tasks and datasets under varying taxonomies?

B. Implementation Details

We evaluated our approach on three music understanding tasks—emotion classification, composer identification, and genre classification—using the EMOPIA [14], Pianist8 [3], and ADL Piano MIDI datasets [15] respectively. For emotion classification task, we adopted Russell's four-quadrant model [16] as the framework for a 4-class classification. The composer task is to perform an 8-class classification between eight modern composers, while the genre classification utilized a simplified version of ADL Piano MIDI covering five representative categories: classical, electronic, jazz, pop, and rock. To compete against state-of-the-art LLM base models and optimize the computational efficiency, we compared our method with conventional few-shot ICL approaches via API services including Deepseek-reasoner, Deepseek-chat [17], Gemini-2.5-Flash, and Gemini-2.5-Pro [18]. We also deployed Qwen3-32B **qwne3** as the local base model with four NVIDIA RTX-A6000 GPUs.

C. Comparison with ICL-Based Methods

In this section, we compare our method with traditional few-shot ICL-based LLM baselines. By examining the outputs of these baselines, we observe that the models do not exhibit hallucination phenomena. Although they can generate analysis content that adheres to formal specifications, their reasoning effectiveness remains poor, demonstrating accuracy close to random classification in most tasks.

Our proposed few/zero-shot TAPA-ICL model achieves significantly higher recognition accuracy than baselines across all subtasks (emotion, composer, and genre), with performance improvements of 22.73 pts, 26.74 pts, and 23.64 pts respectively. This demonstrates the superiority of our TAPA-ICL strategy over traditional few-shot ICL for general music understanding tasks. It also answers our research question: LLMs indeed possess potential few/zero-shot symbolic music understanding capabilities, and our method can effectively utilize and substantially enhance these capabilities.

Through result analysis, we further summarize the following observations: (1) Larger and more powerful base models do not necessarily perform better on few-shot music understanding tasks (e.g., DS-Reasoner vs DS-Chat). We attribute this to randomness having greater impact than the model's inherent reasoning capacity when prompt engineering is insufficient; (2) Zero-shot TAPA performs better on emotion and genre tasks, while few-shot TAPA excels in composer recognition. We hypothesize that LLMs' pretraining gives them stronger prior knowledge about emotion and genre labels, enabling better zero-shot label profile construction. For the more complex composer task with numerous categories, few-shot TAPA can extract basic patterns from provided samples, leading to superior performance.

D. Comparison with Many-Shot Methods

To investigate the practical applicability of TAPA-ICL, we further compare it with existing many-shot approaches. Our comparison includes pre-trained MLMs based on BERT architectures: MidiBERT-Piano [3], MusicBERT [12], and RoAR [19]. These models possess over 100M parameters and strong music modeling capabilities through large-scale pre-training. We also benchmark against specialized models for music emotion recognition (MM_MER [20] and SCMA [21]). Due to our genre dataset being a simplified subset, we exclude genre comparisons for fairness.

On the emotion task, our model demonstrates competitive performance under zero-shot constraints, achieving results within 10.2 pts of the many-shot MM_MER method. For the more challenging 8-class composer recognition task, where few/zero-shot adaptation is inherently more difficult, our approach shows a 25.16 pts performance gap compared to many-shot methods.

TABLE IV: Ablation Study Results on TAPA-ICL Components (Accuracy)

	Proposed Methods		Ablated Implementations		
Task	**TAPA-ICL (zero-shot)**	**TAPA-ICL (few-shot)**	**w/o_TAPA**	**w/o_CoT**	**Qwen3-Based (zero-shot)**
Emotion	**59.09%**	52.27%	38.64%	29.55%	54.55%
Composer	47.73%	**53.41%**	31.82%	15.91%	43.18%
Genre	**61.82%**	56.36%	43.64%	32.73%	49.09%

E. Ablation Studies

To validate the contribution of each component in TAPA-ICL, we designed three ablated versions for comparative experiments: (1) w/o_TAPA, where we removed the TAPA context and only provided the label set; (2) w/o_CoT, where we eliminated the structured CoT prompt while retaining the TAPA context; and (3) Qwen3-Based, where we replaced the API service with a locally deployed Qwen3-8B lightweight model to test the stability of our method on smaller open-source models.

As shown in Table IV, both few-shot and zero-shot TAPA-ICL outperform all ablated versions across three downstream tasks, demonstrating the effectiveness of our proposed modifications. The w/o_TAPA version exhibits smaller performance degradation compared to w/o_CoT, indicating that while the model possesses inherent knowledge about labels, it cannot perform multi-stage reasoning from feature extraction to analysis without proper prompting. The Qwen3-Based version, implemented on the local Qwen3-8B model, shows performance degradation but still maintains significant advantages over traditional few-shot ICL methods, proving the consistent effectiveness of our approach across different base models.

V. Conclusions

This study investigates the potential of utilizing few/zero-shot adaption on generalized music understanding tasks with in-context learning. A novel context generation method TAPA is proposed to produce label-to-feature profile for effective task adaption. The proposed method also contains a detailed chain-of-thought prompt to enforcing a mutli-stage reasoning of music semantics. Experimental results and ablation studies have proven the effectiveness of our proposed strategies.

Our future work will focus on developing more precise generation methods of TAPA context and exploring the potential of TAPA-ICL on other downstream tasks.

VI. Acknowledgement

This work was partially supported by JST FOREST Grant No. JPMJFR2270 and JSPS KAKENHI Grant Nos. 24H00742, 24H00748, 25K22841, and 25H01142.

References

[1] Y.-H. Yang and H. H. Chen, "Machine recognition of music emotion: A review," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 3, pp. 1–30, 2012.

[2] S. Ji, X. Yang, and J. Luo, "A survey on deep learning for symbolic music generation: Representations, algorithms, evaluations, and challenges," *ACM Computing Surveys*, vol. 56, no. 1, pp. 1–39, 2023.

[3] Y. Chou, I. Chen, C. Chang, J. Ching, and Y. Yang, "Midibert-piano: Large-scale pre-training for symbolic music understanding," *CoRR*, vol. abs/2107.05223, 2021. arXiv: 2107.05223. [Online]. Available: https://arxiv.org/abs/2107.05223.

[4] Q. Deng, Q. Yang, R. Yuan, *et al.*, "Composerx: Multi-agent symbolic music composition with llms," in *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024, San Francisco, California, USA and Online, November 10-14, 2024*, B. Kaneshiro, G. J. Mysore, O. Nieto, *et al.*, Eds., 2024, pp. 669–679. DOI: 10.5281/ZENODO.14877425. [Online]. Available: https://doi.org/10.5281/zenodo.14877425.

[5] R. Yuan, H. Lin, Y. Wang, *et al.*, "Chatmusician: Understanding and generating music intrinsically with LLM," in *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, L. Ku, A. Martins, and V. Srikumar, Eds., Association for Computational Linguistics, 2024, pp. 6252–6271. DOI: 10.18653/V1/2024.FINDINGS-ACL.373. [Online]. Available: https://doi.org/10.18653/v1/2024.findings-acl.373.

[6] B. Weck, I. Manco, E. Benetos, E. Quinton, G. Fazekas, and D. Bogdanov, "Muchomusic: Evaluating music understanding in multimodal audio-language models," in *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024, San Francisco, California, USA and Online, November 10-14, 2024*, B. Kaneshiro, G. J. Mysore, O. Nieto, *et al.*, Eds., 2024, pp. 825–833. DOI: 10.5281/ZENODO.14877459. [Online]. Available: https://doi.org/10.5281/zenodo.14877459.

[7] Q. Dong, L. Li, D. Dai, *et al.*, "A survey on in-context learning," in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, Y. Al-Onaizan, M. Bansal, and Y. Chen, Eds., Association for Computational Linguistics, 2024, pp. 1107–1128. DOI: 10.18653/V1/2024.EMNLP-MAIN.64. [Online]. Available: https://doi.org/10.18653/v1/2024.emnlp-main.64.

[8] Z. Wang, Z. Zhao, Y. Lyu, Z. Chen, M. de Rijke, and Z. Ren, "A cooperative multi-agent framework for zero-shot named entity recognition," in *Proceedings of the ACM on Web Conference 2025*, 2025, pp. 4183–4195.

[9] H. Xu, Q. Wang, Y. Zhang, *et al.*, "Improving in-context learning with prediction feedback for sentiment analysis," in *ACL (Findings)*, 2024.

[10] Y.-S. Huang and Y.-H. Yang, "Pop music transformer: Beat-based modeling and generation of expressive pop piano compositions," in *Proceedings of the 28th ACM International Conference on Multimedia*, 2020, pp. 1180–1188.

[11] S. Oore, I. Simon, S. Dieleman, D. Eck, and K. Simonyan, "This time with feeling: Learning expressive musical performance," *Neural Computing and Applications*, vol. 32, no. 4, pp. 955–967, 2020.

[12] H. Zhu, Y. Niu, D. Fu, and H. Wang, "Musicbert: A self-supervised learning of music representation," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 3955–3963.

[13] Z. Zhou, Y. Wu, Z. Wu, *et al.*, "Can llms "reason" in music? an evaluation of llms' capability of music understanding and generation," in *Proceedings of the 25th International Society for Music Information Retrieval Conference, ISMIR 2024, San Francisco, California, USA and Online, November 10-14, 2024*, B. Kaneshiro, G. J. Mysore, O. Nieto, *et al.*, Eds., 2024, pp. 103–110. DOI: 10.5281/ZENODO.14877281. [Online]. Available: https://doi.org/10.5281/zenodo.14877281.

[14] H.-T. Hung, J. Ching, S. Doh, N. Kim, J. Nam, and Y.-H. Yang, "Emopia: A multi-modal pop piano dataset for emotion recognition and emotion-based music generation," *arXiv preprint arXiv:2108.01374*, 2021.

[15] L. N. Ferreira, L. H. Lelis, and J. Whitehead, "Computer-generated music for tabletop role-playing games," AIIDE'20, 2020.

[16] J. A. Russell, "A circumplex model of affect.," *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[17] D. Guo, D. Yang, H. Zhang, *et al.*, "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," *arXiv preprint arXiv:2501.12948*, 2025.

[18] G. Team, R. Anil, S. Borgeaud, *et al.*, "Gemini: A family of highly capable multimodal models," *arXiv preprint arXiv:2312.11805*, 2023.

[19] Z. Li, R. Gong, Y. Chen, and K. Su, "Fine-grained position helps memorizing more, a novel music compound transformer model with feature interaction fusion," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 37, 2023, pp. 5203–5212.

[20] J. Zhao and K. Yoshii, "Multimodal multifaceted music emotion recognition based on self-attentive fusion of psychology-inspired symbolic and acoustic features," in *2023 Asia Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, 2023, pp. 1641–1645. DOI: 10.1109/APSIPAASC58517.2023.10317539.

[21] Y. Xiao, H. Ruan, X. Zhao, P. Jin, and X. Cai, "Music emotion recognition using multi-head self-attention-based models," in *Advanced Intelligent Computing Technology and Applications: 19th International Conference, ICIC 2023, Zhengzhou, China, August 10–13, 2023, Proceedings, Part IV*, Zhengzhou, China: Springer-Verlag, 2023, pp. 101–114, ISBN: 978-981-99-4751-5. DOI: 10.1007/978-981-99-4752-2_9. [Online]. Available: https://doi.org/10.1007/978-981-99-4752-2_9.