Do music LLMs learn symbolic concepts? A pilot study using probing and intervention

Wenye Ma **Xinyue Li** **Gus Xia**
Department of Machine Learning, MBZUAI
wenye.ma@mbzuai.ac.ae, xinyue.li@mbzuai.ac.ae, gus.xia@mbzuai.ac.ae

Abstract

Music large language models (LLMs) have shown impressive capabilities in generating long-term, high-quality music trained on raw audio or token sequences. However, the underlying mechanisms largely remain unexplored. Do these models generate music by simply relying on shallow contextual dependencies, or do they learn symbolic concepts, such as pitch and chord, similar to how the human mind processes music? To address this question, we conducted a pilot study to investigate and manipulate the hidden states of MERT and MusicGen, two state-of-the-art Transformer-based music LLMs. Experiments show that these models indeed acquire the concept of pitch and chord root, with a notable improvement in representational strength in deeper layers. Additionally, we see a strong preference for retaining pitch content over its stylistic counterpart, instrument timbre, and a similar relationship is observed between chord root note and chord quality. These observations offer valuable insights into the inner workings of music LLMs.

1 Introduction

In recent years, Transformer-based large language models (LLMs) have demonstrated remarkable capabilities in generating and understanding music [1, 2, 3, 4]. Despite their impressive achievements, the intrinsic mechanisms of music LLMs—how they represent music internally—remain largely unexplored. Do these models process music using only shallow contextual dependencies, or do they internalize *symbolic* musical concepts such as pitch and chord, akin to human cognitive processes? To a more extreme extent, are they bypassing human perception and learning different yet useful representations that are even unfathomable to humans?

To address this, we investigate and manipulate the hidden states of two state-of-the-art Transformer-based music LLMs, MERT [4] and MusicGen [3], and assess their reliance on conceptual understandings that resemble symbolic music notions that humans often rely on. We choose two Transformers working under different architecture to ensure the experiment results are more comprehensive and less biased. Specifically, we examine whether and where each model acquires the concepts of (1) pitch and timbre of single notes and (2) root and quality of chords. Rather than assessing the models' performance on more complex MIR tasks, we deliberately focused on these fundamental concepts to evaluate the models' capacity for abstract, hierarchical conceptualization. Among the musical concepts, pitch and chord root are the content, which we define to be the more fine-grained details and discrete music information; while timbre and chord quality are the style, which we define to be the more high-leveled and integrated, thus more abstract music information. The definition of content and style can also be understood in other art forms. Looking at the Starry Night by Van Gogh, one might comment that for a single brush stroke, the content is the color blue and the style is oil paint; while for the whole painting, the content is the color scheme of blue and yellow and style is post-impressionism. We propose to investigate if the music LLMs acquire the concepts of content and style differently.

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

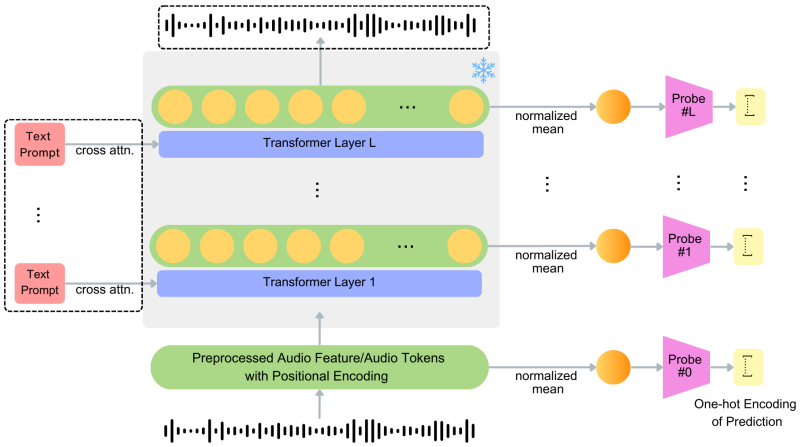

Figure 1: An illustration of probing method based on frozen LLMs. We trained probes starting from the layer 0 (input into the transformer), up until the last layer of the transformer (12 layers for MERT and 24, 48 for MusicGen-small and MusicGen-medium/melody). The graph applies to both the architectures of MERT and MusicGen in terms of probe training, except for the regions bounded by dashed-line boxes which only applies to MusicGen models.

Experiments suggest that these models acquire hierarchical concepts resembling those that humans might use, though the learning patterns between content and style differ significantly. Both MERT and MusicGen achieve high accuracy in recognizing pitch and chord root, with probing accuracy exceeding 0.9, and a clear improvement in representational strength in deeper layers. This indicates that Transformers are effective content learners, and symbolic concepts like pitch and chord root serve as valuable representations for audio prediction in these models.

Second, we see a strong correlation between the representational strength of pitch and chord root, which aligns well with the symbolic music notions in music theory — chord root is a more abstract symbol built on the concept of pitches.

Lastly, the representational strength of pitch is significantly stronger than that of timbre, with timbre probing accuracy dropping in deeper layers.

In summary, our contributions are:

- As far as we know, this is the first in-depth interpretability study of music LLMs using probing and intervention methodology.
- We observe that music LLMs learn pitch and chord root concepts, with representational strength improving in deeper layers. The correlation between pitch and chord root suggests a hierarchical symbolic representation of music.
- We see a strong preference for retaining pitch content over its stylistic counterpart, instrument timbre, and a similar relationship is observed between chord root note and chord quality.

2 Methodology

In this section, we elaborate on the probes targeting music concepts and intervention details.

2.1 Probing

Probing is an approach that uses features extracted from different layers to train classifiers for predicting the original classes. [5] The LLMs under our probing investigation are MusicGen (MusicGen-small, MusicGen-medium, MusicGen-melody) and MERT (MERT-v1-95M).

We use classifier probes to detect music concepts from the hidden states of large language models, focusing on *pitch*, *timbre*, *chord root*, and *chord quality*. For each music concept, we design two simple probes: a linear probe and a two-layer perceptron (MLP) probe.

Given the hidden state sequence $\{\boldsymbol{x}_1^l, \boldsymbol{x}_2^l, ..., \boldsymbol{x}_T^l\}$ for a T-frame audio feature input and layer index l, we firstly transform the sequence of hidden states into a global representation $\boldsymbol{h}^l$ using a normalization layer f: $\boldsymbol{h}^l = f(\sum_{t=1}^{T} \boldsymbol{x}_t^l / T) \in \mathbb{R}^d$.

For each concept, we train separate probes for hidden states across different model layers, and different concept c, as illustrated in Figure 1. The output activation $\hat{\boldsymbol{y}_c^l} \in \mathbb{R}^{n_c}$ of the probe $p_\theta^{l,c}$ (either linear or MLP) is given by:

$$\hat{\boldsymbol{y}_c^l} = \begin{cases} \boldsymbol{W}^{l,c}\boldsymbol{h}^l & \text{if } p_\theta^{l,c} \text{ is linear,} \\ \boldsymbol{W}_2^{l,c}\delta_1(\boldsymbol{W}_1^{l,c}\boldsymbol{h}^l) & \text{if } p_\theta^{l,c} \text{ is MLP.} \end{cases} \tag{1}$$

Here, l is the layer index, n_c is the classification category of concept c, and $\boldsymbol{W}^{l,c} \in \mathbb{R}^{n_c \times d}$, $\boldsymbol{W}_1^{l,c} \in \mathbb{R}^{512 \times d}$, and $\boldsymbol{W}_2^{l,c} \in \mathbb{R}^{n_c \times 512}$ are learnable matrices. δ_1 denotes the ReLU activation function. Let y_c be the ground truth for the given feature input. We optimize $p_{\theta,c}^l$ using cross-entropy loss: $\mathcal{L}_{l,c} = \mathcal{CE}(\delta_2(\hat{\boldsymbol{y}_c^l}), y_c)$,, where δ_2 is the softmax function.

Additionally, we train probes on randomly initialized networks for both models to serve as baselines. This allows us to assess the inherent difficulty of the tasks and better understand what the Transformer models are specifically learning.

2.2 Intervention

While probing is an effective way to test whether the hidden states are informative enough to *classify* the correct label, intervention techniques further examine the causal power of hidden states by testing whether intentionally changing the hidden states (along the gradient of the probing classifiers) would yield expected, altered outputs. Intuitively, intervention is a more difficult task compared to probing, as sometimes learning partial information is good enough for accurate classification, but *generating* an expected output requires a more complete representation.

Before the intervention, we pre-train all the probes and freeze their weights. Let α denotes the scaling factor for intervention. For a given music concept c and layer index l, we update the hidden state at t-th frame $\boldsymbol{x}_t^l$ to $\hat{\boldsymbol{x}}_t^l$ through gradient descent at each intervention step as follows:

$$\hat{\boldsymbol{x}}_t^l \leftarrow \boldsymbol{x}_t^l - \alpha \frac{\partial \mathcal{L}_{l,c}}{\partial \boldsymbol{x}_t^l}. \tag{2}$$

To assess the representational strength of each layer, we perform interventions by modifying only the hidden state of the targeted layer in each experiment. We halt intervention when $\mathcal{L}_{l,c}$ consistently achieves low levels and converges. Following intervention, we feed the modified hidden states into the MusicGen model to generate interpretable output predictions. Note that we only intervene MusicGen, as MERT does not directly generates audio.

3 Experiments

3.1 Pitch and timbre dataset

We used two datasets: NSynth [6] and a manually-synthesized dataset. NSynth contains over 300,000 4-second musical notes with annotations like pitch and instrument family. To manage its size, we used 25% of the training set while keeping the test set intact. For MERT and MusicGen-small, we utilized NSynth due to its 128 pitch classes and 11 timbre classes, making pitch classification more challenging than timbre, providing a clear test of our hypothesis that music LLMs learn content more ef- fectively than style.

For MusicGen-medium and MusicGen-melody, we used a smaller synthetic dataset of 2,257 1-second audio clips, each representing a single note across 37 soundfonts and covering pitches from C2 to C7. This was necessary to handle the computational load of these larger models, which have 48 layers.

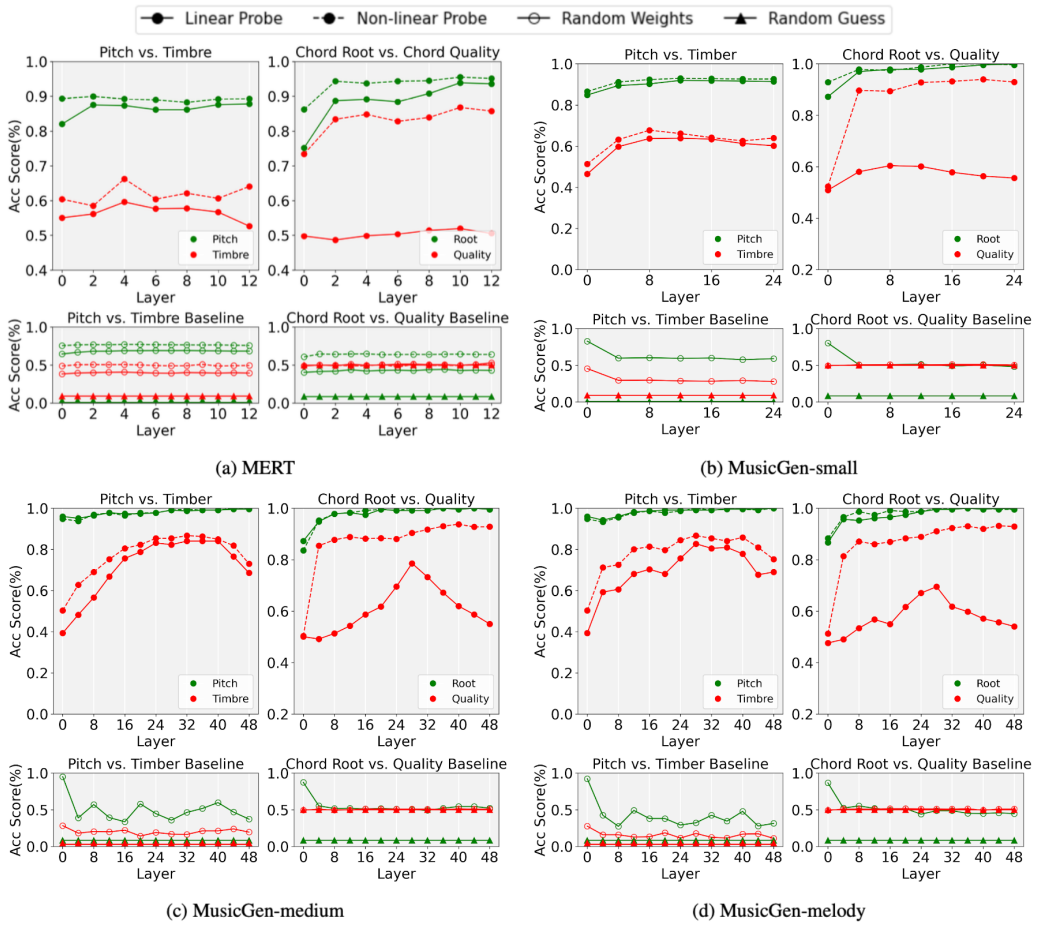

Figure 2: Accuracies of linear and non-linear probes on 4 tasks with (a) MERT, (b) MusicGen-small, (c) MusicGen-medium and (d) MusicGen-melody.

3.2 Chord dataset

We created a synthetic chord dataset with 13,542 audio clips, each featuring a chord with root notes from C2 to C7. Chords are either Major or Minor triads in root position or inversions. This dataset is used for Chord Recognition (ChordR) and Chord Quality (ChordQ) tasks.

Intervention experiments were conducted on MusicGen-small using 150 single-note clips and 300 chord clips from the synthetic datasets. The goal was to evaluate whether manipulating hidden states could shift the pitch or root note, tested using rule-based MIR algorithms.

4 Results and Analysis

4.1 Overall results

Figure 2 shows both linear and non-linear probing accuracy of pitch, timbre, chord root, and chord quality on four different models. Pitch and timbre are a pair of content and style, while chord root and quality can be regarded as another pair of content and style built on the concept of pitch. We also plot the two kinds of baselines: (1) random weights (in hollow circles), which means the probing accuracy is obtained when the LLM's weights are set to be random, and (2) random guess (in triangles).

Overall, we see that for all tasks and all models, the accuracy of both linear and non-linear probes is higher than the baselines. Based on these observations, our interpretation is that music LLMs indeed acquire pitch, timbre, chord root, and quality chord concepts.

4.2 Transformers are active content learner

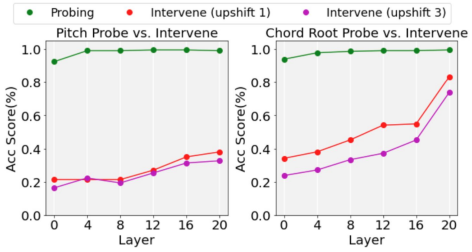

Figure 3: Probing and intervention results on MusicGen with 12 pitch classes and 12 chord root classes.

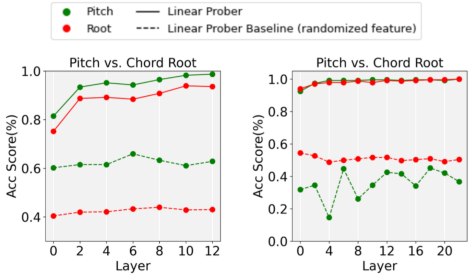

Figure 4: Comparison of linear probing performance on pitch and chord root tasks. The left panel shows results from MERT, while the right panel displays results from MusicGen-small.

As shown in Figure 2 , content-related tasks, such as pitch and chord root prediction, consistently outperform style-related tasks, like timbre and chord quality prediction. The accuracy for content-related tasks tends to be higher at corresponding transformer layers compared to style-related tasks. Additionally, the representational strength of content-related concepts increases steadily in deeper transformer layers, whereas style-related concepts show no such consistent growth.

Figure 3 further illustrates the intervention accuracy for pitch and chord root, both as 12-class classification tasks, alongside probing accuracy. Intervention accuracy also improves in deeper layers, likely because probing requires only partial information from shallower layers, while intervention demands more complete content extraction, which occurs in the deeper layers.

Thus, we conclude that music LLMs are active content learners, progressively extracting more complete symbolic music information in deeper layers.

4.3 Comparison between pitch and chord root

We can also compare the results horizontally, looking at how music LLMs learn the concept of content across different abstraction hierarchy. In our case, pitch in single note is the less abstract musical concept while chord root in chord is the more abstract musical concept. In Figure 4, we can see that MERT comprehends the less abstract content (pitch) with a better capacity while MusicGen-small performs equally good at absorbing musical content at different hierarchy, both in single note and in chord.

5 Conclusion

In our work, we applied probing and intervention method to Music LLMs. Our experiments showed that (1) Transformers are particularly effective at learning content-related concepts, especially in deeper layers, where they consistently outperform their ability to learn style-related concepts.(2) The consistent relationship between pitch and chord show that chord root is a more abstract concept derived from pitches. Several potential avenues for future research emerge from our work. A natural extension would be to expand our probing beyond individual notes and chords to include sequential concepts, such as musical scales. Another promising direction would be to investigate more complex concepts embedded within the models, such as the key of a song. Moreover, it will be interesting if we can monitor and intervene the continuation generated by music LLMs, it can further aid the music generation process and opens up intriguing avenues for refining these models to enhance their interpretive depth and accuracy.

References

[1] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever. Jukebox: A generative model for music, 2020.

[2] Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matt Sharifi, Neil Zeghidour, and Christian Frank. Musiclm: Generating music from text, 2023.

[3] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez. Simple and controllable music generation, 2024.

[4] Yizhi Li, Ruibin Yuan, Ge Zhang, Yinghao Ma, Xingran Chen, Hanzhi Yin, Chenghao Xiao, Chenghua Lin, Anton Ragni, Emmanouil Benetos, Norbert Gyenge, Roger Dannenberg, Ruibo Liu, Wenhu Chen, Gus Xia, Yemin Shi, Wenhao Huang, Zili Wang, Yike Guo, and Jie Fu. Mert: Acoustic music understanding model with large-scale self-supervised training, 2024.

[5] Guillaume Alain and Yoshua Bengio. Understanding intermediate layers using linear classifier probes, 2018.

[6] Jesse Engel, Cinjon Resnick, Adam Roberts, Sander Dieleman, Douglas Eck, Karen Simonyan, and Mohammad Norouzi. Neural audio synthesis of musical notes with wavenet autoencoders, 2017.

[7] Yonatan Belinkov. Probing classifiers: Promises, shortcomings, and advances, 2021.

[8] Ian Tenney, Dipanjan Das, and Ellie Pavlick. Bert rediscovers the classical nlp pipeline, 2019.

[9] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2023.

[10] Rodrigo Castellon, Chris Donahue, and Percy Liang. Codified audio language modeling learns useful representations for music information retrieval, 2021.

[11] Kenneth Li, Aspen K. Hopkins, David Bau, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Emergent world representations: Exploring a sequence model trained on a synthetic task, 2023.

[12] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.

[13] Wes Gurnee and Max Tegmark. Language models represent space and time, 2024.

Appendices

A Feature Selection

We use hidden states after every layer of the transformers as the probe features $\boldsymbol{x}_t^l$ where $t \in \mathbb{Z}^+$ is the timestep and the parameterization of l depends on the probed model. For MusicGen-small, $l \in [0..24]$ indexes the 24 decoder layers, with $l = 0$ representing the input embeddings derived from discrete EnCodec tokens before entering the MusicGen transformer decoder. Similarly, for MusicGen-melody and MusicGen-medium, $l \in [0..48]$ indexes the 48 decoder layers, where $l = 0$ also denotes the input embeddings from EnCodec tokens. For MERT, $l \in [0..12]$, with $l \in [1..12]$ indexing the 12 encoder layers and $l = 0$ representing the input embeddings before the MERT encoder, following the 1D-convolution feature extractor. For all models, when using transformer features, we use the hidden states after the layer normalization of the feed-forward sub-layer as the probe features. Those selected features then serve as the input to our learnable probes.

B Related Works

We review two realms of related works: 1) music large language models (LLMs), especially audio-based LLMs, and 2) existing works on understanding LLMs of other domains.

B.1 Music LLMs

The introduction of Jukebox [[1] by OpenAI laid the foundation for audio-based LLMs, showcasing the ability to model complex musical structures from raw audio. Its hierarchical Transformer architecture, which is facilitated with lyrics condition, highlighted the model's adeptness at discerning complex audio patterns. Building on this, MusicLM [2] improved audio generation with text conditioning, enhancing control and diversity. More recently, MERT [4] and MusicGen [3] achieved state-of-the-art results in music understanding and generation, respectively, using single-Transformer architectures. In this study, we explore the interpretability of these models, focusing on MERT and MusicGen.

B.2 Probing and Intervene

Probing and intervention techniques have become essential for understanding the internal workings of LLMs and gaining finer control over their outputs [5, 7, 8]. Probing analyzes the representations learned across model layers to uncover how information is processed, while intervention involves manipulating internal states to assess their causal influence on decision-making [9]. For example, JukeMIR [10] probed Jukebox and found that the middle layers' hidden states were the most effective for downstream tasks. Similarly, a study on a GPT variant fine-tuned for Othello revealed that the model developed non-linear internal representations of board states without explicitly learning the game's rules [11]. Investigations into models like Llama-2 [12] further explore whether LLMs grasp broader world concepts, such as space and time, beyond merely memorizing data [13].

C Limitation

Our experimental conclusions are specifically limited to the pitch, timbre, chord root, and chord quality concepts explored in this study. Additionally, the datasets used were restricted to NSynth and our own synthesized data. As such, we cannot guarantee that our findings will generalize to other musical concepts or datasets, or to settings outside those tested here. Future work is needed to evaluate the applicability of these conclusions in a broader range of musical contexts and with diverse data sources.