MITSUBISHI ELECTRIC RESEARCH LABORATORIES
https://www.merl.com

UNDERSTANDING AND CONTROLLING GENERATIVE MUSIC TRANSFORMERS BY PROBING INDIVIDUAL ATTENTION HEADS

Koo, Junghyun; Wichern, Gordon; Germain, François G; Khurana, Sameer; Le Roux, Jonathan

TR2024-032 March 21, 2024

Abstract

This paper presents an in-depth analysis of MusicGen, a generative music transformer model, focusing on the capabilities of its self-attention heads in understanding and representing diverse musical elements. We uncover how MusicGen encodes various aspects of music through head-wise probing, from instrument recognition to more complex downstream tasks. Our findings reveal that certain attention heads are particularly adept at discerning specific musical characteristics, suggesting a pathway to highly nuanced music generation. By leveraging techniques for inference-time control, originally developed for large language models, we discuss the potential for achieving additional precise control in text-to-music generation tasks. This approach allows for fine-grained customization beyond basic text prompts, facilitating music generation that more accurately reflects the user's creative intent.

IEEE ICASSP Satellite Workshop on Explainable Machine Learning for Speech and Audio (XAI-SA) 2024

Mitsubishi Electric Research Laboratories, Inc.
201 Broadway, Cambridge, Massachusetts 02139

UNDERSTANDING AND CONTROLLING GENERATIVE MUSIC TRANSFORMERS BY PROBING INDIVIDUAL ATTENTION HEADS

Junghyun Koo[1,2]*, Gordon Wichern*[1]*, François G. Germain*[1]*, Sameer Khurana*[1]*, Jonathan Le Roux*[1]

[1]Mitsubishi Electric Research Laboratories (MERL), Cambridge, MA, USA
[2]Department of Intelligence and Information, Seoul National University, Seoul, South Korea

ABSTRACT

This paper presents an in-depth analysis of MusicGen, a generative music transformer model, focusing on the capabilities of its self-attention heads in understanding and representing diverse musical elements. We uncover how MusicGen encodes various aspects of music through head-wise probing, from instrument recognition to more complex downstream tasks. Our findings reveal that certain attention heads are particularly adept at discerning specific musical characteristics, suggesting a pathway to highly nuanced music generation. By leveraging techniques for inference-time control, originally developed for large language models, we discuss the potential for achieving additional precise control in text-to-music generation tasks. This approach allows for fine-grained customization beyond basic text prompts, facilitating music generation that more accurately reflects the user's creative intent.

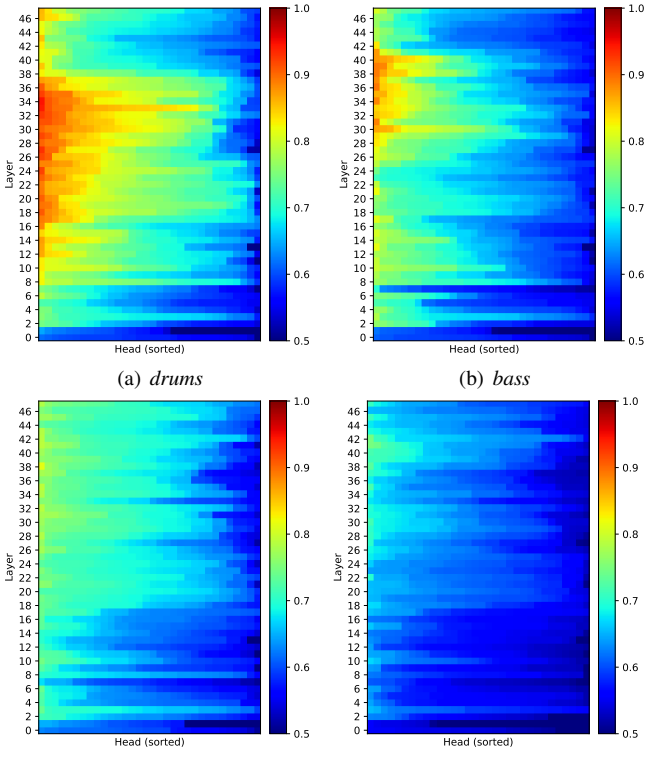
Index Terms— Generative music transformer, classifier probes, multi-head attention, music information retrieval

1. INTRODUCTION

A standard approach for evaluating the type of information captured by a given deep-learning representation is classifier probing [1–3], where simple classifiers (e.g., linear or very small multi-layer networks) are trained on frozen deep representations. If the combination of deep representation and simple classifier performs well for a given classification task, then we have some evidence that the deep representation (and by association the network used to create that representation) has learned something about the classification task. In the audio domain, both the HEAR [4] and SUPERB [5] challenges provide a large set of classification tasks for evaluating learned representations.

In audio as in other domains, the transformer has become the dominant network architecture. Surprisingly, it was shown in [6] that Whisper, a large speech recognition model trained in noisy conditions, is also a strong audio tagger, indicating that the model learned to classify non-speech sounds as a byproduct of the noisy-condition training. In the music domain, training classifier probes on the output of Jukebox, a generative text-to-music transformer, lead to strong performance on a wide range of music information retrieval tasks including tagging, genre classification, emotion recognition, and key detection [7]. However, these works and related ones in the speech domain (e.g., [8–10]) operate on entire transformer layers.

Since a transformer layer's key component is the multi-head self-attention process, where each attention head in a layer operates in parallel, we analyze in this work whether individual attention heads of a generative music transformer model have learned useful representations for downstream classification tasks. Specifically, we analyze MusicGen [11], a recent autoregressive generative music transformer, in terms of a multi-instrument classification task and the music tagging, genre classification, emotion recognition, and key detection tasks from [7]. To the best of our knowledge, this is the first work evaluating downstream music information retrieval task performance on the outputs of individual attention heads.



(a) *drums* (b) *bass*

(c) *guitar* (d) *piano*

Fig. 1: Instrument recognition performance of individual attention head probes from the $\text{MusicGen}_{\text{large}}$ model activations. All colorbars are normalized to the same range.

In addition to providing a more fine-grained understanding of learned transformer representations, individual attention head probes may also be useful for inference-time control of pre-trained transformers. In the case of language models, it was recently shown in [12] that individual attention heads are useful for classifying the accuracy or truthfulness of text, and "steering" the output of only those attention heads sensitive to truthfulness can lead to a more truthful language model, while maintaining overall helpfulness. In this work, in addition to analyzing attention head probe accuracy on music tasks, we also speculate on their use for fine-grained inference-time control.

This work was performed while J. Koo was an intern at MERL.

Table 1: Comparison of multi-label instrument recognition performance for head-wise probing on MusicGen and a supervised model.

Model	Accuracy / F1 Score					Num. Param.
	vocals	*bass*	*drums*	*other*	*avg.*	
$\text{ConvNeXt}_{\text{tiny}}$ [13]	**97.8% / 0.947**	**94.4%** / 0.891	95.1% / 0.914	**93.2% / 0.880**	**95.1%** / 0.906	28.5M
$\text{MusicGen}_{\text{small}}$	92.5% / 0.929	91.6% / 0.920	95.0% / 0.952	87.1% / 0.868	91.5% / 0.917	300M
$\text{MusicGen}_{\text{medium}}$	92.9% / 0.934	91.8% / **0.923**	94.4% / 0.947	87.4% / 0.874	91.6% / 0.919	1.5B
$\text{MusicGen}_{\text{large}}$	92.8% / 0.932	91.7% / 0.920	95.1% / 0.953	85.7% / 0.872	91.8% / 0.919	3.3B
$\text{MusicGen}_{\text{melody}}$	94.1% / 0.945	91.8% / **0.923**	**95.8% / 0.960**	87.9% / 0.876	92.4% / **0.926**	1.5B

2. UNDERSTANDING MUSICGEN

In this section, we seek to investigate and quantify the comprehension of music by each attention head within MusicGen [11]. MusicGen is a pre-trained and publicly accessible generative music transformer that uses EnCodec [14] to create a discrete audio token, an autoregressive transformer to predict the next token, and an EnCodec decoder to output an audio signal. This analysis will provide insights into the model's potential for fine-grained control via attention head steering. Our methodology begins with a probing task designed to assess the model's ability to distinguish musical pieces based on the presence or absence of specific instruments, before broadening our analysis to multiple downstream tasks covering various musical attributes.

2.1. Probing MusicGen

We describe the methodology of probing MusicGen by evaluating the capability of its self-attention heads in recognizing instruments (i.e., determining whether a target instrument is present in the audio stream). We create a dataset by curating data from MusDB [15] and MoisesDB [16], which offer multi-track recordings with isolated instrument stems. For a given target stem, we form a positive class of mixtures where the target stem is present, and a negative class of corresponding mixtures with the target stem removed as follows. First, we remove from every multi-track recording the time regions where the target stem is silent. Then, out of this pruned recording, the mixture of all of its stems is added to the positive class data, while the mixture of all of its stems except its target stem is added to the negative class data. Subsequently, we process 3-second-long segments of these tracks for training (and testing), passing them through MusicGen to extract the intermediate activation $z_{l,h}(T)$ at the last time step T for every self-attention layer l and head h. At time step t, the l-th self-attention layer computes H self-attention heads $z_{l,h}(t) \in \mathbb{R}^D$ from an input vector $x_l(t) \in \mathbb{R}^{DH}$ as

$$z_{l,h}(t) = \text{Att}(W^Q_{l,h} x_l(t), W^K_{l,h} x_l(1\!:\!t), W^V_{l,h} x_l(1\!:\!t)), \quad (1)$$

where $x_l(1\!:\!t) = [x_l(1), \ldots, x_l(t)]$, $W^Q_{l,h}$, $W^K_{l,h}$, and $W^Q_{l,h}$ denote the head-specific query, key, and value projection matrices, all in $\mathbb{R}^{D \times DH}$, and Att denotes the attention operator [17]. This forms the basis for the training (and testing) sets of the probe classifier, wherein a simple logistic regression model is employed to distinguish the presence of the instrument.

The testing accuracy of probes from $\text{MusicGen}_{\text{large}}$ (3.3B parameters) across all self-attention layers l and heads h is illustrated in Fig. 1. We observe that specific subsets of heads outperform others in detecting the presence of each target stem. While certain attention layers show better performance, it is notable that not all heads within each layer result in uniform performance; rather, their effectiveness varies considerably. This variation underscores the utility of head-wise probing in achieving precise control over the transformer's behavior. Furthermore, MusicGen's proficiency varies across instruments; it demonstrates a strong understanding of drums and bass, whereas its accuracy on guitar and piano is comparatively lower. This discrepancy suggests a potential bias in MusicGen's training dataset.

2.2. Analytical Insights from Probing

Following the probing method outlined in Section 2.1, we investigate additional downstream tasks through probing to present more objective results and enhance our understanding. First, we assess MusicGen's probing capabilities by comparing them to ConvNeXt [13], a model trained via supervised learning. Subsequently, we contrast our results with prior work on music probing [7] which uses entire attention layer outputs from Jukebox, another generative music transformer model.

2.2.1. Multi-label Instrument Recognition

Following up on the performance of single-instrument recognition depicted in Fig. 1, we further explore MusicGen's capabilities in the multi-label instrument recognition task: identifying the presence of each instrument of a given music. This is achieved by fitting multiple logistic regression classifiers (probes) for each instrument class using again the extracted intermediate activations at the last time step. For the final evaluation, we select the best-performing probes for each instrument class, noting that these may originate from different attention layers (l) and heads (h). Following the methodology described in [18], we conduct recognition tasks on 3-second music segments from the MusDB dataset [15]. For comparative analysis, we also report the performance of a model trained in a supervised manner specifically for instrument recognition [18], utilizing the 'tiny' configuration of ConvNeXt [13].

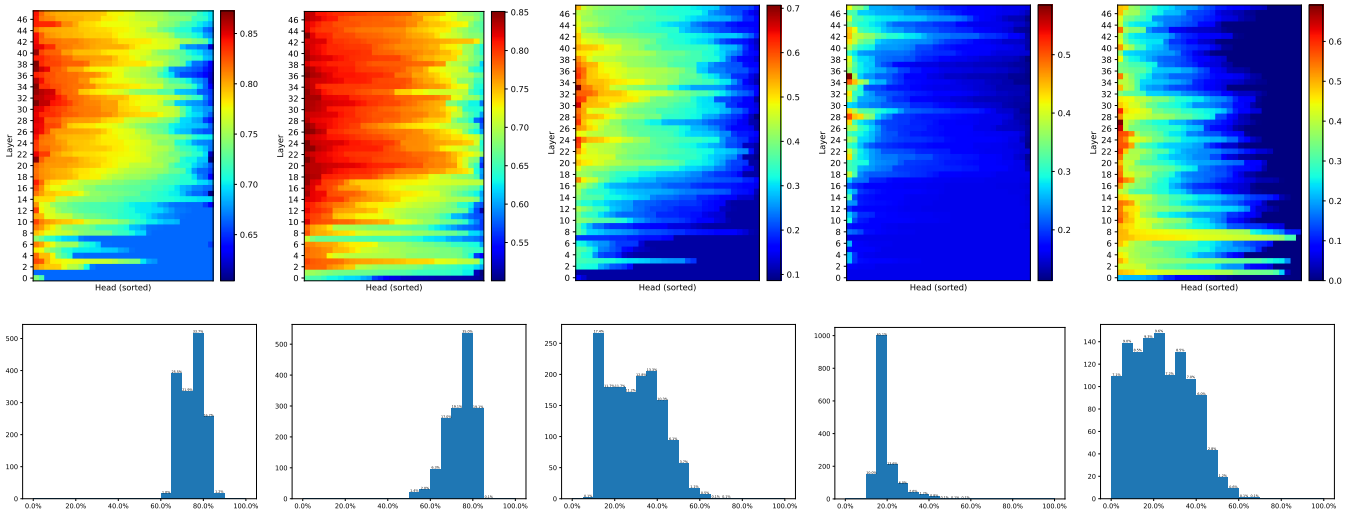
The objective results for multi-label instrument recognition are detailed in Table 1. We find that the performance of MusicGen's probes is comparable to that of the supervised method.However, across all configurations of MusicGen, we note strong performance in terms of the F1 score.Specifically, MusicGen exhibits exceptional performance in recognizing *drums*, whereas its least effective performance is observed in the *other* category. This discrepancy likely arises from the ambiguous definition of the *other* category in the MusDB18 dataset, with the supervised method gaining an advantage by explicitly training on the dataset's label, thereby enhancing its ability to identify *other*.

2.2.2. Various Music Downstream Tasks

To expand our understanding of MusicGen's capabilities beyond instrument recognition, we conducted probing on a set of more general music understanding tasks. These include music tagging (MTT) [19], genre classification (GTZAN) [20], key detection (GS) [21], and

Table 2: Performance comparison of layer-wise and head-wise probing on generative music transformers. For music tagging, we report the performance of the best-performing probe across all classes, with the ensemble result of the best-performing probes for each class indicated within parentheses.

Dataset	**MTT**		**GTZAN**	**GS**	**EMO**				
Task	Tagging		Genre	Key	Emotion				
Metrics	AUC	AP	Acc	$\text{Acc}^{\text{ref.}}$	R2^{A}	R2^{V}	Avg.	**Dim.**	**Num. Param.**
Jukebox [7]	91.5	41.4	79.7	66.7	72.1	61.7	69.9	4.8K	5B
$\text{MusicGen}_{\text{small}}$	85.5 (86.7)	34.1 (37.5)	66.2	46.6	64.2	43.5	55.2	64	300M
$\text{MusicGen}_{\text{medium}}$	85.9 (87.3)	33.9 (38.4)	69.7	57.4	65.3	51.6	59.6	64	1.5B
$\text{MusicGen}_{\text{large}}$	85.1 (87.2)	32.9 (38.5)	71.0	58.5	69.1	49.3	60.3	64	3.3B
$\text{MusicGen}_{\text{melody}}$	85.8 (87.1)	33.3 (38.1)	65.2	62.1	64.7	44.8	58.8	64	1.5B



(a) Multi-label Inst. (F1) (b) Tagging (AUC) (c) Genre (Acc) (d) Key ($\text{Acc}^{\text{ref.}}$) (e) Emotion (R2^{A})

Fig. 2: Probe performance on various music understanding tasks. Each set of figures shows the performance on each task with probes fitted with $\text{MusicGen}_{\text{large}}$ model output activations. Histograms in the bottom row show the distribution of head-wise probes according to their performance.

emotion recognition (EMO) [22]. We adopt appropriate regression models for probing for these tasks: multinomial logistic regression for multi-class classification tasks, multiple logistic regressions for multi-label classification tasks, and linear regression for regression tasks. For these tasks, we set the input duration as 29.0 seconds, and let the probes take the last time step activation output for classification.

Table 2 presents the probing results for various configurations of MusicGen using our attention head probing, and a prior music probing work [7] that relies on a comparative model, Jukebox [23], but uses a different probing methodology:

1. [7] probed Jukebox using entire intermediate attention layers, whereas our approach evaluates performance based on individual attention heads.
2. [7] employed a one-layer MLP with 512 hidden units. In preliminary experiments, using a one-layer MLP for head-wise probes showed limited benefits, possibly because the attention dimension per head that we use is 75 times smaller than that of Jukebox.

Jukebox probing typically shows superior performance, maybe due to the model's larger number of parameters and activation dimension. In music tagging, we present both the performance of the best-performing probe across all classes and the ensemble result of the best-performing probes for each class, indicated within parentheses. We observe an improved performance with the ensemble technique, though the improvement is typically not substantial. Interestingly, this means that we can find a single head capable of understanding multiple attributes simultaneously, a notable feature given the MTT dataset comprises 188 different tags. While a larger number of parameters generally correlates with enhanced performance, the melody configuration outperforms others in key detection. This superior performance is likely attributed to the model's training with a melody conditioning objective [11].

Figure 2 provides a qualitative visualization of the performance of $\text{MusicGen}_{\text{large}}$ probes across various music understanding tasks, alongside histograms representing the distribution of head-wise performance. In tasks such as multi-label instrument recognition and music tagging, there is a noticeable similarity in the trends observed both in the histograms and the layer-wise performance distributions. For key detection, the results indicate that comprehension is limited to only a few heads. On the other hand, emotion recognition demonstrates a relatively uniform performance across the majority of heads in the earlier layers, particularly layer 7. However, the highest-performing head for emotion recognition is located in a middle layer, specifically at the 27^{th} layer. This analysis underscores the nuanced role that different attention heads play in music understanding, highlighting the potential for precise control on each attention head to optimize performance for specific tasks within generative music models.

3. POSSIBLE APPLICATIONS OF HEAD-WISE PROBING

We demonstrated that individual self-attention heads within the pre-trained MusicGen model have indeed learned to capture distinctive musical characteristics. In this section, we discuss the potential for fine-grained control in text-to-music tasks, leveraging our insights to enable more nuanced and targeted music generation.

Leveraging the classifier probes methodology, we have evidence that self-attention heads within MusicGen are capable of encoding musical aspects that users may wish to manipulate. Given the architectural similarities between this class of text-to-music models and LLMs, it is compelling to investigate whether techniques for inference-time control of LLMs—e.g., those aimed at gaining controllability via latent activation manipulation as outlined in [24–26]—could be adapted for audio applications.

A notable study by Li et al. [12] sought to improve the "truthfulness" of text language models. They used classifier probes to identify attention heads that captured the concept of "truthfulness," allowing for precise, targeted modifications to the model's truth-related components. This resulted in a language model with improved performance on the TruthQA [27] benchmark. Inspired by this approach, we could apply similar techniques to MusicGen to build a system that is not only capable of coherent music generation but also offers additional layers of control.

Such applications would offer precise control over the generative process, allowing for customizations beyond textual descriptions and addressing scenarios where text-to-music models may struggle, such as generating classical music with the unconventional inclusion of heavy metal drums. This advanced level of control could allow for more nuanced adherence to a user's creative intent, even when it deviates from standard genre conventions, thereby expanding the boundaries of expressive and adaptable text-to-music generation.

4. CONCLUSION

Our comprehensive exploration of MusicGen has revealed its individual self-attention heads' ability to discern and represent various musical characteristics. The probing method applied to these attention heads not only validated the model's capability in instrument recognition but also demonstrated its proficiency in other complex downstream music informatics tasks. The insights gained suggest that adopting inference-time control techniques could achieve a higher degree of precision in music generation than what is possible with coarse text prompts alone. This would enable intricate customization that adhere more closely to the user's creative vision.

5. REFERENCES

[1] Guillaume Alain and Yoshua Bengio, "Understanding intermediate layers using linear classifier probes," *arXiv preprint arXiv:1610.01644*, 2016.

[2] Yonatan Belinkov, "Probing classifiers: Promises, shortcomings, and advances," *Comput. Linguist.*, vol. 48, no. 1, pp. 207–219, 2022.

[3] Ian Tenney, Dipanjan Das, and Ellie Pavlick, "BERT rediscovers the classical NLP pipeline," in *Proc. ACL*, 2019, pp. 4593–4601.

[4] Joseph Turian, Jordie Shier, Humair Raj Khan, Bhiksha Raj, Björn W. Schuller, Christian J. Steinmetz, Colin Malloy, George Tzanetakis, Gissel Velarde, Kirk McNally, et al., "HEAR: Holistic evaluation of audio representations," in *Proc. NeurIPS Compet. Demonstrations Track*, 2022, pp. 125–145.

[5] Shu-wen Yang, Po-Han Chi, Yung-Sung Chuang, Cheng-I Jeff Lai, Kushal Lakhotia, Yist Y. Lin, Andy T. Liu, Jiatong Shi, Xuankai Chang, Guan-Ting Lin, et al., "SUPERB: Speech processing Universal PERformance Benchmark," *arXiv preprint arXiv:2105.01051*, 2021.

[6] Yuan Gong, Sameer Khurana, Leonid Karlinsky, and James Glass, "Whisper-AT: Noise-robust automatic speech recognizers are also strong general audio event taggers," in *Proc. Interspeech*, 2023, pp. 2798–2802.

[7] Rodrigo Castellon, Chris Donahue, and Percy Liang, "Codified audio language modeling learns useful representations for music information retrieval," in *Proc. ISMIR*, 2021, pp. 88–96.

[8] Puyuan Peng and David Harwath, "Self-supervised representation learning for speech using visual grounding and masked language modeling," in *Proc. AAAI SAS Workshop*, 2022.

[9] Ankita Pasad, Chung-Ming Chien, Shane Settle, and Karen Livescu, "What do self-supervised speech models know about words?," *arXiv preprint arXiv:2307.00162*, 2023.

[10] Bastiaan Tamm, Rik Vandenberghe, and Hugo Van hamme, "Analysis of XLS-R for speech quality assessment," in *Proc. WASPAA*, 2023.

[11] Jade Copet, Felix Kreuk, Itai Gat, Tal Remez, David Kant, Gabriel Synnaeve, Yossi Adi, and Alexandre Défossez, "Simple and controllable music generation," in *Proc. NeurIPS*, 2023, pp. 47704–47720.

[12] Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg, "Inference-time intervention: Eliciting truthful answers from a language model," in *Proc. NeurIPS*, 2023, pp. 41451–41530.

[13] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie, "A ConvNet for the 2020s," in *Proc. CVPR*, 2022, pp. 11976–11986.

[14] Alexandre Défossez, Jade Copet, Gabriel Synnaeve, and Yossi Adi, "High fidelity neural audio compression," *arXiv preprint arXiv:2210.13438*, 2022.

[15] Zafar Rafii, Antoine Liutkus, Fabian-Robert Stöter, Stylianos Ioannis Mimilakis, and Rachel Bittner, "MUSDB18-HQ - an uncompressed version of MUSDB18," 2019.

[16] Igor Pereira, Felipe Araújo, Filip Korzeniowski, and Richard Vogl, "MoisesDB: A dataset for source separation beyond 4-stems," in *Proc. ISMIR*, 2023, pp. 619–626.

[17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin, "Attention is all you need," in *Proc. NeurIPS*, 2017, pp. 6000–6010.

[18] Junghyun Koo, Yunkee Chae, Chang-Bin Jeon, and Kyogu Lee, "Self-refining of pseudo labels for music source separation with noisy labeled data," in *Proc. ISMIR*, 2023, pp. 716–724.

[19] Edith Law, Kris West, Michael I. Mandel, Mert Bay, and J. Stephen Downie, "Evaluation of algorithms using games: The case of music tagging.," in *Proc. ISMIR*, 2009, pp. 387–392.

[20] George Tzanetakis and Perry Cook, "Musical genre classification of audio signals," *IEEE Trans. Speech Audio Process.*, vol. 10, no. 5, pp. 293–302, 2002.

[21] Peter Knees, Ángel Faraldo Pérez, Herrera Boyer, Richard Vogl, Sebastian Böck, Florian Hörschläger, Mickael Le Goff, et al., "Two data sets for tempo estimation and key detection in electronic dance music annotated from user corrections," in *Proc. ISMIR*, 2015, pp. 364–370.

[22] Mohammad Soleymani, Micheal N. Caro, Erik M. Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang, "1000 songs for emotional analysis of music," in *Proc. ACM Workshop Crowdsourcing Multimed. (CrowdMM)*, 2013, pp. 1–6.

[23] Prafulla Dhariwal, Heewoo Jun, Christine Payne, Jong Wook Kim, Alec Radford, and Ilya Sutskever, "Jukebox: A generative model for music," *arXiv preprint arXiv:2005.00341*, 2020.

[24] Nishant Subramani, Nivedita Suresh, and Matthew E Peters, "Extracting latent steering vectors from pretrained language models," in *Proc. ACL*, 2022, pp. 566–581.

[25] Alex Turner, Lisa Thiergart, David Udell, Gavin Leech, Ulisse Mini, and Monte MacDiarmid, "Activation addition: Steering language models without optimization," *arXiv preprint arXiv:2308.10248*, 2023.

[26] Evan Hernandez, Belinda Z. Li, and Jacob Andreas, "Inspecting and editing knowledge representations in language models," *arXiv preprint arXiv:2304.00740*, 2023.

[27] Stephanie Lin, Jacob Hilton, and Owain Evans, "TruthfulQA: Measuring how models mimic human falsehoods," in *Proc. ACL*, 2022, pp. 3214–3252.