Evaluation of pretrained language models on music understanding

Yannis Vasilakis
Queen Mary University of London
i.vasilakis@qmul.ac.uk

Rachel Bittner
Spotify
rachelbittner@spotify.com

Johan Pauwels
Queen Mary University of London
j.pauwels@qmul.ac.uk

Abstract

Music-text multimodal systems have enabled new approaches to Music Information Research (MIR) applications such as audio-to-text and text-to-audio retrieval, text-based song generation, and music captioning. Despite the reported success, little effort has been put into evaluating the musical knowledge of Large Language Models (LLM). In this paper, we demonstrate that LLMs suffer from 1) prompt sensitivity, 2) inability to model negation (e.g. "rock song without guitar"), and 3) sensitivity towards the presence of specific words. We quantified these properties as a triplet-based accuracy, evaluating the ability to model the relative similarity of labels in a hierarchical ontology. We leveraged the Audioset ontology to generate triplets consisting of an anchor, a positive (relevant) label, and a negative (less relevant) label for the genre and instruments sub-tree. We evaluated the triplet-based musical knowledge for six general-purpose Transformer-based models. The triplets obtained through this methodology required filtering, as some were difficult to judge and therefore relatively uninformative for evaluation purposes. Despite the relatively high accuracy reported, inconsistencies are evident in all six models, suggesting that off-the-shelf LLMs need adaptation to music before use.

1 Introduction

The capability of Large Language Models (LLM) to obtain informative context-dependent word embeddings with long-range inter-token dependencies showed that they can be used effectively to encode knowledge from several domains without manually curating datasets.

During the last 5 years, the scientific community combined audio-based Deep Neural Networks (DNN) with LLMs to form audio-text models, leading to improved performance on several music applications such as audio-to-text retrieval and text-to-audio retrieval (Huang et al., 2022; Manco et al., 2022; Wu et al., 2023), music captioning (Gardner et al., 2024; Manco et al., 2021) and text-based song generation (Yu et al., 2022).

LLMs are usually used pretrained and off-the-shelf (Manco et al., 2022; Huang et al., 2022). While datasets for semantic similarity of general language (Ojha et al., 2024) are available, we are not aware of any such datasets for music. Therefore, LLMs haven't been thoroughly evaluated on their musical knowledge and potential issues might be obscured.

In this paper, we quantify musical knowledge in LLMs using triplets obtained through an ontology and report three shortcomings when used off-the-shelf. We leverage Audioset, a hierarchical ontology, to extract the triplets of (anchor, positive, negative) format. The anchor label is chosen arbitrarily from the ontology, a similar label is selected as the positive, and a relatively less similar label as the negative term of the triplet. We quantify the relative similarity using the ontology-based distance between pairs of labels. Thus, we evaluate LLM's musical knowledge by comparing the relative similarity between anchor-positive and anchor-negative labels. We collected 13633 Music Genre and 37640 Music Instrument triplets. We evaluated the sensitivity of LLMs to 20 different musically informed prompts and their inability to model negation. Finally, we report performance improvements when both labels and their definitions are used.

Both code snippets and sets of triplets used are made publicly available for reproducibility reasons [1].

2 Related Work

2.1 BERT

Bidirectional Encoder Representations from Transformers (BERT) (Devlin et al., 2019; Sun et al.,

[1] https://github.com/YannisBilly/Evaluation-of-pretrained-language-models-on-music-understanding

2022) is the backbone for many Natural Language Processing (NLP) applications such as translation (Xu et al., 2021), text summarization (Liu and Lapata, 2019), and others. These systems were trained with unstructured large corpora through masked word and next-sentence prediction, without the need for curated datasets.

BERT provides a context-dependent token-based embedding vector but doesn't calculate independent sentence embeddings. This means that sentence embeddings need to be calculated as a function of the token embeddings at inference time. Obtaining the latter is not straightforward (Choi et al., 2021; Alian and Awajan, 2020) and several different approaches have been proposed. The most frequent, better-performing method is averaging the token embeddings in different layer depths. Another one is using the [CLS] token, obtaining subpar performance (Li et al., 2020). We focus on the first approach as the most prominent but highlight that calculating sentence embeddings is still an active research topic (Xu et al., 2024; Amur et al., 2023).

2.2 Large Language Models in Music Information Research

Transformer-based models have been introduced in several applications. Zero-shot classification utilizes word embeddings to infer a classifier on unseen classes based on the similarity of the new class label with the labels of the known classes (Du et al., 2024). Audio-to-text and text-to-audio retrieval is successful in aligning audio and text embeddings using music/caption pairs (Manco et al., 2022; Huang et al., 2022). Automatic music caption uses music embeddings to condition an LLM (Manco et al., 2021; Gardner et al., 2024) to generate music descriptions. Lastly, sentence similarity has been used to weigh intra-caption similarity in contrastive loss functions (Manco et al., 2022; van den Oord et al., 2018).

3 Evaluation of language models on musical knowledge

As far as we are concerned, a linguistic evaluation dataset of musical knowledge doesn't exist apart from language-based artist similarity (Oramas et al., 2018, 2015).

Information used for semantic similarity is usually scraped from websites and we argue that this information is not directly useable. Generally, these websites highlight the history of the queried label without juxtaposing related concepts, audio attributes or providing slang labels and abbreviations. Also, their massive size can hinder inspection and therefore, reduce their value as evaluation sets.

We argue that an evaluation dataset needs to be cleaned and inspected thoroughly before increasing its size. This hasn't been done in captioning and tagging datasets, as most are weakly annotated and have highly noisy annotations (Choi et al., 2018).

Therefore, we chose to utilize an ontology with less than 200 musical labels which have a manageable size, can be manually inspected and filtered. However, we need to acknowledge that most existing ontologies are far from being exhaustive. We drew inspiration from the Semantic Textual Similarity task (Ojha et al., 2024; Dong et al., 2021; Wahle et al., 2022) that contains pairs of sentences and their degree of similarity but proposed a method of obtaining such sentences automatically leveraging a taxonomy.

We evaluated 6 general-purpose Transformer-based models (Reimers and Gurevych, 2019) for sentence similarity using musical terminology. In detail, a global average pooling layer is appended on top of the final layer and the sentence embedding is calculated as the mean of the respective token embeddings. The models used are *MPNet, Distil-RoBERTa, MiniLM* and *ALBERT* trained on different corpora. More information about the models is provided in appendix section B and tables 1, 2.

3.1 Audioset and its ontology

Large-scale annotated datasets have been essential for Computer Vision. Drawing inspiration from this, Audioset (Gemmeke et al., 2017) was proposed which has $\approx$ 1.79 million 10-second long audio snippets scraped from YouTube, annotated with a hierarchical ontology of 632 audio classes.

The creation of their taxonomy focused on two properties: (1) labels must be recognizable by typical listeners without additional information beyond the label, and (2) the taxonomy must be comprehensive enough to describe most real-world recordings adequately. After finalizing the taxonomy, annotators were given a 10-second audio clip and a label. They had to choose from "present", "not present", or "unsure" to indicate whether the audio and label used were positively, negatively, or uncertainly related, respectively.

In this paper, we use the Audioset sub-tree

of Music [2]. Due to the unitary depth of most child nodes (e.g. Music mood), we will only include the sub-trees of "Musical Instrument" and "Music Genre". A deficiency of using a tree is that inter-category relations cannot be modeled (e.g. "Rock music" and "Guitar"). The triplet-based evaluation methodology can be extended to other graph structures and elaborate ontologies (e.g. WordNet (Miller, 1995)), as well as include intra-category relations (e.g. "Rock music', "Electric guitar", "Viola").

The "Musical Instrument" taxonomy has a maximum depth of 4, encompassing most instrument families, including classical, modern, and non-western instruments. Although it does not separate playing techniques from instruments (e.g., "electric guitar" and "tapping"), omits some instruments (e.g., "viola" from "bowed string instruments") and contains vague concepts ("Musical ensemble"), the taxonomy remains well-defined and free of ambiguous labels.

The "Musical genre" taxonomy has a maximum depth of 3, covering Western music with detailed categorization of contemporary genres (e.g., "Grime music"), as well as folk and non-Western genres. However, it lacks nuance in classical music, only including opera.

3.2 Triplet-based musical knowledge quantification

To curate the music knowledge corpus for LLM evaluation, we leverage the aforementioned subtrees of the Audioset ontology and generate triplets. Specifically, we form triplets of an anchor, a positive and a negative label. The positive and negative labels are defined relative to their semantic similarity with respect to the anchor label. If the anchor is more similar to label 1 than label 2, label 1 is the positive and label 2 is the negative label. This method can encode abstract relationships between labels, including comparisons between non-homogeneous labels (e.g., "happy music", "rock music", "reggae music") but is left for future work as it requires more elaborate ontologies.

We use the distance between the labels based on each tree to quantify their relative similarity. A valid triplet is defined as one where the anchor-positive is less than the anchor-negative distance. After obtaining the valid triplets, we manually inspect them and remove the ones that are ambiguous, vague or too difficult to judge[3].

Finally, we are left with 13633 Genre triplets and 37640 Instrument triplets that will be evaluated separately. Despite the manual inspection, it is important to declare that the dataset is biased toward authors' knowledge of Western music and some triplets might have been erroneously left out.

3.3 Experiments and results

After obtaining the sentence embedding using triplets, cosine similarity will be used to evaluate the relative semantic similarity. Anchor-positive and anchor-negative cosine similarity will be compared and a triplet will be regarded as correct if the first is greater than the second. A thorough analysis of the results is provided in the appendix chapter D. Finally, the accuracy of correct triplets will be calculated and reported.

3.3.1 Prompt sensitivity

Wrapping queried labels in a prompt is useful (Radford et al., 2021) but we are not aware of a thorough analysis of the performance variance concerning different prompts. As a result, we used 20 musically informed prompts. The exact wording of the prompts is provided in appendix C.1. Several words as "music", "recording" or "sound" have been used, to simulate human music captions/descriptions.

The standard deviation reported is relatively high for every case apart from the paraphrased-MiniLM model as presented in table 1. As the prompts do not provide additional information, it can be argued that the models are moderately sensitive to the prompts and "musical" words added can be useful. Lastly, the best model according to model size and performance is paraphrased-ALBERT.

3.3.2 Inability to model negation

Despite the acquired grammatical understanding reported by LLMs, they cannot model negation (e.g. "not rock") (García-Ferrero et al., 2023). To validate if this holds for musical labels, we constructed a separate list of triplets for both "Musical Genre" and "Musical Instruments". For each valid triplet obtained, we extracted unique anchor-positive pairs and introduced a negative label as a negation of the anchor and positive labels. We are left with 3756 and 8284 negative triplets for Genres and Instruments respectively. These were then used alongside 4 negative prompts, listed in appendix C.2.

[2]Visualization: http://www.jordipons.me/apps/audioset/

[3]Removed triplet cases are provided in Appendix table 4

Models	Prompts		Negation	
	Instruments	Genres	Instruments	Genres
mpnet-base	71.3 ± 3.7	76.4 ± 2.3	41.1 ± 3.7	43.2 ± 3.8
distilroberta	62.4 ± 2.4	69.6 ± 2.6	37.2 ± 3.6	42.3 ± 3.4
MiniLM-L12-v2	62.7 ± 2.3	70.9 ± 2.3	33.8 ± 6.5	37.3 ± 6.9
MiniLM-L6	65.8 ± 2.7	70.5 ± 1.6	37.4 ± 5.8	41.4 ± 5.8
Para-albert	69.6 ± 3.2	66.5 ± 1.7	33.4 ± 5.8	35.6 ± 5.7
Para-MiniLM-L3	63.2 ± 2.7	66.9 ± 0.8	29.0 ± 6.7	34.3 ± 5.0

Table 1: Presenting the percentage of correctly inferred triplets for Instruments and Genres respectively. Prompt sensitivity showcased from high standard deviation along 20 prompts. Also, Transformer-based models cannot model negation as the accuracy obtained is worse than random.

Models	Instrument Definitions		Genre Definitions	
	Definition + Label	Definition - Label	Definition - Label	Definition + Label
mpnet-base	83.2 ($\uparrow +11.9$)	72.5 ($\uparrow +1.2$)	84.9 ($\uparrow 8.5$)	72.7 ($\downarrow -3.7$)
distilroberta	75.8 ($\uparrow +13.4$)	73.9 ($\uparrow +11.5$)	71.5 ($\uparrow +1.9$)	69.5 ($\downarrow -0.1$)
MiniLM-L12-v2	81.8 ($\uparrow +19.09$)	72.4 ($\uparrow +9.7$)	79.5 ($\uparrow +8.6$)	70.2 ($\downarrow -0.7$)
MiniLM-L6	80.9 ($\uparrow +15.1$)	72.7 ($\uparrow +6.9$)	79.7 ($\uparrow +9.2$)	69.3 ($\downarrow -1.2$)
Para-albert	79.9 ($\uparrow +10.3$)	68.8 ($\downarrow -0.8$)	80.1 ($\uparrow +13.6$)	74.6 ($\uparrow +8.1$)
Para-MiniLM-L3	81.6 ($\uparrow +18.4$)	67.7 ($\uparrow +4.5$)	76.8 ($\uparrow +9.9$)	70.2 ($\uparrow +3.3$)

Table 2: Results for the experiment showing that models are sensitive towards specific words and cannot properly leverage the context, in the form of a definition. The figures in brackets indicate the difference in accuracy with respect to the experiments with prompts only of table 1.

The performance is worse than random, as shown in table 1, which provides further evidence that LLMs cannot model negation in general and musical terminology. Different prompts lead to considerable differences in accuracy, with the worst performance reported being $\approx 23\%$. This might have potential implications in applications such as captioning, as datasets include negation.

3.3.3 Sensitivity towards the presence of specific words

Using artificially generated definitions of labels instead of generic prompts led to an increased zero-shot image classification accuracy (Pratt et al., 2023). Drawing inspiration from this and leveraging single-sentence definitions provided by Audioset, we evaluate the performance when using the label-free definition and the combination of the label and definition simultaneously.

Excluding the label from the definition leads to a drop in every experiment, meaning that models might be sensitive to labels and not the semantics provided indirectly by the definition. On the other hand, the definition leads to an increment in accuracy in most cases, as shown in table 2.

4 Conclusions and future work

In this paper, we quantified the musical knowledge of six Transformer-based models based on triplet accuracy with musical labels for genres and instruments. We identified three shortcomings: prompt sensitivity, difficulty modeling negation and sensitivity to specific words.

To overcome these shortcomings, we propose using augmentation during training and varying the prompt structures to avoid prompt sensitivity. This approach can utilize definitions to substitute labels with their definitions. To address negation modeling, we suggest multi-task learning that includes tagging negative labels in a caption and maximizing the distance between negative and positive versions of the tags in contrastive losses.

We recommend using lexical databases (e.g. WordNet), which offer more elaborate music concept relationships, instead of using a tree to obtain triplets. We highlight that further filtering needs to be done to form meaningful triplets and produce good-quality evaluation datasets. Lastly, despite reporting increments when definitions are used, further testing is required.

References

Andrea Agostinelli, Timo I. Denk, Zalán Borsos, Jesse Engel, Mauro Verzetti, Antoine Caillon, Qingqing Huang, Aren Jansen, Adam Roberts, Marco Tagliasacchi, Matthew Sharifi, Neil Zeghidour, and Christian Havnø Frank. 2023. Musiclm: Generating music from text. *ArXiv*, abs/2301.11325.

Marwah Alian and Arafat Awajan. 2020. Factors affecting sentence similarity and paraphrasing identification. *Int. J. Speech Technol.*, 23(4):851–859.

Zaira Hassan Amur, Yew Kwang Hooi, Hina Bhanbhro, Kamran Dahri, and Gul Muhammad Soomro. 2023. Short-text semantic similarity (stss): Techniques, challenges and future perspectives. *Applied Sciences*, 13(6).

Hyunjin Choi, Judong Kim, Seongho Joe, and Youngjune Gwon. 2021. Evaluation of BERT and ALBERT sentence embedding performance on downstream NLP tasks. *CoRR*, abs/2101.10642.

Keunwoo Choi, György Fazekas, Kyunghyun Cho, and Mark Sandler. 2018. The effects of noisy labels on deep convolutional neural networks for music tagging. *IEEE Transactions on Emerging Topics in Computational Intelligence*, 2:139–149.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Qingxiu Dong, Xiaojun Wan, and Yue Cao. 2021. Parasci: A large scientific paraphrase dataset for longer paraphrase generation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 424–434.

Xingjian Du, Zhesong Yu, Jiaju Lin, Bilei Zhu, and Qiuqiang Kong. 2024. Joint music and language attention models for zero-shot music tagging. In *ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1126–1130. IEEE.

Iker García-Ferrero, Begoña Altuna, Javier Alvez, Itziar Gonzalez-Dios, and German Rigau. 2023. This is not a dataset: A large negation benchmark to challenge large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 8596–8615, Singapore. Association for Computational Linguistics.

Josh Gardner, Simon Durand, Daniel Stoller, and Rachel Bittner. 2024. Llark: A multimodal instruction-following language model for music. *Proc. of the International Conference on Machine Learning (ICML)*.

Jort F. Gemmeke, Daniel P. W. Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R. Channing Moore, Manoj Plakal, and Marvin Ritter. 2017. Audio set: An ontology and human-labeled dataset for audio events. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 776–780.

Geoffrey E. Hinton, Oriol Vinyals, and Jeffrey Dean. 2015. Distilling the knowledge in a neural network. *ArXiv*, abs/1503.02531.

Qingqing Huang, Aren Jansen, Joonseok Lee, Ravi Ganti, Judith Yue Li, and Daniel P. W. Ellis. 2022. MuLan: A joint embedding of music audio and natural language. In *International Society for Music Information Retrieval Conference*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Bohan Li, Hao Zhou, Junxian He, Mingxuan Wang, Yiming Yang, and Lei Li. 2020. On the sentence embeddings from pre-trained language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 9119–9130, Online. Association for Computational Linguistics.

Yang Liu and Mirella Lapata. 2019. Text summarization with pretrained encoders. *CoRR*, abs/1908.08345.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *ArXiv*, abs/1907.11692.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2021. Muscaps: Generating captions for music audio. In *2021 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8.

Ilaria Manco, Emmanouil Benetos, Elio Quinton, and György Fazekas. 2022. Contrastive audio-language learning for music. In *Proceedings of the 23rd International Society for Music Information Retrieval Conference (ISMIR)*.

Ilaria Manco, Benno Weck, Seungheon Doh, Minz Won, Yixiao Zhang, Dmitry Bogdanov, Yusong Wu, Ke Chen, Philip Tovstogan, Emmanouil Benetos, Elio Quinton, György Fazekas, and Juhan Nam. 2023. The song describer dataset: a corpus of audio captions for music-and-language evaluation. In *Machine Learning for Audio Workshop at NeurIPS 2023*.

George A. Miller. 1995. Wordnet: a lexical database for english. *Commun. ACM*, 38(11):39–41.

Atul Kr. Ojha, A. Seza Doğruöz, Harish Tayyar Madabushi, Giovanni Da San Martino, Sara Rosenthal, and Aiala Rosá, editors. 2024. *Proceedings of the*

18th International Workshop on Semantic Evaluation (SemEval-2024). Association for Computational Linguistics, Mexico City, Mexico.

Sergio Oramas, Luis Espinosa Anke, Francisco Gómez, and Xavier Serra. 2018. Natural language processing for music knowledge discovery. *Journal of New Music Research*, 47:365 – 382.

Sergio Oramas, Mohamed Sordo, Luis Espinosa Anke, and Xavier Serra. 2015. A semantic-based approach for artist similarity. In *International Society for Music Information Retrieval Conference*.

Sarah Pratt, Ian Covert, Rosanne Liu, and Ali Farhadi. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15691–15701.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Conference on Empirical Methods in Natural Language Processing*.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *Preprint*, arXiv:1910.01108.

Xiaofei Sun, Yuxian Meng, Xiang Ao, Fei Wu, Tianwei Zhang, Jiwei Li, and Chun Fan. 2022. Sentence Similarity Based on Contexts. *Transactions of the Association for Computational Linguistics*, 10:573–588.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748.

Jan Philip Wahle, Terry Ruas, Tomáš Foltýnek, Norman Meuschke, and Bela Gipp. 2022. Identifying machine-paraphrased plagiarism. In *Information for a Better World: Shaping the Global Future*, pages 393–413, Cham. Springer International Publishing.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Yusong Wu, Ke Chen, Tianyu Zhang, Yuchen Hui, Taylor Berg-Kirkpatrick, and Shlomo Dubnov. 2023. Large-scale contrastive language-audio pretraining with feature fusion and keyword-to-caption augmentation. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP*.

Haoran Xu, Benjamin Van Durme, and Kenton Murray. 2021. Bert, mbert, or bibert? a study on contextualized embeddings for neural machine translation. In *Conference on Empirical Methods in Natural Language Processing*.

Lingling Xu, Haoran Xie, Fu Lee Wang, Xiaohui Tao, Weiming Wang, and Qing Li. 2024. Contrastive sentence representation learning with adaptive false negative cancellation. *Information Fusion*, 102:102065.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32*, pages 5754–5764. Curran Associates, Inc.

Botao Yu, Peiling Lu, Rui Wang, Wei Hu, Xu Tan, Wei Ye, Shikun Zhang, Tao Qin, and Tie-Yan Liu. 2022. Museformer: Transformer with fine- and coarse-grained attention for music generation. In *Advances in Neural Information Processing Systems*, volume 35, pages 1376–1388. Curran Associates, Inc.

A Acknowledgments

The first author is a research student at the UKRI Centre for Doctoral Training in Artificial Intelligence and Music, supported jointly by UK Research and Innovation [grant number EP/S022694/1] and Queen Mary University of London.

B Language models used

All the models used are pretrained and then fine-tuned for sentence similarity on several corpora of pairs. Paraphrase models share the same fine-tuning dataset and the same happens for the remaining 4, with an additional 50 million sentence pairs for all-distilroberta-v1. More information can be found in the respective papers, Sentence Transformer [4] package documentation and Hugging Face websites [5].

MPNet unifies the Masked Language Modeling (MLM) and Permuted Language Modeling pretasks, used by BERT (Devlin et al., 2019) and XLNet (Yang et al., 2019) respectively, to train a Transformer backbone. The tokens of the input are permuted, a set of them is masked and the objective is to predict the masked section, while the

[4] https://sbert.net/

[5] https://huggingface.co/

positional information of the full sentence is also known.

DistilBERT (Sanh et al., 2020) is a 40% smaller BERT model that is trained on the same regime as BERT but with an additional loss term. The distillation loss (Hinton et al., 2015) is:

$$L_{ce} = \sum_i t_i * log(s_i) \quad (1)$$

where t_i, s_i is the probability for the predicted tokens of the teacher (BERT) and student (DistilBERT) models respectively. This is used to let the student approximate the target probability distribution of the teacher and therefore, learn from the teacher model.

RoBERTa (Liu et al., 2019) is a model based on BERT with removing next-sentence prediction pretraining, increasing the mini-batch size and altering key hyperparamaters. The analysis of the last are out of the scope for this paper. DistilRoBERTa uses RoBERTa and the distillation process described for DistilBERT.

Instead of approximating the target probability distribution, MiniLM (Wang et al., 2020) proposed to "mimic" the last self-attention module between the student and teacher models. In addition to approximating the attention distribution, this system approximates the relations between the scaled dot-products of queries, keys and value embeddings. Therefore, it also models the second-degree associations between the self-attention embeddings, as well as their distribution.

Finally, ALBERT (Lan et al., 2019) utilizes parameter reduction techniques, as well as swapping the Next Sentence Prediction to Sentence Ordering Prediction. Firstly, Factorized Embedding Parametrization is used to decompose the vocabulary embedding matrix into two small matrices. As a result, the size of the hidden layers is decoupled from the size of the token embeddings. Secondly, Cross-Layer Parameter Sharing relaxes the dependency between memory demands and model depth. Lastly, Sentence Ordering Prediction is focused on predicting the sequence of two sentence segments, while Next Sentence Prediction is used to predict if the pair of sentences is from the same document or not.

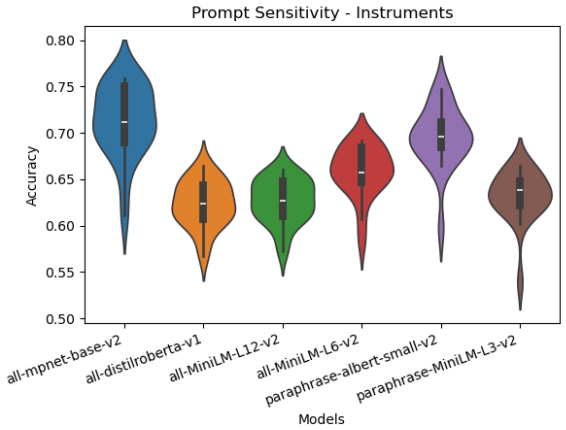

Figure 1: Prompt sensitivity of 6 Transformer-based models with respect to musical instrument terminology.

C Prompts used

C.1 Prompt sensitivity

The prompts used for evaluating the sensitivity towards different musically informed prompts of Transformer-based models are:

1. "*The sound of <label>*"
2. "*Music made with <label>*"
3. "*A <label> track*"
4. "*This is a recording of <label>*"
5. "*A song with <label>*"
6. "*A track with <label> recorded*"
7. "*A music project with <label>*"
8. "*Music made from <label>*"
9. "*Music of <label>*"
10. "*A music recording of <label>*"
11. "*This song is made from <label>*"
12. "*The song has <label>*"
13. "*Music song with <label>*"
14. "*Music song with <label> recorded*"
15. "*Musical sounds from <label>*"
16. "*This song sounds like <label>*"
17. "*This music sounds like <label>*"
18. "*Song with <label> recorded*"
19. "*A <label> music track*"
20. "*Sound of <label>*"

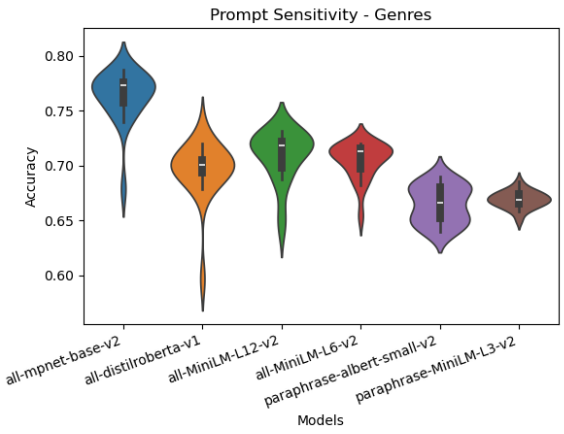

Figure 2: Prompt sensitivity of 6 Transformer-based models with respect to musical genre terminology.

C.2 Negation modeling

The four prompts used to evaluate the inability to model negation:

1. "*No <label>*"
2. "*Not the sound of <label>*"
3. "*Doesn't sound like <label>*"
4. "*Not music from <label>*"

	Instrument Prompts			
Models	**#1**	**#2**	**#3**	**#4**
mpnet-base	45.4	35.9	44.0	39.5
distilroberta	41.4	31.6	39.0	36.7
MiniLM-L12-v2	44.2	30.8	33.6	26.8
MiniLM-L6	46.9	33.2	37.5	32.0
Para-albert	42.7	28.6	28.5	34.0
Para-MiniLM-L3	40.4	23.6	26.8	25.1

(a) Instruments

	Genre Prompts			
Models	**#1**	**#2**	**#3**	**#4**
mpnet-base	49.0	39.5	44.0	40.0
distilroberta	45.6	37.8	45.2	40.1
MiniLM-L12-v2	47.2	35.6	38.7	27.9
MiniLM-L6	49.6	40.7	42.1	33.3
Para-albert	44.8	32.2	29.8	35.7
Para-MiniLM-L3	42.4	32.5	33.0	29.1

(b) Genres

Table 3: Presentation of results for experiment 3.3.2. No model performed on par with the random baseline.

C.3 Examples of removed triplets

As stated in 3.2, there were some triplets of ambiguous quality. We argue that removing these is far more important than building a very big evaluation dataset.

For reference, we present 10 triplets of different ambiguousness levels for each category in table 4.

D Detailed experiment results

D.1 Prompt sensitivity

Generally, prompt sensitivity is evident in every model. The biggest and best model, all-mpnet-base-v2, has the largest and one of the largest variances for instruments (figure 1) and genres respectively (figure 2).

Paraphrase-MiniLM-L3-v2 had the smallest variance for genres, at the expense of a lower accuracy. This might be due to the different distillation process chosen. If an application demands robustness towards prompt sensitivity, that would be the best choice.

Apart from all-mpnet-base-v2, every model had approximately the same variance when the outliers were discarded, as can be seen in figure 1.

D.2 Negation modeling

By far the worst deficiency found is the inability of Transformer-based models to model negation. These failed to surpass random choice in every experiment, while altering the prompt led to a significant decrease in accuracy, up to $\approx 20\%$. This is presented in table 3a.

This result can have large implications on developing or evaluating captioning systems, as datasets (Agostinelli et al., 2023; Manco et al., 2023) contain negation and following these results, can lead to erroneous inference. Also, joint audio-text models, also known as two-tower systems, can be negatively impacted. Further testing is required in the future.

Instruments		
Anchor	Positive	Negative
Musical instrument	Plucked string instrument	Mandolin
Cowbell	Accordion	Flute
Guitar	French horn	Timpani
Electric guitar	Hammond organ	Rhodes piano
Bass guitar	Brass Instrument	Alto saxophone
Tapping (guitar technique)	French horn	Electric piano
Sitar	Cymbal	Rimshot
Keyboard (musical)	Cowbell	Acoustic guitar
Piano	Didgeridoo	Cello
Organ	Trombone	Timpani
Genres		
Anchor	Positive	Negative
Music genre	Rhythm and blues	Swing music
Pop music	Jazz	Swing music
Hip hop music	Classical music	Drum and bass
Rock music	Independent music	Grime music
Heavy metal	Electronic music	Oldschool jungle
Progressive rock	Chant	Oldschool jungle
Reggae	Music of Asia	Cumbia
Jazz	New-age music	Heavy metal
Kuduro	Music for children	Grunge
Funk carioca	Christian music	Electronica

Table 4: Table with examples of removed triplets. The filtering criterion is based on the ambiguity or relative difficulty in determining whether the anchor is more similar to the positive or negative label.