Investigating the Emergent Reasoning Abilities of Large Language Models in Sentiment Analysis of Codified Music

Gabriel Assis, Laura Alvarenga, João Vitor de Moraes,
Lívia de Azevedo, Aline Paes

Instituto de Computação, Universidade Federal Fluminense (UFF)
Niterói, RJ, Brazil

{assisgabriel,l_alvarenga,joaovitormoraes,liviaazevedosilva}@id.uff.br,
alinepaes@ic.uff.br

Abstract. *The perception of positive or negative sentiment in a piece of music is influenced by musical features such as tempo, notes, chords, and rhythms, collectively shaping its emotional character. This paper examines the ability of Large Language Models (LLMs) to account for these features and infer sentiment from symbolically encoded music, exploring both zero-shot and fine-tuned approaches. Our results show that while LLMs exhibit some capability in processing symbolic representation of musical elements, their ability to associate music with sentiment reliably remains limited. The models struggle to align their predictions with human-assigned labels, with accuracy hovering around 0.6. These results suggest that current text-based approaches may not fully capture the complex interplay between musical structure and emotional expression.*

1. Introduction

Musical Computing (MC) is a multidisciplinary field that encompasses a wide range of tasks, including music information retrieval [Yepez et al. 2024], algorithmic generation of melodies and lyrics [Copet et al. 2023, Yuan et al. 2024], genre classification [Seufitelli and Moro 2023, Oliveira et al. 2024], and sentiment analysis [Ferreira and Whitehead 2019, Alves et al. 2024]. However, bridging computational precision with musical expressiveness involves challenges like structural representation, pattern recognition, and capturing emotional nuance, which is central to many music-related tasks [Yuan et al. 2024].

The rise of Large Language Models (LLMs) has revolutionized numerous fields with impressive emergent capabilities [Wolf et al. 2020, Qin et al. 2024]. These capabilities stem from the extensive knowledge acquired during pre-training, which can be accessed through mechanisms such as in-context learning [Brown et al. 2020]. Expanding beyond their native domain of natural language [Qin et al. 2024], recent research explores the application of these models to symbolic music, exploring whether their advanced contextual understanding can be transferred to the structured language of musical notation [Yu et al. 2023, Yuan et al. 2024].

This study directly evaluates whether LLMs can interpret sentiment from music using only symbolic representations and their internal, pre-trained knowledge. By focusing on symbolic encoding, which represents musical information textually, we enable LLMs to analyze music structurally without requiring audio processing. This approach leads us to address two fundamental research questions: *(RQ1) Are LLMs capable of* *"understanding" the encoding of symbolic music?* and, subsequently, *(RQ2) Can LLMs leverage symbolic musical encoding to identify the sentiment expressed in music?*. We aim to explore whether LLMs can capture and interpret musical subjectivity and sentiment, relying solely on their internal representations and prompt-based stimuli.

To answer these questions, our work evaluates cutting-edge language models, including Mistral [Jiang et al. 2023], Llama 3 [Grattafiori et al. 2024], Phi 3 [Abdin et al. 2024], Qwen 2.5 [Yang et al. 2024], Gemma 2 [Riviere et al. 2024], and GPT-4o [Hurst et al. 2024] families. These models are tested for their ability to perform zero-shot sentiment analysis. Furthermore, we compare their outcomes with those of a Llama 3 model specifically adapted through fine-tuning for the task. The experiments use the VGMIDI dataset [Ferreira and Whitehead 2019], a symbolic music collection offering binary sentiment labels (positive vs. negative). Our findings indicate that, while the models demonstrate the ability to interpret the encoded representation of music — linking features such as tempo and pitch to the assignment of sentiments — this alignment falls short of achieving correlation with the labels provided by human listeners.

This paper has seven sections beyond the Introduction. Section 2 reviews the related work. Section 3 explains data preparation, the selected models and strategies employed. Section 4 discusses the experimental setup, including the dataset and implementation details. Section 5 presents and analyzes the results. Lastly, 6 summarizes the findings and discusses potential avenues for future research[1].

2. Related Work

This section discusses prior research related to the computational encoding of music and the analysis of sentiment in musical content.

2.1. Encodable Musical Representation

Representing music computationally is complex due to elements like rhythm, chords, and harmonies [Yu et al. 2023]. Research has explored various methods to enable machine learning (ML) models to analyze and generate music. [Alves et al. 2024] utilize feature-engineered and spectrogram-based representations for music emotion recognition, employing models like Decision Trees and CNNs. [Yuan et al. 2024] leverage the ABC notation system[2], a concise ASCII-based format encoding musical notes and structures.

In industry contexts, MIDI (Musical Instrument Digital Interface) is widely used as a standardized, event-driven representation capturing notes, duration, intensity, tempo, and instruments [Lu et al. 2023, Huang and Yang 2020, Ferreira and Whitehead 2019]. Extensions to MIDI, such as REMI (Revamped MIDI-Derived Events), enhance interpretability by aligning with human perception [Huang and Yang 2020]. [Lu et al. 2023] integrate MIDI and REMI formats into music generation frameworks driven by textual descriptions. [Ferreira and Whitehead 2019] further transform MIDI data into word-like sequences tailored for sentiment analysis in music, a representation aligned with the current study's goals. Additional representation details are discussed in Section 3.1.

[1] Code on `https://github.com/MeLLL-UFF/LLM4CodedMusicSent`

[2] `https://abcnotation.com`

2.2. Sentiment Analysis in Music

Music Emotion Recognition (MER) is a field within computational music research, aiming to identify and analyze perceivable emotions in music, such as anger, joy, and calmness [Han et al. 2022]. In this context, [Santos et al. 2021] present advancements leveraging models such as MLPs and CNNs applied directly to feature representations and spectrograms. These findings align closely with the results reported by [Alves et al. 2024], who conducted a comparative evaluation of various ML models in the task of predicting emotions associated with five primary categories: Happiness, Sadness, Dramaticism, Romanticism, and Aggressiveness.

More aligned with our work on sentiment analysis, [Ferreira and Whitehead 2019] employ a binary valence scale (positive vs. negative) [Russell 1980] in experiments designed to embed polarity perception into Long Short-Term Memory (LSTM) networks. Their method involves training the model as a language model, using textual representations of music for the classification task, and then applying it to generate compositions that convey positivity or negativity. Similarly, [Hung et al. 2021] also apply LSTM-based approaches for the classification of sentiment in text-encoded music within the valence spectrum, demonstrating that such methods can rival — and even outperform — audio-based sentiment analysis.

To the best of our knowledge, no previous work has explored using LLMs' encoded knowledge for sentiment analysis on symbolic music representations.

3. Method

This section outlines the processes employed to evaluate the ability of LLMs to perform sentiment analysis on the symbolic representation of music. Detailed descriptions of the data representation and preparation, the prompt design, and the selection and application of the models are provided in the subsequent sections.

3.1. Data Preparation

Aiming to incorporate features such as melody, harmony, tempo, and timbre — elements associated with the perception of valence in music — our work leverages the representation developed by [Ferreira and Whitehead 2019], utilizing a publicly available associated implementation that transforms MIDI data into the proposed format[3].

The available encoder transforms MIDI files into a linear, token-based textual representation that encapsulates various aspects of musical information. Each token corresponds to a specific musical event or attribute, such as tempo changes (e.g., t_XXX), velocity levels (e.g., v_XXX), note durations (e.g., d_TYPE_DOTS), pitch values (e.g., n_XX), and inter-event timings (e.g., w_XX). This encoding is designed to capture the music's temporal, dynamic, and harmonic characteristics in a structured sequential format, which a language model may process. Further details are depicted in Figure 1.

[3] https://github.com/lucasnfe/music-sentneuron

Details of the symbolic music encoding

- **Tempo Changes (`t_XXX`)**:
  - `XXX`: An integer representing the tempo in beats per minute (BPM).
  - **Range**: Discretized in bins of size 4, such that $XXX \in \{24, 28, 32, \ldots, 160\}$.
  - **Example**: `t_120` represents a tempo of 120 BPM.
- **Velocity Changes (`v_XXX`)**:
  - `XXX`: An integer indicating the note's velocity (loudness).
  - **Range**: Discretized in bins of size 4, such that $XXX \in \{4, 8, 12, \ldots, 128\}$.
  - **Example**: `v_100` sets the velocity to 100.
- **Note Durations (`d_TYPE_DOTS`)**:
  - `TYPE`: The duration type (e.g., `whole`, `quarter`, `eighth`, etc.).
  - `DOTS`: An integer indicating the number of augmentation dots (0 to 3).
  - **Example**: `d_quarter_1` represents a dotted quarter note.
- **Pitch (`n_XX`)**:
  - `XX`: An integer representing the MIDI pitch number.
  - **Range**: $XX \in \{0, 1, \ldots, 127\}$, covering the full MIDI pitch range.
  - **Example**: `n_60` corresponds to Middle C.
- **Wait (`w_XX`)**:
  - `XX`: An integer specifying the number of time steps to wait before the next event.
  - **Example**: `w_4` means a wait of 4 time steps.

Figure 1. Symbolic encoding scheme for musical data.

3.2. Prompt Design

The prompt incorporates the previously defined symbolic representation to evaluate whether LLMs can effectively perform sentiment analysis based on this format. To assess interpretability, models were also instructed to provide brief justifications for their classifications. The goal is to examine whether the output aligns with the intended representation and whether the models establish meaningful associations between musical features and sentiment labels. For instance, this could include associating higher tempos with the positive class. The full prompt used in this evaluation is presented in Figure 2.

Prompt

You will classify music represented in symbolic form as either positive or negative.

Symbolic Representation
d_[duration]_[dots]: Defines the duration of the upcoming notes. The [duration] specifies the type of note (e.g., breve, whole, half, quarter, eighth, 16th, or 32nd). The [dots] indicates the number of dots extending the note's duration, and can be any integer from 0 to 3.
v_[velocity]: Indicates the velocity (or loudness) of the following notes. Velocity is discretized into bins of size 4, allowing values such as 4, 8, 12, up to 128.
t_[tempo]: Changes the tempo of the piece, measured in beats per minute (bpm). Tempo is discretized into bins of size 4, ranging from 24 to 160 bpm. This controls the speed at which the piece is played.
n_[pitch]: Specifies the pitch of a note using its MIDI pitch number. The [pitch] value is an integer ranging from 0 to 127, representing the full range of MIDI pitches. For example, n_60 corresponds to Middle C.
w_[wait]: Specifies the number of time steps (units of waiting) that pass before the next musical event occurs. The value associated with w, such as in w_2 or w_3, represents the number of time steps with no musical events.
\n: Marks the end of the piece.

Music
`{music}`

Your answer must strictly follow this format:
- answer: A string, either "positive" or "negative"
- justify: A brief explanation justifying your classification

Figure 2. Prompt for Symbolic Music Classification.

3.3. Models

This section describes the selected models and the methods used for their application.

3.3.1. Model Selection

In light of hardware constraints (see Section 4) and the objective of evaluating cutting-edge LLMs, we selected models with up to 9 billion parameters from the renowned families Mistral, Llama 3, Phi 3, Qwen 2.5, and Gemma 2. Specifically, the models included Mistral 7B [Jiang et al. 2023], Llama 3 8B [Grattafiori et al. 2024], Phi 3 Small [Abdin et al. 2024], Qwen 2.5 7B [Yang et al. 2024], and Gemma 2 9B [Riviere et al. 2024], all of which were used in their "instruct" options. To evaluate a state-of-the-art closed-source model, experiments were also conducted with GPT4o-mini [Hurst et al. 2024]. This selection examines how diverse models, varying in architecture and pretraining, can integrate symbolic music representations and perform reasoning in the context of sentiment analysis.

3.3.2. Model Application

To assess how the models inherently encode information — stemming from their pre-training phase — and to determine whether their emergent abilities alone are sufficient to capture the notion of sentiment associated with symbolic representations, all models were evaluated using in-context learning in a zero-shot configuration [Brown et al. 2020]. In this setup, no adjustments were made to the models' weights, and no example samples were provided during inference through the prompt.

However, a comparative analysis was conducted to evaluate further whether emergent abilities alone are adequate. Following the evaluation of models performance in the zero-shot configuration, a model from the best-performing family was fine-tuned. This additional step examined how fine-tuning impacts the model's capacity to perform the sentiment analysis task. During the fine-tuning process, the same prompt described in Section 3.2 was employed, except for the instruction to produce a justification, which was omitted due to the lack of ground truth data.

4. Experimental Setup

4.1. Data

The dataset used is VGMIDI [Ferreira and Whitehead 2019][4], comprising 4,050 MIDI piano arrangements of video game soundtracks, specifically tailored for affective music composition and valence analysis. It includes a subset of 204 tracks annotated for emotion using Russell's Circumplex Model [Russell 1980], covering valence (positive/negative) and arousal dimensions. For valence, the focus here, 138 tracks are positive, and 66 are negative. Annotations were crowdsourced from 1,425 U.S.-based participants (average age 31; 55% female, 42% male, 3% other/non-disclosed). [Ferreira and Whitehead 2019] provides binary sentiment labels (positive or negative) based on these annotations. In this

[4] https://github.com/lucasnfe/vgmidi

study, we use only the annotated subset: the full set is used for zero-shot inference, and a 10-fold stratified split is applied during fine-tuning.

4.2. Implementation Details

To ensure reproducibility, experiments were performed on two Nvidia RTX 4090 GPUs (24GB each). Open-source models were implemented using Hugging Face's *Transformers* [Wolf et al. 2020] with generative parameters: `max_new_tokens=500`, `do_sample=True`, `temperature=1` (given the creative nature of the task), and `top_p=0.9`. GPT4o-mini was accessed via OpenAI's public API[5] using same configurations. The fine-tuned model was trained for three epochs with a learning rate of 2×10^{-5} and weight decay of 0.01. Due to hardware limitations, music representations in prompts were truncated to 1,800 characters.

5. Results

This section presents the classification results of the models evaluated for predicting positive (+) and negative (-) sentiment in musics. Additionally, we discuss a qualitative inspection conducted to examine the rationale generated with the label assignments made by the best-performing model, selected based on classification metrics.

5.1. Quantitative Results

Table 1. Macro classification metrics for the evaluated models in the zero-shot setting. The best results are in bold; the second-best is underlined.

Model	Prec.	Rec.	F1	Acc.
Mistral-7B	0.33	0.47	0.39	0.60
Phi-3-small	0.39	0.48	0.40	0.64
GPT4o-mini	0.42	0.49	0.41	**0.66**
Gemma-2 9B	0.33	0.47	0.39	0.64
Qwen-2.5-7B	0.48	0.50	0.42	**0.66**
Llama-3.1-8B	**0.52**	**0.51**	**0.51**	0.60

Table 1 summarizes the performance metrics — precision, recall, F1-score, and accuracy — for the binary sentiment analysis task. Notably, the GPT4o-mini and Qwen-2.5-7B models excel in terms of accuracy. However, the highest scores for the remaining metrics are consistently achieved by the Llama-3.1-8B model. The second-best results across most metrics are attributed to the Qwen-2.5-7B model. It is worth highlighting the F1-score, where Llama-3.1-8B outperforms the Qwen-2.5-7B model — the second-best performer for this metric — by approximately 0.1 points. This distinction is especially significant given the dataset's class imbalance, as the F1-score is a key metric in such contexts. Nonetheless, no model achieved high levels of performance. Models such as Gemma-2 9B and Mistral-7B appear to struggle more with the task, being the only models with at least one metric falling below 0.40. On the other hand, the best results for precision, recall, and F1-score hover around 0.50, while accuracy reaches a maximum of 0.66. This indicates that while certain models exhibit relative strengths, overall performance remains moderate across the results.

[5] https://platform.openai.com/

Given that the best performance observed in Table 1 was achieved using a model from the Llama 3 family, and considering the significant hardware demands associated with fine-tuning — substantially higher than those required for inference — the Llama-3.2-3B model was chosen as the representative of this family for further evaluation. The results (Table 2) indicate a slight improvement in precision and recall, stability in the F1 score, and a minor decrease in accuracy when evaluating this new fine-tuned model.

Table 2. Comparison of macro-averaged metrics between Llama 3.1 8B (zero-shot) and Llama 3.2 3B (fine-tuned). The best results are in bold.

Model	Prec.	Rec.	F1	Acc.
Llama-3.1-8B	0.52	0.51	**0.51**	**0.60**
Llama-3.2-3B	**0.54**	**0.54**	**0.51**	0.56

Table 3. Confusion matrices for Top-3 models (Table 1 and Table 2): Llama-3.2-3B (fine-tuned), Llama-3.1-8B (zero-shot), and Qwen-2.5-7B (zero-shot).

(a) Llama-3.2-3B

	Pred. +	Pred. –
True +	88	50
True –	40	26

(b) Llama-3.1-8B

	Pred. +	Pred. –
True +	110	28
True –	51	15

(c) Qwen-2.5-7B

	Pred. +	Pred. –
True +	133	5
True –	64	2

However, Table 3 reveals notable differences in the predictive behavior of the two approaches analyzed. While the fine-tuned model achieved a more balanced performance across both classes (positive and negative), the top-performing models within the zero-shot approach tended to favor predictions for the positive (and majority) class over the negative class. This observation underscores the superior robustness and generalizability of the fine-tuning approach.

Table 4. Comparison of accuracy between the fine-tuned LLM approach (Llama 3.2 3B) and the LSTM approaches reported by [Ferreira and Whitehead 2019]. The best result is in bold.

Model	Accuracy
Gen. LSTM (Ferreira and Whitehead, 2019)	**0.90**
Sup. LSTM (Ferreira and Whitehead, 2019)	0.60
Llama-3.2-3B (this work)	0.56

Lastly, Table 4 summarizes the results comparing the baseline for the VGMIDI dataset established by [Ferreira and Whitehead 2019] and the fine-tuned LLM approach explored in our study. [Ferreira and Whitehead 2019] evaluated a supervised LSTM-based model alongside their proposed method, which involved adapting the LSTM vocabulary to a symbolic representation for music generation and combining it with logistic regression for feature extraction. While our fine-tuned LLM approach performed comparably to the supervised LSTM, it underperformed relative to the generative LSTM. Furthermore, other LLMs evaluated in our experiments, such as Qwen 2.5, GPT4o-mini, and Phi 3, surpassed the performance of the supervised LSTM, as detailed in Table 1. However, it is essential to note that accuracy, the metric available for comparison, can mask class imbalance issues, as evidenced by the confusion matrix for Qwen 2.5 in Table 3.

5.2. Qualitative Inspection

To investigate the relationship between the outputs generated by the classifications assigned by the models, we conducted a qualitative inspection of the justifications generated by the most successful approach, Llama 3.2 3B fine-tuned. The goal was to determine whether the model could effectively map the encoded meaning of the music to the possible labels. Moreover, we analyzed the importance attributed in the justifications to the encoded attributes within the context of sentiment analysis.

In music studies, specific musical elements such as tempo and key have long been associated with particular emotions and valences. For instance, faster tempos are often linked to emotions like joy and fear [Scherer 1986], while slower tempos evoke feelings of tenderness and sadness [Davitz 1964, Fónagy and Magdics 1963]. Similarly, major keys are typically associated with happiness, and minor keys with sadness [Loveday 2022]. Even though [Helmholtz 1954] demonstrated that minor chords produce more complex sound waves, which are more challenging to process and could feel less harmonious, cultural differences reveal contrasting responses to these features. For example, the Khowar and Kalash tribes of northwestern Pakistan have been found to associate minor chords with positive emotions, while major chords are seen as negative [Loveday 2022].

Table 5. Most frequently used modifiers (word count) to describe musical elements in positive and negative predictions.

Predicted Label	Modifiers Used (Frequency)
Positive	*consistent* (53), *slow* (6), *fast* (26), *fast-paced* (8), *faster* (7), *higher* (23), *high* (14), *strong* (16), *loud* (8)
Negative	*consistent* (9), *slow* (30), *fast* (4), *fast-paced* (1), *faster* (1), *high* (3), *higher* (5), *loud* (1), *low* (5)

In our qualitative evaluations, we observed that the model accurately identified changes in speed, recognizing fast and slow tempos along with other musical features. Furthermore, it grounded its predictions in these elements, many of which align with the literature — for instance, linking fast tempos with happiness and minor chords with sadness. Here are some examples of the model's reasoning for positive valence: *"The use of higher velocities (e.g., v_124) suggests a more energetic and lively performance"* and *"The music piece has a consistent and upbeat tempo with a high velocity of 100, which suggests a positive and energetic mood."* Conversely, here are examples of the model's reasoning for negative valence: *"I noticed that the notes 'n' are predominantly below 60 (comparable to a dissonant minor note)"* and *"slower beat per minute also contribute to the negative overall feel."* The model also linked less variation in tempo to positive valence, as reflected in its reasoning: *"the tempo changes are not extreme"*, while connecting more frequent tempo changes to negative valence, as evidenced by its reasoning: *"the sudden changes in tempo, and the use of rests and wait commands create a sense of tension and uncertainty."* Table 5 further supports this perception, showing that the model most frequently associates modifiers like *'fast'*, *'high'*, and *'higher'* with positive valence, while *'slow'* and *'low'* are more often linked to negative valence.

The model's comparison of musical features and valences was dissociated from humans' final perception of the feelings provoked by the music, as seen by the model's performance in the previous subsection 5.1. While it linked certain musical elements, such as tempo and key, to specific polarity classifications, these associations did not fully align with how the annotators interpret or experience the valences conveyed by the music. However, we only have access to the final valence labels, without the detailed human interpretations that could provide deeper insights. The model itself acknowledged this limitation, stating: *“the exact interpretation of the music can be subjective, and a more thorough analysis would be needed for a definitive classification"*.

This underscores the fact that a range of subjective factors beyond the inherent properties of the sound itself influences the musical valence. The challenge lies in that while models may effectively encode musical features and establish correlations between these features and sentiment — often supported by existing literature — they may still struggle to accurately predict the final label assigned by annotators. In other words, even when a model reflects a widely accepted understanding of how certain features relate to sentiment, it may not account for the nuanced, individual differences in how people perceive and emotionally connect with music. Furthermore, nostalgia, an emotion often evoked by music associated with personal and social memories, could influence the emotional classification of music [Barrett and Janata 2016]. In the context of our corpus — a curated collection of piano arrangements of video game soundtracks — we suggest that nostalgia may explain the prevalence of positive valences, as the music likely triggers affective memories in listeners.

Other variables, such as musical expertise, could further influence emotion perception, but these were difficult to quantify in our analysis. Elements like personal memories, cultural context, gender, and individual differences in emotional processing all play significant roles in how music is emotionally perceived, which the model could not account for. As a result, while the model provided valuable predictions based on the analysis of musical features, it failed to capture the nuanced, human-centric nature of the emotional response to music.

6. Conclusion

This study explored the capabilities of six state-of-the-art LLM families to analyze encoded music and classify sentiments as either positive or negative, employing both zero-shot and fine-tuned approaches. The fine-tuned model notably achieved a more balanced classification, whereas the zero-shot approach tended to favor positive labels. Our results show that, overall, LLMs fall short of achieving satisfactory performance. Nevertheless, our qualitative inspection reveals that the models are generally capable of correctly interpreting the symbolic representation, associating each token with its intended musical concept (*RQ1*). Despite this, such understanding does not consistently lead to accurate classification of the music’s sentiment as positive or negative (*RQ2*).

Although our study had some constraints, such as limiting experiments to models under 9 billion parameters and using fixed prompts and hyperparameters, the findings offer insights into leveraging LLMs for music sentiment tasks. Future research could benefit from exploring larger models, incorporating diverse musical datasets, and experimenting with tokenization approaches tailored explicitly to musical representations, which might enhance the models’ interpretative capabilities, as indicated by previous promising LSTMs frameworks [Ferreira and Whitehead 2019, Hung et al. 2021].

Finally, while LLMs show potential for efficiently automating sentiment analysis in music, we recognize their current limitations in capturing the richness of human emotional experiences. Additionally, given that these models reflect biases from their training data, ongoing critical reflection on their outputs remains essential.

Acknowledgments

We thank the financial support of CNPq (National Council for Scientific and Technological Development), grant 307088/2023-5, FAPERJ – *Fundação Carlos Chagas Filho de Amparo à Pesquisa do Estado do Rio de Janeiro*, processes SEI-260003/002930/2024, SEI-260003/000614/2023, and CAPES (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior) – Finance Code 001. We also thank the support of the National Institutes of Science and Technology IAIA and TILD-IAR.

References

Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., Bahree, A., Bakhtiari, A., Bao, J., Behl, H., Benhaim, A., Bilenko, M., Bjorck, J., Bubeck, S., Cai, M., Cai, Q., Chaudhary, V., Chen, D., Chen, D., Chen, W., Chen, Y.-C., Chen, Y.-L., Cheng, H., Chopra, P., Dai, X., Dixon, M., Eldan, R., Fragoso, V., Gao, J., Gao, M., Gao, M., Garg, A., Giorno, A. D., Goswami, A., Gunasekar, S., Haider, E., Hao, J., Hewett, R. J., Hu, W., Huynh, J., Iter, D., and et al., S. A. J. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. *arXiv:2404.14219*.

Alves, C. F., Mozart, T. G., and Kowada, L. A. B. (2024). Emotion Recognition in Instrumental Music Using AI. In *Proceedings of the 34th Brazilian Conference on Intelligent Systems (BRACIS)*, Belém, Brasil. To appear.

Barrett, F. S. and Janata, P. (2016). Neural responses to nostalgia-evoking music modeled by elements of dynamic musical structure and individual differences in affective traits. *Neuropsychologia*, 91:234–246.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., Hesse, C., Chen, M., Sigler, E., Litwin, M., Gray, S., Chess, B., Clark, J., Berner, C., McCandlish, S., Radford, A., Sutskever, I., and Amodei, D. (2020). Language models are few-shot learners. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Copet, J., Kreuk, F., Gat, I., Remez, T., Kant, D., Synnaeve, G., Adi, Y., and Défossez, A. (2023). Simple and controllable music generation. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Davitz, J. R. (1964). *The Communication of Emotional Meaning*. McGraw Hill, New York.

Ferreira, L. N. and Whitehead, J. (2019). Learning to Generate Music with Sentiment. *Proceedings of the Conference of the International Society for Music Information Retrieval*.

Fónagy, I. and Magdics, K. (1963). Emotional Patterns in Intonation and Music. *STUF - Language Typology and Universals*, 16(1-4):293–326.

Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., and et al., A. S. (2024). The Llama 3 Herd of Models. *arXiv:2407.21783*.

Han, D., Kong, Y., Han, J., and Wang, G. (2022). A survey of music emotion recognition. *Frontiers of Computer Science*, 16(6):166335.

Helmholtz, H. v. (1954). *On the Sensations of Tone as a Physiological Basis for the Theory of Music*. Dover Publications, New York.

Huang, Y.-S. and Yang, Y.-H. (2020). Pop Music Transformer: Beat-based Modeling and Generation of Expressive Pop Piano Compositions. In *Proceedings of the 28th ACM International Conference on Multimedia*, MM '20, page 1180–1188, New York, NY, USA. Association for Computing Machinery.

Hung, H.-T., Ching, J., Doh, S., Kim, N., Nam, J., and Yang, Y.-H. (2021). EMOPIA: A Multi-Modal Pop Piano Dataset For Emotion Recognition and Emotion-based Music Generation. In *Proceedings of the 22nd International Society for Music Information Retrieval Conference*.

Hurst, A., Lerer, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., Mądry, A., and et al., A. B.-W. (2024). Gpt-4o system card. *arXiv:2410.21276*.

Jiang, A. Q., Sablayrolles, A., Mensch, A., Bamford, C., Chaplot, D. S., de las Casas, D., Bressand, F., Lengyel, G., Lample, G., Saulnier, L., Lavaud, L. R., Lachaux, M.-A., Stock, P., Scao, T. L., Lavril, T., Wang, T., Lacroix, T., and Sayed, W. E. (2023). Mistral 7B. *arXiv:2310.06825*.

Loveday, C. (2022). Why are minor chords sad and major chords happy?

Lu, P., Xu, X., Kang, C., Yu, B., Xing, C., Tan, X., and Bian, J. (2023). Musecoco: Generating Symbolic Music from Text. *arXiv:2306.00110*.

Oliveira, A., Carvalho, L., Campos, D., and Mantovani, R. (2024). Music Genre Recognition with Handcrafted Audio Features. In *Anais do XXI Encontro Nacional de Inteligência Artificial e Computacional*, pages 541–552, Porto Alegre, RS, Brasil. SBC.

Qin, L., Chen, Q., Feng, X., Wu, Y., Zhang, Y., Li, Y., Li, M., Che, W., and Yu, P. S. (2024). Large Language Models Meet NLP: A Survey. *arXiv:2405.12819*.

Riviere, M., Pathak, S., Sessa, P. G., Hardin, C., Bhupatiraju, S., Hussenot, L., Mesnard, T., Shahriari, B., Ramé, A., Ferret, J., Liu, P., Tafti, P., Friesen, A., Casbon, M., Ramos, S., Kumar, R., Lan, C. L., Jerome, S., Tsitsulin, A., Vieillard, N., Stanczyk, P., Girgin, S., Momchev, N., Hoffman, M., Thakoor, S., and et al., J.-B. G. (2024). Gemma 2: Improving Open Language Models at a Practical Size. *arXiv:2408.00118*.

Russell, J. A. (1980). A Circumplex Model of Affect. *Journal of Personality and Social Psychology*, 39(6):1161.

Santos, A., Jácome, K. R., and Masiero, B. (2021). Song Emotion Recognition: A Study of the State of the Art. In *Anais do XVIII Simpósio Brasileiro de Computação Musical*, pages 209–212. SBC.

Scherer, K. R. (1986). Vocal affect expression: a review and a model for future research. *Psychol Bull*, 99(2):143–165.

Seufitelli, D. and Moro, M. (2023). From exploration to exploitation: Understanding the evolution of music careers through a data-driven approach. In *Anais do XXXVIII Simpósio Brasileiro de Bancos de Dados*, pages 244–255. SBC.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-Art Natural Language Processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Yang, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Li, C., Liu, D., Huang, F., Wei, H., Lin, H., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Lin, J., Dang, K., Lu, K., Bao, K., Yang, K., Yu, L., Li, M., Xue, M., Zhang, P., Zhu, Q., Men, R., Lin, R., Li, T., Xia, T., Ren, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Wan, Y., Liu, Y., Cui, Z., Zhang, Z., and Qiu, Z. (2024). Qwen2.5 Technical Report.

Yepez, J., Tavares, B., Peres, F., and Becker, K. (2024). Na batida do funk: modelagem de tópicos combinando llm, engenharia de prompt e bertopic. In *Anais do XXXIX Simpósio Brasileiro de Bancos de Dados*, pages 613–625, Porto Alegre, RS, Brasil. SBC.

Yu, D., Song, K., Lu, P., He, T., Tan, X., Ye, W., Zhang, S., and Bian, J. (2023). MusicAgent: An AI Agent for Music Understanding and Generation with Large Language Models. In Feng, Y. and Lefever, E., editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 246–255, Singapore. Association for Computational Linguistics.

Yuan, R., Lin, H., Wang, Y., Tian, Z., Wu, S., Shen, T., Zhang, G., Wu, Y., Liu, C., Zhou, Z., Xue, L., Ma, Z., Liu, Q., Zheng, T., Li, Y., Ma, Y., Liang, Y., Chi, X., Liu, R., Wang, Z., Lin, C., Liu, Q., Jiang, T., Huang, W., Chen, W., Fu, J., Benetos, E., Xia, G., Dannenberg, R., Xue, W., Kang, S., and Guo, Y. (2024). ChatMusician: Understanding and generating music intrinsically with LLM. In Ku, L.-W., Martins, A., and Srikumar, V., editors, *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6252–6271, Bangkok, Thailand. Association for Computational Linguistics.