PROMPT-ENGINEERING FOR MUSIC GENERATION

Guglielmo Fratticioli, Elia Pirrello

Dipartimento di Elettronica, Informazione e Bioingegneria (DEIB), Politecnico di Milano
Piazza Leonardo Da Vinci 32, 20122 Milano, Italy
[guglielmo.fratticioli, elia.pirrello]@mail.polimi.it

ABSTRACT

This study investigates how to optimize prompts for MusicGen[1], a recent AI model capable to generate music from textual descriptions. The main objectives are: Establishing guidelines for writing MusicGen textual prompts; Evaluating the quality of music generated using two distinct approaches for textual prompts: structured, and qualitative. The methodology involves a perceptual AB test, in which participants, given a generic prompt, are asked to pick the most accurate piece between the two generated with the two different approaches. Results show that qualitative prompts produce higher quality music compared to other types, with marked preferences for certain musical genres.

1. INTRODUCTION

Prompt engineering is a critical component in the field of artificial intelligence, mainly to obtain consistent results from generative models. As AI models become increasingly sophisticated the significance of prompt engineering may face great changes from the current importance, however conducting a specific research on actual models will eventually influence future methodologies for text-to-music models optimisation. We can see how music generative models today are conditioned in the generation process by textual descriptions, while, in the future, text-based models might just serve as initial layers to interpret user prompts, with a generation process based on a model trained with music described by more significant music descriptors. Nevertheless, for the time being, the careful crafting of prompts is crucial for achieving the best outcomes.

1.1. Other researches

In other areas, such as image generation, prompt engineering has shown remarkable success. For instance, a study by Liu and Chilton (2022)[2] demonstrates the impact of structured prompts on the quality of generated images, highlighting the necessity of precise prompt formulation in achieving desired results. However it is worth noting that the efficacy of prompt engineering is directly correlated with the labelling used in the training dataset. While in Image generation exhaustive description dataset are usually available in Music captioning (MusicCaps[3]) tracks description are usually shorter and organised with tags.

1.2. Musicgen

MusicGen is an AI model designed to generate original music from a text description, nominally is capable of understanding natural language producing great results with different prompt types. Among several models tested, including MusicLM and AudioGen, MusicGen provided the most accurate but also consistent results, making it the focus of this study. MusicGen's ability to generate high quality music that closely aligns with the provided prompts makes it an interesting subject of study for prompt engineering.

1.3. Objective

The primary objective of this research is to establish guidelines for writing prompts that MusicGen can effectively understand, resulting in more coherent musical outputs. Specifically, we aimed at determining whether having a rigid syntax structure in the text prompt is preferable than a prompt based on qualitative descriptions. To achieve this, we conducted a perceptual AB test where participants were presented with a general song description and asked to choose between two audio clips, each of which was generated following one of the two types of prompt writing approach under test.

This study not only seeks to provide insights into the best prompt engineering practices for MusicGen but also contributes to the broader understanding of text-to-music generation. In the process of exploration of MusicGen we have generated several musical tracks with various kind of prompt, this made us elaborate several hypothesis on the best prompt writing procedures, given the current time and resources needed for the project we had the chance to test the effectiveness of prompt syntax structuring versus qualitative descriptions but eventually we will discuss some of the other hypothesis in the next sections.

This research addresses a significant gap in the field, currently there are no comprehensive public studies on the best practices for writing prompts in music generation. Lastly, we highlight the importance of further research, there is an audience and a need for advancements in AI-driven music generation.

2. BACKGROUND

The MusicGen model[1] is an autoregressive transformer-based decoder that generates music conditioned on text or melody inputs. It utilises a language model over quantized units derived from an EnCodec audio tokenizer that is capable of high-fidelity audio reconstruction from low frame rates. EnCodec employs Residual Vector Quantization (RVQ), producing several parallel streams of discrete tokens from different learned codebooks.

The text encoder, based on the T5 model, transforms text inputs into useful embeddings. These embeddings are then processed by the transformer-based decoder, which generates sequences of audio tokens from the text embeddings. This audio tokens are gathered in codebooks that represent the semantic correlation that allows for consistent music generation. Generally we can say the audio encoder/decoder component, using the EnCodec framework, compresses audio waveforms into tokens, which the model uses to generate music. The audio decoder reconstructs the waveforms from these tokens.

Considering the current state of development in music generative AIs, we have also identified additional text-to-music alternative models. MusicLM[4] and Jukebox[5] are two of the most popular alternative models but achieve worse performance due to the architecture difference and the most likely more exhaustive proprietary dataset MetaAi used to train MusicGen. In recent days, after we conducted the tests, there were released two new text to music services: Suno[6] and Udio[7] seems to provide more coherent and advanced results than MusicGen. However this two systems are fully private and does not provide any public paper making them less attractive for a public research study.

3. TEST METHODOLOGY

3.1. Prompt types definition

To evaluate the effectiveness of different prompt types in generating music with MusicGen, from a generic song description, we defined and tested two types of prompts: structured and qualitative. Here are the specific examples for each type:

1. Generic Description:

   These prompts provide a general description of the song. For example:

   - "A piano-driven ballad with a simple yet powerful melody, backed by strings and percussion. It gradually builds in intensity, culminating in a soaring climax."
   - "An 80's inspired synth-pop track with a pulsating beat and a catchy synth riff. It combines electronic and organic elements, including synthesizers, drums, and bass guitar."

2. Structured Prompts:

   These prompts follow the syntactic form:
   Subject *[Sub]* + Action *[Act]* + Quality *[Qlt]*

   - **A grand piano** *[Sub]* opens softly, **playing** *[Act]* **haunting** *[Qlt]*, **string chords** *[Sub]* **swell** *[Act]* in the background.
   - **Synth keyboards** *[Sub]* **create** *[Act]* a **pulsating** *[Qlt]* melody, **drums** *[Sub]* pound with a **driving beat** *[Qlt]*

3. Qualitative Prompts:

   These prompts express emotional and atmospheric aspects of the music. For example:

   - "Generate a captivating pop song. The instrumentation should be lush and dynamic, featuring rich piano chords, warm strings, and subtle electronic elements to create an atmosphere of intimacy and grandeur. Ensure the chorus is infectious and unforgettable, with a build-up of energy that leaves a lasting impact."
   - "Generate a pop song with pulsating synths, catchy melodies, and an infectious beat. The song should radiate an electrifying energy, capturing themes of longing, desire, and exhilaration. Create a dynamic arrangement that builds tension and release, culminating in an irresistible chorus that sticks in the listener's mind."

3.2. Audio Generation

The audio clips were generated using the MusicGen model hosted through a Gradio webapp onto a Google Colab [8]. We used the following parameters: model=large, duration=15 seconds, top k=250, top p=0, temperature=1, and classifier-free guidance=3.

We have strong confidence that the specified selected length of the audio generation does not affect the quality of the output. In general a well written prompt perform with the same mood and coherency with all of the possible length (up to 60 min as generation limit).

1. Prompt Writing:

   We wrote the prompts in the two formats: structured and qualitative. We provide the full list of the prompt [9].

2. Generate audio with MusicGen:

   - Load the model choosing the MusicGen-large version.
   - Set Generation parameters.
   - Insert prompt in the Gradio text entry.
   - Generate audio.

   Note that each prompt was run three times, and the best output was selected based on coherence and relevance to the prompt.

3.3. Perceptual AB Test

The perceptual AB test was hosted in a Google Form, we provided the total of 12 queries, for each query, participants were presented with a generic description of the songs and two links to the audio clips (A and B), each generated with one of the two different prompting approach. The participants were asked to choose the audio clip that best matched the song description without knowing which prompt type was used. Participants varied in age, gender and music knowledge. This choice was done to obtain results that represented a broader set of people, so actual music listeners, and not necessarily musicians.

1. Participant Instructions:

   Participants were instructed to listen to each pair of audio clips, then, select the one that mostly matched the description for them. They were informed that the clips were generated based on different prompts types but were not told which clip corresponded to which prompt type.

2. Data Collection:

   Participants' choices were recorded with Google Form and exported in a *.csv* sheet, we avoided duplicate answers by requesting the personal email. The sheet was then analized to determine the preferred prompt type.

4. RESULTS

The analysis of the perceptual AB test reveals several key insights into the effectiveness of structured (Option 1) and qualitative (Option 2) prompts in generating high-quality music using MusicGen.

These results along with the related audio files are available in the paper's repository [9]

	Option 1 (Structured)	Option 2 (Qualitative)
Test 1 (Rock)	28.57%	71.43%
Test 2 (Rock)	32.65%	67.35%
Test 3 (Rock)	42.86%	57.14%
Test 4 (Rock)	73.47%	26.53%
Test 5 (Jazz)	10.20%	89.80%
Test 6 (Jazz)	75.51%	24.49%
Test 7 (Jazz)	73.47%	26.53%
Test 8 (Jazz)	20.41%	79.59%
Test 9 (Pop)	26.53%	73.47%
Test 10 (Pop)	30.61%	69.39%
Test 11 (Pop)	42.86%	57.14%
Test 12 (Pop)	53.06%	46.94%

Table 1: Percentage of votes for each test, winner highlighted

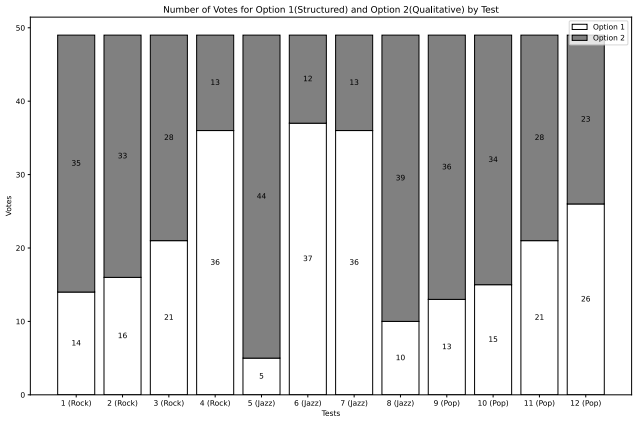

Figure 1: Number of votes for kind of prompt

4.1. Overall Preferences

The test results showed that qualitative prompts (Option 2) were generally preferred over structured prompts (Option 1). Out of the total votes:

- Structured prompts received 42.52% of the votes (250 votes).
- Qualitative prompts received 57.48% of the votes (338 votes).

In terms of the number of victories across different tests:

- Structured prompts won 4 times 33.33%.
- Qualitative prompts won 8 times 66.67%.

4.2. Patterns and Observations

Despite Option 2 winning more tests, the margin of preference in total votes was not extremely large, indicating that both types of prompts had their strengths in different contexts. The detailed analysis of specific genres showed that:

1. Rock Music:
   - Option 1 won 1 test.
   - Option 2 won 3 tests.
2. Jazz Music:
   - Option 1 won 2 tests.
   - Option 2 won 2 tests.
     (we have an even result with more polarized scores)
3. Pop Music:
   - Option 1 won 1 test.
   - Option 2 won 3 tests.

4.3. Results Findings

The examination of the results suggests that the preferences varied significantly across different musical genres. In jazz music, the results were particularly balanced indicating that a structured prompt approach gives almost the same results as the qualitative one. However for the Rock and Pop genres the victory of a qualitative description is significative. This variability might also indicate that in model's training dataset for jazz songs were described in a more objective and schematic way than Pop and Rock tracks or that the trained dataset for jazz music is not as exhaustive as the Rock and Pop one.

However we should address that the audience might be less used to listening and judging jazz music, that has a significant smaller audience than Pop or Rock. This suggest that in that section (Jazz) results may have been given in a more random fashion, leaving room for the possibility of a broader validity of our assumption and result.

4.4. Implications and Hypothesis

The analysis indicates that while qualitative prompts are generally more effective for generating high-quality and more coherent music, structured prompts seems to still hold a value in the jazz genre. The mixed results in jazz music highlight the importance of considering a music genre audience as well as the training dataset characteristics. Unfortunately the latter was unavailable for the current model and we could not perform a more exhaustive analysis on the matter.

In conclusion, the perceptual AB test results affirm the overall preference for qualitative prompts but also underscore the different performance across musical contexts, emphasizing the need for a tailored approach in prompt engineering analysis for music generation.

5. CONCLUSIONS

This study set out to explore the optimal strategies for crafting prompts to be used with MusicGen, a recent AI model designed for generating music from textual descriptions. Our research focused on comparing structured prompts with qualitative prompts, assessing their efficacy through a perceptual AB test involving human participants.

5.1. Results review

The findings revealed a clear preference for qualitative prompts across most musical genres, with participants favoring the music generated from qualitative descriptions in 66.67% of the tests. This preference was particularly pronounced in the rock and pop genres, where qualitative prompts significantly outperformed structured ones.

Interestingly, the results for jazz music were more evenly split, with structured prompts performing as well as qualitative prompts. This suggests that certain genres might benefit from a more objective and schematic prompt structure probably due to the nature of the training data. The mixed results for jazz also highlight the potential influence of the training dataset's of the subjective nature of musical preference.

5.2. Prompt engineering Guidelines

Our research underscores the overall better performance of writing qualitative prompts when using Musicgen. It is advised to focus on the emotional and atmospheric aspects of the desired output. Effective qualitative prompts should vividly describe the mood, energy and emotive aspect of the music, using evocative language and precise adjectives.

A good qualitative prompt is the following:

"catchy pop song with lush, dynamic instrumentation that creates an intimate and grand atmosphere, featuring rich piano chords".

On the other hand it is discouraged to rely on precise description of the musical scenes in which there are described some musical subjects (guitar/piano/strings) that perform some actions as we have tested how this kind of sentences seems to produce tracks that are perceived less coherent and enjoyable.

Structured Prompts like the following are discouraged:

"Funky track. bright synths create an uplifting melody, bass grooves steadily, drums establish a danceable beat, harmonies enrich the chorus, electronic elements add a futuristic vibe. "

The users of the Musicgen model should be aware that usually the available useful context length of the prompt is about 4-5 phrases *(usually those models start generating unpredictable results for very long prompts)*. In this little space we have highlighted how it is more important to focus on choosing wisely the particular adjectives and general descriptions of the music we want to obtain rather than writing in a descriptive way the music instruments we want to be present and their action.

Additionally this means that prior musical knowledge about composition, orchestration and music theory is way less important than a good emotional understanding and evocative language writing skill in the context of Musicgen prompt engineering.

In conclusion, this study contributes valuable insights into prompt engineering for music generation, offering a foundation for future research and practical guidelines for enhancing AI-driven music composition. The preference for qualitative prompts suggests a promising direction for developing more intuitive and coherent AI-generated music, although further exploration is needed to refine these findings across a broader range of genres and for different generative models.

6. REFERENCES

1. Jade Copet, Felix Kreuk, Itai Gat *Simple and Controllable Music Generation*, arXiv:2306.05284
2. Vivian Liu, Lydia B. Chilton *Design Guidelines for Prompt Engineering Text-to-ImageGenerative Models*
3. Andrea Agostinelli, Timo I. Denk *MusicLM: Generating Music From Text*, arXiv:2301.11325
4. Andrea Agostinelli, Timo I. Denk *MusicCaps*, https://www.kaggle.com/datasets/googleai/musiccaps
5. Prafulla Dhariwal, Heewoo Jun *Jukebox: A Generative Model for Music*, arXiv:2005.00341
6. https://suno.com/
7. https://www.udio.com/
8. https://colab.research.google.com/github/camenduru/MusicGen-colab/blob/main/MusicGen_colab.ipynb
9. https://guglielmofratticioli.github.io/MusicGen-Prompt-Engineering/