Automated Music Composition with a Large Language Model – An Exploration

Joel P. S. Breit
University of Pittsburgh

December 2023

Abstract

This project explores the potential for an LLM, specifically OpenAI's GPT-4-Turbo via its assistants API, to perform the task of creating text-based musical compositions from text-based input. It focuses on the creation of valid and pleasing short tunes using ABC notation and explores the technical complications involved. A comparison of potential techniques is discussed, and the results of the utilized approach are evaluated across 20 compositions resulting in average output of fine to good quality. Finally, several directions for future research are discussed.

1 Introduction

1.1 Background

As of 2023, the capabilities of large language models (LLMs) have far outpaced those of generative music AI. While theoretically coherent text generation is a much more difficult task than coherent music notation, few options are available for musicians who want to generate music notation from text descriptions. At the same time, current LLMs are capable of eloquently describing the core ideas of music theory and composition and can describe music in text. This research seeks to bridge the gap between the nominal understanding of music in LLMs and their ability to generate musical scores.

1.2 Objectives

The objective for this project was to utilize the intelligence of an LLM to easily generate pleasant and customizable musical compositions. This objective consisted of many subgoals including the following:

- Restructure user input for optimal use with an LLM
- Prompt the LLM in a way that utilizes the wide range of its musical understanding

- Generate actual music, not just descriptions of music
- Get output from the LLM that makes the music legible
- Provide guide rails to guide the LLM towards composition tasks that it is capable of
- Assure that the musical output is technically well-formed (lacking detrimental mistakes)
- Translate the LLM's output into actual audio/visual output
- Maintain the individuality of user input by creating music that specifically matches their input
- Generate genuinely new music
- Generate music of substantial length
- Generate music with substantial texture
- Generate coherent music
- Generate music with interesting melodic direction
- Generate music with interesting rhythmic structure

2 Methods

2.1 Picking an LLM

In order to address the goals of this research, a large language model and a text-based music notation style needed to be selected. For the large language model, GPT-4 was chosen primarily due to its reputation as being the state-of-the-art language model in October-December of 2023 when this research was conducted [1]. This research also likely could have been conducted using other LLMs such as Claude from Anthropic or Llama from Meta, but such exploration fell outside the scope of this research and preliminary testing showed that other models often struggled to produce even coherent outputs.

2.2 Picking a Notation Style

The following formats were tested for text-based music notation style:

1. ABC notation
2. (note, duration) pairs (e.g., ("A4", 2) where "A4" represents the musical note A4, and 2 represents the duration of 2 beats)
3. Python code utilizing the music21 library

4. MusicXML

In test runs, MusicXML was too verbose and resulted in invalid syntax and a dearth of musical content relative to the volume of text required. Similarly, `music21` code frequently failed to compile making it difficult to test, and (note, duration) pairs were difficult to translate into a usable format. ABC notation and (note, duration) pairs provided the best preliminary results. Next, 8 identical prompts were used to compare the quality of compositions generated using ABC notation and (note, duration). No formal analysis was conducted on these results, but the resulting compositions can be found at `https://www.OrchestrAI.site/juxcompose`. Based on the above, it was determined empirically that ABC notation produced the best results.

2.3 Creating an Interface

After making these determinations, an interface needed to be developed for replicable usage of the LLM and for translation of the LLM's text-based output. To this end, a web application was created utilizing web development languages and libraries including the open-source ABCjs library which allows for in-browser rendering of audio-video enabled sheet music from ABC notation. For each composition, a different mood was requested to be expressed in the music. Upon submission, the prompt *"Compose a tune that expresses the following mood: [provided mood]"* was sent to the API. After completion, the web application extracted the ABC notation, provided it in an editable text field, rendered the ABC notation as playable audio-video sheet music, and separately displayed any accompanying text that the model generated.

2.4 Customizing the Model

For the purpose of this project, a custom GPT-4-Turbo model was created using OpenAI's Assistants API which was available in beta at the time. For customization, a custom prompt was provided which is available in the appendix of this report. The prompt discouraged common failures such as mismatched voices and encouraged proven competencies such as including chords and composing multiple sections.

In addition to a custom prompt, 3 reference materials were added to provide a targeted knowledge base for the GPT. The first document was a web archive of the ABC notation Standards provided at `https://abcnotation.com/wiki/abc:standard:v2.1`. The other two documents were transcripts generated by GPT-4 on what makes a good melody and what makes good harmonization. The knowledge represented in these documents originated entirely from GPT-4.

2.5 Evaluation

In order to evaluate the results of this project, 20 compositions were generated with varying request parameters and each was evaluated on 11 metrics: grade,

notation, melody, harmony, chords, mood, lyrics, text, measures, duration, and voices. Measures was the number of musical measures written in the output, duration was the time that the tune took to play in the ABCjs interface, and voices was the number of voices (1 through 4) included in the composition. All other metrics were assessed qualitatively according to the following rubric:

	5 (Outstanding)	**4 (Proficient)**	**3 (Satisfactory)**	**2 (Needs Improvement)**	**1 (Unsatisfactory)**
Overall Quality	Exceptional composition; would confidently claim authorship.	High-quality composition; reflects a strong understanding of the task.	Adequate composition; suitable for sharing but lacks refinement.	Subpar composition; requires significant revision to be acceptable.	Poor composition; contains fundamental issues, not suitable for presentation.
Notation	Flawless notation with no observable errors.	Minor, non-critical notation errors that do not impede interpretation.	Noticeable notation errors, but they can be easily corrected.	Critical notation errors that compromise the legibility of the composition.	Significant notation errors that obscure the structure and intent of the composition.
Melody	Exemplary melody, demonstrating clear voice leading and rhythmic variation.	Strong melody with sufficient variation and sense of direction.	Adequate melody, though lacking in variation or rhythmic complexity.	Lacks a discernible melody or variation.	Deficient melody, exhibiting structural incoherence or invalid musical constructs.
Harmony	Exceptional harmony, contributing positively to the overall aesthetic and demonstrating intent.	Generally good harmony that supports the overall composition.	Satisfactory harmony with minor flaws or lack of direction.	Weak harmony that detracts from the composition's quality.	Discordant harmony that severely impacts the listenability of the composition.

Chords	Outstanding chord progression; reflects advanced musical understanding.	Chord progressions that fit well and provide structure to the composition.	Adequate chord progression with minor errors.	Chord progressions that are inconsistent across voices.	Chord progressions that conflict with the key, creating dissonance.
Mood Alignment	The composition perfectly aligns with the intended mood.	The composition appropriately reflects the intended mood.	The composition generally fits the intended mood, but lacks precision.	The composition does not effectively convey the intended mood.	The composition conflicts with the intended mood, detracting from the overall effect.
Lyrics	Lyrics are excellently crafted and match the notes with precision.	Lyrics are well-written but may slightly misalign with the notes.	Lyrics are adequate but require refinement to better match the composition.	Lyrics detract from the composition and would benefit from removal.	Lyrics significantly hinder the overall quality of the composition.
Descriptive Text	The descriptive text accurately captures the essence of the composition.	The descriptive text is mostly accurate but may overstate or misinterpret some elements.	The descriptive text is acceptable but includes at least one notable error.	The descriptive text contains several inaccuracies.	The descriptive text introduces confusion or misrepresentation within the composition.

3 Results

The best way to understand the results of this research are to go and view them on its dedicated website: `https://www.OrchestrAI.site/compose`. The site allows visitors to generate new compositions, view them as sheet music, listen and watch them be played back, and edit them. It also allows for saving and sharing generated compositions. A curated selection of successful compositions generated by the model can also be viewed at `https://www.OrchestrAI.site/portfolio`.

For a more analytical review of the results, rubric-graded results are provided below. Average overall quality fell between fine (3) and good (4) quality. The

most satisfying aspects, in order, were the text descriptions, chords, mood, and melody, while harmony, notation, and lyrics left room for improvement. Notably, some compositions turned out well in nearly all aspects while others were crippled by failure to create clean, coherent syntax.

3.1 Evaluation Results

+4cm

	Grade	Notation	Melody	Harmony	Chords	Mood	Lyrics	Text
Tune 1	4	4	5	4	4	5	4	5
Tune 2	2	2	4	3	4	5	4	5
Tune 3	4	4	4	3	4	3	4	3
Tune 4	4	4	3	3	4	3	4	5
Tune 5	1	1	2	2	4	3	1	5
Tune 6	1	2	2	2	4	3	2	5
Tune 7	3	5	3	2	4	3	4	4
Tune 8	3	4	4	4	2	4	4	5
Tune 9	4	4	3	3	3	4	4	4
Tune 10	3	3	5	4	4	5	3	5
Tune 11	4	2	5	5	5	5	5	5
Tune 12	4	5	4	4	4	5	4	5
Tune 13	4	5	4	4	5	5	4	5
Tune 14	5	5	5	5	5	5	5	5
Tune 15	5	4	5	4	5	5	3	5
Tune 16	4	3	5	5	5	5	5	5
Tune 17	4	4	4	5	5	5	4	5
Tune 18	4	4	5	4	5	5	5	5
Tune 19	4	4	5	4	3	5	4	5
Tune 20	4	4	4	3	4	4	4	5
Average	3.5	3.7	4.1	3.7	4.3	4.1	3.0	4.6
C.I.	0.5	0.5	0.4	0.5	0.5	0.3	0.9	0.3
Failure Rate	15%	20%	10%	17%	7%	0%	29%	0%

Table 2: Evaluation of 20 Tunes with Multiple Metrics

4 Discussion

4.1 Limitations

It was difficult to conduct this research without noticing several limitations of the methodology. Mentioned above are several reasons for the use of ABC notation; however, ABC notation has limitations of its own. ABC notation was developed, and is largely used for annotation of folk music, and like nearly all public sources of written music, ABC notation datasets primarily consist of music written at least a century ago. This means that any ABC notation that modern LLMs happen to be trained on is highly skewed towards simplistic and, frankly, uninteresting music. This is also at least partially due to the fact that if music is particularly complicated or intricate, it is unlikely to be represented in ABC notation as opposed to other formats. This may result in LLMs being biased towards generating simplistic musical structure when prompted to write in ABC notation.

While it can be shown that LLMs retain an incredible depth of knowledge of music theory, it is difficult to demonstrate that knowledge through the ABC notation medium. A large portion of the tweaking phase of this research was focused on improving the interestingness of generated compositions and particularly on generating rhythms other than quarter notes and musical direction other than stepwise motion. It is also the case that while GPT-4 had no trouble describing the rules and conventions of good voice leading (the rules that lead to good, or at least not unpleasant, harmonies), it frequently failed to implement these rules in its compositions.

Some other notable and recurring issues included filling measures with too many/too few beats, mismatching beats across voices, including contradicting chords in different voices, failing to put notes in the correct octave (likely due to the convoluted nature of octave representation in ABC notation), crossing voices (an easy-to-spot voice leading mistake), failing to line lyrics up with notes (this is an admittedly difficult task), and otherwise failing to generate consistency across voices.

While there are many limitations to this research, it is the first of its kind and provides a strong foundation for future research. The resulting tool created by this research is also likely the first to accomplish any of the following: generate playable music from unrestricted text input, generate viewable music notation using a machine learning system, generate creative music output not based on a musical corpus, allow unrestricted input for music generation, and allow for composition editing and extension.

4.2 Further Work

Ultimately, while this research attempts to provide a high-level view of the limitations and potential for using large language models in music composition, it provides neither a sufficiently deep inspection of these limitations nor a comprehensive exploration of the possible approaches that this line of research could take. There are many facets of large language models' understanding of written music that could be further scrutinized as well as many other applications of these models in music composition that could be explored.

4.2.1 Prompt and Resource Adjustments

Considerable progress was made in the quality of the output by tweaking the prompt that was provided. Some adjustments have already been implemented since this research was conducted, and there are likely further improvements that can be made in this easy-to-update domain. Marginal improvements are likely to result from a sufficient exploration of prompting techniques.

Some of the more surprising improvements came from providing resources to the custom GPT. Furthermore, the sources provided were selected largely for expedience and were in fact produced directly by the model itself. These resources likely helped in priming the results. Future research may explore iteratively providing professionally produced resources on music theory and notation

standards.

4.2.2 Few-Shot Learning

It was peripherally observed that the model was able to produce better results in contexts where successful examples were provided. While a few-shot learning approach was not implemented in this research, it is likely to be a fruitful avenue for future research. However, given the importance of creativity in music generation, it is possible that this approach could limit the variability of output and thus limit visibility into what the best case scenario generations could be.

4.2.3 Fine-Tuning

Notably, this research did not attempt to test a fine-tuned model. When this research was conducted, fine-tuning was not available for GPT-4, and other language models for which fine-tuning was available proved significantly less capable at generating music. While training data in the form of ABC notation may be a limiting factor, fine-tuning will likely provide a significant advantage in future research.

4.2.4 Synthetic Training Data

A novel approach that this method of music generation allows is the creation of synthetic training data. Standardized training data for music research is among the most difficult to come by, and generating new compositions has, perhaps until now, been infeasible outside of the most well-funded research efforts. Many current generative music systems have been limited to music that is a century or more old, is from low-quality recordings, or that has limited labeling data associated with it. Indeed, some of the best generative music systems are trained on just a few hundred compositions. Music generated by large language models and verified by human judges could be used to create a new dataset of music that has ample metadata, and is public domain and original by nature. This data could be useful for future iterations of this flavor of music generation as well as for other music generation tasks.

4.2.5 Other Text-Based Music Notation Formats

In future work, it may be worth revisiting other text-based music notation formats. This could include music libraries for computer languages like `music21`, file formats like MusicXML or MIDI, or other text representations. There is a wide space to be explored in this domain, and more research could be done on analyzing what text representations large language models are best at comprehending and best at expressing.

4.2.6 Comparative Evaluations

Providing training data based on `music21`, or (note, duration) pairs may improve reliability in which case these formats may prove superior if LLMs are more adept at reasoning in code (quite possible given the comparative abundance of code training data compared to music training data) or in small snippets which makes sense *a priori*. It may also be prudent to create a converter that translates syntax like (note, duration) pairs for ease of experimentation. Similarly, it would be worth exploring how other large language models perform at similar tasks when compared to GPT-4. Other models can be fine-tuned which is likely to provide an important advantage with enough care for crafting good training data. Furthermore, a direct comparison of evaluations between models may be more enlightening than a single evaluation as conducted above.

4.2.7 Future Models

Finally, as large language models continue to improve, these capabilities may need to be reassessed. It is possible that the limitations of this current iteration may soon be addressed by better models with improved reasoning capabilities.

Appendix

Custom Prompt

OrchestrAI is an advanced AI composer specializing in ABC notation, focusing on producing longer musical pieces, with a beginning, middle and end, with a strong emphasis on technical music theory. It avoids creating short, simplistic tunes and ensures structural consistency by preventing mismatched beams and inconsistent pickup beats across voices. It does not include percussion instruments. OrchestrAI crafts compositions with rhythmically varied and harmonically rich elements, harmonizing voices, matching chords, and incorporating counter melodies and bass lines. It uses music theory to enhance compositions and includes lyrics when appropriate. It sets an appropriate tempo and sets the composer to "OrchestrAI". The GPT refrains from unnecessary explanations of ABC notation and does not include any text after the ABC notation, and focuses instead on coming up with interesting music theory ideas and then implementing them in the ABC notation section.

References

[1] Chatbot Arena Leaderboard. `https://lmsys.org/blog/2023-05-03-arena/`