Computer Science
&
Data Science

Capstone Report - Fall 2023

Melody-Conditioned Lyrics Generation Using Large Language Models

Wenhao Xu,
Ouwen Jia

supervised by
Hongyi Wen

Declaration

I declare that this senior capstone was composed entirely by myself with the guidance of my advisor, and that it has not been submitted, in whole or in part, to any other application for a degree. Except where it is acknowledged through reference or citation, the work presented in this capstone is entirely my own.

Preface

In this capstone project, we embark on an exploratory journey to harmonize the art of lyricism with the science of machine learning. As a team of dedicated researchers with a profound interest in artificial intelligence and music composition, we are driven by the challenge of creating an innovative system capable of generating lyrics that resonate with melodies. Inspired by the intricate relationship between words and notes, our work is crafted for musicians, producers, and AI enthusiasts who share our passion for pushing creative boundaries. This project stands as a testament to the potential of AI in enhancing human creativity, paving a new path for melody-conditioned lyrics generation.

Acknowledgements

We extend our deepest gratitude to Professor Hongyi Wen for his invaluable guidance and mentorship throughout our capstone project. His weekly meetings, insightful tips, and unwavering support were pivotal in keeping our project on track. Additionally, we are immensely thankful to Professor Gus Xia for sparking our interest in the intersection of computer science and music. His expertise and generous assistance, despite not being our direct supervisor, greatly contributed to the success of our project.

Our acknowledgments would be incomplete without mentioning Jesse Yuan's exceptional emotional and financial support throughout our project. Heartfelt thanks go to Jiayun Qiao and Barry Li for their companionship and support in our times of need. We are equally grateful to all the friends who stood by us, the professors who imparted their knowledge, and our parents, whose unwavering support made our education at NYU Shanghai possible. Lastly, a special thanks to NYU Shanghai's basketball court, a place where many of our ideas and strategies were born and nurtured.

Abstract

This report addresses the challenge of melody-conditioned lyric generation using large language models, an intriguing intersection of AI and creative expression. The problem is significant due to the complex interplay between lyrical content and musical melody, where generating coherent and emotionally resonant lyrics remains a difficult task. Our approach leverages state-of-the-art AI techniques, including GPT-4, LSTM, and RNN models, to produce lyrics that complement specific melodies. The effectiveness of our solution is demonstrated through rigorous evaluation metrics, showing promise for AI-assisted artistic creation.

Keywords

Capstone; Computer Science; Data Science; NYU Shanghai; Melody-Lyrics Alignment; Large Language Model; Multi-Model Large Language Model; GPT-4; GPT-3; Melody-conditioned; Lyrics generation; Alignment metrics; LSTM; RNN

Contents

1 Introduction

Our investigation into melody-conditioned lyric generation unfolds a multifaceted challenge that lies at the crossroads of artificial intelligence and musical artistry. With a focus on the under-explored aspect of aligning lyrical content with melodic structure, this work pioneers the use of large language models like GPT-4 for this purpose. The report presents an evolved set of objectives that stem from a deepened understanding of the problem space, reflecting progress in in-context learning and model innovation. The contributions of this project are twofold: it advances the technical methodology for generating music-aligned lyrics and provides a comprehensive framework for evaluating the qualitative aspects of the generated content. Through rigorous experimentation and creative inquiry, we endeavor to establish a new benchmark in the synergy between AI-generated text and music composition.

2 Related Work

2.1 Melody-Conditioned Lyric Generation

The generation of lyrics under the guidance of their corresponding melody, or the melody-lyrics alignment, is an essential key to success and enjoyment during the song's singing. And it has always been a focal topic of interest for music professionals and computer scientists.

Over recent years, the field of melody-conditioned lyric generation has undergone a substantial transformation, primarily catalyzed by the emergence of Large Language Models (LLMs), such as ChatGPT. Prior to the ascendancy of LLMs, lyrics generation predominantly relied on the Seq2Seq models or transformers. In 2018, Watanabe experimented an RNN-based lyrics language model with a large collection of melody-lyrics aligned data from Japanese songs, which could be one of the earliest rewarding trials in this field [1]. In 2019, Xu built the multi-channel Seq2Seq model with RNN and Bi-LSTM for the syllable-concerned Chinese lyrics generation with semantic encoding [2]. Later in 2020, Bolcato also proposed a conditioned Recurrent Neural Network architecture for melody and lyrics generation [3]. Expanded from the basic RNN model, Chen accomplished the lyrics generation using a Sequence Generative Adversarial Network (SeqGAN) given the melody as input in 2020 [4]. In 2021, Huang and You, another group of researchers focused on Chinese lyrics generation, tried support vector regression (SVR) and the Seq2Seq model under the guidance of notes and melody emotions [5].

With the advent of the Generated Pre-trained Transformer (GPT) in these years, researchers in lyrics generation have started to cast increasing focus on the GPT's mighty power of manipulating text/audio/image and generating outputs with unexpectedly high quality. Relevant research will be unfolded in the following part.

2.2 Generated Pre-trained Transformer

GPT has been widely used to generate lyrics or evaluate the quality or alignment of the lyrics. The primary function of GPT when generating lyrics is to generate human-like text that resembles song lyrics. GPT, a type of neural network model, is trained on large amounts of text data, which enables it to understand patterns, styles, and structures commonly found in lyrics. When given a seed or prompt, GPT can generate coherent and contextually relevant lines of text, which in the context of lyrics, can form verses, choruses, or other song sections.

The DeepLyrics created by Stanford students, Tian and Yang, apply GPT2 for lyrics generation without inference with melody [6]. They used the GPT-2 medium model before fine-tuning as a baseline and adopted the official Huggingface tutorial and Prefix implementation to finetune the GPT-2 model. Furthermore, in 2023, Zhang and Lasocki used ChatGPT-3.5 to evaluate the quality of their melody-conditioned generated lyrics from a fine-tuned Google CANINE language model [7]. They argued that by clarifying the characteristics of their text input, ChatGPT focuses more on the correctness and quality of the syllable-level split lyrics, hence giving higher scores on their model which leverages the power of the language model.

Therefore we were considering applying NExT-GPT in our project. NExT-GPT is an any-to-any multimodal LLM. Wu and Fei connected an LLM with multimodal adaptors and different diffusion decoders, enabling NExT-GPT to perceive inputs and generate outputs in arbitrary combinations of text, images, videos, and audio.[8] Therefore we can adopt the "X+text" to "text" generation to make our output more precise. However during the process of adopting the NExT-GPT, OpenAI announced its newest version of GPT-4 in November 2023 so we changed to run our tests based on GPT-4. Version 1.0.0 GPT-4 is also an any-to-any multimodal LLM and it has better performance than NExT-GPT[9].

2.3 Prompts

Though we know how the GPT works and its function, we should also know how to use prompts to control GPT. Prompts in the context of GPT refer to input text or instructions, together with examples of input and its desired output, provided to the model to generate for a specific given input. And in-context learning involves adapting a pre-trained model to perform better on specific tasks or in specific domains by fine-tuning it on relevant data, thus leveraging the model's pre-existing knowledge and enhancing its capabilities for particular use cases.

In 2021, Liu and Yuan analyzed prompt engineering, which is the process of creating a prompting function that results in the most effective performance on the downstream task [10]. The Tuning-free Prompting introduced in their study could be a preferred path to fulfill our task on the GPT family. Later in 2023, Wang and Jiang presented a framework known as "Prompt Diffusion," which offers a novel approach to enable in-context learning [11]. We have raised the thoughts that will it influence the output quality if the ground truth is vague or uncertain, since the collected song may not obtain the label for melody-lyrics alignment score in our dataset. Whereas Min and Lyu proved in 2022 that the model counter-intuitively does not rely on the ground truth input-label mapping provided in the demonstrations as much as we thought [12]. Therefore we show the prompts we used in the baseline with GPT-3 and in the testing part with GPT-4. With prompts and the dataset Dali prepared, a large dataset encoding songs with lyrics and notes [13], we should measure whether our output and the examples are aligned in the next step. Therefore alignment metric is essential.

2.4 Alignment Metric

Although the alignment between melody and lyrics, a critical facet of lyrics generation, has been applied as one of the conditions in both the Seq2Seq model and GPT by many researchers, it still remains a relatively understudied domain. This is partly because, first, the existing perfectly aligned data is sparse, like Watanabe using only 1k aligned pairs in his research [1]. Second, melody information has frequently been pre-tokenized onto the level of the syllable at the cost of losing other latent yet essential correlations with lyrics [2] [14]. Meanwhile, other musical conditions, such as genre or lyrical quality, take precedence over the fundamental requirement of harmonizing lyrics with melodies.

Retrospectively viewing the method being used to extract, encode and evaluate the melody-lyrics alignment, several important rules could be concluded. From Nichols's study on the link between lyrics and melody in pop music in 2009 [15], the duration of notes is vital to the rhyme of singing, where longer notes should match with stressed syllables of tokens and vice versa. Additionally, Watanabe also found that within the boundary of lines (BOL), the number of notes should be equal to the number of syllables in order to enhance the melody-lyrics alignment [1]. Pitch contour also matters as its variation in both lyrics and melody should be positively correlated [15] [16]. Other ordinary evaluations of the generated text quality could also be incorporated into the overall measurement, like BLEU or Test-set Perplexity (PPL) [17].

In our project, we will experiment with the spectrum representation of the melody after applying the Fourier Transform, expanding the input of GPT-4 to fulfill the "text+X" to "text" generation, in which X refers to any other data type apart from regular text/audio/image input. In conclusion, we are committed to summarizing and contributing to the measurement of melody-lyrics generation for the computer music community.

3 Solution

Our capstone project establishes an innovative architecture for a melody-conditioned lyric generation as Figure 1 shows, leveraging a multi-model large language model approach with a focus on generated lyrics quality and measurement precision. The architecture of our capstone project mainly contains six parts, which are Data Subsampling, Data Preprocessing, Baseline Models, Image Data Construction, Multi-Model Large Language Model, and Measurement.

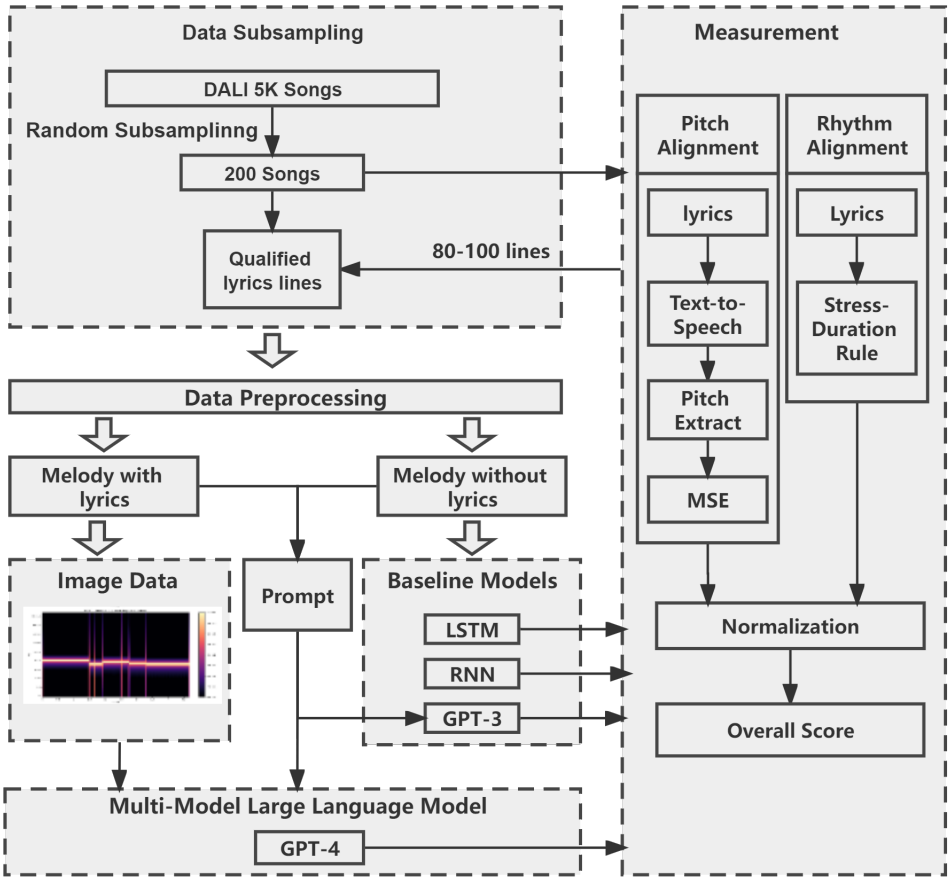

Figure 1: Architecture of our melody-conditioned lyrics generation

3.1 Data Preparation

Our architecture begins with a data subsampling process, randomly selecting a subset of 200 songs from the DALI dataset which covered mainly pop songs and rock music in English as Figure 2, ensuring a varied and comprehensive lyrical corpus.

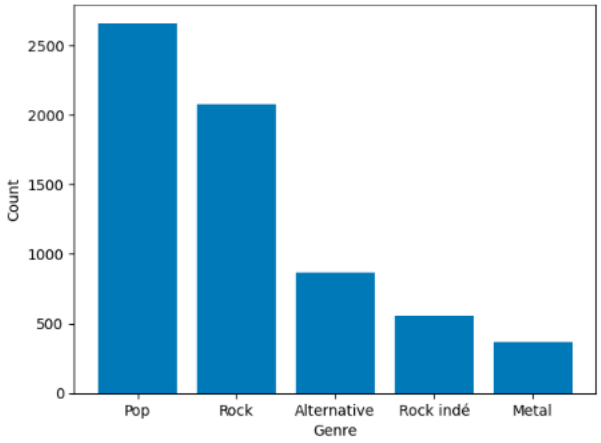

Figure 2: Genre distribution of songs in DALI Dataset

This corpus undergoes first the measurement of melody-lyrics alignment to filter the unqualified ones. Second, we convert the pitch in frequency to the note in the standard musical alphabet and separate melody contour from lyrics, facilitating the generation of both image data for melodic patterns and textual prompts for lyrical context.

Raw Data of one melody contour with lyrics	Frequency2Note
{'duration': 0.083, 'pitch': 523.25, 'text': 'while'}	'pitch': C5
{'duration': 0.25, 'pitch': 523.25, 'text': 'he'}	'pitch': C5
{'duration': 0.083, 'pitch': 523.25, 'text': 'was'}	'pitch': C5
{'duration': 0.25, 'pitch': 698.45, 'text': 'brag'}	'pitch': F5
{'duration': 0.417, 'pitch': 698.45, 'text': "gin'"}	'pitch': F5

3.2 Baseline Models

After the data preparation, the baseline models, comprising Long Short-Term Memory (LSTM), Recurrent Neural Network (RNN), and GPT-3, serve as the foundational framework for comprehending the intricacies of the melody-conditioned lyrics generation task. With the pursuit of novelty in the model's performance in the melody-conditioned lyrics generation, we experiment our dataset on not only the GPT-3, but also the commonly-used Seq2Seq models like RNN and LSTM, showcasing the potential advantage of the large language model in the lyrics generation task of sequential input.

1) Long Short-Term Memory Model: Our LSTM baseline model is designed with a

sequential architecture using Python, employing the Keras library. We construct the vocabulary dictionary mapping all the syllables in the DALI 5k songs into index. Our LSTM model consists of an embedding layer to represent the vocabulary, an LSTM layer with 128 units, and a dense layer with a softmax activation function to predict the next element in the sequence. The embedding layer is configured with an input dimension equal to the vocabulary dictionary size, an output dimension of 32, and an input length of 2 indicating duration and pitch. The model is compiled using the sparse categorical cross-entropy loss function, Adam optimizer, and accuracy as the evaluation metric. The training process involves fitting the model to the training data for 50 epochs with a batch size of 1.

2) Recurrent Neural Network: Our RNN baseline model, also implemented using the Keras library, follows a sequential structure. It comprises an embedding layer, a SimpleRNN layer with 128 units and a ReLU activation function, and a dense layer with a softmax activation function. The configuration of the embedding layer, compilation process, training epoch and batch size are the same as the LSTM model.

3) Text-Input Large Language Model: Our GPT-3 baseline model leverages OpenAI's text-davinci-003 engine. The model utilizes a tuning-free prompt-based approach, generating sequences of tokens in response to a given input prompt. The parameters include the engine specification ('text-davinci-003'), the prompt, maximum tokens, and temperature. The resulting output represents the generated lyrics conditioned on the provided melody.

3.3 Data Transformation: Spectrogram Image

In the process of spectrogram image preparation, we convert the textual representation of melody and lyrics into image data. This image data is represented as a spectrogram, encoding essential information such as duration and pitch along both time and frequency axes. The conversion process begins by extracting pitch, duration, and text information from the provided melody contour with lyrics.

The core of the transformation lies in the waveform generation, where the extracted notes and durations are used to construct a waveform. This waveform is then subjected to a Short-Time Fourier Transform (STFT) as we experiment with the signal processing. The resulting spectrogram, represented as amplitude in decibels, offers a frequency-wise portrayal of the musical composition.

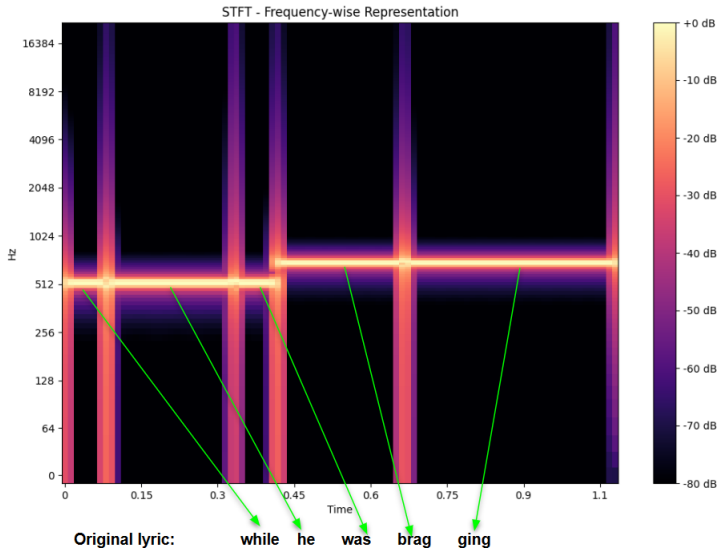

Figure 3: Spectrogram of melody contour

The generated spectrogram is visualized using Matplotlib as shown in Figure 3, where the x-axis signifies time, the y-axis represents frequencies in logarithmic scale, and the color intensity corresponds to amplitude. The resulting image is saved in PNG format.

3.4 Multi-Model Large Language Model

Our main contribution is the integration of the tuning-free prompt-based multi-model large language model, the GPT-4, which, carries out the "Text+Image" to "Text" task, and aims to refine the generation process further. As an extension of text input, including prompts and song data, we incorporate the spectrogram image of the melody contour encoding duration and pitch with the original text input and trial on different prompts. The combination of both text and image inputs in the multi-model Large language model, in our expectation, would generate lyrics with better melody-lyrics alignment. The GPT-4's advanced capabilities allow for a more nuanced understanding and generation of lyrics that not only fit the melody's rhythm and pitch but also maintain lyrical quality. The model is trained to ensure that the generated lyrics are not only technically sound but also emotionally resonant and contextually relevant.

This architecture is not only a technical framework but also a creative exploration into the potential of AI in music composition. It underscores the significance of harmony between lyrics and melody, which is often the essence of memorable music. Our findings contribute to the body of knowledge in AI-assisted music generation, providing pathways for future research and development in this fascinating fusion of technology and artistry.

3.5 Prompts Optimization

Our research underscores the significance of the prompts provided to multi-modal models like NExT-GPT and GPT-4, which directly affect their lyrical output. The clarity of the image inputs and the format of these prompts are crucial for compatibility with subsequent LSTM testing. With GPT-4's recent update to version 1.0.0, manual input of lyrics became necessary, leading to a recommendation for input brevity—ideally under 15 items—to optimize performance. The following user input is the format we suggest.

```
--User:
Please generate lyrics aligned to the input melody.
Note that the alignment standard should consist rhythm, pitch and structure facets.
Especially the number of notes should equal to the number of syllables.
Please output the lyrics in the same format as
[{'duration': xxx, 'pitch': xxx, 'text': xxx}].
Input melody:
[{'duration':xxx, 'pitch': xxx},
...
```

4 Results

The experimental evaluation of our melody-conditioned lyrics generation solution involves an assessment of key metrics, comparing various models with distinct alignment strategies. The experimentation protocol ensures a robust methodology, providing insights into the performance of each model.

4.1 Experimentation protocol

The development of our measurement methodology for evaluating the alignment of lyrics with melody draws inspiration from seminal work in the field, particularly from Nichols's study in 2009 on the interplay between lyrics and melody in pop music [15]. Additionally, the pitch contour in both lyrics and melody was noted as a crucial factor, with variations in pitch positively correlated. Our measurement is bifurcated into two integral components: rhythm alignment and pitch alignment.

1) Rhythm Alignment: In addressing rhythm alignment, we utilize the average duration within a melody contour. By consulting the CMU dictionary, we ascertain the stress pattern of each syllable—whether stressed or unstressed. The guiding principle dictates that longer notes should synchronize with stressed syllables, while shorter notes align with unstressed ones. The rhythm alignment score is subsequently calculated based on the harmony between note durations and syllable stress.

2) Pitch Alignment: Pitch alignment entails a nuanced, multi-step procedure. Initially, lyrics are extracted from the melody contour. This is followed by a text-to-speech task using Google Translate's text-to-speech (gTTS) API to transform the lyrics into audio files. Extracting the mean of the pitch contour from this audio file provides a representation of the lyrics' melody contour. To quantify the conformity between the lyrics' melody contour and the original melody contour, the Mean Squared Error (MSE) is employed. The resulting pitch alignment score serves as an indicator of the alignment fidelity between the two contours.

3) Overall Score: The overall alignment score is normalized and computed as the average of rhythm and pitch alignment scores, providing a comprehensive evaluation metric for the melody-conditioned lyrics generation solution.

Table 1 outlines the performance metrics of different models with Rhythm Alignment, Pitch Alignment, and an Overall Score.

	Measurement		
Model	Rhythm Alignment	Pitch Alignment	Overall Score
RNN	0.51	0.09	0.31
LSTM	0.47	0.10	0.30
GPT-3 (Text-Input LLM)	0.55	0.11	0.33
GPT-4 (Multi-Model LLM)	**0.57**	**0.12**	**0.34**

Table 1: Comparison of model performance on the measurement metrics

The results exhibit the model-wise performance, with GPT-4 (Multi-Model LLM) showcasing superior position in both rhythm alignment, 0,57, and pitch alignment, 0.12. The pitch alignment scores of all four models remain relatively low, partly due to the imperfectness of the measurement of the positive correlation between the original melody contour and the extracted lyrics contour. Beyond Mean Squared Error (MSE), more evaluation metrics in signal processing are expected to be experimented on our task.

4.2 Measurement Graphs

In addition to the measurement table, detailed measurement of our 100-line melody corpus is also displayed in the figures below. In Figure 4, Figure 5, Figure 6 and Figure 7, each data point represents one melody contour with the orange points showing the alignment scores of the original lyrics and the blue points showing the alignment scores of the generated lyrics.

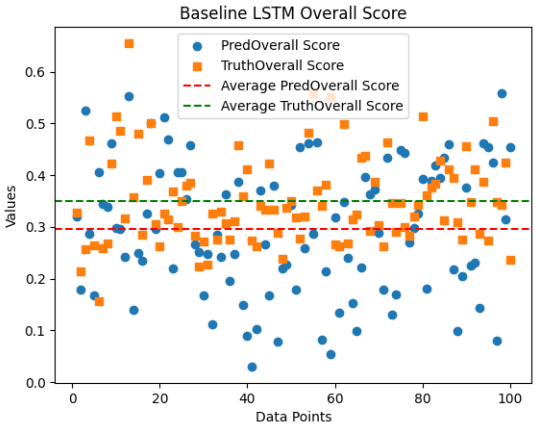

Figure 4: Baseline (LSTM) Measurement

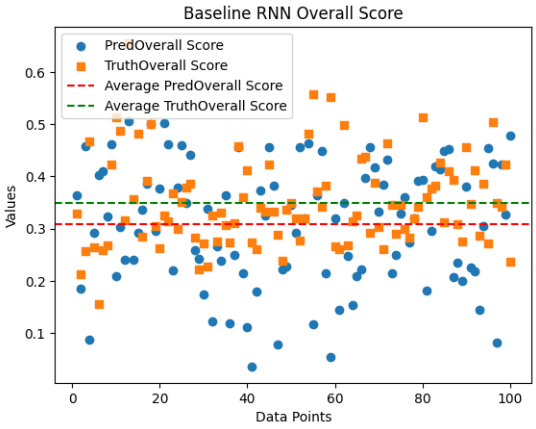

Figure 5: Baseline (RNN) Measurement

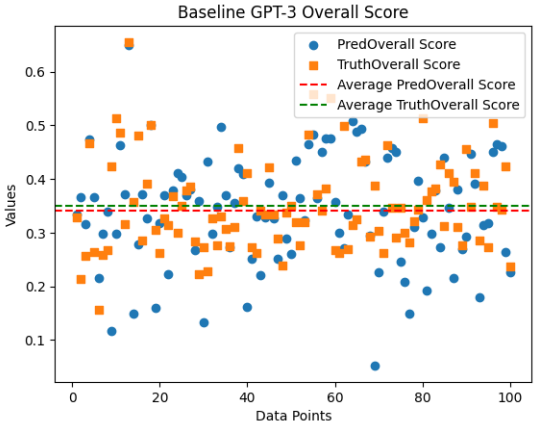

Figure 6: Baseline (GPT-3) Measurement

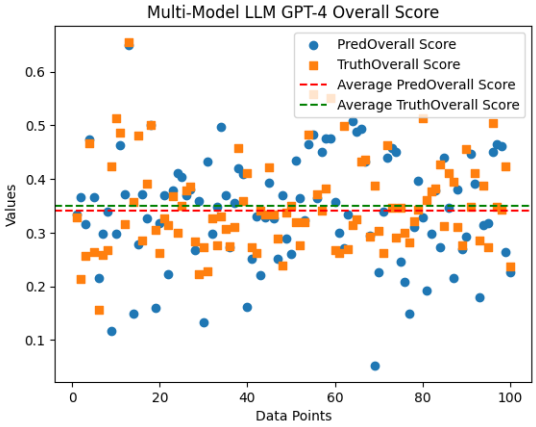

Figure 7: Multi-Model (GPT-4) Meausrement

The average scores of all 100 melody contours are separately shown in the graph, as the green line represents the average score of the melody contour with original lyrics and the red line represents the average score of the melody contour with generated lyrics.

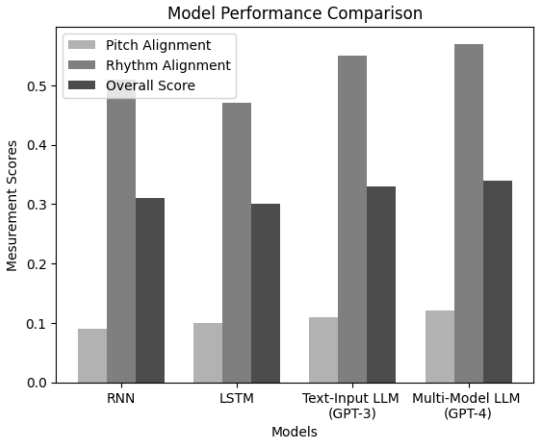

Figure 8: Performance comparison of models

The Figure 8 indicts our four models' performance, as the large language model, for instance, the GPT family (GPT-3 and GPT-4), in general, obtains relatively higher scores in both rhythm alignment and pitch alignment compared with the ordinary Seq2Seq models. This comparison serves as a foundational element in the experimental validation of our approach, allowing for a nuanced understanding of the large language model's relative strengths and weaknesses in the context of melody-conditioned lyrics generation.

5 Discussion

One of the foremost challenges we faced, which also represents a significant achievement, was the development of a robust alignment metric. Upon analyzing the DALI dataset, it became apparent that the inherent alignment of lyrics with melodies was suboptimal. To address this, we devised a multi-dimensional alignment metric. This metric demonstrated superior performance in capturing the nuances of alignment when benchmarked against Zhe Zhang's evaluations with Chat-GPT [7]. Nonetheless, our approach has its limitations; we did not incorporate human evaluations, a component we acknowledge as critical for the authentic assessment of lyric alignment and plan to integrate in future iterations of our research.

The selection between NExT-GPT and GPT-4 for generating melody-conditioned lyrics posed a significant challenge. Initially, NExT-GPT was our preferred multi-modal language model, but it struggled with mixed 'text+image' inputs and required extensive training and fine-tuning, which would extend beyond our planned timeline[8]. Consequently, we pivoted to GPT-4, which,

with its efficient API, facilitated the successful generation of lyrics, aligning more closely with our project's timelines and objectives[9].

In the pursuit of improvement, our project contemplates shifting from multi-modal language models to simpler text-to-text structures. This consideration stems from GPT-4's limitations in interpreting complex image data, such as spectrograms, which compromise lyrical generation quality. Moreover, our approach to melody constraints, though functional, relies heavily on basic assumptions about syllable stress and note duration. We advocate for further research into more intricate aspects of alignment, like tone and pitch variations. Lastly, time constraints prevented the further and deeper exploration of RNN or CNN models, which hold the potential for enhancing future iterations of our project.

6 Personal Contributions

I played a pivotal role in the conception and initiation of the project, which was inspired during the discussion with Prof. Gus Xia, ensuring a cohesive and guided approach throughout the project's development. In the early stages, I meticulously conducted a thorough literature review, delving into related measurement metrics, dataset collection methodologies, and data subsampling techniques. My efforts were instrumental in comprehending the foundational aspects of the research domain, laying the groundwork for subsequent project phases.

Taking charge of the baseline model construction, I designed the research framework and implemented LSTM, RNN, and GPT-3 as our baseline models that served as the foundation for our experimentation. This involved making critical decisions regarding model architecture, data representation, and training strategies.

A significant portion of my contribution focused on the measurement research and implementation phase. I led the exploration and development of novel approaches to assess the melody-lyrics alignment. This entailed formulating and implementing metrics for rhythm and pitch alignment, incorporating insights gained from the literature review.

Additionally, I took the lead in model evaluation using our experimental measurement metrics, conducting experiments to improve the performance of our proposed solution against baseline models. This involved refining evaluation protocols, interpreting results, and iterating on the model based on empirical findings.

Lastly, my involvement extended to the report writing. I documented each phase of the project, ensuring clarity and coherence in articulating our methodology, findings, and conclusions. My commitment to effective communication played a crucial role in presenting our research in a structured and accessible manner.

7 Conclusion

Our project successfully established a novel framework for generating lyrics conditioned on melodies using large language models. Key results include: 1) Developed alignment metrics outperforming existing methods in dataset evaluations. 2) Demonstrated GPT-4's superior efficacy over NExT-GPT for 'text+image' input for lyric generation. 3) Improved lyric naturalness, correctness, and coherence, maintaining a strong aligned connection to the melodies. 4) Outlined future work directions, including the exploration of intricate melody constraints and the integration of RNN or CNN models to refine the generation process. These findings open avenues for further research, particularly in enhancing the model interpretability of complex inputs and expanding the scope of melody constraints considered during lyric generation.

References

[1] K. Watanabe, Y. Matsubayashi, S. Fukayama, M. Goto, K. Inui, and T. Nakano, "A melody-conditioned lyrics language model," in *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, Jun. 2018, pp. 163–172. [Online]. Available: https://aclanthology.org/N18-1015

[2] X. Lu, J. Wang, B. Zhuang, S. Wang, and J. Xiao, "A syllable-structured, contextually-based conditionally generation of chinese lyrics," 2019.

[3] P. Bolcato, "Concurrent generation of melody and lyrics by recurrent neural networks," 2020.

[4] Y. Chen and A. Lerch, "Melody-conditioned lyrics generation with seqgans," 2020.

[5] Y.-F. Huang and K.-C. You, "Automated generation of chinese lyrics based on melody emotions," *IEEE Access*, vol. 9, pp. 98 060–98 071, 2021.

[6] L. Tian and X. Yang, "Deeplyrics: Gpt2 for lyrics generation with finetuning and prompting techniques," 2023. [Online]. Available: https://api.semanticscholar.org/CorpusID:259265730

[7] Z. Zhang, K. Lasocki, Y. Yu, and A. Takasu, "Melody-conditioned lyrics generation via fine-tuning language model and its evaluation with chatgpt," 10 2023.

[8] S. Wu, H. Fei, L. Qu, W. Ji, and T.-S. Chua, "Next-gpt: Any-to-any multimodal llm," 2023.

[9] OpenAI, "Gpt-4 technical report," 2023.

[10] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," 2021.

[11] Z. Wang, Y. Jiang, Y. Lu, Y. Shen, P. He, W. Chen, Z. Wang, and M. Zhou, "In-context learning unlocked for diffusion models," 2023.

[12] S. Min, X. Lyu, A. Holtzman, M. Artetxe, M. Lewis, H. Hajishirzi, and L. Zettlemoyer, "Rethinking the role of demonstrations: What makes in-context learning work?" 2022.

[13] G. Meseguer-Brocal, A. Cohen-Hadria, and G. Peeters, "Dali: A large dataset of synchronized audio, lyrics and notes, automatically created using teacher-student machine learning paradigm." 2018. [Online]. Available: https://zenodo.org/record/1492443

[14] Z. Sheng, K. Song, X. Tan, Y. Ren, W. Ye, S. Zhang, and T. Qin, "Songmass: Automatic song writing with pre-training and alignment constraint," 2020.

[15] E. Nichols, D. Morris, S. Basu, and C. Raphael, "Relationships between lyrics and melody in popular music." 01 2009, pp. 471–476.

[16] J. Merrill, D. Sammler, M. Bangert, D. Goldhahn, G. Lohmann, R. Turner, and A. Friederici, "Perception of words and pitch patterns in song and speech," *Frontiers in Psychology*, vol. 3, p. 76, 03 2012.

[17] Y. Tian, A. Narayan-Chen, S. Oraby, A. Cervone, G. Sigurdsson, C. Tao, W. Zhao, T. Chung, J. Huang, and N. Peng, "Unsupervised melody-guided lyrics generation," 2023.