LyrAIX - tweaking the style

YEVHENII BUDNYK, University of Twente, The Netherlands

Recently, Large Language Models (LLMs) have been tested in different domains. This research examined the capabilities of a fine-tuned LLM to generate lyrics in different styles. A pre-trained medium (355M of learning parameters) size GPT-2 model was fine-tuned with a compiled dataset of song parts and corresponding stylistic labels. The dataset was constructed from lyrics collected from the Genius website, which were later filtered and labeled with the help of unsupervised and rule-based classifiers. The model was tested to generate lyrics defined by such stylistic parameters as affect, topic, rhyme scheme, and content explicitness. Additionally, the model was assessed on the preservation of the author's style originality and distinctiveness. The results have shown that a fine-tuned LLM is more capable of lyrics generation with defined text explicitness and affect, rather than topics and rhyming scheme. Furthermore, a positive indication of the original author's style preservation was discovered with the reported average similarity score of 6.167 on a 1 to 10 Likert scale.

Additional Key Words and Phrases: Transformer Neural Networks, Large Language Models, Style Transfer, Lyrics Generation, Transfer Learning

1 INTRODUCTION

Artificial Intelligence is becoming an omnipresent assistant for humanity. Machine Learning algorithms and Neural Networks are helping people to accomplish different types of tasks such as weather prediction [32] or Text-to-Image generation with control of conditional inputs such as edge maps, keypoints, and segmentation maps [45]. Recent access to Large Language Models (LLM) such as ChatGPT [27] has unarguably broadened the range of opportunities to all their users, even on such complex tasks as medical challenge problems [25]. The expansion of Natural Language Processing (NLP) applications in the domain of creative writing tasks was not long in coming either. For example Roemmele and Gordon [33] have examined how AI can assist writers by suggesting continuing sentences in their text. Moreover, several other research projects aimed to generate lyrics while taking into account the musical component [8, 20, 44]. Chen and Lerch [8] explored the capabilities of Sequence Generative Adversarial Networks on generating lyrics from the input melody. Meanwhile, Xue et al. [44] built a Transformer-based autoregressive language model that produces both rhymes and rhythms for rap music. Another great attempt to develop a lyrics generation assistant is AI-lyricist [20] which creates texts based on input vocabulary and MIDI files. Although music is unquestionably a vital part of any song, not every songwriter has the melody prepared before creating the lyrics. Thus, many other scientific projects [15, 40, 46] aimed to improve exclusively the text-writing part of the song creation process. Watanabe et al. [40] presented novel generation models that are capable of topic modeling and smooth topic transitions with the help of the Hidden Markov Model. Zhang et al. [46] introduced a project that includes an interactive generation mode, enabling users to choose the sentences they find favorable from the generated options. Although the system gives control over multiple parameters for lyrics generation, it is available only in the Chinese language. Other scientific projects either perceive the style as a single attribute [15] or focus only on one stylistic parameter [40]. This approach significantly limits the range of available options to the user during the lyric generation process. Approaching AI as a tool that facilitates and assists, rather than substitutes humans in the process, requires more interactive capabilities. Moreover, such an assistant could help beginner artists with experiments and exploration of different styles. This paper describes LyrAIX, a novel solution for interactive lyrics generation. LyrAIX empowers a pre-trained GPT-2 [31] LLM that is fine-tuned on a constructed dataset of lyrics-generating instructions. Besides several stylistic parameters such as affect, rhyme scheme, topic, and content explicitness the model was also trained with the idea to be aware of the unique author's style. Thus, the author's name was also included in the prompts. As a result of conducted research several important finding were discovered. The accuracy of stylistically conditioned lyrics generation by a fine-tuned GPT-2 LLM appeared to vary from 15.32% to 71.67%, depending on the attribute. This signifies a very limited efficiency in task performance. Moreover, modeling of the topics and rhyme scheme according to the input parameters appeared to be the least efficient, with both accuracy scores being lower than 30% according to the automated evaluation. On the other hand, the generation of lyrics in the style of the specific author as well as defined affect and explicitness level has shown promising results that are worth further exploration. First of all, this paper examined the capabilities of fine-tuned LLMs on the task of stylistically restricted lyrics generation. Secondly, it explored if such LLMs are sensitive enough to grasp the personal style of the author. These research contributions are aimed at answering the following two research questions:

1.1 Research Question 1

To what extent can a fine-tuned LLM generate lyrics with controlled stylistic attributes such as affect, rhyme structure, topics, and explicitness?

Given that the style has primarily been treated as a single, universal entity, it remains uncertain whether this is the most accurate way to represent it. In this research paper, style is approached not as a single parameter, but rather as a complex structure, having 4 different characteristics. A fine-tuned model is examined on the generation of lyrics in distinctive stylistic specifications, for example, a song chorus about life and relationships with high valence and rhyme scheme 0-1-0-1.

1.2 Research Question 2

To what extent can such a LLM generate lyrics in a style of a specific author?

TScIT 39, July 7, 2023, Enschede, The Netherlands

Artists tend to use authentic words in their lyrics, create unique metaphors or sing about unreal objects that become household names. Furthermore, creative texts serve as an embodiment of the author's emotions, cultural background, and life experiences. For some authors, these factors do not deviate significantly from song to song, and hence form a recognizable personal style. Thus, in this research paper authors are perceived as entities with their own unique styles, and are also prompted to the model as a separate attribute during the generation process.

2 RELATED WORK

The following subsections will briefly cover the specifications of the deep neural network architecture that lies at the core of this research. Further, related lyrics generation projects will be discussed. Finally, research papers focused on style transfer and the generation of text with controllable parameters will be reviewed.

2.1 Large Language Models and GPT-2

The increased popularity of LLMs is largely attributed to the recent advancements in neural network architectures. Transformer neural network [38] employs a self-attention mechanism, that has initially been used for machine translation tasks. Lately, Transformers became a central building block of several popular LLMs and has shown its high efficiency in the tasks of Language Generation [4, 9, 31]. GPT-2 [31] is a model that can be imported from the official distribution, making it easily accessible by any user. It exists in several versions, implying different amounts of learning parameters the model contains. According to Radford et al., [31] the model was pre-trained on a WebText dataset containing millions of web pages. Furthermore, several evaluations outlined excellent fine-tuning capabilities of GPT-2 on such tasks as patent claim generation [17] or dialogue generation in RPG games [37]. This paper elaborates on fine-tuning and evaluation of GPT-2 on the task of Lyrics Generation with controllable stylistic parameters.

2.2 Automatic Lyrics Generation

Automatic Lyrics Generation refers to the process of using software programs for lyrics writing. Prior to this, scientists were working on the automatic generation of poetry. According to Oliveira [26] the research on this subject started evolving in the late 1990s.

2.2.1 Rule-Based Systems. One of the earliest systems that were developed in an attempt to generate poems according to the user's preferences was ASPERA [12]. It is a forward-reasoning rule-based software, that considers desired rhyme structure, mood, degree of formality, length of the poem, and the setting to generate a new poem. Besides these parameters, the user is also asked to input a prose paraphrase, that is later used in the planning of the poem draft. During the creation of the text, the system is retrieving other poems from the database and uses them as a starting point. Once generated lyrics are reviewed and validated by the user of ASPERA - the poem is stored in the database for future reuse. Although lyrics and poems both rely on the rhyming of the neighboring lines, they still have distinctive characteristics. First of all, lyrics are made specifically for songs and are supposed to be accompanied by music. Additionally, due to the speed of spread in social media, lyrics of the songs tend to become more viral than poems and thus could be viewed by more people, potentially creating more impact. Considering that ASPERA merely creates a new poem by significantly adjusting another, this approach eventually leads to a lack of originality in the generated outputs. Moreover, with the emergence of Deep Neural Networks (DNN) more robust solutions became available.

2.2.2 Neural Networks. Several scientific groups have concentrated their effort specifically on the research and development of lyrics generating software based on neural networks. Potash et al. [30] proposed a Short-Term Memory language model constructed for the generation of novel rap lyrics in a similar style of a given rapper. The system used Long Short-Term Memory (LSTM) neural network for lyrics writing, which has a main limitation - its inherent sequential nature. LSTM models rely significantly on a hidden state and memory cells to generate text, which makes it less effective if the sequence length grows bigger. Zhang et al. [46] has introduced Youling - an AI-assisted Lyrics Creation System. The researchers pre-trained a GPT-2 model on a 30GB of Chinese literature corpus and later fine-tuned it on a 300K lyrics corpus. Additionally, the system provides a high level of controllability over the process of lyrics generation. First of all, it provides the user with two sets of input parameters, namely content and format-controlling attributes. Meanwhile, the content control is similar to what LyrAIX can offer, the format-controlling parameters are introduced more extensively, including the number of sentences and words per sentence. Second of all, Youling enables its users to modify certain word choices in the generated lyrics. Despite a lot of similar features between my research and Youling, there are also fundamental differences. Firstly, Youling is created to generate lyrics only in the Chinese language, making it an unsuitable option for those who want to write lyrics in other languages. Secondly, Youling is a full-scale system, while LyrAIX is rather a fine-tuned LLM that has no user interface. Thirdly, LyrAIX uses instruction-based fine-tuning, while the aforementioned system gives the style parameters in a raw format. Another point that differentiates LyrAIX from Youling, is the attention it dedicated to the idea of the unique personal style of the songwriter. Similarly to Ghostwriter [30], it assumes that the unique style of the author is an attribute that is beyond any simple composition. At the same time, LyrAIX is not restricted by a single genre of music and provides more control over other stylistic parameters, compared to Ghostwriter. Concluding the aforementioned differentiating characteristics of this research, it aims to generate lyrics in the style of the author but also gives other ways to receive the desired output.

2.3 Style Transfer

Style transfer in lyrics generation is the process of modifying certain stylistic parameters while preserving others. Nikolov et al. [24] trained a Transformer-based denoising autoencoder to generate lyrics from the content of any text. Li et al. [19] proposed a Delete, Retrieve, Generate (DRG) framework for sentiment transfer with content preservation. The system is first deleting a part of a sentence attributed with sentiment, then retrieves the new phrases that correspond to a target value, and employs Recurrent Neural Network (RNN) to fluently combine them. Later, Sudhakar et al.

[34] proposed a more effective way to perform textual style transfer. The researchers enhanced the DRG framework with Transformer architecture, which outperformed state-of-art systems on gender, sentiment, and political slant. However, the software was not tested on the task of style transfer in lyrics generation. Finally, Bucinca et al. [6], have developed a system that is able to preserve the content of the message while manipulating its affect. To perform style transfer, the software collects all the lemmas from the sentence and finds candidate words that are semantically similar, but closer to the target affect. Although LyrAIX is not designed to transfer the style of specific song texts, the aforementioned research projects provide valuable insight into the field. In my project, 4 stylistic parameters can be transferred, but the unique style of the author is aimed to be preserved.

2.4 Style Attributes

This subsection briefly mentions recent developments in affective computing, rhyme modeling, topic modeling, and text explicitness classification.

2.4.1 Affect. According to Munezero et al., [23], 'affect' represents a broad range of subjective experiences and precedes specific feelings and emotions. Before, Bao and Sun [2] developed an Emotional Lyrics and Melody Generator. Although the software can produce both the melody and the lyrics, the emotions spectrum is limited to 2 discrete values: positive and negative. In my research, emotions are conceptualized as a combination of three dimensions: valence, arousal, and dominance - yielding a more intricate representation. Although in the end the values of each dimension are discretized into categories, this approach provides more combinations of affective labels to choose from. Additionally, such representation has proven its efficiency in AffectOn [6], which made it a suitable option for experimentation in the context of my research.

2.4.2 Rhyme Modeling. Rhyme modeling is another dimension of the style uniqueness of each artist and can vary from genre to genre as well as from two different songs of the same author. Xue et al. [44] expand the current knowledge about N-gram rhyme generation in rap songs with left-to-right neural network training. The study also shows that the model can hardly generate a good rap rhyme without rhyme modeling regardless of the size of the N-gram [44]. Although the case study is focused on rap music, the outcomes are applicable to different genres, as many unique styles exist. In my research, rhyme modeling happens only on the level of the last words of each line. Although such an approach does not aim to improve the quality of rhymes, it examines how well LLM can learn and reproduce such rhyming schemes.

2.4.3 Topic Modeling. So far any of the existing lyrics datasets do not provide the ground truth labels of the topics. However, some other projects attempted to define topics with unsupervised learning techniques. For example, Buffa et al. [5] have trained the LDA [21] model to create the labels in the WASABI [5] dataset. The model was trained on more than 1 million songs and resulted in a number of different token clusters. Unfortunately, the WASABI dataset does not have the full lyrics available to the public. Hence, to avoid any potential inconsistencies during the labeling, my research used a similar methodology, but on a custom dataset. The quality of such an approach can not be accurately evaluated, due to the absence of the ground truth.

2.4.4 Lyrics Explicitness classification. There exist both rule-based and machine learning methods of identifying explicit lyrics. Dwiyani et al. [10] have shown an accuracy of 96.3% on a testing dataset. The model was trained with the TF-IDF vectorization method and random forest algorithm. However, such a method requires additional work related to parameter fine-tuning. Additionally, Fell et al. [11] have compared several automated methods of lyrics explicitness classification, including profanities dictionary lookup, dictionary regression, Transformer model, and Textual Deconvolution Saliency. In conclusion, the researchers found that simple dictionary-based models achieve comparable results to deep neural networks. Thus, in my paper, a dictionary lookup method was employed to label explicitness, chosen for its robustness and ease of implementation.

3 METHODS OF RESEARCH

The research involved several important stages such as data collection and cleaning, data labeling, LLM fine-tuning, and a final evaluation of the results.

3.1 Data Collection and Cleaning

For the purpose of this project, a publicly available dataset [7] of web-scraped lyrics from the Genius [18] website was used. Firstly, the lyrics dataset was curated by excluding songs that were not in the English language. Secondly, all the data entries having multiple artists were eliminated, due to the challenge of accurately attributing song parts to their original authors. These two modifications resulted in a dataset consisting of 2,797,631 whole songs. Subsequently, the songs that had a clear textual indication of the start and end of the lyrics part were selected. Lyrics lacking clear division into song parts were removed because they appeared too long for successful rhyme scheme identification. The selected songs were later split into separate entries with the help of regular expression, each representing a text related to a specific part such as Chorus, Refrain, Pre-Chorus, Verse, Hook, and Bridge. However, other song parts such as Intro, Outro, and Interlude were excluded from the scope of this research as those that rarely appear in the rhyming form. Finally, after the labeling process, to reduce the fine-tuning time of the GPT-2 model - all the entries that had less than 2000 online views on Genius at the moment of scraping were excluded too, assuming those songs are not widely known to the audience. The whole data preparation process reduced the dataset size to 1,091,880 entries, as well as significantly decreased the length of each data record. All the preparatory and cleaning procedures resulted in the creation of a completely reformatted dataset, where each data entry included the name of the author, the name of the song part and the corresponding lyrics. Later, these entries were labeled with their predicted topics, affect, rhyme scheme, and content explicitness.

3.2 Data Labeling

This section elaborates on data labeling, including motivation for the design decisions and description of methods, tools, challenges, and results of the process.

3.2.1 Topic Prediction. Songwriters are used to expressing their thoughts on certain topics in their lyrical texts. Thus, LyrAIX aimed to provide its eventual users with the opportunity to choose the topic of the generated text. For this purpose, the lyrics dataset was labeled with the corresponding topics, that were derived by an unsupervised topic modeling method. The pipeline for topic classification is an adapted version of [35]. Each song part was first tokenized and lemmatized with the help of NLTK [3] library. Profanities and stopwords were subsequently eliminated from the corpus using a dictionary lookup method, leveraging both the YouTube Comment Blacklist [28] and the NLTK list of stopwords. The motivation behind this decision was to eliminate elements that introduce excessive noise, as these specific parts are highly likely to occur in the text regardless of the topic. Finally, outlier words that are present in less than 100 songs and more than 80% of entries were also removed, to prevent the inclusion of rare and ubiquitous words that may not contribute significantly to the final goal. The described filtering process aims to improve the quality and relevance of the topic model by focusing on more prevalent terms from the dataset. Further, LDA [21] model was used to cluster the remaining tokens into groups in an unsupervised manner. The model provided two outputs - document-topic and topic-word matrices. The former represents the likelihood of each entry belonging to different topics, while the latter shows the probability of each word belonging to different topics. The number of token clusters was assigned to 4, to achieve a better distribution of the topic labels over the whole dataset. Finally, the labels for each of the 4 word clusters were assigned manually, using the most overarching topic names as the basis. This resulted in the following topic categories: General, Life and Relationship, Money and Authority, and Religion and Society. Table 1 shows the number of song parts labeled with each topic and provides the top 5 salient terms for each of the topic clusters. In this context, a salient term is a word that has the highest probability to appear in the text about a certain topic.

	General	L&R	M&A	R&S
Number of entries	257,491	345,147	410,539	78,704
Salient Term 1	Time	Know	Like	God
Salient Term 2	Take	Love	Man	Soul
Salient Term 3	Life	Never	Money	Free
Salient Term 4	Come	Want	Hit	Hear
Salient Term 5	Day	Like	Back	Fire

Table 1. Number of entries and the most salient terms of each topic

3.2.2 Affect Calculation. This research aimed to provide users with more variability and control over the emotional spectrum of the song. Thus, each lyrical text part was assessed in three dimensions: Valence, Arousal, and Dominance. During labeling lyrical texts were tokenized with NLTK library and each word from every song was looked up in the NRC VAD [22] dictionary, to retrieve corresponding valence, arousal and dominance vectors on a bipolar scale. The vector numerical values were later totaled and divided by the number of words of the song part that appeared in the dictionary, resulting in an average value over the particular piece of text.

$$\frac{\text{sum of all affect vectors of the song part}}{\text{number of words of the song part that are found in affect dictionary}}$$

Next, all vectors were normalized according to the following formula where $x_{\min}$ and $x_{\max}$ resemble the vectors with minimum and maximum values of each dimension.

$$x' = \frac{2 \cdot (x - x_{\min})}{x_{\max} - x_{\min}} - 1$$

This normalization was carried out to maintain consistency in the distribution of values across dimensions. Finally, to adapt calculated numerical values for LLM fine-tuning, they were discretized into categories: Low, Medium, and High. To make the LLM learn from the equal distribution of each category within each dimension special thresholds were determined. The thresholds were established by dividing the values of each dimension into three segments, each representing 33 percentiles. For valence, the lower threshold was 0.087, and the upper threshold was 0.229, for arousal the thresholds were -0.216 and -0.098, and for dominance, they appeared to be -0.100 and -0.004 correspondingly. In the end, each dimension of Valence, Arousal, and Dominance had around 33% of values in each of the discrete categories. This ensured the model had a substantial amount of each category representation in every affective dimension.

3.2.3 Rhyme Scheme Identification. Next, all the lyrical texts were labeled with the corresponding rhyming scheme. This task introduced several challenges to the research. First of all, there are many variations of rhymes such as perfect and imperfect rhymes, internal rhymes, and end rhymes, and those that consist of multiple words. Second of all, deriving the scheme from multisyllabic rhymes is a computationally demanding task. Facing these challenges, for the scope of LyrAIX only single-word end rhymes were considered. Thus, the last word of each line of the lyrics was selected. Further, the rhyming score for each pair of the end words was computed with a modified and extended script of Rap Lyrics Generator [42, 43]. The software is translating words into phonemes taken from the variation of CMU pronouncing dictionary [41]. Practically, the algorithm computes a rhyming score based on vowel score, stress score, and consonant score that are derived from normalized log-odds scores, based on frequency statistics [13]. In the context of current research, all the words that had a rhyme score higher than 2 were considered as rhyming. This threshold was manually established by experimental trials with the labeling software, to reduce the amount of falsely identified rhymes. Finally, each rhyming line was assigned the same number, resulting in the following type of scheme.

$$\text{...hello} - \text{...name} - \text{...trello} - \text{...fame}$$

$$0 - 1 - 0 - 1$$

3.2.4 Content Explicitness Classification. An important requirement of the system was to give users the freedom to choose the level of explicitness of generated lyrics. Some profanities were included in the NRC VAD dictionary [22], but their presence in the text did not play a significant role in the final affect score. Moreover, profanities were excluded from the training data of the LDA [21] model, as they were introducing more noise to the data. As far as none of the aforementioned stylistic parameters could outline the level of explicitness, a separate attribute was added. A publicly available YouTube Comment Blacklist [28] was used to perform a dictionary lookup classification. All the lyrics containing more than

0 profanities were marked as "Explicit Content", and others were classified as "Non-Explicit Content".

3.3 Model Training

The GPT-2 pre-trained LLM was selected as the most suitable for the goals of this research paper due to its availability, comparatively low demand for GPU computing resources, and high speed in fine-tuning. To make the model learn how to respond to a particular task the prepared dataset was reconstructed into a text file containing user instructions mentioning the labels and the corresponding lyrics as a desired output. This method has proven its efficiency in fine-tuning the process of other LLMs [39]. In the text file each instruction starts with <INSTRUCTION> token and each subsequent data entry is marked with <OUTPUT> token.

<INSTRUCTION> Generate a Chorus of song lyrics in a style of JAY-Z about life and relationships, with High valence, High arousal, and High dominance, the rhyming scheme should be 0-1-2 and there should be Non-Explicit content.

<OUTPUT>Ge-ge-geyeahhh
Can I live?
Can I live?[14]

The instructions dataset was split into 90% of training data and 10% of validation data. Later, nanoGPT [16] codebase was customized to specifically fine-tune the lyrics generation with controllable parameters. A medium size version of the model with 355M of parameters was used. The dropout rate was set to 0.15 to prevent it from overfitting the data. The 355M GPT-2 model was fine-tuned with 9 thousand iterations, which resembles 3 learning epochs correspondingly. The context window of the model remained unchanged and was equal to 1024 tokens. The gradient accumulation steps value was adjusted to 48, making the model learn from the batches of 49152 tokens in every iteration.

3.4 Evaluation

To assess the performance of a fine-tuned LLM according to the research questions both machine and human evaluation methods were used.

3.4.1 Automated evaluation. To examine how well the model can generate the lyrics according to the given parameters, 300 instructions were created with an automated script. These instructions consisted of randomly constructed sets of parameters including song part, author, valence, arousal, dominance, rhyming scheme, explicitness, and topic, selected from the training dataset. Further, the generated commands were prompted to the fine-tuned model, and the outputs were labeled with the same annotation scripts used for the training data. Despite the limitations imposed by the accuracy of labeling tools, this method remains the most pragmatic approach, considering the time constraints of this research. This approach allows for evaluating the capacity of a fine-tuned GPT-2 LLM to generate the text according to the provided input parameters. Finally, accuracy scores for each attribute were computed individually with the help of Scikit-Learn [29] library and confusion matrices were utilized to visually represent the distribution of True and Predicted labels. True labels mean those that were initially prompted to the model in the form of instructions.

3.4.2 Human evaluation. To evaluate the generation of lyrics in an author-specific style 10 participants were recruited and each of them confirmed to have a strong command of the English language. 7 people identified themselves as male, and 3 - as female. The mean age of the sample is μ=21.6 years old, and the variance of the sample's age is σ^2=2.267 years. Additionally, only 6 out of 10 participants have tried writing their own lyrics before. Each person was arranged with a 20-minute time slot for an online video call to be interviewed. During the interview, each participant was asked to choose the author whose lyrics style they can easily recognize. Later, participants were explained what each of the stylistic attributes resembles and what potential impact it could have on the final output. Further, the screen of the interviewer's computer was translated and participants were able to construct the prompt of their choice by giving the instructions to the interviewer. Each participant had 3 attempts to try different combinations of parameters, however, they were not allowed to change the author. After each generation round participants were requested to assess the extent to which the created text represented the style of the prompted author by providing a rating on a scale of 1 to 10. Finally, they were also asked open-end questions to elaborate on what made the lyrics less or more similar to the style of the chosen artist. This format resembles a survey as a part of a semi-structured interview. Subsequently, the result of the experiment was derived as an average score on the Likert scale, without any statistical testing due to the time constraints. The answers to open questions were reviewed and the key points of each interview were later outlined in the brief summary.

4 RESULTS

The results of each experiment are stated in this section according to the selected evaluation metrics.

4.1 Generation of style conditioned lyrics

Results of the automated evaluation have shown that LLM can produce lyrics according to the input stylistic parameters with varying accuracy scores. The highest accuracy was achieved in controlling the level of explicitness in the text, with 71.67% of the generated lyrics receiving the same label as the input value. However, the model struggled with the generation of Explicit lyrics. In Table 2 it is visible that the model has generated 70 Non-Explicit lyrics, while was prompted to do the opposite. Subsequently, an accuracy score of 44% was reported for generating lyrics with a specific dominance level, 38.33% for arousal level, and 32% for valence level. In Tables 3-5 the confusion matrices for each of the affective dimensions show that the Medium category was predicted the least often for all valence, arousal, and dominance. Furthermore, 25.33% of accuracy was reached in the generation of lyrics according to the specific topic. Additionally, Table 6 demonstrates that 140 lyrics were labeled as those resembling the General topic, while were supposed to be about other themes. Finally, the least accuracy score was achieved with the generation of text according to the defined rhyming scheme - 15.32%. Rhyming schemes that were generated accordingly to the prompted label had an average length of 4.2 lines.

In conclusion, the GPT-2 model fine-tuned on a custom instructions dataset has shown underwhelming accuracy in generating stylistically restricted lyrics according to the automated evaluation. Tables 2-6 show the confusion matrices for each of the stylistic parameters with True Values that were prompted on the left and the labeled Predicted Values at the top. The numbers highlighted in bold font represent the instances that were accurately identified.

		Predicted Value	
		Explicit	Non-explicit
True Value	Explicit	**10**	70
	Non-explicit	15	**205**

Table 2. Confusion Matrix - Explicit Content

		Predicted Value		
		Low	Medium	High
True Value	Low	**25**	31	42
	Medium	31	**24**	38
	High	39	23	**47**

Table 3. Confusion Matrix - Valence

		Predicted Value		
		Low	Medium	High
True Value	Low	**54**	30	24
	Medium	43	**28**	29
	High	40	19	**33**

Table 4. Confusion Matrix - Arousal

		Predicted Value		
		Low	Medium	High
True Value	Low	**48**	21	35
	Medium	34	**28**	35
	High	21	22	**56**

Table 5. Confusion Matrix - Dominance

4.2 Preservation of author's style

In this section, the results of people evaluating the generated lyrics on adherence to the style of the author of their choice are presented in a quantitative and qualitative manner.

4.2.1 Survey results. In the end, 30 data points were collected, 3 from each of the interview participants. The mean of the sample ratings appeared to be μ=6.167 with a variance of σ^2=2.902. The lowest rating of similarity reported is 2, which appeared only once in the sample, while the highest rating reported is 9 and it was used 2 times. The distribution of the ratings is visualized in Figure 1.

		Predicted Value			
		L&R	M&A	General	R&S
True Value	L&R	**15**	11	62	11
	M&A	30	**6**	69	7
	General	10	4	**53**	4
	R&S	7	0	9	**2**

Table 6. Confusion Matrix - Topic

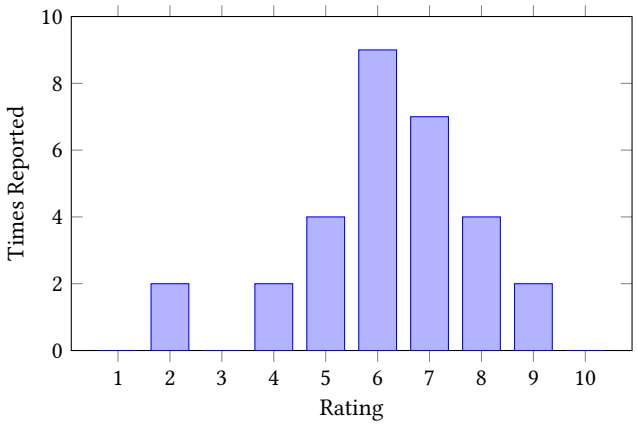

Fig. 1. Survey Results

4.2.2 Interview results. During the interviews, all participants were asked to explain their choice of rating of the lyrics on style similarity adherence. First of all, 4 interviewees reported that elaborating on their decisions appeared to be a tough task, as style similarity is an aspect that can not be explicitly evaluated with any specific metric. However, 8 people also stated that they have a feeling of the style of their author, that can not be described in words, but only intuitively perceived. 4 participants stated that the generated lyrics were more similar to the genre in which the artist creates music, rather his or her specific style. All of the interviewees reported that the rhyming scheme had never been ideally matched, and in most of the cases, it appeared also to be unusual for the author's style. 3 people stated that the generated lyrics correspond less to the style of the author when generated with the stylistic parameters that are atypical for his or her texts. In 12 out of 30 generated lyrics the fine-tuned GPT-2 model failed to provide the requested level of content explicitness. Additionally, 7 out of 10 participants reported the generated lyrics being mostly consistent with prompted valence, arousal, and dominance values. One interviewee stated that "the topic of the song defines the style to a very big extent, and thus it is hardly possible to imagine the author singing about something else than life struggles". 2 respondents also mentioned that changing the stylistic parameters to the opposite of what their artists sing can not just adjust the style, but also generate the lyrics in a different genre. This way the generated verse of "Five Finger Death Punch" a heavy-metal band with the prompted topic of Money and Authority was assessed as more similar to rap music by one of the interviewees.

5 DISCUSSION

This section aims to discuss and highlight the key findings from the results of the research. The least controllable parameter appeared to be the rhyming scheme, a conclusion confirmed by both human and automated evaluation. Although the selected context window during the training and the generation was deemed sufficient for the model to capture comprehensive text segments, it exhibited a limited ability to learn this feature representation. This observation proves the need for alternative approaches such as marking the end token of each line [30] or reverse-order language generation model [44]. Further, a slightly better result was shown by the generation of lyrics with a defined topic. This finding supports the notion that the older methods such as LDA still outperform the capabilities of LLMs in topic modeling [1]. Additionally, the research discovered that topics significantly influence how humans perceive the unique style of the author. The automated evaluation revealed an improved performance in generating songs with controlled affect values compared to the previous two parameters, yet it remained below any noteworthy thresholds for achieving satisfactory outcomes. However, the majority of interview participants reported that their expectations in affective lyrics generation were met in most of the cases. These conflicting results imply that humans have a more sophisticated perception of emotions, and assess affect of the lyrics differently from the machines. The highest accuracy during automated evaluation appeared to be in the preservation of content explicitness. However, many interviewees reported the model being unable to generate the lyrics according to the requested explicitness level, especially for the authors who are not using any profanities in their original texts. This suggests that explicitness can potentially be more effectively generated on the format level similar to defining the specific rhymes of the lyrics [46]. A fine-tuned GPT-2 model demonstrates the ability to generate lyrics with some control over content explicitness, to a lesser extent with control over three affective dimensions, and is less effective in adhering to defined topics and rhyming schemes. Additionally, a choice of substantially altered style attributes can lead to the generation of lyrics that are less specific to the author. Despite the model's tendency to generalize the generated lyrics to the author's genre, the lyrics produced during the interviews were reported to maintain a certain level of similarity to the author's style. This implies that despite the weak ability of a fine-tuned GPT-2 model to generate the lyrics precisely according to the artificial labels, it still has the potential to satisfy the need for lyrics generation according to the specific author's style.

6 LIMITATIONS AND FUTURE WORK

In this section, the limitations of the conducted research will be mentioned and explained.

6.1 Limitations

The evident limitations of the research, its methods, and its evaluation are explained in the following subsections.

6.1.1 Data Labeling Limitations. As far as the labeling methods are based on unsupervised or rule-based tools, the quality of the resulting dataset introduces several concerns. First of all, the topic labels were derived in the absence of ground truth. Moreover, the names of the topic labels were assigned manually by the researcher based on the most salient terms, introducing a certain level of bias. Furthermore, in Table 1 several words such as "Take", "Know" or "Like" relate stronger to some specific topics rather than others, despite being generic common words that can be found in the text of any theme. Another limitation is that the rhyming schemes were based purely on the rhyming score of the end words of each line, ignoring the fact that rhyme can consist of multiple words. Furthermore, since GPT-2 LLM lacks phoneme awareness, the generation of rhymes depended on the model's capacity to learn the higher likelihood of certain words appearing at the end of subsequent lines compared to others. Moving forward, valence, arousal, and dominance labeling considered only unigrams of words, neglecting the word negations and intensifiers. Hence, the song that had the word combination "not happy" repeated several times - would have an incorrect VAD vector, as it would only take into account the word "happy" but omit "not". Finally, text explicitness classification did not consider the evaluation of latent semantic explicit content, which made it less precise in some cases. Such a method is not able to grasp the violent or sexual content from the words in isolation, however, explicitness arises from the context [11].

6.1.2 Model Training Limitations. Due to the limited computational and time resources this research could not afford to train a LLM from scratch as well as fine-tune another bigger model as LLaMA [36]. According to Radford et al., [31] larger versions of GPT-2 are superior in question answering and reading comprehension than a 355M model. Moreover, fine-tuning GPT-2 on a bigger dataset with more epochs appeared an impractical mission either, for the same reasons.

6.1.3 Evaluation limitations. The quality of automated evaluation is significantly dependent on the precision of labeling methods. However, in the context of this research, accurately estimating the precision of labeling tools without any ground truth is an unrealistic task. If the LDA topic model exhibits bias by misclassifying 100K song parts about Society and Religion as General, it would erroneously label the generated lyrics accordingly, despite the LLM potentially producing the correct output. Additionally, one of the main purposes of LyrAIX is to assist lyrics writers in the creation of new texts. Hence, to properly evaluate to what extent a fine-tuned model can serve this goal - the most reliable assessment can be made only by the direct users. Unfortunately, due to time constraints, this research did not involve an extensive human evaluation that could have covered all the aspects of the model's capabilities and limitations.

6.2 Future Work

There are multiple aspects that can be further explored or adjusted in future studies.

6.2.1 Improving dataset quality. There are several ways for possible improvement of the data quality. Firstly, the creation of a smaller, manually annotated dataset could provide a significant enhancement to the project. Wang et al. [39] proved that a set of 53K instructions can be considered enough for LLM fine-tuning to answer different types of queries. As soon as this project involves a single type of

instruction that can have multiple variations with different parameters, the size of the instructions dataset can be reduced even more. Secondly, several alternative approaches can be considered for labeling according to the current parameters. For example, similarly to Youling, keywords can be used as an alternative or an addition to the existing topic label, as it can give advanced control over the context. Subsequently, a more robust method of VAD vector values calculation involving bigrams or trigrams of words can be used during the affect labeling. Moreover, a combination of dictionary lookup and unsupervised methods can be used for the derivation of the explicitness level. Finally, with access to more processing power, rhyme modeling can be improved too, by comparing the scores of 2 or 3 endwords during the process.

6.2.2 Text Formatting. Another possible dimension for future work is to involve text formatting as a separate step of the generation. Following the work of Zhang et al. [46] the user can achieve a higher level of control over the output lyrics with the help of format attributes.

6.2.3 Extensive human evaluation. More extensive and robust evaluation of human opinions about the generated lyrics is required to enhance the positive indication of its efficiency with more data and the corresponding statistical test. Firstly, more participants should be involved in the study, to receive a bigger sample of responses. Later, one of the options would be to ask each participant to indicate the author whose lyrics they can recognize the best in advance. Further, during the interview, a set of 10 song parts will be presented to the person, 5 of them will be generated by the model and the rest will be taken from the original texts. The texts will be shown in a randomized order to minimize potential biases. Additionally, some options can contain the texts of other artists, to allow a more comprehensive comparison of results and exploration of new insights. During the interview, with intended deception, the person will rate the similarity of the lyrics to the style of their author on the Likert scale from 1 to 10, thinking that all the lyrics were produced by the LLM. Later, in case the assumption of normality holds for a sample of responses, a paired sample t-Test can be used to compare the means of the ratings of original and generated lyrics. Otherwise, a non-parametric Wilcoxon signed-rank test may be a suitable alternative. The Null Hypothesis of the experiment can be the following H_0: "There is no significant difference between the mean rankings of the original lyrics and the mean rankings of generated lyrics". The level of significance can be set to α=0.05, to avoid the Type I error. This approach can provide scientifically stronger results than the current research.

7 CONCLUSION

This work introduced a novel approach to lyrics generation using a fine-tuned GPT-2 model, providing new insight into the potential of LLMs in the field of creative writing. To fine-tune the model a dataset of lyrics was cleaned and artificially labeled with the help of unsupervised and rule-based methods. Each data point incorporated the name of the song part, corresponding lyrics, and the derived topics, rhyme schemes, explicitness labels, and affective values. Subsequently, LyrAIX was developed for the generation of lyrics with controllable stylistic parameters and style preservation of the author. Later, the fine-tuned model was evaluated with the help of automated and human-related methods. The software demonstrated the ability to generate lyrics with some control over content explicitness, a relatively limited level of control over affect of the text, and exhibited less effectiveness in adhering to predefined topics and rhyming schemes. At the same time, the model's outputs have shown a promising indication of adherence to the unique personal style of the author. Although some generated lyrics appeared more similar to the genre, rather than specifically to the author, a fine-tuned LLM has shown potential to accomplish this task. The capacity of LyrAIX to generate lyrics that reflect a specific author's style is enhanced when input stylistic parameters are more in alignment with the distinct stylistic nuances inherent to that author. Thus, a fine-tuned GPT-2 LLM has shown its capacity to generate lyrics in author's specific style to a moderate extent, that needs to be more accurately evaluated in the future works. This leads to the conclusion that LyrAIX can potentially become a valuable lyrics writing assistant that can help young artists to discover their own styles.

8 ACKNOWLEDGMENTS

The author of this paper would like to acknowledge the supervisor of this research Dr. Lorenzo Gatti as a valuable advisor and effective co-manager of the project. His extensive scientific expertise facilitated the successful completion of the project.

REFERENCES

[1] Henrik Axelborn and John Berggren. 2023. Topic Modeling for Customer Insights: A Comparative Analysis of LDA and BERTopic in Categorizing Customer Calls.
[2] Chunhui Bao and Qianru Sun. 2022. Generating Music with Emotions. *IEEE Transactions on Multimedia* (2022), 1–1. https://doi.org/10.1109/TMM.2022.3163543
[3] Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python.* O'Reilly Media Inc.
[4] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. arXiv:2005.14165 [cs.CL]
[5] Michel Buffa, Elena Cabrio, Michael Fell, Fabien Gandon, Alain Giboin, Romain Hennequin, Franck Michel, Johan Pauwels, Guillaume Pellerin, Maroua Tikat, and Marco Winckler. 2021. The WASABI Dataset: Cultural, Lyrics and Audio Analysis Metadata About 2 Million Popular Commercially Released Songs. In *The Semantic Web*, Ruben Verborgh, Katja Hose, Heiko Paulheim, Pierre-Antoine Champin, Maria Maleshkova, Oscar Corcho, Petar Ristoski, and Mehwish Alam (Eds.). Springer International Publishing, Cham, 515–531.
[6] Zana Buçinca, Yücel Yemez, Engin Erzin, and Metin Sezgin. 2023. AffectON: Incorporating Affect Into Dialog Generation. *IEEE Transactions on Affective Computing* 14, 1 (2023), 823–835. https://doi.org/10.1109/TAFFC.2020.3043067
[7] CarlosGDCJ. 2023. Genius song lyrics. https://www.kaggle.com/datasets/carlosgdcj/genius-song-lyrics-with-language-information
[8] Yihao Chen and Alexander Lerch. 2020. Melody-Conditioned Lyrics Generation with SeqGANs. In *2020 IEEE International Symposium on Multimedia (ISM).* 189–196. https://doi.org/10.1109/ISM.2020.00040
[9] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805 [cs.CL]
[10] Luh Kade Devi Dwiyani, I Made Agus Dwi Suarjaya, and Ni Kadek Dwi Rusjayanthi. 2023. Classification of Explicit Songs Based on Lyrics Using Random Forest Algorithm. *Journal of Information Systems and Informatics* 5, 2 (2023), 550–567.
[11] Michael Fell, Elena Cabrio, Michele Corazza, and Fabien Gandon. 2019. Comparing automated methods to detect explicit content in song lyrics. In *RANLP 2019-Recent Advances in Natural Language Processing.*

[12] Pablo Gervás. 2001. An expert system for the composition of formal spanish poetry. In *Applications and Innovations in Intelligent Systems VIII: Proceedings of ES2000, the Twentieth SGES International Conference on Knowledge Based Systems and Applied Artificial Intelligence, Cambridge, December 2000*. Springer, 19–32.

[13] Hussein Hirjee and Daniel G Brown. 2009. Automatic Detection of Internal and Imperfect Rhymes in Rap Lyrics.. In *ISMIR*. 711–716.

[14] Jay-Z. 1996. can I live.

[15] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing Modern Language Using Copy-Enriched Sequence-to-Sequence Models. arXiv:1707.01161 [cs.CL]

[16] Andrej Karpathy. 2023. Karpathy/nanogpt: The simplest, fastest repository for training/finetuning medium-sized gpts. https://github.com/karpathy/nanoGPT

[17] Jieh-Sheng Lee and Jieh Hsiang. 2020. Patent claim generation by fine-tuning OpenAI GPT-2. *World Patent Information* 62 (2020), 101983.

[18] Tom Lehman, Ilan Zechory, and Mahbod Moghadam. 2009. Genius. https://genius.com Accessed: 2023-05-03.

[19] Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, Retrieve, Generate: A Simple Approach to Sentiment and Style Transfer. arXiv:1804.06437 [cs.CL]

[20] Xichu Ma, Ye Wang, Min-Yen Kan, and Wee Sun Lee. 2021. AI-Lyricist: Generating Music and Vocabulary Constrained Lyrics. In *Proceedings of the 29th ACM International Conference on Multimedia* (Virtual Event, China) *(MM '21)*. Association for Computing Machinery, New York, NY, USA, 1002–1011. https://doi.org/10.1145/3474085.3475502

[21] Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *Proceedings of the 34th International Conference on Machine Learning (Proceedings of Machine Learning Research, Vol. 70)*, Doina Precup and Yee Whye Teh (Eds.). PMLR, 2410–2419. https://proceedings.mlr.press/v70/miao17a.html

[22] Saif Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Melbourne, Australia, 174–184. https://doi.org/10.18653/v1/P18-1017

[23] Myriam Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. 2014. Are They Different? Affect, Feeling, Emotion, Sentiment, and Opinion Detection in Text. *IEEE Transactions on Affective Computing* 5, 2 (2014), 101–111. https://doi.org/10.1109/TAFFC.2014.2317187

[24] Nikola I. Nikolov, Eric Malmi, Curtis G. Northcutt, and Loreto Parisi. 2020. Rapformer: Conditional Rap Lyrics Generation with Denoising Autoencoders. arXiv:2004.03965 [cs.CL]

[25] Harsha Nori, Nicholas King, Scott Mayer McKinney, Dean Carignan, and Eric Horvitz. 2023. Capabilities of GPT-4 on Medical Challenge Problems. arXiv:2303.13375 [cs.CL]

[26] Hugo Oliveira. 2009. Automatic generation of poetry: an overview. *Universidade de Coimbra* (2009).

[27] OpenAI. 2023. ChatGPT. https://chat.openai.com. Accessed: 2023-05-07.

[28] James Parker. [n. d.]. YouTube Blacklist Words Free and YouTube comment moderation. https://www.freewebheaders.com/youtube-blacklist-words-free-and-youtube-comment-moderation/

[29] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research* 12 (2011), 2825–2830.

[30] Peter Potash, Alexey Romanov, and Anna Rumshisky. 2015. Ghostwriter: Using an lstm for automatic rap lyric generation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 1919–1924.

[31] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI blog* 1, 8 (2019), 9.

[32] Xiaoli Ren, Xiaoyong Li, Kaijun Ren, Junqiang Song, Zichen Xu, Kefeng Deng, and Xiang Wang. 2021. Deep Learning-Based Weather Prediction: A Survey. *Big Data Research* 23 (2021), 100178. https://doi.org/10.1016/j.bdr.2020.100178

[33] Melissa Roemmele and Andrew Gordon. 2018. Linguistic Features of Helpfulness in Automated Support for Creative Writing. In *Proceedings of the First Workshop on Storytelling*. Association for Computational Linguistics, New Orleans, Louisiana, 14–19. https://doi.org/10.18653/v1/W18-1502

[34] Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. Transforming Delete, Retrieve, Generate Approach for Controlled Text Style Transfer. arXiv:1908.09368 [cs.CL]

[35] Tdenzl. 2021. TDENZL/Lyricslda: Generating and visualizing a topic model of Song Lyrics. https://github.com/tdenzl/LyricsLDA

[36] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. arXiv:2302.13971 [cs.CL]

[37] Judith van Stegeren and Jakub Myśliwiec. 2021. Fine-tuning GPT-2 on annotated RPG quests for NPC dialogue generation. In *Proceedings of the 16th International Conference on the Foundations of Digital Games*. 1–8.

[38] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. arXiv:1706.03762 [cs.CL]

[39] Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, Alisa Liu, Noah A. Smith, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. Self-Instruct: Aligning Language Models with Self-Generated Instructions. arXiv:2212.10560 [cs.CL]

[40] Kento Watanabe, Yuichiroh Matsubayashi, Kentaro Inui, and Masataka Goto. 2014. Modeling structural topic transitions for automatic lyrics generation. In *Proceedings of the 28th Pacific Asia conference on language, information and computing*. 422–431.

[41] Robert Weide et al. 1998. The Carnegie Mellon pronouncing dictionary. *release 0.6, www. cs. cmu. edu* (1998).

[42] Matthias Wentink. 2023. *Creating and Evaluating a Lyrics Generator Specialized in Rap Lyrics with a High Rhyme Density*. Ph. D. Dissertation. TScIT 38. https://essay.utwente.nl/94362/1/Wentink_BA_EEMCS.pdf

[43] Matthias Wentink. 2023. Matthiaswentink1/Rap-lyrics-generator. https://github.com/MatthiasWentink1/Rap-Lyrics-Generator

[44] Lanqing Xue, Kaitao Song, Duocai Wu, Xu Tan, Nevin L. Zhang, Tao Qin, Wei-Qiang Zhang, and Tie-Yan Liu. 2021. DeepRapper: Neural Rap Generation with Rhyme and Rhythm Modeling. arXiv:2107.01875 [cs.SD]

[45] Lvmin Zhang and Maneesh Agrawala. 2023. Adding Conditional Control to Text-to-Image Diffusion Models. arXiv:2302.05543 [cs.CV]

[46] Rongsheng Zhang, Xiaoxi Mao, Le Li, Lin Jiang, Lin Chen, Zhiwei Hu, Yadong Xi, Changjie Fan, and Minlie Huang. 2022. Youling: an AI-Assisted Lyrics Creation System. arXiv:2201.06724 [cs.CL]