ANALYSIS OF MUSICAL COMPOSITION THROUGH QUANTIZED FINE-TUNING OF PRE-TRAINED MODELS

Mario Luis Chiaparini Neto

27 June 2024

Abstract

The case study addresses the generation of song lyrics conditioned on the fine-tuning process among different authors. This technique aims to introduce the concept of solution in specific tasks, such as creating lyrics from the title of the work. This association intends to meet the demand for overcoming creativity blocks that composers may face. To achieve this, the musicality derived from different musical styles is inferred through a mapping process based on the adjustment of pre-trained weights in large language models. The system uses a collection of foundation language models, LLaMA, which undergoes fine-tuning with parameter optimization known as QLoRA. This optimization, involving quantization and low-rank adaptation, minimizes computational resources.

1 Introduction

Artificial intelligence has become prevalent in various aspects of humanity. Machine Learning and Neural Networks are responsible for assisting in the successful completion of tasks, such as the recent development of Large Language Models (LLMs), including the creation of the GPT model [5]. These advancements have provided numerous opportunities for complex tasks. For example, studies by Roemmele and Gordon have examined the efficiency of these language processing models in aiding the creative process of book authors [6].

Pre-trained language models achieve better state-of-the-art performance in specific tasks when they undergo fine-tuning. This is because pre-trained weights, imbued with semantic and syntactic knowledge, are adjusted to a particular vocabulary and context of the task at hand [7]. Fine-tuning allows models to leverage the learning obtained from large volumes of general data and specialize in smaller, more specific datasets, providing superior results in applications such as language translation, sentiment analysis, and text generation.

However, this approach also presents some challenges. Fine-tuning can be computationally intensive and requires large amounts of high-quality labeled

data for each new task, which is not always available. Additionally, there is a risk of overfitting, where the model becomes excessively tailored to the specific training data and loses the ability to generalize to new data. Careful selection of hyperparameters and regularization techniques are essential to mitigate these challenges. Despite these limitations, fine-tuning remains a powerful tool for adapting pre-trained language models to a wide variety of specific tasks.

Methodologies like QLoRA ensure efficient memory reduction needed for adjusting pre-trained weights in models with up to 65 billion parameters, using a single 48GB GPU, without compromising fine-tuning quality [2]. QLoRA (Quantized Low Rank Adapters) introduces the 4-bit NormalFloat (NF4) data structure, using double quantization to reduce memory consumption. This technique stores model parameters in a 4-bit format, preserving precision. Combining NF4 with Low Rank Adapters (LoRA), QLoRA adjusts a small number of new low-dimensional parameters, saving memory and computation. This allows large-scale models to be fine-tuned efficiently on accessible hardware, maintaining fine-tuning effectiveness.

Through these techniques of reducing hyperparameters for adjusting pre-trained models, the goal of this project is to fine-tune these models to generate texts that align with the knowledge base of musical composition. Tests will be conducted with different artists and various musical approaches. In the analysis of the fine-tuning quality characterization, a comparison will be made with a composer selected from Vagalume dataset and with the unadjusted model. Metrics to ensure the quality monitoring of the model's fine-tuning process include ROUGE n-grams score, BLEU score, and the reduction of the loss function as the pre-trained weights are adjusted, ensuring that the proposed generation effectively accomplishes the task of generating music from a title.

In this report, we will verify whether it is possible to use language models, such as LLaMA-13B, for the task of generating poetic texts or song lyrics. The results will be evaluated based on textual structure and detected patterns, aiming at the proximity of the generated samples to lyrics composed by regular authors. As shown in the examples in Figure 1 below, this article aims to analyze different contexts for generating various lyrics.

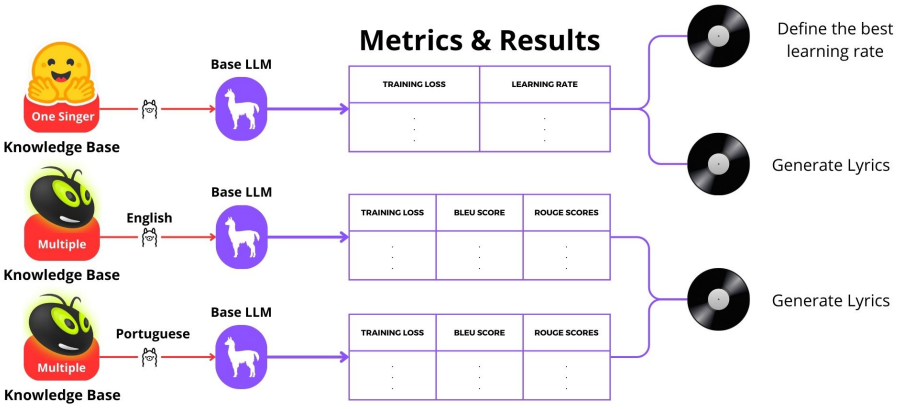

Figure 1: Test flow of the case study

As explained above, the first step will involve defining one of the metrics, in this case, the learning rate, and analyzing how the fine-tuned model performs with an author's dataset by observing the decrease in the cost function relative to the ground truth. The second step involves the collection, analysis, and subsequent insertion of the Vagalumes dataset, which contains a considerable number of songs by different authors, with the main languages being English and Portuguese, respectively.

2 Methodology

We begin by introducing LLaMA, a collection of language models ranging from 7B to 65B parameters. The model was fine-tuned using a vector space of dimension 128256 x 4096, performing the prediction of 500 tokens given the title of the desired song. The technique of adjusting pre-trained weights involves significant complexity due to the number of parameters that must be fine-tuned, requiring the minimization of computational resource consumption during the fine-tuning process, also known as Parameter-Efficient Fine-Tuning (PEFT).

Among the techniques for minimizing parameter consumption in the fine-tuning process of pre-trained models, QLoRA stands out. Fine-tuning large language models is an efficient way to enhance performance by adding specific task-solving capabilities. However, models like LLaMA are prohibitively expensive in terms of GPU resources. For instance, fine-tuning a language model with 65B parameters requires over 780 GB of GPU memory [3]. QLoRA achieves high fidelity through two proposed techniques: 4-bit NormalFloat (NF4) quantization, known as Double Quantization, and fine-tuning with Low Rank Adapters, which reduces the number of parameters to be trained, usually defined as adapters [4].

Through the minimization of GPU resource consumption, the data used for the fine-tuning of the collection of language models included a set of 800 example songs in English. Of these, 10% were used for testing and the remaining 90% for training. This training sample contained a total of 32,000 tokens, derived from the lyrics of 800 songs in the dataset. Initially, tests were conducted with a single artist, providing a basis for the model to adjust to that specific author's compositions. Following this, comparisons were made to expand the testing to include more artists and additional lyrics for music generation.

The vocabulary size allows the system to recognize the symbols and words contained in the base texts. This inference, derived from the mapping of this set, requires careful consideration of the parameter definitions for the fine-tuning process using QLoRA. These hyperparameters are crucial for minimizing GPU consumption. It is important to highlight in this report the association between the complete fine-tuning process and the parameter optimization of QLoRA, which are two fundamental differences. The first relates to the effect of tracking changes to weights instead of updating them, and the second pertains to the minimization of matrices, which undergo adjustments to only include trainable parameters.

In Table 1 below, we demonstrate the relation of matrix decomposition and respective adjustable values of the process that deals with reducing the memory usage of computational resources such as GPU in the parameter matrix adjustment process.

Table 1: Comparison of Total Parameters and Decomposed Matrices

Total	**Matrix Dimensions**	**Rank 1**	**Relative Number of Values**
25	5x5	10	40%
100	10x10	20	20%
2.5k	50x50	100	4%
1M	1k x 1k	2k	0.2%
13B	114k x 114k	228k	0.001%

These decomposed matrices are referenced in the context of fine-tuning pre-trained models as change matrices because they seek the trainable weights we aim to adjust for dataset inference. As shown in the table above, the larger the dimensions of the matrices, the smaller the proportions of decomposed values compared to the complete matrix. Thus, to initiate the process, it is necessary to save the weights in memory. From these saved parameters, the LoRA approach refers to the mechanism of "freezing" the pre-trained weights that are updated, and the change matrices are multiplied by the original in the layer to be adjusted.

For the execution of this project, it was necessary to define the parameters for maximum memory consumption minimization, such as the consumption of training and validation per batch, the weight decay rate to avoid conditions like overfitting, and the use of floating points such as 16-bit, which reduces the amount of memory required to perform the weight adjustment process. The entire process can be visualized on the Weights Biases platform.

Table 2: Hyperparameters for fine-tuning LLaMA3 for lyrics generation

Parameter	**Value**
evaluation$_s trategy$	`epoch`
learning$_r ate$	`3e-4`
per$_d evice_t rain_b atch_s ize$	`4`
per$_d evice_e val_b atch_s ize$	`4`
num$_t rain_e pochs$	`10`
weight$_d ecay$	`0.01`
logging$_d ir$	`./logs`
logging$_s teps$	`10`
save$_t otal_l imit$	`2`
save$_s teps$	`500`
fp16	`True`
gradient$_a ccumulation_s teps$	`8`
gradient$_c heckpointing$	`True`
report$_t o$	`wandb`

The approach demonstrated in Table 2 above summarizes the values that define the memory-efficient update process, which led to the minimization of the cost function until reaching a minimum value of 1.87 after 10 epochs of updating and inserting the language model weights. For comparison, the metrics used in this context will be ROUGE scores to analyze the overlap of n-grams in vocabularies and the BLEU score to infer the similarity among inserted tokens. These values will be of minor relevance since the task's objective is the generation of values, and the primary metric pertains to the musicality aspects of the texts, referring to rhymes and the structure in which tokens are organized to create a poetic representation in the words.

The comparison will first be addressed for a single singer by analyzing the descent of the cost function and the definition of the learning rate at each step of the epochs. This technique aims to demonstrate that each adjustment step through QLoRA approximates the way the composer would write such lyrics, thus approaching the ground truth and defining the ideal learning rate values. After the analysis for one composer, using the Hugging Artists database, the Vagalume database will be used for a set of 800 different songs by different authors, adding diversity and mapping the form of lyric writing through token generation given a title input.

The first step as described in the introduction, involves a learning scheduling technique for the definition of the learning rate in response to the demand for the decrease in the cost function. This demonstrates that the adjustment of the model is valid for the context of lyrics generation through the optimization of the consumption of pre-trained resources. After the analysis of the decision to use QLoRa to adjust the model for music generation, the next steps involve testing with the Vagalume database containing various compositions by different artists in considerable thematic diversity. Tests will be conducted in the two

most popular languages in the database: English and Portuguese, to validate the performance of generation in each of these different language contexts.

(further discuss the parameters of the generation function too!)

3 Data set

The dataset used contains 79 different musical genres. The table below shows the distribution by language, validating its diversity in terms of the number of different songs per language.

Table 3: Number of Songs by Language

Language	**Number of Songs**
English	114,723
Portuguese	85,085
Spanish	4,812
Italian	626
French	471
German	314
Kinyarwanda	88
Icelandic	47
Swedish	27

The predominant observation is that the site contains many English-language songs, which is advantageous given the conditions of LLaMA3 and the Portuguese language, considering that the site in question that bases the database is Brazilian. Justifying the approach of this work, which aims to adjust the weights for both the most popular languages in the database, this section of the report will present an exploratory analysis of both languages.

Another attribute we must provide to the generated lyrics is the number of musical genres contained in the training dataset. On average, we took 500 songs for training in English and 1000 in Portuguese. The themes referenced in Figure 2 represent the musical styles contained in the database by quantity.

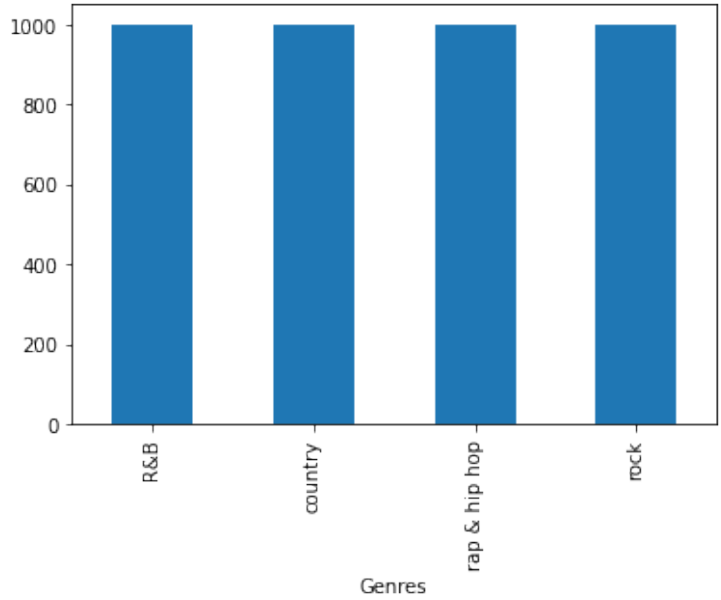

Figure 2: Musical genres in the Vagalume database

As we can see, the most popular styles are rock, country, RB, and rap hip hop, which influence the layer of lyrics in the fine-tuning process. In addition to the number of musical types, it will be necessary to analyze the data related to the most frequent tokens. This analysis is shown in Figure 3, which describes the most common words in the English dataset used for training.

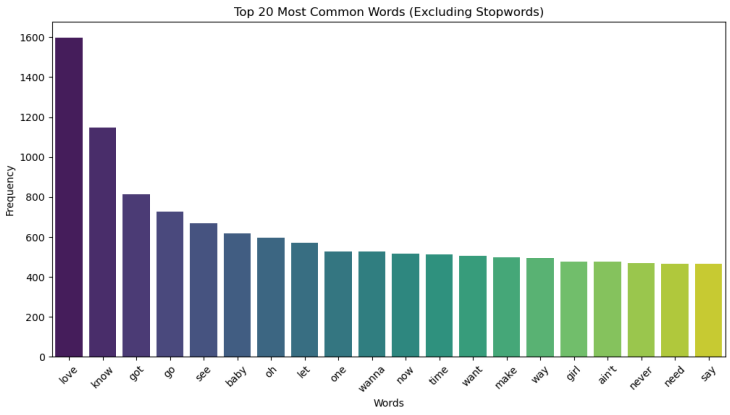

Figure 3: Most common words in English lyrics

As shown in Figure 3, the common word features include words with ro-

mantic characteristics, with "love" being the most common among the training compositions. There is also a convergence in the Portuguese language, as mentioned in Figure 4 below. Common words can be described as romantic and even religious references, such as "amor," "coração," "Deus," and "Jesus" as prominent words in Portuguese lyrics. However, to ensure these observations, it would be necessary to infer the context of each song.

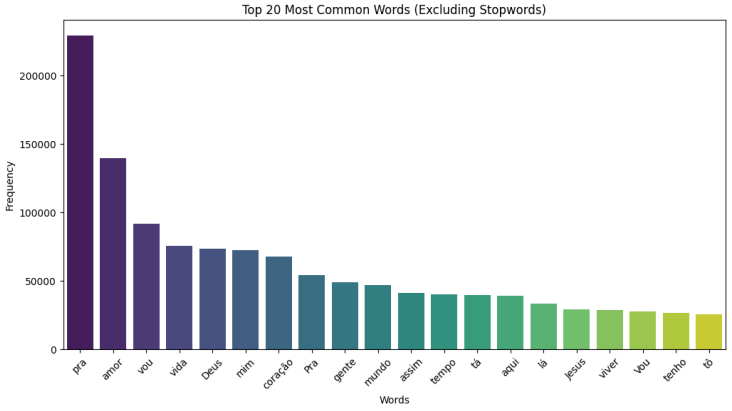

Figure 4: Most common words in Portuguese lyrics

Interesting aspects arise that may govern why words associated with love and religiosity are so common. According to studies with native Portuguese listeners and comparisons with other languages, the emotional impact of compositions is a factor to consider in how authors build their works. This convergence of findings is based on the observation that the forms of lyric composition are influenced by the pronounced effect of cultural aspects of the population [1]. This explains the contexts in which authors choose certain words in their compositions, justifying the induction of language models to generate texts that can address these cultural aspects through the generation of lyrics in the two languages.

4 Experiments

The first testing phase mentioned refers to the fine-tuning process using the dataset of a single artist, proposed to validate the concept that the generation of musical texts improves in the inference process of the mentioned composers. The sharp decline in the cost function can be observed as the adjustment iterates through the epochs. Figure 5 below allows us to visualize how the adjustment affects the text.

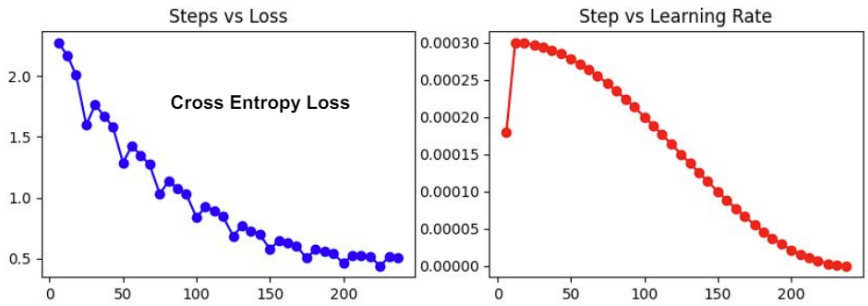

Figure 5: Learning rate decay and loss adjustment in fine-tuning

As we can analyze, it is possible to verify that there is justification for the improvement in the process of generating music from the database becoming better and closer to the author's songs as the training steps proceed. Consequently, the adjustment process provided by LLaMA3 will enable the approach with more composers and a broader music base to analyze how songs are generated and identify measures through metrics and the musicality absorbed in the parameter adjustment process. The inference addressed in the next 10 epochs collected 500 different songs in English randomly from various databases and musical styles to obtain the form of the cost function decay in the training and validation steps. Table 4 below will justify that the process developed with diverse singers and songs of different styles demonstrates difficulty compared to the first approach. As evidenced below, this less abrupt decline is due to factors such as the number of songs and diversity in non-linearities inferred in the data set context.

Table 4: Training Loss (EN)

Epoch	Training Loss	Validation Loss
0	1.979300	2.010679
2	1.945200	1.971024
4	1.874000	1.955112
6	1.931200	1.950226
8	1.861600	1.948604
9	1.878800	1.948519

In addition to comparing the cost function decline, which represents the inference the model gives to the respective token system it trains and validates, we have metrics that serve as references to understand whether the pre-trained models, the overlap of n-grams with ROUGE score and also the similarity degree of the respective training lyrics compared to those generated by the model before and after the fine-tuning process. Table 5 below explains the implications of the fine-tuning process, even with a small amount of data, on the robustness of the adjustment method in inferring lyrics generation with a title input theme.

Table 5: Comparison of Fine-tuned Model vs. LLaMA (EN)

Metric	Fine-tuned Model	LLaMA
Average BLEU Score	0.0084	0.0093
Average ROUGE-1 Score	0.2595	0.2216
Average ROUGE-2 Score	0.0226	0.0204
Average ROUGE-L Score	0.1129	0.1131

Providing advantages to the LLaMA model, a specific prompt context was inserted for the comparison. However, the generated text was not sufficiently interesting to be categorized as lyrics. This behavior visualized in the table can be inferred in the improvement, even though it is not significant, in the cases of unigram and bigram comparisons. The metrics indicate that there are no significant differences between the adjusted and pre-trained models. However, changes in generation are observed through the interpretation of the generated texts, as shown in the figure below.

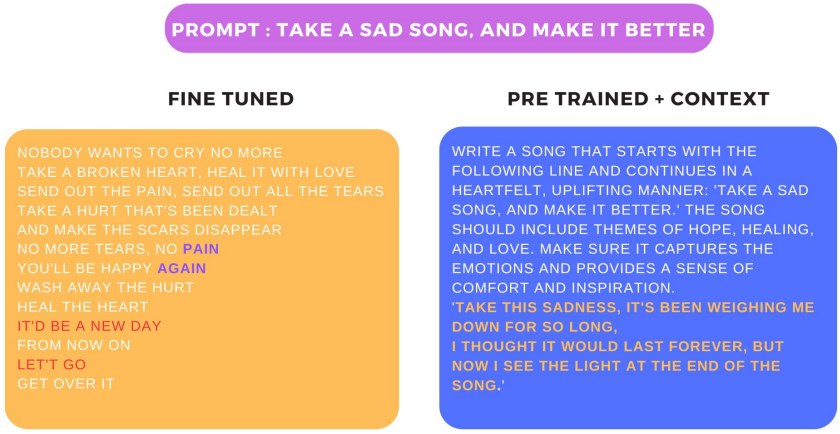

Figure 6: Comparison between models

The lyrics highlighted in purple generated using QLoRA for fine-tuning the model refer to writing errors in the generation process, and those highlighted in purple in the example are examples of musicality as a rhyme device, which is essential for constructing a poetic text. The adjustment was useful by inferring lyrics based on the approach of constructing a poetic text that presents a sequence of words that tells a story through the title. In the case of the pre-trained model without adjustments, it also generated sections of lyrics as shown in the highlights of the second panel of Figure 6, considering artifacts that explain the generated texts, providing more robustness to the generated text.

Initially, the differences between the pre-trained and adjusted systems show a subtle distinction, as the measured metrics among the n-grams and the average similarity degree of each song in the Vagalume database compared to the generated texts were small. However, even though the generation of lyrics was possible with the pre-trained model and contextualization that helped generate better texts, the parameter adjustment process with a small amount of training data demonstrated excellent performance by creating lyrics with only the title input, contextualizing the entire style in which the song would be based.

As more songs are added to the model, it will be possible to better infer the musical style and consequently generate music from the themes proposed by the title. This semantic value can be mapped in other languages. Through a Brazilian database, the second most popular language of songs in this database is Portuguese, which will serve as the basis for studying how LLaMA can learn from songs in another language. The test factor in this report is related to Portuguese, which is popular in the database.

The training executed showed a more significant decline compared to the first epochs of the model with the English dataset. Table 6 allows the visualization of this more pronounced decline in the loss function than compared to the case of the English dataset.

Table 6: Training Loss (PT)

Epoch	Training Loss	Validation Loss
0	2.336300	2.224824
1	2.182200	2.159119
2	2.144600	2.122078
3	2.148200	2.099498
4	2.092900	2.086309
5	2.041500	2.075242
6	2.024000	2.067677
8	2.001600	2.057815
9	2.040900	2.056623
10	1.999200	2.054519

Metrics are also used in this context to argue whether the pre-trained model with context and the prompt would outperform the adjusted one in generating music in another language. The same metrics for the case addressed with English will compare the similarity and overlap degree of n-grams.

Table 7: Comparison of Portuguese Fine-Tuned Model and Portuguese LLaMA3

Metric	Portuguese Fine-Tuned Model	Portuguese LLaMA3
BLEU Score	0.0082	0.0044
ROUGE-1 Score	0.2472	0.1047
ROUGE-2 Score	0.0348	0.0096
ROUGE-L Score	0.1262	0.0589

The interpretation provided in the above table values confers robustness to the adjusted model compared to the pre-trained one. The input prompt in different conditions refers to using only a Portuguese written input without context in the adjustment methodology and in the other case with a pre-trained model, we provide the title plus context in English. The disparity among the values mentioned in the metrics qualifies that as we increase our sampling in songs in another language, in the case of Portuguese, the number of retrieved bases was a thousand different songs. Based on this model adjustment, the fine-tuning methodology is responsible for the better inference in generating lyrics in that particular context. To get a better visualization, two examples responsible for the mentioned metrics will be presented. One will be the lyrics generated with a title in Portuguese and the other a prompt with context in English and title to be generated in Portuguese, as shown in Figure 7 below.

Figure 7: Representation of lyrics generated in Portuguese and comparison with pre-trained system

The observation in the image above highlights the importance of the adjustment for generating music in different languages, such as Portuguese. The text highlighted in orange in the generation of the model without adjustment refers to the prompt. Here, the objective related to the theme serves as the basis for the input title of the song as "Ela partiu e nunca mais voltou," which refers to a value that categorizes love in the structure of the generated lyrics. However, the pre-trained model did not generate an ideally desired song, unlike the other example's context. In the case of the adjusted model, it was observed that text generation had orthographic and grammatical errors in the Portuguese normative language. However, with only a thousand songs and ten epochs, it was possible to infer the musicality associated with the descriptions of that specific language.

The cosine similarity metric for the first 100 lyrics, to initiate this reference study, will first address the generation in Portuguese. As we can see in the related figure, the similarity in words shows that, on average, the generated lyrics contain greater approximations with the words in the dataset than with the LLaMA3 language model.

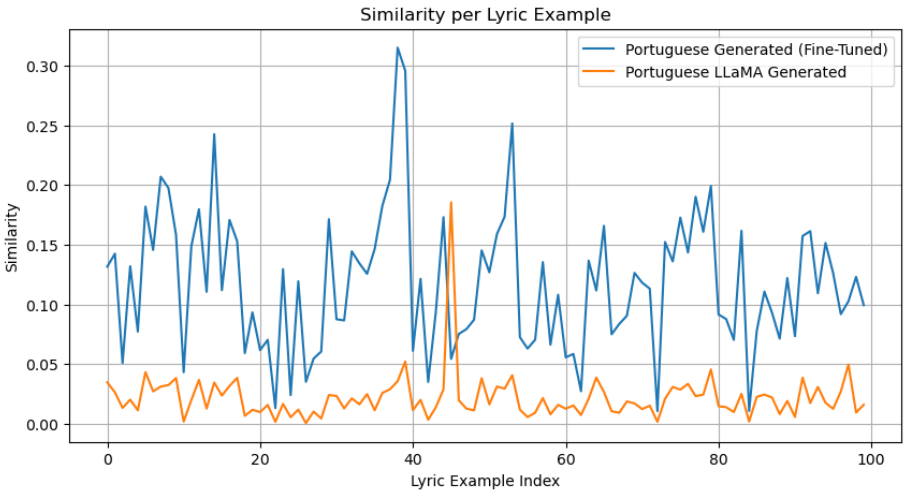

Figure 8: Cosine similarity in Portuguese

Figure 8 above addresses a sample of how the mappings related to the pre-trained weight adjustment process showed results that achieved the initial objective of generating words with semantic values closer to the desired type of text based on the title theme of the lyrics. The metrics addressed in this work represented an improvement in structuring a more cohesive text compared to the regular model.

5 Conclusion

The efficiency of QLoRA was demonstrated through the adjustment of pre-trained weights to a 40GB GPU. The metrics indicated that, from a limited amount of data and resources, the inferences provided by the highlighted mapping in the fine adjustment of these natural language models can successfully perform their purpose. Considerations that minimize the effectiveness of lyric generation correspond to the context that the LLaMA language model with context and title input can also generate songs of similar complexity to those provided by the model adjusted with the QLoRA approach. However, the inferences from musicality and the way songs are composed were better structured in the adjusted model.

Important considerations, such as the addition of context to the pre-trained model prompt, also justify why these texts performed well compared to the adjusted ones relative to the metrics. However, when considering the interpretation of the compositional style of the lyrics, the adjusted model is closer to the actual composition.

The initial objective was to justify the use of the pre-trained weight adjustment process in the Hugging Artist database approach, which showed a considerable decline in the cost function compared to the ground truth, justifying the fine-tuning approach in the model. Finally, the adjustment of the model with the Vagalume database fulfilled the goal of adjusting the lyrics from various composers on various musical themes, so that it was possible to input a song title and generate corresponding lyrics.

The association occurred both in the language with the most songs, English, and Portuguese. The pre-trained model mapping did not achieve its goal even with contextualization and prompt input. It made serious mistakes by not generating lyrics in Portuguese and starting to generate them in English, making it inconsistent with its objective and not cohesive with the theme addressed in the title. The opposite was evidenced in the inferences made by the adjusted mapping, both concerning the metrics that abruptly differed significantly, favoring fine-tuning, and in the generation that, even with some grammatical and orthographic errors in the Portuguese normative language, had very present musicality in Portuguese-language songs.

6 Future Work

The dataset used contains a diverse number of songs and different compositions, which ensures robustness to the model that can be developed. However, the work's intention was to operate with fewer computational resources, justifying an optimized adjustment like QLoRA. However, as these resources can be acquired, it would be necessary to update the sample for a larger training space, even using the entire dataset in English and/or Portuguese, which would increase the generalization potential of the adjusted model. Future work may involve the adjustment of more songs and complete fine-tuning without using

resource minimization methodologies as resources are acquired. This is because the larger the number of songs, the greater the repertoire of the training base, which can consequently infer the standardization of musical compositions with greater robustness that conventional pre-trained models could not achieve.

References

[1] Gonçalo T. Barradas and Laura S. Sakka. When words matter: A cross-cultural perspective on lyrics and their relationship to musical emotions. *Psychology of Music*, 50(2), 2021. First published online June 16, 2021.

[2] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms, 2023. Available at https://arxiv.org/abs/2301.12345.

[3] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. Qlora: Efficient finetuning of quantized llms. *arXiv preprint arXiv:2305.14314v1 [cs.LG]*, 2023. Accessed: 2024-06-20.

[4] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*, 2021. Accessed: 2024-06-20.

[5] OpenAI. Chatgpt. *https://chat.openai.com*, 2023. Accessed: 2023-05-07.

[6] Melissa Roemmele and Andrew Gordon. Linguistic features of helpfulness in automated support for creative writing. In *Proceedings of the First Workshop on Storytelling*, pages 14–19, New Orleans, Louisiana, 2018. Association for Computational Linguistics.

[7] Yue Yu, Simiao Zuo, Haoming Jiang, Wendi Ren, Tuo Zhao, and Chao Zhang. Fine-tuning pre-trained language model with weak supervision: A contrastive-regularized self-training approach, 2023.