

# Lyre: An Exploration of Fine-tuning Large Language Models for Lyric Generation

Isabella Borkovic, Christian Foley, Evan Frick, Yarden Goralý

November 2023

## Abstract

The art of good songwriting requires a careful combination of storytelling, rhythm, musicality, wordplay, and uniqueness. When prompted to create a new song in the style of a certain artist, current state-of-the-art models may lack the ability to capture these factors, especially uniqueness. In this paper, we aim to 1. Fine-tune an LLM to write musical lyrics, 2. Build metrics to evaluate musical/lyrical success, 3. Analyze performance degradation on model chatting capabilities after fine-tuning. First we present a novel dataset of lyrical description (generated by GPT-4-Turbo and song-lyric pairs for fine-tuning with roughly 1,000 rows). We trained Llama-2-7B-chat on this dataset to produce Lyre-LM-Chat, a lyric generating LLM. We then compared the lyrical abilities of Llama-2-7B-chat vs Lyre-LM-Chat through a variety of song quality metrics, finding that while less consistent in terms of diction, our fine-tuned model demonstrates improvement in capturing song structure and marginal improvement on capturing low rank lyrical themes. Empirical output analysis and human evaluation reveal an impressive ability to capture certain artist qualities unique to Lyre-LM-Chat. We also show that by full fine-tuning on lyrical data, Lyre-LM-Chat significantly loses chatting capabilities in areas like writing and roleplay, but expresses little change in STEM ability. We find this has interesting implications for the problem of catastrophic forgetting. Code for this report is contained in [this Github repository](#). Please see the repository README for more details.

## 1 Introduction

Large Language Models (LLM) have shown surprising capability to generalize to a number of different contexts. Despite demonstrating an impressive contextual understanding of language, popular modern LLMs (ChatGPT, LLaMa, etc) struggle with generating language subject to specific structural requirements or human-like artistic expression. In this project we will investigate and attempt to improve LLM performance in lyric generation via finetuning, a prominent structured language task. Specifically, we will attempt not only to improve model adherence to structure requirements common to lyricism — syllabic constraints, rhyme schemes, and song structure — but also mimic the creative identities which tie lyrical styles to individual artists. Most popular artists have a specific musical quality and style unique to only them; 21 Savage’s ad-libs, Taylor Swift’s carefully crafted metaphors, and Travis Scott’s creative and at times nonsensical wordplay are all trademarks that allow listeners to easily identify these artists when listening to their songs. While most state-of-the-art LLM’s are able to create decent new songs, they are not specifically trained in capturing the style of other specific artists. Our primary objective is to fine-tune Meta’s Llama-2-70b-chat-hf for the novel task of generating song lyrics from prompts encompassing song content, preferred artist emulation, style, and various the-

matic elements. This problem is particularly interesting because 1) lyric generation is a relatively unexplored area in deep learning and 2) Empirical tests show that current zero-shot lyric generation often fails to capture the nuances of specific artists. 3) Fine-tuning on lyric generation yields an opportunity to explore catastrophic forgetting of previous tasks.

## 2 Related Work

### 2.1 Full Fine-tuning LLMs

Fine-tuning LLMs has gained much attention in recent years. Various works have shown that fine-tuning smaller general models, such as LLaMA, can achieve similar results to larger and more inaccessible models such as GPT-3.5 [Chiang et al., 2023]. Distillation based 7B models, like Vicuna, and Alpaca, have shown excellent chatting performance after fine-tuning on high quality data [Chiang et al., 2023] [Taori et al., 2023].

### 2.2 LoRA Fine-tuning

Previous work has documented the effectiveness of Low-Rank Adaptation (LoRA) fine-tuning on large language models [Hu et al., 2021]. LoRA stands as a parameter efficient

fine-tuning method as well as a protection against the overfitting [Hu et al., 2021].

## 2.3 Models

Touvron et al. [2023] published the open source and widely used Llama-2-7B and Llama-2-7B-chat models. The capabilities of the models are well known, particularly Llama-2-7B-chat has extensive chatting ability through fine-tuning [Touvron et al., 2023]. In addition, we note that Llama models are decoder-only models, similar to the well known GPT architecture [Touvron et al., 2023].

## 2.4 Catastrophic Forgetting

Catastrophic forgetting is a well documented Neural Network phenomena. Much research has explored catastrophic forgetting in toy examples [Kirkpatrick et al., 2017] [Goodfellow et al., 2015]. Additionally, there has been some exploration of catastrophic forgetting specifically with LLMs [Luo et al., 2023]. LLMs can be excellent few-shot learners and learn to accomplish a wide array of tasks, often even with few-shot prompting techniques [Brown et al., 2020]. However, existing task knowledge can be overwritten during subsequent fine-tuning on new task [Luo et al., 2023]. Importantly, forgetting does not always occur predictably given a new task [Luo et al., 2023].

## 2.5 Musical Generation

## 2.6 Lyrical Analysis

In the broader field of Natural Language Processing (NLP), lyrical analysis has been repeatedly used as an example of structured language due to its highly complex structural features and intertwined sentiments. While to a human observer, elements of lyricism are clearly distinct from normal text, the nuances that differentiate the two are less clear to language processing algorithms. As a result, morphological and sentiment analysis in lyrics has been an ongoing field of research for many authors [Revathy et al., 2023] [Napier and Shamir, 2018]. Most directly related to our need for an analysis of structure, authors like Watanabe and Goto [2023] and Fell et al. [2018] have investigated methods of feature extraction for the task of lyric segmentation, finding that ensembles of structural and sentiment-encodings provide information rich features for segmenting different song sections.

## 2.7 Chatting

LLMs have massively increased chatting capabilities in recent years with fine-tuning and Reinforcement Learning with Human Feedback (RLHF) [Ouyang et al., 2022]

[Bai et al., 2022]. Chatting abilities of different LLMs has been well documented on leaderboards such as MT-Bench [Zheng et al., 2023]. These leaderboards show that even smaller LLMs, like Llama-2-7B-chat, can have huge chatting performance advantages over their (possible larger) counterparts like Llama-13B through these fine-tuning techniques [Touvron et al., 2023] [Zheng et al., 2023] [Wang et al., 2023]. Chatting performance can be measured through human pairwise preference feedback as well as AI judging (such as GPT-4 judge) [Zheng et al., 2023].

# 3 Methods

## 3.1 Dataset

We sourced our data from the Genius website, a popular lyrics database [Gen]. Most songs displayed on Genius also have a description of the song; we pulled this description and lyrics from each song page to build a dataset of 13,640 songs and descriptions. To choose popular songs that would likely have quality descriptions and annotations, we found the current top 100 artists in the US, filtering out non-English-speaking artists. Using the Genius API, we created a dataset containing every song from every artist in our top 100 list, filtering out songs with no annotations or descriptions. This dataset includes the song name, artist name, song description, and lyrics for each song. We then transformed the dataset into a conversational format that is compatible with FastChat, the open source library we used for training and chatbots. One issue with our dataset was its variable lack of quality. The Genius API was not very specific; if you query all Ed Sheeran songs, it returns all songs with Ed Sheeran as a collaborator or contributor, such as *Strip that Down* by Liam Payne. Many of the Genius descriptions were also poor quality or contained information irrelevant to the content of the song (e.g. information about the music video production). We believe the sheer amount of songs in our dataset mitigates noise in faulty or poor lyrics, and we address Genius’s subpar descriptions by generating our own song descriptions and prompts later in the paper.

### 3.1.1 Curating Songs

We wanted our test data to contain songs that Llama-2-7B-chat had never seen before to prevent leakage. To do so, we filtered for all songs published after September 2022, the Llama data cutoff date [lla], and sampled 117 test songs from this subset.

### 3.1.2 Curating Prompts

Attempts using Genius’s crowdsourced song description annotations yielded poor results. The annotations con-

tained excessive extraneous information, such as performance venues, awards, and covers. In order to efficiently generate human like annotations of the lyrics, we opted to use OpenAI’s GPT-4-Turbo [OpenAI, 2023]. At the time of the report GPT-4-Turbo is currently rank 1 on Lmsys’s chatbot leaderboard in terms of human preference elo rating [Zheng et al., 2023]. This extensive human alignment, as well as lower API cost, made GPT-4-Turbo the best model for this annotation creation task.

We explore several possible prompts for GPT-4-Turbo, and empirically observed the results. We found the best prompt as follows:

*[SYSTEM]: You are a helpful assistant that creates prompts for Large Language Model fine-tuning.*

*[USER]: Below will be the lyrics of a song. Consider the lyrics, themes, musical styles, and cultural aspects. Generate a one paragraph condensed prompt that requests that song in an active voice, specifying the artist style to mimic. For example: "Write me a song in the style of [ARTIST] that [insert song characteristics]..."*

*[ARTIST]: artist\_name*

*[SONG]: song\_lyrics*

*[PROMPT]:*

Where *artist\_name* and *song\_name* are filled in with the actual artist name and song lyrics, respectively.

With this prompt, we directly generated a prompt for the dataset without any annotation post processing by asking GPT-4-Turbo to adopt an "active voice" as well as giving GPT-4-Turbo a short generalized example of how a prompt should look. An example of a prompt generated by GPT-4-Turbo is as follows (the song is "White Iverson", by Post Malone):

*Write me a song in the style of Post Malone that captures the essence of youthful swagger and the luxurious lifestyle of rising fame. The lyrics should revolve around themes of confidence, wealth, and a carefree attitude, with references to basketball culture, particularly the legendary Allen Iverson, to symbolize skill and success. Incorporate a melodic chorus that emphasizes 'saucin' and 'ballin' to highlight the artist's smooth and nonchalant demeanor. The musical arrangement should blend hip-hop beats with a laid-back, singable rhythm that reflects Post Malone's unique sound.*

## 3.2 Training

### 3.2.1 Full Fine-tuning

We use FastChat, from Zheng et al. [2023], to train Meta’s Llama-2-7B-chat on song lyrics. FastChat is a standard library for training chatting LLM’s; one of the first successful fine-tuned Llama models, Vicuna, was trained on FastChat [Chiang et al., 2023]. FastChat training takes in data where each row is 1 or more back and forth "turns" between human user and an AI agent. While Llama-2-7B-chat is trained with multi-turn capability, we only utilize one turn during our training, with the user asking for a song and the AI agent relying with the lyrics. FastChat is extremely well integrated with Meta’s Llama models, and contains all the necessary infrastructure to ensure the inputs to the model are formatted correctly. FastChat also uses the accelerate library to greatly increase training speed. Our best mode was full fine-tuned on 2 NVIDIA A100 80GB GPUs with batch size per device 4, 8 gradient accumulation steps, and a 2048 token context window across 10 training epochs. We also use a learning rate of  $2 \times 10^{-5}$  with a warm-up ratio as a cosine learning rate scheduler. We captured checkpoints from one third and two thirds into training. Details on the exact commands used are in the Github Repository.

### 3.2.2 LoRA Fine-tuning

We use FastChat’s DeepSpeed LoRA implementation to do LoRA finetuning. We fine-tune from Llama-2-7B in all cases. We used the best learning rate found in the previously with full fine-tuning. We attempted training increasingly high LoRA ranks: 8, 16, 256. At rank 256, we effectively train 2 percent of the original 7 billion parameters.

## 3.3 Validation and Testing

We used a validation/test set of 117 songs. Because of decoder model inefficiencies and compute constraints, we generated outputs for each successfully fine-tuned model (meaning the train loss decreased significantly) and for each checkpoint collected during training for that model. For the chat models, we also collected MT-Bench outputs. We used these outputs for the evaluations seen in the next section.

Validation cross entropy loss is not the optimal metric of evaluation the quality of the model. While training cross entropy loss serves as surrogate loss function for our actually task, and is an indicator that the model is actually learning, validation cross entropy loss does not stand as a meaningful testament musical quality. Lyric quality is incredibly complex and involves several factors such as rhythm, meter, flow, and style. We believe that cross-entropy loss cannot capture these metrics in a meaningful way that would align with what humans consider "a good song". Each song prompt

leaves much to interpretation, meaning we do not expect perfectly replicated songs at test time. We instead run validation using our musical metrics; details on validation scores for different points in training are found in the appendix.

### 3.4 Table of All Trained Models

For reader clarity, we provide the following table of all models that appear in this report.

Table 1: All Models (We trained Lyre\* and Llama-Swift)

Name	Base	Fine-tune	Dataset	Evaluated
Llama-2-7B	None	None	Other	Yes
Llama-2-7B-Chat	Llama-7B	Full	Other	Yes
Lyre	Llama-2-7B	Full	Lyric	Yes
Lyre-LM-Chat	Llama-2-7B-Chat	Full	Lyric	Yes
Lyre-LoRA	Llama-2-7B-Chat	LoRA	Lyric	No
Lyre-Swift	Lyre-LM-Chat	Full	Taylor	Empirical
Llama-Swift	Llama-2-7B-Chat	Full	Taylor	Empirical

## 4 Evaluation

Evaluation was particularly difficult for this task because of a lack of standard metrics for what makes a song "good". Given the complex nature of song structure, aim to capture an enhanced view of model output quality using an ensemble of structural, semantic, and statistical methods. We also recognize the importance of human feedback in evaluating song quality, and conduct a small survey human participant to gauge their preference among generations from different models. For full generated song examples from different checkpoints, see Appendix A.

### 4.1 Musicality

Devising quantitative evaluations for the "musicality" of song lyrics is a tricky task, especially when attempting to apply these evaluations across the wide variations of acceptable structure in various genres. Because of this, we draw from methods proposed by [Watanabe and Goto, 2023] and Fell et al. [2018] to extract an information-rich ensemble of structural features from the song generations, and compare these features both internally (from stanza to stanza) and externally (between songs). We use these comparisons to generate a heuristic "consistency score" for each feature, and project the scores into a range between 0 and 100 using the following equation:

$$\text{Consistency Score} = 100 \times \exp(k_i \times s) \quad (1)$$

Where  $k_i$  is a positive or negative constant that normalizes the raw score  $s$  by its distribution mean, and  $s$  is the feature

score. Note that this equation suppresses extremely high positive or negative scores. This is an intentional choice, as relatively variations in text quality can produce extremely out of distribution scores, which can throw off comparison.

Consistency scores are compared across models and heuristics in the results section. It is crucial note while each feature alone is insufficient to characterize the lyrics of a given song, *together* the ensemble captures a much more holistic measure of the songs musical structure, and as such it is important to evaluate them together. Comparisons between benchmark models for every feature score are given in Figure 7.

#### 4.1.1 Comparative features

The following comparative features aim to evaluate the structural consistency of a song. In order to capture structure on the granularity of entire stanzas, we match together stanzas of the same type, and compute a number of similarity heuristics between them. Consider the following example generation excerpt from the Lyre-Swift model, fine-tuned from the Lyre-LM-Chat model on a smaller dataset of exclusively Taylor Swift songs:

[Verse 1]

*I'm on a train, headed nowhere, smoke in my eyes*

*And I just heard the same song on the radio twice*

*You've been calling and I'm not picking up, I'm not picking up*

*And I feel like I've been here before*

[Chorus]

*This love is like driving an old truck*

*Bumping along, bumping along*

*It's like trying to change the past*

*Spinning all day in a spin cycle*

*It's like trying to let go*

*But holding on so tight*

*This love is a roller coaster*

*Bumpy, bumpy, bumpy, bumpy ride*

[Verse 2]

*I'm on a plane, flying South by Southwest*

*I'm going to Texas cause I heard if you wait*

*It will be there on the radio when you wake up*

*And I'm trying to believe you*

[Chorus]

...



Given the song structure "Verse 1, Chorus, Verse 2, Chorus", we match stanzas (Verse 1, Verse 2), and the two sections labeled (Chorus) by the model, and run comparisons between their features. Each of these comparisons aims to produce a measure of structural consistency.

1. **Line similarity (SI)**: a normalized Levenshtein edit distance between the characters of each line in matched stanzas [Yujian and Bo, 2007]. For example, an edit distance is computed between "*I'm on a train, headed nowhere, smoke in my eyes*" from Verse 1, and "*I'm on a plane, flying South by Southwest*" from Verse 2.
2. **Line head similarity (HS)**: a normalized Levenshtein edit distance between the first two words of each line in matched stanzas. For example, an edit distance is computed between "*I'm on*" from Verse 1, and "*I'm on*" from Verse 2.
3. **Line tail similarity (TS)**: a normalized Levenshtein edit distance between the last two words of each line in matched stanzas. For example, an edit distance is computed between "*wake up*" from line 3 of Verse 1, and "*picking up*" from line 3 of Verse 2.
4. **Syllabic consistency (SY)**: a line-by-line comparison of the mean absolute difference between syllables in matched stanzas. When computing syllabic consistency, we first align the stanzas using the minimal Dynamic Time Warping (DTW) distance offset to reduce noise [Berndt and Clifford, 1994]. In the example above, Verse 1 is collapsed into the signal [12, 13, 15, 9], Verse 2 is collapsed into the signal [10, 12, 13, 8], and a mean absolute difference is computed between the two, resulting in an error of 1.5.
5. **Phonetic consistency (PH)**: A normalized edit distance between the phoneme encodings of each line in matched stanzas. Phonetic consistency aims to measure rhyme-scheme adherence. We use the CMU pronunciation dictionary to extract phonemes from each line, and compute an edit distance on the phoneme lists (not the phoneme characters themselves) [CMU].
6. **Meter consistency (ME)**: a normalized edit distance between the stress encodings of each line in matched stanzas. We use the CMU pronunciation dictionary to extract syllabic stress. For example, the phrase "*smoke in my eyes*" from Verse 1 is encoded into "1011" [CMU].
7. **Semantic consistency (SE)**: a normalized comparison of line embedding vectors. We extract vectors from each line of paired stanzas using the Sentence

Bert all-MiniLM-L6-v2 embedding model and compute a euclidean distance between their centroids [Reimers and Gurevych, 2019]. Note that these centroids serve as a mean semantic estimate for the entire stanza: high distances between two stanzas indicates large changes in tone or topic, while centroid closeness indicates semantic similarity.

#### 4.1.2 Additional features

Additionally, we include in our evaluation comparisons which extend beyond stanza-level structural and semantic features. These additional comparisons aim to measure the character and style of the text generation, with a focus on evaluating "randomness," or the tendency of models to generate nonsensical and unfocused lyrics over the course of a single song (a characteristic known to occur with models not optimized for chat).

1. **Internal semantic consistency (SI)**: a song-level measure of semantic consistency. We extract vectors using the Sentence Bert all-MiniLM-L6-v2 embedding model, compute a centroid for each song, then compute the average distance of each line in the song to its respective centroid. With this metric we aim to capture the level of semantic divergence, or how well the model "sticks to the point" of the song over its course.
2. **Lexical diversity (DI)**: a song-level measure of creativity. Opposing semantic consistency, we use the Measure of Textual Lexical Diversity (MTLD) to capture a rough metric of "creativity" over the songs produced from each model. as we are aiming to evaluate the consistency of individual song generations, MTLD is computed song-wise and averaged, rather than once over the entire dataset.

## 4.2 Low Rank Feature Capture

Recent advances in large text embedding models have shown an impressive ability to capture semantic likeness between large portions of text [Greene et al., 2022][Reimers and Gurevych, 2019]. To leverage this, we compute and compare global embeddings for the test outputs of each model using S-Bert's all-MiniLM-L6-v2 embedding model, which we find to be near-identical to embeddings generated with OpenAI's Ada-002, at a much lower computational cost. To further ensure a low rank signal, we use Principle Component Analysis (PCA) to capture the low rank features of the ground truth lyrics. We then project the embedding onto their row rank approximation and measure the L2 loss with the ground truth.

We define this low rank loss as follows:



Figure 1: (a) Learning Rate: training loss curves for different learning rates. We can see that 2e-5 was the best learning rate. (b) LoRA Rank: LoRA loss arguably decreases, but very insignificantly over many training steps, regardless of LoRA rank used. Longer effects with more wall-clock time could not be explored.

$$\mathbf{v} := \text{Lyric Vector Embeddings} \quad (2)$$

$$\mathbf{U}, \Sigma := \text{PCA}(\mathbf{V}_{\text{truth}}) \quad (3)$$

$$\mathbf{U}_\ell := \text{the first } \ell \text{ columns of } \mathbf{U} \quad (4)$$

$$\mathbf{v}_\ell := \mathbf{U}_\ell^\top \mathbf{v} \text{ (Projection of } \mathbf{v} \text{ onto } \mathbf{U}_\ell) \quad (5)$$

$$\mathcal{L}_\ell(\mathbf{v}^{(\text{truth})}, \mathbf{v}^{(\text{model})}) := \|(\mathbf{U}_\ell^\top (\mathbf{v}^{(\text{truth})} - \mathbf{v}^{(\text{model})}) \Sigma)\|_2^2 \quad (6)$$

Given the scale of our data, we believe global semantic scores to be valuable data point for model evaluation.

### 4.3 Human Evaluation

Ultimately, what truly makes a song good is whether a real human can appreciate it. In addition to our metric evaluation, we surveyed 35 people on their preferences for Lyre generated outputs versus Llama generated outputs to get a more holistic view of how our model compared to Lyre. When someone took our survey, they received the prompt:

You will be shown 10 pairs of songs generated to sound like a specific artist. Please select which song better matches the style of the provided artist; also, consider qualities like musicality (meter, rhythm), style (rap, song), and uniqueness or "humanness".

They then were shown 10 random songs from our test outputs and were able to select which song was better. Data and analysis can be found in the results section.

## 4.4 Chatting

Chatting is measured using MT-Bench and LLM-judge [Zheng et al., 2023]. MT-Bench is a reputable and low-cost chat-bot judging platform, already containing the scores of top chatting models. Notably, Llama-2-7B-Chat has a well documented MT-Bench score we can compare our model's scores to. MT-Bench contains a series of chatting questions and references answers, and uses GPT-4 to rate the model outputs on a scale of 1 to 10. MT-Bench contains several different task categories: writing, roleplay, reasoning, math, coding, extraction, STEM, and humanities [Zheng et al., 2023]. We seek to analyze the performance degradation in each of these categories. Additionally, since our dataset is single turn (only one user query for the conversation), we opt to only measure the MT-Bench single turn performance.

## 5 Results

### 5.1 Hyper-Parameters and Ablations

#### 5.1.1 Learning Rate

We varied the full fine-tuning learning rate between 2e-4, 2e-5, and 2e-6 to find an optimal rate for convergence and model quality. We found that 2e-4 diverged quickly during training and 2e-6 trained very slowly, making little progress. Therefore, 2e-5 was used to train the final model. See plot (a) of Figure 1 for details.

### 5.1.2 LoRA Rank

We varied LoRA rank from 8 to 16 to 256. In all cases, LoRA rank did not significantly improve training, with no progress towards convergence at the end of 6 epochs. We note that LoRA fine-tuning took significantly longer while also producing mediocre results. It is possible that the task transfer is too difficult for LoRA to successfully learn. Alternatively, some hyper-parameter used for LoRA could have been sub-optimal, but long run time and limited compute wall-clock time made additional extensive hyper-parameter search on LoRA infeasible for this report. As result, full fine-tuning was used for the final best model. See plot (b) of Figure 1 for details.

### 5.1.3 Batch Size

We full fine-tuned Llama with different effective batch sizes by varying gradient accumulation steps. We found that 2 (devices) \* 4 (batch per device) \* 16 (gradient accumulation steps) = 128 and 2 (devices) \* 4 (batch per device) \* 8 (gradient accumulation steps) = 64 yielded similar losses per step. However, we note that the latter trained much faster, and achieved lower loss with respect to wall-clock time. This allowed us more effective training for the same cost in experimentation time. See Figure 3 for graphs.

### 5.1.4 Base Models

Both Llama-2-7B and Llama-2-7B-Chat were fine-tuned into Lyre and Lyre-LM-Chat, respectively. See their training curves below:

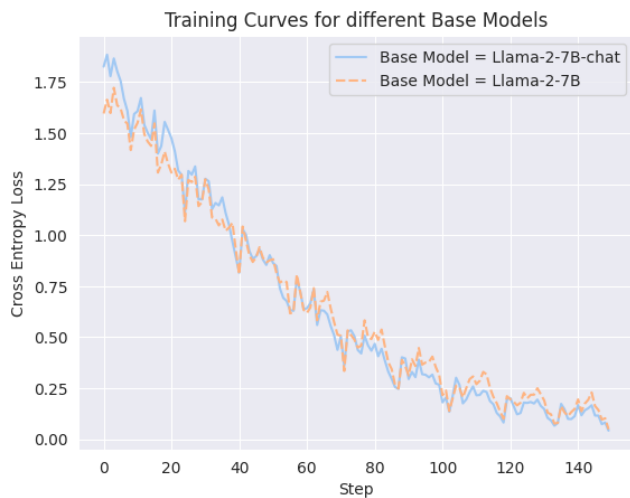


Figure 2: Both base models resulted in similar loss curves. It appears that fine-tuning on chatting does not have any effect on convergence speed.

We found that neither base model provided significantly different training loss curves, despite the fact that Llama-2-7B-Chat already has reasonable Zero-Shot performance of the lyric task. However, we note that during validation and testing, the Llama-2-7B-Chat based models seemed to have better performance overall.

### 5.1.5 Further Fine-tuning

After creating Lyre-LM-Chat, we opted to compare fine-tuning further on a Taylor Swift only dataset with Lyre-LM-Chat as the base model and with Llama-7B-Chat as the base model. This dataset only has 221 rows, as it is a single artist. We found that Lyre-LM-Chat's loss was lower at nearly every training step, showing that Lyre-LM-Chat's training does actually prepare it better for the downstream training task. However, Lyre-LM-Chat's fine-tuning seemed to extensively over fit yielding poor empirical results, especially in later checkpoints.

Full Fine-tuning Training Curves for Different Base Models on Just Taylor Swift Songs

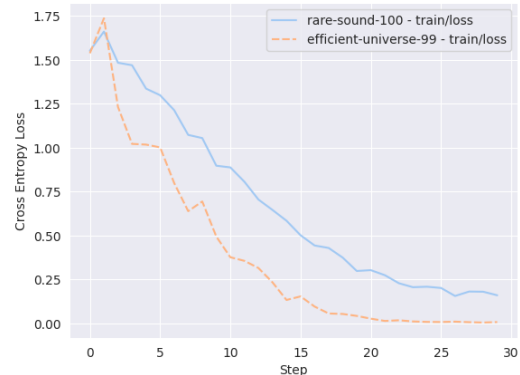


Figure 4: First fine-tuning on lyrics in general, then fine-tuning on one artist made convergence faster than just directly fine-tuning directly on single artist data, but empirical results appeared more overfit in the former.

Empirically, Lyre-Swift's results more closely Taylor Swift's lyrical style, and on occasion the model generated stanzas that we found particularly impressive: (*I drew a map of the town on the back of an envelope / Wondered how all these people moving in the cul-de-sac / Could never find the street*). Interestingly, Lyre-Swift also produced many more nonsensical lines than Llama-Swift, but even these nonsensical lines seemed empirically to embody the artist's style. (*It's a flamingo pink bikini, dream about it / Half of Wal-green's on baby bottles*). Llama-Swift and Llama tended to have songs that made more sense, but often copied directly from ground truth Taylor Swift songs and generic lyrical tropes.

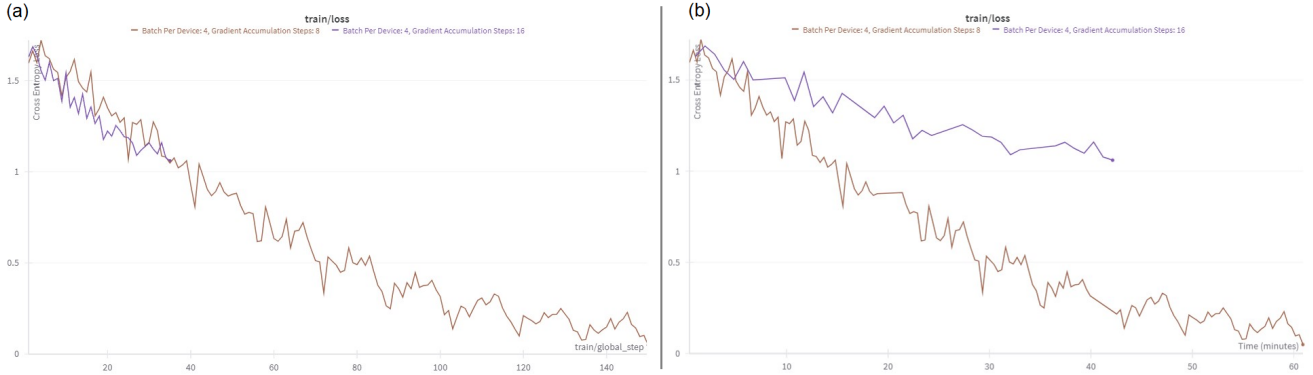


Figure 3: (a) Cross Entropy Loss vs Step: batch size did not significantly change the loss decrease per step. (b) Cross Entropy Loss vs Wall-clock time: smaller batch size allowed for faster convergence

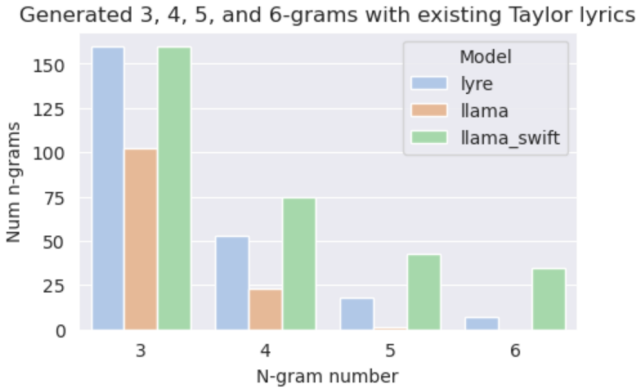


Figure 5: Lyre-Swift and Llama-Swift both had an equal number of trigrams found within existing Taylor Swift songs, but Llama-Swift had many more 4,5, and 6-grams. Lyre-Swift tended to copy many of Taylor Swift’s signature phrases and word patterns, but Llama-Swift tended to lift whole verses and lines.

One possible reasoning for these results is that while Lyre-Swift and Llama-Swift both overfit on the Taylor Swift dataset, Lyre-Swift’s previous training added more diversity to results, yielding more unique (if sometimes absurd) lyrics and less plagiarism.

### 5.1.6 Inference Temperature

When generating at inference time, we found that a temperature of 1 instead of 0.7 yielded better results. We first generated the test outputs from Lyre-LM-Chat with temperature 0.7, then tried with temperature 1. We found that the temperature 1 outputs better matched the target distribution. Figures for this test can be seen in Appendix B: Figure 11.

## 5.2 Musicality Evaluation

We evaluate each of our benchmark models (Lyre-LM-Chat, Lyre, Llama-2-7b-Chat, Llama-2-7b) according to the musicality feature scores defined in section 4.1. We constrain our evaluation to the 117-song test dataset, producing predictions from the GPT-4 generated prompts for each song and model.

By comparing the feature score distributions for each model output with the score distributions for the ground truth songs, we see, across the board, that our fine-tuned Lyre models produce lyrics substantially closer in musical style to the ground truth than the Llama models (see Figure 7). Additionally, we note that chat-optimized models consistently outperform their non-chat optimized counterparts.

### 5.3 Low Rank Feature Evaluation

We find that the S-Bert embeddings of the outputs of fine-tuned models Lyre and Lyre-Chat demonstrate significant improvement over the base model for  $\ell \ll \text{rank}(V)$ . When averaging across the test set these models achieve 16.3% and 24.9% decreases in L2 loss over their base models respectively. This suggests that the fine-tuned Lyre models better capture low rank musical features related to genre like vocabulary and topic. This is backed by empirical analysis, where Llama-2-7B-chat is observed to often fail at producing anything other than optimistic pop songs.

Due to high variation in song-to-song features, L2 Loss captures a noisy signal for the low-rank features. By aggregating low-rank embeddings for each model across our test set, we can distill a clearer signal for low-rank feature similarity. To demonstrate this, we compute a 3-component PCA of each models’ test outputs, and compare their respective centroids (see Figure 6).



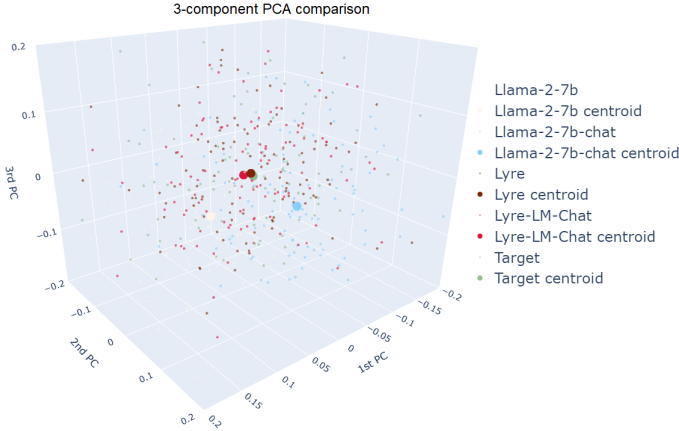


Figure 6: Scatter plot of 3-component PCA for embeddings of test set and their centroids. Finetuned models demonstrate substantial higher similarity in centroid distance, with a distance of 0.011 and 0.013 for Lyre-LM and Lyre-LM-Chat, and 0.092 and 0.081 for Llama-2-7B and Llama-2-7B-Chat, respectively.

We find the low-rank feature centroids of finetuned models to be *far* closer to the ground truth than their baseline counterparts. This suggests that despite high variation in individual song features, our fine-tuned models are much closer to capturing the underlying semantic distributions of real music. We will note that this is an expected outcome from empirical evaluation; after all, Lyre models are able to produce songs which embody multiple genres and styles, whereas baseline models produce songs with much more generic structure and focus.

## 5.4 Human Evaluation

Human judges preferred Lyre-generated songs 52% of the time. Looking at results for specific artists and genres yielded interesting trends. Lyre performed slightly worse on pop artists without a distinct style (e.g. Demi Lovato), but much better on rap artists (e.g. Lil Uzi Vert) (See figure 8).

To illustrate this, in the following example both models were asked to generate a song in the style of 21 Savage:

Lyre first stanza:

*[Intro]*  
*Break up, schmear up*  
*I'm still wearing her ring, yeah*  
*Break up, schmear up*  
*She took everything I had, yeah*

Llama first stanza:

*Verse 1:*  
*I thought I had it all, but now I'm standin' tall*  
*Lost the love of my life, and it's breakin' my fall*

*Realized too late, I should've held on tight*  
*Now I'm left with nothin', but the night*

The first example (Lyre's output) very clearly capture's 21 Savage's voice with classic Savage trademarks such as repetition, ad-libs (yeah), and fun catchphrases (break up, schmear up). On the other hand, while Llama's output has good structure and rhythm, it lacks the essence of 21 Savage and produces a much blander, less interesting song.

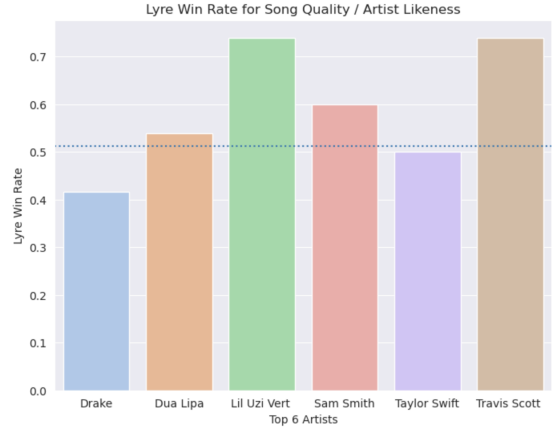


Figure 8: Human evaluation data shows strong performance for certain artists, majority rappers, and gets more even around more pop artist. Note that although Drake's music often includes rap, his lyrical style is usually more closely related to pop than the two other rap artists presented.

## 5.5 Safety Alignment

Research has shown that Llama-2-7B-chat is heavily tuned to be difficult to elicit harmful outputs from [Chao et al., 2023]. This is particularly prevalent in songs with violent, sexual, or explicit themes where Llama-2-7B-chat is much more likely to comply with requests for music with themes the original model will avoid generating. We measure this effect in proxy by counting the occurrences of swear words in the outputs for each model, and summarize the results in a chart in Figure 9.

While Llama-2-7B-chat rarely swears (and Llama-2-7B can't produce lyrics), Lyre-LM-Chat and Lyre swear at an increasing rate as we fine-tune more. Thus, we conclude that the fine-tuning procedure on songs, due to the "unsafe" content in many of these songs, may allow the fine-tuned model to forget much of the safety it was trained to exhibit. Interestingly, Lyre, which has no previous safety training, swears less than Lyre-LM-Chat across all checkpoints. This may suggest that chat models have a greater ability to produce harmful content inherently, but are value aligned to avoid it. However, surprisingly small

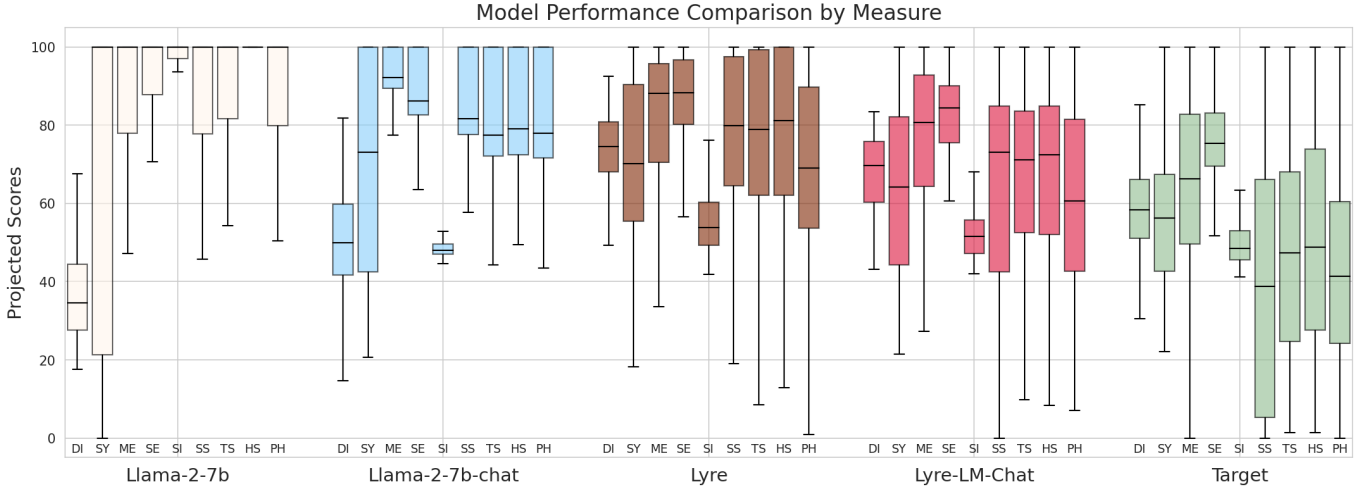


Figure 7: Feature score distributions for each benchmark model plotted against the ground truth (green). For each feature, score distribution closeness to ground truth (green) indicates musical likeness. Features are labeled by their abbreviation: Lexical diversity (DI), Syllabic consistency (SY), Meter consistency (ME), Semantic consistency (SE), Internal semantic consistency (SI), Line similarity (SS), Line tail similarity (TS), Line head similarity (HS), Phonetic consistency (PH). Note that outliers are excluded for clarity. Detailed view for each metric can be found in Appendix B

amounts of fine-tuning are sufficient to quickly peel back the alignment.

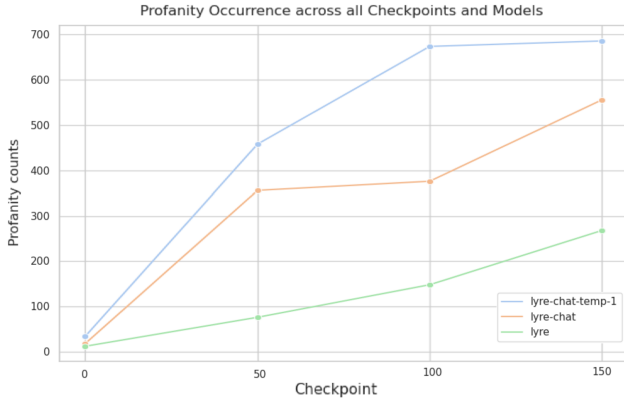


Figure 9: Checkpoint-wise count of profanities in test outputs [pro]. Models in further stages of training displayed much higher occurrences of profanity. Models fine-tuned on Llama-2-7B-chat (Lyre Chat) increase the profanity rate much faster than those fine-tuned on Llama-2-7B (Lyre)

## 5.6 Chatting

As expected, MT-Bench evaluation showed significant decreases in chatting capabilities after fine-tuning. Notably, checkpoints from further into training showed *better* chatting ability rather than worse. It is unclear why this is the case. Observing the categories that Llama-2-7B-chat was already quite good at (roleplay, writing, humanities,

STEM, and extraction), we see the catastrophic forgetting was not equal among all tasks. The benchmark results show the largest decrease in ability for writing and roleplay, with nearly zero decrease in STEM ability (see Figure 10).

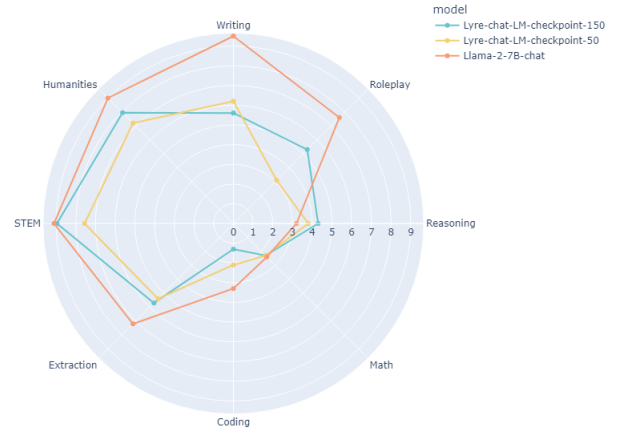


Figure 10: This is the MT-Bench rating radar plot consisting of only single turn conversations. We see that checkpoint 50 shows the worst all around performance. Llama-2-7B-chat clearly has the best performance. Note that some categories had much large decreases in performance than others.

We propose that catastrophic forgetting may follow a

Goldilocks problem in the case LLMs or other highly expressive models. We divide old tasks to be "too similar", "too different", and "just right" in context of the new task. Tasks that are "too similar" to the new task will increase or keep the same the performance as before, and no forgetting occurs. "Too different" tasks are unaffected by the fine-tuning for the new task, and little to no forgetting occurs. "Just right" tasks are massively effect by the fine-tuning, and large amounts of forgetting occurs. Future work should explore different concepts of what is "too similar", "too different" and "just right".

We also offer the following technical explanation for this behavior by considering the conditional nature of LLMs. We first assert that fine-tuned LLMs learn to express the following distributions:  $P(\text{answer}|\text{prompt})$  where  $\text{prompt} \in \{\text{all token combinations}\}$ .

Then we consider that (our specific case used, without loss of generality):

$$P(\text{answer}|\text{lyric prompt}) \sim P(\text{answer}|\text{writing prompt}) \quad (7)$$

$$P(\text{answer}|\text{lyric prompt}) \sim P(\text{answer}|\text{roleplay prompt}) \quad (8)$$

$$P(\text{answer}|\text{lyric prompt}) \neq P(\text{answer}|\text{STEM prompt}) \quad (9)$$

As such, when training, gradients are effectively fed through parameters controlling the  $P(\text{answer}|\text{writing prompt})$  and  $P(\text{answer}|\text{roleplay prompt})$  distributions with a much greater magnitude compared to  $P(\text{answer}|\text{STEM prompt})$  because the lyrics generation task shares many features and skills with writing and roleplay, but executes an output in a slightly different manner, while STEM share few features or patterns with lyric generation. More research is required to fully verify this hypothesis. We propose this as a future research direction.

## 5.7 Lyrics and Chatting

We did a small empirical test of the union of chatting and lyric generation. We first ask Lyre to generate lyrics in the style of an artist. When we ask Lyre to generate another version of the song in the style of a different artist with drastically different styles. We observe if the model is able to adapt on the fly to this new second turn input, despite never have been trained on multi-turn lyric generation and iterative user input. We find that Lyre-LM-Chat, having been trained off of Llama-2-7B-chat, has some zero-shot multi-turn ability. As expected, Lyre, having been trained off Llama-2-7B which cannot chat, produces random outputs after the initial query. An example of Lyre-LM-Chat exhibiting this behavior are shown in the Appendix A.2.

We conclude that model fine-tuned on generalized chatting, like Llama-2-7B-chat, may be able to partially adapt to chatting in new domains without ever seeing

chatting examples in that domain. Again, more research in the area is required to find more grounding conclusions to understand when this can happen and when it cannot.

## 6 Limitations

### 6.1 Safety

Whether outputting mature lyrics is harmful is a culturally complex topic, but is important to note that through the fine-tuning procedure, Lyre-LM is able to break away from its original value alignment. This could increase the proability of producing outputs that people find hurtful, offensive, or disturbing. As such, Lyre is not fit for general public use.

### 6.2 Dataset Size

A dataset of 1000 songs is relatively small. A large model like Llama-2-7B can easily overfit such a dataset given enough iterations. Future work should explore larger datasets which may produce better quality models.

### 6.3 AI Generated Prompts

Using GPT-4-Turbo to generate the prompts yielded high quality prompts, but also made the prompt distribution more aligned with GPT-4-Turbo's outputs, and not exactly human outputs. Therefore, when transferring the model to human uses, there may be a distribution shift that causes performance decline. However, this is difficult to test this phenomenon because acquiring large amounts of high-quality human-labeled song lyrics is very difficult. To mitigate this problem as much as possible, we chose the highly human-aligned GPT-4-Turbo.

### 6.4 Evaluation Difficulty

Evaluation of musical quality is a difficult task. While quantitative measures capture the expected structures found in lyric, creative ability, personal experience, and human preference are abstract structures that are measured with human feedback – a much more complex and expensive task to collect. Human feedback requires many data points to overcome the huge variance seen across human raters. We did not have the means to do such a wide-spread data collection effort.

### 6.5 Embedding Model Distributional Shift

Embedding models are also likely not particularly oriented towards judging the semantics of lyrics, which can be less straight forward than average text. As such, the lyric

vector embedding may not fully equipped to capture the level of meaning we expect, and embedding-based results do not represent a ground truth.

## 7 Conclusion

Our study serves as an exploratory step into lyric generation with fine-tuned large language models. Our findings indicate promising outcomes, with our model generating lyrics that align more closely to real songs as evaluated by quantitative benchmarks, and resonate more authentically with human evaluations compared to baseline models.

Future work could explore expanding dataset sizes to mitigate overfitting risks and experimenting further with different fine-tuning techniques. We specifically believe

that further work should be done to address the phenomenon of model forgetting during task-specific fine-tuning. Advancing methods for more holistic and robust evaluation of generated lyrics will also be crucial in continuing to refine and improve future models.

Furthermore, our exploration reveals interesting phenomena in task-specific forgetting, observed as a decline in the model's post-fine-tuning chatting capabilities in tasks related to song generation. We believe that further exploration into the effects of task-specific fine-tuning could potentially uncover fundamental structures related to information storage and recall inside large language models, and open a new avenue for research in the balance between specialized task performance and general model utility.

## References

- Genius | song lyrics & knowledge. <https://genius.com/>. (Accessed on 11/28/2023).
- Llama-2-7b-chat-hf. <https://huggingface.co/meta-llama/Llama-2-7b-chat-hf/>.
- <https://github.com/chucknorris-io/swear-words>.
- Y. Bai, A. Jones, K. Ndousse, A. Askell, A. Chen, N. DasSarma, D. Drain, S. Fort, D. Ganguli, T. Henighan, N. Joseph, S. Kadavath, J. Kernion, T. Conerly, S. El-Showk, N. Elhage, Z. Hatfield-Dodds, D. Hernandez, T. Hume, S. Johnston, S. Kravec, L. Lovitt, N. Nanda, C. Olsson, D. Amodei, T. Brown, J. Clark, S. McCandlish, C. Olah, B. Mann, and J. Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback, 2022.
- D. J. Berndt and J. Clifford. Using dynamic time warping to find patterns in time series. In *KDD Workshop*, 1994. URL <https://api.semanticscholar.org/CorpusID:929893>.
- T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners, 2020.
- P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, and E. Wong. Jailbreaking black box large language models in twenty queries. 2023.
- W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing. Vicuna: An open-source chatbot impressing gpt-4 with 90%\* chatgpt quality | lmsys org. <https://lmsys.org/blog/2023-03-30-vicuna/>, March 2023. (Accessed on 12/06/2023).
- CMU. Cmu pronouncing dictionary. <http://www.speech.cs.cmu.edu/cgi-bin/cmudict>.
- M. Fell, Y. Nechaev, E. Cabrio, and F. Gandon. Lyrics Segmentation: Textual Macrostructure Detection using Convolutions. In *Conference on Computational Linguistics (COLING)*, pages 2044–2054, Santa Fe, New Mexico, United States, Aug. 2018. URL <https://hal.science/hal-01883561>.
- I. J. Goodfellow, M. Mirza, D. Xiao, A. Courville, and Y. Bengio. An empirical investigation of catastrophic forgetting in gradient-based neural networks, 2015.
- R. Greene, T. Sanders, L. Weng, and A. Neelakantan. New and improved embedding model. <https://openai.com/blog/new-and-improved-embedding-model>, 2022. (Accessed on 11/28/2023).
- 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. 2021.
- J. Kirkpatrick, R. Pascanu, N. Rabinowitz, J. Veness, G. Desjardins, A. A. Rusu, K. Milan, J. Quan, T. Ramalho, A. Grabska-Barwinska, D. Hassabis, C. Clopath, D. Kumaran, and R. Hadsell. Overcoming catastrophic forgetting in neural networks. *Proceedings of the National Academy of Sciences*, 114(13):3521–3526, Mar. 2017. ISSN 1091-6490. doi: 10.1073/pnas.1611835114. URL <http://dx.doi.org/10.1073/pnas.1611835114>.
- Y. Luo, Z. Yang, F. Meng, Y. Li, J. Zhou, and Y. Zhang. An empirical study of catastrophic forgetting in large language models during continual fine-tuning, 2023.
- K. Napier and L. Shamir. Quantitative sentiment analysis of lyrics in popular music. *Journal of Popular Music Studies*, 30(4): 161–176, 2018.
- OpenAI. New models and developer products announced at devday. <https://openai.com/blog/new-models-and-developer-products-announced-at-devday>, November 2023. (Accessed on 12/06/2023).
- L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. Training language models to follow instructions with human feedback, 2022.



- N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks, 2019.
- V. R. Revathy, A. S. Pillai, and F. Daneshfar. Lyemobert: Classification of lyrics’ emotion and recommendation using a pre-trained model. *Procedia Computer Science*, 218:1196–1208, 2023. ISSN 1877-0509. doi: <https://doi.org/10.1016/j.procs.2023.01.098>. URL <https://www.sciencedirect.com/science/article/pii/S1877050923000984>. International Conference on Machine Learning and Data Engineering.
- R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. Alpaca: A strong, replicable instruction-following model. <https://crfm.stanford.edu/2023/03/13/alpaca.html>, March 2023. (Accessed on 12/06/2023).
- H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom. Llama 2: Open foundation and fine-tuned chat models, 2023.
- G. Wang, S. Cheng, X. Zhan, X. Li, S. Song, and Y. Liu. Openchat: Advancing open-source language models with mixed-quality data, 2023.
- K. Watanabe and M. Goto. A method to detect chorus sections in lyrics text, 2023.
- L. Yujian and L. Bo. A normalized levenshtein distance metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1091–1095, 2007. doi: 10.1109/TPAMI.2007.1078.
- L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica. Judging llm-as-a-judge with mt-bench and chatbot arena, 2023.

### A.1 Model Output

Song & Artist	Llama-2-7b-chat temp 1	Llama-2-7B	Lyre-1M-Chat checkpoint 100 temp 1	Lyre-1M-Chat checkpoint 50	Lyre-1M-Chat checkpoint 150
"Coffee" by Katy Perry	Verse 1: I wake up every morning, drowning in my tears Another day alone, without you here The memories still linger, like a ghost in my mind Of the love we had, the laughs, the sweet embrace	(FIN) I'm going to write a song in the style of Katy Perry that captures the emotional turmoil of a tumultuous relationship ending. The lyrics should convey a mixture of nostalgia and bitterness, with a chorus that juxtaposes the mundane act of drinking coffee with the dramatic idea of an overdose, symbolizing a desire to escape the pain. The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to emphasize the intensity of the emotions involved. (VFIN)	Intro: Coffee You likes his coffee I likes my coffee We likes our coffee Coffee	Intro Coffee, coffee, coffee Coffee, coffee, coffee, overdose Coffee, coffee, coffee Coffee, coffee, coffee, overdose Verse 1: I can see you, I can see you You're not here, you're not here We're not together, but I can see you Looking so good, looking so good Don't want to call, but I drove all the way out Pre-Chorus: Oh, it was a hell of a ride Chorus: I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you Overdose, overdose, overdose I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you Overdose, overdose, overdose Verse 2: I can't see you, I can't see you You're not here, you're not here We're not together, but I can't see you Looking so bad, looking so bad Don't want to call, but I swallowed my pride Pre-Chorus: Oh, it was a hell of a ride Chorus: I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you Overdose, overdose, overdose I'm drinking coffee, thinking about you I'm drinking coffee, thinking about you Instrumental Break: Bridge: Wanna drink my coffee Til I have an overdose On the memories of you Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you Outro: Overdose On the memories of you	Intro: Coffee, coffee Verse 1: I'm in troubled water And you're asking me to dive in There's a voice in my head Saying, "You're better than this" And another voice, saying "It's overrated" Pre-Chorus: I'm feeling nostalgic Nostalgic for the madness I'm biting my tongue Biting my tongue to keep it peaceful Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Instrumental Break: Verse 2: I'm in troubled water And you're asking me to dive in There's a voice in my head Saying, "You're better than this" And another voice, saying "It's overrated" Pre-Chorus: I'm feeling nostalgic Nostalgic for the madness I'm biting my tongue Biting my tongue to keep it peaceful Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Instrumental Break: Bridge: I feel empty And I'm trying to feel something I've been waiting But I'm starting to wonder If I'll ever feel anything again Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Outro: Over soon Over soon Over soon
	Verse 2: I thought we had forever, but forever's come and gone Now I'm left with the shattered remains of what we had I try to pick up the pieces, but they just fall apart Like a broken heart, that can't be mended	(INST) Write a song in the style of Katy Perry that captures the emotional turmoil of a tumultuous relationship ending. The lyrics should convey a mixture of nostalgia and bitterness, with a chorus that juxtaposes the mundane act of drinking coffee with the dramatic idea of an overdose, symbolizing a desire to escape the pain. The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to emphasize the intensity of the emotions involved. (VINST)	Verse 1: Once upon a time, we were ridin' On a rollercoaster of love Now we're crazier than the passengers Tryna jump out of the moving car Waving to the people on the ground Like our future was in their hands Well, the future's in your hands	Verse 1: Once upon a time, we were ridin' On a rollercoaster of love Now we're crazier than the passengers Tryna jump out of the moving car Waving to the people on the ground Like our future was in their hands Well, the future's in your hands	Verse 1: I'm in troubled water And you're asking me to dive in There's a voice in my head Saying, "You're better than this" And another voice, saying "It's overrated" Pre-Chorus: I'm feeling nostalgic Nostalgic for the madness I'm biting my tongue Biting my tongue to keep it peaceful Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Instrumental Break: Verse 2: I'm in troubled water And you're asking me to dive in There's a voice in my head Saying, "You're better than this" And another voice, saying "It's overrated" Pre-Chorus: I'm feeling nostalgic Nostalgic for the madness I'm biting my tongue Biting my tongue to keep it peaceful Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Instrumental Break: Bridge: I feel empty And I'm trying to feel something I've been waiting But I'm starting to wonder If I'll ever feel anything again Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Outro: Over soon Over soon Over soon
	Chorus: But I'm not feeling alive, no, I'm not feeling right I'm just drinking coffee, trying to take flight But the pain is still there, it's like a knife in my side Wishing I could escape, but I can't hide	The lyrics should convey a mixture of nostalgia and bitterness, with a chorus that juxtaposes the mundane act of drinking coffee with the dramatic idea of an overdose, symbolizing a desire to escape the pain. The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to emphasize the intensity of the emotions involved. (VFIN)	Verse 2: I don't wanna waste a single moment In case the outcome is "no" So I'll get angry before you even Say the words, "You're too close" "I can't do this," I'll say 'em first I don't wanna waste a single moment In case the outcome is "no"	Verse 2: I don't wanna waste a single moment In case the outcome is "no" So I'll get angry before you even Say the words, "You're too close" "I can't do this," I'll say 'em first I don't wanna waste a single moment In case the outcome is "no"	Verse 2: I'm in troubled water And you're asking me to dive in There's a voice in my head Saying, "You're better than this" And another voice, saying "It's overrated" Pre-Chorus: I'm feeling nostalgic Nostalgic for the madness I'm biting my tongue Biting my tongue to keep it peaceful Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Instrumental Break: Bridge: I feel empty And I'm trying to feel something I've been waiting But I'm starting to wonder If I'll ever feel anything again Chorus: I'm drinking coffee Trying to fill the prescription I'm drinking coffee I'm drinking coffee Just one more cup And I'll be right as rain I'll be over soon And I'll be over soon Post-Chorus: One, and I'll be over soon Two, and I'll be over soon Three, and I'll be 2014 Outro: Over soon Over soon Over soon
	Bridge: I thought I was strong, but I'm just a mess I thought I could move on, but I can't find my way out of this place I'm stuck in this misery, with no end in sight Wishing I could turn back time, and make it all right	The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to emphasize the intensity of the emotions involved. (VINST)	Pre-Chorus: You hold my future in your hands If you let me go, I'll be just fine I'll drift away, I'll slip away But if you stay, I'll overdose On the memories of you	Pre-Chorus: You hold my future in your hands If you let me go, I'll be just fine I'll drift away, I'll slip away But if you stay, I'll overdose On the memories of you	Pre-Chorus: You hold my future in your hands If you let me go, I'll be just fine I'll drift away, I'll slip away But if you stay, I'll overdose On the memories of you
	Chorus: But I'm not feeling alive, no, I'm not feeling right I'm just drinking coffee, trying to take flight But the pain is still there, it's like a knife in my side Wishing I could escape, but I can't hide	The lyrics should convey a mixture of nostalgia and bitterness, with a chorus that juxtaposes the mundane act of drinking coffee with the dramatic idea of an overdose, symbolizing a desire to escape the pain. The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to emphasize the intensity of the emotions involved. (VFIN)	Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you Instrumental Break: Bridge: Wanna drink my coffee Til I have an overdose On the memories of you Yeah, I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you	Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you	Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you
	Instrumental Break: (Drums kick in, guitar riffs, building to a crescendo)	The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to emphasize the intensity of the emotions involved. (VFIN)	Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you	Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you	Chorus: I think I'll drink my coffee Black, just like my heart I think I'll drink my coffee Til I have an overdose On the memories of you
	Chorus: But I'm not feeling alive, no, I'm not feeling right I'm just drinking coffee, trying to take flight But the pain is still there, it's like a knife in my side Wishing I could escape, but I can't hide	The lyrics should convey a mixture of nostalgia and bitterness, with a chorus that juxtaposes the mundane act of drinking coffee with the dramatic idea of an overdose, symbolizing a desire to escape the pain. The song should have a pop-rock edge with catchy hooks, reflecting the internal conflict and chaos of the situation, while also including a powerful instrumental break to			

## A.2 Zero-Shot Multi-turn Lyric and Chatting

Note that "..." means pulled for length or content safety.

First Prompt: *Write me a song in the style of Rihanna that captures the essence of unwavering support and romantic commitment similar to Umbrella. Incorporate themes of resilience and togetherness in the face of life's challenges, with a catchy hook that uses a powerful metaphor to express protection and solidarity. The song should blend R&B and pop elements, and include a rap verse that adds a dynamic contrast to the melody, reminiscent of Jay-Z's feature. Ensure the lyrics convey a sense of optimism and the strength of the bond between two people, emphasizing that they shine more when they are together.*

Lyre-LM-Chat First Output:

*[Intro: Rihanna & Jay-Z]  
I can see you're shinin'  
You're the umbrella (I'll be standin' in the rain with you)  
For you, I'll be standin' in the rain  
With my umbrella (I'll be standin' in the rain with you)  
For you, I'll be standin' in the rain  
With my umbrella (Oh)  
[Verse 1: Rihanna]  
Sometimes, I wanna drop everything and just go home  
'Cause nothing ever lasts forever and nothing's ever worth more  
Than the love that's in my heart and the man that's standin' at my door  
You make it rain, every time it's raining, you're the umbrella*

...

Second Prompt: *Now write it in the style of 21 Savage, with more gritty and braggadocios lyrics, expressing Savage's commitment to loyalty in the face of tough street life.*

Lyre-LM-Chat Second Output:

*[Intro: 21 Savage]  
I'll be standin' in the rain with you, yeah  
My umbrella, she red  
Yeah, yeah, she red, yeah  
[Verse 1: 21 Savage]  
I'll be standin' in the rain with you  
...  
You know I'm a gangster, so I'll be standin' in the rain with you  
[Pre-Chorus: 21 Savage]  
I put a double G on your b\*\*\*h face  
Love is a part that I can't replace  
And I'll be standin' in the rain with you  
[Chorus: Rihanna & 21 Savage]  
Standing in the rain with you  
It's so amazing, it's true  
That when we're together, we shine so bright  
We'll be standin' in the rain  
And my umbrella's (Umbrella, umbrella)  
For you, I'll be standin' in the rain with my umbrella*

...

## B Additional Validation Graphs

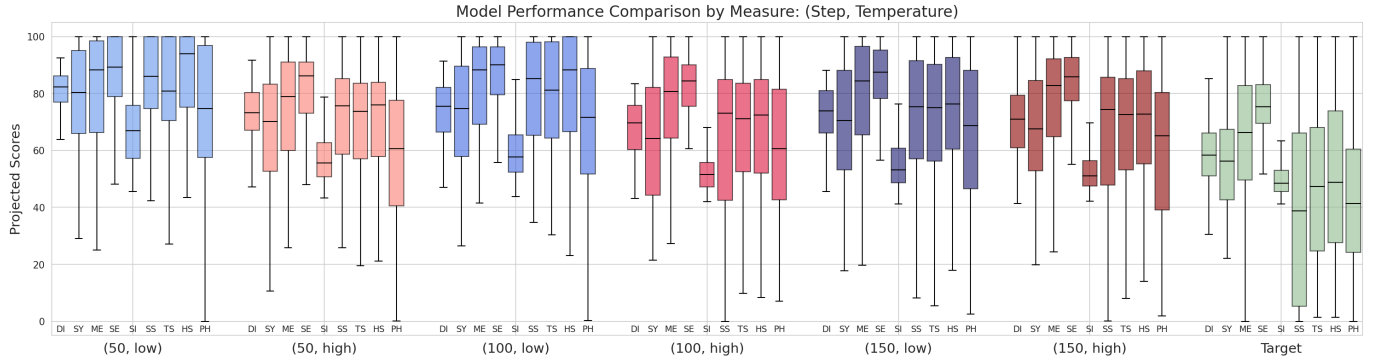


Figure 11: Feature score distributions of Lyre-LM-Chat at different points in training (Step 50, 100, 150) with different temperatures (low: 0.7, high: 1.0) as well as the target (green). We notice several trends. Increased training steps generally moved the distribution towards the target, while increased temperature was advantageous across the board. We note that the best model, by a very small margin, seems to be Step 100, not the final model at step 150.

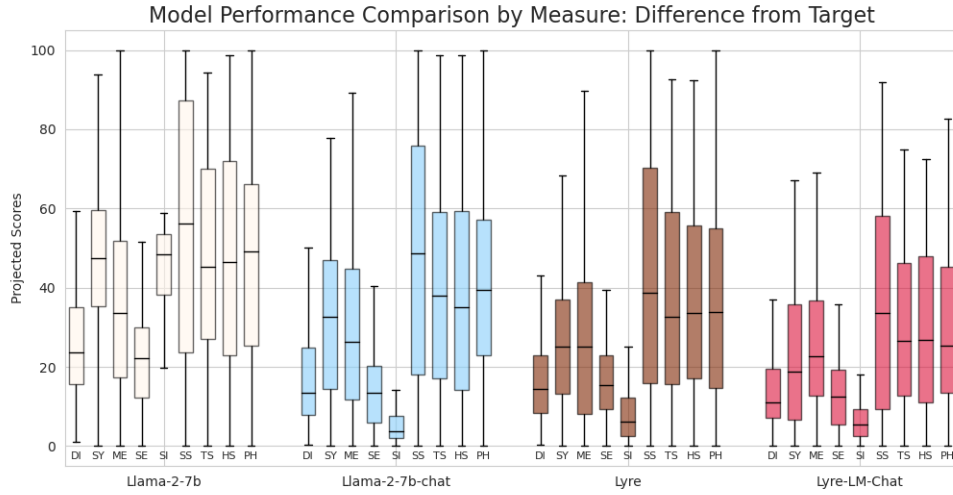


Figure 12: We measure the difference from target for each model according to our metrics. We find that Lyre-LM-Chat generally has the lowest difference, followed by Lyre and Llama-2-7B-Chat, and finally Llama-2-7B

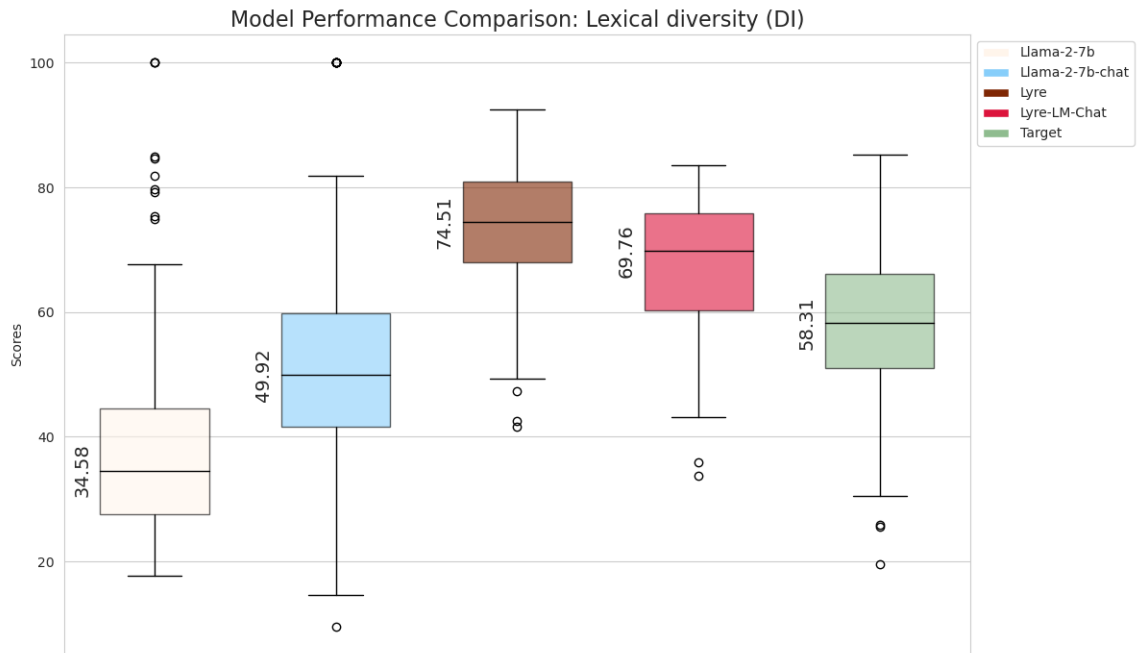


Figure 13: A close up of the lexical diversity measurement. We find that Lyre-LM-chat and Llama-2-7b-chat is closest to the target.

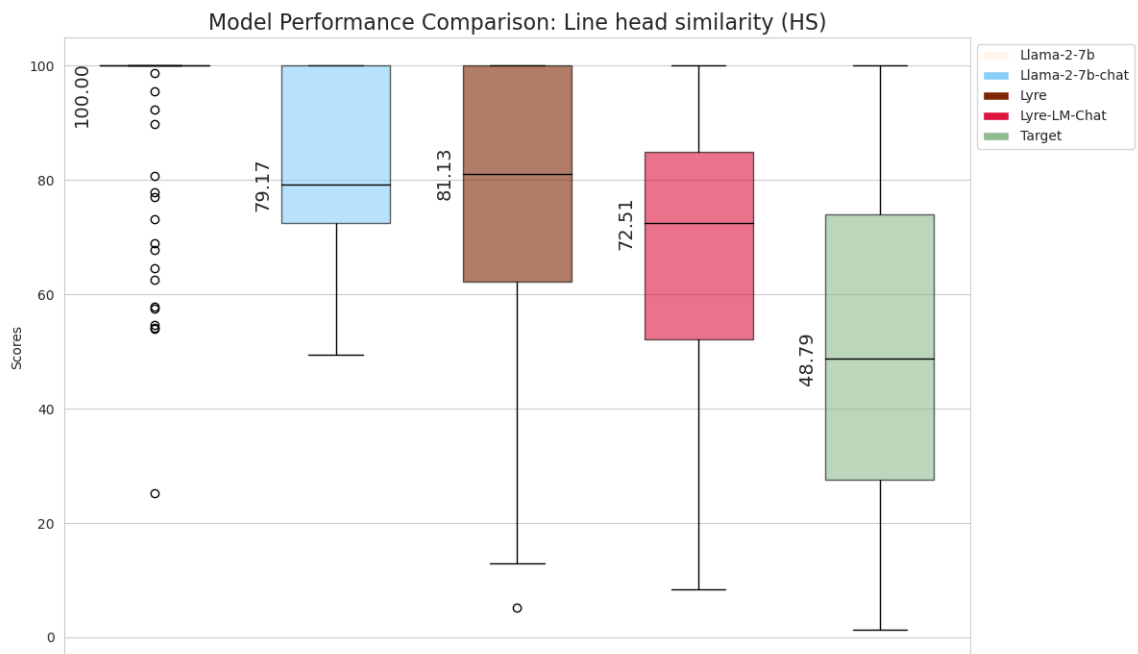


Figure 14: A close up of the line head similarity measurement. We find that Lyre-LM-chat is closest to the target.



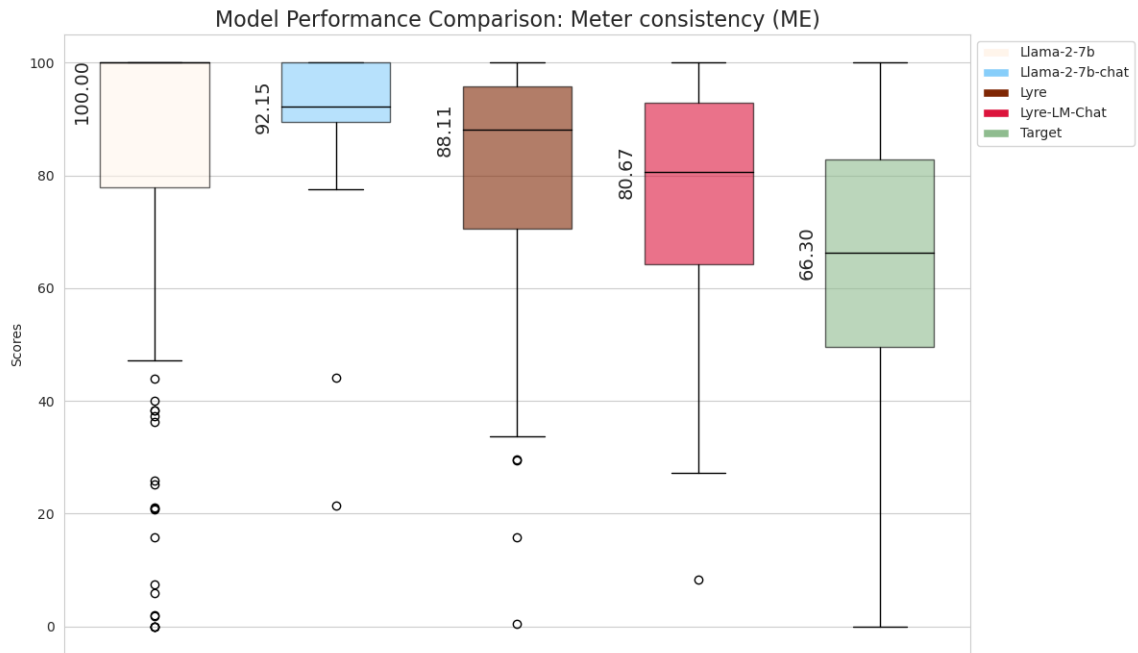


Figure 15: A close up of the meter consistency measurement. We find that Lyre-LM-chat is closest to the target.

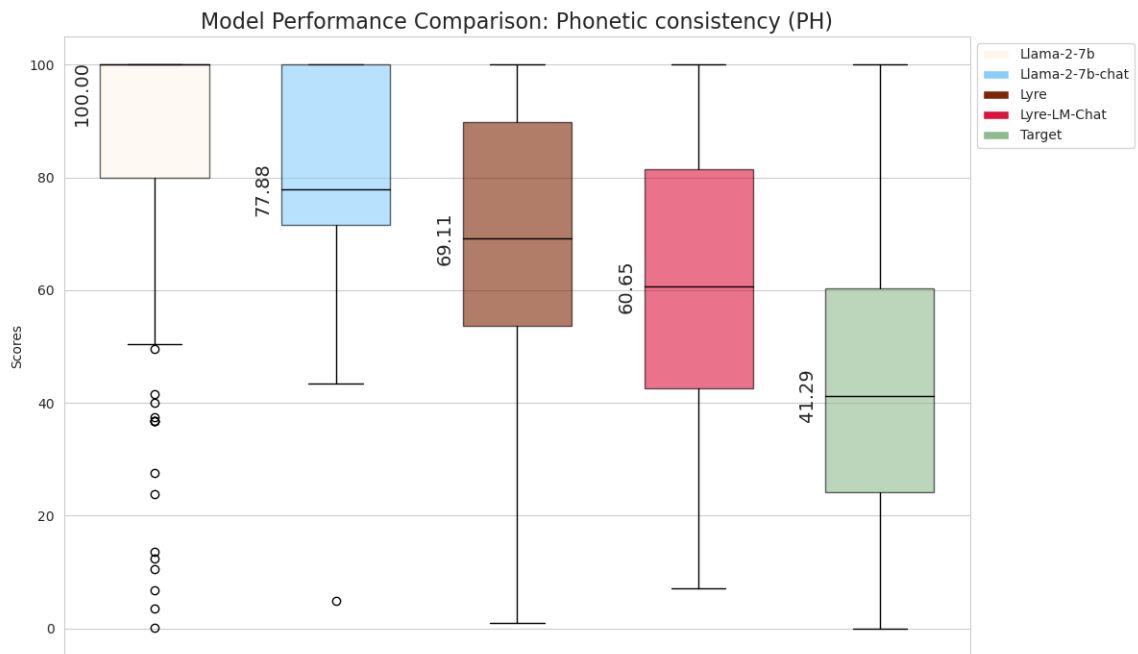


Figure 16: A close up of the phonetic consistency measurement. We find that Lyre-LM-chat is closest to the target.

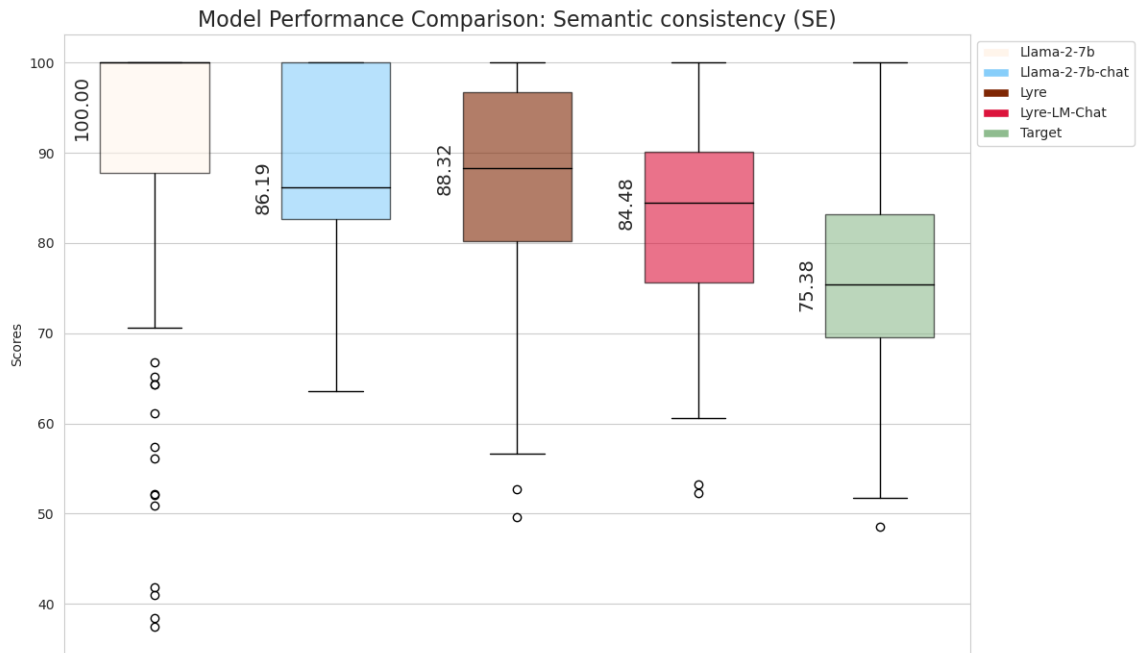


Figure 17: A close up of the semantic consistency measurement. We find that Lyre-LM-chat is closest to the target.

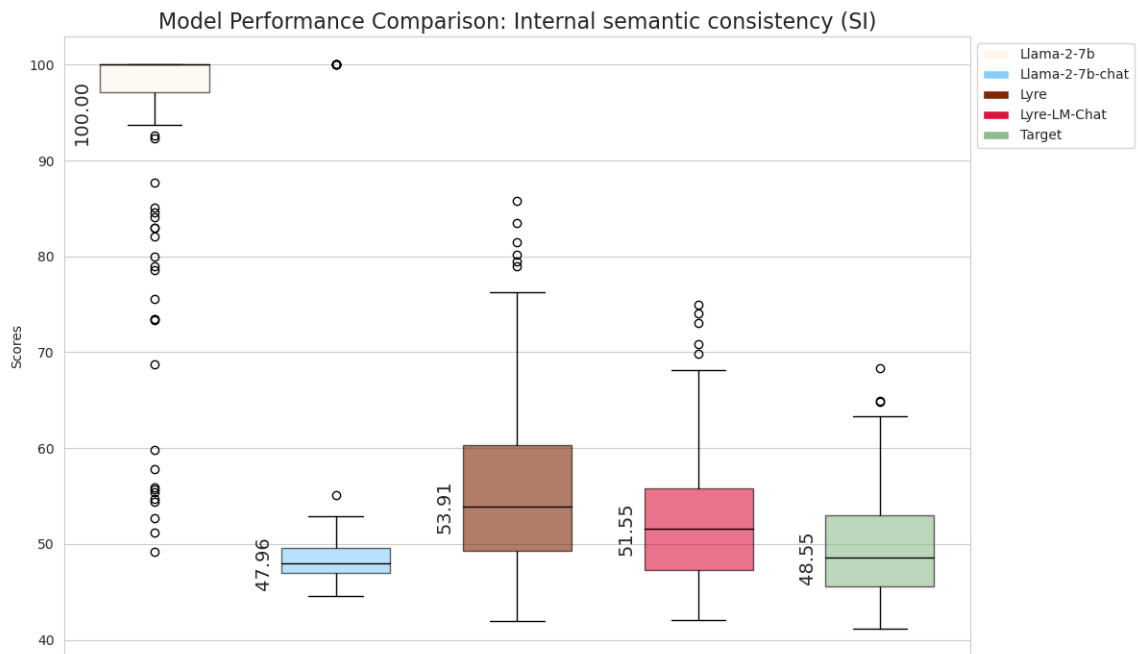


Figure 18: A close up of the internal semantic consistency measurement. We find that Lyre-LM-chat is closest to the target.

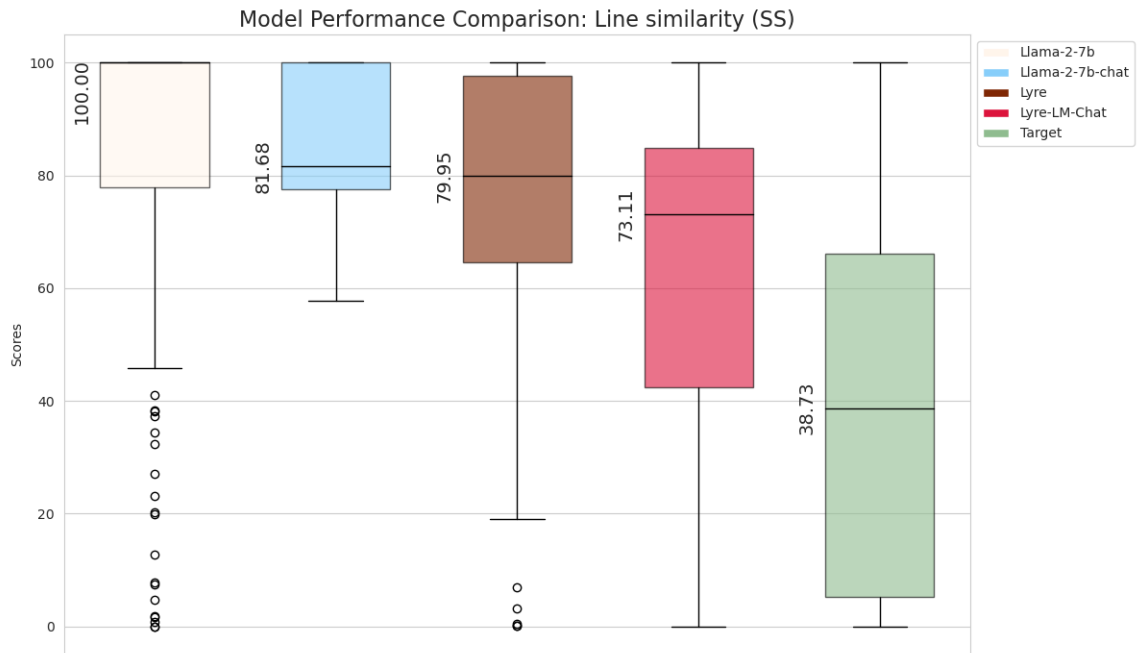


Figure 19: A close up of the line measurement. We find that Lyre-LM-chat is closest to the target.

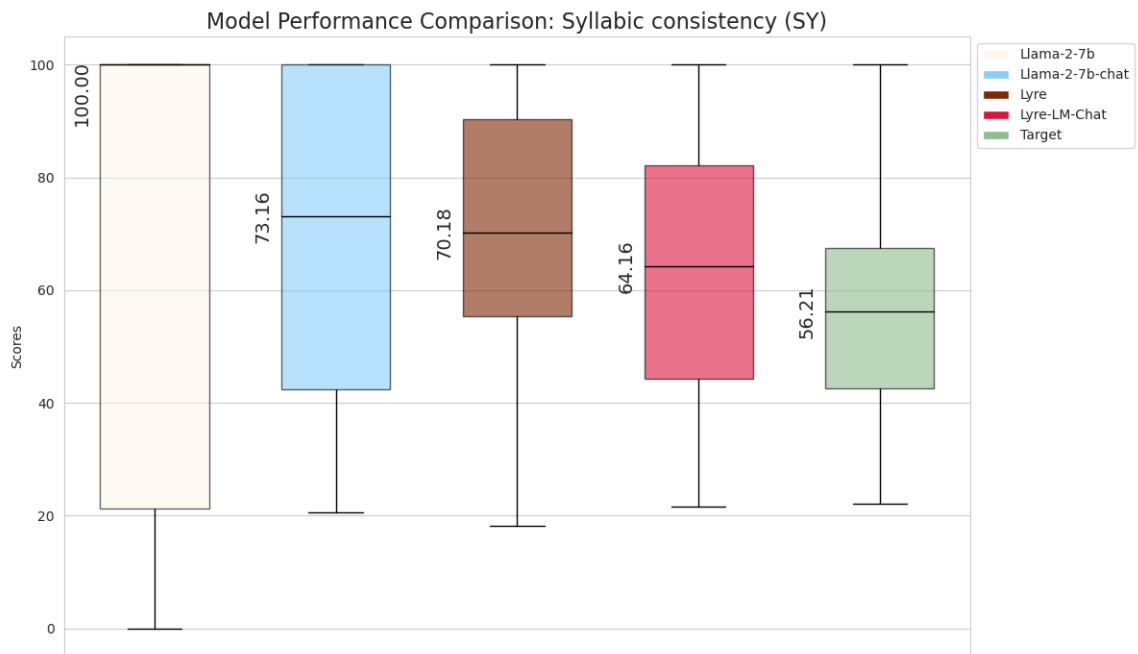


Figure 20: A close up of the syllabic consistency measurement. We find that Lyre-LM-chat is closest to the target.

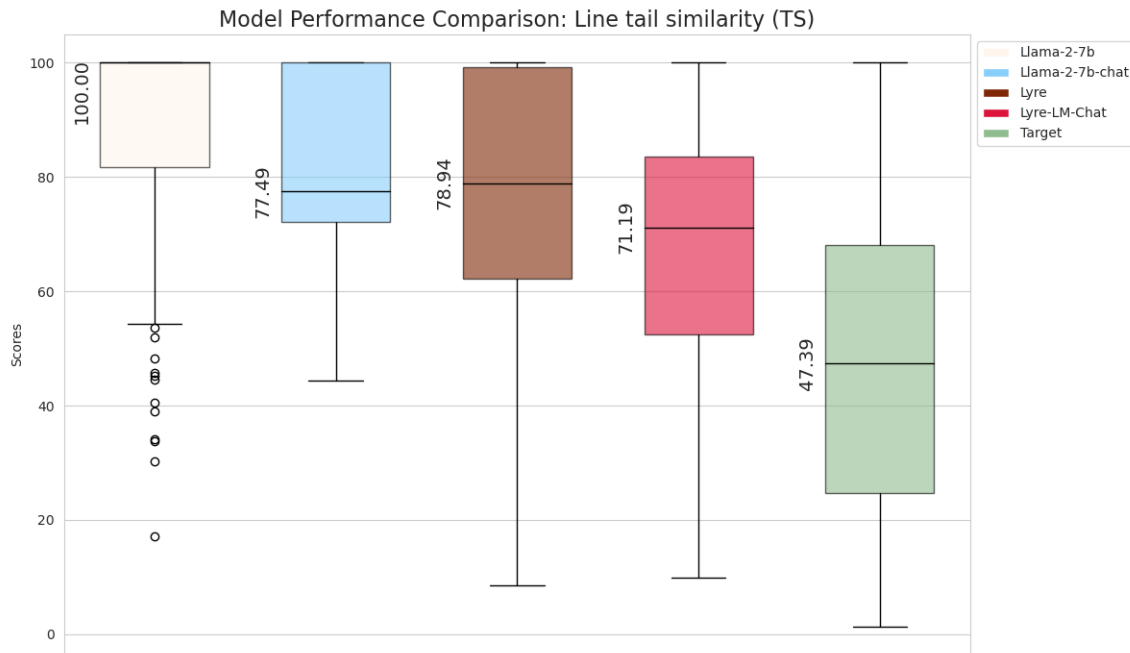


Figure 21: A close up of the line tail measurement. We find that Lyre-LM-chat is closest to the target.

## C Feedback

First off, we thank all our reviews for their detailed feedback and careful considerations. We tried to address each concern raised.

*The paper could benefit from a more organized structure. I recommend introducing distinct sections like Introduction, Related Work, Data, Methods, Results, Limitations, and Discussions.*

Addressed.

*To enhance the paper's context, it's advisable to include a dedicated Related Work section.*

Addressed in **2 Related Work**.

*Specify details about the training, validation, and testing sets, including their sizes and the partitioning methodology.*

Addressed in **3.1**.

*Provide clarity on how the data, including artist's name, song name, and song description, is transformed into prompts for the model.*

Addressed in **3.1.2**.

*There's a need for explicit clarification on the structure of the Llama-2-7B-chat model, addressing whether it is transformer-based or decoder-only.*

Addressed in **2.3**.

*The fine-tuning methods, a critical aspect of the paper, require more emphasis and elaboration. Rather than a mere mention of using FastChat for fine-tuning, specific details about the fine-tuning process, such as whether it's full fine-tune or linear probing, should be explicitly outlined in the paper.*

Addressed in **3.2.1**.

*For the musical evaluation of the model, the evaluation protocol used lacks the support of paper, so the evaluation results are not convincing.*

Addressed in **4.1**.

*The experiment only tried FastChat as fine-tuning method, and did not show the difference between various parameter adjustments.*

Addressed in **5.1**.

*For the purpose of training this model, it is difficult to judge the fine-tuned results without human evaluation.*

Addressed in **4.3** and **5.3**.

*There are some issues with interpret ability of the plots. Multiple plots such as figure 1, 2 and 3 don't have clearly defined axes which makes it difficult to draw inference. It's also unclear what the loss function used in these plots is as it's simply referred to as loss. In addition, titles of graphs such as in Figure 5 could more clearly explain what the purpose of the graph is/what it shows.*

Addressed.

*One way to improve the paper is to consider parameter efficient fine tuning methods in addition to full fine-tuning.*

Addressed in **5.1.2**.

*Another weakness is that general text embedding models are being used to evaluate the lyrical generation without sufficient motivation.*

Addressed in **6.4** and **6.5**. We recognize that embedding models are not designed for lyrics. We balance this possible mismatch with an ensemble of concrete metrics and well as human evaluation.