Reese Carlton, Mia Kobayashi

DATASCI 266

8 December 2024

Exploring AI-Driven Lyric Generation: Emulating Artist Styles

Abstract

This project investigates the application of natural language processing (NLP) techniques to generate song lyrics that emulate the distinct styles of iconic artists, including Elvis Presley, Johnny Cash, and Adele. Utilizing the GPT-2 model, fine-tuned with curated lyric datasets from these artists, we aim to capture their unique linguistic patterns, thematic elements, and stylistic nuances. The baseline model is based on GPT-2, with further exploration into domain adaptation techniques such as low-rank adaptation (LoRA) and specialized prompting to enhance the model's ability to faithfully replicate each artist's style. By doing so, we aim to develop a tool that not only demonstrates the potential of AI in artistic expression but also provides practical applications for songwriters and producers, offering new avenues for creative inspiration and collaboration. This research contributes to the field of computational creativity, highlighting the role of AI in augmenting human artistic endeavors and expanding the possibilities of music creation.

Introduction

The music industry is celebrated for its diverse artistic expression, where unique styles and lyrical voices define each artist's identity. Song lyrics are crucial for conveying emotions, narratives, and cultural messages. Emulating an artist's lyrical style through AI presents both challenges and opportunities. Recent advances in natural language processing and machine learning have enabled creative applications in music and lyrics generation. However, authentically capturing an artist's style remains complex due to the need for linguistic fluency and preservation of thematic elements.

This project aims to develop a system that generates song lyrics in the styles of iconic artists like Elvis Presley, Johnny Cash, and Adele, spanning country music storytelling to contemporary pop ballads. We fine-tune a pre-trained GPT-2 model using each artist's lyrics, leveraging its transformer architecture to replicate their style. The model adapts from general language to a specialized understanding of each artist's voice. Beyond a baseline model, we explore domain adaptation strategies to enhance stylistic accuracy, including LoRA (low-rank adaptation) and specialized prompting. These aim to capture surface features and deeper emotional themes unique to each artist.

Successfully generating artist-specific lyrics has significant implications. For the music industry, it offers valuable tools for songwriters and producers, fresh inspiration, and educational insights into an artist's stylistic evolution. This research also contributes to the discourse on computational creativity, exploring the limits of machine-generated artistic expression.

Background

The use of AI in song lyrics generation is an exciting research area in natural language processing, aiming to emulate human creativity and assist artists. One key advancement is DeepLyrics, which leverages GPT-2's in-context learning, or tuning-free prompting. Described by Tian and Yang, this approach addresses the resource demands of traditional NLP methods. By using well-crafted prompts, DeepLyrics achieves comparable or superior performance to models that require extensive fine-tuning, making AI-driven lyric generation more efficient and accessible.

Another important contribution is from Vechtomova et al., who integrate audio and text in their system for generating lyrics conditioned on an artist's style. Their system uses a variational autoencoder (VAE) with artist embeddings pre-trained on MEL spectrograms, bridging musical and lyrical elements. This innovative method shows that initializing artist embeddings with spectrogram-derived representations can improve stylistic accuracy in generated lyrics.

These studies showcase the potential of combining advanced NLP techniques with multi-modal data to advance lyric generation. By integrating musical and textual insights, researchers can create systems that align with an artist's linguistic style and artistic identity. Building on these foundations, our project uses GPT-2 as a baseline and explores domain adaptation techniques, such as low-rank adaptation (LoRA) and specialized prompting to enhance stylistic fidelity. We fine-tune the model on curated lyrics from artists like Elvis, Cash, and Adele to capture their unique styles.

Methods

The project began with acquiring and preprocessing song lyrics data for model training. Lyrics were scraped using the Genius API through the lyricsgenius Python library, focusing on artists like Elvis, Cash, and Adele. The search_artist function gathered a list of songs where the artist was the primary performer or a collaborator, with include_features set to True for a comprehensive dataset. Essential details, such as song ID, title, primary artist, featured artists, release date, and lyrics, were extracted.

Data cleaning involved removing incomplete entries, normalizing text by stripping non-essential characters and annotations (e.g., [Chorus], [Verse 1]) using regex, and eliminating duplicates. To ensure the dataset contained only songs, entries were cross-referenced with known song titles. For compatibility with GPT-2's input constraints, lyrics were concatenated into a single text string, with songs separated by newline characters. This text was segmented into chunks with a maximum length of 1024 characters, adhering to GPT-2's sequence length limit. Any chunk exceeding the token count was truncated accordingly.

Prompt engineering was crucial for generating coherent and contextually relevant lyrics emulating specific artists' styles. The first sentence of each song served as the prompt, capturing narrative depth and thematic essence. These prompts were tokenized and fed into the fine-tuned GPT-2 model, which predicted subsequent text based on learned language patterns and stylistic nuances. The effectiveness of prompts was evaluated using semantic similarity measures, comparing generated lyrics to original songs starting with the prompt phrases. This evaluation assessed the model's alignment with the intended artistic style and the efficacy of the prompt engineering process.

Baseline Model

The baseline model in this project uses the pre-trained GPT-2, renowned for generating coherent, contextually relevant text. Fine-tuning was performed straightforwardly, without advanced techniques like Low-Rank Adaptation (LoRA), using a learning rate of 5e-5 to ensure balanced convergence and stability.

For lyric generation, the model.generate() function was employed with parameters optimized for quality and diversity. The max_length was set to 500 to allow substantial lyrical content per sequence. The num_return_sequences was set to 1, focusing on quality over quantity.To enhance text diversity and avoid repetition, no_repeat_ngram_size was set to 2. Nucleus sampling was used with top_k at 20 and top_p at 0.7, balancing randomness and coherence. A temperature of 0.7 was chosen to control randomness, balancing conservatism and creativity. Early stopping was enabled to end generation once a coherent output was achieved, rather than filling the maximum length.

Lora & Lora Fine Tuned

The LoRA (Low-Rank Adaptation) model builds on the baseline GPT-2 by introducing lightweight fine-tuning to emulate distinct lyrical styles with minimal computational overhead. Both models used consistent initial parameters: max_length of 500 for substantial output, num_return_sequences of 1 for focused quality, and no_repeat_ngram_size of 2 to prevent redundancy. Generation settings included nucleus sampling with top_k: 20, top_p: 0.7, and temperature: 0.7, balancing coherence and diversity. Early stopping was employed to end generation upon achieving a coherent sequence. This approach allows for domain-specific tuning without retraining the full model, preserving each artist's stylistic essence.

The fine-tuned LoRA model further refined the stylistic alignment by adjusting key text generation parameters through a grid search process. This optimization targeted top_k: [10, 20, 30], top_p: [0.6, 0.7, 0.8, 0.9], and temperature: [0.6, 0.7, 0.9] to enhance the model's ability to replicate lyrical diversity and thematic consistency. These parameters were selected for fine-tuning as they directly influence the diversity and coherence of the generated text. Adjusting these parameters helps balance creativity and precision—top_k limits the model to the most probable word choices, top_p ensures selections remain contextually relevant, and temperature controls randomness. This approach emphasized contextual depth and artistic fidelity, aiming to address stylistic variations that the default LoRA parameters could not fully encapsulate.

Results & Discussion

Table 1 below presents sample outputs from various configurations of our models trained to emulate Johnny Cash's lyrical style, using the song "Rusty Cage" as a basis for comparison with the "Rusty Cage - Actual" column providing the original lyrics. The "Rusty Cage - Baseline", “Rusty Cage - LoRA”, “Rusty Cage - Lora fine-tuned” and “Rusty Cage - GPT 3.5” output shows an attempt to replicate these themes, though it veers towards more abstract imagery, reflecting each model's initial ability to capture thematic elements but not specific phrasing.

Table 1

Actual	Baseline	LoRA	LoRA, fine-tuned	Chat GPT 3.5
You wired me awake And hit me with a hand of broken nail You tied my lead and pulled my chain To watch my blood begin to boil But I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run…	You wired me awake I've been to the moon And I've seen the stars But I never saw the sun The moon is just a little bit dimmer When I look back on the past It's just another day Then I'll be back With a smile that's like a dream That's all I ever dreamed For I'm a man of God …	That you've been seen, you're not alone You've seen me, I've heard you I've known you, and I know you'll be there …	n/a	*You wired me awake,* With the buzz of the night, Your voice like a train, Rolling steady through the fight. The moonlight was faded, But your shadow stayed bold, A spark in the darkness, In the stories you told…

Table 2 below shows a comparison of different model’s performance in replicating the lyrical styles of various artists using metrics like Semantic Similarity, BLEU, Rouge-1, and Rouge-L score. It highlights the effectiveness of a baseline GPT-2 model and its adaptations, including LoRA and fine-tuning, in capturing stylistic and thematic elements. The results show that for the artists in question, Johnny Cash, Adele and Elvis, the baseline model has a solid foundation and the use of domain adaptation can yield varying results, with the baseline outperforming in many of the metrics.

Table 2

Arist	Model	Semantic Similarity	BLEU	ROUGE-1	ROGUE-L	ChatGPT 3.5
Johnny Cash	baseline	0.49312	0.04082	0.30814	0.15783	1.25 / 5
	LoRA	0.47421	0.04123	0.30734	0.16109	1.25 / 5
	LoRA, fine-tuned	x	x	x	x	x
Elvis	baseline	0.52193	0.02786	0.27435	0.14252	1 / 5
	LoRA	0.51959	0.02920	0.28128	0.14910	2 / 5
	LoRA, fine-tuned	0.47507	0.02924	0.27610	0.156145	1.75 / 5
Adele	baseline	0.57179	0.03632	0.37257	0.17246	2.25 / 5
	LoRA	0.36788	0.01370	0.23778	0.15249	2.6 / 5
	LoRA, fine-tuned	0.35352	0.01492	0.25978	0.15212	2.4 / 5

Baseline Model Performance

The baseline model's performance was evaluated using Semantic Similarity, BLEU, ROUGE-1, and ROUGE-L scores to assess its ability to emulate the lyrical styles of Cash, Presley, and Adele.

Johnny Cash: The model achieved a semantic similarity score of 0.48699, indicating moderate alignment with reference lyrics. The BLEU score was low at 0.03992, suggesting challenges in capturing exact phrasing. However, it performed better with ROUGE-1 and ROUGE-L scores of 0.30367 and 0.15395, reflecting its ability to capture Cash's lyrical structure.

Elvis: The semantic similarity score was higher at 0.54069, showing stronger alignment. The BLEU score was 0.03078, highlighting difficulties in precise replication. ROUGE-1 and ROUGE-L scores were 0.27518 and 0.14251, indicating the model captures some stylistic elements but requires refinement to better mirror Elvis's intricacies.

Adele: The model had the highest semantic similarity score at 0.60635, indicating good alignment. The BLEU score was 0.03469, showing better phrasing capture. ROUGE-1 and ROUGE-L scores were 0.37711 and 0.17329, demonstrating strong performance in replicating Adele's lyrical structure and content.

Overall, the baseline model shows potential in capturing thematic and stylistic elements but needs further refinement for precise lyrical replication. These results establish a benchmark for evaluating improvements with advanced techniques like LoRA and additional fine-tuning.

LoRA

The LoRA (Low-Rank Adaptation) model introduces a lightweight fine-tuning approach to adapt the baseline model for better emulation of the lyrical styles of Cash, Presley, and Adele. The initial LoRA model used the same parameters as the baseline; however, it demonstrated a slight decrease in performance across most metrics.

Cash: While the model was intended for evaluation, it could not complete running due to GPU memory limitations. However, BLEU and ROUGE-L scores for partial runs indicated slight improvements (0.04123 and 0.16109), suggesting that LoRA may help refine the capture of lyrical phrasing and structure, though not dramatically.

Elvis: The LoRA model closely mirrored the baseline in semantic similarity (0.51959 vs. 0.52193), with a slight improvement in BLEU and ROUGE-L scores (0.02920 and 0.14910). This suggests LoRA can subtly enhance stylistic replication.

Adele: The LoRA model showed a notable drop in semantic similarity (0.36788 from 0.57179) and BLEU (0.01370), reflecting challenges in thematic alignment and phrasing. However, the ROUGE-L score (0.15249) remained comparable, suggesting that structural elements of Adele's style were better preserved than thematic ones.

These results show that LoRA introduces nuanced adjustments but may not consistently outperform the baseline across artists.

LoRA Fine-Tuned

The fine-tuned LoRA models showed deterioration from the baseline LoRA model, likely due to overfitting. This process may have led to parameters that overly fit the training data, hindering broader stylistic capture. Despite slight improvements for Presley in BLEU (0.02924 vs. 0.02920) and ROUGE-1 (0.27610 vs. 0.28128), Adele's model lagged in semantic similarity, suggesting trade-offs between thematic consistency and stylistic flexibility.

For Presley, the fine-tuned parameters (top_k: 20, top_p: 0.6, temperature: 0.6) resulted in slight improvements in BLEU (0.02924 vs. 0.02920) and ROUGE-1 (0.27610 vs. 0.28128), suggesting a modest enhancement in capturing stylistic phrasing and structure. For Adele, the best parameters (top_k: 20, top_p: 0.7, temperature: 0.7) revealed notable challenges, as semantic similarity dropped, reflecting trade-offs between thematic consistency and stylistic flexibility.

Possible reasons for the fine-tuned model's subpar performance include suboptimal parameter choices that didn't generalize well, loss of stylistic balance leading to incoherent lyrics, insufficient data diversity causing overfitting, and negative interactions between hyperparameters like top_k, top_p, and temperature. Additionally, the GPU memory issue with the Cash model underscored the computational constraints of exploring a broader parameter space. The fine-tuning process affected each artist differently, highlighting the need to tailor adjustments to individual stylistic nuances.

Conclusion

This study explored the potential of AI-driven lyric generation to emulate the distinctive styles of Johnny Cash, Elvis Presley, and Adele using GPT-2 models with baseline, LoRA, and fine-tuned LoRA configurations. The results indicate that no single model is inherently superior across all metrics or artists. Instead, each approach offers unique strengths and challenges. The baseline model provided a solid foundation, demonstrating moderate alignment with the original styles. LoRA introduced lightweight adaptations, enhancing certain stylistic features, while fine-tuned LoRA models had varying performance depending on the artist, reflecting the complexity of optimizing hyperparameters for stylistic fidelity.

The models' limitations highlight the dependence on the size and quality of the training data. A more extensive and diverse dataset could enable better thematic and stylistic replication, particularly for artists with rich, multifaceted discographies. Additionally, the reliance on textual data alone constrains the models' ability to fully capture the musicality and performative nuances of an artist's style.

Future research should explore multi-modal approaches that integrate text with audio data, such as incorporating vocal tone, rhythm, and instrumentation. By leveraging techniques like artist-conditioned embeddings derived from spectrograms or integrating models trained on both lyrics and audio, AI systems could achieve a more holistic emulation of an artist's creative identity. Such advancements would not only enhance computational creativity but also offer more practical tools for songwriters, producers, and educators in the music industry.

Bibliography

1. Tian, L., & Yang, X. (n.d.). *DeepLyrics: GPT2 for Lyrics Generation with Finetuning and Prompting Techniques*. DeepLyrics: GPT2 for lyrics generation with finetuning and prompting techniques
2. Chen, Y., & Lerch, A. (2020). *Melody-Conditioned Lyrics Generation with SeqGANs*. https://arxiv.org/pdf/2010.14709
3. Vechtomova, O., Bahuleyan, H., Ghabussi, A., & John, V. (n.d.). *Generating Lyrics with Variational Autoencoder and Multi-Modal Artist Embeddings*. https://arxiv.org/pdf/1812.08318
4. Zhang, Z., Lasocki, K., Takasu, A., & Yu, Y. (2024). *Syllable-Level Lyrics Generation from Melody Exploiting Character-Level Language Model*. https://arxiv.org/pdf/2310.00863

Appendix

Table 1. Rusty Cage Complete Songs

Rusty Cage - Actual	Rusty Cage - baseline	Rusty Cage - LoRA	Rusty Cage - LoRA fine tuned	Rusty Cage - GPT 3.5
You wired me awake And hit me with a hand of broken nail You tied my lead and pulled my chain To watch my blood begin to boil But I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run Too cold to start a fire I'm burning diesel, burning dinosaur bones I'll take the river down to still waters And ride a pack of dogs I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run When the forest burns along the road Like God's eyes in my headlights When	You wired me awake I've been to the moon And I've seen the stars But I never saw the sun The moon is just a little bit dimmer When I look back on the past It's just another day Then I'll be back With a smile that's like a dream That's all I ever dreamed For I'm a man of God A man who's always been a part of me My love for you is so strong You're the only one who ever loved me and I love you Your love is the reason I can't die Because I know that you'll always be with me, and that I will always love If I could just get over the feeling Of being alone There's no way I'd ever love another man Like you, I just can not get along Just like you	That you've been seen, you're not alone You've seen me, I've heard you I've known you, and I know you'll be there [Chorus] Oh, my God, oh, God My God! And I'm going to be here But I'll never be alone, No, no, not ever Never, ever, never Just a little bit of love A little little love, a lot of joy That's all I can say Well, it's just a matter of time Then I will be back There's no way I won't be It's a long way to go Until I get back to my home So I hope you all love me Love, love love you too		*You wired me awake,* With the buzz of the night, Your voice like a train, Rolling steady through the fight. The moonlight was faded, But your shadow stayed bold, A spark in the darkness, In the stories you told. I walked down the line, Where the blacktop would burn, And the lessons you left me, Were sharp as they turned. The bottle was empty, But the fire wouldn't die, Your words kept on burning, Like stars in the sky. *Chorus:* Oh, you wired me awake, Like a match to the flame, I can't find no escape, From the weight

the dogs are looking for their bones And it's raining icepicks on your steel shore I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run I'm gonna break I'm gonna break my I'm gonna break my rusty cage and run	don't get around to loving me anymore So I want to tell you a story About a time or two ago			of your name. Oh, the echo, it lingers, Like the tremble of strings, You wired me awake, Now I'm tied to these things. Through the miles and the shadows, Your whisper's my guide, It cuts through the stillness, Like a thief in the night. The road keeps on calling, But it won't heal the ache, For the spark that you lit, When you wired me awake.

Johnny Cash, LoRA ChatGPT Score Explanation

1. Thematic Alignment (1/5):

- **Original:** "Rusty Cage" is raw and intense, centered on themes of entrapment, rebellion, and breaking free. It uses visceral imagery ("hand of broken nails," "burning dinosaur bones") to convey a sense of struggle and liberation.
- **Generated:** The themes are scattered and unclear. It vaguely touches on love, regret, forgiveness, and some form of resilience, but these ideas lack focus and cohesion. It fails to capture the rebellious, defiant essence of the original.

2. Emotional Tone (2/5):

- **Original:** Powerful, aggressive, and liberating. It evokes a visceral response with its dark, gritty tone and vivid imagery.
- **Generated:** While it tries to evoke emotions (love, hurt, and resilience), the tone is inconsistent, ranging from sentimental to nonsensical. The emotional depth is diluted by the lack of coherence and abrupt shifts.

3. Structure and Flow (1/5):

- **Original:** The song has a clear and repetitive structure, with verses and a chorus that build momentum toward the cathartic refrain, "I'm gonna break my rusty cage and run."
- **Generated:** Disorganized and rambling. It begins with a semblance of a chorus but devolves into fragmented and nonsensical phrases. The long list of disconnected words and phonetic fragments ("ivin' in ixiv xiv") breaks any sense of flow.

4. Lyrical Flow and Imagery (1/5):

- **Original:** Poetic and evocative, with strong imagery and metaphors that create a vivid mental picture (e.g., "forest burns along the road," "raining icepicks on your steel shore").
- **Generated:** Lacks poetic devices and imagery. The nonsensical, filler-like phrases (e.g., "ivin' in ixiv xiv") distract from any intended meaning and undermine lyrical integrity.

Final Score: 1.25/5

The generated lyrics fail to match the original in theme, tone, structure, and quality of writing. "Rusty Cage" is cohesive, intense, and purposeful, whereas the generated lyrics are fragmented, inconsistent, and lack the defiant spirit that defines the original.

Elvis, LoRA Fine-Tuned ChatGPT Score Explanation

1. Thematic Alignment (2/5):

- **Original:** Focuses on the magical and transformative nature of love when "a boy meets a girl." It's simple, romantic, and cohesive in its celebration of a singular, life-changing moment.
- **Generated:** Rambles across various themes, including love, friendship, childhood, familial relationships, and self-reflection. It lacks the singular romantic focus of the original, diluting thematic alignment.

2. Emotional Tone (2/5):

- **Original:** Consistent in tone—optimistic, pure, and whimsical, emphasizing joy and enchantment.
- **Generated:** Emotionally erratic, moving from longing to regret to gratitude. It tries to evoke emotion but lacks the focus and simplicity that makes the original effective.

3. Structure and Flow (1/5):

- **Original:** Follows a clear structure with concise verses and repeated themes, making it easy to follow and sing.
- **Generated:** Disorganized, with unclear verses and choruses, interruptions ("Choruses," "Verb," "Guitar"), and fragmented ideas. This lack of structure makes it hard to perceive as a cohesive song.

4. Lyrical Flow (2/5):

- **Original:** Smooth, rhythmic, and poetic, with an effective use of repetition ("When a boy like me meets a girl like you").
- **Generated:** Lacks rhyme and rhythm, with awkward phrasing and extraneous words that interrupt the flow. Some parts are confusing or incomplete ("I don’ know what I want to do").

Final Score: 1.75/5

While the generated lyrics attempt to convey emotion, they lack the romantic focus, clear structure, and lyrical elegance of the original. The thematic scope is too broad, and the disorganized presentation makes it a weak match.

Adele, LoRA Fine-Tuned ChatGPT Score Explanation

1. Thematic Alignment (3/5):
   - **Original:** A heartfelt plea for intimacy and connection during a potentially final, fleeting moment of love. It focuses on vulnerability and the fear of losing love forever.
   - **Generated:** Shares some thematic overlap with the original in expressing emotional vulnerability and commitment, but it diverges into a more general and less urgent tone of friendship and steadfastness. The stakes feel lower compared to the original's poignant finality.
2. Emotional Tone (2.5/5):
   - **Original:** Deeply emotional and bittersweet, capturing the urgency of seizing a significant moment with someone special.
   - **Generated:** Lacks the depth and complexity of the original. Lines like “You will not be a burden to me” and “My heart will never stop” sound reassuring but fail to evoke the same level of intensity or emotional conflict.
3. Structure and Flow (2/5):
   - **Original:** Progresses smoothly from introspection to a heartfelt chorus, building emotional intensity and maintaining focus.
   - **Generated:** Feels more like a fragmented thought or statement than a structured verse. It lacks a clear narrative arc or climax, making it less compelling as a standalone piece.
4. Lyrical Quality and Imagery (2/5):
   - **Original:** Uses vivid, evocative imagery (“Hold me like I’m more than just a friend,” “Give me a memory I can use”) to create an emotional connection with the listener.
   - **Generated:** Relies on simple, generic phrases that lack the same poetic or descriptive quality. Lines like “You will not be a burden to me” are clear but lack the nuance and richness of the original.
5. Narrative Consistency (2.5/5):
   - **Original:** Consistent in its exploration of love, vulnerability, and the desire to create a lasting memory.
   - **Generated:** While consistent in tone, it doesn’t build a strong emotional narrative or progression. It hints at loyalty and steadfastness but doesn’t delve deeper into the relationship or context.

Final Score: 2.4/5

The original lyrics are a masterful balance of vulnerability and urgency, rich with emotional depth and poetic imagery. The generated lyrics touch on themes of companionship and commitment but lack the emotional gravity, vivid imagery, and structured storytelling that make the original compelling. While both aim to express deep feelings, the generated version feels more surface-level and less impactful.