Writing Style Text Transfer

Exploring tools to maintain writing style in the age of LLMs

Final Report for Final Project

By Nathan Harris

CSPB 3832 Natural Language Processing – Summer 2025

Table of Contents

Abstract

This project explored Text Style Transfer from neutral to stylized writing by fine-tuning GPT-2 Medium using open source Hugging Face transformers on Google Colab. The dataset consisted of 300 text chunks from Herman Melville's three most popular novels paired with corresponding neutralized texts generated using ChatGPT. Evaluation using BLEU, ROUGE, and perplexity scores revealed that while the model successfully transferred stylistic characteristics, it struggled with content preservation—findings that align with existing research showing the inherent trade-off between style accuracy and semantic fidelity in text style transfer tasks.

Introduction

If the majority of internet users are utilizing ChatGPT, or similar LLMs, to generate blog posts, social media posts, write emails and so forth, we all might start writing and sounding like an LLM. This could be described as homogenization of text online. If someone wants to write with a distinct writing style while also leveraging LLMs, an AI-powered writing style assistant could provide support for cultivating and maintaining that writing style. This problem falls under Text Style Transfer in Natural Language Processing which focuses on transferring writing style from one text to another text while preserving the core meaning. This could include transferring from a casual style to a formal academic writing style or from a positive sentiment to a negative sentiment. In this project I explored transferring writing style from a neutral writing style to a targeted writing style. For this project, I focused specifically on transferring neutral text into the distinctive style of Herman Melville, whose complex prose in his novels provides a rich test case for style transfer techniques. By experimenting with fine-tuning approaches, this work aims to understand both the potential and limitations of current methods for maintaining distinctive authorial voices in AI-assisted writing.

Related Work

A Google research team found that professional writers benefit from AI assistants for idea generation and brainstorming, but these tools struggle to maintain distinctive authorial voice, defaulting to bland language due to training on broad internet data (Ippolito et al., 2022). This foundational work establishes the core problem that motivates style transfer research.

Current style transfer approaches fall into three main groups: prompting techniques, fine-tuning, and memory-augmented LLMs. Fine-tuning generally outperforms prompting

techniques, and while larger models perform better, there reaches a point where improvement diminishes with increasing size (Toshevska & Gievska, 2025). This survey confirms the approach taken in this project while highlighting the inherent trade-offs in the field.

StyleTunedLM addresses the growing concern of language homogenization by using Parameter-Efficient Fine-Tuning (PEFT) with Low-Rank Adaptation (LoRA) on LLaMA-2-7b models trained on ten Project Gutenberg authors (Liu et al., 2024). While this approach is limited to well-known authors with substantial available data, it leverages open source tools readily available on Hugging Face, making it accessible for replication.

In contrast, ASTRAPOP focuses on low-resource authorship style from non-famous authors, using a two-stage framework that first removes the original author's style, then applies policy optimization with reward functions (Liu et al., 2024). This work introduces the novel concept of using an additional AI agent as a writing coach that provides feedback to the model during training, addressing the challenge of limited training data.

Panza demonstrates a practical implementation requiring only 50-100 emails, combining both prompt engineering and fine-tuning approaches by having the model create prompts in the first phase before fine-tuning (Nicolicioiu et al., 2025). This hybrid approach shows promise for real-world applications where users have limited personal writing samples.

This project builds on the fine-tuning approaches established in prior work while exploring the fundamental trade-offs between style transfer quality and content preservation using a well-established literary corpus.

Data

The initial dataset consisted of 100 paragraphs extracted from Herman Melville's *Moby Dick*, obtained through Project Gutenberg. For an additional iteration of the project, I expanded the dataset by adding 200 more paragraphs from two other Melville novels: *Bartleby the Scrivener* and *Billy Budd*, resulting in a total of 300 paragraphs across three works. I segmented the text at paragraph boundaries rather than sentence level, recognizing that sentence structure is a vital component of writing style that should be preserved during the style transfer process.

To create the paired training data required for supervised fine-tuning, I used ChatGPT to generate neutralized versions of each stylized paragraph. This neutralization process involved manually copying and pasting portions of Melville's text into ChatGPT with prompts designed to convert the ornate, complex prose into simpler, more neutral language while preserving the core semantic content.

The final dataset was formatted for Hugging Face Transformers with three fields:

Author: "Herman Melville" (consistent across all samples)

Input: Neutralized text (simple, accessible language)

Target: Original stylized text (Melville's distinctive prose)

This structure enabled the model to learn the mapping from neutral style to Melville's characteristic writing style during fine-tuning.

The dataset construction process had several limitations. The manual neutralization process using ChatGPT was time-intensive and potentially inconsistent, as the quality of neutralized text varied depending on the complexity of the original passage. Additionally, the relatively small dataset size (300 paragraph pairs) may have limited the model's ability to fully capture the nuances of Melville's writing style across different works.

Methodology

Approach

This project employed supervised fine-tuning to learn text style transfer from paired examples of neutral and stylized text. The approach treated style transfer as a sequence-to-sequence task, similar to machine translation, where neutral text serves as input and stylized text as the target output.

Model Selection

I tested several pretrained models including Google Flan T5 (small and large), TinyLlama, and GPT-2 (small and medium). GPT-2 Medium was the only model to produce usable results, with its autoregressive nature and parameter count well-suited for learning the neutral-to-stylized mapping.

Implementation

The project used Hugging Face Transformers on Google Colab with a two-notebook workflow:

Neutralize Text: Processed Melville's corpus and generated the paired dataset

Style Imitation: Fine-tuned GPT-2 Medium and conduct evaluation

The fine-tuning process used standard transformer techniques, minimizing loss between predicted and target stylized text. Initially, I attempted text generation using only Melville's original text, but this proved insufficient, the model required explicit neutral-to-stylized examples.

Evaluation

Three metrics assessed different aspects of style transfer:

BLEU Score: Style transfer quality (similarity to target style)

ROUGE-1 Score: Content preservation (semantic overlap)

Perplexity: Text fluency and coherence

These metrics capture the fundamental trade-off between style accuracy and content preservation in style transfer tasks.

Results

I evaluated the model using diverse inputs: original neutralized paragraphs from *Moby Dick*, simple sentences ("I am hungry"), paragraphs from other public domain novels by different authors, and a paragraph from a baseball news article.

BLEU scores were frequently 0, indicating strong stylistic similarity to Melville's writing. While the variety in other metrics raised questions about BLEU's reliability in this context, qualitative assessment confirmed the model successfully imitated Melville's characteristic winding, run-on prose style. Though word choice differed from source material, the overall rhythm and sentence structure felt authentically Melvillesque.

ROUGE-1 scores consistently tested very low, revealing that core meaning was often lost during style transfer. This aligns with existing research identifying content preservation as the primary challenge in text style transfer tasks.

Initial tests on smaller datasets showed low perplexity scores, indicating fluent, coherent text. However, after code updates and expanded test cases, perplexity scores increased significantly, suggesting reduced coherence. Manual evaluation confirmed this assessment. While stylistic elements transferred successfully, semantic content often became garbled.

The model demonstrated clear success in capturing Melville's distinctive writing style but struggled with the fundamental challenge of preserving meaning during transfer. The results highlight the inherent trade-off between style accuracy and content preservation in style transfer approaches.

Discussion

This project proved more complex than initially anticipated for the given timeframe. While I originally planned to compare multiple techniques (prompting versus fine-tuning), explore different authors, and optimize parameters across various metrics, the reality was that

achieving any measurable style transfer, rather than generating gibberish or unchanged text, became the primary accomplishment.

I initially imposed constraints to use only open-source tools that could run locally, motivated by privacy concerns about sharing writing content with commercial LLMs that might incorporate user data into training. However, the quality limitations of open-source neutralization tools ultimately required using ChatGPT for text neutralization, as it produced significantly better results than available transformer alternatives.

The automated pipeline for dataset creation represents a key limitation. With additional time, human verification of the neutralized texts would ensure accurate paraphrasing while maintaining neutral style. ROUGE scores could also be systematically applied during dataset creation to confirm content similarity between original and neutralized versions.

Inconsistent results posed significant troubleshooting challenges. Even in the final implementation, multiple notebook runs are sometimes required to achieve measurable results. This variability hindered systematic optimization and made it difficult to isolate the effects of specific changes.

Despite these challenges, the project provided valuable experience with open-source transformers and fine-tuning techniques. The core problem, maintaining distinctive writing styles in an era of AI homogenization, remains compelling and worthy of continued investigation.

Conclusion & Future Work

This project successfully demonstrated that fine-tuning GPT-2 Medium can capture distinctive authorial style characteristics, specifically replicating Herman Melville's complex prose patterns. However, the results also confirmed the fundamental challenge in text style transfer: the trade-off between style accuracy and content preservation. While the model effectively transferred stylistic elements, semantic content was often lost in the process.

Improving dataset quality through human verification of neutralized texts would strengthen the foundation for style transfer. Expanding to include multiple well-known authors, anonymous writers, and diverse formats (blog posts, emails, social media) would test usability across different writing contexts.

Deeper exploration of fine-tuning parameters and implementation of Parameter-Efficient Fine-Tuning (PEFT) with LoRA adapters could improve efficiency and results. These approaches, mentioned in my research, warrant systematic investigation in a larger time frame.

Prompt engineering and Retrieval Augmented Generation (RAG) may achieve comparable results with less computational overhead. RAG systems could compress and cache writing style patterns, offering a more practical solution for real-world applications.

The core problem, maintaining distinctive writing voices amid AI homogenization, remains compelling and increasingly relevant as LLM adoption continues to grow.

Bibliography

Ippolito, Daphne, et al. "Creative Writing with an AI-Powered Writing Assistant: Perspectives from Professional Writers." arXiv preprint arXiv:2211.05030v1, 9 Nov. 2022

Liu, Shuai, et al. "Authorship Style Transfer with Policy Optimization." *arXiv preprint arXiv:2403.08043* (2024).

Liu, Xinyue, et al. "Customizing Large Language Model Generation Style using Parameter-Efficient Finetuning." arXiv preprint arXiv:2409.04574, 6 Sept. 2024,

Nicolicioiu, Armand, et al. "Cheap and Effective Personalization of Foundation Language Models for Imitating a User's Writing Style." ICLR 2025 Workshop on Foundation Models in the Wild, 2025.

Toshevska, Martina, and Sonja Gievska. "LLM-Based Text Style Transfer: Have We Taken a Step Forward?" IEEE Access, vol. 13, 2025, pp. 44707-44721, doi:10.1109/ACCESS.2025.3548967.

Sample Outputs

Stylized Target: Once more. Say you are in the country; in some high land of lakes. Take almost any path you please, and ten to one it carries you down in a dale, and leaves you there by a pool in the stream. There is magic in it. Let the most absent-minded of men be plunged in his deepest reveries—stand that man on his legs, set his feet a-going, and he will infallibly lead you to water, if water there be in all that region. Should you ever be athirst in the great American desert, try this experiment, if your caravan happen to be supplied with a metaphysical professor. Yes, as every one knows, meditation and water are wedded for ever.

Neutral Input: If you are in an area with lakes, most paths eventually lead downhill to a pool of water. Even someone lost in thought will unknowingly walk toward water if it is nearby. If you are ever thirsty in a desert and traveling with a philosopher, try this as a method to find water. Thought and water are closely connected.

Generated Text: But even though one may be tempted by his own thoughts at times when he wishes for some fresh air or something else from afar—even so far back as where there are no other human beings anywhere near him, yet all that such feelings can accomplish; what would have been done had they not happened? Would any man then wish without hesitation upon entering a lake-streams could get himself out into the

---

Input: The sky is blue…

Generated: I am a man, and an old friend of mine. You are too young to understand me now that I have written you this letter; but when we met here at last night it was my wish to meet again with one who knows all about myself…

Input: Shohei Ohtani, a player for the Los Angeles Dodgers, hit his 40th home run of the season on Saturday in a game against the Toronto Blue Jays, giving the Dodgers a 3-0 lead. The Dodgers won the game 9-1. It was their first consecutive home wins since they defeated the Chicago White Sox in three games from July 1 to July 3. Ohtani's home run traveled 417 feet to center field.

Generated: Now this afternoon at noon I heard that there had been a scoreless inning between two teams here yesterday and today; but as soon then as my eyes were drawn over it again I began thinking how many more like them would have gone down by now if not such an incident happened! And yet some things happen which are beyond our power to control—and though we cannot know what is going upon with these