Shifting Voices: Teaching Transformers to Rewrite Across Genres

STATISTICAL SOFTWARE (16:954:577:01)

Ansh Datir Sai Lohith Chimbili Jayesh Kasbe

Group 18

Abstract

Genre-conditioned rewriting (genre transfer) is an emerging NLP problem where a model must rewrite a paragraph in a target genre while preserving the original meaning, events, and named entities. This is substantially more challenging than tasks like sentiment transfer because successful rewrites require maintaining narrative content while altering tone, vocabulary, imagery, and stylistic markers. We develop a genre-conditioned text rewriting system for six genres (Romance, Gothic Horror, Detective Mystery, Fantasy, Science Fiction, and Comedy) using a cleaned and balanced dataset of 9,575 paragraphs from Project Gutenberg and openAI's API of Chat GPT-4 (self-mode and transfer-mode). We compare three systems: (i) GPT-4 prompting, (ii) Qwen 2.5–1.5B–Instruct base model generation, and (iii) Qwen 2.5–1.5B–Instruct fine-tuned using LoRA. Evaluation is performed using a composite metric combining style accuracy, content preservation, named entity preservation, and fluency. Results show that GPT-4 performs best overall, particularly on out-of-domain data, while LoRA fine-tuning yields small but consistent in-domain improvements over the base model. Importantly, our findings highlight that genre transfer is feasible when supported by high-quality human-like rewrites and well-aligned parallel training data, underscoring the central role of supervision quality in stylistic text generation.

1 Introduction

1.1 Problem Statement

We address the NLP problem of **genre-conditioned text rewriting**, where the goal is to transform an input narrative paragraph into a specified **target genre** while preserving:

- **Meaning and main events** (logical sequence, actions, outcomes),
- **Named entities** (character names, locations, identities, relationships),
- while modifying **style** (tone, narration, vocabulary, atmosphere, imagery) to match the target genre.

We focus on six genres: Romance, Gothic Horror, Detective Mystery, Fantasy, Science Fiction, and Comedy.

1.2 Why This Problem Is Challenging

Genre transfer requires a model to hold two objectives in tension: (1) maintain semantic fidelity and entity consistency, and (2) produce strongly genre-consistent stylistic signals. In contrast to simpler transfer tasks (e.g., sentiment flipping), genre transfer involves broad stylistic shifts and risks **hallucination** (adding new events/characters) or **content drift** (losing key plot details).

1.3 Project Approach Overview

Our approach contains three major components:

1. **Dataset construction:** A balanced dataset of 9,575 paragraphs built from Project Gutenberg and GPT-4. We create two modes: self-mode (original paragraphs) and transfer-mode (GPT-4 rewrites into target genres).
2. **Modeling:** We generate genre-conditioned rewrites using the Qwen 2.5–1.5B–Instruct model and apply LoRA fine-tuning to obtain a domain-adapted variant, which we compare against the base model.
3. **Evaluation:** We design an automatic evaluation suite combining (a) style accuracy via a genre classifier, (b) content preservation via embedding cosine similarity, (c) named entity overlap via spaCy NER, and (d) fluency via GPT-2 perplexity.

1.4 Baselines, Architectures, and Results Summary

Baselines: GPT-4 prompted rewrites and Qwen Instruct base model rewrites.
Architectures used: Qwen 2.5–1.5B–Instruct for generation (base + LoRA), and an MPNet-embedding classifier (baseline logistic regression and a neural classifier).

Headline result: While GPT-4 remains the strongest model overall, especially out-of-domain, consistent in-domain improvements from LoRA fine-tuning demonstrate the viability of building dedicated genre-transfer models.

2 Data Collection

2.1 Data Source

We use **Project Gutenberg**, a public-domain corpus of classic novels, as our primary source of narrative paragraphs. This provides rich stylistic variety and legal accessibility.

2.2 Data Size and Construction

Our final dataset contains **9,575 paragraphs** of length approximately **80–220 words**, balanced across six genres. It contains:

- **Self-mode (6,000):** Original paragraphs labeled by their source genre.
- **Transfer-mode (3,575):** The same paragraphs rewritten into a different target genre using GPT-4 (synthetic supervision).

2.3 Cleaning and Preprocessing

We applied cleaning steps to make the dataset suitable for modeling:

- Removed non-narrative artifacts (headers/footers, chapter titles, metadata noise where applicable),
- Standardized paragraph lengths (filtering to 80–220 words),
- Ensured class balance across all six genres,

2.4 Time Period of Data

Project Gutenberg primarily contains classic literature spanning the 19th and early 20th centuries. Our in-domain distribution therefore reflects older narrative style, which becomes relevant when evaluating out-of-domain generalization.

2.5 Train/Validation/Test Splits and Evaluation Sets

We first split the tokenized dataset into an 90/05/05 train/validation/test split using a fixed random seed (42).

- **Train:** 8617 examples
- **Validation:** 479 examples

- **Test:** 479 examples

Because genre transfer is the core task, we focus evaluation on **transfer-mode** rows. In the 479-example test set, **193 examples are transfer-mode**, and we report generation metrics primarily on these transfer examples (self-mode rows are not true cross-genre rewrites).

Two Complementary Evaluation Datasets: In addition to the held-out in-domain test split above, we evaluate generalization using an external out-of-domain set:

- **Held-out in-domain test set (transfer subset):** 193 transfer-mode samples drawn from the 479-example test split, separated from training/validation for LoRA fine-tuning.
- **External out-of-domain validation set:** 400 samples constructed by generating input paragraphs using GPT-4, used to test distribution shift beyond Project Gutenberg.

2.6 Training Sampling Strategy (Up-weighting Transfer-Mode)

To ensure the model learns stronger cross-genre transformation behavior, we up-weight transfer-mode examples during training using a weighted sampler. Specifically, we assign higher sampling weight to transfer rows than self-mode rows (1.0 vs. 0.25), encouraging the model to observe cross-genre supervision more frequently per epoch.

2.7 Class Distribution and Characteristics

We maintain a balanced dataset across the six genres to prevent the evaluation classifier from biasing toward majority classes and to ensure fair comparisons.

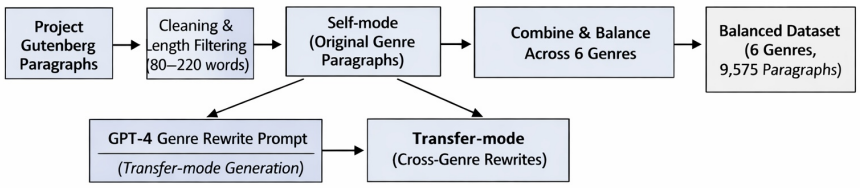

Figure 1: **Dataset construction pipeline.**

Field	Example (truncated)
Source Style	Comedy
Target Style	Detective Mystery
Input	"No, no, no! What yer got? Give me somethin' ..."
Output	"In the dim, flickering light of the neglected room ..."
Mode	Transfer

Table 1: Structure of a transfer-mode datapoint with truncated text.

3 Experiments

3.1 Models and Baselines

We compare three generation systems:

1. **GPT-4 (prompted rewrites):** A strong external baseline generating rewrites directly from an instruction prompt using openAI's API.

2. **Qwen 2.5–1.5B-Instruct:** An instruction-tuned variant of the pretrained Qwen 2.5–1.5B model, used for genre-conditioned text rewriting under the same task framing as GPT-4. The model is obtained from the Qwen open-source release (`https://huggingface.co/Qwen/Qwen2.5-1.5B-Instruct`). It is a decoder-only Transformer language model consisting of 24 Transformer layers with 16 self-attention heads per layer. Each layer includes multi-head self-attention and a feed-forward network using the SiLU (Swish) activation function, with a final linear projection and softmax layer for next-token probability prediction.

3. **Qwen 2.5–1.5B–Instruct + LoRA:** Fine-tuned model trained on self-mode + transfer-mode to improve genre transfer within domain.

We also train style classifiers used for evaluation:

- **Baseline style classifier:** MPNet embeddings + logistic regression.
- **Final style classifier:** MPNet embeddings + neural classifier (hidden dim 256).

Other models used in Evaluations:

- **Sentence Transformer Model (MPNet):** We use the pretrained Sentence Transformer `all-mpnet-base-v2` model (`https://huggingface.co/sentence-transformers/all-mpnet-base-v2`) to obtain fixed-length sentence embeddings. The model is based on a Transformer encoder with **12 layers**, **12 self-attention heads per layer**, and a hidden size of 768. Feed-forward layers use a **GELU** activation function. Mean pooling over token embeddings produces a single 768-dimensional sentence representation.
- **spaCy Named Entity Recognition Model:** Named entity preservation is evaluated using spaCy's pretrained English NER pipeline, `en_core_web_trf` (`https://spacy.io/models/en#en_core_web_trf`). This pipeline uses a Transformer-based encoder followed by a transition-based NER component to identify entities such as persons and locations. The model is applied without task-specific fine-tuning and is used solely to measure overlap between entities extracted from source and generated text.

3.2 How We Used the Architecture for Our Problem (No Re-Teaching)

We do not modify the underlying Transformer architecture. Instead, we operationalize genre transfer as **instruction-following conditional generation**:

- Inputs are formatted as chat-style messages containing the paragraph and explicit genre transfer instructions.
- Outputs are complete rewritten passages that preserve entities and events while shifting style.
- LoRA is used to efficiently adapt Qwen to our genre rewriting distribution under limited compute.

3.3 Fine-Tuning Strategy (LoRA)

We fine-tuned Qwen 2.5–1.5B–Instruct using Low-Rank Adaptation (LoRA) to efficiently adapt the pretrained generator to the genre transfer task under limited computational resources. Our fine-tuning strategy emphasizes learning strong cross-genre stylistic transformations while preserving semantic content and named entities. We rely on the model's pretrained token embeddings and do not introduce separate or task-specific word embeddings.

Stability-focused optimization: Training is performed for **2 epochs** with gradient accumulation and a polynomial learning-rate schedule.

LoRA Configuration: We apply LoRA with rank $r = 16$ and scaling factor $\alpha = 32$, enabling sufficient adaptation capacity while limiting the number of trainable parameters. A LoRA dropout of 0.05 is used to regularize the low-rank updates and reduce overfitting.

Optimization and Scheduling: We use a polynomial learning-rate schedule with warmup, decaying from 7×10^{-6} to 7×10^{-7} with a warmup ratio of 0.03. Training uses an effective batch size of 16 via gradient accumulation.

Regularization and Stability: Multiple regularization mechanisms are employed:

- **LoRA dropout** (0.05) to regularize adapter updates.
- **Weight decay** (0.1) via AdamW to discourage large parameter updates.
- **Gradient clipping** with max norm 1.0 to stabilize optimization.
- **Gradient checkpointing** to reduce memory usage and enable stable training.

Step	Training Loss	Validation Loss
1000	1.652400	1.642911

Figure 2: **LoRA fine-tuning training/validation loss.**

3.4 Evaluation Metrics

We evaluate each system on style and faithfulness using four metrics.

Style Accuracy (Genre Consistency): We evaluate stylistic correctness using a genre classifier trained on MPNet sentence embeddings. For each generated paragraph, the classifier outputs a probability distribution over the six genres. Rather than using a strict top-1 accuracy alone, we adopt a graded scoring scheme to account for near-miss predictions:

- score = 1.0 if the target genre is ranked first,
- score = 0.5 if the target genre is ranked second,
- score = 0.25 if the target genre is ranked third,
- score = 0 otherwise.

This approach captures the strength of genre cues while reducing sensitivity to classifier uncertainty.

Content Preservation: We compute cosine similarity between Sentence Transformer embeddings of the original paragraph and the rewritten output. Higher similarity suggests better semantic preservation.

Named Entity Preservation: We use spaCy Named Entity Recognition (NER) to extract entities from input and output and compute overlap (entity retention), approximating preservation of characters and places.

Fluency: We compute perplexity using GPT-2 and convert it to a bounded fluency score:

$$\text{Fluency} = \frac{1}{1 + \text{Perplexity}}$$

This rewards grammaticality and naturalness in generated outputs.

3.5 Final Composite Metric

We aggregate metrics into a single weighted score:

$$\text{Score} = 0.3 \cdot \text{Style} + 0.3 \cdot \text{Content} + 0.3 \cdot \text{NER} + 0.1 \cdot \text{Fluency}$$

This weighting emphasizes genre correctness and faithfulness, while still rewarding fluent writing.

3.6 Style Classifier Details (Used for the Problem)

Our evaluation depends on a genre classifier:

- **Input:** 768-dimensional MPNet sentence embeddings
- **Baseline:** Logistic Regression (F1 $\approx$ 0.73)

- **Final classifier:** Neural classifier over embeddings (hidden dim 256, dropout, Adam) achieving F1 $\approx$ 0.81 on original novel paragraphs

Genre	Prec.	Rec.	F1	Supp.
Comedy	0.81	0.84	0.83	320
Detective Mystery	0.80	0.85	0.82	320
Fantasy	0.77	0.81	0.79	320
Gothic	0.80	0.77	0.79	320
Romance	0.77	0.80	0.79	320
Science Fiction	0.91	0.78	0.84	315
Accuracy			**0.81**	1915

(a) Neural classifier (MPNet + RNN)

Genre	Prec.	Rec.	F1	Supp.
Comedy	0.72	0.75	0.73	320
Detective Mystery	0.72	0.74	0.73	320
Fantasy	0.76	0.72	0.74	320
Gothic	0.70	0.72	0.71	320
Romance	0.72	0.69	0.71	320
Science Fiction	0.79	0.77	0.78	315
Accuracy			**0.73**	1915

(b) Baseline classifier (MPNet + Logistic Regression)

Table 2: **Style classifier performance.** Comparison of the neural and baseline classifiers.

3.7 Results

We report results on:

- **Held-out in-domain test set (193 samples)**
- **External out-of-domain validation set (400 samples)**

Key finding (Style Classifier Stress Test): Classifier performance drops from approximately 81% F1 on original novel paragraphs to approximately 41% F1 on GPT-4 style-transferred outputs, where inputs are target genres and outputs are GPT-4–rewritten paragraphs indicating that although GPT-4 preserves content well, its genre-conditioned rewrites often contain weaker or more subtle genre cues.

This suggests that even strong, widely deployed models such as GPT-4 may not consistently produce sufficiently distinct stylistic transformations for genre transfer, resulting in reduced genre separability under automatic classification.

Setting	Model	Content	Style	NER	Fluency	Overall
Unseen	GPT-4	0.784	0.484	0.640	0.0099	**0.573**
	LoRA-Qwen	0.764	0.309	0.450	0.0149	0.458
	Qwen-Instruct-Base	0.773	0.289	0.474	0.0146	0.462
Held-Out	GPT-4	0.803	0.380	0.725	0.0159	**0.574**
	LoRA-Qwen	0.768	0.259	0.587	0.0226	0.486
	Qwen-Instruct-Base	0.782	0.247	0.572	0.0223	0.483

Table 3: **Model comparison across evaluation metrics.**

Overall results summary: On in-domain data, LoRA-Qwen shows small but consistent improvements over Qwen Instruct base. On out-of-domain data, GPT-4 clearly outperforms both Qwen models across the composite metric.

4 Conclusions

4.1 Conclusions from Experiments

Across evaluations, GPT-4 remains the strongest system overall, particularly out-of-domain, due to robust content and named entity preservation; however, style classifier stress tests reveal that its stylistic transformations are often subtle, resulting in weaker genre separability.

Fine-tuning Qwen 2.5–1.5B–Instruct with LoRA yields small but consistent in-domain improvements over the base model, indicating that lightweight adaptation helps capture dataset-specific stylistic patterns, though gains remain modest due to limited training data and soft genre cues in GPT-4–generated targets. Out-of-domain generalization remains weak for both Qwen models.

Notably, Qwen models achieve higher fluency under GPT-2–based perplexity scoring, highlighting their potential as efficient backbones for future in-domain fine-tuning.

4.2 Limitations

Our results are shaped by practical resource constraints. Generating high-quality ground-truth genre rewrites requires costly GPT-4 API calls, and LoRA fine-tuning remains computationally demanding, which limits the scale and diversity of supervision. Additionally, automatic evaluation of genre transfer relies on proxy metrics, such as classifier-based style accuracy, which may not fully capture nuanced stylistic differences.

The absence of human-written or human-verified target paragraphs further constrains the strength of stylistic supervision. A more robust genre classifier and human evaluation would likely provide more reliable signals for both training and evaluation.

4.3 How to Improve Results

Future improvements are likely to depend more on the quality and diversity of supervision than on model size alone:

- Expand training data with more diverse and modern narrative sources to reduce domain bias,
- Incorporate human-written or human-verified transfer-mode targets to strengthen stylistic supervision,
- Use stronger evaluation frameworks, including human judgment rubrics and improved genre classifiers with higher discriminative power.

4.4 What We Learned

We learned that effective genre transfer requires carefully balancing stylistic strength with semantic and named entity preservation, and that evaluation for style transfer demands multi-dimensional metrics rather than reliance on a single proxy. Our results from the in-domain improvements of LoRA-fine-tuned Qwen Instruct over the base Instruct model further show that genre transfer is feasible within an instruction-following generation framework, but achieving strong and clearly separable stylistic transformations remains challenging.

This observation highlights the relevance of our work: Despite the availability of powerful general-purpose language models like GPT-4, the field still lacks strong and controllable systems for genre transfer. Our results provide clear evidence that targeted fine-tuning, combined with high-quality novel-based training data, can lead to measurable improvements in genre fidelity beyond off-the-shelf prompting.

5 Related Work

Paper chosen: Fu et al. (2018), *"Style Transfer in Text: Exploration and Evaluation"*.

Fu et al. survey neural text style transfer and emphasize the central challenge of separating "style" from "content" in natural language. They describe how many approaches either preserve meaning but fail to produce strong stylistic signals, or produce stylistic text that drifts semantically. The paper also highlights that evaluation is difficult: metrics like classifier-based style accuracy can be misleading if the classifier overfits superficial cues, and semantic preservation metrics often miss subtle meaning changes.

This perspective directly relates to our findings. We observed that GPT-4 preserves content and named entities extremely well, but automatic style classification suggests its genre cues can be subtle - mirroring Fu et al.'s discussion of the style/content trade-off and evaluation limitations. Our multi-metric evaluation framework (style, content similarity, NER overlap, fluency) is motivated by the same need for balanced evaluation rather than relying on any single metric.

References

- Fu, Z., Tan, X., Peng, N., Zhao, D., & Yan, R. (2018). *Style Transfer in Text: Exploration and Evaluation.* (Add venue details if required by your course.)