# Authorship Style Transfer with Inverse Transfer Data Augmentation

Anonymous ACL submission

## Abstract

Authorship style transfer aims to modify the style of neutral text to match the unique speaking or writing style of a particular individual. While Large Language Models (LLMs) present promising solutions, their effectiveness is limited by the small number of in-context learning demonstrations, particularly for authorship styles not frequently seen during pre-training. In response, this paper proposes an inverse transfer data augmentation (ITDA) method, leveraging LLMs to create (neutral text, stylized text) pairs. This method involves removing the existing styles from stylized texts, a process made more feasible due to the prevalence of neutral texts in pre-training. We use this augmented dataset to train a compact model that is efficient for deployment and adept at replicating the targeted style. Our experimental results, conducted across four datasets with distinct authorship styles, establish the effectiveness of ITDA over traditional style transfer methods and forward transfer using GPT-3.5. For further research and application, our dataset and code are openly accessible at `https://github.com/AnonymousRole/ITDA`.

## 1 Introduction

Text style transfer, a technique that rewrites text into a specific style while preserving content, has garnered considerable attention in recent years. Most existing methods excel at style attribute transfer, which entails shifting text along particular style dimensions such as sentiment, formality, and politeness. We refer to styles with well-defined attributes as polar styles. In contrast, authorship style (Xu et al., 2012; Carlson et al., 2018) constitutes a unique category describing an individual's writing or speaking style. It is characterized by word choice, structure, quirks, and topics, but lacks well-defined attributes, making it challenging to categorize as positive/negative or polite/impolite. Figure 1 (a) presents examples that illustrate how authorship style encompasses more intricate and indefinable elements compared to polar styles.

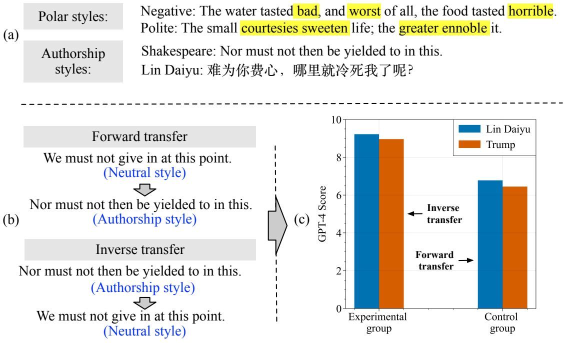

Figure 1: Illustration of (a) polar style with highlighted stylized words and authorship style; (b) forward transfer and inverse transfer; (c) inverse transfer demonstrating more promising performance than forward transfer.

This paper investigates authorship style transfer, which aims to transform neutral style text into text that aligns with a specific author's writing style, a topic previously addressed in studies like (Syed et al., 2020) and (Patel et al., 2022). This problem offers diverse applications, including creating personalized digital assistants that communicate in a user's chosen style, aiding students and researchers in understanding different authors' unique writing styles—important for literary studies and education—and improving privacy by altering an individual's writing style to conceal their identity, particularly useful for sensitive documents.

Recently, Large Language Models (LLMs) such as GPT-3.5 (Ouyang et al., 2022) and GPT-4 (OpenAI, 2023) have been utilized for their strong generalization abilities to infuse desired styles into generic neutral texts—a process known as **forward transfer**—through in-context learning with a few demonstrations. The limited input length of LLMs constrains the number of feasible demonstrations, impeding comprehensive instruction on a target author's style, especially for less-covered authorship styles in LLM pre-training. Research, like (Reif et al., 2022; Patel et al., 2022), suggests integrating descriptive adjectives into prompts to capture the target author's style. While this approach reduces the need for extensive demonstrations, condensing an author's distinct style into a few words remains challenging.

Instead of relying on in-context learning with limited examples to guide LLMs in authorship style transfer, we propose an alternative method: training a compact, specialized model using abundant examples augmented from existing stylized texts. This method is more effective for dealing with uncommon authorship styles and also cuts down on costs related to model deployment and inference. The crucial part of this approach involves creating high-quality pairs of neutral and stylized text for training our compact model. Leveraging LLMs, we've developed **Inverse Transfer** Data Augmentation (ITDA) to remove specific styles from texts, transforming them into neutral texts. These transformed texts are then utilized in reverse – from neutral to stylized – to train our compact model. The "inverse" data augmentation method often outperforms the conventional "forward" approach, as LLMs typically excel at creating neutral rather than highly stylized texts due to the prevalence of neutral texts in pre-training. We illustrate this concept using diagrams in Figure 1 (b) and have conducted a pilot study, detailed in Section 4, demonstrating the effectiveness of this inverse approach. The results, presented in Figure 1 (c), exhibit an impressive approximately 40% increase from forward to inverse transfer in terms of GPT-4 score (OpenAI, 2023).

In implementing ITDA, our focus includes dynamic prompting and stylized text augmentation. Dynamic prompting aims to identify the most appropriate demonstrations for stylized text, effectively aiding in style removal process. This is achieved by clustering the stylized corpus and annotating neutral texts for the most representative stylized text in each cluster, enabling dynamic prompting with minimal human labeling efforts. Additionally, to address the challenge of limited available stylized texts in less common styles, we utilize LLMs to generate new texts in these specific styles. The key contributions are summarized as follows:

- We propose ITDA, an inverse transfer data augmentation method designed to address authorship style transfer. Leveraging LLMs, we perform inverse transfer to convert stylized texts into neutral texts, resulting in a corpus that trains a compact and deployable model.

- We introduce a clustering-based dynamic prompt selection method to bolster the performance of inverse transfer. We also leverage LLMs to synthesize new texts in the target style to mitigate data scarcity.

- Through comprehensive experiments conducted on four authorship-stylized datasets in both Chinese and English, we demonstrate the advantages of ITDA compared to traditional style transfer approaches and direct forward transfer by GPT-3.5.

## 2 Related Work

Style transfer methods can be roughly classified into three categories: original representation revision, latent representation revision, and in-context learning on LLMs. The first two are primarily utilized for style attribute transfer, with several works also applying to authorship style transfer.

Original representation revision (Sudhakar et al., 2019; Reid and Zhong, 2021) follows a "delete-generate" framework (Li et al., 2018), in which the original stylized words are removed and the desired stylized words are added. While offering excellent interpretability by modifying original words, this approach struggles with authorship style transfer, as identifying stylized words within the authorship-stylized text is challenging.

Latent representation revision (Wang et al., 2019; Xu et al., 2020; Xiao et al., 2021) involves revising the original text's latent representation within a Euclidean space, guided by content and style loss, and then decoding to generate the target-stylized text. (Syed et al., 2020; Riley et al., 2021) explore its application in authorship style transfer. However, directly manipulating the latent representation may lead to a low-density region, resulting in unpredictable and low-quality text output. Moreover, directly modifying the latent representation lacks nuanced control over the target style, as discussed in (Jin et al., 2022).

In-context learning using LLMs is currently a favored method for style transfer. A prime example is the Prompt-and-Rerank technique with GPT-2 (Suzgun et al., 2022), which generates multiple outputs for each input and ranks them based on factors like textual similarity, style, and fluency. Researches like (Patel et al., 2022) and (Reif et al.,

2022) incorporate descriptive adjectives extracted from stylized texts into prompts for GPT-3.5 to mimic a target author's style. The former applies the same demonstrations across different styles, while the latter varies them according to the style. However, distilling an author's style into a few words is complex, and the limited demonstrations may not fully capture the nuances of less common styles. While the latter also uses inverse transfer, their focus is on automating demonstrations rather than data augmentation to provide a compact model with more extensive training examples.

## 3 Problem Definition

**Authorship Style.** Neutral text, devoid of a particular style, is common across various articles and platforms. This is precisely why we choose it as the transfer subject. Stylized text, on the other hand, contains distinctive expressive elements, such as sentiment and formality. Authorship style is a special type of stylized text which embodies an individual author's unique word choices, writing structures and emotional inclinations. However, unlike other well-defined styles, the authorship style lacks clearly defined attributes, making it challenging to summarize its characteristics in a few words.

**Authorship Style Transfer.** Given a target authorship style $s$, and an input text $x$ with the neutral style, our objective is to transform it into text $y$ that exhibits the style $s$. We refer to this conversion process as **forward transfer**. Conversely, the process of converting $y$ back to $x$, where the style $s$ is removed from $y$, is termed **inverse transfer**. We use the notation $D^s$ to represent a collection of texts that exhibit an authorship style $s$.

## 4 Pilot Study

As analyzed in Section 1, LLMs are more skilled at inverse transfer rather than forward transfer. We design the following controlled experiments to validate this assumption.

**Datasets.** We prepare two distinct authorship-stylized datasets. The first style embodies the essence of "Lin Daiyu", an iconic figure from Chinese ancient literature, while the latter style captures the essence of "Trump", a renowned American enterpriser and politician. These two datasets consist of 500 and 2,000 textual pieces respectively.

**Experimental Protocol.** We design an experimental group for inverse transfer, where the style is removed from a stylized text, and a control group for forward transfer, where a target style is added to a neutral text. We employ the few-shot prompting technique on GPT-3.5 to validate our hypothesis. For each authorship-stylized dataset, we randomly select eight (stylized, neutral) pairs from the test sets depicted in Table 2 as demonstrations for inverse transfer. These pairs are then inverted to form eight (neutral, stylized) pairs, which serve as demonstrations for forward transfer.

The stylized inputs for the experimental group and the neutral inputs for the control group are paired and sampled from the (stylized, neutral) pairs in the test sets, excluding those selected as demonstrations. With their respective eight demonstrations, we prompt GPT-3.5 to output the corresponding counterparts for the stylized or neutral inputs of the two groups.

**Observation.** We measure the performance of inverse and forward transfer using GPT-4, the most advanced commercial LLM, as the evaluator for an objective evaluation. GPT-4 evaluates the output comprehensively based on three dimensions: content preservation, style transfer strength, and text fluency, assigning scores from 1 to 10. Evaluation metrics in detail can been seen in Section 6.1.

Figure 1 (c) illustrates that, in comparison with control group for forward transfer, the experimental group for inverse transfer outperforms around 40% in terms of GPT-4 score. We conjecture that neutral text, with its simpler form, is relatively easy to learn. During pre-training, LLMs are exposed to a greater volume of neutral text than specific authorship style text. This increased exposure augments the ability of LLMs to generate neutral text. Guided by this observation, we craft our inverse transfer data augmentation for authorship style transfer.

## 5 ITDA

**Framework Overview.** The basic idea of ITDA is to augment data by inverse transfer (*i.e.*, stylized to neutral) on LLMs and then fine-tune a small model based on these augmented pairs. This idea surpasses the direct few-shot prompting for forward transfer, which primarily restricted by the input length of LLMs. Given the intricate nature of the authorship style, effectively transferring arbitrary neutral text demands a sufficient number of $\{(x, y)\}$ demonstrations to facilitate a comprehensive understanding of the authorship style by

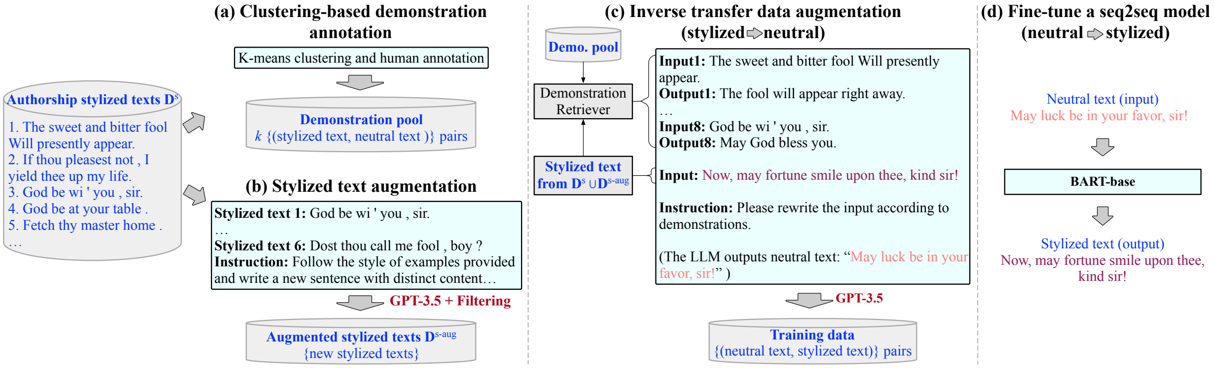
(a) Clustering-based demonstration annotation
K-means clustering and human annotation
Demonstration pool
k {(stylized text, neutral text )} pairs
Authorship stylized texts D^s
1. The sweet and bitter fool Will presently appear.
2. If thou pleasest not , I yield thee up my life.
3. God be wi ' you , sir.
4. God be at your table .
5. Fetch thy master home .
…
(b) Stylized text augmentation
Stylized text 1: God be wi ' you , sir.
…
Stylized text 6: Dost thou call me fool , boy ?
Instruction: Follow the style of examples provided and write a new sentence with distinct content…
GPT-3.5 + Filtering
Augmented stylized texts D^s-aug
{new stylized texts}
(c) Inverse transfer data augmentation (stylized ⇨ neutral)
Demo. pool
Demonstration Retriever
Stylized text from D^s ∪D^s-aug
Input1: The sweet and bitter fool Will presently appear.
Output1: The fool will appear right away.
…
Input8: God be wi ' you , sir.
Output8: May God bless you.
Input: Now, may fortune smile upon thee, kind sir!
Instruction: Please rewrite the input according to demonstrations.
(The LLM outputs neutral text: “May luck be in your favor, sir!” )
GPT-3.5
Training data
{(neutral text, stylized text)} pairs
(d) Fine-tune a seq2seq model (neutral ⇨ stylized)
Neutral text (input)
May luck be in your favor, sir!
BART-base
Stylized text (output)
Now, may fortune smile upon thee, kind sir!

Figure 2: The ITDA framework, featuring four key components: (a) We cluster $D^s$ and annotate representative texts to establish the demonstration pool. (b) We augment stylized texts $D^s$ to create $D^{s-aug}$. Based on both, (c) we execute inverse transfer from stylized texts $D^s \cup D^{s-aug}$ to neutral texts by dynamically selecting demonstrations from the pool. Finally, (d) we fine-tune a compact seq2seq model with the augmented (neutral, stylized) data.

LLMs. Unfortunately, the length limitation prevents the inclusion of a large number of examples, potentially prompting LLMs to draw style inferences from their pre-existing knowledge beyond the limited demonstrations. For instance, if the target is to transfer text into the style of “Lin Daiyu”, LLMs may inadvertently mirror a classical Chinese style rather than the specific style of “Lin Daiyu”. Similarly, when aiming to emulate a “Shakespeare” style, LLMs may unintentionally reflect an archaic English style. Unlike the direct forward transfer, we opt for (c) the better inverse transfer process, as evidenced by the pilot study in Section 4, to generate a large number of $\{(x, y)\}$ pairs by LLMs and (d) train a compact model to gain exposure to a sufficient number of training examples. Note we train a separate compact model for each style $s$.

In addition to the main components (c) and (d), we introduce two enhancement strategies, (a) and (b), for inverse transfer data augmentation. The first strategy involves establishing a demonstration pool with minimal human labeling effort to enable dynamic demonstration retrieval for in-context learning during the inverse transfer data augmentation process. This pool is created by clustering the samples in $D^s$ and selecting the most representative text samples for labeling. The second strategy focuses on augmenting $D^s$. Recognizing the often limited availability of collected authorship-stylized text $D^s$, we utilize LLMs to produce additional authorship-stylized text $D^{s-aug}$, thereby enhancing the diversity of final produced (stylized, neutral) data pairs. We illustrate the four components in Figure 2 and explain them as below. Note although ITDA is proposed to address authorship style transfer, it can certainly be leveraged for broader style transfers, such as sentiment or formality transfer.

**(a) Clustering-based Demonstration Annotation.** We utilize dynamic prompting for inverse transfer data augmentation. However, dynamic prompting necessitates annotating additional (stylized, neutral) pairs for demonstrations. To minimize human labeling efforts while creating the most representative demonstration pool, we propose a clustering-based approach: (1) We initially employ Sentence-BERT (Reimers and Gurevych, 2019) to represent each sentence $y \in D^s$ and then apply the k-means algorithm to cluster them into $k$ categories. The determination of $k$ (*e.g.*, 40 for Lin Daiyu) is based on the silhouette coefficient metric (Dinh et al., 2019), with specific details provided in Appendix A.1. (2) Subsequently, we select the sample closest to the center of each cluster as the representation of that cluster, resulting in $k$ representative texts. (3) We annotate the counterpart in neutral style for each representative text by leveraging LLMs initially and then validating through human annotation.

While the clustered demonstration pool is smaller than $D^s$, it's meticulously designed to encapsulate the given authorship style, facilitating an effective and efficient retrieval solution. In addition to the benefit of reducing human labeling efforts, the clustering-based prompting, as utilized by (Zhang et al., 2022; Li et al., 2023), confirms an additional advantage: demonstrations selected from different clusters exhibit diversity, thus aiding in the inference of a wide range of new inputs.

**(b) Stylized Text Augmentation.** Collecting adequate text in a specific authorship style can be challenging, especially when the style is scarce or unavailable as open-source datasets. To overcome this limitation, we leverage LLMs to augment the corpus of author-stylized $D^s$ into $D^{s-aug}$. We randomly sample six sentences from $D^s$ and combine them with the instruction like "*Follow the style of examples provided and write a novel sentence with distinct content. The newly generated text needs to cover a wide range of topics across various fields.*" This prompt guides the LLM to replicate the given style and generate new texts.

Replicating pure style text is considerably less challenging than style transfer, as the content of simulated texts can be freely expressed without the requirement for alignment with input texts. Existing products such as Character AI[1] and research projects like RoleLLM (Wang et al., 2023) and Character-LLM (Shao et al., 2023) also demonstrate efforts to enable LLMs to generate dialogues with specific styles, providing evidence of LLMs' capabilities in replicating styles. Moreover, to improve the stylistic quality of the synthesized text, we conduct style examination using a binary style classifier (0-1) to filter out texts with inappropriate styles. This classifier is also employed to evaluate the performance of authorship style transfer, and its functionality is elaborated on in Section 6.1.

**(c) Inverse Transfer Data Augmentation.** Using the prepared demonstration pool and augmented authorship-stylized text corpus $D^s \cup D^{s-aug}$, we dynamically select the most relevant demonstrations to perform inverse transfer for each stylized text $y \in D^s \cup D^{s-aug}$, converting it into its neutral counterpart $x$. To accomplish this, we evaluate the similarity between $y$ and each $y'$ in the demonstration pool using Sentence-BERT. We then select the eight most similar demonstrations, forming pairs $\{(y', x')\}$, as dynamic demonstrations. Detailed prompts are available in Appendix A.7.

**(d) Fine-tune a Compact Model for Forward Transfer.** The resulting pairs $\{(y, x)\}$ are reversed to create $\{(x, y)\}$ corpus, based on which we fine-tune a BART model (Chipman et al., 2010) for forward transfer, enabling the transformation of any new neutral input text into authorship style $s$.

# 6 Experiment

## 6.1 Experimental Settings

**Dataset.** We create four authorship-style datasets, encompassing the styles of "Shakespeare", "Trump", and "Lyrics" in English, as well as "Lin Daiyu" in Chinese. Among them, the dataset "Shakespeare" consists of sentences written by Shakespeare, as published by He et al. (2019). The dataset "Lyrics" features sentences from modern lyric poetry, as published by Krishna et al. (2020). "Donald Trump" encompasses speeches made by Trump and is collected from the publicly available websites[2]. "Lin Daiyu" consists of sentences spoken by the character Lin Daiyu, extracted from the Chinese novel "The Dream of Red Mansion".

For each authorship style, we partition the collected stylized texts into a training data (original) and a test set. Subsequently, we augment the stylized texts in the original training data using the stylized text augmentation step in the proposed ITDA, resulting in the augmented training data, as depicted in Table 2. Each stylized text in the three sets is paired with a neutral text. In the two training data, the corresponding neutral texts are generated by the proposed ITDA, while those in the test set are annotated by humans to ensure their correctness for evaluation. Specifically, we engage three language experts, with two independently writing neutral text for each stylized text in the test set following annotation criteria aimed at preserving the content while removing the style. The third expert then selects the superior neutral text that adheres to the criteria from the two annotations. If neither text meets the criteria, the process is repeated with re-annotation. Both stylized and neutral texts in the training data are used for training. In the test sets, neutral texts are fed into different models to predict stylized texts, and the corresponding original stylized texts serve as the ground truth for evaluation.

**Style Classifier.** We train a style classifier for two main purposes: (1) to filter out texts with inappropriate styles during the stylized text augmentation step described in Section 5, and (2) to evaluate the style transfer capabilities of various comparison methods. For English datasets, we initialize the classifier with BERT[3], while for Chinese datasets, we utilize ChineseRoBERTa[4]. To ensure its quality, we choose to train it using the original training data consisting of collected stylized texts and corresponding generated neutral texts. It's important to note that, similar to having a distinct style transfer model for each authorship style, we also train a distinct classifier for each style. Across four distinct

[1] https://beta.character.ai/

[2] https://www.nytimes.com; https://edition.cnn.com

[3] https://huggingface.co/bert-base-cased

[4] https://huggingface.co/uer/chinese_roberta_L-12_H-768

| Approach | Lin Daiyu |  |  |  | Shakespeare |  |  |  | Trump |  |  |  | Lyrics |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | BLEU | BS | SC | GPT-4 | BLEU | BS | SC | GPT-4 | BLEU | BS | SC | GPT-4 | BLEU | BS | SC | GPT-4 |
| **Original Representation Revision** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| DRG (Delete-Only) | - | - | - | - | 0.21 | 0.63 | 0.54 | 1.93 | 0.16 | 0.71 | 0.34 | 1.69 | 0.24 | 0.73 | 0.38 | 2.28 |
| DRG (Delete-and-Retrieve) | - | - | - | - | 0.33 | 0.67 | 0.48 | 3.15 | 0.24 | 0.73 | 0.39 | 2.37 | 0.42 | 0.79 | 0.33 | 3.04 |
| Transform DRG (Delete Only) | 0.33 | 0.62 | 0.21 | 4.28 | 0.51 | 0.70 | 0.36 | 3.89 | 0.21 | 0.74 | 0.12 | 3.04 | 0.59 | 0.85 | 0.27 | 3.70 |
| **Latent Representation Revision** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| CTAT | 0.26 | 0.54 | 0.38 | 2.16 | 0.36 | 0.69 | 0.41 | 2.93 | 0.30 | 0.73 | 0.50 | 3.28 | 0.30 | 0.75 | 0.40 | 3.19 |
| CP-VAE | - | - | - | - | 0.25 | 0.64 | 0.38 | 2.47 | 0.14 | 0.71 | 0.47 | 2.52 | 0.29 | 0.73 | 0.39 | 2.75 |
| TSST | 0.48 | 0.67 | 0.45 | 4.75 | 0.46 | 0.73 | 0.54 | 4.11 | 0.44 | 0.80 | 0.51 | 4.76 | 0.62 | 0.83 | 0.43 | 3.14 |
| **Few-shot Prompting on LLMs** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| Prompt-and-Rerank (GPT-2) | 0.24 | 0.53 | 0.41 | 3.49 | 0.67 | 0.85 | 0.17 | 3.49 | 0.37 | 0.76 | 0.35 | 3.90 | 0.58 | 0.87 | 0.39 | 3.61 |
| Few-shot (GPT-3.5) | 0.66 | 0.81 | 0.44 | 6.78 | 0.65 | 0.88 | 0.47 | 7.17 | 0.67 | 0.89 | 0.40 | 6.45 | 0.67 | 0.91 | 0.42 | 5.32 |
| **Our methods** |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| ITDA (w/o dynamic prompts) | 0.70 | 0.87 | 0.61 | 7.52 | 0.77 | 0.91 | 0.62 | 7.63 | **0.83** | 0.94 | 0.54 | 7.13 | 0.78 | 0.93 | 0.46 | 6.27 |
| ITDA | **0.72** | **0.89** | **0.74** | **8.16** | **0.78** | **0.92** | **0.73** | **8.35** | 0.80 | **0.95** | **0.62** | **7.62** | **0.84** | **0.96** | **0.58** | **6.83** |

Table 1: Overall evaluation. BLEU and BS (BERTScore) measure content preservation, SC measures style transfer strength, and GPT-4 measures overall performance. Values in bold signify the best performance.

| Dataset | Language | #Train data (Original) | #Train data (Augmented) | #Test set |
|---|---|---|---|---|
| Shakespeare | English | 4,000 | 50,000 | 2,000 |
| Trump | English | 4,000 | 30,000 | 2,000 |
| Lyrics | English | 4,000 | 100,000 | 2,000 |
| Lin Daiyu | Chinese | 1,000 | 50,000 | 500 |

Table 2: Dataset statistics. Each sample consists of a (stylized, neutral) pair, where the stylized texts are either collected or augmented, and the neutral texts are generated by ITDA in the training data and annotated by humans in the test set.

datasets, our trained classifiers achieve an average accuracy of 98% on their corresponding test sets, highlighting their reliability and effectiveness.

**Evaluation Metrics.** We follow previous studies to evaluate the quality of model's predictions on three dimensions: content preservation, style transfer strength, and text fluency.

For content preservation, we employ the BLEU (Papineni et al., 2002; Rao and Tetreault, 2018) and BERTScore metrics (Zhang et al., 2019). This assessment involves comparing the similarity between the models' output and the ground-truth stylized text in the test set.

To assess the strength of style transfer, previous studies typically rely on a pre-trained style classifier (Fu et al., 2018; Kashyap et al., 2022; Reif et al., 2022) to make a binary judgment on the style of the model's output. However, unlike conventional stylized texts characterized by distinctive expressive elements, authorship style lacks clearly defined attributes, making it potentially more influenced by text's content. We aim to minimize the impact of content and focus solely on measuring the strength of style change. To achieve this, we introduce a new metric called Style Change (SC). Specifically, We utilize the previously introduced style classifier to compute the probability of belonging to the target style for both the input text and the model's output text, denoted as $s^o$ and $s^i$ respectively. We then calculate their difference $s^o - s^i$ to represent the style change of each sample. Finally, we average $s^o - s^i$ over the samples in test set, thereby evaluating the model's ability to transfer style. A high average Pearson correlation coefficient of 0.89 (Cohen et al., 2009) between human evaluations confirms the reliability of the SC metric, with details provided in Appendix A.5.

For text fluency, some prior studies utilize perplexity (PPL) scores (Logacheva et al., 2022). However, selecting an appropriate language model to compute the PPL score can be challenging because these language models are unlikely to encounter texts with the target authorship style during pretraining, potentially resulting in high PPL scores for stylized texts. Instead, we substitute PPL with the GPT-4 score, which falls within the range [0,1] and evaluates the overall quality of the model's predictions. The prompt for computing the GPT-4 score is provided in Appendix A.7.

Note that BLEU and BERTScore require comparison with the ground-truth stylized texts, whereas SC and GPT-4 scores directly measure the predicted texts without relying on ground truth.

**Baselines.** As outlined in Section 2, we classify the baselines into three main groups: original representation revision, latent representation revision, and few-shot prompting on LLMs. In the first category, we examine **DRG** (Li et al., 2018) and

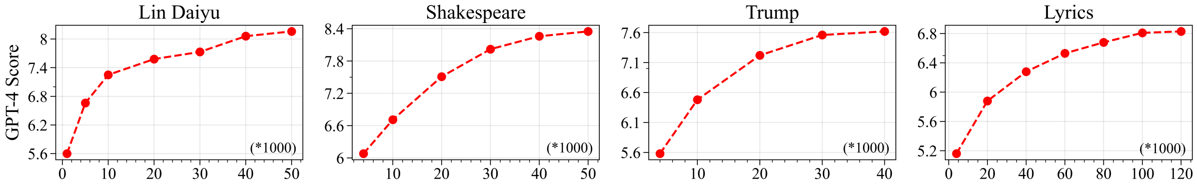

Figure 3: Correlation between the GPT-4's score and the sizes of the datasets used for training BART.

**Transform DRG** (Sudhakar et al., 2019). The second category includes **CTAT** (Wang et al., 2019), **CP-VAE** (Xu et al., 2020), and **TSST** (Xiao et al., 2021). In the third category, we evaluate **Prompt-and-Rerank (GPT-2)** (Suzgun et al., 2022) alongside the direct forward transfer achieved through the few-shot prompting of GPT-3.5, identified as **Few-shot (GPT-3.5)**. While recent studies by Patel et al. (2022) and Reif et al. (2022) employ GPT-3.5 for style transfer, they primarily concentrate on summarizing descriptive adjectives, which may not be suitable for describing unclearly-defined authorship style. More information about baselines is detailed in Appendix A.3. The proposed ITDA employ GPT-3.5 for inverse transfer and train BART-base for forward transfer. Other implementation details are elaborated in Appendix A.2.

### 6.2 Overall Evaluation

In Table 1, ITDA consistently outperforms other methods across all metrics and datasets. Notably, CP-VAE and DRG, relying on language-specific tools, face limitations when applied to Chinese datasets.

Methods that revise latent representations can inadvertently navigate through low-density regions of the language space, risking original content distortion. Original representation revision techniques, focusing on token-level edits like removing stylized words, fall short in authorship styles lacking obvious stylized terms. Both of them highly probably alter the original contents. A very low BLEU score below 0.4 or BERTScore below 0.6 indicates a failure to adequately retain the original content, deeming the method ineffective in those cases.

Then, both Prompt-and-Rerank (GPT-2) and Few-shot (GPT-3.5) approaches utilize few-shot learning on LLMs, achieving better performance. Our ITDA outperforms Few-shot (GPT-3.5) in all aspects, especially the SC scores. This advantage stems from our method generating a high-quality dataset via inverse transfer, thus providing the smaller BART model with a broader array of training samples.

**Human Evaluation.** We enlist the assistance of eight human annotator to assess the predictions across four test sets, evaluating content preservation, fluency, and style transfer strength. These human evaluation results shown in Table 6 in Appendix A.4 closely align with the above automated assessments, showcasing consistency in our method's advanced performance. For a comprehensive breakdown of the setting and results, please refer to Appendix A.4.

### 6.3 Ablation Studies

**Dynamic Prompting.** During inverse transfer, we replace the dynamic prompting strategy with fixed demonstrations and show its performance in Table 1 (refer to "ITDA (w/o dynamic prompts)"). We can find that dynamic prompting outperforms static prompting across almost all datasets and metrics. This advantage arises from dynamic prompting's ability to offer more analogous demonstrations for each input, enhancing LLMs' capacity to better perform inverse transfer.

**Stylized Text Augmentation.** Figure 3 illustrates the relationship between the GPT-4 scores and the data size used for training BART. The results suggest a positive correlation between the GPT-4 score and dataset size. However, the GPT-4 score exhibits a slow increase beyond a certain dataset scale across the four datasets. This plateau is attributed partly to BART-base, a smaller model, quickly reaching its data requirement limit, and partly to the augmented data starting to duplicate the existing dataset due to the capacity limitations of GPT-3.5. Datasets of different authorship types also show varied augmentation needs. For example, the "Trump" dataset, with its everyday language, sees optimal results with about 30,000 augmentations. Meanwhile, "Lin Daiyu" and "Shakespeare" datasets, reflecting classical Chinese and old English, benefit from around 50,000 augmentations. The "Lyrics" dataset, known for its poetic style and significant deviation from neutral text, requires the most data augmentation, around 100,000 instances.

| Style | Input (neutral) | Output of ITDA | Output of few-shot (GPT-3.5) | Output of TSST |
|---|---|---|---|---|
| Shakespeare | I didn't want you to leave me to be murdered. | I did not wish for thee to depart and leave me to be slain. | I would not have you to leave me and get murdered. | I did not you you to leave me to leave me to be beloved. |
| Lyrics | You're such a waste. | Your such a waste. | You're such a waste of time. | You 're such a waste of song. |
| Trump | I experienced some losses, but then I won, and the policy was implemented. | I lost, and then I lost again, but then I won, and we have the policy. | I suffered some losses, but then I prevailed, and the policy was put into effect. | I have some believed but then I campaigned and the went was. |

Table 3: Comparative analysis between our proposed ITDA and two baselines.

| | |
|---|---|
| **Input (neutral)** | The shale pieces look really nice when they're closed up. |
| **Shakespeare** | And those shale pieces, when they're shut up, be marvellous good. |
| **Trump** | Close up, the shale pieces look rather lovely. |
| **Lyrics** | The pieces of shale do show a fair picture when viewed up close. |
| **Input (neutral)** | I can feel a change will happen today. |
| **Shakespeare** | I can sense a transformation shall come to pass this day. |
| **Trump** | I can tell you that's going to change today. |
| **Lyrics** | Now a change is gonna come, I can feel it in the wind today. |
| **Input (neutral)** | I am depressed in my mind. |
| **Shakespeare** | My heart is heavy. |
| **Trump** | I am feeling down in my mind. |
| **Lyrics** | Blues wrapped around my head. |

Table 4: Cases that transforms a neutral text into three distinct styles by ITDA.

## 6.4 Out-of-Distribution Evaluation

Currently, both the training and test sets for each style typically cover similar topics in their contents. However, in real-world scenarios, user-provided neutral text often spans a range of topics. When the topics involved in test data significantly differ from those in training data, achieving high-quality forward transfer becomes challenging. This challenge is referred to as out-of-distribution evaluation.

To evaluate this, we introduce a new test set comprising solely neutral texts spanning diverse topics, resulting in out-of-distribution topics compared to the training data. Specifically, we collect a broad range of neutral texts from literature, finance, education, current politics, and other fields, encompassing a total of 15 topic categories. These out-of-distribution test sets are compiled for both Chinese and English, consisting of 500 samples for Chinese and 2,000 for English. The objective is to transfer them into the four previously tested authorship styles. We primarily compare with the best-performed baseline, Few-shot (GPT-3.5). Since there is no ground truth output (i.e., the corresponding stylized text for each neutral text), we exclude BLEU and BERTScore metrics here. The results are presented in Table 5. Compared with the performance on same-distribution topics in Table 1, both Few-shot (GPT-3.5) and our ITDA 's performance decrease, indicating the challenges of this out-of-distribution test set. Nonetheless, our ITDA still outperforms Few-shot (GPT-3.5) on these out-of-distribution topics, demonstrating its robustness across a spectrum of topics.

| Approach | Lin Daiyu | | Shakespeare | | Trump | | Lyrics | |
|---|---|---|---|---|---|---|---|---|
| | SC | GPT-4 | SC | GPT-4 | SC | GPT-4 | SC | GPT-4 |
| Few-shot (GPT-3.5) | 0.35 | 5.29 | 0.37 | 5.47 | 0.28 | 4.89 | 0.34 | 4.06 |
| ITDA (w/o DP) | 0.52 | 6.34 | 0.59 | 6.72 | 0.40 | 5.24 | 0.49 | 4.87 |
| ITDA | **0.58** | **6.67** | **0.66** | **7.16** | **0.51** | **5.82** | **0.62** | **5.33** |

Table 5: Out-of-distribution evaluation using input neutral texts from various topics. Values in bold signify the best performance. DP represents "dynamic prompts".

## 6.5 Case Studies

Table 3 shows some style transfer results from ITDA, few-shot (GPT-3.5), and traditional TSST method. In the first case, our method accurately preserves content, but both GPT-3.5 and TSST misinterpret the object of "murder". In the second case, GPT-3.5 and TSST introduce new elements like "waste of time" or "waste of song", deviating from original text's meaning. In the last case, ITDA adeptly adjusts sentence structures to fit the desired style, unlike GPT-3.5's superficial changes and limited emulation of complex styles like Trump's. TSST scores lowest in BLEU, indicating problems with repetition, errors, or omissions. Table 4 shows ITDA's ability to transform a single neutral text into various styles, demonstrating its effectiveness in both wording and structural adaptation. More cases are shown in Appendix A.6.

# 7 Conclusion

We propose an inverse transfer data augmentation approach for authorship style transfer, primarily using few-shot prompting with LLMs to revert authorship-stylized texts to neutral texts. These paired corpora are utilized to train a compact model capable of forward transfer, converting neutral texts into the specified authorship style. Experiments show that inverse transfer outperforms forward transfer by GPT-3.5, owing to the prevalence of neutral texts in its pre-training. The resulting compact model shows enhanced performance compared to GPT-3.5, benefiting from a larger volume of exposed training examples of the target style.

## Limitation

When utilizing LLMs for stylized text augmentation, the style of the generated text can be specified, but the content remains uncontrollable. While we aim to encourage LLMs to produce diverse texts by providing various demonstrations as prompts, it is inevitable that some similar texts may be generated, leading to a less efficient use of training resources. Furthermore, if the security of LLMs is inadequate, biased or toxic text may be generated during data augmentation, which could influence the distilled model to a certain degree. In practice, we could leverage the most advanced commercial LLM, such as GPT-4, for such generation, and explore more meticulous data filtering methods designed to ensure the safety, impartiality, and high quality of data synthesized through LLMs.

## Ethical consideration

Regarding **Intellectual Property**, the four authorship-style datasets we use are all publicly accessible. Regarding **Data Annotation**, we invite eight annotators with language backgrounds to label the test sets and for human evaluation. All annotators are briefed on the annotation criteria and are fairly compensated for their efforts. Regarding **Intended Use**, the proposed ITDA is aimed at adding styles to neutral input texts, with the intention of creating personalized digital assistants that communicate in a user's chosen style, aiding students and researchers in understanding different authors' unique writing styles—important for literary studies and education, improving privacy by altering an individual's writing style to conceal their identity, etc. Regarding **Misuse Risks**, there is a potential for misuse through imitation, distortion, plagiarism, and more. For instance, it could be used to generate fake negative reviews or political statements that mimic the styles of various authors. Regarding **Misuse Control** , we make our model checkpoint and code available to the open-source community, allowing users to gain a deeper understanding of our methodology and mitigate the risk of misuse. Our goal is to effectively communicate the potential risks to the public to increase awareness regarding the possible misapplication of this technique and restore its original academic intent.

## References

Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the bible. *Royal Society open science*, 5(10):171920.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. 2010. Bart: Bayesian additive regression trees. *The Annals of Applied Statistics*, 4(1).

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Duy-Tai Dinh, Tsutomu Fujinami, and Van-Nam Huynh. 2019. Estimating the optimal number of clusters in categorical data clustering by silhouette coefficient. In *Knowledge and Systems Sciences: 20th International Symposium, KSS 2019, Da Nang, Vietnam, November 29–December 1, 2019, Proceedings 20*, pages 1–17. Springer.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. 2018. Style transfer in text: Exploration and evaluation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2019. A probabilistic formulation of unsupervised text style transfer. In *International Conference on Learning Representations*.

Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. 2022. Deep learning for text style transfer: A survey. *Computational Linguistics*, 48(1):155–205.

Abhinav Ramesh Kashyap, Devamanyu Hazarika, Min-Yen Kan, Roger Zimmermann, and Soujanya Poria. 2022. So different yet so alike! constrained unsupervised text style transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 416–431.

Kalpesh Krishna, John Wieting, and Mohit Iyyer. 2020. Reformulating unsupervised style transfer as paraphrase generation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 737–762.

Juncen Li, Robin Jia, He He, and Percy Liang. 2018. Delete, retrieve, generate: a simple approach to sentiment and style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1865–1874.

Yuang Li, Yu Wu, Jinyu Li, and Shujie Liu. 2023. Prompting large language models for zero-shot domain adaptation in speech recognition. *arXiv e-prints*, pages arXiv–2306.

Varvara Logacheva, Daryna Dementieva, Sergey Ustyantsev, Daniil Moskovskiy, David Dale, Irina Krotova, Nikita Semenov, and Alexander Panchenko. 2022. Paradetox: Detoxification with parallel data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6804–6818.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Gray, et al. 2022. Training language models to follow instructions with human feedback. In *Advances in Neural Information Processing Systems*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Ajay Patel, Nicholas Andrews, and Chris Callison-Burch. 2022. Low-resource authorship style transfer with in-context learning. *arXiv e-prints*, pages arXiv–2212.

Sudha Rao and Joel Tetreault. 2018. Dear sir or madam, may i introduce the gyafc dataset: Corpus, benchmarks and metrics for formality style transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 129–140.

Machel Reid and Victor Zhong. 2021. Lewis: Levenshtein editing for unsupervised text style transfer. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3932–3944.

Emily Reif, Daphne Ippolito, Ann Yuan, Andy Coenen, Chris Callison-Burch, and Jason Wei. 2022. A recipe for arbitrary text style transfer with large language models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 837–848.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David C Uthus, and Zarana Parekh. 2021. Textsettr: Few-shot text style extraction and tunable targeted restyling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3786–3800.

Yunfan Shao, Linyang Li, Junqi Dai, and Xipeng Qiu. 2023. Character-llm: A trainable agent for role-playing. *arXiv preprint arXiv:2310.10158*.

Akhilesh Sudhakar, Bhargav Upadhyay, and Arjun Maheswaran. 2019. “transforming” delete, retrieve, generate approach for controlled text style transfer. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3269–3279.

Mirac Suzgun, Luke Melas-Kyriazi, and Dan Jurafsky. 2022. Prompt-and-rerank: A method for zero-shot and few-shot arbitrary textual style transfer with small language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 2195–2222.

Bakhtiyar Syed, Gaurav Verma, Balaji Vasan Srinivasan, Anandhavelu Natarajan, and Vasudeva Varma. 2020. Adapting language models for non-parallel author-stylized rewriting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9008–9015.

Ke Wang, Hang Hua, and Xiaojun Wan. 2019. Controllable unsupervised text attribute transfer via editing entangled latent representation. *Advances in Neural Information Processing Systems*, 32.

Zekun Moore Wang, Zhongyuan Peng, Haoran Que, Jiaheng Liu, Wangchunshu Zhou, Yuhan Wu, Hongcheng Guo, Ruitong Gan, Zehao Ni, Man Zhang, et al. 2023. Rolellm: Benchmarking, eliciting, and enhancing role-playing abilities of large language models. *arXiv preprint arXiv:2310.00746*.

Fei Xiao, Liang Pang, Yanyan Lan, Yan Wang, Huawei Shen, and Xueqi Cheng. 2021. Transductive learning for unsupervised text style transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2510–2521.

Peng Xu, Jackie Chi Kit Cheung, and Yanshuai Cao. 2020. On variational learning of controllable representations for text without supervision. In *International Conference on Machine Learning*, pages 10534–10543. PMLR.

Wei Xu, Alan Ritter, William B Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*, pages 2899–2914.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. *arXiv preprint arXiv:1904.09675*.

Zhuosheng Zhang, Aston Zhang, Mu Li, and Alex Smola. 2022. Automatic chain of thought prompting in large language models. In *The Eleventh International Conference on Learning Representations*.

# A Appendix

## A.1 Cluster Number $k$

In clustering-based demonstration annotation, to determine the appropriate value of the cluster count $k$, we employ the silhouette coefficient to measure the effectiveness of clustering. Figure 4 presents the values of the silhouette coefficient for varying cluster count $k$ across four datasets. The results generally indicate a positive correlation between the silhouette coefficient and the cluster count $k$. However, after $k$ reaching a certain scale, the silhouette coefficient no longer exhibits a significant growth for $k$, but rather fluctuates within a certain range. Based on the results presented in Figure 4 and considering a balance between clustering effectiveness and the cost of manual annotation, we set the value of $k$ as 40 for the "Lin Daiyu" dataset and 80 for the other three English datasets.

## A.2 Implementation Details

We employ GPT-3.5 (text-davinci-003) for inverse transfer and train BART-base for forward transfer. The value of $k$ is set as 40 for the "Lin Daiyu" dataset and 80 for other English datasets. These are determined empirically by the silhouette coefficient, which assesses the clustering outcomes. Detailed empirical analyses are available in Appendix A.1. Both static and dynamic few-shot prompting employ a set of eight demonstrations, while stylized data augmentation involves the use of six demonstrations. LLMs baselines use the same eight demonstrations as the proposed ITDA(Static). For each test set, we execute the distilled BART-base model multiple times to obtain averaged evaluation results.

English compact model initializes from Bert-base-cased[5], and Chinese compact model initializes from Bart-base-chinese[6]. The hyperparameters we use for fine-tuning BART-base are as follows. We fine-tune the model for 12 epochs using AdamW optimization. We gradually increase the learning rate from zero to 4e-5 over 5% of the total training steps, followed by a cosine decay to zero towards the end. The batch size is fixed at 64, and the maximum length of the context window is set to 512 tokens. Training completes in approximately five hours utilizing an NVIDIA RTX A6000 48G GPU.

## A.3 Baselines

We compare our method with three types of baselines: latent representation revision, original representation revision, and few-shot prompting based on language models. The first approach alters the latent representation of the original input to conform it to the given style. The second type follows a "delete-generate" framework that initially removes the stylized words in the original text and then incorporates the specific style through generation. The third type leverages the robust in-context learning ability of LLMs, utilizing few-shot prompting specifically for style transfer. Below, we elaborate on the details of these specific baselines. Importantly, none of the baselines rely on the annotated parallel data that translates from neutral text to stylized text.

- **Delete, Retrieve, Generate (DRG) (Li et al., 2018)** is categorized under the first type. It operates by deleting the style words using a predefined dictionary, which contains words that occur much more frequently within $D^S$ than in other arbitrary neutral texts. The method then generates the target stylized text based on the remaining content words and auxiliary information. We evaluate two variants of this method. The first, known as Delete-only, removes the style words. The second, Detete-and-Retrieve, also identifies similar sentences of the desired target style, extracting stylized words from them to serve as the auxiliary information. The generation process in both cases is handled through an RNN model.
- **Transforming Detete, Retreve, Generate (Transform DRG) (Sudhakar et al., 2019)** falls into the first style category. This method adheres to the delete-retrieve-generate framework but introduces a transform-based classifier for style work removal. Additionally, it replaces the traditional generation model with the GPT model.
- **Controllable Text Attribute Transfer**

[5] https://huggingface.co/bert-base-cased

[6] https://huggingface.co/fnlp/bart-base-chinese

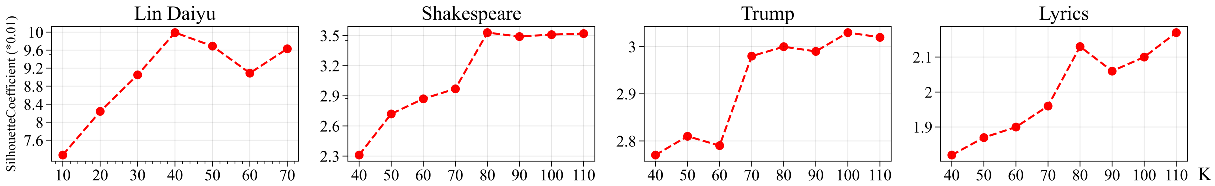

Figure 4: Correlation between the number of clusters $k$ and the Silhouette Coefficient.

| Approach | Lin Daiyu |  |  | Shakespeare |  |  | Trump |  |  | Lyrics |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | **Con** | **Flu** | **Style** | **Con** | **Flu** | **Style** | **Con** | **Flu** | **Style** | **Con** | **Flu** | **Style** |
| **Original Representation Revision** |  |  |  |  |  |  |  |  |  |  |  |  |
| DRG (Delete-Only) | - | - | - | 1.2 | 1.2 | 2.7 | 1.8 | 3.7 | 2.3 | 2.4 | 3.1 | 2.2 |
| DRG (Delete-and-Retrieve) | - | - | - | 2.6 | 1.5 | 2.6 | 2.5 | 1.2 | 2.1 | 3.5 | 2.8 | 1.9 |
| Transform DRG (Delete Only) | 2.6 | 3.4 | 1.7 | 3.8 | 3.7 | 2.1 | 2.2 | 4.0 | 1.6 | 4.1 | 3.7 | 2.1 |
| **Latent Representation Revision** |  |  |  |  |  |  |  |  |  |  |  |  |
| CTAT | 2.3 | 3.2 | 2.1 | 2.7 | 3.5 | 2.5 | 3.1 | 3.3 | 2.8 | 2.9 | 3.2 | 1.8 |
| CP-VAE | - | - | - | 2.4 | 3.3 | 2.3 | 1.9 | 3.7 | 2.9 | 2.6 | 3.1 | 2.1 |
| TSST | 2.5 | 3.1 | 2.5 | 3.2 | 2.9 | 3.3 | 3.4 | 2.8 | 3.1 | 3.9 | 3.4 | 2.5 |
| **Few-shot Prompting on LLMs** |  |  |  |  |  |  |  |  |  |  |  |  |
| Prompt-and-Rerank (GPT-2) | 1.5 | 3.3 | 2.4 | 4.0 | 4.1 | 2.0 | 2.6 | 4.2 | 2.6 | 3.8 | 4.1 | 2.3 |
| Few-shot (GPT-3.5) | 3.9 | 4.3 | 3.1 | 3.9 | **4.2** | 3.3 | 4.2 | 4.3 | 3.0 | 4.2 | 4.4 | 2.6 |
| **Our methods** |  |  |  |  |  |  |  |  |  |  |  |  |
| ITDA (w/o dynamic prompts) | 4.2 | 4.3 | 3.5 | 4.0 | 4.1 | 3.9 | **4.6** | 4.1 | 3.3 | 4.3 | 4.2 | 2.8 |
| ITDA | **4.6** | **4.4** | **4.0** | **4.2** | **4.2** | **4.5** | 4.5 | **4.4** | **3.8** | **4.6** | **4.3** | **3.4** |

Table 6: Human evaluation across four datasets. Values in bold signify the best performance.

**(CTAT)** (**Wang et al., 2019**) is categorized under the second type. It employs a transformer-based autoencoder to learn the representation of an input text. After that, a style classifier is trained, and the latent representation is subsequently modified through the iterative gradient back-propagation of attribute classification loss, continuing until the latent representation can be classified as possessing the desired target style.

- **Constrained Posterior VAE (CP-VAE)** (**Xu et al., 2020**) falls into the second category, focusing on learning the representation of text using VAE. To address the latent vacancy problem in text, CP-VAE restricts the posterior mean to a learned probability simplex and subsequently manipulates this simplex.

- **Transductive Style Transfer (TSST)** (**Xiao et al., 2021**) is classified under the second type. It identifies the most similar stylized text to the given input text and represents them together, aiding in the transfer of the input text' style. By employing adversarial style loss, the representation is guided to approximate the target style.

- **Prompt-and-Rerank (GPT-2)** (**Suzgun et al., 2022**) represents the the third type. It employs few-shot prompting on GPT-2 to generate multiple diverse outputs for each input. The method then re-ranks the outputs, taking into account a combination of factors such as the textural similarity between input and output, the strength of the output style, and the fluency of the output.

- **Few-shot (GPT-3.5)** constitutes the third type. In this method, we use eight handcrafted examples the same as ITDA (static) as the few-shot prompts for GTP-3.5. The prompt is shown in Table 15.

### A.4 Human Evaluation

We invite eight volunteers with strong language proficiency to assess the model's style transfer effectiveness across the four datasets. These volunteers have diverse educational backgrounds and span various age groups. We randomly sample 500 neutral texts from each test set. Then for each corresponding predicted stylized text, we hide the method of its generation and ask volunteers to rate it on a scale of 1 to 5 for content preservation (Con), fluency (Flu), and style transfer strength (Style). A higher score indicates a greater agreement with this aspect. The average scores given by the volunteers were taken as the final results and presented in Table 6.

The results of human evaluation generally coincide with the automated assessment metrics. Traditional transfer methods exhibit more issues in terms of content preservation and grammatical correctness in human evaluation. Those traditional methods with relatively low BLEU scores or BERTScore scores sometimes exhibit a phenomenon of piling up style-related words without adhering to grammar rules. Our method demonstrates high quality in three aspects, particularly excelling in content preservation and style transfer strength surpassing all other methods.

### A.5 Pearson Coefficient

To validate the reliability of the proposed style change (SC) metric, we conduct a meta-evaluation. Specifically, we calculate the Pearson correlation coefficient between the SC scores shown in Table 1 and the Style scores shown in Table 6 across all models on different test sets. For the test sets Lin Daiyu, Shakespeare, Trump, and Lyrics, their respective Pearson correlation coefficients are 0.96, 0.89, 0.91, and 0.83, with an average of 0.89. The high correlation coefficient confirms the reliability of the SC metric.

### A.6 Additional Case Studies

We select several relatively well-performing traditional methods and showcase their transfer examples on different datasets. Specific examples can be found in Table 7, Table 8 and Table 9. It is evident that traditional methods exhibit issues such as missing content, addition of irrelevant content, and various grammar errors when transferring authorship styles.

### A.7 The Employed Prompts

#### Dynamic Prompts for Inverse Transfer

We present the dynamically selected demonstrations of (stylized, neutral) pairs during inverse transfer on different datasets, as illustrated in Table 10, Table 11, Table 12 and Table 13.

#### Prompts for Few-shot (GPT-3.5)

We present the prompts used for direct forward transfer with GPT-3.5 for each dataset, as shown in Table 14 and Table 15.

#### Prompts for Evaluation with GPT-4

We present the evaluation prompt templates for forward transfer in Table 16 and inverse transfer in Table 17. When evaluating, we provide GPT-4 with eight demonstrations of conversion between author-stylized text and neutral text, which are the same as those used in forward transfer and inverse transfer with GPT-3.5.

| Input | Output of TSST | Output of CTAT |
|---|---|---|
| 会议马上就开始了。 | 紫鹃,就早了. | 嗳哟,我们都有那么高的。今年第二二次第二次第二次世界大战设备? |
| 我就是不喜欢喝药。 | 我就是不在家大作. | 我不一样。这27日成功,到底是由他一同西安出来！ |
| 就是呢,你过去吧,不需要过来了。 | 就是呢,你在内罢,不子过来了。 | 就是呢,你不需要过去吧！配,只是方才宝玉多的地方怎么看 |
| 那么为什么伤心起来呢。 | 那么你你起来呢。 | 什么为什么那么多人来过戛纳,马云发现爱情的主力不出来吗? |
| 你现在跑过来干什么。 | 你我去过了。 | 你现在过来！你家女孩女孩子才有什么系列7小时，拒绝速度? |
| 我熬不过这一天了，仅仅等你的消息就很烦。 | 我怕这一天了,我等你的,就说话。 | 我这不过一天，就有了参与的去处好去处，你老师号呢? |
| 我没有别人有趣，你心里没有我。 | 我没有别人呢,你心里没有我。 | 我有心里不好，没有梦想过亿，你也从金刚金刚经? |
| 你如果觉得我无聊，明天又觉得别人有趣，真伤我心。 | 你你这我杨妃,还又别人别人接客,真只管接客。 | 你觉得我又一项，如果可以有效机器机器人，但是我支持的那么开心。 |
| 这样说反而是我不对了，好像我在斤斤计较，不如别人。 | 这样说,是我不杨妃,为我在去,不如别人 | 我是这样走在心上，对了不需要与男神，原来堵在自家了。 |
| 平时我和你说的，都当耳旁风了，别人说的你就听话。 | 今儿我和你说的,都浮躁了,别人说的你就奚落. | 平时都说我的家，你和别人的五种这五种方法，你也被嘲了！ |

Table 7: Examples of style transfer outputs using TSST and CTAT on the “Lin Daiyu” dataset.

| Style | Input | Output of DRG (Delete-and-Retrieve) |
|---|---|---|
| Shakespeare | You’re gonna put your face on the face! | I put your pleasure , then , sir . |
| | Unfortunately, I’m sorry, but. | Ay , sir , the of yours . |
| | The confession is riddled with riddles, but the confession is not. | The service is furnished with ’t ; but the art is decreed . |
| | Madam, I’m here. | Why , then , the is a whore , And let us the only . |
| | Good-bye, my dear. | My lord , my lord ! , my ! |
| | Doctor, come back, I’ll remember you. | No , I’ll come , and I’ll forget come . |
| | Without his roe, like a dried fish. | Without his , like a wings . |
| | He’s not telling the truth. | He’s not , sir , the truth. |
| | I’ll hide here, too. | No , then , I’ll be a little ere Come , and I’ll bring them . |
| | 0 Not life, but love in death. | Not so , in thee , I will not lose thee in ’t . |

Table 8: Examples of style transfer outputs using DRG (Delete-and-Retrieve) on the “Shakespeare” dataset.

| Style | Input | Output of Prompt-and-Rerank |
|---|---|---|
| Trump | No, i did not vote for her. | I did not have faith in her, i would have loved her more. |
| | Consider, please, what they do. | That is very serious. I am sorry to have to remind you. |
| | Good morning, i awoke feeling quite pleasant. | Greetings. |
| | Chfis, where are you? We needs you. | There will be no shortage of trouble if you come. Stay where you are. |
| | That was an excellent result, mike. | That was an unfortunate result, to be sure, but not surprising either. |
| | I complete various tests and return home. | You will find that I do return home. |
| | They no longer show disrespect towards us, dear followers. | They do not have a show mind. |
| | Instead, we have the beast. | That’s a great idea, and I’ll talk to you soon. |
| | They just keep it here, right here. | In your book, you were a bit of a hard-on. |
| | After persistent efforts for four years, hhey ultimately gave up. | And it was just for a few hours. |

Table 9: Examples of style transfer outputs using Prompt-and-Rerank on the “Trump” dataset.

| **Lin Daiyu** |
|---|
| **Stylized Input:**你也不用再这样唠叨了，要恼就直接说出来，何必再绕弯子呢。<br>**Dynamic Prompts:**<br>原句：我也好了许多,谢你一天来几次瞧我,下雨还来.这会子夜深了,你且请回去,明儿再来.<br>目标句：我也好了不少，感谢你每天看我几次，下雨也过来。现在夜深了，你先回去，明天再来。<br>原句：果然如此, 我们的虽不好, 亦可以带好了.<br>目标句：真的这样，我们的虽然不好，也可以处理好了。<br>原句：你不理我,我还在那里做什么!<br>目标句：既然你不理我，我还在那里干嘛。<br>原句：此时想也快天亮了,到底要歇息歇息才是.<br>目标句：这时天也快亮了，还是要歇息歇息的。<br>原句：正是呢,不必过来了.<br>目标句：就是呢，不需要过来了。<br>...<br>请按照示例将句子重写为中立文本。<br>**Neutral Output:** 你不用再唠叨了，要有意见就直接说出来，干嘛再绕弯子呢。 |
| **Stylized Input:**若是你不肯信，怕是要把我的心碎了。<br>**Dynamic Prompts:**<br>原句：你大抵是倦了，竟回我这般敷行.<br>目标句：你倦了，回我敷行。<br>原句：何苦来,又招我!<br>目标句：又招惹我。<br>原句：你不理我,我还在那里做什么!<br>目标句：既然你不理我，我还在那里干嘛。<br>原句：我原以为我自是跟旁的不一样,想来是我自作主张了。<br>目标句：我原来认为自己和别人不一样，看来我是自作主张了。<br>原句：果然如此, 我们的虽不好, 亦可以带好了.<br>目标句：真的这样，我们的虽然不好，也可以处理好了。<br>...<br>请按照示例将句子重写为中立文本。<br>**Neutral Output:** 你不用再唠叨了，要有意见就直接说出来，干嘛再绕弯子呢。 |

Table 10: Dynamic prompts used for inverse transfer on the “Lin Daiyu” dataset.

| **Shakespeare** |
|---|
| **Stylized Input:**Thyself shalt see the act.<br>**Dynamic Prompts:**<br>Input:Fair youth , I would I could make thee believe I love .<br>Output:Young boy , I wish I could make you believe that I'm in love .<br>Input:If thou pleasest not , I yield thee up my life .<br>Output:If not , you can kill me .<br>Input:And I do believe your Majesty takes no scorn to wear the leek upon Saint Tavy's day .<br>Output:I do believe your Majesty takes no shame in wearing the leek on Saint Davy's Day .<br>Input:Tis well for thee That , being unseminared , thy freer thoughts May not fly forth of Egypt .<br>Output:It's a good thing for you that , being castrated , you can better concentrate on my needs .<br>Input:Make your vaunting true , And it shall please me well .<br>Output:Make your boasts come true , and I'll be thrilled .<br>...<br>Please rewrite the sentence as neutral text according to the examples.<br>**Neutral Output:** You will witness the act. |
| **Stylized Input:**The Queen shall then have courtesy , so she Will yield us up ?<br>**Dynamic Prompts:**<br>Input:For the best turn i' th' bed .<br>Output:For the favor of sleeping in the bed .<br>Input:And I do believe your Majesty takes no scorn to wear the leek upon Saint Tavy's day .<br>Output:I do believe your Majesty takes no shame in wearing the leek on Saint Davy's Day .<br>Input:I'll seal to such a bond , And say there is much kindness in the Jew .<br>Output:I'll agree to those terms and even say that Jews are nice .<br>Input:Would you praise Caesar , say "Caesar." Go no further .<br>Output:Oh , you If you want to praise Caesar , just say his name , that's all the praise that's necessary .<br>Input:Nor must not then be yielded to in this .<br>Output:Then we won't agree to his demands .<br>...<br>Please rewrite the sentence as neutral text according to the examples.<br>**Neutral Output:** Will the Queen then show us courtesy and surrender? |

Table 11: Dynamic prompts Used for inverse transfer on the “Shakespeare” dataset.

| **Trump** |
|---|
| **Stylized Input:**I have middle of the road, I have poor, I have everybody.<br>**Dynamic Prompts:**<br>Input:Look, 300% in certain very bad crimes, New York.<br>Output:300% of some very serious crimes come from new york.<br>Input:Build a wall, build a wall, true.<br>Output:Build a wall.<br>Input:I don't know how many people here, but there's a lot.<br>Output:There are a lot of people.<br>Input:Everyone makes mistakes, but it's what you do with them and what you learn from them that matters.' Midas Touch.<br>Output:Everyone makes mistakes, but what matters is how you treat them and what you learn from them.<br>Input:Your congressmen, all of your Congresspeople, men, wonderful people, they're at a place called Congress right now.<br>Output:Your congressman is now in a place called Congress.<br>...<br>Please rewrite the sentence as neutral text according to the examples.<br>**Neutral Output:** I have people from all walks of life. |
| **Stylized Input:**I did that heavy, heavy Pocahontas deal.<br>**Dynamic Prompts:**<br>Input:This guy did the swine flu, right, it was a catastrophe.<br>Output:This guy has swine flu, which is a disaster.<br>Input:Give you your tax cuts, I gave them to you.<br>Output:I have given you tax cuts.<br>Input:Hunter walked out of the plane, had a quick meeting, walked away with one and a half billion dollars.<br>Output:Hunter spent \$1.5 billion on a quick meeting by plane.<br>Input:I have to say this very, very unfair to my family.<br>Output:I find it unfair to my family.<br>Input:I kept my promise, recognized the true capital of Israel and opened the American Embassy in Jerusalem.<br>Output:I recognized the real capital of Israel and opened the American Embassy in Jerusalem.<br>...<br>Please rewrite the sentence as neutral text according to the examples.<br>**Neutral Output:** I handled the difficult Pocahontas situation. |

Table 12: Dynamic prompts used for inverse transfer on the "Trump" dataset.

| **Lyrics** |
|---|
| **Stylized Input:** Hate it or love it, the underdog's on top.<br>**Dynamic Prompts:**<br>Input:My heart is all in tatters, I ain't nobody's saint.<br>Output:I'm all torn up, and I'm not a saint.<br>Input: Blues wrapped around my head.<br>Output: I am depressed in my mind.<br>Input: Love is a mine of gold.<br>Output:Love is very precious.<br>Input:But the last wall standing's fell, daddy kicked it down.<br>Output:But the last wall fell, and Dad kicked it down.<br>Input: No part of this road feels wrong.<br>Output: This road feels all right.<br>...<br>Please rewrite the sentence as neutral text according to the examples.<br>**Neutral Output:** The underdog is in a position of power. |
| **Stylized Input:** Looking back on when we first met.<br>**Dynamic Prompts:**<br>Input: Never look back, walk tall, act fine.<br>Output: Keep your chest up to walk forward and don't look back.<br>Input: I get him hot and bothered.<br>Output: I make him irritable.<br>Input:You my babe, I got my eyes on you.<br>Output: You are my baby and I would always pay attention on you.<br>Input:Everything I ever had to lose.<br>Output:Everything I've ever lost.<br>Input: When you run back to your wife?<br>Output: It's time for you to find your wife.<br>...<br>Please rewrite the sentence as neutral text according to the examples.<br>**Neutral Output:** Remembering when we first met. |

Table 13: Dynamic prompts used for inverse transfer on the "Lyrics" dataset.

| **Lin Daiyu** |
|---|
| **Fixed Prompts of Lin Daiyu:** |
| **Dynamic Prompts:** |
| 原句：你倦了，回我敷行。 |
| 目标句：你大抵是倦了，竟回我这般敷行 |
| 原句：没有别的妹妹有趣，哥哥心里没有我。 |
| 目标句：没有别的妹妹有趣，终究哥哥心里没有我 |
| 原句：疼爱的只有你母亲，今天见到了你，我怎么能不伤心！ |
| 目标句：所疼者独有你母,今见了你,我怎不伤心! |
| 原句：经常服用的是什么药,为什么不赶紧做疗治? |
| 目标句：常服何药,如何不急为疗治? |
| 原句：这些人每个都像这样的恭肃严整，来的人是谁，放诞无礼到这样的地步? |
| 目标句：这些人个个恭肃严整如此,这来者系谁,这样放诞无礼? |
| 请将句子重写，将目标作者风格提炼为中性输入，作为上述示范的成功转化。 |
| 原句：也还算便宜。 |
| 目标句：倒也便宜. |
| 请根据以上成功转化的示例，重写句子，将目标作者风格注入中性输入文本。 |

Table 14: Fixed prompts used for forward transfer with GPT-3.5 on the “Lin Daiyu” dataset.

**Fixed Prompts of Shakespeare:**

Input:I have half a mind to hit you before you speak again.
Output:I have a mind to strike thee ere thou speak'st.
Input:And he's friendly with Caesar.
Output:And friends with Caesar.
Input:I'm going to make you a rich man.
Output:Make thee a fortune from me.
Input:No , I didn't say that.
Output:I made no such report.
Input:What did you say to me?
Output:What say you?
Input:You say he's friendly with Caesar , healthy , and free.
Output:He's friends with Caesar , In state of health , thou say'st , and , thou say'st , free.
Please rewrite the sentence to inject target authorship style into the neutral input according to the successful transformation of above demonstrations.

**Fixed Prompts of Trump:**

Input:I find it unfair to my family.
Output:I have to say this very, very unfair to my family.
Input:We can't let it happen.
Output:Right? Can't let it happen, folks.
Input:They are just a form.
Output:Look it, they just form.
Input:We love our nation that is still great today.
Output:We love our nation, our nation is great today.
Input:We killed the number one terrorist.
Output:He was vehemently‘Ă‘ę We killed this number one, terrorist.
Input:I have to prove that they are liars.
Output:I had to because I had to show they're liars.
Please rewrite the sentence to inject target authorship style into the neutral input according to the successful transformation of above demonstrations.

**Fixed Prompts of Lyrics:**

Input:You know our relationship.
Output:Yeah, yeah, you know how me and you do.
Input:I have your arms open.
Output:Your arms are open for me.
Input:It's at least until tomorrow.
Output:So far at least until tomorrow.
Input:Everything I've ever lost.
Output:Everything I ever had to lose.
Input:I'm sure he'll kill him.
Output:And I promise its going to kill.
Input:People are on the street.
Output:And people on the streets.
Please rewrite the sentence to inject target authorship style into the neutral input according to the successful transformation of above demonstrations.

Table 15: Fixed prompts used for forward transfer with GPT-3.5 on English datasets.

**Instruction:**
Please act as an impartial judge and evaluate the model's ability to infuse target authorship style into neutral text. You will be provided with some demonstrations of successful migration to the target authorship style.
You should make a comprehensive assessment and consider factors such as the style transfer strength, content preservation and fluency of the response.
You must first provide your explanation, then rate the response on a scale of 1 to 10.
**[The Start of Demonstrations]**
{Neutral Text:... Corresponding Author-stylized Text:...}
{Neutral Text:... Corresponding Author-stylized Text:...}
......
**[The End of Demonstrations]**
**Original Neutral Text:**......
**Transferred Text:**.....

Table 16: Prompt template for evaluating forward transfer quality with GPT-4.

**Instruction:**
Please act as an impartial judge and evaluate the model's ability to remove target author-stylized features from original stylized text to generate neutral text.
You will be provided with some demonstrations of successful removal
of the target authorship style.
You should make a comprehensive assessment and consider factors such as
the style transfer strength, content preservation and fluency.
You must first provide your explanation, then rate the response on a scale of 1 to 10.
**[The Start of Demonstrations]**
{Authorship-stylized Text:... Corresponding Neutral Text:...}
{Authorship-stylized Text:... Corresponding Neutral Text:...}
......
**[The End of Demonstrations]**
**Original Stylized Text:**......
**Generated Neutral Text:**.....

Table 17: Prompt template for evaluating inverse transfer quality with GPT-4.