

# RT-SFT: Text Style Transfer from Non-Parallel Corpora by Roundtrip Translation

**Ruoxi Liu**

Department of Computer Science  
Johns Hopkins University  
rliu79@jh.edu

**Philipp Koehn**

Department of Computer Science  
Johns Hopkins University  
phi@jhu.edu

## Abstract

Text style transfer (TST) is naturally a supervised task — rewrite a sentence in a target style while preserving its meaning — yet the parallel corpora that supervision requires exist for only a handful of style domains. A common workaround is to *normalize* an input into a style-agnostic intermediate and then *stylize* it into the target style, but the normalizer is typically a lightweight, task-specific paraphraser applied only at test time, feeding a correspondingly small stylizer. We observe that a style-stripping normalizer already exists at scale: neural MT systems trained on hundreds of millions of general-domain sentence pairs preserve content while regressing toward generic phrasing, so roundtrip translation through a pivot language strips stylistic signal without any task-specific training. This turns normalization from an inference-time patch into a data-generation tool. Roundtrip-translating a monolingual in-style corpus yields a pseudo-parallel corpus on which we LoRA-finetune an instruction-tuned LLM as the stylizer (RT-SFT); applying the same normalizer to test queries keeps that stylizer in-distribution. We show that across four style domains, RT-SFT outperforms state-of-the-art methods, such as few-shot in-context learning, by considerable margins. We also report on effective retrieval augmentation methods for expert style domains with strict terminology and naming conventions.

## 1 Introduction

Text Style Transfer (TST) is the task of rephrasing text by modifying stylistic attributes while preserving its core, attribute-independent semantics and intent (Shen et al., 2017; Toshevskaa and Gievska, 2024). These attributes encompass formality, attitude, verbosity, preferred terminology, and other characteristics inherent to the text. A central obstacle in TST is the scarcity of annotated parallel corpora, which prevents the use of fully supervised

training or finetuning (Pan et al., 2024) in most style domains.

One established approach to this problem is *normalize-then-stylize*: a normalizer maps an input to a style-agnostic form, either a latent encoding or text, and a stylizer regenerates that form in the target style. STRAP (Krishna et al., 2020), for instance, normalizes by paraphrasing the input to strip its style before decoding with a style-specific generator. The advantage of this framing is that the normalizer defines a single shared domain through which every input passes, whatever style it arrives in, so the stylizer never has to accommodate unpredictable source styles.

Robustness to diverse, unpredictable input styles therefore depends on the training data for the normalizer. Most current normalizers are trained on curated *paraphrase corpora* such as ParaFusion (Jayawardena and Yapa, 2024). Such corpora are reported to offer limited lexical and syntactic diversity (Hu et al., 2019; Krishna et al., 2020; Jayawardena and Yapa, 2024), and paraphrasers trained on them degrade on stylistic features unseen during training. A second line of work prompts an instruction-tuned LLM to paraphrase without any parameter update (Roy et al., 2023; Sun et al., 2025). We argue that this is less principled: the stylizer unavoidably carries the stylistic features and biases of the prompted LLM.

We propose roundtrip translation as the normalizer: a sentence is translated into a pivot language and then back into the original language, a technique previously used to assess MT robustness and generation quality (Somers, 2005; Moon et al., 2020). This requires neither paraphrase supervision nor a prompted LLM (Figure 1). We first train two Marian NMT models (Junczys-Dowmunt et al., 2018) on general-domain parallel corpora (Kocmi et al., 2025), and compose them into an English–pivot–English roundtrip pipeline. Applying this pipeline to a non-parallel corpus in the

target style produces an RT-normalized counterpart for each sentence, and the two sides form a pseudo-parallel corpus that restores a supervised TST setting in a domain with no bitext. We then finetune an instruction-tuned LLM on this corpus as the stylizer (an RT-SFT-LLM), with retrieval augmentation applied to both the finetuning and the inference prompts (§3.3). At test time we pass the query through the same roundtrip pipeline before generation, so that the stylizer receives the RT-normalized input it was trained on (§3.1). We show in §4.3 that this alignment raises style accuracy in all four domains.

We evaluate our method (§4) on four style domains with distinctive stylistic features (§4.1.1), against two established baselines: few-shot In-Context Learning (ICL) (Reif et al., 2022; Suzgun et al., 2022) and Automatic Post-Editing (APE) (Simard et al., 2007; Moon et al., 2022). Following prior work, we measure transfer quality with BERT-based style classifiers trained on held-out data and content preservation with BLEU (Subramanian et al., 2018; Wan et al., 2023; Aycock and Bawden, 2024).

Our **main contributions** are:

- **Pseudo-parallel dataset construction** (§3.1). We use roundtrip translation through general-domain MT as a style normalizer, synthesizing the parallel corpus that supervised finetuning requires in domains that lack bitext.
- **Retrieval augmentation for finetuning and inference** (§3.3). We apply retrieval augmentation to the **finetuning** prompts as well as the inference prompts, and show that its benefits extend beyond prompting by aligning the retriever with the generator.
- **Systematic evaluation of RT-SFT-LLMs** (§4). We benchmark RT-SFT-LLMs across four style domains using four LLMs as base model, three prompt templates, and several retrieval and inference configurations, against established baselines.

## 2 Related Work

**Unsupervised and semi-supervised TST** Because corpora that align the same content across styles are scarce, most work performs style transfer from non-parallel data. Early neural approaches disentangle content from style in a latent space

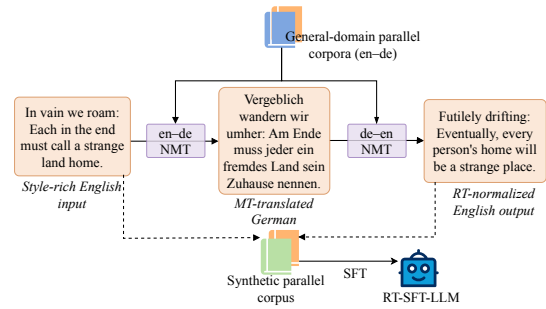


Figure 1: Our proposed workflow for finetuning large language models (LLMs) for text style transfer (TST) using only non-parallel data in the target domain. A bilingual general-domain parallel dataset is used to train a pair of neural machine translation (NMT) models capable of translating between English and a pivot language. We then obtain RT-normalized texts of the original in-domain texts by roundtrip translating the in-domain set with the NMT models. This enables supervised finetuning of LLMs for TST, where we finetune LLMs for RT-normalized-domain to target-domain transfer using the synthetic parallel corpus.

and recombine the content with a target-style representation (Shen et al., 2017; Nangi et al., 2021; Voigt et al., 2018), typically using adversarial objectives and auxiliary losses that preserve content through the transfer (Tian et al., 2018; Nouri, 2022). Semi-supervised methods further reduce the reliance on task-specific data by exploiting generic pretrained resources and reward models (Lai et al., 2021). More recent work abandons explicit disentanglement, instead synthesizing pseudo-parallel data with LLMs and refining generation with reinforcement learning: Pan et al. (2024) couple LLM prompting with attention masking, while intensity-controllable methods first supervise-finetune a base transfer model and then apply reinforcement learning with hierarchical rewards to modulate stylistic strength (Gu et al., 2026). Our method also assumes only non-parallel, target-style data, but recovers a supervised training signal through normalization (§3).

**Normalize-then-stylize** A strategy for non-parallel TST is to first map the input to a style-agnostic intermediate—to *normalize* it—and then generate in the target style—to *stylize* it. The normalizer has taken many forms: a learned diverse paraphraser (Krishna et al., 2020), roundtrip or back-translation through a pivot language whose output is assumed to retain content but not style (Prabhumoye et al., 2018; Rabinovich et al., 2017;

Zhang et al., 2018; He et al., 2020), denoising and latent-variable encoders (Subramanian et al., 2018; Riley et al., 2021), abstract-meaning-representation graphs (Jangra et al., 2022), and, most recently, diffusion paraphraser (Horvitz et al., 2024a). In these methods the normalizer is typically applied *at inference time* and paired with a comparatively small stylizer: STRAP (Krishna et al., 2020) paraphrases the input at test time to strip its style and then decodes with a style-specific GPT-2; Horvitz et al. (2024b) reconstruct text from a neutralized paraphrase conditioned on an authorship embedding; conversational methods reduce an utterance to a style-free form and rewrite it (Roy et al., 2023); and synthesize-then-decode prompting amplifies the target style only at decoding (Sun et al., 2025). We adopt this paradigm but realize it with two components trained at scale—a general-domain neural machine translation (NMT) model as the normalizer (through roundtrip translation) and an instruction-tuned LLM, finetuned to map normalized text to the target style, as the stylizer—together with retrieval augmentation at both training and inference (§3). This trained, LLM-scale realization is precisely what prior, largely inference-time methods lack.

**Prompt engineering and in-context learning** A parallel line of work performs TST with LLMs purely through prompting, without updating any parameters. Reif et al. (2022) frame transfer as instruction-guided rewriting with augmented zero-shot prompts; other methods rerank sampled candidates by style strength and fluency (Suzgun et al., 2022), supply retrieved in-domain exemplars as few-shot “shots” (Liu et al., 2024a; Chen, 2024; Bhandarkar et al., 2024; Aycock and Bawden, 2024), or optimize the prompt template and its retrieved context (Zhang et al., 2024; Wan et al., 2023; Tao et al., 2024). The most recent methods steer generation through register analysis (Yang and Carpuat, 2025), structure long-document transfer (Wu and Deng, 2025), and probe how far prompting alone can go (Toshevskaa and Gievska, 2025; Mukherjee et al., 2024). Such approaches require no training, but they are inherently prompt-bound: they cannot adapt model parameters to a target domain, which limits terminology control and long-range stylistic consistency. We instead finetune the stylizer (§3).

**Supervised TST and parameter-efficient finetuning** When aligned data are available, style

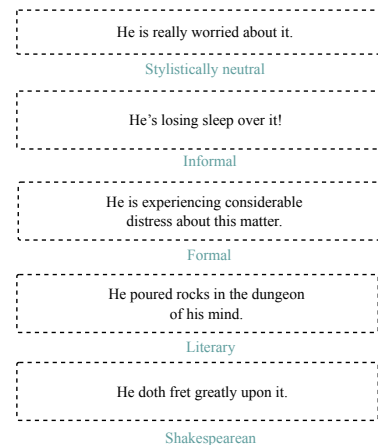


Figure 2: Example sentences illustrating semantically equivalent content in various styles. Outputs of our roundtrip translation pipeline are considered RT-normalized.

transfer can be cast as supervised sequence-to-sequence learning, as in Jhamtani et al. (2017)’s copy-augmented model for Shakespearean English. The bottleneck is data: parallel style corpora exist for only a handful of domains—modern-to-Shakespearean English (Xu et al., 2012), formality (Rao and Tetreault, 2018; Briakou et al., 2021), biblical prose (Carlson et al., 2018), expert-versus-layman medical text (Cao et al., 2020), and, more recently, parallel detoxification corpora (Dementieva et al., 2025; Protasov et al., 2025)—and new benchmarks continue to target these same well-resourced settings (Xie et al., 2025; Jangra et al., 2025). Parameter-efficient finetuning of LLMs for TST has begun to appear (Liu et al., 2024b; Mukherjee et al., 2024) but remains confined to domains with existing parallel data. We remove this constraint: by normalizing a monolingual target-style corpus, we synthesize the parallel data that supervised finetuning requires and extend it to domains with no bixtext (§3).

## 3 Methods

### 3.1 Roundtrip Translation

Unlike most state-of-the-art methods that instantiate the *normalize-then-stylize* paradigm (Krishna et al., 2020), we train both components at scale. The normalizer is a pair of neural machine translation (NMT) models, trained with Marian (Junczys-Dowmunt et al., 2018) on a large general-domain

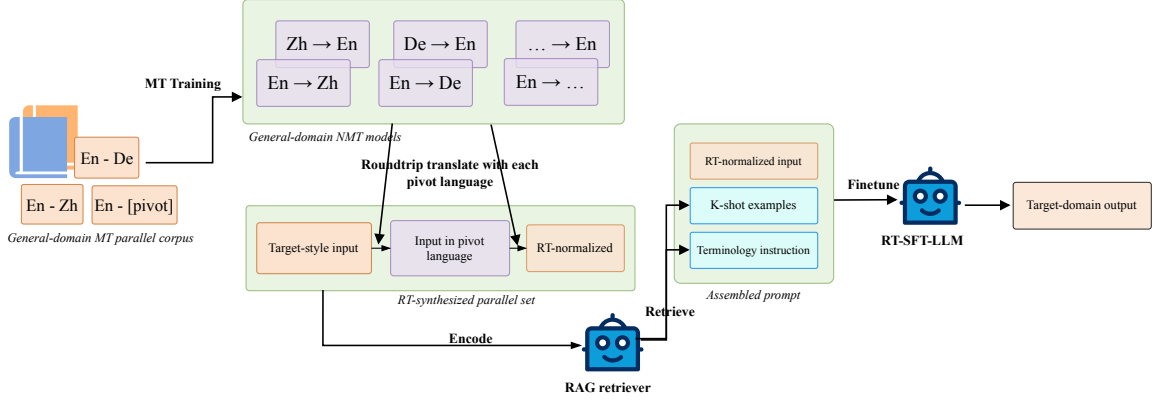


Figure 3: **Our proposed workflow.** We train MT systems using general domain parallel data for Chinese-English and German-English two-way translations, and roundtrip-translate target style text to obtain RT-normalized counterparts. We roundtrip translate every record with both pivot languages and mix the synthetic corpus. The synthetic parallel corpus is then used for finetuning a stylizer LLM, with retrieved information injected in the prompt.

corpus between English and a pivot language; composed into a roundtrip translation pipeline, it strips pronounced stylistic features while preserving content, yielding *RT-normalized* text; Figure 2 shows RT-normalized outputs alongside the same content rendered in each target style. Applying it to a monolingual in-style corpus pairs every sentence with its RT-normalized counterpart, producing the pseudo-parallel corpus on which we parameter-efficiently finetune an instruction-tuned LLM as the stylizer (an *RT-SFT-LLM*; Figure 1). At test time the two run in sequence, the normalizer followed by the stylizer (Figure 3).

This sequencing resolves a distribution mismatch: during finetuning the stylizer sees only RT-normalized inputs, whereas an arbitrary query at test time may carry pronounced, unseen style. Passing the query through the same roundtrip pipeline first mirrors the training-time construction exactly, and is the analogue of STRAP’s inference-time paraphrasing (Krishna et al., 2020), realized here by an NMT normalizer. We compare RT-first inference against feeding the raw query directly in §4.3, and find that it improves target-style conformity, especially for stylistically diverse or out-of-domain inputs.

### 3.2 RT-SFT with Parameter Efficient LoRA Finetuning

LoRA (Low-Rank Adaptation of Large Language Models) is an efficient approach that reduces computational and memory costs by using low-rank approximation techniques (Hu et al., 2021). The LoRA approach involves freezing the pre-trained

model’s weight matrices and introducing trainable low-rank decomposition matrices into the model’s layers. This approach allows us to finetune the 7B and 8B LLMs with 2 NVIDIA A100 GPUs, each with 80GB of memory. Hyperparameters and configurations we used are given in Appendix B.

### 3.3 RAG Retrievers for RT-SFT-LLMs

Lewis et al. (2020) proposed the Retrieval Augmented Generation (RAG) framework in which a retriever-generator model is trained end-to-end to enhance coherence between the pre-trained retriever and generator subsystems. In a similar spirit, our RAG system injects retrieved information into both finetuning and inference-time prompts for RT-SFT-LLM, unlike previous RAG-based TST methods where RAG is primarily considered a prompting technique (Liu et al., 2024a; Wan et al., 2023; Aycock and Bawden, 2024; Zhang et al., 2024). In §4.2, we present the performance of our system with increasing levels of retrieval augmentation compared against our method without RAG (RT-SFT 0-shot) and state-of-the-art inference-time RAG without training (5-shot In-Context Learning). We show that RAG introduced at training time produces considerable improvements compared to both RT-SFT without RAG and inference-time RAG.

#### 3.3.1 Relevant Example Retrieval Augmentation

The first type of retrieval augmentation we report on obtains sentence pairs as instructions for how we would like the query to be transferred. Results demonstrate that examples relevant to the query are

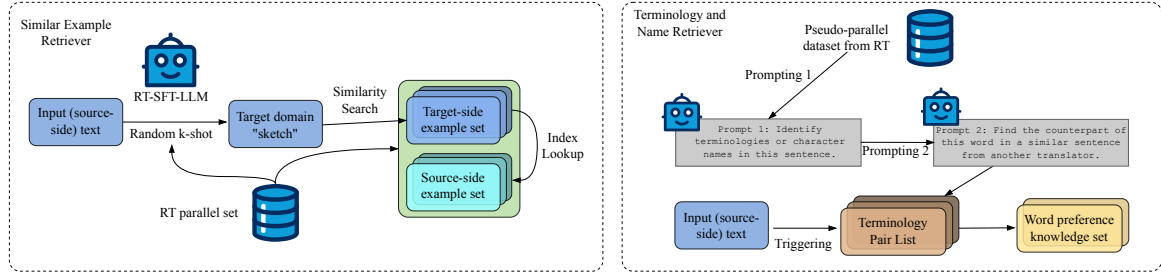


Figure 4: **Retrieval augmentation workflow. Left (a):** Relevant example retriever. We encode the **target-side texts** of the parallel synthetic datasets for nearest-neighbor search. For each query, we first do k-shot inference with the RT-SFT-LLM to obtain an “in-domain” sketch, which is used as search query in the target-side dataset to obtain the k most relevant pairs. **Right (b)** Terminology and name retriever: For each instance in the synthetic parallel datasets, the first LLM call extracts relevant words from the source side, then the second call matches them with their counterparts in the target side, yielding a terminology pair list for each domain.

more instructive and bring stronger performance improvements (§4.2).

We vectorize the pseudo-parallel corpus and use Faiss (Douce et al., 2024) (Figure 4) to retrieve nearest neighbors, following the original RAG schema (Lewis et al., 2020). Interestingly, we found that searching with the target side (in-domain) retrieves better instances than searching with the source (out-of-domain) side, giving slightly better results. We thus use retrieved examples with relevant target sides for example prompt injection. Specifically, we exclude the current instance itself during retrieval to avoid leakage.

For inference-time RAG where the target side text is unavailable and is what we want to predict, we take a “sketch-first” approach following Wang et al. (2022)’s two-round retrieval schema (Figure 4). We first perform few-shot inference with **randomly selected** examples to generate a draft “sketch” output that resembles the in-domain transferred generation, though of limited quality due to the random shots. We then use this draft as the query to retrieve the most relevant examples from the Faiss vector bank for a refined second-round inference. In §4.3, we show that relevant example augmentation at inference time through this approach yields better performance than randomly selected examples and inference without any examples (Table 3).

### 3.3.2 Terminology and Name List Retrieval

Diction and word preference are important aspects of text style domains. The same concept or object can be referred to by different terminologies in different domains. In literary translation domains, there is a similar issue of **naming consistency**, where machine-translated works may use seman-

tic translations and direct translations in different contexts to refer to the same characters, causing confusion (Matusov, 2019).

We improve our RT-SFT-LLM’s terminology correctness and long-context consistency by retrieving a **terminology and name list** from our pseudo-parallel corpus (Figure 4). For each record in the pseudo-parallel corpus, we first prompt an LLM with the source side of the paired sentences and ask it to identify any domain-specific terms or names in it. Then, we do a second round of prompting with the target side of the sentence pairs alongside the retrieved domain-specific terms, and ask the LLM to find their counterparts. Through this pipeline, we construct a list of source domain to target domain preferred terminology pairs. If any of the source-side words are present in the query, we add a one-sentence instruction in the inference prompt that provides terminology and name transfer guidance. The prompts we used are in Appendix A.

### 3.4 Evaluation: BLEU and Style Classification Accuracy Score

We primarily evaluate two aspects of our models, namely style-transfer quality and content preservation. For each style domain we train a BERT-based (Devlin et al., 2018) style classifier on held-out in-domain data, following Liu et al. (2024b,a); Mukherjee et al. (2024). The classifier scores a given text between 0 and 1, where we treat 1 as “in-domain” and 0 as “out-of-domain”. We apply the trained classifier to the output of our RT-SFT-LLMs to yield a style-classification accuracy that measures how well the output matches the target style. BLEU is computed between the generation and the reference in-domain text, measuring how

well the transferred output preserves the intended content.

## 4 Experiments

### 4.1 Experiment Setup

#### 4.1.1 Datasets, Synthetic Data Generation, and Baselines

| Dataset       | Training             | Dev          | Test         |
|---------------|----------------------|--------------|--------------|
| WMT26-eng-zho | <u>241,429,382</u>   | <u>3,000</u> | –            |
| WMT26-deu-eng | <u>1,128,176,715</u> | <u>3,000</u> | –            |
| MSD           | 130,199              | <u>150</u>   | <u>675</u>   |
| Shakespeare   | 24,852               | <u>1,313</u> | <u>1,293</u> |
| GYAFC         | 103,562              | <u>1,000</u> | <u>2,748</u> |
| ParaDetox     | 19,466               | <u>150</u>   | <u>150</u>   |

Table 1: **Datasets.** Underlined counts are parallel (aligned) sentence pairs; plain counts are monolingual sentences. The WMT26 General Machine Translation corpora (Kocmi et al., 2025) are used to train the roundtrip-translation normalizers; we use only the parallel portion of the released splits and manually hold out 3,000 sentence pairs per language pair as dev set. The four style datasets are used as training data for our RT-SFT method. Note that although 3 of the 4 style domains provide parallel training data, we only use the in-domain side of them during training, as our system only requires in-domain non-parallel training data. We use the provided, human-annotated parallel sentences only for the dev and test sets.

**Machine-translation data** We build the roundtrip translation normalizer from two general-domain parallel corpora released for the WMT26 General Machine Translation shared task (Kocmi et al., 2025), covering English–Chinese and German–English. We use only the parallel portion of the released splits to train the Marian NMT models (Junczys-Dowmunt et al., 2018), holding out 3,000 sentence pairs per language pair as a development set; detailed configurations are given in Appendix B. Every in-style sentence is roundtrip translated through both the Chinese and the German pivot, and the two resulting RT-normalized-to-target pairs are pooled into a single synthetic corpus for finetuning. Corpus sizes are reported in Table 1.

**Style-transfer datasets.** We evaluate on four publicly available benchmarks, each targeting a specific style and providing a human-aligned parallel test set. The domains are: expert medical text (MSD; Cao et al., 2020), Shakespearean English (Shakespeare; Xu et al., 2012), formal text

(GYAFC; Rao and Tetreault, 2018), and detoxified text (ParaDetox; Logacheva et al., 2022). Split sizes are listed in Table 1, where underlined counts denote parallel data and plain counts denote monolingual data.

**Baselines:** First, we set up an **In-Context Learning (ICL)** baseline method by prompting the LLMs with the same prompt from our own method, with the addition of few-shot examples retrieved based on similarity and the absence of finetuning. The second baseline we adopt is **Automatic Post-Editing (APE)**: we train an additional MT model for each domain on the output (RT-normalized) versus text input (target style), so that it performs English-to-English “translations” that bring RT-normalized texts to in-domain. This is a widely-used practice that accompanies Machine Translation systems (Simard et al., 2007; Moon et al., 2022), where the additional APE module learns to correct systematic errors in the MT system through NMT training. Our variant is monolingual (source-free), post-editing the roundtrip output directly to match the target text, which is the input to the roundtrip translation pipeline.

#### 4.1.2 TST Prompt Templates

We experimented with three potential prompt templates for TST finetuning. Prompt details and experiments on prompts are in Appendix A. After testing and careful evaluation, we decided to use the following prompt template for our method:

**Rewrite the given sentence into the style of [style name].**  
 Here are [n] examples:  
 Input: [example input i]. Output: [example output i].  
 .....  
 Note that you may want to rewrite "[input term]" to "[output term]" for contextual consistency.  
 Now go ahead: Input: [query input]. The [style name] output:

## 4.2 Overall Performance Comparison

This subsection compares roundtrip-translation supervised finetuning (RT-SFT) against three reference points across the four style domains (Table 2). We use Llama-3.1-8B-Instruct as the base model and are finetuned on the synthetic corpus of §4.1.1, which pools roundtrip translations through both pivots (Chinese and German); an experiment on various LLMs as the base model is reported in §4.4. **COPY** offers a calibration where we measure the

| Method                            | MSD          |              | Shakespeare  |              | GYAFC        |              | ParaDetox    |              |
|-----------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                                   | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     |
| COPY                              | 0.107        | 14.01        | 0.009        | 15.35        | 0.149        | 16.54        | 0.012        | 20.03        |
| 5-shot In-Context Learning (2022) | 0.396        | 9.58         | 0.771        | 12.79        | 0.521        | 19.34        | 0.792        | <b>21.35</b> |
| Automatic Post-Editing (2022)     | 0.672        | 12.21        | 0.693        | 11.90        | 0.435        | 17.74        | 0.720        | 17.35        |
| RT-SFT 0-shot                     | 0.706        | 11.32        | 0.841        | 11.47        | 0.591        | 21.33        | 0.820        | 18.40        |
| RT-SFT random-3-shot              | 0.715        | 13.35        | 0.809        | 13.21        | 0.620        | 22.41        | 0.893        | 18.11        |
| RT-SFT random-5-shot              | 0.732        | 13.61        | 0.794        | 13.42        | 0.611        | 22.22        | 0.891        | 18.09        |
| RT-SFT relevant-3-shot            | <b>0.797</b> | 13.59        | 0.841        | 14.65        | 0.642        | <b>24.79</b> | <b>0.935</b> | 17.91        |
| RT-SFT relevant-5-shot            | 0.782        | 13.53        | 0.839        | 14.77        | <b>0.644</b> | 4.32         | 0.931        | 19.35        |
| + terminology and name retrieval  | 0.763        | <b>13.79</b> | <b>0.842</b> | <b>17.35</b> | 0.612        | 24.17        | 0.912        | 19.91        |

Table 2: **Main results.** TST performance of RT-SFT against baselines across the four style domains (base model: Llama-3.1-8B-Instruct; synthetic corpus pooled from German and Chinese roundtrip pivots). *random* and *relevant* indicate whether the  $k$  in-context example pairs are sampled at random or retrieved for relevance; + *terminology and name retrieval* adds domain-specific term guidance when triggered (§3.3.2). Best result per column in bold, excluding COPY: returning the input unchanged attains very high BLEU at near-zero style accuracy, a known artefact of BLEU-based TST evaluation (Pang, 2019). RT-SFT improves style accuracy over both baselines in all four domains and in every configuration tested, and relevant-example retrieval improves accuracy over random examples in all eight paired comparisons.

BLEU and accuracy of the input itself against the target text. Among RT-SFT variants,  $k$ -**shot (random)** supplies  $k$  randomly drawn RT-normalized-to-target pairs and  $k$ -**shot (relevant)** the  $k$  most relevant pairs, retrieved by cosine-distance search at finetuning time and by the procedure of §3.3.1; the final variant adds **terminology and name retrieval** (§3.3.2). Results are single runs, and we do not interpret accuracy differences below 0.02.

Style accuracy is measured by a BERT classifier trained per domain (§3.4), so the COPY row calibrates it: it reports how often an untransferred input is already labelled in-domain. The near-zero values on Shakespeare (0.009) and ParaDetox (0.012) show that these boundaries are almost linearly separable, whereas MSD (0.107) and GYAFC (0.149) are harder to classify, and their scores must be read against a higher floor and a correspondingly lower ceiling (0.797 and 0.644, against 0.842 and 0.935). Against this calibration, RT-SFT improves accuracy over both baselines in all four domains and in every configuration tested—24 of 24 cells. Neither half of the method suffices alone: zero-shot RT-SFT, which uses no retrieval at all, beats ICL by +0.310, +0.070, +0.070 and +0.028, and beats APE by +0.034, +0.148, +0.156 and +0.100. Retrieval augmentation therefore adds to these gains rather than producing them. APE leads ICL only on MSD (0.672 vs. 0.396), the domain with the most specialized lexicon.

BLEU must be read jointly with accuracy, since the two trade off systematically in TST: it is widely

reported by prior work that untransferred input routinely outscores most systems (Pang, 2019; Mukherjee et al., 2023). Our COPY row reproduces this—considerably high BLEU across all four domains, and the highest BLEU on MSD (14.01) at only 0.107 accuracy—and explains the single column in which a baseline leads: 5-shot ICL takes the top ParaDetox BLEU (21.35) at a comparatively low style accuracy (0.792, against 0.912–0.935 for the relevant-shot RT-SFT variants); consistent with the style–content trade-off, it under-edits toward the input. A BLEU drop relative to COPY is thus expected wherever source and reference already overlap heavily. Where they genuinely diverge the trade-off does not bind: on GYAFC and Shakespeare the best RT-SFT variant leads on both metrics, improving BLEU by +5.45 and +4.56 over the stronger baseline.

Retrieval improves both metrics, but unevenly. Relevant examples beat random ones on accuracy in all eight paired comparisons (mean +0.043) and on BLEU in six of eight (mean +1.06), with gains concentrated where a length- and vocabulary-matched exemplar is available: +2.38 on GYAFC and +1.44 on Shakespeare at  $k=3$ , against +0.24 on MSD. The number of shots has no consistent effect, 3-shot winning five of eight cells by margins well below the relevant-vs-random gap. Terminology and name retrieval behaves as the lexical intervention it is: added to relevant-5-shot it raises BLEU in three of four domains—most on Shakespeare (+2.58, the largest single gain in the table),

| Inference methods                                  | MSD          |              | Shakespeare  |              | GYAFC        |              | ParaDetox    |              |
|----------------------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                                                    | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     |
| COPY                                               | 0.107        | 14.01        | 0.009        | 15.35        | 0.149        | 16.54        | 0.012        | 20.03        |
| 5-shot In-Context Learning (2022)                  | 0.396        | 9.58         | 0.771        | 12.79        | 0.521        | 19.34        | 0.792        | <b>21.35</b> |
| Automatic Post-Editing (2022)                      | 0.672        | 12.21        | 0.693        | 11.90        | 0.435        | 17.74        | 0.720        | 17.35        |
| 0-shot inference                                   | 0.730        | 13.31        | 0.791        | 17.21        | 0.595        | 23.96        | 0.909        | 17.07        |
| random 5-shot inference                            | 0.759        | 11.53        | 0.841        | 17.01        | 0.575        | 21.17        | 0.893        | 18.63        |
| relevant 5-shot inference without RT normalization | 0.671        | <b>14.55</b> | 0.731        | <b>18.69</b> | 0.556        | <b>26.62</b> | 0.902        | 20.21        |
| relevant 5-shot inference                          | <b>0.763</b> | 13.79        | <b>0.842</b> | 17.35        | <b>0.612</b> | 24.17        | <b>0.912</b> | 19.91        |

Table 3: TST performance under different inference procedures. All groups use one Llama-3.1-8B-Instruct RT-SFT-LLM, finetuned with 5-shot relevant examples and terminology retrieval (§4.2); baseline rows are repeated from Table 2 for reference. All RT-SFT rows apply RT-first inference except *without RT normalization*, which feeds the raw query. Best result per column in bold, excluding COPY. RT-first inference raises style accuracy and lowers BLEU in all four domains; relevant-example retrieval raises BLEU in all four while leaving accuracy largely unchanged.

where the style is carried by a fixed vocabulary and by character names—but on GYAFC, where formality is realized syntactically, it lowers both metrics. Averaged over the four domains it trades 0.017 accuracy for 0.82 BLEU. We therefore recommend it for domains with a stable terminology or naming convention, and relevant-example retrieval alone elsewhere.

### 4.3 Inference-time methods

We next perform ablation studies on the inference procedure (Table 3). All groups use the same Llama-3.1-8B-Instruct RT-SFT-LLM, finetuned with 5-shot relevant examples and terminology retrieval (§4.2), and differ only in what is done at test time: **0-shot** prompts with the task description alone, while **random** and **relevant 5-shot** append examples sampled at random or retrieved by the procedure of §3.3.1. The **without RT normalization** row feeds the raw query directly into the finetuned stylizer LLM without being normalized by the RT pipeline, and provides the controlled contrast isolating RT-first inference.

RT-first inference yields a strictly monotone trade: style accuracy rises in all four domains (+0.092, +0.111, +0.056, +0.010) and BLEU falls in all four (−0.76, −1.34, −2.45, −0.30). Aligning the query with the finetuning distribution thus costs some semantic drift through the pivot. The gain is largest on MSD and Shakespeare, whose source registers are the most strongly marked, and negligible on ParaDetox, where source and target differ by a small lexical edit and little style remains for the normalizer to strip.

At inference time, relevant examples act on content rather than style: against random 5-shot they

raise BLEU in all four domains (mean +1.72) but leave accuracy unchanged in three of four, reversing the pattern of Table 2, where finetuning-time relevance was primarily an accuracy lever. Random examples are actively harmful—0-shot inference exceeds random 5-shot on BLEU in three of four domains. That 0-shot inference here also exceeds the 0-shot-*finetuned* model of Table 2 on BLEU in three domains, neither prompt containing any example, indicates that the benefit of retrieval-augmented finetuning is retained in the weights.

### 4.4 Experiments on Pretrained LLMs

To demonstrate the generalizability of RT-SFT, we repeated the experiment with three other pretrained LLMs under the best configuration of §4.2—relevant 5-shot examples with terminology and name retrieval (Table 4). Style accuracy is again given by the per-domain BERT classifier of §3.4, and COPY provides the untransferred floor.

Every model clears that floor in all sixteen cells, by margins of 0.403 to 0.900, so the synthetic corpus supplies a usable training signal irrespective of the base LLM. Llama-3.1-8B-Instruct is strongest on both metrics—leading style accuracy in all four domains and BLEU in three, with Gorilla-openfunctions-v2 ahead by 0.02 on ParaDetox—and we therefore adopt it for §4.2 and §4.3. Gorilla and Mistral-7B are indistinguishable from each other on both metrics (average BLEU 17.32 vs. 17.22, accuracy 0.710 vs. 0.700). Scale is not the binding constraint: opt-2.7b, roughly a third the size of the others, trails Llama by only 0.095 in average accuracy and still exceeds the 5-shot ICL baseline of Table 2 in three of four domains. As in §4.2, all four models fall below the input itself

| Pretrained LLMs                      | MSD          |              | Shakespeare  |              | GYAFC        |              | ParaDetox    |              |
|--------------------------------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|--------------|
|                                      | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     |
| COPY                                 | 0.107        | 14.01        | 0.009        | 15.35        | 0.149        | 16.54        | 0.012        | 20.03        |
| meta-llama/Llama-3.1-8B-Instruct     | <b>0.763</b> | <b>13.79</b> | <b>0.842</b> | <b>17.35</b> | <b>0.612</b> | <b>24.17</b> | <b>0.912</b> | 19.91        |
| gorilla-llm/gorilla-openfunctions-v2 | 0.661        | 11.07        | 0.733        | 17.11        | 0.573        | 21.15        | 0.874        | <b>19.93</b> |
| mistralai/Mistral-7B                 | 0.692        | 12.21        | 0.749        | 15.97        | 0.552        | 22.23        | 0.806        | 18.45        |
| facebook/opt-2.7b                    | 0.631        | 11.39        | 0.701        | 13.33        | 0.606        | 21.27        | 0.811        | 17.21        |

Table 4: TST finetuning performance across base LLMs (relevant 5-shot finetuning with terminology and name retrieval; synthetic corpus pooled from German and Chinese roundtrip pivots). The Llama-3.1-8B-Instruct row is the best configuration of Table 2, repeated here for comparison. COPY returns the input unchanged and gives the untransferred floor for each domain’s style classifier. Best result per column in bold, excluding COPY. All four base models exceed that floor in every domain; Llama-3.1-8B-Instruct leads on style accuracy in all four and on BLEU in three.

(COPY) on MSD and ParaDetox BLEU, where source and reference already overlap heavily.

## 5 Conclusion

We presented a method for Text Style Transfer that addresses the scarcity of parallel corpora by using roundtrip translation as a style normalizer to generate parallel training data. Because general-domain NMT preserves content while regressing toward generic phrasing, roundtrip-translating a monolingual in-style corpus yields the pseudo-parallel data that supervised finetuning requires, turning normalization from an inference-time patch into a data-generation tool. LoRA-finetuning an instruction-tuned LLM on this corpus improves style-classification accuracy over both In-Context Learning and Automatic Post-Editing in all four domains and in every configuration tested, and does so without either half of the method being sufficient alone: zero-shot RT-SFT, which uses no retrieval, already beats ICL—the same model and prompt without the synthetic corpus—and APE, the same corpus without the LLM stylizer. BLEU behaves as expected under the style–content trade-off (Mukherjee et al., 2023; Pang, 2019), and RT-SFT leads on both metrics in the two domains where source and reference genuinely diverge.

Retrieval augmentation applied at training time, rather than only in the prompt, improves style accuracy in all eight paired comparisons against randomly sampled examples, and its benefit is retained in the weights: a retrieval-augmented model prompted with no examples outperforms one finetuned without retrieval and likewise prompted with none. Terminology and name retrieval trades a small amount of accuracy for content preservation, and helps most where the target style is carried by

a fixed vocabulary.

Roundtrip-translating the query at inference time to match the finetuning distribution raises style accuracy in all four domains and lowers BLEU in all four, a trade whose size tracks how strongly marked the source register is. Quantifying that trade-off under stylistically diverse and genuinely out-of-domain queries is the natural next step, along with varying the pivot language and extending a single RT-SFT-LLM to multi-style and non-binary transfer.

## Limitations

The main limitations of our work are as follows:

- **Semantic drift and error propagation.** Our method relies on machine translation models to generate parallel finetuning datasets. As a result, its performance depends on the quality of the underlying NMT systems and their training data. We observed that when these models introduce errors or cause semantic drift during roundtrip translation, such inaccuracies become embedded in the synthetic parallel corpus used for finetuning. We applied post-processing steps to mitigate such effects, and further efforts could also be made to test various NMT methods or architectures to find the most suitable configuration for the TST task. These improvements and post-editing works, however, are beyond the scope of this study.
- **Alternatives for the Current Roundtrip Translation Pipeline.** In this work, we primarily used Marian to train the NMT models and did not explore alternative methods or

workflows for performing roundtrip translation. An intriguing potential alternative is to employ large language models to perform machine translation, either by ICL or finetuning, which might yield better results compared to the current Marian-based approach. However, we did not test these alternative approaches in the current study owing to constraints on time and space.

- **Limits to domains with available corpus.**

Due to data availability constraints, our experiments are conducted on four style domains, which may not fully capture the range of stylistic variations encountered in real-world scenarios. This limitation could introduce biases into our analysis and potentially restrict the generalizability of our methods. We selected benchmarks that are as diverse and distinctive as possible—spanning Shakespearean English, formality, expert–layman medical communication, and toxicity—in an effort to enhance the overall robustness and applicability of our method. We strive to enhance the generalizability of our experiments and demonstrate the effectiveness of our method in different domains and conditions.

## References

- Seth Aycock and Rachel Bawden. 2024. [Topic-guided example selection for domain adaptation in llm-based machine translation](#).
- Avanti Bhandarkar, Ronald Wilson, Anushka Swarup, and Damon Woodard. 2024. [Emulating author style: A feasibility study of prompt-enabled text stylization with off-the-shelf llms](#).
- Eleftheria Briakou, Di Lu, Ke Zhang, and Joel Tetreault. 2021. Olá, bonjour, salve! XFORMAL: A benchmark for multilingual formality style transfer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*.
- Yixin Cao, Ruihao Shui, Liangming Pan, Min-Yen Kan, Zhiyuan Liu, and Tat-Seng Chua. 2020. Expertise style transfer: A new task towards better communication between experts and laymen. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Keith Carlson, Allen Riddell, and Daniel Rockmore. 2018. Evaluating prose style transfer with the Bible. *Royal Society Open Science*, 5(10).
- Jianlin Chen. 2024. [Lmstyle benchmark: Evaluating text style transfer for chatbots](#).
- Daryna Dementieva, Nikolay Babakov, Amit Ronen, Abinew Ali Ayele, Naqee Rizwan, Florian Schneider, Xintong Wang, Seid Muhie Yimam, Daniil Moskovskiy, Elisei Stakovskii, Eran Kaufman, Ashraf Elnagar, Animesh Mukherjee, and Alexander Panchenko. 2025. Multilingual and explainable text detoxification with parallel corpora. In *Proceedings of the 31st International Conference on Computational Linguistics (COLING)*.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina N. Toutanova. 2018. [Bert: Pre-training of deep bidirectional transformers for language understanding](#).
- Matthijs Douze, Alexandr Guzhva, Chengqi Deng, Jeff Johnson, Gergely Szilvasy, Pierre-Emmanuel Mazaré, Maria Lomeli, Lucas Hosseini, and Hervé Jégou. 2024. [The faiss library](#). *Preprint*, arxiv:2401.08281 [cs].
- Shuhuan Gu, Wenbiao Tao, Xinchun Ma, Kangkang He, Ye Guo, Xiang Li, and Yunshi Lan. 2026. [Unsupervised text style transfer for controllable intensity](#). *Preprint*, arXiv:2601.01060.
- Junxian He, Xinyi Wang, Graham Neubig, and Taylor Berg-Kirkpatrick. 2020. [A probabilistic formulation of unsupervised text style transfer](#).
- Zachary Horvitz, Ajay Patel, Chris Callison-Burch, Zhou Yu, and Kathleen McKeown. 2024a. ParaGuide: Guided diffusion paraphrasers for plug-and-play textual style transfer. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
- Zachary Horvitz, Ajay Patel, Kanishk Singh, Chris Callison-Burch, Kathleen McKeown, and Zhou Yu. 2024b. [TinyStyler: Efficient few-shot text style transfer with authorship embeddings](#). *Preprint*, arXiv:2406.15586.
- Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.
- J. Edward Hu, Abhinav Singh, Nils Holzenberger, Matt Post, and Benjamin Van Durme. 2019. [Large-scale, diverse, paraphrastic bitexts via sampling and clustering](#). In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 44–54, Hong Kong, China. Association for Computational Linguistics.
- Anubhav Jangra, Preksha Nema, and Aravindan Raghuvier. 2022. T-STAR: Truthful style transfer using AMR graph as intermediate representation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

- Anubhav Jangra, Bahareh Sarrafzadeh, Silviu Cucerzan, Adrian de Wynter, and Sujay Kumar Jauhar. 2025. [Evaluating style-personalized text generation: Challenges and directions](#). *Preprint*, arXiv:2508.06374.
- Lasal Jayawardena and Prasan Yapa. 2024. [ParaFusion: A large-scale LLM-driven English paraphrase dataset infused with high-quality lexical and syntactic diversity](#). *Preprint*, arXiv:2404.12010.
- Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. [Shakespeareizing modern language using copy-enriched sequence to sequence models](#). In *Proceedings of the Workshop on Stylistic Variation*, pages 10–19, Copenhagen, Denmark. Association for Computational Linguistics.
- Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. [Marian: Fast neural machine translation in c++](#).
- Tom Kocmi, Ekaterina Artemova, Eleftherios Avramidis, Rachel Bawden, Ondřej Bojar, Konstantin Dranch, Anton Dvorkovich, Sergey Dukanov, Mark Fishel, Markus Freitag, Thamme Gowda, Roman Grundkiewicz, Barry Haddow, Marzena Karpinska, Philipp Koehn, Howard Lakouga, Jessica Lundin, Christof Monz, Kenton Murray, and 10 others. 2025. [Findings of the WMT25 general machine translation shared task: Time to stop evaluating on easy test sets](#). In *Proceedings of the Tenth Conference on Machine Translation*, pages 355–413, Suzhou, China. Association for Computational Linguistics.
- 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)*.
- Huiyuan Lai, Antonio Toral, and Malvina Nissim. 2021. [Thank you BART! rewarding pre-trained models improves formality style transfer](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 484–494, Online. Association for Computational Linguistics.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. [Retrieval-augmented generation for knowledge-intensive NLP tasks](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.
- Qingyi Liu, Jinghui Qin, Wenxuan Ye, Hao Mou, Yuxuan He, and Keze Wang. 2024a. [Adaptive prompt routing for arbitrary text style transfer with pre-trained language models](#).
- Xinyue Liu, Harshita Diddee, and Daphne Ippolito. 2024b. [Customizing large language model generation style using parameter-efficient finetuning](#).
- 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.
- Evgeny Matusov. 2019. [The challenges of using neural machine translation for literature](#). In *Proceedings of the Qualities of Literary Machine Translation*, pages 10–19, Dublin, Ireland. European Association for Machine Translation.
- Hyeonseok Moon, Chanjun Park, Jaehyung Seo, Sugyeong Eo, and Heuiseok Lim. 2022. [An automatic post editing with efficient and simple data generation method](#). *IEEE Access*, 10:21032–21040.
- Jihyung Moon, Hyunchang Cho, and Eunjeong L. Park. 2020. [Revisiting round-trip translation for quality estimation](#). *CoRR*, abs/2004.13937.
- Sourabrata Mukherjee, Zdeněk Kasner, and Ondřej Dušek. 2023. Balancing the style-content trade-off in sentiment transfer using polarity-aware denoising. In *Text, Speech, and Dialogue (TSD)*.
- Sourabrata Mukherjee, Atul Kr Ojha, and Ondřej Dušek. 2024. [Are large language models actually good at text style transfer?](#)
- Sharmila Reddy Nangi, Niyati Chhaya, Sopan Khosla, Nikhil Kaushik, and Harshit Nyati. 2021. Counterfactuals to control latent disentangled text representations for style transfer.
- Nasim Nouri. 2022. Text style transfer via optimal transport.
- Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. 2024. [Unsupervised text style transfer via llms and attention masking with multi-way interactions](#).
- Richard Yuanzhe Pang. 2019. [The daunting task of real-world textual style transfer auto-evaluation](#). *Preprint*, arXiv:1910.03747.
- Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation.
- Vitaly Protasov, Nikolay Babakov, Daryna Dementieva, and Alexander Panchenko. 2025. [Evaluating text style transfer: A nine-language benchmark for text detoxification](#). *Preprint*, arXiv:2507.15557.
- Ella Rabinovich, Raj Nath Patel, Shachar Mirkin, Lucia Specia, and Shuly Wintner. 2017. [Personalized machine translation: Preserving original author traits](#). In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1074–1084,

- Valencia, Spain. Association for Computational Linguistics.
- 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, New Orleans, Louisiana. Association for Computational Linguistics.
- 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)*.
- Parker Riley, Noah Constant, Mandy Guo, Girish Kumar, David Uthus, and Zarana Parekh. 2021. Textsettr: Few-shot text style extraction and tunable targeted restyling.
- Shamik Roy, Raphael Shu, Nikolaos Pappas, Elman Mansimov, Yi Zhang, Saab Mansour, and Dan Roth. 2023. Conversation style transfer using few-shot learning. In *Proceedings of the 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL)*.
- Tianxiao Shen, Tao Lei, Regina Barzilay, Tommi Jaakkola, and Mit Csail. 2017. [Style transfer from non-parallel text by cross-alignment](#).
- Michel Simard, Cyril Goutte, and Pierre Isabelle. 2007. Statistical phrase-based post-editing. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT)*, pages 508–515.
- Harold Somers. 2005. Round-trip translation: What is it good for? In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 127–133.
- Sandeep Subramanian, Guillaume Lample, Eric Michael Smith, Ludovic Denoyer, Marc’Aurelio Ranzato, and Y-Lan Boureau. 2018. [Multiple-attribute text style transfer](#).
- Han Sun, Zhen Sun, Zongmin Zhang, Linzhao Jia, Wei Shao, and Min Zhang. 2025. [SynDec: A synthesize-then-decode approach for arbitrary textual style transfer via large language models](#). *Preprint*, arXiv:2505.12821.
- 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 (EMNLP)*.
- Zhen Tao, Dinghao Xi, Zhiyu Li, Liumin Tang, and Wei Xu. 2024. [Cat-llm: Prompting large language models with text style definition for chinese article-style transfer](#).
- Youzhi Tian, Zhiting Hu, and Zhou Yu. 2018. [Structured content preservation for unsupervised text style transfer](#).
- Martina Toshevskas and Sonja Gievska. 2024. [Large Language Models for Text Style Transfer: Exploratory Analysis of Prompting and Knowledge Augmentation Techniques](#).
- Martina Toshevskas and Sonja Gievska. 2025. LLM-based text style transfer: Have we taken a step forward? *IEEE Access*, 13:44707–44721.
- Rob Voigt, David Jurgens, Vinodkumar Prabhakaran, Dan Jurafsky, and Yulia Tsvetkov. 2018. [Rtgender: A corpus for studying differential responses to gender](#).
- Zhen Wan, Yating Zhang, Yexiang Wang, Fei Cheng, and Sadao Kurohashi. 2023. [Reformulating domain adaptation of large language models as adapt-retrieve-revise: A case study on chinese legal domain](#).
- Yifan Wang, Zewei Sun, Shanbo Cheng, Weiguo Zheng, and Mingxuan Wang. 2022. [Controlling styles in neural machine translation with activation prompt](#).
- Yusen Wu and Xiaotie Deng. 2025. [Implementing long text style transfer with LLMs through dual-layered sentence and paragraph structure extraction and mapping](#). *Preprint*, arXiv:2505.07888.
- Yifei Xie, Jiaping Gui, Zhengping Che, Leqian Zhu, Yahao Hu, and Zhisong Pan. 2025. TSTBench: A comprehensive benchmark for text style transfer. *Entropy*, 27(6):575.
- Wei Xu, Alan Ritter, Bill Dolan, Ralph Grishman, and Colin Cherry. 2012. Paraphrasing for style. In *Proceedings of COLING 2012*.
- Xinchen Yang and Marine Carpuat. 2025. [Steering large language models with register analysis for arbitrary style transfer](#). *Preprint*, arXiv:2505.00679.
- Chiyu Zhang, Honglong Cai, Yuezhong, Li, Yuexin Wu, Le Hou, and Muhammad Abdul-Mageed. 2024. [Distilling text style transfer with self-explanation from llms](#).
- Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. [Style transfer as unsupervised machine translation](#).

Table 5: Prompts for TST finetuning

| Prompt Template Index | Prompt Template Text                                                                                                                                                                                                                                                                                     |
|-----------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| I                     | Rewrite the following sentence into the style of [style name]. Here are [n] examples: Input: [example input i]. Output: [example output i]. Note that you may want to rewrite "[input term]" to "[output term]" for contextual consistency. Now go ahead: Input: [query input]. The [style name] output: |
| II                    | Rewrite the following sentence into the style of [style name]: Input: [query input]. Here are [n] examples: Input: [example input i]. Output: [example output i]. Note that word [input term] should be rewritten to [output term] for contextual consistency.                                           |
| III                   | Note that word [input term] should be rewritten to [output term] for contextual consistency. General domain: [example input i]. [style name] domain: [example output i]. general domain: Input: [query input]. [style name] domain:                                                                      |

Candidate prompt templates for LLM style transfer finetuning.

## A Prompt Templates

### TST finetuning prompts:

We experimented with three potential prompt templates for RT-SFT (Table 5). These prompts organize the query input sentence and several example sentence pairs into a prompt, with proper task descriptions and guidance for the generation. Templates (I) and (II) explicitly state the rewriting task, but have different orders of the example and query content. Template (III) is a classic Machine Translation prompt template with demonstrated effectiveness for Machine Translation with LLM. By changing language name to style domain name, we adapt it to guide LLM for text style transfer task.

In this experiment (Table 6), we conducted relevant 5-shot finetuning with terminology retrieval on Llama-3.1-8B-Instruct with the different templates, while leaving other conditions unchanged. Template (I) has the overall highest style accuracy score in the two tested domains. We thus used Template (I) throughout.

#### Terminology retrieval prompts:

We retrieved domain-specific term and name pair lists for each domain to enhance TST performances, by calling the LLM twice for each instance in the synthetic parallel corpus. The prompts we used are shown in Table 7.

| Template     | MSD          |              | Shakespeare  |              |
|--------------|--------------|--------------|--------------|--------------|
|              | Acc.         | BLEU-ref     | Acc.         | BLEU-ref     |
| COPY         | 0.107        | <b>14.01</b> | 0.009        | 15.35        |
| APE (2022)   | 0.672        | 12.21        | 0.693        | 11.90        |
| Template I   | <b>0.763</b> | 13.79        | <b>0.842</b> | 17.35        |
| Template II  | 0.614        | 12.91        | 0.773        | 16.69        |
| Template III | 0.721        | 13.63        | 0.751        | <b>18.01</b> |

Table 6: BLEU and acc. score across the MSD and Shakespeare domains for three potential templates. Template (I) has consistently higher style accuracy score, indicating robust ability in transferring to target style. Both Template (I) and (III) have stably high BLEU score.

Table 7: Prompts for terminology RAG

| Prompt type  | Prompt Text                                                                                                                                                                                    |
|--------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|
| First round  | Identify terminologies or character names in the sentence and return in comma separated format, without any additional explanation. Sentence: [source-side sentence]. Terminologies and names: |
| Second round | Find the counterpart of the word [source-side retrieved word] in the following sentence and return a single word, without any additional explanation. Sentence: [target-side sentence]:        |

**Prompts for terminology RAG.** The first prompt retrieves a list of terminologies and names from the source side sentence of each parallel instance, and for each of these retrieved words, the second prompt retrieves its counterpart in the corresponding target side sentence.

## B Hyperparameters and experiment configurations

### LoRA finetuning hyperparameters:

We set the learning rate to 2e-4, rank for the low-rank approximation is set to 512, the scaling factor is set to 256, and we use float16 data type. A dropout rate of 0.05 is applied. We save and evaluate the model every 2000 steps.

### Marian Configurations:

We used Marian for the Neural Machine Translation model pairs in the roundtrip translation system. In our system we used the Transformer architecture with R2L Reranking, with learning rate 0.0001, 49500 BPE operations, and step size 20000.

### Decoding configuration:

All generation from the RT-SFT-LLMs uses greedy decoding with beam size 1 and temperature 0. Generation is run with vllm in float16. The sketch-first

retrieval pass of §3.3.1 is decoded with the same configuration, so the first-round sketch, and hence the neighbour set it retrieves for the second round, is fixed for a given query and checkpoint.