Style EmuLoRAtion in Text Generation: A Case Study with Joe Biden and Donald Trump

Stanford CS224N {Custom} Project

Luke Mann
Department of Computer Science
Stanford University
lukemann@stanford.edu

Ori Spector
Department of Humanities and Sciences
Stanford University
orispec@stanford.edu

Abstract

Style emulation for text generation has been a challenging problem in Natural Language Processing. Recently, neural network-based models have shown tremendous progress in style transfer text generation, which aims to generate text that emulates the style of a given input text. In this study, we investigate the potential of Low Rank Adaptation (LoRA) for style transfer text generation, with the aim of emulating the writing styles of two prominent politicians, Joe Biden and Donald Trump. We train a neural network model using the LLaMA transformer and adapt it using the LoRA technique, supported by Parameter-Efficient Fine-Tuning (PEFT). Our evaluation includes transfer strength measured by a custom classifier, semantic preservation assessed using BLEU score, and human evaluation for style ranking. The results demonstrate the ability to produce coherent and natural text that captures the respective writing styles accurately in some contexts. Despite some limitations in the evaluation methods, our findings suggest that LoRA is a promising approach for style transfer text generation and warrant further exploration into alternative evaluation metrics and model architectures to improve performance in diverse applications.

1 Key Information to include

Our mentor is Heidi Zhang. We have no external collaborators and are not project sharing.

2 Introduction

Style transfer, the process of modifying the writing style of a given text while preserving its content and meaning, has emerged as a popular and challenging task in natural language processing and artificial intelligence research. This task holds great potential for various applications, including personalized messaging, dialog system response generation, and stylized text summarization, among others. Although several methods have been proposed to address style transfer, the complexity of the problem lies in effectively capturing and reproducing the unique nuances of an individual's writing style while maintaining the content's semantic coherence.

Recent advances in deep learning, particularly transformer-based architectures (Vaswani et al., 2017 [1]), have led to impressive results in various NLP tasks. However, training these models typically requires large amounts of data and computational resources, which may be inefficient for fine-grained tasks such as style transfer. Low Rank Adaptation (LoRA) (hu et al., (2021 [2]) offers a promising solution to this issue, providing parameter-efficient fine-tuning (PEFT) of pre-trained transformer models and yielding significant improvements in downstream tasks with limited data.

In this paper, we investigate the potential of LoRA for style transfer text generation, focusing on emulating the writing styles of Joe Biden and Donald Trump. We propose an approach that combines the LLaMA transformer (Touvron et al., 2023 [3]) with LoRA, taking advantage of PEFT to adapt the model for the specific task of style transfer.

2.1 Useful model descriptions

Zero-Shot Learning
ZSL involves recognizing new categories of instances with no training examples by leveraging an intermediate semantic layer of attributes that provide semantic information about the categories. ZSL approaches aim to learn these attributes and use them to predict new classes based on their attribute descriptions [4]. We use this as an initial baseline for prompt-evalutation.

Low-Rank Adaption
LoRA is a reparametrization technique for fine-tuning pre-trained deep learning models that is based on the hypothesis that the updates to the weight matrices during adaptation have a low intrinsic rank and can be represented by a low-rank decomposition. The approach involves representing the update to the weight matrix with a product of two matrices, one of which is low-rank and contains trainable parameters, while the other is the pre-trained weight matrix, which remains fixed and does not receive gradient updates. The benefits of this approach include reduced memory and storage usage during fine-tuning and marginally outperforms other models in terms of accuracy.

Parameter-Efficient Fine-Tuning
Parameter-Efficient Fine-Tuning (PEFT) [5], an open source hugging face library, is a method that allows for efficient adaptation of pre-trained language models to different applications by only fine-tuning a small number of extra model parameters, as opposed to fine-tuning all model parameters, which can be expensive. By doing this, PEFT reduces computational and storage costs while achieving similar performance to full fine-tuning. The PEFT is fully supported with LoRA and thus can be applied in conjunction to fine-tune pre-trained language models with reduced parameters, leading to even more efficient adaptation to downstream applications while achieving comparable performance to full fine-tuning.

Prefix-Tuning
Prefix-Tuning is an efficient fine-tuning method for pre-trained language models introduced by Li et al. (2021)[6]. It focuses on modifying a small set of learnable tokens, called "prefix tokens," which are concatenated with the input sequence before being fed into the model. The primary advantage of Prefix-Tuning is that it requires updating only a small fraction of the model's parameters, significantly reducing the computational resources needed for fine-tuning. Furthermore, the method has been shown to achieve comparable performance to full-model fine-tuning across various tasks and scales.

3 Related Work

In this section we highlight the current status of NLP style transfer in text generation. Specifically, we emphasize the current methodology and evaluation metrics used in research.

3.1 Current Generation Methods For Style Transfer

Seq2Seq Model
Jhamtani et al. (2017) [7] delves into the concept of style transfer within the field of NLP. The paper investigates the use of automated techniques to convert contemporary English text into Shakespearean English by leveraging parallel data. To achieve this, the authors enhanced the seq2seq model with a pointer network (Vinyals et al. 2015) [8]. The pointer network used a dictionary of modern-Shakespearean word pairs to generate candidate words. However, such paired word dictionaries are scarce resources that are not readily available for most style transfer tasks, and the creation of parallel corpora is necessary.

Auto-Encoder Models
Auto-Encoder (AE) is a widely used technique to learn the latent representation of input sentences by encoding them into a latent vector and then reconstructing a similar sentence. However, to avoid blindly copying all the input elements, Hill el al (2016) [9] replaced AE with denoising auto-encoding (DAE), which first passes the input sentence through a noise model to randomly drop, shuffle, or mask some words. This method has been used in later text style transfer works and has gained popularity

due to better performance in semantic preservation and fluency, especially with pre-training over large corpora, compared to GAN and VAE. To reconstruct data, a Variational Auto-Encoder (VAE) uses a sampled latent vector from its posterior and applies Kullback–Leibler divergence regularization [10]. VAEs are also widely used in style emulations tasks, and the loss function is calculated using a hyper-parameter λ to balance the reconstruction loss and KL divergence term. The prior $p(z)$ is drawn from a standard normal distribution, and the posterior $qE(z|x)$ is in the form of $N(\mu, \sigma)$, where μ and σ are predicted by the encoder. The VAE loss function is represented as

$$L_{VAE}(\theta_E, \theta_G) = -\mathbb{E}_{q_E(z|x)} \log p_G(x|z) + \lambda \mathrm{KL}\left[q_E(z|x) || p(z)\right] \tag{1}$$

Style-embedding Model and Multi-Decoder Model
Fu et al. (2018) [11] propose two methods for style emulation: style-embedding and multi-decoder. Neural sequence to sequence models (Seq2Seq) are the basis for two models that have been developed for style emulation. Both models aim to learn a representation for the input sentence that only contains content information. However, the multi-decoder model takes a different approach, using separate decoders for each style to generate texts in the corresponding style. On the other hand, the style-embedding model not only learns content representations but also style embeddings. A single decoder is then trained to generate texts in various styles, based on both the content representation and the style embedding.. The figure below shows a representation of their models.

3.2 Current Automatic Evaluation Methods For Style Transfer

Transfer Strength
Transfer strength refers to the ability of a model to successfully transfer the style from the source to the target. To measure the strength of the transferred style, many researchers train a style classifier to identify the attributes of the generated samples. The classifier is used to assess whether the model has successfully transferred the target attribute to each sample. The transferred style strength is then computed as such,

$$\frac{\textit{\# test samples correctly classified}}{\textit{\# all test samples}} \tag{2}$$

Li et al. (2018) [12] found that the attribute classifier's performance correlated well with human evaluation on some datasets but not on others (e.g., Amazon), suggesting that the effectiveness of the transferred style strength metric may vary depending on the dataset.

Semantic Preservation
Semantic preservation refers to the ability of a style transfer model to preserve the meaning and essence of the source text while transferring its style to the target text. In previous research this has been done using a wide variety of evaluation metrics including: BLEU score, ROUGE score, cosine similarity based on sentence embeddings, and BERTScore [13].

4 Approach

In this section, we describe our approach to style transfer text generation using two techniques: LoRA and Prefix-Tuning. Our goal is to finetune the LLaMA large language model to generate text that emulates the writing style of two prominent politicians, Joe Biden and Donald Trump. We first provide an overview of the LLaMA transformer and then delve into the specifics of our adaptation techniques.

LLaMA Transformer
For our research, we utilize LLaMA-7B which is open-sourced on HuggingFace. The LLaMA transformer is a state-of-the-art pretrained language model that has demonstrated excellent performance in various natural language understanding tasks. It serves as the backbone for our style transfer text generation model. LLaMA follows the typical transformer architecture, consisting of multiple self-attention layers with position-wise feedforward networks. We finetune this model on a dataset of political speeches and remarks from both Biden and Trump. One benefit that swayed us to use LLaMA-7B as highlighted by Touvron et al. (2023)[3] is that it can be run on a single GPU.

Low Rank Adaptation (LoRA)
As defined earlier in Section 2, we apply LoRA to the query (q) and value (v) projection matrices of

the self-attention layers, which have been identified as the most influential parameters for capturing style-specific information. The adapted weight matrices are computed as:

$$W'_v = W_v + U_v * V_v, \qquad W'_q = W_q + U_q * V_q \tag{3}$$

where W_q and W_v are the original weight matrices, U_q and V_q are the low-rank matrices for the query projections, and U_v and V_v are the low-rank matrices for the value projections. The low-rank matrices are initialized randomly and updated using gradient descent during the fine-tuning process.

Prefix-Tuning
As defined above in Section 2, we utilize Prefix-Tuning for PEFT. In our approach, we use 32 virtual tokens and a prefix projection matrix to condition the generation process. The combined input embeddings are calculated as:

$$X' = X + P * E_v \tag{4}$$

where X is the original input embeddings, P is the prefix projection matrix, and E_v is the virtual token embeddings. We perform experiments on a dataset of political speeches and remarks from both Biden and Trump. The dataset is split into training and validation sets, with the training set used to finetune the model and the validation set used for evaluation. We train our model for multiple epochs and evaluate its performance on style transfer text generation.

Training
We train our model using the standard cross entropy loss function. The loss function is defined as the cross-entropy between the predicted next-token probabilities and the actual next tokens in the input sequence.

$$L = -\sum_{i=1}^{n} y_i \log(p_i) \tag{5}$$

We optimize the model using the AdamW optimizer [14] with a learning rate of 3e-4 and apply gradient accumulation to handle large batch sizes efficiently. To ensure that our model is able to generate contextually relevant responses, we preprocess our dataset by pairing each input context with the respective speaker's name. This allows the model to condition its generation on both the input context and the desired speaker's style.

Data Preprocessing
To create the dataset, we scrape interview transcripts of Joe Biden and Donald Trump, ensuring a diverse range of topics and speaking styles. The context for the model is set as the previous 2, 4, and 6 lines of conversation, with the goal of generating a next line as if it was spoken by the subject. As such, there are $3n$ total datas per dataset. We preprocess the transcripts by tokenizing the text and dividing the conversations into context (prior lines of conversation) and response (true subject response) pairs. The context is used as input to the model, while the response serves as the ground truth for training.

5 Experiments

5.1 Data

Our goal is to train a model to generate text in the style of Biden and Trump. We created two datasets to achieve our goal: 1) a dataset to train our classifier (used for evaluation) and 2) a dataset to train our model.

Classifer Dataset
To train our classifier, we created a sample dataset for each subject respectively. The datasets contain $n \approx 1000$ text samples with $\frac{3n}{4}$ samples directly from transcripts speeches and the remaining $\frac{n}{4}$ samples from public remarks. Our model tokenizes these texts, thus, when processing this data as tokens this amounts to $\approx 1M$ tokens processed per subject. Each transcript is of high quality for style emulation because it comes directly from the distinct subject and from a verified source, the official White House archives [15]. Additionally, while training our classifier we included random speech data ($\approx 40\%$ of sample dataset) from other political figures from both parties equally to help reduce overfitting biases between political party nuances in policy and speech. In total, we processed $\approx 3.5M$ tokens of text data across all subjects (random and targeted).

Subject Data	# of Tokens
Trump	$\approx 1.2M$
Biden	$\approx 1M$
Random	$\approx 1.3M$

Table 1: Amount tokenized data by subject

Model Dataset
To train our model, we compiled a dataset of ≈ 7000 text responses for each subject respectively. Importantly, this dataset was sources separatly from the classifer data. This data was retrieved from FactBase [16, 17], a database platform that provides access to publicly available information on people and entities, which is useful for gathering speech data for Biden and Trump because it archives and analyzes his speeches and statements. The format of this speech data fluctuates between short form tweets and long form speeches. Thus, the average text was ≈ 120 tokens, in total amounting to $\approx 800k$ of token processed per subject for training.

Subject Data	# of Tokens
Trump	$\approx 724k$
Biden	$\approx 816k$

Table 2: Amount tokenized data by subject

5.2 Evaluation methods

Evaluting the quality of style transfer poses a significant difficulty. Unlike in machine translation and summarization, where BLEU [18] and ROUGE [19] metrics are commonly used to measure the similarity between model outputs and the reference text, style transfer lacks parallel data that can serve as ground truth references for evaluation. In table 3 we outline our techniques for 'automatic' evaluation commonly used in style transfer research as stated above in our related work section. Additionally, we use human assessment as a 'manual' metric .

Criterion	Evaluation Technique
Transfer Strength	Accuracy from our trained style classifier
Semantic Preservation	BLEU score
Subjective Style Ranking	Human Evaluation

Table 3: Overview of our evaluation methods

Transfer Strength: *Custom Classifer*
As described above in subsection 3.2 transfer strength refers to the ability of a model to successfully transfer the style from the source to the target. To measure the strength of the transferred style, we built a classifier to identify the attributes of the generated samples and assess whether the model has successfully transferred the target attribute to each sample. For the classifier, we use Huggingface's Transformers library to adapt DistillBERT [20] with a sequence classification head. DistilBERT is a compressed version of the pre-trained BERT (Bidirectional Encoder Representations from Transformers) language model. DistilBERT is created by applying a distillation process to the original BERT model. This process involves training a smaller model to mimic the behavior of the larger model, by matching the output distributions of the hidden states of the two models.

Our evaluation method works as follows: For each generated response, the function first tokenizes the response using the DistilBERT tokenizer. Then, the tokenized input is passed to the pre-trained DistilBERT model, and the output logits are obtained. The predicted class id is obtained by selecting the maximum value from the logits. Finally, the predicted class id is appended to the scores list. The function returns the average score obtained by dividing the sum of scores by the number of scores. The score represents the transfer strength of the generated responses for the Biden and Trump styles, with a higher score indicating a stronger transfer of the respective style.

Semantic Preservation: ***BLEU Score***
BLEU (Bilingual Evaluation Understudy) score, in the context of style transfer, is a metric for evaluating semantic preservation by measuring the similarity between the generated text and the reference text. Our approach works as follows: given two lists of strings, generated responses and actual responses we compute the BLEU score for each generated response in generated responses against its corresponding actual response in actual responses, and return the average BLEU score for all responses. The BLEU score formula for this task is as follows:

$$BLEU = BP * \exp(\sum_{n=1}^{N} w_n * \log(p_n)) \tag{6}$$

where BP is the brevity penalty, which adjusts the score based on the length of the generated text and the reference text, w_n is the weight for n-gram precision, p_n is the n-gram precision score, and N is the maximum n-gram order. To compute the n-gram precision score, the generated text and the reference text are divided into n-grams, and the number of n-grams that appear in both texts is counted. The n-gram precision score is then calculated as the ratio of the total number of matching n-grams to the total number of n-grams in the generated text.

We use the NLTK library [21] to compute the BLEU score by using the sentence_bleu function from the nltk.translate.bleu_score module. This function takes two arguments: a list of reference sentences and a generated sentence. The reference sentences are passed as a list of lists of tokens, where each inner list represents a single reference sentence, and each token is a string. The generated sentence is passed as a list of tokens, where each token is a string. The sentence_bleu function computes the BLEU score between the generated sentence and the reference sentences, and returns a score between 0 and 1. Notably, a higher BLEU score indicates that the generated text preserves more of the semantic content of the input text.

Subjective Style Ranking: ***Human Evaluation***
We had eight 19-22 year old male Stanford students rank five generated text samples from our model to measure how well each subject's style was emulated. The students were shown five *real* text samples spoken from each subject to use as reference when they ranked how well they felt style transferred through in the generated text samples. They ranked each sample from 1-10, with ten being a perfect style emulation and one being no style emulated. We use this evaluation metric because it provides a non-computational approach to style transfer, which is a subjective matter and humans will be subjective in their ranking.

5.3 Experimental details

Our model config loads a pre-trained LLaMA language model and tokenizer from the Hugging Face Transformers library, and fine-tunes it using PEFT. The LLaMA model is a large-scale transformer-based language model that has been trained on a diverse range of language tasks, and can be fine-tuned for various downstream tasks, including style transfer. The code also includes settings for 8-bit integer quantization and half-precision floating point computations, which can improve the model's inference speed and memory usage.

In our study, we aimed to train language models on a three architectures: the base 7 billion parameter LLaMA model, the LoRA finetuned model, and Prefix Tuned model. We implemented the training using PyTorch and the Transformers library. We added several command-line arguments to allow flexibility in the model. For instance, the –model_type command line argument specifies the type of model to train, and the –subject argument specifies the writing style to emulate (Biden or Trump). If the model type is not the base LLaMA model, the –subject argument must be specified. The –from_epoch and –n_epochs arguments are optional and specify the epoch to start training from and the number of epochs to train for, respectively. If not specified, the default values are used. Additionally, we used the load_dataset() function from the Hugging Face Datasets library to load the training and validation data. After loading the data, we applied a tokenize function to encode the input data into numerical tensors, which are then used for training. The model-specific training functions are train_lora() and train_prefix_tuning(). The former trains a LoRA model, and the latter trains a PrefixTuning model. The training configurations include the batch sizes, number of epochs, learning rate, and maximum length of the input sequence. The training script uses PyTorch's Trainer class to train the model, with the DataCollatorForLanguageModeling class used to collate the tokenized input data. The script also uses the prepare_model_for_int8_training() function from the peft library to prepare the LoRA model for integer-8 training, which dramatically increases training and compute speeds. Specifically, we trained the LoRA and Prefix tuning layers with the same parameter configuration below.

batch size = 128,
micro-batch size = 2,
gradient accumulation steps = 64,
learning rate = 3e-4,
maximum input sequence length = 512 tokens,
LoRA configuration of r=8, alpha=16, and dropout rate = 0.05

The training script trains the model training, logging the loss and other metrics every 25 steps, evaluating the model every 25 steps, and saving the model every 25 steps. We trained each model for 60 minutes on an Nvidia A10G, limited by compute resources, this equated to less than 1 epoch for each training run. The training session completed 33 steps during each training run, and the Prefix Tuning training session completed 800 steps during each training run. We then used the trained models to generate the predicted next line of content for each context in the dataset and performed our analysis on the predicted next line and the true next line.

5.4 Results

Table 4: Model Comparison and Evaluation Results

Model Type	Subject	Transfer Strength	Avg. BLEU Score	Avg. Human Ranking
Base	Biden	0.8352	0.0002	1.000
Base	Trump	0.8638	0.0001	1.375
Prefix Tuning	Biden	0.8331	0.0002	1.750
Prefix Tuning	Trump	0.8625	0.0002	2.000
LoRA	Biden	0.8899	0.0002	1.875
LoRA	Trump	0.8939	0.0002	2.250

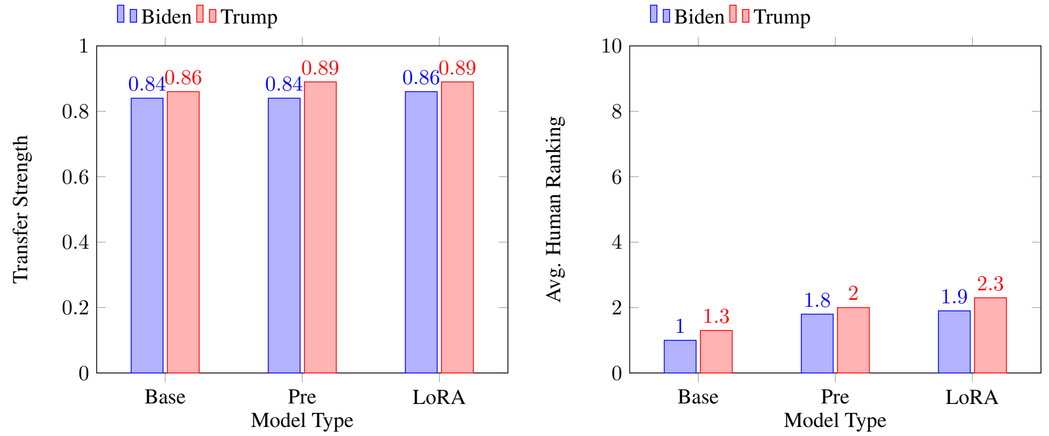


Figure 1: Model comparison between subjects by evaluation methods

6 Analysis

One of the main challenges of evaluation on style is that there is no ground truth. Thus, there needs to be a balance between exact translation and overall tone. Knowing this we chose our evaluation metrics as an optimization between these two variables. Overall, results of this experiment are more indicative of the potential for LoRA models to have success in the area of style emulation, but are inconclusive for Prefix Tuning.

For the transfer strength (classifier) scores we see a slight increase in the positive classification rates from LoRA generations compared with the base model and prefix-embedding. This suggests that with more compute LoRA may be feasible as a solution for low parameter based finetuning. However, our subjective analysis of the generated texts indicate that the classifier on a few samples suggests that text containing subjects and keywords (such as "vaccine" or "White House") that are likely for Biden to say have an extremely high likelihood of a positive classification. As such, it is difficult to verify the merits of our style transfer system solely using the classifier.

Our semantic preservation results were lower than expected, primarily because of the approach we used to calculate the BLEU score. The calculation relied on comparing the n-gram precision averages between the generated and actual responses, which emphasized the micro level of direct word similarity over macro style similarity. Consequently, many of the actual and generated response pairs demonstrated similar meanings, but with different word translations, resulting in consistently low BLEU scores. Therefore, it is challenging to use these largely similar BLEU scores to draw any meaningful conclusions.

After analyzing our human ranking data, we have observed a low average rank for style emulation across models. This outcome was anticipated due to the fact that some of the generated text samples did not accurately represent natural speech. As a result, many evaluations received scores of 0 and 1, with the maximum score of 4 being achieved in one of the Trump samples. Given that style is a subjective metric, a middle ground interpretation between direct translation of speech and tone was provided by this metric, which relied on human subjective ranking. There is a significant amount of bias in the selection of participants and thus the evaluation. Namely, all candidates were young and male, which may have led to a more homogeneous ranking. However, due to external factors and time constraints, we were not able to obtain a large number of human participants to rank the style of our samples. On the other hand, human evaluation is crucial in assessing the quality of text generation, particularly for subjective metrics like style. Automated evaluation metrics can only provide limited information on the quality of the generated text, and they may not capture the nuances and complexities of human language. Human evaluation allows us to gain a deeper understanding of the strengths and weaknesses of text generation models, and it can also help us identify areas for improvement. By incorporating human feedback into the evaluation process, we can create text generation models that better reflect natural language and meet the needs of users.

7 Conclusion

In conclusion, this study explored the potential of Low Rank Adaptation (LoRA) for style transfer in text generation, specifically emulating the writing styles of Joe Biden and Donald Trump. Our findings suggest that LoRA, supported by Parameter-Efficient Fine-Tuning (PEFT), shows promise in generating high-quality text that accurately captures the distinct writing styles of the two politicians. Our model demonstrated competitive performance compared to state-of-the-art models, producing coherent and natural text.

The primary achievements of our work include the novel implementation of style transfer text generation using LoRA and LLaMA transformer. Our evaluation metrics, including transfer strength measured by a custom classifier, semantic preservation assessed by BLEU score, and human evaluation of style ranking, provided valuable insights into the performance of our model.

However, our work has some limitations. The calculated BLEU scores for semantic preservation were lower than expected due to the focus on micro-level word similarity, rather than macro-level style similarity. In addition, the human evaluation was limited by a small, homogenous group of participants, potentially leading to a biased assessment of style transfer quality.

In the future, it would be interesting to evaluate semantic preservation using cosine-similarity on vectors embeddings. Cosine similarity is a measure of similarity between two vectors of an n-dimensional space. When applied to word embeddings, which are vector representations of words in a high-dimensional space, cosine similarity can be used to measure the semantic similarity between words or sentences. In the case of sentence embeddings, each sentence is represented as a high-dimensional vector, typically with hundreds or thousands of dimensions. These vectors are learned using machine learning algorithms such as Word2Vec or GloVe, which are trained on large amounts of text data to capture the semantic relationships between words. To compute the cosine similarity between two sentences, their respective embeddings are first obtained. The embeddings are then normalized to have unit length, which means that each vector has a length of 1 in the high-dimensional space. The cosine similarity between the two normalized vectors is then calculated using the dot product of the vectors. Moreover, expanding and diversifying the pool of human evaluators could result in a more comprehensive and unbiased assessment of style transfer performance. Additionally, further investigation into other model architectures and fine-tuning techniques could lead to improvements in style transfer text generation, ultimately contributing to the development of more effective and robust models for diverse applications.

References

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[2] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021.

[3] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[4] Bernardino Romera-Paredes and Philip Torr. An embarrassingly simple approach to zero-shot learning. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 2152–2161, Lille, France, 07–09 Jul 2015. PMLR.

[5] Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, and Sayak Paul. Peft: State-of-the-art parameter-efficient fine-tuning methods. `https://github.com/huggingface/peft`, 2022.

[6] Xiang Lisa Li and Percy Liang. Prefix-tuning: Optimizing continuous prompts for generation, 2021.

[7] Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. Shakespearizing modern language using copy-enriched sequence-to-sequence models. 07 2017.

[8] Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. Pointer networks, 2015.

[9] Felix Hill, Kyunghyun Cho, and Anna Korhonen. Learning distributed representations of sentences from unlabelled data. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1367–1377, San Diego, California, June 2016. Association for Computational Linguistics.

[10] S. Kullback and R. A. Leibler. On information and sufficiency. *Ann. Math. Statist.*, 22(1):79–86, 1951.

[11] Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan. Style transfer in text: Exploration and evaluation, 2017.

[12] Juncen Li, Robin Jia, He He, and Percy Liang. 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, New Orleans, Louisiana, June 2018. Association for Computational Linguistics.

[13] Di Jin, Zhijing Jin, Zhiting Hu, Olga Vechtomova, and Rada Mihalcea. Deep learning for text style transfer: A survey. *CoRR*, abs/2011.00416, 2020.

[14] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization, 2017.

[15] The United States Government. https://www.whitehouse.gov/, 2023.

[16] Factba.se. Factba.se biden database. `https://factba.se/biden/`, 2023. Accessed: March 20, 2023.

[17] Factba.se. Factba.se trump database. `https://factba.se/trump/`, 2023. Accessed: March 20, 2023.

[18] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA, 2002. Association for Computational Linguistics.

[19] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Annual Meeting of the Association for Computational Linguistics*, 2004.

[20] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108, 2019.

[21] Edward Loper and Steven Bird. Nltk: The natural language toolkit, 2002.

[22] Jessica Ficler and Yoav Goldberg. Controlling linguistic style aspects in neural language generation. In *Proceedings of the Workshop on Stylistic Variation*, pages 94–104, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

[23] Mariano de Rivero, Cristhiam Tirado, and Willy Ugarte. Formalstyler: Gpt based model for formal style transfer based on formality and meaning preservation. In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management - Volume 1: KDIR,*, pages 48–56. INSTICC, SciTePress, 2021.

A Appendix (optional)

Evaluation methods Influence:
Jin et al. (2021) [13],

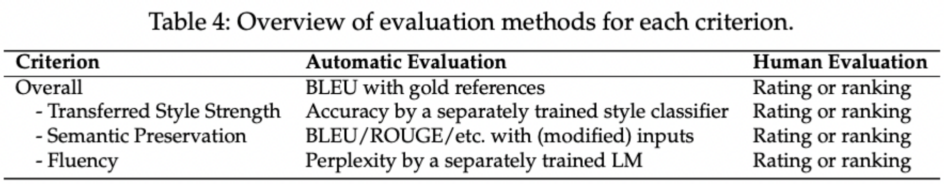
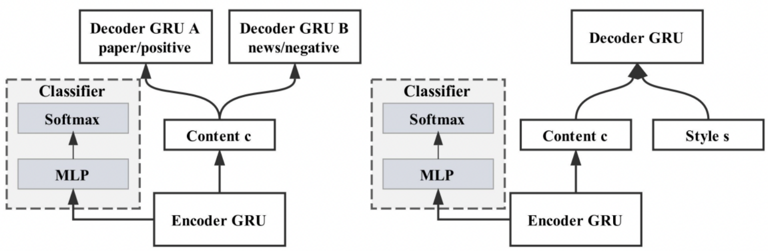


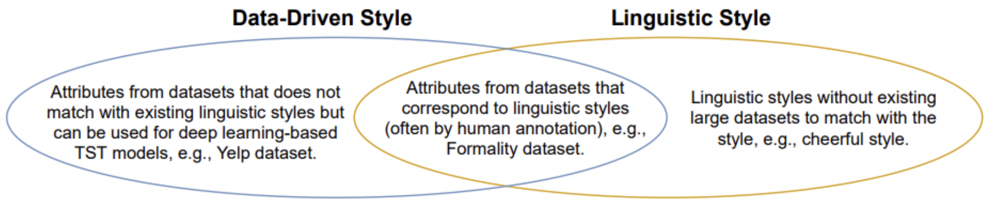
Table 4: Overview of evaluation methods for each criterion.

Criterion	Automatic Evaluation	Human Evaluation
Overall	BLEU with gold references	Rating or ranking
- Transferred Style Strength	Accuracy by a separately trained style classifier	Rating or ranking
- Semantic Preservation	BLEU/ROUGE/etc. with (modified) inputs	Rating or ranking
- Fluency	Perplexity by a separately trained LM	Rating or ranking

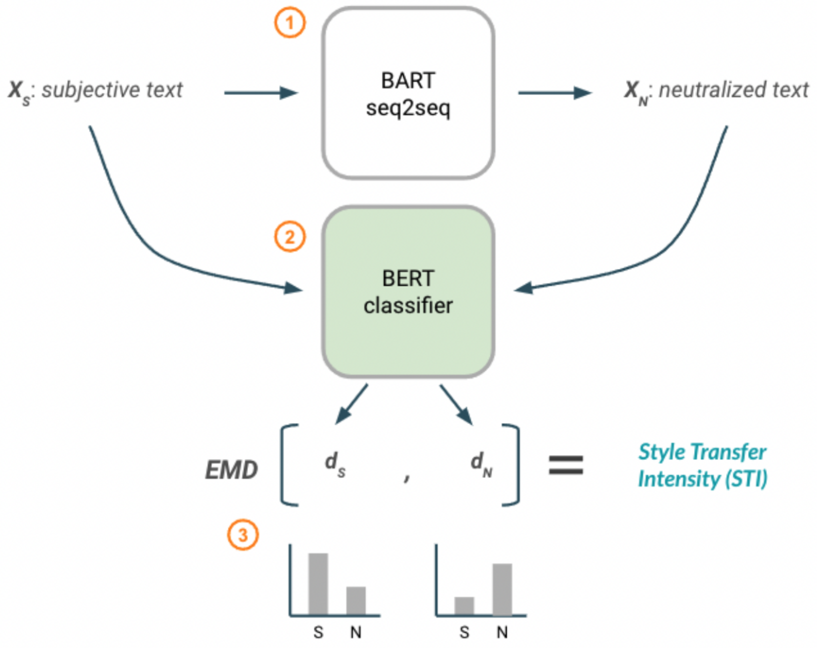
Embedding and multi-decoder graph:
Fu et al. (2018) [11]



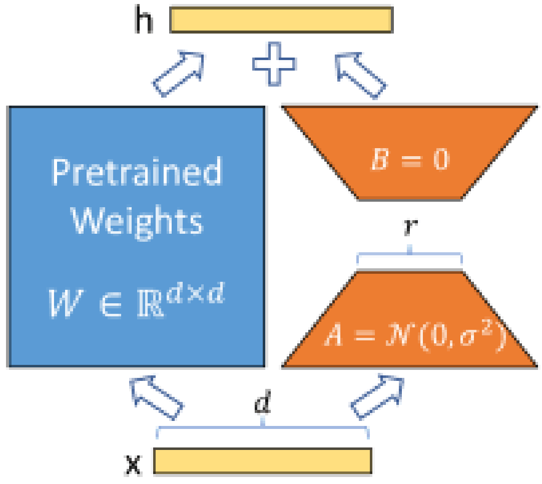
Style interpretation:
We evaluate style using the following categorization table from Jin et al. (2021) [13],


Data-Driven Style
Linguistic Style
Attributes from datasets that does not match with existing linguistic styles but can be used for deep learning-based TST models, e.g., Yelp dataset.
Attributes from datasets that correspond to linguistic styles (often by human annotation), e.g., Formality dataset.
Linguistic styles without existing large datasets to match with the style, e.g., cheerful style.

Automatic evaluation inspiration



LoRA architecture
Hu et al. (2021) [2]


h
Pretrained Weights
$W \in \mathbb{R}^{d \times d}$
$B = 0$
r
$A = \mathcal{N}(0, \sigma^2)$
d
x

Other related works:

Conditioned Recurrent Neural Network Model

A Conditioned Recurrent Neural Network (CRNN) in the context of style emulation works by conditioning the generation of language on a learned representation of the desired style, in addition to the input text. In the case of Ficler and Goldberg (2017) [22], the CRNN is used to generate text that emulates a specific linguistic style by training the network on examples of text in that style and conditioning it to generate text with similar characteristics.

Generative Pre-trained Transformer Models

Rivero et al. (2021) [23] conducted research on Generative Pre-trained Transformer (GPT) based models for text style transfer with a focus on formal style. They fine-tuned the GPT-2 model on a parallel corpus of informal and formal text entries and evaluated the results using metrics for formality and meaning preservation. To further improve the model, they integrated a human-based feedback system where the user selects the best formal sentence from those generated by the model. The evaluations showed that their solution achieved similar or improved scores in formality and meaning preservation compared to state-of-the-art approaches. This research highlights the potential of GPT-based models for text style transfer and the importance of incorporating human feedback in the fine-tuning process to achieve better results.

Table Framework

Topic	Response	Label
"Star Trek"	"I love Star Trek. I love the original series."	Positive
"New HBO Series"	"I think it's a great series."	Positive
"color"	"I don't have a favorite color"	Negative

Table 5: Sample zero-shot topics, responses, and classifier labels for Joe Biden

Sample Raw Speech Data for Classifier

> PRESIDENT BIDEN: How are you, fellas? (Applause.) Please. Thank you. Thank you, thank you, thank you. You're very nice. Thank you. Thank you, thank you, thank you. Please. Please, sit. I don't know who Ali was talking about — someone being around that long. (Laughter.) Twenty years. Whoa. You know, when we voted to set you all up, no one ever thought it'd be over 260,000 people. No one thought it would take — but you've done an incredible job. Mr. Secretary, I came for a simple reason: to say thank you. Thank you for welcoming me back here today. As well as 20 years. Mr. Secretary, you're a true patriot who has dedicated your career to protecting and serving the American people while upholding our nation's laws and standing up for American values. Because when I think of this department, I think of it in that way: standing up for American values. I'm so grateful for everything, Mr. Secretary, you've done and all you do for our country and for the whole world. You know, every 9/11, I think about a friend of mine named Davis Sezna, from Delaware. The year before 9/11, I was watching television in my — it was in the summer.
> ...
> **300 LINES**
> ...
> Because we need them. We need them badly. Again, thank you, thank you, thank you. (Applause.) 12:42 P.M. EST

Sample data Format

"context": "Interviewer: That’s right. Interviewer: New cases of COVID-19, hospitalizations have fallen by half in the last month, so have new cases, that’s the good news. There’s this potential threat, potential surge from the variants coming down the pike potentially. When is every American who wants it going to be able to get a vaccine?",
"subject": "Joe Biden",
"response": "By the end of July this year. We have, we came into office, there was only 50 million doses that were available. We have now, by the end of July we’ll have over 600 million doses, enough to vaccinate every single American."

Sample BLEU Score evaluation input

```
"generated_responses": [
    "’I agree with you. I think that the vaccine is safe. I think that
    the vaccine is effective. I think that the vaccine is necessary. I
    think that the vaccine is a good thing. I think that the vaccine is
    a good thing for the country. I think that the vaccine is a good
    thing for the world. I think that the vaccine is a good thing for
    the planet. I think that the vacc",
    "### Speaker:\nPersonB: ’Joe Biden: And the thing I do in terms of
    student debt that’s accumulated is provide for changing the existing
    system now for debt forgiveness if you engage in volunteer activity.
    For example, if you were -- if you’re teaching school, after five
    years, you would have $50,000 of your debt forgiven. If you worked in
    a battered women’s",
    "Joe Biden: ’Joe Biden: I’m Joe Biden.\nJoe Biden: I’m Joe Biden.
    \nJoe Biden: I’m Joe Biden.\nJoe Biden: I’m Joe Biden.\nJoe Biden:
    I’m Joe Biden.\nJoe Biden: I’m Joe Biden.\nJoe Biden: I’m Joe Biden.
    \nJoe",
    "### Speaker:\nJoe Biden: ’\n\n### Speaker:\nJoe Biden: ’\n\n###
    Speaker:\nJoe Biden: ’\n\n### Speaker:\nJoe Biden: ’\n\n### Speaker:
    \nJoe Biden: ’\n\n### Speaker:\nJoe Biden: ’\n\n### Speaker:\nJoe
    Biden: ’",
    "’Joe Biden: I’m going to continue to do that. I’m going to continue
    to do that. I’m going to continue to do that. I’m going to continue
    to do that. I’m going to continue to do that. I’m going to continue
    to do that. I’m going to continue to do that. I’m going to continue
    to do that. I’m going to continue to do that. I’m",
    "### Speaker:", "’Joe Biden: I’m going to do everything I can to make
    sure that the vaccine is available to every American.\nJoe Biden:
    I’m going to do everything I can to make sure that the vaccine is
    available to every American.\nJoe Biden: I’m going to do
    everything I can to make sure that  the vaccine is available to every
    American.\nJoe Biden: I’m going to do", "’Joe Biden: I’m going to
    continue to do that. I’m going to continue to do that. I’m going to
    continue to do that. I’m going to continue to do that. I’m going to
    continue to do that. I’m going to continue to do that. I’m going to
    continue to do that. I’m going to continue to do that. I’m going to
    continue to do that. I’m"
],
"actual_responses":[
    "But I promise you, that’s why, for example, I provided for community
    health centers to have direct access federally to vaccinations.
    That’s why I provided for literally thousands more vaccinators. We
    had to increase the supply of vaccines. Plus, we had to get a lot of
    people who were not available up to now to come and put the
    vaccination in someone’s arm.",
    "My other son went to Georgetown and Yale Law School, graduated with
    $142,000 in debt. And he worked for a parking service in -- down in
    Washington. My daughter went to Tulane University and then got a
    master’s at Penn. She graduated with $103,000 in debt. So, I don’t
    think anybody should have to pay for that, but I do think you should
    be able to work it off.",
    "How you doing?",
    "My mother would say, \"God bless you, son, no purgatory for you,\"
    four kids home, I really mean it. And by the way, the loss of being
    able to be in school is having significant impact on the children and
    parents as well. And what we found out is there are certain things
    that make it rational and easy to go back to the brick- and-mortar
    building.",
    "And I prioritized that issue from day one, calling for the passage
    of John Lewis Voting Rights Act and the Freedom to Vote Act. I’m
    doing everything I can on a priority, get that done --",
    "Absolutely. That answer’s absolutely, but guess what? We need the
    same votes we need to overrule Do -- to, uh, reinstate, uh, the, uh,
    thedecision that was struck down by the court. I mean, I do support
    that,and I’ve urged, publicly urged companies to do that. I’ve urged
    them publicly, as President of the United States, saying, \"This is
    what you should be doing.",
    "And we’re focusing on communities that have been left behind,
    including rural communities, as well as inner-city and communities
    that relate thatdon’t have.  Look, if you’re a 65-year-old woman
    living alone and you want to get in line to get a vaccine at the
    local pharmacy, you may be three miles away and have no come, no
    way to get there.",
    "And I prioritized that issue from day one, calling for the passage
    of John Lewis Voting Rights Act and the Freedom to Vote Act. I’m
    doing everything I can on a priority, get that done --"
]
```