RoSA Text Style Transfer & Evaluation

Stanford CS224N Custom Project

Ayaan Naveed Malik
Dept. of Computer Science
Stanford University
ayaan04@stanford.edu

Arnav Gupta
Dept. of Computer Science
Stanford University
navgup@stanford.edu

MacVincent Agha-Oko
Dept. of Computer Science
Stanford University
maghaoko@stanford.edu

Abstract

Given the large size of modern LLMs, researchers have focused on creating parameter-efficient fine-tuning (PEFT) techniques that allow models to perform specific tasks at a much lower compute and time cost. Low-rank adaptation (LoRA) has been a widely popular PEFT technique. However, a recently released method, Robust Adaptation (RoSA) promises better performance on complex tasks relative to LoRA in resource-contained conditions. This study analyzes the performance of these PEFT techniques on a difficult task, text style transfer (TST) from modern English to Shakespearean style, by finetuning a Google Multilingual T5 (mT5) base model using fine-tuning methods like RoSA and LoRA, as well as zero-shot and few-shot prompting. Our studies demonstrate the efficacy of the RoSA method on the Shakespearean TST task when compared to other model adaptation methods.

1 Key Information to include

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

Team Contributions:

- **Ayaan:** Devised & implemented evaluation pipeline & fine-tuned and ran experiments on all LoRA & RoSA models for Shakespearean TST. Completed sections of report.
- **Arnav:** Created and trained TST classifier, debugged model code. Completed sections of report. Created poster and ethics statement.
- **MacVincent:** Implemented evaluation pipeline for model output. Trained Quantized LoRA. Assisted with the training of models for fine-tuning tasks. Completed sections of the report.

2 Introduction

Text-style transfer (TST), is a family of problems focused on transforming the style or appearance of text while preserving its content (Shen et al., 2017). Examples of TST problems include sentiment transfer (translating a positive sentence to a negative one) or deciphering text. LLMs have shown impressive performance on TST tasks: but modern LLMs like GPT-4 are incredibly large, making full finetuning (FFT) prohibitively time and cost-intensive, necessitating the use of parameter-efficient finetuning methods (PEFT), like Low-Rank Adaptation (LoRA). While LoRA is many times faster than FFT, it generally fails to achieve similar performance on more complex tasks (Hu et al., 2021). RoSA is a recently-released PEFT method that has shown much-improved performance over techniques like LoRA in complex generation tasks while still remaining efficient (Nikdan et al., 2024). However, RoSA has not been tested outside of the original paper and its downsides are not apparent, so RoSA's performance on many tasks is still unclear. In an attempt to answer this question, our project evaluates RoSA's performance on the task of TST and compares it to LoRA.

Furthermore, evaluating TST is itself an area of research – recent work Ostheimer et al. (2023) shows that LLMs may be an effective evaluation method. We explored this idea in this paper and proposed a novel technique utilizing LLM evaluation intertwined with generic metrics and trained classifier models.

3 Related Work

Existing research on TST shows a wide range of neural models have been deployed on these problems: one early paper, Jhamtani et al. (2017) used a bidrectional LSTM to translate modern English sentences into Shakespearian English (a task that we tested our model on as well). LLMs have, of course, been used for this task as well, but much of the existing work like Reif et al. (2022) and Pan et al. (2024) focuses on prompt-based approaches (few and zero-shot learning) and do not use any form of finetuning for this task.

There are also a diversity of methods used to evaluate TST. Semantic evaluation metrics like BLEU and BERTScore are inexpensive, but focus purely on semantic meaning (Zhang et al., 2020). Transfer strength, developed by Fu et al. in 2018 specifically for TST is based on training a classifier model to evaluate what proportion of sentences were of the desired style. Recent research has shown that LLMs with few-shot learning are also highly accurate at evaluating text style transfer (Ostheimer et al., 2023). In our experiments, we deployed a number of these metrics.

4 Approach

As explained previously, training large language models (LLMs) is memory and computationally expensive. Improving their performance for particular tasks can be done by training on the pre-trained model weight through full finetuning (FFT). However, this is memory and computationally just as prohibitive. In this section, we describe two major parameter-efficient fine-tuning (PEFT) methods we deployed to the style transfer task to approximate FFT updates even when training with limited data. Specifically, we discuss Low-Rank Adaptation (LoRA) and Robust Adaptation (RoSA). We describe the strengths and drawbacks of each of these approaches as well as prompting and quantization approaches to style transfer.

Low-Rank Adaptation (LoRA): LoRA is based on the idea that a model's learnings can be expressed on a low intrinsic dimension. The changes in a model's weights during finetuning can also be encapsulated in such a low-rank dimension. In LoRA, dense layers are frozen and the model is adapted by optimizing the rank-decomposition matrices associated with each dense layer. For a frozen dense layer $W_0 \in \mathbb{R}^{d\times k}$ we can represent its updates as $W_0 + \Delta = W_0 + AB$ where $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times k}$, and rank $r \ll \min(d, k)$. B is initialized using Gaussian initialization while A is initialized to zero. This ensures that the initial value of the adaptation is zero. For an input matrix X the output of the dense layer after adaptation becomes:

$$O = W_0X + \Delta X = W_0X + A(BX) \tag{1}$$

LoRA modules can also be merged with the original model's weight. This reduces the need for extra steps during inference. While LoRA-style methods work well in practice, they fail to achieve the accuracy levels of FFT on tasks with complex targets like code generation or mathematical reasoning, and we expected that trend to hold for TST (Hu et al., 2021).

Robust Adaptation (RoSA): RoSA (Nikdan et al., 2024) fixes the complex task performance problem by proposing a method that combines low-rank approximations, sparse matrices, and quantization to match or surpass the performance of FFT methods on complex tasks while maintaining the computational efficiency of LoRA-based approaches. The RoSA system, depicted in Figure 1 consists of a sparse adapter and a low-rank adapter. Typical LoRA approaches apply the low-rank adapter when approximating FFT updates. The issue with the low-rank adapter is that it fails to adequately represent the outlier components needed by the LLMs to support complex task targets. To fix this, RoSA utilizes sparse matrixes to obtain this representation. For a frozen dense layer $W_0 \in \mathbb{R}^{d\times k}$ we can represent its updates as:

$$W_0 + \Delta^S + \Delta^L = W_0 + \Delta^S + AB \tag{2}$$

Where $A \in \mathbb{R}^{d\times r}$, $B \in \mathbb{R}^{r\times k}$, $\Delta^S \in \mathbb{R}^{d\times k}$, and rank $r \ll \min(d, k)$.

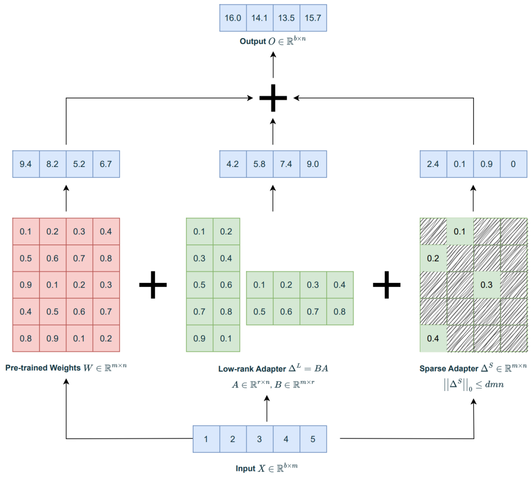

16.0 14.1 13.5 15.7
Output $O \in \mathbb{R}^{b\times n}$
9.4 8.2 5.2 6.7
4.2 5.8 7.4 9.0
2.4 0.1 0.9 0
0.1 0.2 0.3 0.4
0.5 0.6 0.7 0.8
0.9 0.1 0.2 0.3
0.4 0.5 0.6 0.7
0.8 0.9 0.1 0.2
0.1 0.2
0.3 0.4
0.5 0.6
0.7 0.8
0.9 0.1
0.1 0.2 0.3 0.4
0.5 0.6 0.7 0.8
0.1
0.2
0.3
0.4
Pre-trained Weights $W \in \mathbb{R}^{m\times n}$
Low-rank Adapter $\Delta^L = BA$
$A \in \mathbb{R}^{r\times n}, B \in \mathbb{R}^{m\times r}$
Sparse Adapter $\Delta^S \in \mathbb{R}^{m\times n}$
$||\Delta^S||_0 \leq dmn$
1 2 3 4 5
Input $X \in \mathbb{R}^{b\times m}$

Figure 1: Applying RoSA to a Fully Connected Layer (Nikdan et al., 2024)

For an input matrix X the output of the dense layer after adaptation becomes:

$$O = (W_0 + \Delta^S)X + \Delta^L X = (W_0 + \Delta^S)X + A^L(B^L X) \tag{3}$$

The sparse matrix adapter is based on the intuition that the sum of all FFT gradient updates can be viewed as a robust PCA problem. The RoSA method identifies a highly performant sparsity mask, provides GPU kernels to support efficient sparse backward pass, and a mechanism that guarantees the parallel training and convergence of both the sparse and low-rank adapters. Finally, they demonstrate that RoSA closes the accuracy gap between full finetuning and adaptation methods on complex tasks.

Quantized Low-Rank Adaptation (QLoRA): A major drawback of the LoRA and particularly the RoSA method discussed above is that they still require significant compute power to hold the adaptation modules and this increases linearly with the number of layers adapted. QLoRA introduced in Dettmers et al. (2023) utilized page optimizers, quantizing of quantization constants to reduce overall memory footprint, and a new datatype they found to be optimal for representing normally distributed weights. The end result was a method that approached the performance of full-finetuning for a fraction of the memory costs. We will be applying these quantization techniques in our paper.

Zero Shot Prompting: In zero-shot prompting, we adopt a language model to make predictions using input data for which the model was not explicitly trained. When applying this method to the style transfer task, we design prompts that instruct the model to modify an input sentence in a way most similar to the new prompt. An example of a sample prompt is:

```
This is an input text in modern English: Loving yourself, my king, isn't as
      bad as neglecting yourself. Modify this to sound Shakespearean:
```

The advantage of this method is that the model can be generalized in resource-constrained situations where we lack the data or compute needed to finetune the model. We can also apply the same model to a variety of tasks by updating the input prompt. The drawback, however, is that performance may be limited since output is heavily influenced by the biases of the pre-trained model and as such would not approach of the performance of models fine-tuned on task-specific data.

Few Shot Prompting: Few-shot prompting offers more of a middle ground. In this method, we still adopt a language model to make predictions using input data for which the model was not explicitly trained. However, rather than designing a prompt that contains only transfer instructions, we also include a few examples of input sentences and the equivalent output with style transferred. An example of a few-shot prompt is:

```
    incorrect: talk tuoba listening to the elpoep
    correct: talk about listening to the
    correct: horrible
    incorrect: unfortunately that ’s tuoba erehw things went
```

With more guidance, the model can improve its performance relative to the zero-shot prompting. However, we still retain the generalizable ability of zero-shot in resource-constrained situations where we lack the data or compute needed to finetune the full model. Regardless, performance may still be limited when compared to a fully finetuned model. We verify this claim through experiments.

5 Experiments

5.1 Data

We used a dataset of plain English sentences translated into Shakespearean-styled text curated in Jhamtani et al. (2017) for our experiments. The train set consisted of 18395 sentence pairs, the validation set consisted of 1218 sentence pairs, and the test set consisted of 1462 sentence pairs. Examples of sentence pairs in our dataset include:

Modern English	**Shakespearean Style**
Oh , poor Romeo !	Alas , poor Romeo !
A jumbled confession can only receive a jumbled absolution .	Riddling confession finds but riddling shrift .
There’s still a stain on your cheek from an old tear that hasn’t been washed off yet .	Lo , here upon thy cheek the stain doth sit Of an old tear that is not washed off yet .

5.2 Evaluation method

A core challenge in the field of text style transfer lies in establishing robust evaluation methods. Traditional metrics such as BLEU, ROUGE, and METEOR are often inadequate for this domain. Style encompasses a range of linguistic elements beyond semantic meaning, including formality, sentiment, lexical choice, and syntactic complexity. Therefore, our approach is to divide the evaluation task into smaller, focused evaluation tasks. We will be evaluating the following tasks:

Style Classifier Accuracy: We evaluated the output of each model against a finetuned DistilBERT Sanh et al. (2020) classifier hosted on the HuggingFace transformers library Wolf et al. (2020) and trained to produce classification scores indicating whether a sentence follows the Shakespeare style or modern English. The scores from this classifier are in the range $[0, 1]$. The higher the score, the more likely our sentence is Shakespearean.

Semantic Preservation: For semantic preservation, we utilize **Word Mover’s Distance (WMD)**. Unlike traditional similarity measures, WMD utilizes word embeddings to capture nuanced relationships between words by mapping them into a continuous vector space where semantically similar words are in closer proximity. WMD considers the semantic meaning encoded within word embeddings, allowing it to detect similarity even when synonyms or different word forms are used. Unlike metrics that rely on strict word order, WMD is less sensitive to changes in word sequence. This is valuable in style transfer, where the goal is often to alter stylistic elements while preserving sentiment.

Let two text documents be represented as:

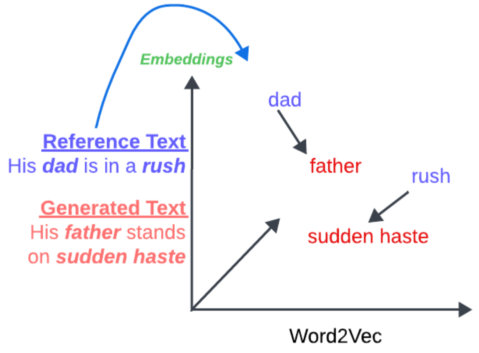
$$\text{Document A: } \{w_{a_1}, w_{a_2}, ..., w_{a_n}\} \quad ; \quad \text{Document B: } \{w_{b_1}, w_{b_2}, ..., w_{b_m}\}$$

WMD conceptualizes this as an optimization problem derived from the Earth Mover’s Distance.

1. **Word Embeddings:** Each word w_i is mapped to its corresponding word embedding vector $\mathbf{x}_i$.

2. **Transportation Matrix:** We define a transportation matrix $\mathbf{T} \in \mathbb{R}^{n \times m}$, where $\mathbf{T}_{ij}$ represents the amount of "word mass" transported from word w_{a_i} in document A to word w_{b_j} in document B.

3. **Distance Calculation:** WMD formulates the distance as:



$$WMD(A,B) = \min_{\mathbf{T} \geq 0} \sum_{i=1}^{n} \sum_{j=1}^{m} \mathbf{T}_{ij} \|\mathbf{x}_{a_i} - \mathbf{x}_{b_j}\|_2$$

where $\|\cdot\|_2$ denotes the Euclidean distance.

For our implementation, we employed the pre-trained 'google.bin.gz' word2vec model, providing rich word vector representations. It includes word vectors for a vocabulary of 3 million words, phrases that they trained on 100 billion words from a Google News dataset, and has a vector length of 300 features. Texts were then tokenized using NLTK's 'word_tokenize' function. Lastly, we removed common stop words as well as a custom list of Shakespearean-era stop words to focus on semantically meaningful terms. (This step increased accuracy by 73% against human evaluation for a small test dataset.)

Secondly, we utilize **BERTScore**. BERTScore quantifies the degree of semantic similarity between a reference BERT sentence embedding (x) and a candidate sentence embedding ($\hat{x}$) by generating contextual word embeddings for both the reference and candidate sentences. These high-dimensional embeddings encapsulate a word's meaning informed by its surrounding context within the sentence. BERTScore then computes the cosine similarity between the embeddings of words in the reference sentence and the candidate sentence. The cosine similarity **CosSim**$_{BERT}$ between the two BERT sentence embeddings $\hat{x}$ and x, which we report in our evaluation is given by:

$$CosSim_{BERT}(x, \hat{x}) = \frac{x \cdot \hat{x}}{||x||||\hat{x}||}$$

BERTScore consists of three metrics: **Recall, Precision, and F1 Score**. Recall measures the extent to which words in the reference sentence have a strong semantic match in the output sentence. Precision indicates the proportion of words in the output sentence that have synonyms in the reference sentence. F1 Score provides a balanced overall similarity metric by computing the harmonic mean of recall and precision. We report the F1 scores in our experiments:

$$R_{BERT} = \frac{1}{|x|} \sum_{x_i \in x} \max_{\hat{x}_j \in \hat{x}} x_i^T \hat{x}_j \quad ; \quad P_{\text{BERT}} = \frac{1}{|\hat{x}|} \sum_{\hat{x}_j \in \hat{x}} \max_{x_i \in x} x_i^T \hat{x}_j$$

$$F1_{BERT} = \frac{2 \cdot (R_{BERT} \cdot P_{BERT})}{R_{BERT} + P_{BERT}}$$

Holistic Evaluation: To address the limitations of individual metrics, we developed a holistic evaluation framework inspired by the principles of Retrieval-Augmented Generation (RAG) models. RAG models combine the strengths of information retrieval and generative language modeling. RAG retrieves relevant documents to the prompt and conditions the generation process on this retrieved-context, providing a richer informational basis for their outputs. We adapt this concept by providing LLMs with:

1. **Contextualization:** A detailed prompt establishes the task of text style transfer, the criteria for evaluation (style transfer accuracy, semantic preservation, fluency, etc.), and explicitly emphasizes the potential limitations of individual metrics.

2. **Diverse Metrics:** Computed metric values for WMD, BERTSCORE, Perplexity and Style Classification on the input-output pairs which LLM's cannot compute.

The LLM is tasked with interpreting these metrics in light of the provided context and generating a holistic score on a scale of 0-100. By employing multiple LLM models, we mitigate potential training biases and obtain a more nuanced evaluation. Particularly, we utilize Mixtral-8x7B, LLaMA-2-7b, and Gemma-7b, and average the three evaluations.

Before running our experiments, we ran an evaluation on our proposed holistic evaluation pipeline. To do so, we developed a list of 100 normal texts and their style-transferred Shakespearean version. These 100 data points contained a range of style transfers to do a thorough evaluation. Examples include:

1. **Poor (0 - 30):** Reflects significant deviations from the intended style or substantial alterations in the text's original sentiment.
2. **Medium (30 - 70):** Demonstrates an attempt at stylistic change and maintains sentiment, but exhibits grammatical flaws.
3. **Good (70 - 100):** Exhibits successful stylistic transformation while preserving sentiment and maintaining grammatical structure.

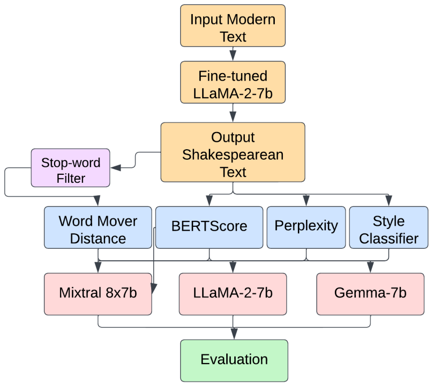


Figure 2: Evaluation Pipeline

We ranked each text of a dataset of 100 sentence pairs on a scale from 0 to 100, reflecting the categories you outlined. We then averaged the score for each sentence pair. To compare the outputs of your automated pipeline with the established human perception, you employed the Mean Squared Error (MSE) metric. The MSE for each input-output pair was calculated as follows:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(\text{human\_score}_i - \text{model\_score}_i)^2 \tag{4}$$

Our evaluation pipeline provides an MSE score of 7.3444, which is competitive to state-of-the-art. It typically fails on very low evaluation scores (1̃5), where it overestimates evaluation scores by an average of 6.3 points.

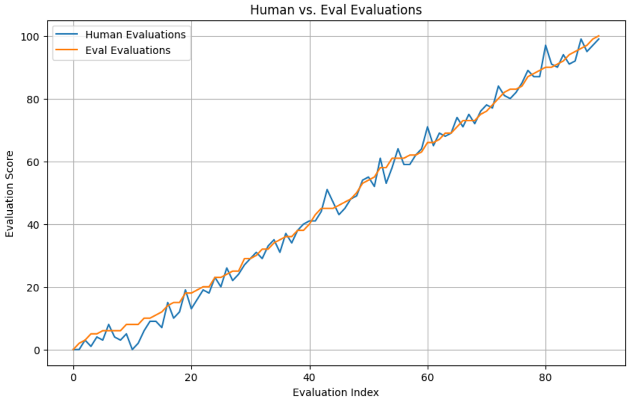


Figure 3: Evaluation Pipeline Results

5.3 Experimental details

Our experiments were run on a variant Google Multilingual T5 (mT5) based model Muennighoff et al. (2022), which is a 580M parameter text-to-text transformer model fine-tuned for following human instructions. We finetuned our RoSA models on a single A100 and fine-tuned our QLoRA and LoRA models on a single V100. We utilized a learning rate of $5e-3$ and a single epoch over all experiments. We varied the rank LoRA rank r for each method and Sparse density d during our experiments to understand their impact on the quality of our fine-tuned output. We also performed experiments on an 8-bit quantized version of our LoRA model to understand the impact of quantization on fine-tuned model output and report those results as well.

5.4 Results

Method	**WMD ↓↓**	**F1**	**CosSim ↑**	**Classifier ↑**	**LLM ↑**
Zero-Shot Learning	**0.7939**	**0.8802**	0.5097	0.1299	39.8
Few-Shot Learning	1.2627	0.8227	0.2081	0.0494	43.2
QLoRA $r = 16$	1.0827	0.8505	0.4690	0.0726	78.4
LoRA $r = 16$	0.9973	0.8547	0.4299	0.2191	83.3
LoRA $r = 8$	0.9967	0.8548	0.4303	0.2190	76.5
LoRA $r = 4$	0.9332	0.8703	0.5344	0.1811	79.8
RoSA $r = 16, d = 0.6\%$	1.0248	0.8512	0.4112	**0.2239**	**85.1**
RoSA $r = 8, d = 0.3\%$	0.9332	0.8512	0.4112	**0.2239**	79.3
RoSA $r = 4, d = 0.15\%$	0.9329	0.8696	**0.5348**	0.2005	81.1

Table 1: Evaluation Scores on the Shakespeare Text Style Transfer (TST) task

We initially set out to perform our experiments using the LLaMA-2-7b model. However, due to budget and compute constraints, we ended up using a much smaller 500M parameter model, Google Multilingual T5 (mT5). This explains our less-than-impressive raw scores. Regardless, this does not hamper our ability to understand the extent to which the various model adaptation methods help improve the performance of a base model on the TST task. Our experimental results in Table 2 above demonstrate the efficacy of the RoSA method on the Shakespearean TST.

Classifier Metric: Relative to the other methods $RoSA\ r = 16$ and $RoSA\ r = 8$ had the highest classifier scores. A high classifier score indicates that these ROSA models are able were able to effectively translate modern English text to its Shakespearean equivalent. However, it says nothing about whether the translated sentence still retains its meaning between styles.

F1 and WMD: Both WMD and F1 are measures of semantic preservation – the similarity of the inputted sentence to the model's output. Unsurprisingly, the zero-shot model performed the best in both these metrics. We suspect that this was because the model did not see any examples of Shakespearan vocabulary, so it did not attempt to "force" those words where they did not belong and thus made less semantic errors. The LoRA and RoSA models performed similarly on these metrics, with both beating out the Few-Shot model.

CosSim Cosine similarity on BERT embeddings on the other hand helps us determine the extent to which semantic meaning is maintained across styles. On this metric another RoSA variant, $RoSA\ r = 16$, $RoSA\ r = 16$ performed best. This once again proves RoSA's capabilities. Zero-shot learning, however, was the second best. This indicates that Zero-Shot Learning produced results extremely similar to the modern English text in terms of semantic meaning but not similar to the expected output in terms of Shakespearean style. Our RoSA scores, however, maintained high semantic meaning across styles which still translated to the expected style as demonstrated by relatively high $CosSim_{BERT}$ and classifier scores.

Quantization: We also explored the degree to which quantization would affect performance. Our results indicate that quantizing our fine-tuning from $LoRA\ r = 16$ in 32-bit down to its 8-bit equivalent in $QLoRA\ r = 16$, there was significant degradation in performance across most metrics. It remains to be seen how much of this degradation would also be observed in a larger model. In a much larger model, the gains in compute and training time may be worth the performance decrease.

LLM Evaluation: Despite wavering performance across our quantitative metrics, RoSA performs the highest amongst all other TST approaches. We see that zero-shot and few-shot learning do poorly on this. This is likely due to cluttered outputs. The model did not learn to output just the style-transferred text, but had other text before and after, reducing evaluation accuracy. Our holistic LLM evaluation showcases that RoSA TST not only maintains sentiment, but is also stylistically and grammatically correct.

6 Analysis

In specific instances, our models displayed an impressive understanding of Shakespearean vocabulary. However, there were lots of very obvious translation fails. We include some examples of failed translations in the sentence pairs shown in the Table below:

Modern English	**Expected Output**	**Model Output**
I love rich Capulet's daughter	Then plainly know my heart's dear love is set On the fair daughter of rich Capulet.	I love rich Capulet's daughter. Or Shakespeare would write.
Have you given up so quickly on Rosaline , whom you loved so much ?	Is Rosaline , whom thou didst love so dear , So soon forsaken ?	Madam, he's well. And friends with Caesar. You are an honest man.
The cat chased the mouse around the garden.	Thine cat didst pursue the mouse 'round yon garden.	The feline, in its wild pursuit, didst chase the mouse around the garden's greenery.

The first example, for instance, the fine-tuned model rather than learning to perform the translation, simply repeats its input prompt. This clearly shows that the model at times lacks an understanding of the training task for which it is being trained for. A longer training epoch and a more diverse dataset would have helped the model improve its performance. The second failure case, as demonstrated in the second example, indicates that the models simply hallucinated a possible output with very little relationship with the prompt except sounding Shakespearean. While the generated output is in the correct style, it does not capture the semantic meaning of its input.

The model also struggled with handling modern concepts and figurative language. This highlights several key findings. Firstly, lack of context. Word-for-word substitution often fails as the system needs a deeper grasp of how meaning is conveyed within Shakespearean style. Our training dataset was sentence pairings, and had no context which is crucial in these tasks. Secondly, the model's output is fundamentally limited by the vocabulary and sentence patterns found in its training data. Words such as artifical intelligence did not exist in Shakespearean time, so our model would do word-to-word replacement which is not ideal. Given the budget and space constraints, these results are very promising, especially the 85.1% LLM Evaluation score. If we run this approach on larger models such as Claude-3, we are confident in SOTA scores.

In general, the model did show a good command over Shakespearean vocabulary. In the third example, for instance, the model does word-to-word substitution of cat to feline, and also adds to the context with "the wild pursuit", a characteristic often found in Shakespearean English.

7 Conclusion

Though far from perfect, the system evaluated represents a promising step in the challenging task of TST. Our achievements include a novel implementation of RoSA for TST which performs incrementally better than other approaches. We also propose a novel evaluation metric for TST tasks. As has been seen in many works relating to TST, we were limited by evaluation metrics. Automated metrics, designed to capture semantic preservation (ie. BERTScore) were not successful indicators. Specifically, a translation may flow and be semantically correct but still lacks the prose of Shakespeare's writing. Limitations in computing also forced us to use an underperforming base model, which weakened our results. In general, our findings suggest several paths toward an improved TST system including training on larger corpora and perhaps incorporating a knowledge base of Shakespearean concepts. We believe that these steps will help the model's understanding of vocabulary and improve accuracy.

References

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. In *Advances in Neural Information Processing Systems*, volume 36, pages 10088–10115. Curran Associates, Inc.

Zhenxin Fu, Xiaoye Tan, Nanyun Peng, Dongyan Zhao, and Rui Yan1. Style transfer in text: Exploration and evaluation. 32.

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.

Harsh Jhamtani, Varun Gangal, Eduard Hovy, and Eric Nyberg. 2017. Shakespearizing 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.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng-Xin Yong, Hailey Schoelkopf, et al. 2022. Crosslingual generalization through multitask finetuning. *arXiv preprint arXiv:2211.01786*.

Mahdi Nikdan, Soroush Tabesh, Elvir Crnčević, and Dan Alistarh. 2024. Rosa: Accurate parameter-efficient fine-tuning via robust adaptation.

Phil Ostheimer, Mayank Nagda, Marius Kloft, and Sophie Fellenz. 2023. Text style transfer evaluation using large language models.

Lei Pan, Yunshi Lan, Yang Li, and Weining Qian. 2024. Unsupervised text style transfer via llms and attention masking with multi-way interactions.

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.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2020. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter.

Tianxiao Shen, Tao Lei, Regina Barzilay, and Tommi Jaakkola. 2017. Style transfer from non-parallel text by cross-alignment.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Huggingface's transformers: State-of-the-art natural language processing.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert.

A Appendix (Prompts)

Our prompt is as follows:

```
You are an LLM evaluator for Text Style Transfer. In specific, the task at hand
is evaluating how well style was transfer from a sentence of normal, modern style
to Shakespearean styled text. In specific, you will evaluate on a metric of 0 -
100, taking into account a few factors: how accurate the style was transferred,
how accurately the content/sentiment was preserved, how fluent the shakespearen
text is, how grammatically correct it is, and general naturalness of the text in
shakespearean style. For example, you should give a 0 if the sentence is exactly
the same, with no attempt of changing it. You should give a score of 90 - 100 if
it is the best possible shakespearean version of saying the input sentence. You
have a few metrics available to you, each on a scale of 0 - 1:

Word Mover Distance: {wmd}
BERTSCORE (F1): {bert}
Perplexity: {pp}
Classifier: {classifier}

It is very much possible that these scores can be high while actually, the actual
transfer is low on human evaluation (the most accurate evaluation. It is also very
much possible that these scores can be low while actually, the actual transfer is
high on human evaluation. You are provided with these metrics to make a thorough,
more informed evaluation. Please output just your evaluation. The sentences are:

Normal Text: {reference}
Shakespearean Text: {generated}
```