# Antiqu-ator

*What's past hast cometh back!*

**Final Report for the Course Project in**

**Creative Applications of Natural Language Processing**

**ECE 1786**

Rohan Deepak Ajwani (1008640179)

Jingxuan Su (1003276162)

**Word count:** 1991 words

**Penalty:** 0%

## 1. Introduction

Language is a powerful tool and style is an essential part of any writing. The style of writing varies depending on the author, region, time-period, etc. Shakespeare's style was the tradition of the time, however, it is rarely used today.

The goal of this project is to create a system capable of translating modern English texts into Shakespeare-style texts. We explore the capabilities of our transformer architecture for style transfer and fine-tune existing state-of-the-art transformer models for translating modern English to Shakespeare English texts. Finally, we demonstrate the application of our models through a simple chatbot application.

## 2. Illustration

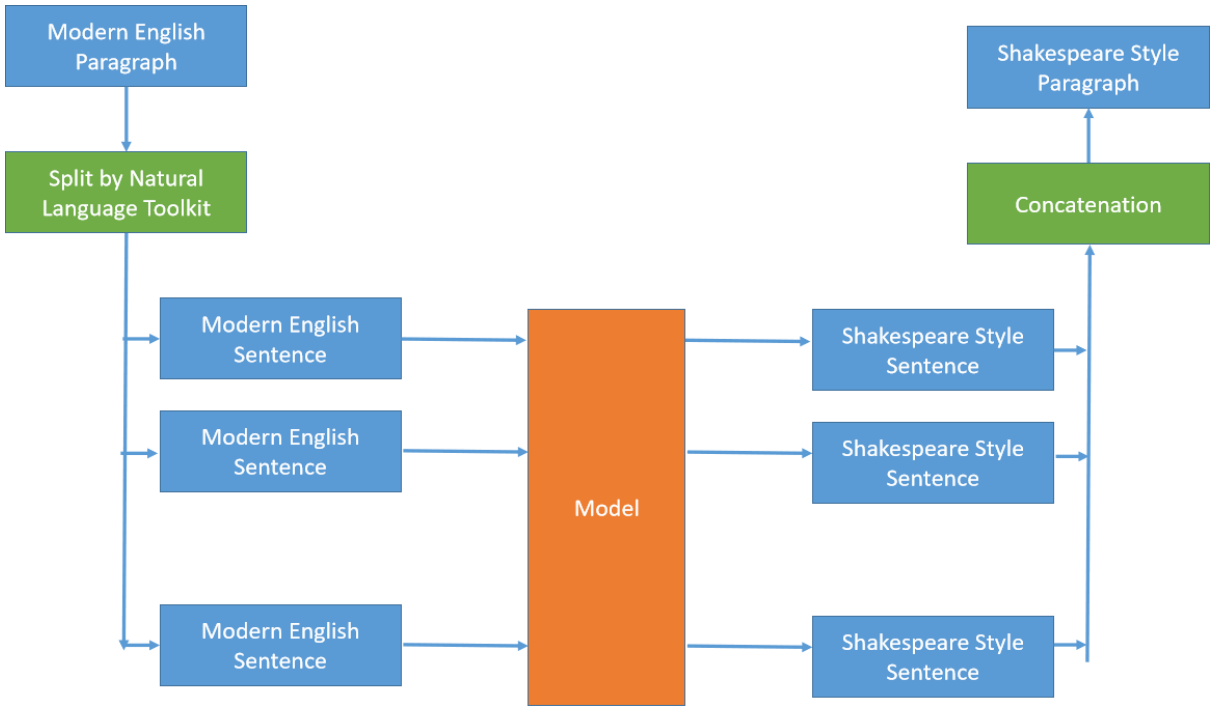



**Figure 1: System diagram**

Figure 1 illustrates the entire system. The orange block is the fine-tuned transformer model.

## 3. Background

Several studies have reported the use of statistical and neural network based approaches for machine translation and style transfer. Xu et. al. [1] proposed the use of phrase-based machine translation techniques for translating between Shakespearean and modern English.

Sutskever et. al. [2] leveraged the ability of the LSTM architecture [3] to store information over long time intervals and introduced the Seq2Seq architecture. In [4], the authors present a sequence-to-sequence neural machine translation model with a global attention mechanism.

Recent advances in transformers [5] have shown much better results than trivial LSTM based Seq2Seq models and have produced state-of-the-art results for sequence-to-sequence tasks including translation, question answering, summarization, and building chatbots.

## 4. Data and Data Processing

The target dataset should be a sentence-to-sentence dataset. For each entry in the dataset, the input data is a modern English sentence and the output label is the corresponding Shakespeare style sentence.

We collect our dataset from Litchart Shakescleare [6]. We collected 31 PDF files, each containing the original and modern English translations of Shakespeare's plays. All the text, including annotation, is divided into paragraphs, separated by the speaker's name in capital letters or a number of annotations. An example of the original PDF text is shown in Figure 2.

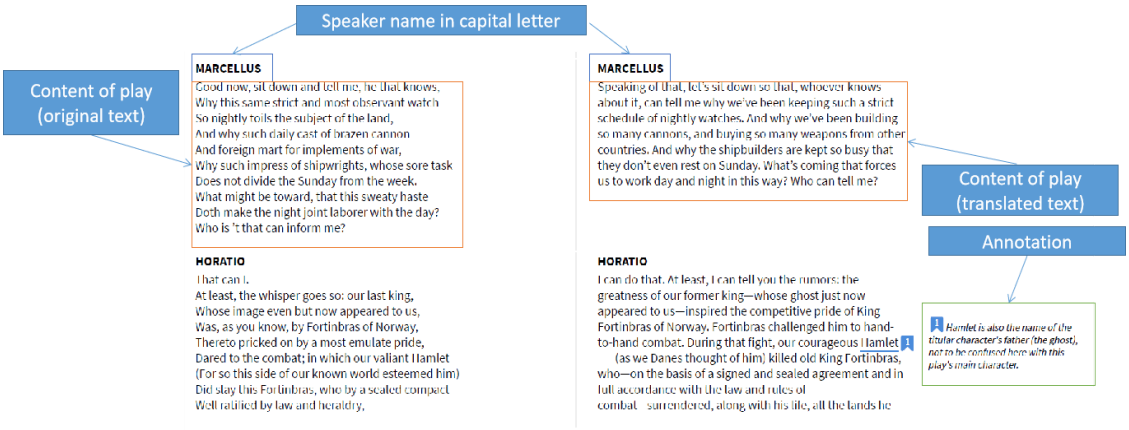



**Figure 2: Example of raw text in PDF**

We use python to collect the raw data from PDF. Then we remove the text for PDF header and footer to make cleaner data. An example of the processed text is shown in Figure 3.

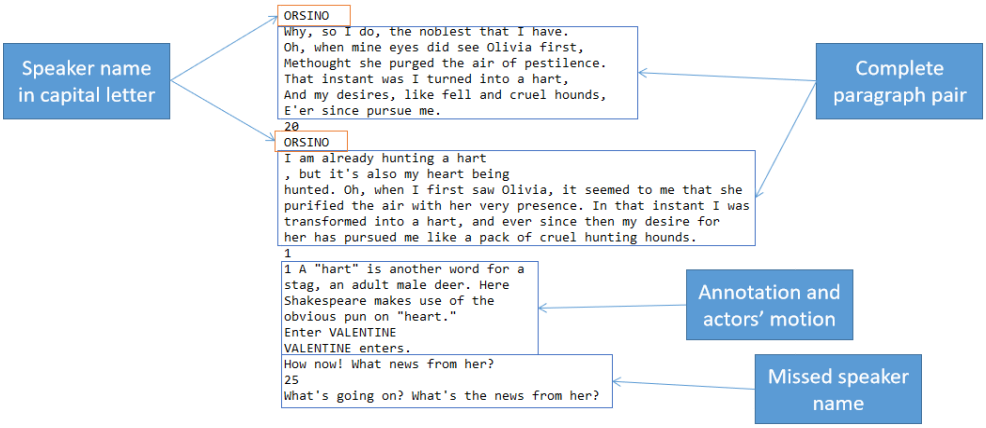



**Figure 3: Example of processed text**

Each paragraph containing the original text is immediately followed by a paragraph containing the corresponding translated modern text. We identify each paragraph using the speaker's name in capital letters, and use the order of paragraphs described above to match them into paragraph pairs. During this process, we manually spot-checked the result to remove mismatched paragraph pairs due to issues in the original text, such as missing or extra speaker names. We also remove words used as annotations and used to describe the actor's actions.

Finally we break the paragraph into sentences, since we observed poorer results on paragraph-to-paragraph translation than sentence-to-sentence translation. However, grammatical issues in the original text such as missing or wrong special symbols caused inaccurate separation. On manually inspecting paragraph pairs, we found that accurately translated sentence pairs should have similar word counts, and decided to remove the sentence pair if the number of words in that sentence pair differed by more than 6. The resulting dataset has 42,535 sentence pairs. We split it into 80:20 for training and validation.

## 5. Architecture

We use GPT2 [7] (Small and Medium) and T5 [8] (Small and Base) as our final architectures.

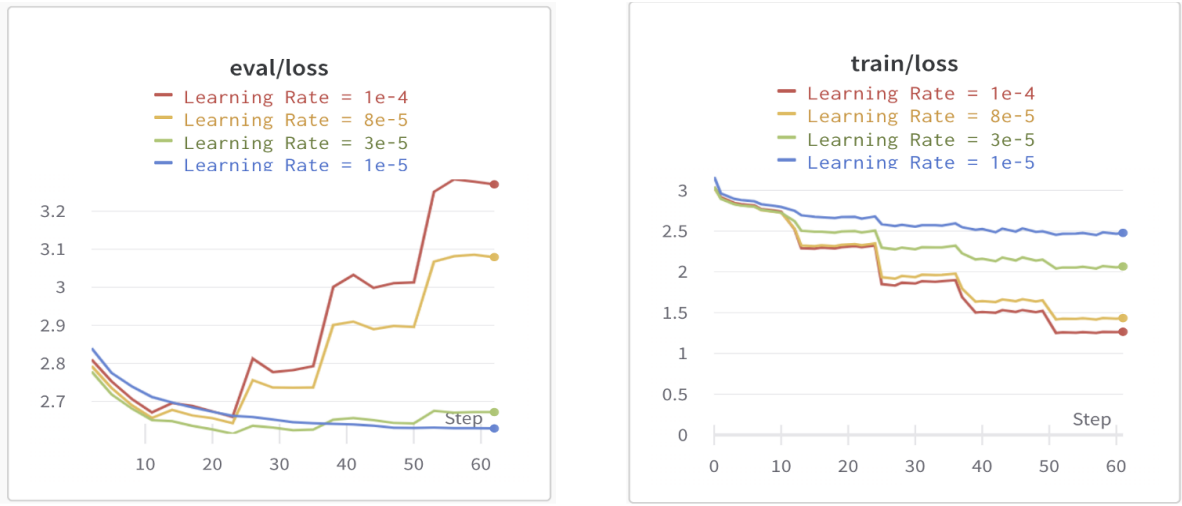
The GPT2 model is fine-tuned using causal language modeling to predict the next token in a given input [9]. To facilitate this, we concatenate the modern and the Shakespearean English sentences with an '=' character. We then pass the encoded text to both the inputs and labels of the model and the model uses the first n tokens of the input and the next token of the output while fine-tuning ($1 \leq n < tokenized\_input\_length$).

The T5 model is fine-tuned in a sequence-to-sequence fashion with teacher forcing [10]. We feed the encoded modern English sentence to the inputs and the corresponding encoded Shakespeare style sentence to the labels.

## 6. Quantitative Results

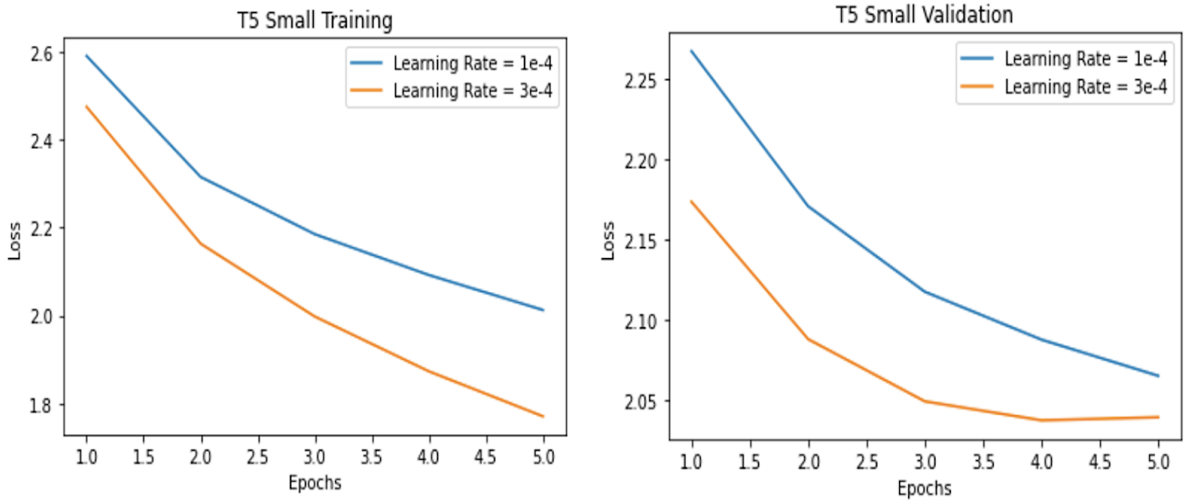
The graphs below show the training and validation loss during training of all models. We see that most of them work well since validation loss steadily decreases. We notice overfitting when the learning rate is too large for the larger models (GPT-2 Medium and T5-Base) .


**(a)**

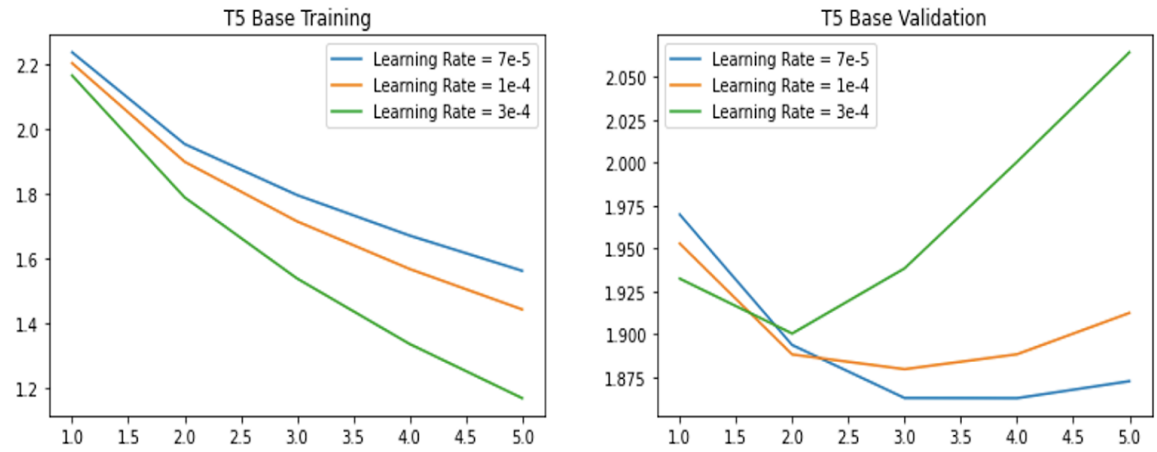



**(b)**

**Figure 4: Validation (left) and Training (right) curves for (a) GPT2 and (b) GPT2-Medium made with Weights&Biases [11]**




**(a)**




**(b)**

**Figure 5:Training (left) and Validation (right) curves for (a) T5-Small and (b) T5-Base**

We evaluate the greedy decoding using BLEU scores [12]. We first calculate the geometric mean of the precisions ($GM_{prec}$) starting from 1-gram upto N-gram precision (N is specified in BLEU-N) between the candidate and the reference texts. Next, we calculate a brevity penalty (*BP*), which penalizes very short output sentences. Finally, we calculate the BLEU score by multiplying $GM_{prec}$ with *BP*.

In our evaluations, we calculate the BLEU-3 scores for all the models across all hyperparameter settings. We calculate BLEU-1 and BLEU-2 for outputs of length 1 and 2 respectively. We calculate the BLEU-3 for all other output sizes.

| **Model** | **Hyperparameters** | | **BLEU-3** | |
|---|---|---|---|---|
| | **Learning Rate** | **Epochs** | **Training** | **Validation** |
| GPT 2 | 3e-5 | 3 | 0.1249 | 0.1204 |
| | 3e-5 | 5 | 0.1375 | 0.1212 |
| | 8e-5 | 3 | 0.1077 | 0.0958 |
| | 8e-5 | 5 | 0.1158 | 0.0985 |
| GPT 2 Medium | 1e-5 | 3 | 0.2048 | 0.1918 |
| | 1e-5 | 5 | 0.212 | 0.1868 |
| | 3e-5 | 3 | 0.2268 | 0.1794 |
| | 8e-5 | 3 | 0.2746 | 0.1702 |

| | | | | |
|---|---|---|---|---|
| T5 Small | 1e-4 | 3 | 0.2290 | 0.2005 |
| | 1e-4 | 5 | 0.2416 | 0.2044 |
| | 3e-4 | 3 | 0.2487 | 0.2088 |
| | 3e-4 | 5 | 0.2801 | 0.2082 |
| T5 Base | 1e-4 | 3 | 0.2606 | 0.2164 |
| | 1e-4 | 5 | 0.3158 | 0.2070 |
| | 3e-4 | 3 | 0.3137 | 0.2181 |
| | 3e-4 | 5 | 0.4125 | 0.2019 |

**Table 1: BLEU-3 scores for all the models with various hyperparameter settings**

Table 1 shows that T5 models have higher BLEU score than GPT2 models which indicates better performance (on using greedy decoding). At the same time, the larger models for both GPT2 and T5 have higher BLEU scores which implies that larger models perform better than smaller models (on using greedy decoding). Higher learning rates and more epochs increase the BLEU score on the training set, but the BLEU scores for the validation first increase and then decrease, indicating overfitting.

## 7. Qualitative Results

We randomly select inputs from the training and validation set, and some common English phrases to inspect the quality of the generation.

**a. T5 Models**

We observed correct but monotonous outputs on using greedy decoding, where the sentence structure did not change much.

Thus, we used sampling with the temperature set to 0.9. Most outputs preserved the meaning with some interesting changes to the sentence structure. Some good examples are mentioned below:

- Input: I can see some reluctant confession in your eyes.
  Output: When I look back on your eyes I discern some naive confession.

- Input: Wake up the lively and swift spirit of fun.
  Output: Awake the livelyness and tempo of the jest.

- Input: Why are you looking for me?

Output: The question, wherefore look you for me?
- Input: I am hanging out with my dog.
  Output: With my dog I am out.

We also observed some outputs that were written in Shakespeare style, but did not retain the meaning of the original sentence, possibly due to overfitting. Following are some illustrations:

- Input: Can you hear me, forest dweller?
  Output: Mark you, mine own?

- Input: What I see and hear is like a dream!
  Output: I take my breath for bread and water, and leave a candle for it.

**b. GPT-2 Models**

There were many instances where GPT2 did not produce any output, or only produced punctuations on using greedy decoding. This could be the reason for the low BLEU scores calculated using greedy decoding. We then used sampling with the temperature set to 0.9.

We select the output starting from the '=' token until the first end of sentence punctuation is produced, as GPT2 does not stop producing output, and keeps writing until the max_new_tokens is reached. We observed some good outputs and some with slight distortion in meaning, as shown below:

- Input: Wake up the lively and swift spirit of fun.
  Output: The lively and swift spirit Of fair amusement awakens!

- Input: It may be the last time you do so.
  Output: If it be the last, your Grace, to do so.

- Input: Why are you looking for me?
  Output: Why seek you this sight of me?

- Input: I'll only confine myself to these clothes I'm wearing.
  Output: To these garments I cannot have my mind.

- Input: Do not cry over spilt milk.
  Output: Do not lament a spelt milk.

Some outputs completely differed in meaning from the input, for example:

- Input: It may be the last time you do so.
  Output: 'Twere done then.

- Input: Many congratulations to both of you!

Output: O happy king!

## 8. Discussion and Learnings

Our results show that the models perform well, especially on sampling in decoder. However, there are a few instances where the outputs entirely differ in meaning, and also a few instances where the output is the same as the input.

It is interesting to note that the GPT2 automatically appends the '=' token to the input, even if we do not provide it at the end of the input.

A surprising finding in GPT2 is that if we provide it with an incomplete sentence, it first completes the sentence, appends an '=' token, then produces the corresponding Shakespeare style sentence. This could be attributed to the fact that GPT2 was trained on the concatenation of the modern and Shakespearean sentences. Following are two such outputs for the same input: "Can you"

- Output: Can you **take me home, Sir Benedick? = Can you take me home, Signor Benedick?**
- Output: Can you **explain how you are so sure that I'm not a traitor? = Art thou convinced, Sir John, that I am not a traitor?**

We used T5 as an alternative to GPT2 as we saw poor quantitative results with GPT2. However, we realized that BLEU scores can be deceptive, and the model could show great performance using sampling when decoding output. If we were to work on a similar project, we would try evaluating using metrics other than BLEU score. We could also try using other sources of data that might have better (i.e. closer in meaning) modern translations of Shakespeare's plays.

We create two applications using gradio [13] to demonstrate the models' performance. The first app allows the user to select any model and specify the temperature, and produces the output in Shakespeare's style.

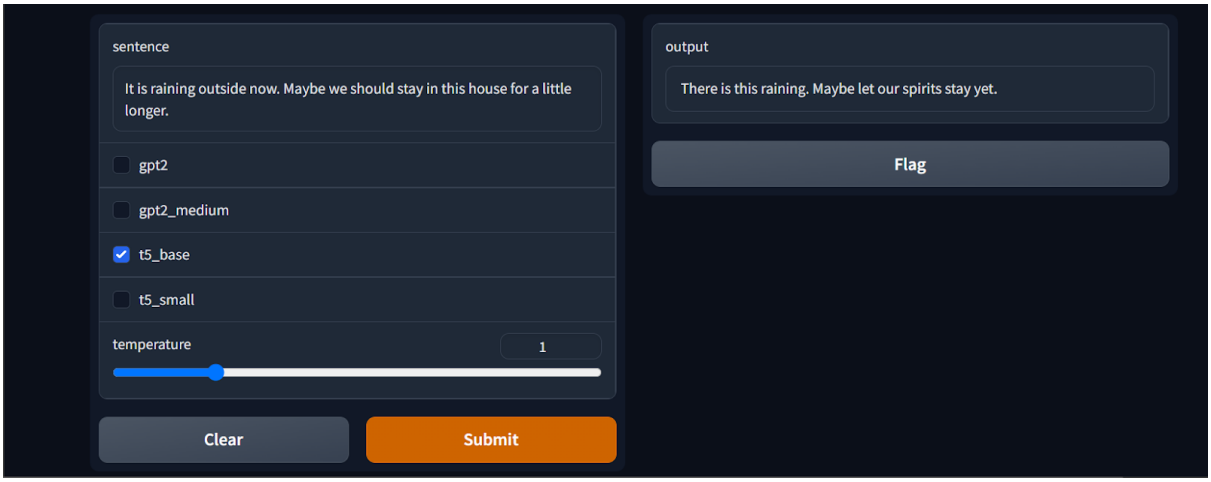



**Figure 6: Example from Modern English to Shakespeare app**

The second app is a chatbot that chats with the user in Shakespeare's style. It uses a pre-trained chatbot model (facebook/blenderbot-400M-distill) [14] from the huggingface hub, converts each output sentence to Shakespeare style using T5-Base and outputs the concatenated output. Some sample conversations are shown below.

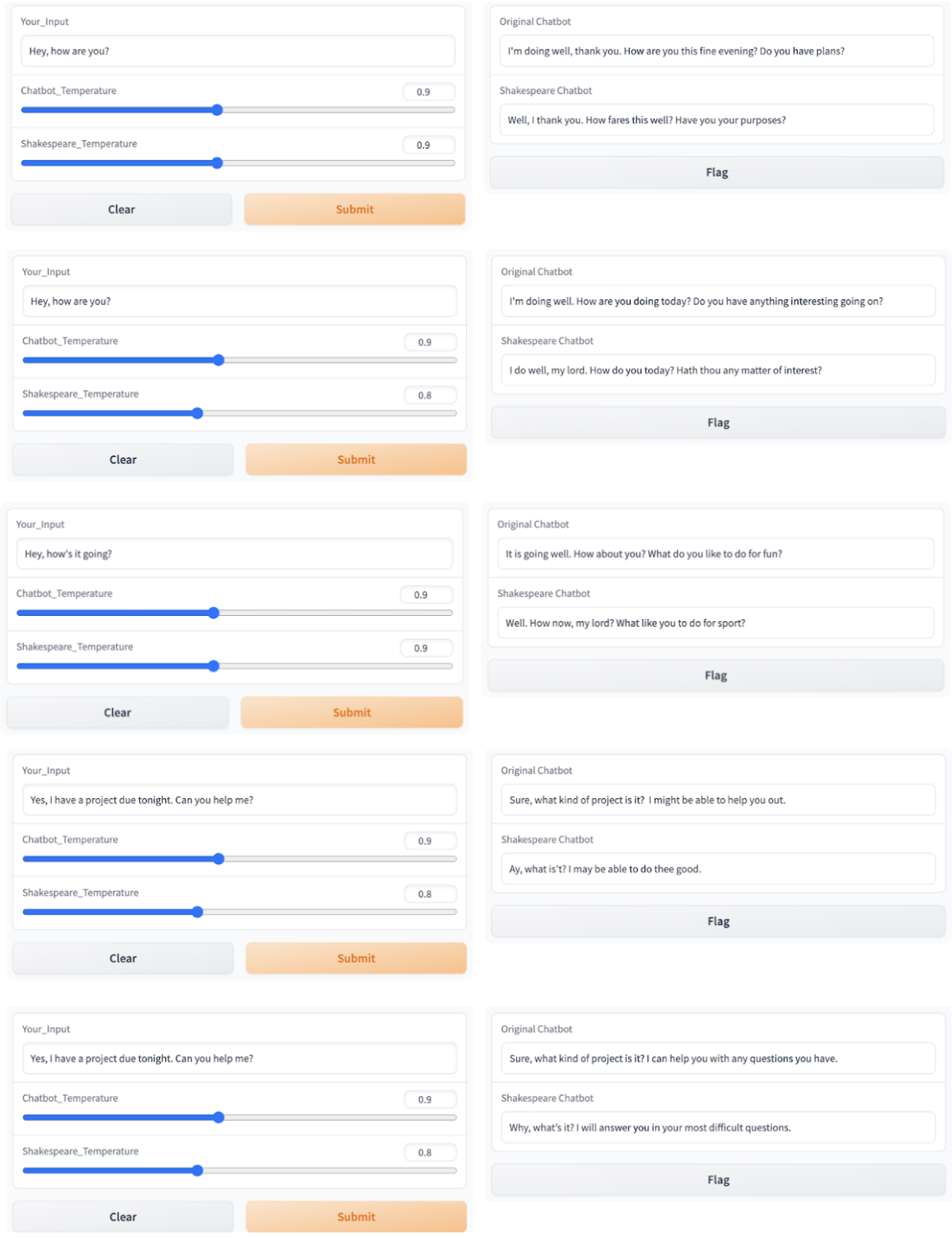



**Figure 7: Conversations from the Shakespeare chatbot app**

Figure 7 demonstrates one of the many potential applications of our modern to Shakespeare English model(s). A possible application of this chatbot could be in fantasy video games, where dialogues and responses are artificially generated.

## 9. Individual Contributions

**Rohan Ajwani**

1. In charge of fine-tuning GPT2 models and saving results to Weights&Biases.
2. Collaborated towards fine-tuning T5-small and T5-base.
3. In charge of quantitative evaluation using BLEU score and qualitative evaluation using sampling.
4. In charge of coding the Shakespeare style chatbot.

**Jingxuan Su**

1. In charge of collecting raw data from Litchart.
2. In charge of processing data to create dataset.
3. Participated in fine-tuning of T5-small and T5-base models.
4. In charge of building the full system based on the developed model.

## References


[1] W. Xu et al., 'Paraphrasing for Style', in *Proc. of COLING*, Mumbai, 2012, pp. 2899–2914.

[2] I. Sutskever, O. Vinyals, and Q. V. Le, 'Sequence to Sequence Learning with Neural Networks'. arXiv, 2014.

[3] S. Hochreiter and J. Schmidhuber, 'Long Short-Term Memory', *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 11 1997.

[4] Se Won Jang, Jesik Min, & Mark Kwon. (2017). Writing Style Conversion using Neural Machine Translation [Online]. Available: https://web.stanford.edu/class/archive/cs/cs224n/cs224n.1174/reports/2757511.pdf

[5] A. Vaswani et al., 'Attention is All you Need', in *Advances in Neural Information Processing Systems*, 2017, vol. 30.

[6] "Litcharts Shakescleare." [Online]. Available: https://www.litcharts.com/shakescleare/shakespeare-translations.

[7] A. Radfordt et. al., 'Language Models are Unsupervised Multitask Learners', 2019.

[8] C. Raffel et al., ‘Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer’. arXiv, 2019.

[9] “OpenAI GPT2.” [Online]. Available: https://huggingface.co/transformers/v3.0.2/model_doc/gpt2.html#overview

[10] “T5.” [Online]. Available: https://huggingface.co/transformers/v3.0.2/model_doc/t5.html#training

[11] L. Biewald, “Experiment Tracking with Weights and Biases,” Weights & Biases. [Online]. Available: http://wandb.com/.

[12] K. Papineni et al., ‘Bleu: a Method for Automatic Evaluation of Machine Translation’, in *Proc. of the 40th Annual Meeting of the Association for Computational Linguistics*, 2002, pp. 311–318.

[13] A. Abid et al., ‘Gradio: Hassle-Free Sharing and Testing of ML Models in the Wild’. arXiv, 2019.

[14] S. Roller et al., ‘Recipes for building an open-domain chatbot’. arXiv, 2020.

## Permissions

| Member | Post Video? | Post Final Report? | Post Source Code? |
|---|---|---|---|
| Rohan Ajwani | Wait till see | Yes | Yes |
| Jingxuan Su | No | Yes | Yes |