Topic: Exploring the Information Loss in Translation after Iterations using Large Language Models (LLMs)

Introduction

This project is based on the breakdown points of the text where the fidelity diminishes of after iterations and the information loss. The breakdown point is based on the language. The specific LLM that I have used puts a mark on especially the regional languages where it's showing an early breakdown due to the less representations in the dataset.

The expected outcome is to gain insights about the fidelity loss and how much the breakdown point values are differing from language to language. The impact that can be seen can be developed into fine robust translation system.

Implementation

For the Large Language Model, I have used Helsinki-NLP models- Marian MT from Hugging Face Pretrained Models for translating between language pairs. I have used these because these can be dynamically accessed based on the source. Specifically (opus-mt-en-fr) is one of them for English to French.

Translation Process:

In the code section, I have used a function: translate (text, src_lang, tgt_lang) because it loads Model and tokenizer. Then encodes the input text into token IDs. It does use the model to generate the translation of the input text, and then again it turns tokens into text.

Example code - def translate (text, src_lang, tgt_lang):

```
model_name = f'Helsinki-NLP/opus-mt-{src_lang}-{tgt_lang}'
```

Iteration Setup:

In the code, I have made it to 1- 10 iterations back and forth. For the odd ones, source to target language and even ones it's the vice versa. This is leading to degradation in quality.

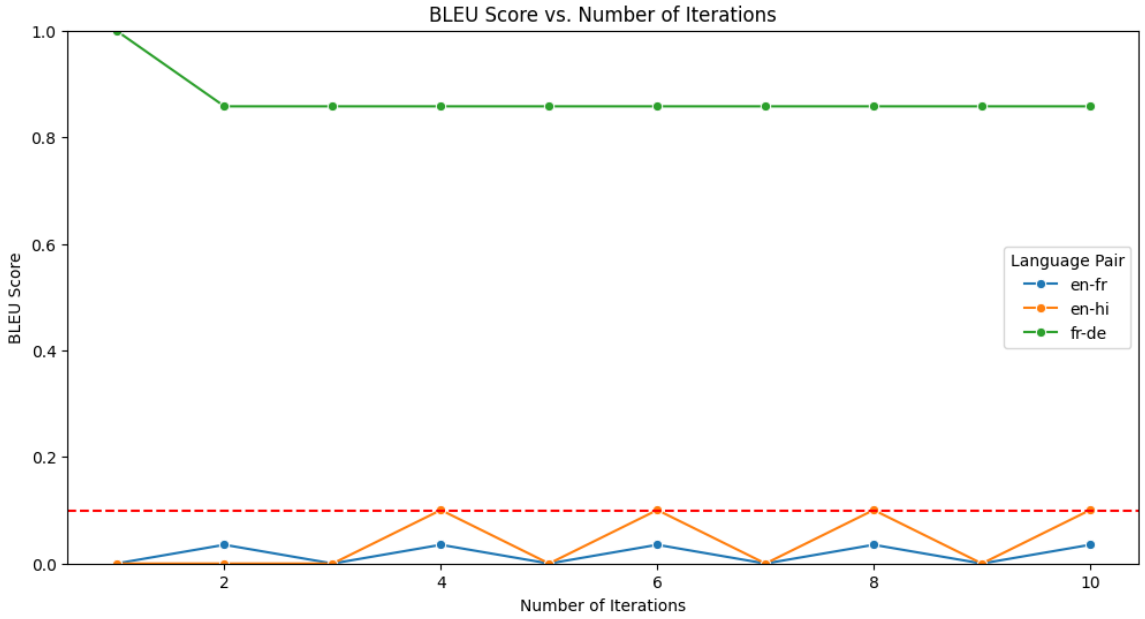
Component: BLEU score – After the iterations as expected there is a significant breakdown as indicated by the extremely low BLEU score. The nature of language pair influences on how quickly the information is lost.

Data Collection:

The original sentence was: "Hey, Rajarshi, how are you doing today? Wanna grab some coffee later? Today, Boss has given us a day off."

For each iteration I have calculated the BELU score between the final and the original output.

Results and Analysis:



There is a significant drop of BLEU scores which had dropped nearly to zero after every iteration. The sharp decline of en –fr and en-hi pairs. The scores are oscillating slightly around a very low value but the quality drops at 1-2 iterations while this is interesting

that the fr-de pair consistently shows BLEU scores a better value , demonstrating a minimal degree of loss in transition quality.

The Drive link for every document , the output CSV files , Code files storing -