Solution Building

Baseline: Zero-Shot Learning on Llama2 7B Chat

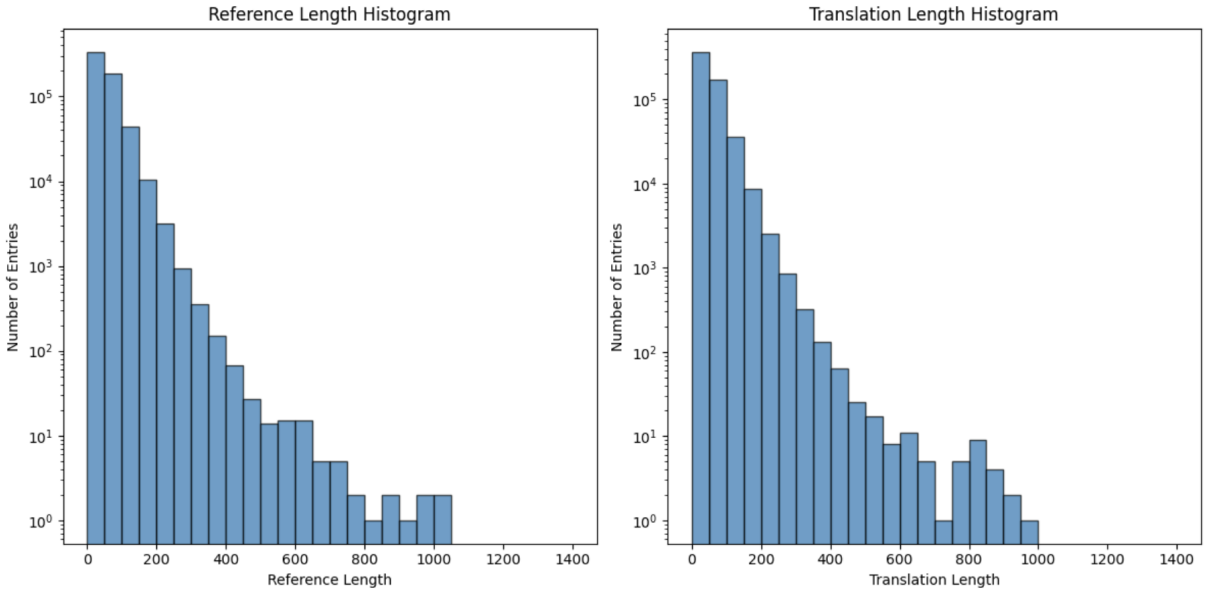
The idea was to use language modelling with prompt for text detoxification. I decided to use Llama2 7B chat for the problem as it is supported by Hugging Face's transformers library, and it is the largest model that can fit into the hardware provided by Colab and Kaggle.

To fit the model into the provided GPUs and make the fine-tuning process more efficient, I applied quantization and Low-Rank Adaptation of Large Language Models (LoRA).

Hypothesis 1: Llama2 7B Chat Fine-Tuning

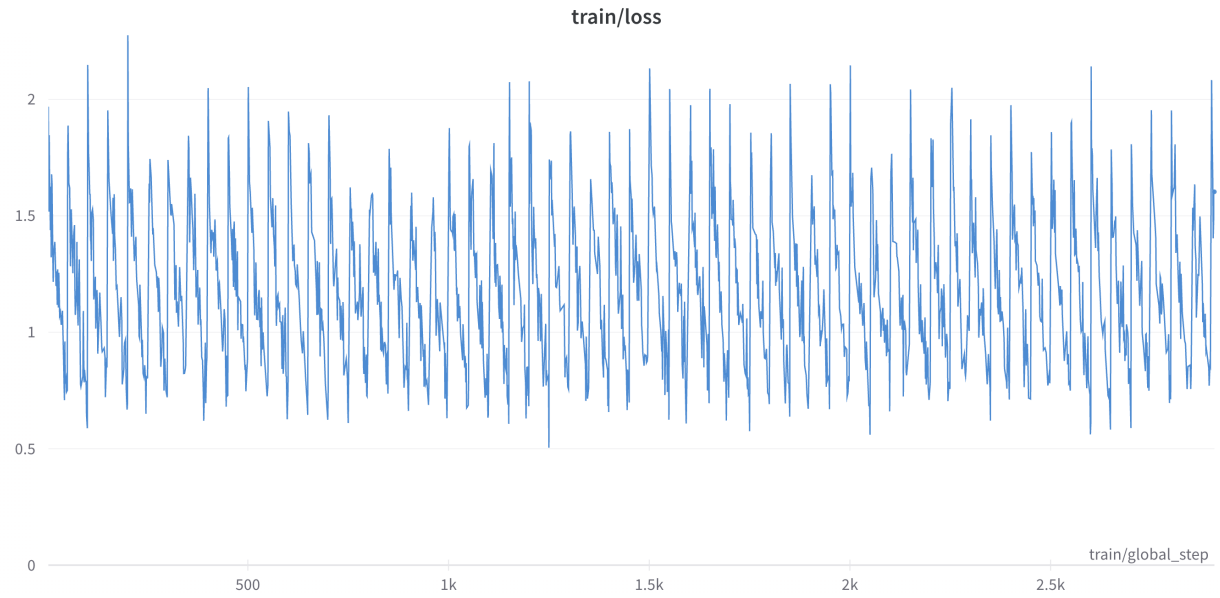
The next idea was to fine-tune the model on the filtered ParaNMT-detox corpus. There were two main problems: limitation in GPU memory and severe loss fluctuations during training.

Since there still was an issue with GPU memory, I explored the lengths of references and translations. The decision was to take the dataset entries which had references with length less or equal to 250 characters.



Number of references and translations in length ranges with step equal to 50 characters.

The next issue I had was the loss fluctiation during the training. I tried different learning rates, maximum gradient norms for gradient clipping, optimizers. I did not find a solution to this problem.



Loss fluctuations during Llama2 7B chat fine-tuning.

Hypothesis 2: T5v1.1 Large Fine-Tuning

Results