The models are variants of the Text-To-Text Transfer Transformer (T5) architecture, developed by Google, renowned for its versatility in natural language processing tasks. This report outlines experiments designed to evaluate and compare the performance of these models across various dimensions including training efficiency, evaluation accuracy along with inference quality and speed.

Experimental Setup:

The experiments were conducted using the same configuration to ensure a fair comparison. Both models were trained and evaluated on the same dataset, with identical hyperparameters. The training dataset was split into batches of size 4, and each model was trained for 5 epochs using the Adam optimizer with a learning rate of 1e-4. Additionally, a separate testing dataset was used to assess the models' performance using perplexity scores. Inference tests were performed on sample inputs to compare the models' output quality and generation speed.

Experiment 1: Training Performance

In this experiment, we trained both the T5-small and T5-base models on a common dataset for a fixed number of epochs. The training losses were recorded and compared across epochs to assess the convergence behavior and relative training efficiency of the two models. Additionally, we monitored the training time for each epoch and the overall duration of training to evaluate the computational cost associated with training each model. This experiment aimed to provide insights into the training dynamics and efficiency of the T5-small and T5-base models.

Experiment 2: Evaluation Performance

For evaluation, both models were assessed using a separate test dataset to calculate the perplexity scores. The perplexity scores served as a metric to quantify the performance of the models in predicting the next word in the sequence. By comparing the perplexity scores obtained for T5-small and T5-base models, we aimed to gauge their relative performance in terms of language modeling capabilities. Here, we see that the model #####

Experiment 3: Inference Performance

In this experiment, the inference performance of both models was evaluated on sample inputs. The time taken for generating outputs was recorded to assess the inference speed of each model. Furthermore, the quality of generated outputs was qualitatively compared to identify any differences in the model's ability to produce coherent and contextually relevant responses. This experiment aimed to provide insights into the efficiency and effectiveness of the T5-small and T5-base models in generating outputs for given input texts.