# Chapter 1

# Resources

## 1.1 Models and Tokenizers

The models used in this project come from two different sources. The GPT2-XL model is based on the OpenAI's GPT2 model [4]. The GPT2-XL model has 48 attention modules as compared to the 12 found in the base GPT2 model. This gives the GPT2-XL model more capability in terms of language modelling tasks but also requires significantly more computing resources to train and work with [3]. This translates to 117 million parameters in the GPT2 model vs 1.6 billion parameters in the GPT2-XL model. Given that the COMET-DISTIL models are based on the GPT2-XL model, we opted to utilize the naive GPT2-Xl model as our baseline for comparison. We downloaded the GPT2-Xl model by using the Hugginface model library [1]. As our tokenizer, we used the pre-trained tokenzier also available from the hugginsface model library [1].

The second model we used was the COMET-DISTIL model [7] that was trained on the ATOMIC-10X knowledge graph. The COMET-DISTIL model was downloaded from the author's provided website that can be found on their githhub page [2]. The tokenizer used with this model was also downloaded from the same source. The tokenizers for both models were adjusted to add EOS tokens to ensure that they could work with the datasets that were used for evaluationg the models.

In order to train and test the models, both models had different language modelling heads (dependening on the task) connected to the models. The heads were then trained on the training set of the datasets and then evaulted on the evaulation set of the datasets.

## 1.2 Datasets

In order to test the performances of the respective models, several different standard datasets were used. This included the widely recognized Glue dataset. The second benchmark was based on the CommonsenseQA dataset

[6] that tests a model's ability to answer a question based on prior knowledge. The SuperGLUE benchmark on the other hand is made up of several different datasets that test a model on a variety of tasks including BoolQ, CB, COPA and MultiRC.

These two benchmarks were choosen based on their wide variety of tasks and the difficulty associated with each of these tasks. In addition, the original glue benchmark was used to test the performance of models to add an additional comparison point to the study.

## 1.3 Compute Resources

To train and evaluate the performance of the respective models, Lambda Compute Instances at the University of Alberta were utilized. The machines included two Nvidia N100 Gpus with 24GB of memory. Due to the large size of the GPT2-XL Model (5.2 Billion Parameters), the DeepSpeed[5] library from Microsoft was also utilized to help fit the models within the GPU memory.