Assignment 3

1. Pre-trained Transformer models and knowledge access

d) We implemented the miniGPT model and trained it on the birth_places_train.tsv dataset without pretraining. As expected accuracy on the evaluation data was very low, around 1.8%. Just predicting “London” as a birth place in the evaluation data gave higher accuracy, which was 5%.

e) The accuracy of our **vanilla model** on the dev set was 20.6%.

g) The accuracy of our **perceiver model** on the dev set was 8.8%. The time complexity of the perceiver model is O($L \cdot m^2$), where L is the number of layers in the transformer and m is the bottleneck dimension. When stating the time complexity, we did not consider the first cross attention computation, which needs to be performed between the learned queries and original input sequence. If we also consider that computation (possibly with the optional cross attention computation between learned queries and input sequence at end of the model), the time complexity would become O($L \cdot m^2 + m \cdot l$), where l is the input sequence length. The time complexity of the vanilla model is O($L \cdot l^2$). As m is usually much lower than l, a perceiver model is much faster than a vanilla model.

2. Considerations in pre-trained knowledge

a) Our pretrained model was able to achieve higher accuracy because of the fact that it had prior knowledge about many properties of human language and specific information including birthplaces of some people in the fine-tuning dataset from being trained on the wiki data. However, the non-pretrained model did not have any awareness of human language and no knowledge about the birthplaces of people, having to learn the specific information from scratch, which led it to have low accuracy.

b) The first reason that the pretrained NLP models can cause concern because of their “hallucination” is that they might increase misinformation in a society, which can lead to negative consequences. For example, we might have a LLM that is made to retrieve and deliver news about political / economic events in an easy-to-digest manner to people. Such a LLM has some possibility of wrongly delivering the information, and thus perpetuating falsehood in a society.

The second reason that the pretrained language models can cause worries is that it might result in big-scale economic issues. For example, in the future a large sector of tech industry and other machine-operating industries might me powered by LLMs, and if somewhere an important “hallucination” problem occurs that results in serious consequences such as somebody being killer or unjustly going for prison, then the public might become afraid of AI and stop using products related to it, causing large-scale recession in a society.

c) A language model can predict answers for factual questions such as the birthplaces even if it did not see the exact answer in its pretraining and fine-tuning datasets. The reason for this is that such models were trained to predict the next token and if you give them a factual question they will answer it, no matter what, unless they are trained to answer, “I don’t know”. If they did not see the answer in the training set, they most likely base their answer on associations and correlations in the data.

For example, if a LLM gets a question about predicting somebody's birth date but it did not have that data during training, it will probably give the birth date of a person that it has seen during the training and has the most similar name to the person’s name in the question. Such a response can cause future concerns because of the reasons mentioned in b).