CSCI 6040: Computational Analysis of Natural Languages Spring 2025
Homework 5 – Self-Attention, Transformers, and Pretraining
Simone Mayers

A. (2 point) Succinctly explain why the pretrained (vanilla) model was able to achieve an accuracy of above 10%, whereas the non-pretrained model was not.
   - The pretrained (vanilla) model achieved higher accuracy because it had already learned general language structure and token patterns from the large pretraining corpus (wiki.txt). This gave it a meaningful initialization before fine-tuning on the birthplace task. In contrast, the non-pretrained model started from random weights and had to learn both the structure of language and the prediction task from scratch, resulting in lower accuracy.

B. (4 points) Take a look at some of the correct predictions of the pretrain+finetuned vanilla model, as well as some of the errors. We think you'll find that it's impossible to tell, just looking at the output, whether the model *retrieved* the correct birth place, or *made up* an incorrect birth place. Consider the implications of this for user-facing systems that involve pretrained NLP components. Come up with two **distinct** reasons why this model behavior (i.e., unable to tell whether it's retrieved or made up) may cause concern for such applications, and an example for each reason.

   - Trust and misinformation
     - Users may trust the model's output as fact, even when it is fabricated. Since the text is fluent and confident, there's no clear signal to the user that a prediction may be incorrect.
     - Example: A chatbot for historical research states, "Marie Curie was born in Vienna." This looks plausible, but it's incorrect. If unverified, this kind of output can spread misinformation, particularly in educational or journalistic contexts.
   - Accountability and verifiability
     - The model does not provide sources for its predictions, so users can't trace answers back to data. This is especially problematic in sensitive domains.
     - Example: A medical assistant tool outputs, "Dr. Jane Smith was born in Chicago." If this is wrong, there's no way for users to verify where the information came from. The model simply made it up, which is dangerous in domains like healthcare or law.

C. (4 points) If your model didn't see a person's name at pretraining time, and that person was not seen at fine-tuning time either, it is not possible for it to have "learned" where they lived. Yet, your model will produce *something* as a predicted birth place for that person's name if asked. Concisely describe a strategy your model might take for predicting a birth place for that person's name, and one ethical concern this raises for the use of such applications. (While 2b discussed the problems that could arise from made up predictions, 2c asks for a mechanism the model could be using for generating birth places of people not seen at fine-tuning time and why such a mechanism could be problematic.)

   - Strategy: The model may rely on statistical associations between character patterns in names and common birthplaces from the training data. For example, it might associate "Giovanni" with "Rome" or "Yuki" with "Tokyo" due to co-occurrence patterns learned during pretraining and fine-tuning.
   - Ethical Concern: This can reinforce cultural or racial stereotypes. Automatically linking a name to a certain country or city may ignore individual identity and propagate biased or discriminatory assumptions — a serious issue in applications involving real users, personal data, or decisions that affect people.