CS685 Quiz 1: *Transformers*
Released 2/28, due 3/8 on Gradescope (please upload a PDF!)
Please answer both questions in 2-4 sentences each. Make sure to also fill out the AI disclosure!

1. Assume we are building a Transformer sequence-to-sequence model to solve a machine translation task. Why don't we need to use masking when implementing cross attention?

We compute the cross-attention between the decoder final level token representation and encoder final level token representation. We get the decoder final-level token representation after computing the MASKED multi-head attention of the prefix or the already produced decoder outputs. So as we have already done MASKED multi-head attention, we don't need to do masking again for cross-attention. The current output final level token representation only has information of auto-regressive tokens before it, so no need of masking again and we want all the final level token representation of decoder to pay attention to the entire sequence of encoder.

2. Why can't the hidden state computations of Transformer language models be parallelized at test time?

During test time, we generate the output in an auto-regressive nature with one word at a time. So we need the prefix to generate the next word, hence at test time we cannot parallelize the prediction.

AI Disclosure

AI1: Did you use any AI assistance to complete this homework? If so, please also specify what AI you used.
Your answer here

(only complete the below questions if you answered yes above)

AI2: If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which problem is associated with which prompt.

- *Your response here*

AI3: (*Free response)* For each problem for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good answer, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to get the answer or check your own answer?

- *Your response here*