CS685 Quiz 5: ***text generation / evaluation***
Released 10/18, due 10/22 on Gradescope (please upload a PDF!)
Please answer both questions in 3-4 sentences each.

1. **The task of *long-form question answering* requires models to generate a paragraph-length answer to a given question. A popular dataset for this task is ELI5, which contains questions such as this one. Describe how you would train a model to solve this task! You may want to browse a few threads on the ELI5 subreddit to get a feel for the data first.**
   The ELI5 task would need to contain two main components - a document collection component (where a set of reference documents would be collected based on which the answer could be constructed) and the training component (which could be a kind of text summarization task based on the collected documents). Basically, it would be a combination of information extraction and natural language processing. The text summarization component could be evaluated using the ROUGE score.

2. **Let's say you've built your ELI5 QA model and now want to evaluate it. Unfortunately, automatic evaluations such as ROUGE are not reliable metrics of answer quality for this task. Instead, you turn to Amazon Mechanical Turk to evaluate your generated answer quality. You take 500 random questions from the test set, paired with your model's generated answers, and ask Mechanical Turk workers to rate the correctness of the answer on a scale of 1 to 5. Describe some problems with this evaluation setup.**
   The problems with using Amazon Mechanical Turk in annotating long-form question answers of this form are -
   1. These questions may require high levels of domain knowledge. Mechanical Turk is crowd-sourced and there is no guarantee that the annotators would have the required knowledge.
   2. Long answer questions would require a longer time to read through, compare and evaluate. MTurkers who are made to read and evaluate multiple answers may get bored and speed through the annotation process, leading to sub-par and inaccurate annotations.