## Description of Improvement

The plan for Part 4 is to implement a two-stage Retrieval-Augmented Generation (RAG) pipeline by introducing a Context Re-ranking stage between BM25 retrieval and LLaMA generation. The pipeline will include the following. I will try 3 context re-rankers on the wikipedia.jsonl dataset for RAG, where we retrieve k = 1 to 10 passages. I will use 3 cross-encoder models to determine which one improves accuracy for more contexts. This change directly addresses the primary failure mode identified in Part 3d: the Context Overload and Distraction problem. BM25 is effective at finding potentially relevant documents, but its ranking is based purely on keyword overlap, often placing irrelevant passages near the top. By using a semantic cross-encoder, we will drastically improve the Signal-to-Noise Ratio of the input to LLaMA, leading to more accurate and focused answers, which should translate directly into higher F1 and EM scores. Afterward, I will compare the best performing RAG + re-ranking combination with the same model including a better prompting technique. I believe that part of the reason there are small EM and F1 scores is because the responses from the LLM are far too verbose. I want to shrink those through prompting. Finally, I want to try and use other techniques for evaluating the performance of the RAG system. The EM and F1 scores are good but do not capture context which I believe is a key part of this.

## Experiments and Measurements

We plan to run the following experiments:

1. Baseline Comparison: Compare the new 3 re-ranked RAG performance against our best result from Part 3, the best n, 1 to 10, without re-ranking.
2. Optimal Cutoff: Experiment with different cutoff values for k, number of passages to include after the re-ranking, to determine the ideal “context length” that maximizes performance while minimizing noise.
3. Re-Rank vs Re-Rank and improved prompting: I think for me the big thing is improving the prompting strategy will lead to the greatest increases in EM and F1 scores. I want to test the best re-ranked version versus the same with better prompting.
4. Metrics: The success of the improvement will be measured using Exact Match and F1 Score on the dev set, focusing on the improvement in EM, which directly indicates fewer incorrect answers, but ensuring there is an improvement in F1 as well. I will try to include the RAGAS testing element as well which will include context measurements to the results.