Homework 9

11-967

Due date: 03/30/2025 11:59 PM EST

Language models struggle to perform tasks that require complex reasoning, knowledge that they did not see during training, or interactions with the real world. There are two strategies to improve language model capabilities at these complex tasks: tool-use and retrieval augmentation. Tool-use involves training an LM to call a computer program that runs externally to the LM. Retrieval-augmentation involves conditioning a language model's generations on retrieved information. In this homework, you will build a RAG system.

Note: You will use the retrieval model you trained in HW8. The starter code of this homework is exactly the same as HW8. You can choose to re-download or start working on your existing code.

*Note: There is **only one** Gradescope submission page this time for your `report.pdf`. There is **no** autograder coding submission on Gradescope even though you need to write a few lines of code.*

Problem 1: Building a RAG System

Retrieval augmentation is a useful technique to improve LLMs' performance in answering factoid queries, especially those regarding niche domains. In this problem, you will use the retrieval model you trained in the previous homework to build a small RAG system.

[Question 1.1] (***Writing, 5 points***) You will be using questions from TOFU, and an instruction-tuned Pythia 6.9B model.

DELIVERABLES FOR Q1.1

A. Do you expect the chosen language model to be able to answer questions from the TOFU dataset? Why?

B. Run `retriever/scripts/generate_without_rag.sh` and present the questions and respective completions.

[Question 1.2] (***Writing, Coding, 5 points***) The code under `retriever/driver/rag.py` is built from the inference code for Pythia on your previous homework. You need to extend its functionality by implementing a function that augments prefixes with relevant passages.

DELIVERABLES FOR Q1.2

A. Implement the missing functions, and run `retriever/scripts/generate_with_rag.sh`. Present the queries and respective completions.

B. Look for the answers in the TOFU dataset, and comment on the effectiveness of the approach.

[Question 1.3] (***Writing, 5 points***) The performance of RAG systems is conditioned on the quality of the retrieval step. Look at the top-N passages retrieved for each query. In this case, you should see the relevant passage very close to the top.

DELIVERABLES FOR Q1.3

A. In more difficult scenarios where the retriever won't be as effective, using a re-ranker is a common step of the pipeline. Besides the cross-encoders introduced in the previous exercise, LLMs as list-wise re-rankers have also been studied. Refer to this paper for more details. Name one advantage and one disadvantage of cross-encoders when compared to LLM list-wise re-rankers.

B. In Question 1.2, the prefix was augmented with the top-1 passage. Present and comment on changes in the generation results if we augment the questions:

- With the top-10 passages in-order.
- With the top-10 passages shuffled.