ELL881/AIL821

Large Language Models: Introduction and Recent Advances

Semester I, 2024-25

Quiz 2

Answer the questions in the spaces provided. No extra pages will be given. Write proper justifications for every answer.

Course Code: _____

Name: _____

Entry Number: _____

Total Marks: 20 **Time: 40 minutes**

Questions	Marks	Score
T/F Questions	10	
Retrieval-Augmented Language Model Pre-training	7	
A Real-world Use-case	3	
Total	**20**	

True-False Questions

Identify whether the following statements are **True** or **False**. Give a short, one-line justification for your answer. *No marks will be given without justification.* Each question carries **2 marks**.

$2 \times 5 = 10$ marks

1. GRACE is a global optimization-based method for knowledge editing in LLMs.

2. Distillation is a lossless model compression technique.

3. Chain-of-Thought prompting technique elicits backward reasoning in LLMs.

4. Retrieval augmented fine-tuning improves the model's ability to answer questions in *closed-book* settings.

5. The self-attention mechanism in transformers scales quadratically with the context length, making it inefficient for processing extremely long input sequences.

Question 1: Retrieval-Augmented Language Model Pre-training

Retrieval-Augmented Language Model pre-training (REALM) augment language model pre-training with a latent *knowledge retriever*. REALM decomposes $p(y|x)$ into two steps: *retrieve*, then *predict*. Given an input x, first, the possibly helpful documents z are retrieved from a knowledge corpus $\mathcal{Z}$. Then, the generation of the output y is conditioned on both the retrieved z and the original input x – modeled as $p(y|z, x)$.

1. Which distribution is modeled by the *neural knowledge retriever* in REALM?

(1 mark)

2. Assume that $f(x, z)$ gives the *relevance score* for a document z, with respect to the input x. What can be a possible choice for such a function f?

(1 mark)

3. Compute the gradient of the log-likelihood pre-training objective in REALM, with respect to the parameters of the knowledge retriever, θ, i.e., $\nabla_\theta \log p(y|x)$. Express $\nabla_\theta \log p(y|x)$ in terms of $p(y|x)$, $p(y|z, x)$, $p(z|x)$ and $\nabla_\theta f(x, z)$.

(5 marks)

Question 2: A Real-world Use-case

Suppose you are tasked with developing an automated assistant to resolve the doubts of students of an educational platform, particularly for the subject *Physics*. You have access to a pre-trained LLM (which is also instruction-tuned and aligned using human feedback, using general conversational data – *no domain-specific data is used for fine-tuning*).

Assume that you also have access to ample computing resources and the whole internet. Students can also ask queries on numerical problems (remember that LLMs are *really* bad at calculations!) or on very recent Physics discoveries that may be absent in the pre-training data of the LLM in use.

Considering all these (along with any other possible challenges you can think of), outline the steps and strategies you will take to design an LLM-based assistant for this scenario. Give proper justification for your decisions, and mention the objective functions if you desire to go ahead with any further custom training/fine-tuning of the pre-trained LLM you have.

(3 marks)