National University of Computer and Emerging Sciences

Lahore Campus

GenAI (AI4009)

Date: April 12th 2025

Course Instructor:

Dr. Hajra Waheed

Sessional-II Exam

Total Time (Hr):	**1**
Total Marks:	**20**
Total Questions:	**2**

Attempt all questions in the provided sequence on your answer sheets. Marks to be deducted otherwise.

CLO 2: Analyze the architectures and pre-training methodologies of large language models.
CLO 4: Discuss advanced topics such as prompt engineering, retrieval-augmented generation (RAG), fine-tuning techniques, quantization.

Q1: Answer the following short questions. **[3 + 3.5 + 4 marks]**

a) Mistral 7B has 32 layers. If each token stores 16 bytes of key-value (KV) cache per layer, how much memory (in MB) is needed to store the KV cache for a batch of 8 sequences, each with 512 tokens?

b) How is prefix tuning different from prompt tuning? Explain.

c) What is the QLoRA's double quantization phenomenon and how is it beneficial?

CLO 2: Analyze the architectures and pre-training methodologies of large language models.

Q2: You are given a batch of 2 sequences, each of 22 tokens.
Each token has a 6-dimensional embedding.
You process them using chunked attention with the following setup:

- Chunk size: 4 tokens per chunk
- For each sequence, chunks are processed independently

For all chunks:

- Q is computed individually for each chunk
- K_all, V_all are accumulated across chunks within a sequence
- If a chunk has fewer than 4 tokens (e.g., remainder), process it as-is (no padding)

Considering the above scenario, **answer the following questions.** **[2 + 3.5 + 2 + 2 marks]**

a) How many chunks will be created per sequence?
b) For Chunk 3 of Sequence 1, what are the dimensions of:
   - Q 1 X 6
   - K 4 X 6
   - V 4 X 6
   - K_all
   - V_all
c) Suppose now that chunk 4 only takes the context from the immediate previous chunk for K_all, V_all. What would their dimensions be in Chunk 4?
d) What would happen if we do not store K_all and V_all across chunks? What impact it has?