0009

National University of Computer and Emerging Sciences

Lahore Campus

GenAI (AI4009)

Final Exam

Date: May 22nd 2025

Course Instructor:

Dr. Hajra Waheed

Total Time (Hr):	**3**
Total Marks:	**52**
Total Questions:	**7**

[redacted]
Roll No

[redacted]
Section

[redacted]
Student Name

Attempt all questions in the provided sequence on your answer sheets. Marks will be deducted (2 marks per question) for violating this rule.

Attach the paper with the answer sheet.

***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.*

Q: Answer the following questions. **[37 marks]**

For all **numeric** questions **make sure to mention each step**, for **conceptual** questions **be precise and to the point.** No marks will be awarded for irrelevant details.

✓ **1)** A Stable Diffusion model takes 50 denoising steps per image. Each step takes 50 ms on GPU. You process a batch of 10 images.

a) What is the total time in seconds? **[2]**

b) How much speed-up would 4× parallelism offer in theory? **[2]**

✓ **2)** Quantization:

a. Convert the following FP32 values into INT4 using the quantized formulas discussed in class: [0.5, - 0.3, -1.0, -0.4, 0.6, 1.0]
Mention formulas wherever required. **[5]**

b. In QLoRA, double quantization compresses model weights using 4-bit quantization and 8-bit quantization constants. If a model has 64,000 parameters, and one 8-bit constant is shared across every 64 weights, compute the total memory used.
Give your answer in both bytes and kilobytes. **[4]**

✓ **3)** Suppose you're fine-tuning a Transformer model using LoRA. You freeze the base weights and only train the LoRA adapters.
If the original layer has a weight matrix of size [1024 × 4096], and LoRA uses rank r = 8, what is the total number of trainable parameters added per LoRA module? **[2]**

✓4) A Vision Transformer (ViT) model is configured to process medical images with a resolution of 512×512 pixels by first dividing each image into non-overlapping patches measuring 32×32 pixels. Each resulting patch is then mapped to a 1024-dimensional embedding space to capture its features. To retain spatial information, the model incorporates positional embeddings, which are stored using float16 precision, with each value occupying 2 bytes.

Given this configuration of the ViT answer the following questions:

a. How many patches are generated from one image? **[2]**

b. Calculate the memory (in MB) required to store positional identifiers for the entire batch. **[3]**

5) In a large language model employing a Mixture of Experts (MoE) architecture, the routing mechanism is configured to use Top-3 routing, where each input token is directed to the top three most relevant experts based on gating scores. The MoE layer consists of 12 experts, each constrained by a fixed capacity of 384 tokens. To encourage a more uniform distribution of tokens across experts and prevent overloading of a few, an auxiliary load balancing loss is incorporated into the training objective. This loss penalizes imbalance in expert utilization and is scaled by a weighting coefficient/constant α = 0.015, promoting equitable load distribution during routing. Answer the following questions and mention each step and formula where required:

a. In the gating mechanisms of MoE why the softmax is applied after the top-k selection? **[3]**

b. Expert 10 receives 450 token assignments. How many tokens will be dropped? **[1]**

c. Calculate the auxiliary loss if the assignment counts across 12 experts are as follows: [240,242,244,246,248,250,252,254,256,258,260,322] **[5]**

d. What percentage of assignments do the top 3 experts handle? **[2]**

✓6) Your company is building a scientific assistant that summarizes long, interconnected research topics (e.g., climate change models or protein folding pathways). The queries often require reasoning over multiple interlinked facts from different documents. However, occasionally the retrieval quality drops, leading to hallucinations in the final response.

Task: You are considering GraphRAG and CRAG to improve the factual grounding of your system.

a. Explain which technique is more suitable when the retrieved passages are topically rich but not tightly connected. Justify your answer using their core mechanisms. **[3]**

b. Which method offers more protection against hallucinations from unreliable documents, and why? **[3]**

CLO 1: *Foundational principles of Generative AI models.*
CLO 3: *Evaluate the optimization methods and evaluation metrics*
CLO 4: *Discuss advanced topics such as prompt engineering, retrieval-augmented generation (RAG), fine-tuning techniques, quantization.*

✓**Q7.** **Answer the following short questions.** **[3*5=15 marks]**

a. In denoising diffusion models, the forward process deliberately corrupts input data by gradually adding noise over several steps.

What is the purpose of this progressive noising process, and how does it contribute to the model's ability to generate high-quality samples during the reverse denoising phase?

b. What are the advantages of using Reinforcement Learning for reasoning compared to supervised learning?

c. DeepSeek-R1-Zero is trained using reinforcement learning without any initial supervised fine-tuning. Explain how the use of a *rule-based reward system* in DeepSeek-R1-Zero contributes to its ability to learn general reasoning abilities without instruction tuning. What are the potential advantages and disadvantages of using a rule-based reward system in this context?

d. Domain-specific applications often require both extensive knowledge and robust reasoning. How does the RARE framework address the challenges posed by these requirements, particularly in the context of limited model capacity?

e. You're implementing the query reformulation step in CRAG. Which parts of the user interaction should your keyword extractor focus on, based on the original CRAG paper?