

## Generative AI (AI4009)

Date: November 5<sup>th</sup> 2025

### Course Instructor

Dr. Akhtar Jamil

## Sessional-II Exam

**Total Time (Hrs):** 1

**Total Marks:** 50

**Total Questions:** 3

---

Roll No

---

Section

---

Student Signature

**Do not write below this line**

**Attempt all the questions.**

[ CLO 1-3]

**Question No. 1. MCQ [1 x 25 = 25]**

**Answer the MCQs on the given answer sheet attached at the end of the question paper. Answers marked on the question paper will not be evaluated.**

1. In DeepSeek, what does context extension mean?  
A) Activating more experts per token  
B) **Allowing longer input sequences**  
C) Reducing training parameters  
D) Using RLHF for alignment
2. In a Mixture-of-Experts (MoE) layer, the Router plays a crucial role. Which of the following statements best describes its function?  
A) It normalizes the hidden vector using RMSNorm before passing it to the experts  
B) It reduces the dimensionality of the input vector to minimize computation  
C) It computes scores for token and removes the top-K tokens from the input sequence  
D) **It computes scores for experts and selects the top-K experts to process each token**
3. Which of the following is **not** an encoder-only model?  
A) BERT  
B) RoBERTa  
C) ELECTRA  
D) **GPT-5**
4. Which of the following tasks would most naturally benefit from an **encoder-decoder** model rather than an encoder-only or decoder-only model?  
A) Sentiment analysis of product reviews  
B) Predicting the next word in a story  
C) **Summarizing long research papers into short abstracts**  
D) Generating embeddings for a search engine

5. In Large Language Models (LLMs), what is a jailbreak?
  - A) A prompt that improves model accuracy on reasoning benchmarks
  - B) An adversarial input designed to make the model bypass its safety or refusal policies**
  - C) A method used to reduce the number of activated experts in MoE models
  - D) A fine-tuning process to align the model with human feedback
6. CycleGAN eliminates the need for paired datasets by using cycle-consistency loss. However, under which condition can the cycle-consistency constraint lead to poor domain mapping?
  - A) When both generators have insufficient capacity to learn inverse mappings.
  - B) When the two domains have overlapping distributions, making the discriminators too weak.
  - C) When the cycle-consistency loss dominates it causes generators to learn identity mappings that minimize reconstruction error without meaningful style transfer.**
  - D) When the learning rate of discriminators is higher than that of generators.
7. In Mixture-of-Experts (MoE) models like DeepSeek, routing collapse can occur during training. Which of the following outcomes best indicates that routing collapse has taken place?
  - A) Most tokens are routed evenly across all experts, resulting in balanced utilization.
  - B) The router consistently sends tokens to only a few experts while others remain largely inactive.**
  - C) The number of activated experts per token increases, improving computational efficiency.
  - D) Experts begin sharing parameters, reducing the number of trainable weights.
8. In GPT-5's unified system, the real-time router improves response quality by:
  - A) Increasing the number of parameters activated per token to enhance reasoning power
  - B) Fixing model selection rules at training time to avoid unpredictable behavior during inference
  - C) Replacing the deeper reasoning model with the smart & fast model whenever users ask for quick answers
  - D) Selecting the most suitable model based on conversation type, complexity, tool needs, and explicit user intent**
9. Why does BERT replace only 80% of the selected tokens with [MASK] instead of 100% during pre-training?
  - A) To make the model robust to unseen tokens and reduce the pre-train/fine-tune mismatch**
  - B) To increase the total number of masked tokens and speed up training
  - C) To balance the number of [MASK] tokens and positional encodings
  - D) To ensure every word in the sentence is eventually masked at least once
10. In BERT's fine-tuning process for Question Answering (QA), what changes are made compared to the pre-trained model?
  - A) The architecture of BERT is completely retrained from scratch
  - B) The [MASK] prediction head is reused for answer span detection on QA dataset
  - C) Only a task-specific output layer is added and trained on the QA dataset**
  - D) The model's embeddings are frozen and not updated during training
11. GPT uses both left and right context during inference.
  - A) True
  - B) False**

**National University of Computer and Emerging Sciences**  
**Islamabad Campus**

- 12.** In Masked Language Modeling (MLM), what happens if the masking ratio is set too high?
- A) The model sees too little context, reducing its ability to predict masked tokens accurately.**
  - B) The model sees too large context which can cause overfitting and slowing convergence.
  - C) Training becomes computationally expensive due to dense masking.
  - D) The model fails to apply positional encodings effectively.
- 13.** In the paper titled “Attention Is All You Need”, how was English Constituency Parsing used?
- A) As a pre-training objective for learning syntactic structures.
  - B) As an unsupervised evaluation of word embeddings.
  - C) As a fine-tuning task to improve its translation capabilities.
  - D) As a fine-tuning task to test the Transformer’s ability to generalize beyond translation.**
- 14.** In StyleGAN, what is the purpose of introducing stochastic variations during image generation?
- A) To randomly adjust model weights and improve convergence
  - B) To add fine-grained randomness (e.g., freckles, hair strands) that increases image diversity without changing overall structure**
  - C) To control global attributes such as pose and lighting
  - D) To replace the latent vector with a learned mapping network
- 15.** In Adaptive Instance Normalization (AdaIN), given:  
 $\mu(x_i) = 6.67$ ,  $\sigma(x_i) = 2.49$ ,  $y_{s,i} = 3$ , and  $y_{b,i} = 1$ ,  
what is the output of AdaIN for  $x_i = 10$ ?
- A) 3.80
  - B) 4.00
  - C) 5.00**
  - D) 6.00
- 16.** In StyleGAN, what problem does Mixing Regularization help prevent?
- A) The generator focusing only on one latent code, causing features to become entangled**
  - B) The discriminator learning features too quickly
  - C) The loss function ignoring fine details in images
  - D) The mapping network producing identical latent vectors
- 17.** During BERT’s fine-tuning stage, what happens to the model parameters?
- A) Only task-specific layers are trained
  - B) Parameters are re-initialized before training
  - C) Encoder layers remain frozen
  - D) All pre-trained parameters are updated using labeled downstream data**
- 18.** In a diffusion model, the noisy sample is generated as
- $$x_t = \sqrt{\alpha_t} x_0 + \sqrt{1 - \alpha_t} \varepsilon,$$
- If  $x_0 = 1.0$ ,  $\alpha_t = 0.81$ ,  $\varepsilon = 0.5$  and  $\varepsilon = 0.5$ , what is  $x_t$  approximately?
- A) 1.90
  - B) 2.90
  - C) 1.81
  - D) 0.81**

19. Which expression correctly represents the Markov property in the forward diffusion process?
- A)  $q(x_t | x_{t-1})$
  - B)  $q(x_t | x_{t-2}, x_{t-1})$
  - C)  $q(x_t | x_0, x_1, \dots, x_{t-1})$
  - D)  $q(x_t | x_{t-3}, x_{t-2})$
20. In a CycleGAN training setup between domains X (horses) and Y (zebras), both the cycle-consistency loss and adversarial loss are used. Suppose the cycle-consistency loss weight  $\lambda$  is set extremely high (e.g.,  $\lambda = 1000$ ). What is the most likely consequence during training?
- A) The generators will focus mainly on fooling discriminators and ignore reconstruction accuracy.
  - B) The model will perfectly preserve input content but fail to generate realistic images in the target domain.**
  - C) The discriminators will collapse due to unstable updates.
  - D) The model will achieve faster convergence with better domain mapping.
21. Which of the following correctly represents the loss function used to train the reverse diffusion process?
- A)  $L(\theta) = E[\|x_t - \hat{x}_t\|^2]$
  - B)  $L(\theta) = E[\|\varepsilon - \varepsilon_\theta(x_t, t)\|^2]$**
  - C)  $L(\theta) = E[\|x^0 - x_t\|^2]$
  - D)  $L(\theta) = E[\|\beta_t - \alpha_t\|^2]$
22. Why is the KL divergence term important for inference in diffusion models?
- A) It ensures the learned reverse process closely matches the true posterior, allowing accurate denoising during generation.**
  - B) It reduces noise accumulation in the forward diffusion process.
  - C) It speeds up sampling by skipping intermediate diffusion steps.
  - D) It prevents overfitting to the training dataset.
23. A Transformer encoder block uses a feed-forward network (FFN) with two linear layers configured as follows:
- Input dimension: 32
  - Hidden dimension: 128
  - Output dimension: 32
- Each linear layer includes one bias term per output neuron. How many total **learnable parameters** are in this FFN?
- A) 8,192
  - B) 8,256
  - C) 8,352**
  - D) 8,384

**National University of Computer and Emerging Sciences**  
**Islamabad Campus**

- 24.** During training, you observe that a Transformer model using Batch Normalization performs inconsistently with small batch sizes, while another version with Layer Normalization remains stable. Which explanation best describes this behavior?
- A) Layer Normalization performs better because it has fewer learnable parameters than Batch Normalization.  
B) Both methods are equally affected by batch size but differ in gradient scaling.  
**C) Batch Normalization depends on batch-level statistics that become unreliable with small batches, whereas Layer Normalization computes statistics within each sample, remaining independent of batch size.**  
D) Batch Normalization fails because it normalizes each feature within a single sample, while Layer Normalization uses the whole batch to stabilize training.
- 25.** In BERT's feature-based approach, how is the model used for downstream tasks?
- A) BERT's parameters are fine-tuned for each new task  
B) BERT's embeddings are replaced with one-hot word vectors  
C) Only the final classification layer of BERT is retrained  
**D) BERT is frozen and used to generate contextual embeddings as input features for another model**

**Question No 2.**

**Write short answers to the following questions. [3 x 5=15]**

**Q.1.** How Mixture-of-Experts (MoE) enables large models to increase their parameter capacity without proportionally increasing computational cost.

Mixture-of-Experts (MoE) allows models to scale up their total parameter capacity by organizing parameters into multiple “experts.” For each input token, only a small subset of these experts is activated, while the rest remain inactive. This selective activation ensures that the model can handle very large numbers of parameters but only uses a fraction of them during computation. As a result, MoE achieves high model capacity while keeping the computational cost and inference time relatively low.

**Q.2.** Suppose  $x = [3, 3, 3]$  and  $\gamma = [0.5, 0.5, 0.5]$

Compute the RMSNorm output vector as used in DeepSeek v3 paper.

- Step 1: Compute mean of squares:  $(3^2 + 3^2 + 3^2)/3 = (9 + 9 + 9)/3 = 27/3 = 9$
- Step 2:  $\sqrt{(\text{mean}(x^2))} = \sqrt{9} = 3$
- Step 3: Normalize:  $[3/3, 3/3, 3/3] = [1, 1, 1]$
- Step 4: RMSNorm:  $[1, 1, 1] \times [0.5, 0.5, 0.5] = [0.5, 0.5, 0.5]$

**Q.3.** You are tasked with building a system for a multinational company that needs to:

# National University of Computer and Emerging Sciences

## Islamabad Campus

1. **Translate** customer support tickets from multiple languages into English.
2. **Summarize** long complaint descriptions into short actionable notes for agents.
3. **Classify** the urgency level of each ticket (low, medium, high).
4. **Automatically generate** follow-up responses to customers.

Based on the Transformer model types (Encoder-only, Decoder-only, Encoder–Decoder), which architecture(s) would be most suitable for each of these four tasks, and why?

- Translation + Summarization → **Encoder–Decoder** (e.g., T5, BART) because these are text-to-text generation tasks.
- Classification (urgency detection) → **Encoder-only** (e.g., BERT, ELECTRA) because they are strong at representation learning.
- Response generation → **Decoder-only** (e.g., GPT models) because they excel at autoregressive text generation.

Q.4. What two specific regularization methods are used in the Transformer architecture as described in the paper?

The paper uses dropout on sub-layer outputs and embedding sums, and label smoothing (with  $\epsilon = 0.1$ ) as its regularization methods.

Q.5. What is the key difference between disentanglement and style transfer in generative models?

### Answer:

Disentanglement focuses on controlling independent factors of variation (like shape, color, or pose) within a model's latent space, while style transfer combines the content of one image with the style or texture of another.

### Question No 3. Long questions. [2 x 5=10]

Q.1. Assume that in your DeepSeek Model, you are given a router with 4 experts, and the following formula for calculating expert scores:

$$s_i = W_i \cdot u$$

- Input hidden vector:  $u = [0.5, -1.0, 2.0]$  is the,
- $W_i$  is the weight vector for expert (i).

The router selects the Top-2 experts with the highest scores and the weights for the 4 experts are:

$$W_1 = [1, 0, 1], \quad W_2 = [0, 1, 2], \quad W_3 = [2, -1, 0], \quad W_4 = [-1, 1, 1]$$

Compute scores

$$S_i = W_i \cdot u = w_{i_1 \cdot u_1} + w_{i_2 \cdot u_2} + w_{i_3 \cdot u_3}$$

#### Expert 1:

$$s_1 = (1)(0.5) + (0)(-1.0) + (1)(2.0) = 0.5 + 0 + 2.0 = 2.5$$

# National University of Computer and Emerging Sciences

## Islamabad Campus

**Expert 2:**

$$s_2 = (0)(0.5) + (1)(-1.0) + (2)(2.0) = 0 - 1 + 4 = 3.0$$

**Expert 3:**

$$s_3 = (2)(0.5) + (-1)(-1.0) + (0)(2.0) = 1 + 1 + 0 = 2.0$$

**Expert 4:**

$$s_4 = (-1)(0.5) + (1)(-1.0) + (1)(2.0) = -0.5 - 1 + 2 = 0.5$$

Compare scores and select Top-2 experts:

**Expert 2 (highest: 3.0)**

**Expert 1 (second: 2.5)**

Q.2. You are given the following matrices and  $d_k = 3$ :

$$Q = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix}, \quad K = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}, \quad V = \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

Compute the attention output using the formula of scaled dot product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

**Solution**

$$QK^T = \begin{bmatrix} 1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 1 \end{bmatrix}$$

$$\frac{QK^T}{\sqrt{3}} = \frac{\begin{bmatrix} 2 & 0 & 1 \end{bmatrix}}{\sqrt{3}} = \begin{bmatrix} 1.1547 & 0 & 0.57735 \end{bmatrix}$$

$$\text{softmax}(\begin{bmatrix} 1.1547 & 0 & 0.57735 \end{bmatrix}) = \begin{bmatrix} 0.5329 & 0.1679 & 0.2992 \end{bmatrix}$$

$$\begin{bmatrix} 0.5329 & 0.1679 & 0.2992 \end{bmatrix} \begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 3.2988 & 4.2988 & 5.2988 \end{bmatrix}$$

$$\text{Attention}(Q, K, V) = \begin{bmatrix} 3.2988 & 4.2988 & 5.2988 \end{bmatrix}$$

**Answer Sheet MCQs**

Fill the correct option. Only one option must be selected. Selection of multiple options or overwriting will result in ZERO marks.

CORRECT METHOD				WRONG METHOD						
<input type="radio"/>	<input checked="" type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="radio"/>	<input type="radio"/>			
NAME:				ROLL NO:						
<input type="checkbox"/>	Roll No			<input type="checkbox"/>	A	B	C	D	<input type="checkbox"/>	
	<input type="text"/>			11	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
0	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	12	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	13	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	14	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		
<input type="checkbox"/>	3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	15	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="checkbox"/>	
	4	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>		A	B	C	D	
	5	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	16	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	6	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	17	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	7	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	18	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
<input type="checkbox"/>	8	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	19	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="checkbox"/>
	9	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	20	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	MCQs					A	B	C	D	
	Section1				21	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	A	B	C	D	22	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
<input type="checkbox"/>	1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	23	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="checkbox"/>
	2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	24	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	3	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	25	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	
	4	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						
	5	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						
<input type="checkbox"/>	A	B	C	D	<input type="checkbox"/>					<input type="checkbox"/>
	6	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						
	7	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						
	8	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						
	9	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>						
<input type="checkbox"/>	10	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="checkbox"/>					<input type="checkbox"/>

**National University of Computer and Emerging Sciences  
Islamabad Campus**