

Generative AI (AI4009)

Final Exam: Part-I

Date: December 31, 2025

Course Instructor

Dr. Akhtar Jamil

Maximum Time (Hrs): 1.5

Total Marks: 50

Total Questions: 1

Roll No

Section

Student Signature

Do not write below this line

Part-I MCQ

Maximum time allowed is 1.5 (one hour thirty minutes). However, you can take Part-II after you complete and return Part-I.

Part-I has 50 MCQs while Part-II has three questions.

[CLO 1-2]

Question No. 1. MCQ [1 x 50 = 50]

Answer the MCQs on the given MCQ Answer Sheet attached at the end of the question paper.

Answers marked on the question paper will not be evaluated.

1. Consider a teacher model producing logits [5, 1] for a binary classification task. When the temperature T is increased from 1 to 10 during distillation, which change most directly improves the student's learning signal?
A) The student receives sharper gradients for the correct class
B) The probability mass collapses to the dominant class
C) The relative confidence gap between classes is reduced
D) The cross-entropy loss becomes independent of logits
2. A large language model is fine-tuned on a narrow domain task and shows strong performance on that task but degraded performance on previously learned tasks. Which mechanism best explains this behavior?
A) Overfitting caused by insufficient regularization
B) Excessive gradient updates overriding previously learned representations
C) Vanishing gradients in lower layers
D) Inference-time distribution shift
3. Why is feature-based knowledge distillation often preferred over response-based distillation when the student architecture differs significantly from the teacher?
A) Feature-based distillation does not require labels
B) Logits cannot be computed for heterogeneous architectures
C) Response-based distillation requires identical output layers
D) Intermediate representations encode transferable structural information

National University of Computer and Emerging Sciences
Islamabad Campus

4. In the Scaled Dot-Product Attention mechanism, why is the scaling factor $1/\sqrt{d_k}$ mathematically applied to the dot product of QK^T prior to the softmax operation?
- A) To normalize the embeddings to a unit variance of 1.0 for cross-entropy stability.
 - B) To counter the growth of the dot product magnitude in high dimensions, preventing the vanishing gradients.**
 - C) To ensure the resulting attention matrix is symmetric and positive definite for optimized GPU computation.
 - D) To reduce the total memory complexity of the subsequent matrix multiplication with the Value (V) matrix.
5. Assume two adaptation strategies are considered for a 10B-parameter language model across 20 downstream tasks:
- Strategy X: Full fine-tuning per task
 - Strategy Y: LoRA-based adaptation per task
- Which one of the following statements is most technically accurate and suitable for this scenario?
- A) Strategy X is preferred as it minimizes catastrophic forgetting across tasks
 - B) Strategy X is preferable when inference latency is not critical
 - C) Strategy Y is preferred as it reduces parameter storage while preserving previous knowledge**
 - D) Strategy Y eliminates the need for task-specific supervision
6. When transferring knowledge between a large teacher network and a smaller student network, an additional transformation layer is sometimes introduced between their intermediate representations. What is the primary technical reason for this?
- A) To match incompatible feature dimensions**
 - B) To regularize the student model
 - C) To reduce computational complexity
 - D) To stabilize the optimizer state
7. Two neural networks are trained simultaneously, each using ground-truth labels while also learning from the predictions of each other. Which loss formulation is most suitable to enable effective mutual knowledge transfer during training?
- A) Mean squared error applied only to the final logits
 - B) Cross-entropy loss with hard labels only
 - C) A combination of cross-entropy loss and a KL-divergence loss**
 - D) Hinge loss applied to intermediate feature representations
8. A model is instruction-tuned using (instruction, input, output) triplets rather than task-specific labels. What is the primary benefit of this approach?
- A) Reduced training time due to fewer parameters
 - B) Improved instruction following and better generalization across diverse tasks**
 - C) Elimination of the need for evaluation metrics
 - D) Guaranteed robustness to distribution shift

National University of Computer and Emerging Sciences

Islamabad Campus

9. DeepSeek-V3 utilizes an auxiliary-loss-free load balancing strategy. If a specific expert i is determined to be overloaded after a training batch, how is the routing mechanism adjusted?
- A) The expert is temporarily removed from the Top-K selection for the duration of the next epoch.
 - B) The softmax temperature for that expert's gating score is increased to flatten the selection probability.
 - C) The router's weight matrix for that specific expert is increased by a hyperparameter γ .
 - D) The routing bias term b_i for that expert is decreased by a factor of γ .**
10. During BERT pre-training, the Masked LM (MLM) task involves a specific replacement strategy for the 15% of tokens selected for prediction. What is generally more appropriate technical distribution of this strategy?
- A) 100% replacement with the [MASK] token to maximize the bidirectional learning objective.
 - B) 80% replacement with [MASK] and 20% replacement with the [SEP] token to facilitate Next Sentence Prediction.
 - C) 50% replacement with [MASK], 25% replacement with a random token, and 25% remains unchanged.
 - D) 80% replacement with [MASK], 10% replacement with a random token, and 10% remains unchanged.**
11. In Variational Autoencoders (VAEs), the Reparameterization Trick is implemented to resolve which fundamental structural training limitation?
- A) The vanishing gradient problem inherent in extremely deep residual encoder layers.
 - B) The inability to backpropagate gradients through a stochastic sampling.**
 - C) The high computational complexity of calculating the KL-Divergence analytically for non-Gaussian priors.
 - D) The tendency of the latent space to collapse into a single point during Evidence Lower Bound (ELBO) optimization.
12. In Knowledge Distillation, the Temperature (T) parameter is applied to the modified softmax function. What is the specific effect of setting $T > 1$?
- A) It generates a smoother probability distribution.**
 - B) It produces a sharper distribution.
 - C) It shifts all logit values into a negative range to facilitate faster L2 loss convergence.
 - D) It forces the student model to ignore the logits and learn exclusively from the teacher's hint layers.
13. Within the Model Context Protocol (MCP), which feature allows an LLM to actively access data from databases or call external APIs?
- A) Resources
 - B) Prompts
 - C) Tools**
 - D) Roots

National University of Computer and Emerging Sciences

Islamabad Campus

14. Chain-of-Thought (CoT) prompting is equally effective for small and large language models, since reasoning quality depends only on the prompt structure and not on model capacity.
- A) True
 - B) False**
15. How does LoRA (Low-Rank Adaptation) maintain parameter efficiency while adapting a frozen Large Language Model?
- A) By injecting and training two small rank decomposition matrices while keeping the original pre-trained weights frozen.**
 - B) By selective pruning of non-essential weights in the Feed-Forward layers of the transformer.
 - C) By adding soft prompts to the input embedding layer that act as the only trainable parameters.
 - D) By quantizing the entire model to 4-bit precision and training only the quantization scale factors.
16. The DeepSeek-V3 architecture utilizes a Mixture-of-Experts (MoE) configuration. For a model with 671B total parameters, what is the activation count per token?
- A) All 671B parameters are activated for every token to maximize reasoning depth.
 - B) It activates 37B parameters for each token to maintain computational efficiency.**
 - C) It activates only the shared expert Ns unless the router detects a complex query.
 - D) Parameters are activated based on a 4-of-64 expert configuration, totaling 236B parameters.
17. In Prompt Engineering, how does Top P (Nucleus) sampling technically restrict the token selection pool?
- A) It masks out any token with a probability lower than a fixed constant P.
 - B) It enforces a frequency penalty on tokens that have already appeared in the current sequence.
 - C) It divides the logits by a factor of P to increase the diversity of the output.
 - D) It limits the sampling pool to the smallest set of tokens whose cumulative probability exceeds the threshold P.**
18. In Multi-Teacher Knowledge Distillation, how is the collective knowledge typically synthesized to guide the student model?
- A) The student trains on a different subset of the dataset for each individual teacher model.
 - B) Only the teacher with the highest KL-divergence relative to the student is used for the loss calculation.
 - C) The student mimics the average response (logits or feature representations) across the ensemble of teachers.**
 - D) The teacher models are trained online while the student model's parameters remain fixed.
19. Chain-of-Thought (CoT) prompting is categorized as an emergent ability because:
- A) It is hard coded into the attention masks of modern Transformer architectures.
 - B) It significantly improves performance primarily as model scale increases**
 - C) It allows the model to update its internal weights during the inference pass via backpropagation.
 - D) It relies on the model's ability to use Scratchpads for storage outside the context window.

National University of Computer and Emerging Sciences

Islamabad Campus

20. Suppose you are designing a large-scale Multimodal RAG system for an enterprise knowledge base containing millions of images and documents. To reduce inference latency, the system applies aggressive embedding compression and combines modality-specific retrieval results only at the final response stage, without enforcing tight cross-modal coupling during representation learning.
- Which risk is most likely to increase as a direct consequence of these design decisions?
- A) **Semantic mismatch between retrieved visual and textual information**
 - B) Text extraction failures caused by degraded optical character recognition
 - C) Increased sensitivity to background artifacts present in images
 - D) Higher computational overhead during offline embedding generation
21. Technically, why does training an LLM require significantly more GPU VRAM than performing inference with the same model?
- A) Inference requires storing the entire history of the optimizer's momentum states.
 - B) Training requires the use of multiple context windows for every attention head simultaneously.
 - C) Inference models are typically required to be FP64, while training is performed at INT8.
 - D) **Training requires saving intermediate activations for the backward pass and storing gradients for every parameter.**
22. Given a BERT-Large architecture with 340 million parameters, calculate the minimum GPU VRAM required to store the model weights using FP32 precision (where 1 parameter = 4 bytes).
- A) **1.36 GB**
 - B) 2.68 GB
 - C) 3.72 GB
 - D) 4.08 GB
23. When implementing a RAG retriever, Semantic Embeddings are technically preferred over Sparse Embeddings because:
- A) Sparse embeddings are computationally more expensive to calculate due to high-dimensional zero-vectors.
 - B) **Semantic embeddings capture deeper contextual meanings rather than relying on exact lexical matching.**
 - C) Sparse embeddings are limited to a context window of 128 tokens.
 - D) Semantic embeddings use the [CLS] token to bypass the vector database search entirely.
24. The Self-Consistency technique improves CoT reasoning by:
- A) Forcing the model to generate the exact same answer path every time using greedy decoding.
 - B) Penalizing the model if its output exceeds the context window's token limit.
 - C) Training a secondary verifier model to check the logic of each intermediate step in real-time.
 - D) **Sampling a diverse set of reasoning paths and selecting the final answer via a majority vote of consistent outputs.**
25. A RAG-based question-answering system consistently returns fluent but factually incorrect answers, even though relevant documents exist in the vector database. Which component is MOST likely responsible for this failure?
- A) **Ineffective chunking leading to loss of fine-grained semantic matches**
 - B) Use of cosine similarity instead of Euclidean distance
 - C) Over-parameterization of the generator LLM
 - D) Excessive prompt length exceeding the LLM context window

National University of Computer and Emerging Sciences

Islamabad Campus

26. In the BERT architecture, what is the specific technical role of the [CLS] token?
- A) To act as a delimiter separating Sentence A from Sentence B in the input stream.
 - B) To encode the positional index of the first token in the 3D embedding space.
 - C) To provide a target for the Masked Language Modeling task in the hidden layers.
 - D) To aggregate information from all tokens for sequence-level classification.**
27. Quantizing a model from FP32 to INT8 precision generally results in which technical outcome?
- A) A 25% reduction in memory usage with a linear increase in latency.
 - B) A 75% reduction in memory usage and faster integer operations on modern hardware.**
 - C) A 50% increase in model accuracy due to the reduction of floating-point noise.
 - D) The model becomes incompatible with standard Transformer attention mechanisms.
28. Least-to-Most prompting improves a model's Compositional Generalization by:
- A) Providing the model with the most complex example first to set a performance ceiling.
 - B) Reducing the context window size to force the model to generate more concise answers.
 - C) Decomposing a complex task into a sequence of simpler sub-problems and solving them incrementally.**
 - D) Using a low temperature for the first half of the response and a high temperature for the second half.
29. A voice-based assistant performs well in quiet indoor environments but fails significantly in real-world mobile settings. Which component of the voice AI pipeline is the most likely bottleneck?
- A) Language modeling due to limited vocabulary
 - B) Decoder beam width selection
 - C) Text normalization during post-processing
 - D) Acoustic modeling sensitivity to background noise and reverberation**
30. Quantizing a BERT-Base model (110M parameters) from FP32 (4 bytes per parameter) to INT8 (1 byte per parameter) results in a memory saving of how many megabytes?
- A) 110 MB
 - B) 220 MB
 - C) 330 MB**
 - D) 440 MB
31. A system must answer questions about rapidly changing financial data (e.g., daily stock prices) while maintaining transparency by referencing original documents. Which approach is MOST appropriate?
- A) Fine-tuning the LLM weekly with updated data
 - B) Using RAG with a non-parametric external knowledge store**
 - C) Increasing model size to reduce hallucinations
 - D) Applying self-distillation on financial reports
32. In a Nucleus (Top-p) sampling strategy with $p = 0.80$, five tokens have the following probabilities: A (0.45), B (0.30), C (0.15), D (0.05), and E (0.05). Tokens are sorted in descending order of probability and added until the cumulative probability reaches or exceeds p . Which tokens are included in the nucleus set?
- A) Tokens A only
 - B) Tokens A and B
 - C) Tokens A, B, and C**
 - D) Tokens A, B, C, D, and E

National University of Computer and Emerging Sciences
Islamabad Campus

33. An organization is evaluating the use of an Agentic AI system that performs multi-step reasoning, tool invocation, and environment interaction. In which of the following scenarios would an agent-based approach be least appropriate, despite the availability of high-quality models?
- A) A compliance monitoring system that periodically queries regulations, checks policy updates, and revises recommendations over time
 - B) A conversational data-analysis assistant that iteratively queries databases and refines hypotheses based on intermediate results
 - C) A latency-critical signal processing pipeline that must produce deterministic outputs within microsecond-level deadlines**
 - D) A decision-support agent that explores alternative strategies, executes simulations, and adapts actions based on observed outcomes
34. A production AI assistant dynamically selects relevant conversation history, retrieves user-specific preferences, queries external documents, and invokes tools before forming the final prompt sent to the language model. The system is designed to support many tasks rather than a single fixed prompt. Which concept best describes this approach?
- A) Prompt engineering
 - B) In-context learning
 - C) Fine-tuning
 - D) Context engineering**
35. Based on relation-based knowledge distillation using Maximum Mean Discrepancy (MMD), if the average teacher feature value is 2.33 and the average student feature value is 2.20, what is the MMD value?
- A) 0.0030
 - B) 0.2800
 - C) 0.0169**
 - D) 0.1178
36. In a Transformer architecture utilizing Multi-Head Attention with $d_{\text{model}} = 512$ and $h = 8$ heads, what are the specific dimensions of the weight matrices W^Q , W^K , and W^V for each individual head?
- A) 512×512
 - B) 512×64**
 - C) 64×512
 - D) 64×64
37. A company must frequently update factual information (e.g., policies, manuals) without retraining its language model. The system must also allow explicit inspection of the source documents used to answer queries. Which design choice and knowledge type best satisfies these requirements?
- A) Fine-tuning the model, relying on parametric knowledge only
 - B) Fine-tuning combined with weight freezing, relying on semi-parametric knowledge
 - C) Retrieval-Augmented Generation using non-parametric knowledge accessible at inference time**
 - D) Pretraining from scratch with larger model capacity to improve memorization

National University of Computer and Emerging Sciences
Islamabad Campus

38. DeepSeek-V3 utilizes a Mixture-of-Experts (MoE) gating mechanism where the scores for selected experts are normalized into probabilities (gates). If the top-2 routed experts for a token have logit scores of $s_1 = 2.0$ and $s_2 = 1.0$, what is the approximate normalized gating score g_1 for the first expert? (Assume $e^2 \approx 7.39$ and $e^1 \approx 2.72$)
- A) 0.500
 - B) 0.667
 - C) 0.739**
 - D) 0.269
39. In the Model Context Protocol (MCP), a single MCP client can maintain one shared connection and simultaneously interact with multiple MCP servers without requiring separate client instances.
- A) False**
 - B) True
40. According to the “Born-Again Networks” paradigm in self-distillation, what is the observed relationship between the teacher (T) and successive student generations (S_k) using identical architectures?
- A) Network Accuracy $T > S_1 > S_2 > S_k$
 - B) Network Accuracy $T = S_1 = S_2 = S_k$
 - C) Network Accuracy $T < S_1 < S_2 < S_k$**
 - D) Network Accuracy is randomized across generations
41. When implementing Multimodal RAG, how does the Early Fusion approach differ technically from Late Fusion?
- A) Late fusion merges embeddings before passing them to the LLM
 - B) Early fusion merges embeddings of different modalities before they are processed by the fusion model**
 - C) Early fusion only supports text-to-text retrieval
 - D) Late fusion requires a shared embedding space like CLIP
42. Within the Model Context Protocol (MCP), the Elicitation flow allows a server to request information from a user. Which participant in the MCP architecture is responsible for presenting the elicitation UI to the human user?
- A) The MCP Server
 - B) The MCP Client
 - C) The MCP Host (AI Application)**
 - D) The Remote Database
43. A Transformer-based language model has a vocabulary size of 50,000 tokens. During autoregressive text generation, the model predicts the next token at each time step. How many probability values are produced by the model’s final softmax layer for each next-token prediction?
- A) 512
 - B) 1
 - C) 50,000**
 - D) Equal to the number of attention heads

National University of Computer and Emerging Sciences
Islamabad Campus

44. In a Transformer model with $d_{\text{model}} = 1024$ and 16 attention heads, what is the value of the scaling factor used in the Scaled Dot-Product Attention mechanism?
- A) 4
 - B) 8**
 - C) 16
 - D) 32
45. 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.**
46. 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**
47. In the ReAct prompting framework, an agent follows a structured reasoning cycle when solving a task. Which sequence best represents this cycle?
- A) Action → Thought → Observation → Answer
 - B) Observation → Thought → Action → Planning → Output
 - C) Thought → Observation → Action → Memory Update
 - D) Thought → Action → Observation → Reflection → Repeat**
48. In a Vision Transformer (ViT), image patches are flattened and passed through a linear projection before being processed by the Transformer encoder. What is the primary purpose of this linear embedding step?
- A) To convert image patches into fixed-size, lower-dimensional representations suitable for self-attention**
 - B) To introduce spatial convolutional inductive bias into the model
 - C) To perform nonlinear feature extraction from raw pixel values
 - D) To reduce the number of attention heads required in the Transformer
49. An organization uses a large language model to perform a newly defined text classification task. The model is not retrained. Instead, the prompt includes a natural-language task description and one carefully chosen input–output demonstration, after which the model is asked to classify unseen inputs.
- Which in-context learning strategy is being applied in this scenario?
- A) Zero-shot learning
 - B) One-shot learning**
 - C) Few-shot learning
 - D) Fine-tuning-based learning

National University of Computer and Emerging Sciences
Islamabad Campus

50. In a language model, decoder outputs are decomposed into a reasoning path r_i and a final answer a_i . Which of the following cases demonstrates a minor commonsense reasoning inconsistency, where the reasoning appears plausible but does not logically support the final answer?

A)

Q: Would a glass cup crack if dropped onto concrete?

r_i : Concrete is a rigid surface that can transfer high impact forces to fragile objects.

a_i : Yes, the glass cup would likely crack.

B)

Q: Would a candle continue burning without oxygen?

r_i : Combustion requires oxygen to sustain a flame.

a_i : No, the candle would not continue burning.

C)

Q: Would a metal spoon feel cold at room temperature?

r_i : Metal conducts heat efficiently, allowing it to transfer heat away from the skin faster than other materials.

a_i : No, the metal spoon would feel warm to the touch.

D)

Q: Would wet clothes dry faster on a windy day?

r_i : Wind increases evaporation by removing moist air from the surface of the clothes.

a_i : Yes, wet clothes dry faster on a windy day.

Rough Work

National University of Computer and Emerging Sciences

Islamabad Campus

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 type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>	<input checked="" type="radio"/>
NAME:		ROLL NO:	
<div style="text-align: center;">Roll No</div> <div style="text-align: center;"> <input type="text"/> <input type="text"/> <input type="text"/> <input type="text"/> </div>			
0	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3	<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"/>
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"/>
MCQs			
Section1			
A	B	C	D
1	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
2	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
3	<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"/>
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"/>
10	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
A	B	C	D
11	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
12	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
13	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
14	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
15	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
16	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
17	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
18	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
19	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
20	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
21	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
22	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
23	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
24	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
25	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Generative AI (AI4009)

Final Exam Part-II

Date: December 31, 2025

Time (Hrs): 1.5

Course Instructor

Total Marks: 50

Dr. Akhtar Jamil

Total Questions: 3

Roll No

Section

Student Signature

Do not write below this line

Part-II

Attempt all questions.

[CLO 1-3]

Question No 2. [3 x 10=30]

Write short answers to the following questions.

Q.1. A teacher model produces logits $z_T = [6, 2, 0]$ for a 3-class classification problem. Compute the softmax probabilities when the temperature $T = 1$.

Answer:

$$p_i = \frac{\exp\left(\frac{z_i}{T}\right)}{\sum_{j=1}^3 \exp\left(\frac{z_j}{T}\right)} \quad p_i = \frac{e^{z_i}}{\sum_j e^{z_j}}$$
$$\sum_j e^{z_j} = e^6 + e^2 + e^0 = 403.43 + 7.39 + 1 = 411.82$$
$$p = \left[\frac{403.43}{411.82}, \frac{7.39}{411.82}, \frac{1}{411.82} \right] \approx [0.979, 0.018, 0.002]$$

Q.2. In mutual learning, what happens when KL divergence is replaced by an L2 loss on logits for prediction alignment?

Answer: KL divergence measures the distance in terms of probability distributions. When changed to the logits, training will become sensitive to the absolute scale and offsets of logits. As a result, the learning process becomes less stable or less informative knowledge transfer compared to KL-based alignment.

Q.3. What is soft prompting, and how does it differ from traditional (hard) prompting in large language models?

Answer: Soft prompting is a parameter-efficient adaptation technique in which a small set of continuous, trainable prompt embeddings is learned and prepended to the input of a large language model while keeping the model weights frozen. In contrast, traditional (hard) prompting relies on manually designed natural-language text prompts composed of discrete tokens.

National University of Computer and Emerging Sciences

Islamabad Campus

The key difference is that soft prompting optimizes prompt representations via gradient descent and is not human-readable, whereas hard prompting uses fixed textual instructions to guide model behavior without introducing new trainable parameters.

Q.4. A 7B-parameter model is stored using FP32 precision. Estimate the memory required to store the model weights.

$$\begin{aligned}\text{Memory} &= 7 \times 10^9 \times 4 \\ &= 28 \times 10^9 \text{ bytes} \\ &= 28 \text{ GB}\end{aligned}$$

Answer:

Q.5. In the Model Context Protocol (MCP), why are Resources categorized as Application-controlled while Tools are Model-controlled?

Answer: Resources are passive data sources provided as read-only context, whereas the model actively decides when to invoke a Tool to perform an action

Q.6. Technically, how does a Presence Penalty differ from a Frequency Penalty?

Answer: Presence penalty applies a constant fixed penalty regardless of how many times a token appears, while frequency penalty scales the penalty based on the total count of repetitions.

Q.7. What is the consequence of a Sliding Window in a context window that has reached its maximum token capacity?

Answer: The oldest tokens are dropped/discarded to make room for new ones, leading to a loss of long-term conversation memory.

Q.8. Explain the concept of Elastic Weight Consolidation (EWC).

Answer: Elastic Weight Consolidation (EWC) is a continual learning technique designed to reduce catastrophic forgetting when a neural network is trained sequentially on multiple tasks. It works by identifying parameters that are important for previously learned tasks and penalizing large changes to those parameters during training on new tasks. This allows the model to learn new tasks while preserving essential knowledge from earlier ones.

Q.9. Explain the concept of additive adapters in Fine-Tuning Transformer-based models.

Answer: Additive adapters are small, trainable neural modules inserted into the layers of a pre-trained Transformer model. Instead of updating the entire model, only these adapter modules are trained for a specific task. This enables parameter-efficient fine-tuning, allowing the model to adapt to new tasks while preserving the original pre-trained parameters and reducing computational and storage costs.

National University of Computer and Emerging Sciences

Islamabad Campus

Q.10. What happens to the diversity of generated text when the Top-p (nucleus sampling) value is decreased, and why?

Answer: When the Top-p value is decreased, the diversity of generated text decreases because the model samples from a smaller set of high-probability tokens, limiting variation in the output.

Question No 3. [5+5]

Q.1 What is Parameter-Efficient Fine-Tuning (PEFT)? Briefly describe how the following PEFT techniques adapt a pre-trained model while limiting the number of trainable parameters: BitFit, LayerNorm tuning, top-layer tuning, output head tuning, and Low-Rank Adaptation (LoRA).

1. **BitFit**

BitFit fine-tunes only the bias parameters of a pre-trained model while keeping all other weights frozen, enabling adaptation with minimal parameter updates.

2. **LayerNorm Tuning**

LayerNorm tuning updates only the scaling (γ) and shifting (β) parameters of normalization layers, allowing the model to adapt while preserving learned representations.

3. **Top-Layer Tuning**

Top-layer tuning fine-tunes only the final layers of a model, assuming that lower layers capture general features and higher layers encode task-specific information.

4. **Output Head Tuning**

Output head tuning trains only the final output layer of the model, mapping fixed internal representations to new task-specific outputs.

5. **Low-Rank Adaptation (LoRA)**

LoRA introduces trainable low-rank matrices into selected weight layers, approximating full weight updates with significantly fewer parameters for efficient fine-tuning.

Q.2 A machine translation system at **FAST-NUCES** is evaluated on an English sentence translation task. The two sentences are given below. Using only unigram precision, determine the final BLEU score.

Candidate translation:

“FAST NUCES offers quality education”

Reference translation:

“FAST NUCES offers high quality education”

Candidate tokens:

FAST, NUCES, offers, quality, education

Candidate length ($c = 5$)

Reference tokens:

FAST, NUCES, offers, high, quality, education

Reference length ($r = 6$)

National University of Computer and Emerging Sciences

Islamabad Campus

Matched unigrams = 5

Total candidate unigrams = 5

Unigram precision:

$$p_1 = 5/5 = 1.0$$

$$BP = \exp(1 - r/c)$$

$$r/c = 6/5 = 1.2$$

$$BP = \exp(1 - 1.2)$$

$$= \exp(-0.2) \approx 0.8187$$

$$BLEU = BP \times p_1$$

$$BLEU = 0.8187 \times 1.0 = 0.8187$$

Question No 4. [5+5]

Q.1. With the help of a diagram explain the Retrieval-Augmented Generation (RAG) pipeline.

Describe each component of the pipeline. Additionally, discuss how RAG helps mitigate limitations of standalone LLMs. Also list down disadvantages of RAG based systems.

Retrieval-Augmented Generation (RAG) Pipeline

Retrieval-Augmented Generation (RAG) is a framework that combines information retrieval with large language models to generate responses grounded in external knowledge sources. As illustrated in the diagram shown in class, the RAG pipeline is composed of three main stages: ingestion, retrieval, and synthesis.

Ingestion Phase

In the ingestion phase, raw documents such as PDFs, reports, web pages, or domain-specific files are collected and prepared for retrieval. They are divided into smaller chunks to make semantic matching more effective and to fit within model context limits. Each chunk is then converted into a dense vector representation using an embedding model that captures its semantic meaning. These embeddings are stored in a vector index.

Retrieval Phase

In the retrieval phase, the query is first transformed into an embedding using the same embedding model. This query embedding is then compared against the stored document embeddings in the index using a similarity metric such as cosine similarity. Based on this comparison, the system retrieves the top-K most relevant document chunks that are semantically closest to the query.

Synthesis Phase

In the synthesis phase, the retrieved document chunks are injected into the prompt of the large language model along with the user query. The language model then generates a response by synthesizing its parametric knowledge with the retrieved contextual information. This ensures that the generated output is more accurate, context-aware, and aligned with the external data, reducing the likelihood of unsupported or fabricated responses.

Mitigating Limitations of Standalone LLMs

- RAG helps address several limitations of standalone large language models by providing access to external knowledge at inference time.

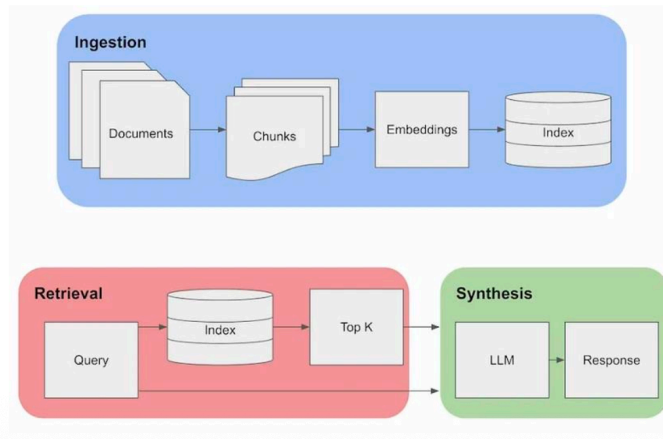
National University of Computer and Emerging Sciences

Islamabad Campus

- Because responses are grounded in retrieved documents, the risk of hallucination is significantly reduced and factual accuracy is improved.
- RAG also allows the system to incorporate up-to-date and domain-specific information without retraining the model, making it more flexible and cost-effective.

Disadvantages of RAG-Based Systems

- RAG-based systems introduce additional complexity compared to standalone language models.
- The system requires multiple components such as document preprocessing pipelines, embedding models, vector indexes, and retrieval mechanisms, which increases engineering and maintenance overhead.
- Retrieval adds latency to inference, and the overall quality of responses heavily depends on the effectiveness of chunking and retrieval strategies.
- The limited context window of language models restricts how much retrieved information can be used at once, and maintaining an up-to-date and consistent index requires continuous effort.



Q.2. Consider the following task:

“A logistics company operates warehouses in multiple cities. A shipment starts in City A, travels to City B, then City C, and finally City D. The cost of transportation between each city pair is given. The total budget is constrained, and the shipment must choose the cheapest valid route that satisfies a delivery time constraint.”

Using the scenario above answer the following questions:

(a) Write a Standard Prompt that directly asks the LLM model to solve the task. Explain briefly why this prompt is likely to fail on complex reasoning.

Standard Prompt: Given the transportation costs and delivery times between City A, B, C, and D, determine the cheapest valid shipment route that satisfies the delivery time constraint.

Failure Reason: This prompt asks for the final answer directly without guiding the model on how to reason. The model may:

- Skip intermediate comparisons
- Ignore constraints
- Produce a plausible but incorrect route
- Hallucinate assumptions about costs or time

National University of Computer and Emerging Sciences

Islamabad Campus

(b) Rewrite the prompt using Chain-of-Thought (CoT) prompting.

CoT Prompt: Given the transportation costs and delivery times between City A, B, C, and D, determine the cheapest valid shipment route that satisfies the delivery time constraint.

- Please solve the problem step by step.
- First, list all possible routes from City A to City D.
- Next, calculate the total cost and total delivery time for each route.
- Then, discard any route that violates the delivery time constraint.
- Finally, select the route with the lowest total cost and provide the final answer.

(c) Rewrite the prompt using Least-to-Most Prompting strategy.

Step 1: Identify Routes (Least Complex)

Prompt 1:

List all possible shipment routes from City A to City D, assuming the shipment must pass through City B and City C.

Step 2: Compute Costs and Times

Prompt 2:

For each route identified earlier, compute the total transportation cost and total delivery time. Present the results in a structured format.

Step 3: Apply Constraints

Prompt 3:

From the computed routes, remove any route that violates the delivery time constraint. Clearly state which routes are invalid and why.

Step 4: Final Decision (Most Complex)

Prompt 4:

Among the remaining valid routes, select the one with the minimum total cost. Provide the final route and justify your selection.