MID-SEMESTER EXAMINATION, November-2025
Large Language Models (CSE 4357)

Programme: B.Tech(CSE)
Full Marks: 30

Semester:7th
Time: 2 Hours

Subject/Course Learning Outcome	*Taxonomy Level	Ques. Nos.	Marks
Understand the fundamental concepts of language models, including tokenization and the representation of text as vector embedding for language processing	L2	1(a,b,c) 2(a,b,c)	12
Understand and explain the core mechanisms of the Transformer architecture used in modern large language models	L2	3(a,b,c)	6
Develop skills to categorize and cluster text data using large language models for text classification and clustering tasks	L2	4(a,b,c), 5(a,b,c)	12
Apply dense retrieval, reranking and retrieval augmented generation methods to enhance traditional keyword-based search systems			
Develop the ability to work with multimodal LLMs by understanding image-to-vector transformations and applying them to visual reasoning tasks			
Understand and implement end-to-end adaptation of LLMs—including data preparation, task-specific fine-tuning, and performance assessment			

*Bloom's taxonomy levels: Remembering (L1), Understanding (L2), Application (L3), Analysis (L4), Evaluation (L5), Creation (L6)

Answer all questions. Each question carries equal mark.

1. (a) What are Large Language Models (LLMs)? What makes a large language model "large"? 2

   (b) Explain how encoder-only architectures differ from decoder-only architectures in design and functionality. 2

(c) What is the Bag of Words model in Natural Language Processing, and how does it represent text data numerically for the following document collections? 2

D_1: NLP models process language
D_2: Language models learn patterns
D_3: Deep learning models process data

2. (a) Define subword tokenization and outline the Byte Pair Encoding procedure. 2

(b) Generate subword tokens from the phrase 'Large Language' using the WordPiece algorithm. 2

(c) Differentiate between static and contextual embedding. 2

3. (a) Discuss the importance of attention mechanisms in LLMs. 2

(b) Compute attention output using scaling with $d_k = 2$ of word "Large" in "Large language Models". Assume input embedding is $[0.4, 0.1, 0.3; 0.2, 0.5, 0.6; 0.9, 0.7, 0.1]$, learned weights $W_q = [0.1, 0.2; 0.8, 0.4; 0.3, 0.5]$, W_k and $W_v = [0.6, 0.5; 0.4, 0.1; 0.2, 0.9]$ 2

(c) How does self-attention differ from multi-headed and masked attention mechanisms? 2

4. (a) Differentiate between representation models and generative models for text classification. 2

(b) Examine the following four text documents, and each value represents the frequency of a term in the document. 2

$D_1(Sports) \rightarrow [3, 2, 7, 0, 1]$

$D_2(Sports) \rightarrow [2, 1, 9, 1, 0]$

$D_3(Politics) \rightarrow [8, 1, 3, 2, 0]$

$D_4(Politics) \rightarrow [6, 0, 2, 3, 1]$

Determine the predicted class (politics or sports) of a new document $D_{new} \rightarrow [2, 1, 5, 0, 4]$ based on the cosine similarity scores.

(c) Explain zero-shot and few-shot learning in LLMs with example. 2

5. (a) Compare and contrast text clustering and topic modelling in terms of their objectives and methodologies. 2

hard soft

(b) What is dimensionality reduction and why it is required before clustering? 2

(c) How does BERTopic extract representative keywords for clusters? 2

End of Questions