ASSIGNMENT
Large Language Models (CSE 4357)

Programme: B. Tech (CSE) **Semester: 7th**
Full Marks: 20 **Date of Submission: 20-12-2025**

Subject/Course Learning Outcome	*Taxonomy Level	Question Nos.	Marks
Understand the fundamental concepts of language models, including tokenization and the representation of text as vector embedding for language processing	L2	1	2
Understand and explain the core mechanisms of the Transformer architecture used in modern large language models	L2	2	2
Develop skills to categorize and cluster text data using large language models for text classification and clustering tasks.	L2	3, 4	4
Apply dense retrieval, reranking and retrieval augmented generation methods to enhance traditional keyword-based search systems	L3	5, 6, 7	6
Develop the ability to work with multimodal LLMs by understanding image-to-vector transformations and applying them to visual reasoning tasks	L2	8	2
Understand and implement end-to-end adaptation of LLMs—including data preparation, task-specific fine-tuning, and performance assessment	L2	9, 10	4

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

Answer all questions. Each question carries equal mark.

- **Assignment scores/markings depend on neatness, clarity and date of submission.**
- **Write your answers with enough detail about your approach and concepts used, so that the grader will be able to understand it easily.**
- **You are allowed to use only those concepts which are covered in the lecture class till date.**

1. Consider a small training corpus with word types and frequencies: low (5), lowest (2), newer (6), and wider (3). Using a SentencePiece tokenization procedure, perform 2 merge iterations and generate the sub-words**.** At each iteration:

    a. Compute Log- likelihood loss computation.
    b. Find the candidates for removal.
    c. Update the corpus representations and recompute the frequencies for the next iteration.

2. Explain the Transformer architecture, focusing on the self-attention mechanism. Discuss the role of the Query (Q), Key (K), and Value (V) vectors in computing the attention score for a token.

3. Consider three document embeddings:

D1 = [0.1, 0.5, 0.9]
D2 = [0.2, 0.6, 0.8]
D3 = [0.9, 0.2, 0.4]

a. Calculate pairwise Cosine similarity, Euclidean distance, and Dot product between all documents.
b. Explain which similarity metric is generally preferred for clustering text embeddings in high-dimensional spaces.

4. Explain how prompting affects the performance of large language models and provides examples of different types of prompting.

5. Compare semantic search with keyword-based search in terms of scalability, accuracy, and computational cost. Provide one example where keyword search may still be preferable. What is the role of embeddings in semantic search pipelines? How do they capture meaning beyond keywords?

6. How does the perplexity metric help in evaluating the quality of text generated by a language model? Explain briefly. Additionally, a language model assigns probabilities of 0.25, 0.10, 0.20, and 0.05 to a four-token sequence. Using these probabilities, calculate the perplexity of the model for this sequence and show the formula and final numerical value.

7. The figure below shows the output of an information retrieval system on two queries. Crosses correspond to the relevant documents, dashes to non-relevant documents. Let the two documents contain 3 and 6 relevant documents, respectively, but only those shown in the figure are retrieved by the system, not the others.

Rank	1	2	3	4	5	6	7	8	9	10
Q1	X			X	X					
Q2		X		X		X	X		X	

a. Calculate precision at K (where K=5)
b. Compute the average precision(AP)
c. Compute the mean average precision(MAP)
d. Normalized Discounted Cumulative Gain (nDCG) (where relevance grade is binary)

8. Explain what makes a language model “multimodal.” How does it differ from unimodal text-only LLMs?

9. State and compare sentence-level embeddings with document-level embeddings. Which would be more suitable for a legal search engine, and why?

10. Explain the difference between fine-tuning, instruction-tuning, and RLHF. How do these methods improve LLM performance and safety?