JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

Make-up Examination-Nov-2025

COURSE CODE (CREDITS):25B1WCI511(2) MAX. MARKS: 25

COURSE NAME: PROMPT ENGINEERING

COURSE INSTRUCTORS:VANI SHARMA MAX. TIME: 1 Hour 30 Minutes

***Note:Note:**(a)All questions are compulsory.*

(b) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

<table>
<tr><th>Q.No</th><th>Question</th><th>CO</th><th>Marks</th></tr>
<tr><td>Q1</td><td>A language model predicts the following next-token probabilities:<table><tr><td>Token</td><td>Probability</td></tr><tr><td>A</td><td>0.28</td></tr><tr><td>B</td><td>0.22</td></tr><tr><td>C</td><td>0.18</td></tr><tr><td>D</td><td>0.15</td></tr><tr><td>E</td><td>0.10</td></tr><tr><td>F</td><td>0.07</td></tr></table>(a) If greedy decoding is used, which token will be selected?
(b) If top-k sampling with $k = 3$ is used, identify the candidate set and compute the renormalized probabilities.
(c) Explain how temperature scaling could affect the diversity of the generated output.</td><td>[2]</td><td>[7]</td></tr>
<tr><td>Q2</td><td><b>(a)</b> Explain in detail how positional encoding enables Transformers to understand and process sequential data.Discuss why it is necessary despite the model's parallel processing nature, and describe the mathematical or conceptual mechanism used to represent positional information.

(b) Describe thoroughly how cross-attention facilitates encoder–decoder interaction in models such as T5 or BERT2BERT. Explain the role of query, key, and value vectors in this process and how this interaction allows the decoder to generate context-aware outputs.</td><td>[2]</td><td>[3+3]</td></tr>
<tr><td>Q3</td><td>(a) What are system prompts, and how do they influence model behaviour compared to user prompts?

(b)An LLM is tasked with solving a math word problem:
“If a shop sells 8 pencils for ₹40, how much will 15 pencils cost?”
You notice that the model sometimes gives direct, wrong answers.
To improve accuracy, you modify the prompt:</td><td>[3]</td><td>[2+3]</td></tr>
</table>

113

	"Think step by step before answering." Explain how *chain-of-thought prompting* changes the reasoning behavior of the model. What does this approach reveal about the model's internal reasoning process?		
Q4	(a) Compare the architecture and functionality of LSTM and GRU networks in detail. Explain how their gating mechanisms control information flow and address the vanishing gradient problem. (b) Write the mathematical equations of all gates used in LSTM (input, forget, output) and GRU (update, reset). Highlight the differences in how each model updates its hidden state. (c) Which of the two — LSTM or GRU — is generally considered more efficient, and why?	[2]	[3+3+1]