Roll No								

EndSemesterExamination 2024

Name of the Course: BTech(CSE)

Semester: 6th

NameofthePaper: ***Large Language Models and Generative AI***

Paper Code: TCS692

Time:3Hour's

Maximum Marks: 100

Note:

(i) All Questions are compulsory.
(ii) Answer any two sub questions among a,b and c in each main question.
(iii) Total marks in each main question are twenty.
(iv) Each question carries 10 marks.

Q1	**(10 X2 = 20 Marks)**	**CO1**
(a)	Explain the Model Architecture of the Transformer.	
(b)	What is the significance of multi-head attention in Transformers?	
(c)	Outline the steps involved in the encoding and decoding process within the Transformer model?	
Q2	**(10 X2 = 20 Marks)**	**CO2**
(a)	What is retrieval augmentation generation? Please highlight the approach for creating RAG application.	
(b)	Explain Text Splitting, Chunking and Embedding, with examples.	
(c)	What is Vector database and explain k-nearest neighbor algorithm.	
Q3	**(10 X2 = 20 Marks)**	**CO3**
(a)	Explain the concept of Prompt programming languages in NLP?	
(b)	Discuss the trade-offs between fine-tuning pre-trained language models and designing Prompts from scratch when approaching a new NLP task.	
(c)	Explain Chain-of-thoughts and Tree-of-thoughts prompt engineering techniques.	
Q4	**(10 X2 = 20 Marks)**	**CO4**
(a)	Explain the Life cycle of Generative AI project.	
(b)	Write a short note on Security and Ethical Concerns on Generative AI.	
(c)	What is Hallucination w.r.t Large Language model and possible ways to control it?	
Q5	**(10 X2 = 20 Marks)**	**CO5**
(a)	Explain Quantization in Generative AI.	
(b)	Write sHort note on quantization techniques.	
(c)	Write note on quantization methods viz. GGUF, GPTQ, NF4 and GGML.	