

25HCST402	LARGE LANGUAGE MODEL	Category	L	T	P	R	Credit	Year of Introduction
		VAC	4	0	0	0	4	2025

Preamble: This course covers the foundations, architectures, and applications of Large Language Models, including their evolution from early NLP to modern transformer-based systems. It explores prompting, model adaptation, alignment methods like RLHF, and key challenges such as hallucination, bias, toxicity, and privacy. Students also learn RAG and practical LLM deployment using evaluation, benchmarking, and LLMops to build responsible, scalable applications.

Prerequisite: Nil

Course Outcomes: After the completion of the course, the student will be able to:

CO	Course Outcomes	Knowledge Level
CO1	Explain the evolution of NLP, transformer architecture, attention mechanisms, and pre-trained LLM design principles.	K2 Understand
CO2	Apply prompt-based learning techniques, including prompt engineering and multi-prompt inference, to solve language tasks.	K3 Apply
CO3	Explain alignment techniques including RL-based approaches and apply strategies to mitigate hallucination, bias, toxicity, and privacy issues.	K3 Apply
CO4	Apply Retrieval-Augmented Generation (RAG) techniques and implement LLM-based applications using suitable evaluation metrics and deployment practices.	K3 Apply

Mapping of course outcomes with program outcomes and program specific outcomes

	PO1	PO2	PO3	PO4	PO5	PO6	PO7	PO8	PO9	PO10	PO11	PSO1	PSO2
CO1	3	3	3	-	-	-	-	-	-	-	2	3	1
CO2	3	3	3	3	-	-	-	-	-	-	2	3	2
CO3	3	3	3	3	-	-	-	-	-	-	2	3	2
CO4	3	3	3	3	-	-	-	-	-	-	2	3	3
Avg	3	3	3	3	-	-	-	-	-	-	2	3	2

Mapping of course outcomes with Knowledge and Attitude Profile (WK)

Wks Mapped with the course

	WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9
CO1	*	*	*	*	*			*	
CO2	*	*	*	*	*			*	
CO3	*	*	*	*	*			*	
CO4	*	*	*	*	*			*	
WK1	WK2	WK3	WK4	WK5	WK6	WK7	WK8	WK9	
*	*	*	*	*			*		

Mark Distribution:

Total Marks	CIE Marks	ESE Marks	ESE Duration
100	40	60	2.5 hrs

Continuous Internal Evaluation Pattern:

Attendance	: 5 marks
Continuous Assessment Test	: 20 marks
Continuous Assessment Assignment	: 15 marks

SYLLABUS**Module 1 (12 hrs)**

Large Language Models:- Introduction, Natural Language, NLP and Language Models Evolution, The Era of Large Language Models, Large Language Models in Practice. Language Models Pre-training– Encoder-Decoder Architecture, Attention Mechanism, Transformers, Data, Pre-trained LLM Design Choices, Commonly Used Pre-trained LLMs (BERT, T5, GPT, Mixtral 8x7B).

Module 2 (14 hrs)

Prompt-Based Learning– Introduction, Basics of Prompt-based Learning, Prompt Engineering, Answer Engineering, Multi-Prompt Inference. LLM Adaptation and Utilization – Introduction, Instruction Tuning, Parameter-Efficient Fine-Tuning, Compute-Efficient Fine-Tuning, End-User Prompting.

Module 3 (15 hrs)

Tuning for LLM Alignment :- Alignment Tuning, The Reinforcement Learning Framework, Mapping the RL Framework to LLMs with Human Feedback, Evolution of RLHF, Overcoming RLHF Challenges.

LLM Challenges and Solutions :- Hallucination, Bias and Fairness, Toxicity, Privacy.

Module 4 (15 hrs)

Retrieval-Augmented Generation– Introduction, Basics of RAG, Optimizing RAG, Enhancing RAG, Evaluating RAG Applications.

LLMs in Production :– Introduction, LLM Applications, LLM Evaluation Metrics, LLM Benchmark Datasets, LLM Selection, Tooling for Application Development, Inference, LLMOps.

Text Books:

1. Kamath, U., Keenan, K., Somers, G., & Sorenson, S. (2024). *Large language models: A deep dive*. Bridging Theory and Practice, Cham: Springer Nature.
2. Alammar, J., & Grootendorst, M. (2024). *Hands-on large language models: language understanding and generation*. " O'Reilly Media, Inc."

Reference Books:

1. Raschka, S. (2024). *Build a large language model (from scratch)*. Simon and Schuster.
2. Ozdemir, S. (2023). *Quick start guide to large language models: strategies and best practices for using ChatGPT and other LLMs*. Addison-Wesley Professional.
3. Chakraborty, T. (2024). *Introduction to large language models* (1st ed.). Wiley.

END SEMESTER EXAMINATION – MODEL QUESTION PAPER

Course Code	25HCST402
Course Name	LARGE LANGUAGE MODEL
Max. Marks: 60	Duration: 2.5 Hours

Part A <i>(Answer all questions. Each question carries 3 marks)</i>				
Qn No.	Question	Marks	CO	BL
1	Explain the evolution of NLP from statistical language models to transformer-based LLMs, highlighting key milestones.	3	1	L2
2	Describe the encoder-decoder transformer architecture and explain the role of the attention mechanism in capturing contextual relationships.	3	1	L2
3	Explain zero-shot, few-shot, and chain-of-thought prompting strategies with a suitable example for each.	3	2	L2
4	A company wants to adapt a Large Language Model (LLM) for customer support but has limited computational resources. Apply Parameter-Efficient Fine-Tuning (PEFT) using LoRA (Low-Rank Adaptation) to customize the model and explain how LoRA reduces the computational cost during fine-tuning.	3	2	L3
5	A company wants to align its LLM to avoid harmful outputs. Use the RLHF framework to design the feedback collection and reward modelling steps for this use case.	3	3	L3

6	Explain hallucination in LLMs, its causes, and two techniques used to mitigate it.	3	3	L2
7	A support team wants to automate responses using an LLM. Apply the RAG pipeline to design the document ingestion, retrieval, and answer generation stages for their FAQ system.	3	4	L3
8	Explain the LLM evaluation metrics - perplexity, BLEU, ROUGE, and BERTScore and their significance in model assessment.	3	4	L2
Part B (Answer any one full question from each Module. Each carry 9 marks)				
Module I				
Qn No.	Question	Marks	CO	BL
9	Explain the transformer architecture covering multi-head self-attention, positional encoding, and pre-training design choices. Briefly describe the BERT, GPT, and T5 models.	9	1	L2
OR				
10	Describe the evolution of NLP from rule-based to transformer-based LLMs. Compare encoder-only, decoder-only, and encoder-decoder architectures with suitable examples.	9	1	L2
Module II				
Qn No.	Question	Marks	CO	BL
11	A company wants to use a Large Language Model (LLM) for answering customer queries. Apply prompt-based learning by designing suitable prompt templates, performing answer engineering, and using multi-prompt inference to improve response quality. Explain how instruction tuning differs from end-user prompting in this application.	9	2	L3
OR				
12	Given a pre-trained LLM and limited hardware resources, apply an appropriate adaptation technique (LoRA, prefix tuning, or adapters) to fine-tune the model for a customer-support application. Explain how the selected method reduces the number of trainable parameters while maintaining performance.	9	2	L3
Module III				
Qn No.	Question	Marks	CO	BL
13. a	Construct an RLHF training pipeline for a customer support	5	3	L3

	chatbot: identify the data collection step, design a reward model, and outline the PPO-based policy optimization process.			
13. b	Given an LLM-based customer service application, apply suitable bias detection and mitigation strategies to ensure fair and unbiased responses. Explain how different sources of bias can affect the model's behavior and how they can be addressed.	4	3	L3
OR				
14. a	A text summarization model trained using Reinforcement Learning from Human Feedback (RLHF) generates repetitive summaries because of reward hacking. Apply Direct Preference Optimization (DPO) and Reinforcement Learning from AI Feedback (RLAIF) to improve the alignment process and explain how each method can be used to reduce repetitive outputs.	5	3	L3
14. b	A company deploys a Large Language Model (LLM) for customer support and discovers that the model occasionally generates toxic responses and reveals sensitive information from its training data. Apply suitable techniques to detect and mitigate toxicity and privacy risks, and explain how these techniques prevent data leakage caused by memorization. .	4	3	L3
Module IV				
Qn No.	Question	Marks	CO	BL
15. a	A legal firm wants to build a Retrieval-Augmented Generation (RAG) system to answer questions from legal documents. Apply an appropriate chunking strategy, retrieval method, and re-ranking technique to process the documents and explain how each component contributes to generating accurate answers.	6	4	L3
15. b	A company is deploying a Large Language Model (LLM)–based customer support system in production. Apply LLMOps practices to manage the deployment and explain how model serving, monitoring, and versioning can be used to ensure reliable operation. Also describe how these practices differ from those used in traditional MLOps.	3	4	L3
OR				
16. a	A medical chatbot built on RAG gives inconsistent answers. Apply faithfulness, answer relevancy, and context recall metrics to evaluate its sparse and dense retrieval configurations and recommend the better approach.	6	4	L3
16. b	A company wants to select an appropriate Large Language	3	4	L3

	Model (LLM) for developing an educational question-answering system. Apply benchmark datasets such as MMLU, HellaSwag, and TruthfulQA to evaluate candidate models and explain how the evaluation results can be used to choose a suitable LLM for the application.			
--	---	--	--	--
