Generative AI and Large Language models

Course Code: **CS5202**

Credits: **3-0-0**

Course Pre-requisites: **Python, Neural Networks, NLP**

This course provides a comprehensive introduction to Generative AI and large language models covering core techniques like Autoencoders, GANs, VAEs, and Diffusion Models. It also explores advanced tools such as Stable Diffusion, ControlNet, LoRA, and visual-language models like DALL-E, CLIP, and BLIP. Students will engage with recent progress in multimodal learning, including video and vision-language models and examine the key challenges associated with LLMs.

Course Outcomes:

After successful completion of course, students will be able to:

CO1: Explain the fundamental concepts and architectures of generative AI and Large Language Models including embeddings, transformers, and training objectives.

CO2. Demonstrate knowledge of prompt engineering techniques and Retrieval-Augmented Generation (RAG) pipelines to enhance GenAI models.

CO3. Develop multimodal generative AI systems using models such as CLIP, BLIP, Stable Diffusion, and ControlNet for various applications.

CO4: Design and implement small-scale GenAI projects, applying learned techniques to real-world problems.

Course Content:

Unit I: Introduction to Generative Artificial Intelligence

Introduction to Generative AI: Definition, History, Principles, Types of Generative models: Mixture models, Latent variable models, Variational inference, Autoencoders, Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Normalizing flows, Diffusion, Score-based Models, Energy-based models, Denoising Diffusion Probabilistic Model (DDPM).

Unit II: Introduction to Large Language Models

Introduction to large language models (LLMs), Transformers architecture, Concepts- Self-Attention mechanism, tokenization, Masked language modeling, Types of LLMs, Pretraining learning objectives, Key LLM parameters: mixture of experts, Context length, temperature, Sampling strategies; Continual pre-training, Supervised finetuning (SFT), Parameter-efficient finetuning (LoRA), Instruction tuning, Fine-tuning and Inference, RLHF: Overview, Reward modeling, PPO and variants, DPO, LLM reasoning, Scaling laws, LLM evaluation-intrinsic vs extrinsic metrics, LLM as a judge- pairwise evaluation, rubric-based scoring.

Unit III: Prompt Engineering and Retrieval Augmented Generation (RAG)

Introduction to prompts, Prompt engineering frameworks and principles, Types of prompting: zero-shot, few-shot, instruction prompting, Prompting strategies: in-context learning, Chain of thoughts, Self-consistency, Retrieval Augmented Generation (RAG), Advanced RAG, Applications of RAG, Challenges and Limitations, Comparative analysis of generative model families- GPT, Claude, Gemini, DeepSeek, Open source and commercial models.

Unit IV: **Multimodal Generative AI and LLMs**

Generative AI in vision, Vision-language models like DALL·E, Contrastive learning, CLIP, and BLIP, Diffusion-based image generation: Stable Diffusion, latent diffusion; ControlNet and controllable image generation; Introduction to Multimodal LLMs, Emerging multimodal LLMs and applications, Generative Models for different tasks such as Summarization, Q&A, Translation, healthcare, finance, manufacturing, and legal, Large Language Model Applications (e.g., conversational systems, code generation).

Unit V: Ethical, Societal, and Safety Considerations

Ethical considerations in AI development and deployment, Bias, fairness, explainablity, transparency, interpretability, Hallucination, Issues in LLMs, guidelines on generative AI usage, Misuse: deepfake, misinformation; Societal impact and safety considerations, Privacy and Data protection, evaluation of ethical failures, case studies in GenAI.

Grading criteria:

Quiz/In-class activity- 20%

Minor-I and Minor-II- 30%

Endterm- 20%

Projects- 30%

Resources:

1. All content (slides, assignments, quiz, suggested readings, research papers, code, etc.) will be shared during the course.

2. Inspired from the following courses:

    i. https://stanford-cs324.github.io/winter2023/syllabus/

    ii. https://www.cse.iitk.ac.in/pages/CS787.html

Suggested readings:

Foster D. Genera 6. Deep Generative Modeling, Jakub M. Tomczak, Springer 2022

Building a large language model from scratch -Generative deep learning. " O'Reilly Media, Inc.; 2023