MANIPAL SCHOOL OF INFORMATION SCIENCES
MANIPAL
(A constituent unit of MAHE, Manipal)

Workshop on Generative AI with LLMs

Summary:

Generative AI is at the forefront of technological innovations, with applications across diverse industries like healthcare, retail, sales, legal, and entertainment. This 3-day workshop will equip the participants with the essential skills to put to use cutting-edge Generative AI models for impactful solutions. The workshop provides a comprehensive exploration of Generative AI and Large Language Models (LLMs), combining foundational theory, practical implementation, and advanced applications. Participants will gain hands-on experience with the tools, techniques, and frameworks that are revolutionizing natural language processing (NLP) and AI-driven solutions.

Learning Objectives:

1. Learn fundamental concepts of word embeddings and transformers.
2. Explore techniques to utilize and fine-tune LLMs, both cloud-based and offline.
3. Learn and apply best practices in prompt engineering, retrieval-augmented generation (RAG), and agent-based systems.
4. Familiarize with tools and frameworks like Hugging Face, TensorFlow2, PyTorch, and llama-index.
5. Work on real-world project applications, including chatbots, automation tools, and data analysis systems.

Schedule:

Friday, December 6, 2024 through Sunday, December 8, 2024 (3 days) with daily schedule 9AM-1PM and 2-6PM in the Data Science Lab at MSIS. Project work will be executed post workshop under instructors' guidance.

Instructors:

(1) **Yeshwanth Sadum**
Data Scientist, VuNet Systems, Bengaluru
E-mail: sadumyeshwanth@gmail.com

(2) **Sudarsan N.S. Acharya**
Assistant Professor, Manipal School of Information Sciences
E-mail: sudarsan.acharya@manipal.edu, acharyan@alumni.stanford.edu

Topics:

Part-1 on Friday, December 6, 2024, from 9AM-1PM: **Introduction to Word Embeddings and Transformers**	• Implementing and exploring word embedding models. • Understanding the building blocks of the transformer model and implementing them.
Part-2 on Friday, December 6, 2024, from 2-6PM: **Introduction to LLMs**	• Accessing LLMs using cloud-based API (Groq) • Prompt techniques: zero shot, few shot, and chain of thoughts • Sentence embeddings • Similarity searches and re-rankers • Retrieval Augmented Generation (RAG): basic implementation
Part-3 on Saturday, December 7, 2024, from 9AM-1PM and 2-6PM: **Offline Models**	• Handling data types in PyTorch • Loading models from Hugging Face • Quantization, model size, and resource requirement estimation • Fine-tuning models: PEFT + QLoRA • Offline LLM implementations: Ollama
Part-4 on Sunday, December 8, 2024, from 9AM-1PM and 2-6PM: **Llama-index, LLM Observability, and Evaluation**	• Implementing RAG using llama-index • Evaluating an RAG framework • Tools and Agents: ReAct Agent using Llama-index

Project Ideas:

1. Tech support chatbot
2. HR assistant tool (e.g., indexing HR-related books)
3. Legal advisor (working with legal datasets)
4. News summarizer and QA tool (processing live data)
5. Sales automation tool (e.g., automating calculations)
6. Medical text classification
7. Information extraction from YouTube reviews