I-STEM®

Linking Researchers and Resources

Nadi

Transform with Nadi

Fundamentals, Fine-Tuning, Quantization, RAG, and Offline Deployment

🌐 www.istem.gov.in ☎ 1800-425-3281

✉ nodal.office@istem.co.in

SCAN TO REGISTER

Apply now

I-STEM
Linking Researchers and Resources

Nadi

Large Language Models (LLMs)

Transformers Fundamentals, Nadi Platform
Fine-Tuning, Quantization, RAG and Offline Deployment

Course Duration: 2 Months

Course Overview:

Course is designed to provide a comprehensive understanding of Generative AI, with a focus on the foundations of transformers, fine-tuning, quantization, RAG, offline deployment along with Python fundamentals with the help of our computer application called Nadi. Nadi is an offline Large Language Model (LLM) based AI PC application designed for AI education with a custom integrated curriculum for continuous and personalized learning.

Target Audience:

- This course is designed for individuals who want to explore the realm of generative AI who may or may not have a programming background.
- Individuals who are passionate about AI and want to delve deeper into the field of Generative AI.
- Professionals already working in the field of AI who want to expand their knowledge and expertise in Generative AI.
- Individuals with a programming background who want to explore the applications of Generative AI in their work.
- Undergraduate, postgraduate, PhD students pursuing degrees in computer science, AI, or related fields who want to gain practical experience in Generative AI.

Hardware Requirement:

- Any Laptop/PC with a Stable Internet Connection.

Classes and Hours

Class will commence from **10-05-2024**, and will be on **Wednesdays** from **7- 9 PM IST.** The program is conducted online using MS Teams.

SCAN TO REGISTER

Apply now

www.istem.gov.in
nodal.office@istem.co.in

1800-425-3281

What is the Benefit of the Course/ What Job I would Get

Students will have LLM fine-tuning, RAG and offline deployment experience after completion of the course, which gives an edge in interviews and the job best pursued after completion of the course is machine learning engineer or with additional qualifications students can pursue data science and become a data scientist. Getting equipped with latest tools on AI and technologies help you to stay ahead in the technology game and unlock the job opportunities out there in the unexplored domain of AI and GenZ.

Course Coding Environment:

In this course, we will be using Google Colab as the coding environment. Google Colab provides a free, cloud-based Python programming environment with pre-installed libraries and resources. This eliminates the need for participants to install Python or any specific IDE on their local machines.

With Google Colab, you can write and execute Python code directly in your web browser, making it convenient and accessible for all participants. It also allows for easy collaboration and sharing of code and notebooks.

We will also explore how Nadi will help you to learn AI better. During the course, we will provide detailed instructions on how to set up and use Google Colab, including accessing the notebooks, running the code, and saving your work. Please ensure that you have a Google account to access Google Colab and join the coding sessions seamlessly.

Course Curriculum:

Module 1:
Introduction to Generative AI, Foundations of Transformers, Nadi platform

- Understanding the basics of Generative AI.
- Introduction to Nadi and how it helps in learning generative AI.
- Understanding the architecture and components of transformers.
- Self-attention mechanism.

Module 2:
Basic Python Programming

- Python Foundations. Variable, data types, operators, control flow, functions, lists, strings, dictionaries, file handling, error handling, regular expression.
- PyTorch deep learning library fundamentals.

Module 3:
Open-Source LLM

- LLAMA-2/LLAMA-3 LLM introduction.
- Grasp the concept of fine-tuning: adapting a pre-trained LLM to a specific task.
- Learn about freezing vs. fine-tuning different layers of the pre-trained model.
- Understand choosing the right pre-trained model for your fine-tuning task.
- Explore data preparation techniques for fine-tuning LLMs.
- Instruction fine-tuning introduction.
- Practice instruction fine-tuning on LLAMA-2 on a custom dataset.
- Fine-tuning hyper parameters and mathematics.

SCAN TO REGISTER

Apply now

Course Curriculum:

Module 4:

Advanced Techniques - Quantization, PEFT, RAG and Offline Deployment

- Understand the concept of model compression: reducing LLM size for efficiency.
- Explore quantization techniques: reducing the number of bits used for model parameters.
- Learn about Q-LoRA (Quantized Low-Rank Adapters) for efficient fine-tuning.
- Grasp Parameter-Efficient Fine-Tuning (PEFT) for memory-constrained training.
- Retrieval Augmented Generation(RAG) and Vector Database with QnA on documents such as PDFs, Docs, Excel and so on using Langchain and LLamaIndex.
- Explore Supervised Fine-Tuning for efficient fine-tuning LLM.
- Learn about LLM deployment techniques: Offline or edge deployment of LLM.
- 4 bit quantization of LLM using LLaMA CPP and converting pre-trained or fine-tuned LLM to GGUF version for offline deployment in CPU/GPU ARM, X86/X64 or Metal. We will be looking into X86 deployment.