NVIDIA NIM for Beginners

Duration: 05 days (40 hours)

Labs: NVIDIA UI will be used for model consumption via User Interface tools
NVIDIA NIM free credits will be used for development via SDK

Pre-requisite: Fundamentals of Python and interested to use Generative AI models for business case scenarios

Module 01: Introduction of NVIDIA Platform

- Tour of NVIDIA Omniverse Cloud
- Working of NVIDIA NIM Architecture
- Key Features and Components
- Process of Deploying Generative AI model with NVIDIA NIM: https://developer.nvidia.com/blog/a-simple-guide-to-deploying-generative-ai-with-nvidia-nim/
- NIM for LLM Benchmarking Guide: https://docs.nvidia.com/nim/benchmarking/llm/latest/overview.html

Module 02: Deepdive into NVIDIA NIM (UI Platform)

- Reasoning Language LLMs
- Vision Language Models (VLM)
- Specialized Foundation Models
- Working with NVIDIA Edify
- Working with Open Diffusion Models
- Generate Embeddings for Text Retrieval
- Working with Reranking Models
- Working with Speech to Text Models
- Working with Convert Text to Speech Models
- Working with Translation Models
- Working with Protein Language Models
- Integrate Generative AI into OpenUSD Workflows
- Animation and Rendering
- Simulate and Optimize Real World Outcomes
- Chat With Your Industry Domain Expertise
- Drug Discovery, Medical Imaging & Genomics

Module 03: Deepdive into NVIDIA NIM using SDK

- Reasoning Language LLMs
- Vision Language Models (VLM)
- Specialized Foundation Models
- Working with Open Diffusion Models
- Generate Embeddings for Text Retrieval
- Working with Reranking Models
- Working with Speech to Text Models
- Working with Convert Text to Speech Models
- Working with Translation Models

Module 04: NVIDIA AI Enterprise Solution

- AI Chatbot with Retrieval Augmented Generation
- Cybersecurity AI Workflow
- Recommender Systems AI Workflow
- Retail Shopping Advisor AI Workflow
- Route Optimization AI Workflow
- Speech AI Workflows

- https://docs.nvidia.com/ai-enterprise/index.html#ai-workflows
- https://github.com/NVIDIA/GenerativeAIExamples

- https://docs.nvidia.com/
- https://developer.nvidia.com/

free nvidia courses to be launched:
https://www.nvidia.com/en-in/training/online/

Teaching Kits:
https://developer.nvidia.com/teaching-kits

- Introduction to Quantization
- Optimization of model weights (data types)
- Modes of Quantization
- Fine tuning LLMs (Meta's Llama / Alibaba's Qwen / Google's Gemma)
- Labs

Module 07: Evaluation of Open Source Models using MLflow

- Introduction to MLflow
- Build a machine learning model using MLflow
- MLflow Deployment Servers
- LLM Evaluation using MLflow
- Lab: Evaluate a Hugging Face LLM