NVIDIA Partner Expert Program

AI Expert Self-Paced Learning - Introductory

Suggested public content to watch before joining the AI Masterclasses

Getting Started with NVIDIA GPU Cloud Catalogue
NVIDIA NGC is a cloud platform that provides access to GPU-optimized software containers and pre-trained deep learning models. Learn how to generate an NGC API key in order to use the NGC service through the Docker client or through NGC CLI.

NVIDIA GPU Cloud – CLI Container Download
Quick demo of a workflow how to use the NGC CLI to download a container.

Simplify AI Workflows with Collections from NVIDIA NGC
NGC Collections makes it easy to discover containers, models, Jupyter notebooks, documentation and other resources to get started with your AI use cases.

AI Inference
Where to begin with Triton Inference Server? A powerful, open-source inference solution that standardizes model deployment and enables fast and scalable AI in production.

Recommenders
Large-scale recommender systems play a key role in on-line activities, influencing 35% of shopping, and as much as 75% of movie selection. Deep learning recommenders also provide better prediction at scale than traditional commercial recommenders. Learn how NVIDIA Merlin is an easy to use, open source, GPU accelerated platform that helps build, scale, optimize and deploy a deep learning recommender system.

Get Started with NVIDIA TAO Toolkit
NVIDIA TAO, built on TensorFlow and PyTorch, is a low-code AI solution that abstracts away the AI and deep learning framework complexity. TAO lets you use the power of transfer learning to fine-tune NVIDIA pretrained models with your own data. This video walkthrough will show you how to install the TAO Toolkit, followed by a step-by-step guide process of fine-tuning a pretrained model in a Jupyter notebook.

Build a Retrieval-Augmented Generation Chatbot in 5 Minutes
Large language models (LLMs) can be developed and deployed for AI chatbot applications — without needing your own GPU infrastructure— in under 5 minutes and with only 100 lines of Python code. This demo showcases how to design and create an enterprise-grade retrieval-augmented generation (RAG) pipeline using NVIDIA AI Foundation models.