

**Rapid Application Development Using Large Language Models**

**Course Number:** NVDA-108EC  
**Duration:** 1 day

**Overview**

This NVIDIA Rapid Application Development training course teaches attendees how to build cutting-edge applications fueled by Large Language Models (LLMs). Participants learn training, optimizing, and deploying strategies to unlock the full potential of LLMs.

**Prerequisites**

* Introductory deep learning, with comfort with PyTorch and transfer learning preferred
* Intermediate Python experience, including object-oriented programming and libraries

**Materials**

All attendees receive official courseware from NVIDIA in electronic format.

**Software Needed on Each Student PC**

The class will be conducted in a remote environment that Accelebrate will provide; students will only need a local computer with a web browser and a stable Internet connection. Any recent version of Microsoft Edge, Mozilla Firefox, or Google Chrome will be fine.

**Objectives**

* Grasp the motivations and challenges of training massive language models
* Master LLM integration and rapid application development workflows
* Understand resource limitations with advanced model reduction and optimization techniques
* Deploy LLM-powered apps to the real world
* Master the latest prompt engineering techniques

**Outline**

* Introduction to Training of Large Models
  + Learn about the motivation behind and key challenges of training large models
  + Get an overview of the basic techniques and tools needed for large-scale training
  + Get an introduction to distributed training and the Slurm job scheduler
  + Train a Megatron-LM-based GPT model using data parallelism
  + Profile the training process and understand execution performance
* Model Parallelism: Advanced Topics
  + Increase the model size using a range of memory-saving techniques
  + Get an introduction to tensor and pipeline parallelism
  + Go beyond natural language processing and get an introduction to DeepSpeed
  + Auto-tune model performance
  + Learn about mixture-of-experts models
* Inference of Large Models
  + Understand the challenges of deployment associated with large models
  + Explore techniques for model reduction
  + Learn how to use NVIDIA
  + Learn how to use Triton Inference Server
  + Understand the process of deploying GPT checkpoint to production
  + See an example of prompt engineering
* Conclusion