Rapid Application Development Using Large Language Models (RADLLM)

ID RADLLM **Price** 995.— €(excl. tax) **Duration** 1 day

Course Overview

Recent advancements in both the techniques and accessibility of large language models (LLMs) have opened up unprecedented opportunities for businesses to streamline their operations, decrease expenses, and increase productivity at scale. Enterprises can also use LLM-powered apps to provide innovative and improved services to clients or strengthen customer relationships. For example, enterprises could provide customer support via AI virtual assistants or use sentiment analysis apps to extract valuable customer insights.

In this course, you'll gain a strong understanding and practical knowledge of LLM application development by exploring the open-sourced ecosystem, including pretrained LLMs, that can help you get started quickly developing LLM-based applications.

Please note that once a booking has been confirmed, it is non-refundable. This means that after you have confirmed your seat for an event, it cannot be cancelled and no refund will be issued, regardless of attendance.

This course is part of the following Certifications

NVIDIA-Certified Associate: Generative AI LLMs (NCA-GENL)

Prerequisites

- Introductory deep learning, with comfort with PyTorch and transfer learning preferred. Content covered by DLI's Getting Started with Deep Learning or Fundamentals of Deep Learning courses, or similar experience is sufficient.
- Intermediate Python experience, including object-oriented programming and libraries. Content covered by Python Tutorial (w3schools.com) or similar experience is sufficient.

Course Objectives

By participating in this workshop, you'll learn how to:

- Find, pull in, and experiment with the HuggingFace model repository and the associated transformers API
- Use encoder models for tasks like semantic analysis, embedding, question-answering, and zero-shot classification
- Use decoder models to generate sequences like code, unbounded answers, and conversations
- Use state management and composition techniques to guide LLMs for safe, effective, and accurate conversation

Detailed Course Outline

Introduction

- Meet the instructor.
- Create an account at courses.nvidia.com/join

From Deep Learning to Large Language Models

- Learn how large language models are structured and how to use them:
  - Review deep learning- and class-based reasoning, and see how language modeling falls out of it.
  - Discuss transformer architectures, interfaces, and intuitions, as well as how they scale up and alter to make state-of-the-art LLM solutions.

Specialized Encoder Models

- Learn how to look at the different task specifications:
  - Explore cutting-edge HuggingFace encoder models.
  - Use already-tuned models for interesting tasks such as token classification, sequence classification, range prediction, and zero-shot classification.

Encoder-Decoder Models for Seq2Seq

- Learn about forecasting LLMs for predicting unbounded sequences:
  - Introduce a decoder component for autoregressive text generation.

- Discuss cross-attention for sequence-as-context formulations.
- Discuss general approaches for multi-task, zero-shot reasoning.
- Introduce multimodal formulation for sequences, and explore some examples.

Decoder Models for Text Generation

- Learn about decoder-only GPT-style models and how they can be specified and used:
  - Explore when decoder-only is good, and talk about issues with the formation.
  - Discuss model size, special deployment techniques, and considerations.
  - Pull in some large text-generation models, and see how they work.

Stateful LLMs

- Learn how to elevate language models above stochastic parrots via context injection:
  - Show off modern LLM composition techniques for history and state management.
  - Discuss retrieval-augmented generation (RAG) for external environment access.

Assessment and Q&A

- Review key learnings.
- Take a code-based assessment to earn a certificate.

About Fast Lane

Fast Lane is a global, award-winning specialist in technology and business training as well as consulting services for digital transformation. As the only global partner of the three cloud hyperscalers- Microsoft, AWS and Google- and partner of 30 other leading IT vendors, Fast Lane offers qualification solutions and professional services that can be scaled as needed. More than 4,000 experienced Fast Lane professionals train and advise customers in organizations of all sizes in 90 countries worldwide in the areas of cloud, artificial intelligence, cyber security, software development, wireless and mobility, modern workplace, as well as management and leadership skills, IT and project management.

Fast Lane Services

- ✓ High End Technology Training
- ✓ Business & Soft Skill Training
- ✓ Consulting Services
- ✓ Managed Training Services
- ✓ Digital Learning Solutions
- ✓ Content Development
- ✓ Remote Labs
- ✓ Talent Programs
- ✓ Event Management Services

Training Methods

- ✓ Classroom Training
- ✓ Instructor-Led Online Training
- ✓ FLEX Classroom – Classroom & Online Hybrid
- ✓ Onsite & Customized Training
- ✓ E-Learning
- ✓ Blended & Hybrid Learning
- ✓ Mobile Learning

Technologies & Solutions

- ✓ Digital Transformation
- ✓ Artificial Intelligence
- ✓ Cloud
- ✓ Networking
- ✓ Cyber Security
- ✓ Wireless & Mobility
- ✓ Modern Workplace
- ✓ Data Center

Worldwide Presence
with high-end training centers around the globe

Multiple Awards
from vendors such as AWS, Microsoft, Cisco, Google, NetApp, VMware

Experienced SMEs
with over 19.000 combined certifications

Germany
Fast Lane Institute for Knowledge Transfer GmbH
Tel. +49 40 25334610
info@flane.de / www.flane.de

Austria
ITLS GmbH
(Partner of Fast Lane)
Tel. +43 1 6000 8800
info@itls.at / www.itls.at

Switzerland
Fast Lane Institute for Knowledge Transfer (Switzerland) AG
Tel. +41 44 8325080
info@flane.ch / www.flane.ch