Hi everyone, welcome to the course "Hands-on Generative AI Engineering with Large Language Models"! My name is Quang Duong. I am excited to have you enrolled in this course.

This course is designed to equip you with the essential skills and knowledge to work effectively with Generative AI, particularly with large language models (LLMs) and develop applications using this advanced technology.

**Big Picture**

Firstly, I would like to give you a big picture. Generative AI is a sub-class of AI, machine learning, and deep learning. Generative AI models are trained on vast amounts of data. They take user instructions in various formats, such as text, voice, or images, as inputs. Then, they produce new content, including new text, images, and audio.

The scope of this course focuses on large language models, dealing with text data. These models take user text instructions to generate new text content. This process will be used to empower various AI applications, such as chatbots, AI assistants, and advanced information retrieval systems, and so on.

The objective of this course is to give you hands-on experience working with generative AI and large language models, and to build applications based on them.

**Introduction to the course**

In this introduction, I would like to present three main points to you: the course objectives, the course structure, and the learning paths.

**Course objectives**

Regarding the course objectives,

This course focused on hands-on engineering for Generative AI using Large Language Models (LLMs).

You will gain an in-depth understanding of how to build, implement, train, and perform inference with LLMs, such as Transformers.

The course will cover the essential components required to develop an LLM-based application, explore various tools and frameworks for building these applications, and provide practical knowledge on serving and deploying LLM-based apps.

Additionally, the course will offer guidance on advanced engineering topics in Generative AI, equipping you with the skills needed to excel in this rapidly evolving field.

**Course structure**

Regarding the course structure

The course is divided into seven parts, guiding you through hands-on, crucial aspects of generative AI, large language models, and components to build and deploy LLM-based applications.

In the first part, we will understand the software prerequisites for Python projects, set up your development environment, and get comfortable with the tools and libraries you'll be using throughout the course.

Next, in parts 2 and 3, we will dive into the "Attention Is All You Need" paper to understand the theoretical underpinnings of Transformers, how they revolutionized NLP, and the mechanics behind their operation. Then, we will translate your theoretical understanding into practice by implementing Transformers from scratch with PyTorch. This hands-on approach will solidify your understanding and give you practical experience.

In part 4, we will learn how to leverage the powerful tools and libraries provided by Hugging Face. This section will teach you how to use pre-trained models, datasets, and other resources to streamline your generative AI projects.

Finally, we will look at the last three parts, 5, 6, and 7:

In part 5, we will understand the building blocks required for creating LLM-based applications, including vector embeddings, prompt engineering, and integrating various LLM frameworks.

In the part 6, we will apply your knowledge to build concrete web applications. This part of the course will guide you through creating various AI assistants, chatbots, information retrieval system, agent applications, etc.

Then, in the part 7, we will learn how to deploy your applications using Docker, ensuring they are scalable and can communicate efficiently between frontend and backend components.

To sum up, these 7 parts will give you a comprehensive understanding of generative AI and large language models, from the foundational concepts to the development and deployment of LLM-based applications.

**Learning path**

Now you know the course structure. How to adapt it with your own learning path?

I designed the course in 7 parts, but we can group it into 4 groups:

Group 1: We Set up development environment and tools that serve all practical aspects of the course.

Group 2 is dedicated to Mastering Transformer models from theoretical to practical implementation.

Group 3 gives you a guide to leveraging the Hugging Face ecosystem to build advanced generative AI projects.

Group 4 provide Hands-on experience in building and deploying LLM-based applications in practice.

You can learn all of them one by one, or learn them separately, or skip the parts that you are already familiar with.

This course design will help you have a flexible learning path, depending on your experience and your needs. I hope it will provide you a flexible and effective learning experience.

Happy learning to you!