Course Glossary – Fundamentals of Building AI Agents using RAG and LangChain

Welcome! This alphabetized glossary contains many terms used in this course. Understanding these terms is essential when working in the industry, participating in user groups, and participating in other certificate programs.

Term	Definition
Bidirectional and Auto-Regressive Transformers (BART)	Sequence-to-sequence large language model (LLM) that follows an encoder-decoder architecture. It leverages encoding for contextual understanding and decoding to generate text.
Bidirectional Representation of Transformers (BERT)	An open-source, deeply bidirectional, unsupervised language representation pretrained using a plain text corpus.
Bradley-Terry model	A probability model for the outcome of pairwise comparisons between items, teams, or objects
Chain-of-thought (CoT)	An AI technique that simulates human-like reasoning by breaking down complex tasks into logical steps.
Chat model	A model designed for efficient conversations. It means that it understands the questions or prompts and responds to them like a human.
Context encoder	A neural network architecture used for image inpainting.

Contextual embeddings	A type of embedding that aptly describes how the transformer processes the input word embeddings by accounting for the context in which each word occurs within the sequence.
Data leakage	An organization faces challenges in exposing sensitive information.
Dense Passage Retrieval (DPR)	A set of tools that fetches relevant passages with respect to the question asked based on the similarity between the high-quality, low-dimensional continuous representation of passages and questions.
Facebook AI Similarity Search (Faiss)	It is a library developed by Facebook AI Research that offers efficient algorithms for searching through large collections of high-dimensional vectors.
Faiss index	A data structure that facilitates efficient similarities between vector searches.
Few-shot prompt	A technique where the model provides a small number of examples, usually between two and five, to adapt new examples from the previous objects.
Fine-tuning	A supervised process that optimizes the initially trained GPT model for specific tasks, like QA classification.
Generative pre-trained transformer (GPT)	A self-supervised model that involves training a decoder to predict the subsequent token or word in a sequence.
GitHub	A developer platform to create, store, manage, and share codes.
Graphic processing unit (GPU)	A process that helps to render graphic smoothly.
Hugging Face	Platform that offers an open-source library with pretrained models and tools to streamline the process of training and fine-tuning generative AI models.

In-Context learning	A technique in which task demonstrations are integrated into the prompt in a natural language format.
LangChain	An open-source interface that simplifies the application development process using LLMs. It facilitates a structured way to integrate language models into various use cases, including natural language processing or NLP and data retrieval.
LangChain – Core	A LangChain Expression Language and is the base for abstractions.
LangChain chains	Sequences of calls
Language model	A model that predicts words by analyzing the previous text, where context length acts as a hyperparameter.
Large language models (LLMs)	Foundation models that use AI and deep learning with vast data sets to generate text, translate languages, and create various types of content. They are called large language models due to the size of the training data set and the number of parameters.
Machine learning	Machine learning is a data analysis method for automating analytical model building.
Model inference	In machine learning, model inference refers to the operationalization of a trained ML model.
Natural language processing (NLP)	The subfield of artificial intelligence (AI) that deals with the interaction of computers and humans in human language. It involves creating algorithms and models that will help computers understand and comprehend human language and generate contextually relevant text in human language.

Prompt engineering	A process of creating effective prompts to enable AI models to generate responses based on the given inputs.
Prompt template	A predefined structure or a format that can be filled with specific content to generate prompts.
Python	A programming language.
PyTorch	A software-based open-source deep learning framework used to build neural networks, combining Torch's machine learning library with a Python-based high-level API.
PyTorch tensors	A fundamental data structure that is useful to represent a multi-dimensional array.
Retrieval-augmented generation (RAG)	RAG is an AI framework that helps optimize the output of large language models or LLMs. RAG uses the capabilities of LLMs in specific domains or the internal database of an organization without retraining the model.
Scoring function	Measures the summary for the evaluation of the point prediction. It means it predicts a property or a function.
Self-consistency	A technique for enhancing the reliability and accuracy of outputs.
Tokenization	The process of converting the words in the prompt into tokens.
Text Classifier	A machine learning technique that assigns a set of predefined categories to open-ended text.
Vector averaging	Process of calculating mean vector from a set of vectors.
watsonx.ai	A platform that allows developers to leverage a wide range of large language models (LLMs) under IBM's own series.

WatsonxLLM	A wrapper of IBM watsonx.ai foundation models.
Zero-shot prompt	A prompt in natural language processing (NLP) where a model can generate results for tasks that have not been trained explicitly.