Generative AI

Course Glossary: Architecture and Data Preparation for LLMs

Welcome! This alphabetized glossary contains many of the terms in this course. This comprehensive glossary also includes additional industry-recognized terms not used in course videos. These terms are essential for you to recognize when working in the industry, participating in user groups, and in other certificate programs.

Estimated reading time: 4 minutes

Term	Definition
Bidirectional and Auto-Regressive Transformers (BART)	Sequence-to-sequence large language model (LLM), which follows an encoder-decoder architecture. It leverages encoding for contextual understanding and decoding to generate text.
Bidirectional Encoder Representations from Transformers (BERT)	Large language model (LLM), which utilizes an encoder-only transformer architecture. It is exceptional at understanding the context of a word within a sentence, which is crucial for nuanced tasks like sentiment analysis.
Data analysis learning with language model for generation and exploration (DALL-E)	AI model developed by OpenAI, known for generating complex and creative images from textual descriptions using deep learning techniques
Data loader	Application component that enables efficient batching and shuffling of data, which is essential for training neural networks. It allows for on-the-fly preprocessing, which optimizes memory usage.
Data set	Collection of data samples and their labels
Diffusion model	Probabilistic generative AI model commonly used for image generation. A diffusion model is trained to generate images by learning to remove noise or reconstruct examples from its training data that have been distorted beyond recognition.
Fine-tuning	Adjusting a pretrained model to improve performance for a specific task or data set. This makes the model generate more accurate and contextually relevant content.
Generative adversarial network (GAN)	Generative AI model that can generate images from random input vectors or seed images. It consists of a generator and a discriminator, which work in a competitive mode.
Generative AI	Deep-learning models that can generate high-quality text, images, and other content based on the data they were trained on. These models are developed and trained to understand patterns and structures within existing data and apply the understanding to produce new and relevant data.
Generative pre-trained transformers (GPT)	Generative AI model based on transformer architecture. It has been pretrained on large amounts of text data and can predict and generate text sequences based on the patterns learned from its training data.
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.
Iterator	An object that can be looped over. It contains elements that can be iterated through and typically includes two methods: iter and next.
LangChain	Open-source framework that helps in streamlining AI application development using large language models (LLMs)
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.
Natural language processing (NLP)	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.
NLTK	Python library used in natural language processing (NLP) for tasks such as tokenization and text processing
Pydantic	Python library that helps streamline data handling. It can be used to parse and validate your data.
PyTorch	Dynamic deep learning framework developed by Facebook's AI Research lab. It is a Python-based library well-known for its ease of use, flexibility, and dynamic computation graphs.
Recurrent neural networks (or RNNs)	Artificial neural networks that use sequential or time series data. RNNs are used to solve data-related problems with a natural order or time-based dependencies.
SentencePiece	Subword-based tokenization algorithm that segments text into manageable parts and assigns unique IDs
spaCy	Open-source library used in natural language processing. It provides tools for tasks such as tokenization and word embeddings.
TensorFlow	Open-source machine learning framework. It provides a set of tools and libraries to facilitate the development and deployment of machine learning models.
Text-to-Text Transfer Transformer (T5)	Transformer-based large language model, which uses a text-to-text framework. It leverages encoding for contextual understanding and decoding to generate text.
Tokenization	Breaking text into smaller pieces or tokens. The tokens help a generative AI model understand the text better.
Tokenizer	Program that breaks down text into individual tokens
Transformers	Deep learning models that can translate text and speech in near-real-time. They take data, such as words or numbers, and pass it through different layers, with information flowing in one direction.
Unigram	Subword-based tokenization algorithm that breaks text into smaller pieces. It begins with a large list of possibilities and gradually narrows down based on how frequently they appear in the text.
variational autoencoders (VAEs)	Generative AI model that operates on an encoder-decoder framework. The encoder network first compresses input data into a simplified, abstract space that captures essential characteristics. The decoder network then uses this condensed information to recreate the original data.
WaveNet	Generative AI model designed for generating audio content. It can be used for tasks such as speech synthesis.
WordPiece	Subword-based tokenization algorithm that evaluates the benefits and drawbacks of splitting and merging two symbols to ensure its decisions are valuable.