Glossary: Generative AI Engineering and Fine-Tuning Transformers

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.

Estimated reading time: 5 minutes

Term	Definition
AdamW optimizer	A stochastic optimization method that helps in modifying the implementation of weight decay from gradient update.
Additive fine-tuning	A method that involves adding new task-specific layers or components to the pre-trained model.
AG News dataset	A subdataset of AG corpus of news articles.
Bidirectional representation of transformers (BERT)	An open-source model that offers deeply bidirectional, unsupervised language representations, pretrained on a plain text corpus.
Bidirectional representation of transformers (BERT) tokenizer	An important component for processing input data before inserting it into the BERT model.
ChatGPT	An artificial chatbot developed by OpenAI.
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.
Data loader	A utility in a machine learning framework that collects operational data from data sources at regular intervals.
Datapoint	An identifiable element in the dataset.
Direct preference optimization (DPO)	A successful fine-tuning strategy for aligning large language models (LLMs) with human preferences without training a reward model or employing reinforcement learning.
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.
Global vector for word representation (GloVe) dataset	An unsupervised learning algorithm for obtaining vector representation for word.
Graphic processing unit (GPU)	A process that helps to render graphic smoothly.
Hugging Face	A platform that offers an open-source library with pretrained models and tools to streamline the process of training and fine-tuning generative AI models.
IMBD dataset	A collection of information about movies, TV shows, and video games.
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 is the base for abstractions.
LangChain community	LangChain community is a third-party integrations that implement the base interfaces defined in LangChain Core, making them ready for use in any LangChain application.
Large language models (LLMs)	Foundation models that use AI and deep learning with vast datasets to generate text, translate languages, and create various types of content. They are called large language models due to the size of the training dataset and the number of parameters.
Llama	A large language model (LLM) trained by Meta AI understands and responds to human input and generates human-like text.
Low-rank adaptation (LoRA)	A technique that quickly adapts machine learning models.
Low-rank transformations	Techniques that approximate large matrices by smaller matrices in order to make computations more efficient.
Machine learning	Machine learning is a data analysis method for automating analytical model building.
Natural language processing (NLP)	The subfield of artificial intelligence (AI) 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.
Neural network	Computational models inspired by the structure of the human brain. A neural network model comprises an input layer, one or more hidden layers, and an output layer.
Parameter-efficient fine-tuning (PEFT)	A method that adapts large pre-trained language models for new tasks with less computational costs.
Prefix tuning	The prefix tuning adds a sequence of learnable embeddings to the key and value vectors of the attention mechanism in each transformer layer.

Term	Definition
Prompt injection	Prompt injection inserts task-specific tokens or embeddings into the input text at multiple positions.
Prompt tuning	Prompt tuning is a set of continuous prompt embeddings that are pretended to the input text.
P-tuning	P-tuning incorporates learnable prompt embeddings directly into the input embedding space that are task-specific and learned during fine-tuning.
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 tokenizer	Converts character strings into tokens understood by different PyTorch models.
Quantized low-rank adaptation (QLoRA)	Combines low-rank adaptation with quantization, reducing the model's memory footprint and computational requirements.
Reinforcement learning from human feedback (RLHF)	A model that represents a fine-tuning approach and enhances model performance on specific tasks, proving particularly effective in chatbot development.
Reparameterization-based methods	Leverage the concept of reparametrizing network weights using low-rank transformations.
Selective fine-tuning	Updates only a subset of layers or parameters and works for other networks.
Soft prompts	Soft prompts are an advanced concept that involves modification in the input data to guide the pre-trained language models to achieve desired outputs.
Supervised fine-tuning (SFT)	A method commonly used in machine learning, especially when working with pre-trained models in transfer learning.
Supervised fine-tuning (SFT) trainer	A technique to enhance the performance of pre-trained AI agents.
Transformer model	A model that can translate text and speech in near real-time.
Tokenization	The process of converting the words in the prompt into tokens.
Vector	A mathematical object represented by a group of numbers commonly used in machine learning algorithms.
Weight-decomposed low-rank adaptation (DoRA)	Adjusts the rank of the low-rank space based on the magnitude of the components, optimizing the model's performance and efficiency.