GLOSSARY

Generative AI: Working with Large Language Models

With Jonathan Fernandes

Use the terms and definitions below to understand concepts taught in this course.

Transcript Search: note that you can search for terms spoken by the instructor during the course. To search videos, switch to the Transcript tab, then search for keywords using the In this video or In this course option.

Term	Definition
BERT	Bidirectional Encoder Representations from Transformers; A machine learning model for natural language processing introduced in 2018 by researchers at Google AI Language
BLOOM	BigScience Large Open-science Open-access Multilingual Language Model: A transformer-based large language model that is currently the largest multilingual open-access AI model
Chinchilla	A family of large language models created in 2022 by Google's AI research team, DeepMind, trained to investigate the scaling laws of large language models; it is a 70B parameter model trained as a compute-optimal model with 1.4T tokens
GLaM	A mixture of experts model that can be thought of as having different experts, where each is specialized for various inputs; the experts in each layer are controlled and activated by a gating network based on the input data
Gopher	An AI-powered search engine that leverages the power of machine learning and natural language processing to deliver accurate and relevant search results
GPT-3	Generative Pre-trained Transformer 3; A large language model that generates text using algorithms; its objective is to predict the next token given the preceding tokens
LLM	Large language model; A type of artificial intelligence model that has been trained to recognize and generate vast quantities of written human language
multi-head attention	A module for attention mechanisms that runs through an attention mechanism several times in parallel and combines knowledge of the same attention, pooling via different representation subspaces of queries, keys, and values

OPT	Open Pre-trained Transfer Language Models by Meta AI; A series of open-sourced large casual language models that perform similar to GPT-3
PaLM	Pathways Language Model; A transfer-based language model developed by Google AI and made up of two components, a transformer layer for the upper block and a mixture of experts layer for the bottom block
scaling laws	Rules that explain how the performance of AI systems changes as models get larger or work with more data; these laws assist with understanding how computer power, the amount of data used, the size of AI models, and other factors affect how well AI systems work
self-attention	A mechanism used in machine learning, particularly in natural language processing, that allows a model to weigh the importance of different parts of an input sequence when making predictions or generating outputs
transfer learning	An AI process that allows a pre-trained machine learning model to be used as a starting point for training a new model