The Generative AI GLOSSARY

Discover the key technical terms associated with generative AI and their meaning

Analytics Vidhya

A

Terms	Description
Activation Function	Function used in neural networks to introduce non-linearity, such as ReLU, sigmoid, and Tanh.
Agents	A software program that can interact with its environment, collect data, and use the data to perform self-determined tasks to meet predetermined goals.
AI Ethics	The branch of ethics that examines the moral implications and responsibilities associated with the creation and use of artificial intelligence.
API	Refer to Application Programme Interfaces that enable the flow of information between applications. An API would allow an AI chatbot to connect to a LLM by a third-party service provider.
Attention	A technique used in machine learning and artificial intelligence to improve the performance of models by focusing on relevant information.
Auto Merging Retriever	A technique used to break the document into multiple chunks of text and further breaks the "parent" chunks into smaller "child" chunks.

B

Terms	Description
Backpropagation	An algorithm used to train neural networks by adjusting weights based on the error rate obtained in the previous epoch.
Backward Diffusion	A process where a model learns to reverse the noise added during the forward diffusion process, reconstructing data from noise.
BARD	A conversational AI model developed by Google for generating human-like text
BERT	BERT or Bidirectional Encoder Representations from Transformers, is based on transformers, a deep learning model in which every output element is connected to every input element, and the weightings between them are dynamically calculated based upon their connection.

Terms	Description
Bias	Systematic error introduced into data or algorithms that leads to inaccurate or unfair outcomes.

C

Terms	Description
Chain of Density	A prompt engineering technique that focuses on increasing the informational density of the generated text.
Chain of Dictionary	A method that structures prompts to utilize dictionary-like definitions for precise and context-rich responses.
Chain of Emotion	A technique that structures prompts to elicit emotionally resonant and contextually appropriate responses.
Chain of Explanation	A prompt strategy that guides the model to provide detailed explanations or reasoning for a given topic or question.
Chain of Knowledge	A method that leverages structured knowledge prompts to elicit responses based on specific facts and information.
Chain of Numerical Reasoning	A method that structures prompts to engage the model in logical and mathematical reasoning for solving numerical problems.
Chain of Question	A prompt engineering strategy that involves a series of related questions to guide the model through a logical sequence of responses.
Chain of symbol	A technique that employs symbolic representations within prompts to generate structured and symbolically meaningful outputs.
Chain of Thought	A technique where a model reasons step-by-step through a problem, breaking it down into smaller, more manageable parts. This method allows the model to simulate human-like thinking patterns and improve its problem-solving abilities.

D

Terms	Description
Data Parallelism	Data Parallelism is a way of performing parallel execution of an application on multiple processors. focuising on distributing data across different nodes in the parallel execution environment.
DDPG	A DDPG is a reinforcement learning agent that searches for an optimal policy that maximizes the expected cumalative long-term reward.
DDPM	Stands for Denoising Diffusion Probablistic Model, learns to generate data by reversing a gradual diffusion process. Noise is added to the data in small increments until it becomes pure noise, and the model learns to denoise this data step-by-step, ultimately generating realistic data samples.
Decoder	A component of sequence-to-sequence models that processes input sequences into a fixed representation
Deep Learning	Deep learning is a method in artificial intelligence (AI) that teaches computers to process data in a way that is inspired by the human brain.
Deep Speed	An optimization library for deep learning applications, enhancing training efficiency and speed.
Denoising Autoencoder (DAE)	A type of autoencoder used to remove noise from data, often utilized in the backward diffusion process.
Diffusion Model	A Diffusion model are generative models, meaning that they are used to generate data similar to the data on which they are trained.

E

Terms	Description
Embedding Layer	The layer in a neural network that converts categorical data, such as words, into continuous vector representations.

Terms	Description
Encoder	A component of sequence-to-sequence models that processes input sequences into a fixed representation.
Epoch	A full pass through the entire training dataset during the training process of a neural network.

F

Terms	Description
Faithfulness	An algorithm used to train neural networks by adjusting weights based on the error rate obtained in the previous epoch.
Falcon AI	A tool used to scan reports to provide AI-powered analysis of the business listings.
Few Shot Prompting	A technique in NLP where a model is given a few examples of a task within the input prompt to guide its responses.
Fine-tuning	The process of taking a pre-trained model and making small adjustments to it on a new, specific task or dataset to improve performance.
Forward Diffusion	A process where noise is gradually added to data over a series of steps, transforming it into a noise distribution.
Foundation Model	A form of generative artificial intelligence (generative AI). They generate output from one or more inputs (prompts) in the form of human language instructions.
Fully sharded data parallelism	A technique in distributed training of deep learning models where both the model's parameters and optimizer states are sharded (divided) across multiple devices or nodes.

G

Terms	Description
GANs	GANs, a generative adversarial network is a class of machine learning framework and a prominent framework for approaching generative AI. It is a framework in which two neural networks compete to generate realistic data.
Gaussian Noise	A type of statistical noise with a probability density function equal to that of the normal distribution, commonly added during the forward diffusion process.
Gemini	A project or technology by Google DeepMind that aims to combine advanced techniques in large language models and reinforcement learning, potentially enhancing the capabilities and applications of AI systems.
Generative AI	A subset of artificial intelligence focused on generating new data that is similar to existing data. This includes generating text, images, music, and more, using models like Generative Adversarial Networks (GANs), Variational Autoencoders (VAEs), and Transformer-based models such as GPT-3 and BERT. Generative AI is widely used in applications like chatbots, content creation, drug discovery, and creative arts.
GLIDE	Stands for Guided Language to Image Diffusion for Generation and Editing, a generative model developed by Open AI that uses diffusion processes to generate and edit images based on textual descriptions.
GPT	Generative Pre-trained Transformer, a type of LLM developed by Open AI, known for generating coherent and contextually relevant text.
Gradient Descent	An optimization algorithm used to minimize the cost function by iteratively adjusting model parameters.

H

Terms	Description
Hidden state	The state in a recurrent neural network (RNN) or transformer that contains information about the input sequence seen so far.

Terms	Description
Hit rate	A metric used to measure the accuracy of a predictive model or recommendation system, defined as the ratio of the number of relevant items successfully retrieved to the total number of relevant items.
Hugging Face	Hugging Face is a machine learning (ML) and data science platform and community that helps users build, deploy and train machine learning models.
Hybrid Fusion Retriever	An advanced information retrieval technique that combines multiple retrieval methods, such as dense and sparse retrieval, to improve the accuracy and relevance of search results.
Hyperparameter	Settings or configurations used to control the training process of a machine learning model, such as learning rate or batch size.

I

Terms	Description
IPEX	Stands for Intel® Extension for PyTorch*, which is a library that optimizes PyTorch performance on Intel hardware, including CPUs and GPUs.
Image Recognisition	Identifies which object or scene is in an image.
Indexing	The process of organizing data to enable efficient retrieval of information. In the context of databases and search engines, indexing involves creating a data structure, that allows for quick searches and access to relevant relevant records or documents.

J

Terms	Description
Joint attention	In models like BERT, the ability to attend to the left and right context of a word simultaneously.

L

Terms	Description
Langchain	LangChain provides AI developers with tools to connect language models with external data sources. It is open-source and supported by an active community.
LangChain Legacy Syntax	Refers to the original or earlier version of the syntax used in the LangChain framework, which is designed for building applications with large language models (LLMs).
LangGraph	A framework or tool designed to structure and visualize relationships and dependencies between different language models and their components. LangGraph helps in understanding and managing complex interactions within multi-model systems
LangServe	A service or platform designed to deploy and manage large language models (LLMs) in production environments.LangServe provides infrastructure and tools to serve LLMs efficiently, ensuring scalability, reliability, and ease of integration
LangSmith	A tool or platform designed to enhance the development and deployment of applications using large language models (LLMs). LangSmith provides features for optimizing, fine-tuning, and integrating LLMs into various applications,
Large Language model	A model that assigns probabilities to sequnece of words and can generate text based on learned patterns
Latent Diffusion	A generative modeling technique where the diffusion process is applied in a latent space rather than directly on the data. This approach involves encoding data into a lower dimension latent space to generate new data points.
Latent Space	A lower-dimensional representation of data where similar data points are closer together, often used in generative models.
LIMA	A process or technique used to adapt pre-trained language models to specific tasks or domains by fine-tuning them on a smaller, task-specific dataset.
LlamaIndex	LlamaIndex is a simple, flexible data framework for connecting custom data sources to large language models (LLMs).

Terms	Description
LLMops	Stands for Large Language Model Operations and refers to the specialized methods and processes meant to accelerate model creation, deployment, and administration over its entire lifespan.
LoRA	Low-Rank Adaption is a techniquew desinged to refine and optimize large language model. It focuses on adapting only specific parts of the neural network.
LSTMs	Stands for Long Short-Term Memory, a type of recurrent neural network (RNN) architecture designed to effectively capture long-term dependencies in sequential data.

M

Terms	Description
Markov Chain	A stochastic process where the probability of each event depends only on the state attained in the previous event, often used in modeling the steps of diffusion proces
Midjourney	An AI tool used for generating images based on text descriptions. It leverages advanced neural networks and deep learning techniques to create high-quality, photorealistic images from natural language prompts
MLops	A set of practices that aim to deploy and maintain machine learning models in production reliably and efficiently. MLOps combines aspects of machine learning (ML), data engineering, and DevOps to streamline the model lifecycle, from development and training to deployment
Model Architecture	The structure and design of a machine learning model, including the arrangement and interactions of its components such as layers, nodes, and connections.
Model Parallelism	A technique in deep learning where different parts of a model are distributed across multiple devices (such as GPUs or CPUs) to parallelize the computation.

Terms	Description
MRR	Stands for Mean Reciporal Rank, a metric used to evaluate the effectiveness of a search or recommendation system. It calculates the average of the reciprocal ranks of the first relevant result for a set of queries, providing a measure of how quickly the system retrieves relevant information.
Multi-document agents	AI systems or tools designed to handle, analyze, and generate responses based on information from multiple documents. These agents integrate data from various sources to provide comprehensive answers, insights, or summaries,

N

Terms	Description
NLP	A field of artificial intelligence that focuses on the interaction between computers and humans through natural language. NLP involves the development of algorithms and models to understand, interpret, and generate human language
No-code AI	Empowers non-technical users to rapidally develop and deploy AI solutions without extensive coding knowledge.
Noise Schedule	A predefined sequence of noise levels applied to data during the forward diffusion process.

O

Terms	Description
One Shot Prompting	Shows the model one clear, descriptive example of what you'd like it to imitate. When this prompt is run, the model's response will be to classify 'It doesn't work' as positive or negative
Output Parsers	Output parsers are responsible for taking the output of an LLM and transforming it to a more suitable format. This is very useful when you are using LLMs to generate any form of structured data.

P

Terms	Description
Parallel paradigm	In the data parallel paradigm, there are many different data and the same operations (instructions in assembly language) are performed on these data at the same time. Parallelism is achieved by how many different data a single operation can act on.
PEFT	Stands for Parameter-efficient fine-tuning, used in a scenario where computational resources are limited or where large pre-trained models are involved.
Pipeline Parallelism	Pipeline parallelism extends on simple task parallelism, breaking the task into a sequence of processing stages. Each stage takes the result from the previous stage as input, with results being passed downstream immediately.
Positional Encoding	Positional encoding is used to provide positional information to the model. In detail, a position-dependent signal is added to each word embedding for each input sequence to help the model incorporate the order of words.
Prompt Engineering	Prompt engineering is the practice of designing inputs for AI tools that will produce optimal outputs. It involves experimenting with different prompts to guide the model and achieve desired responses or outputs.
Pyspark	PySpark is the Python API for Apache Spark, an open source, distributed computing framework and set of libraries for real-time, large-scale data processing.

Q

Terms	Description
QLoRA	An efficient finetuning approach that reduces memory usage enough to finetune a 65B parameter model on a single 48GB GPU while preserving full 16-bit finetuning task performance. QLoRA backpropagates gradients through a frozen, 4-bit quantized pretrained language model into Low Rank Adapters (LoRA).
Quantization	Improves performance by reducing memory bandwith requirement and increase cache utilization. With an LLM model, quantization process at different precision levels enables a model to be run on wider range of devices.

Terms	Description
Query Interface	Query interface is a type-safe way to achieve a safe downcasting and to allow interfaces to be aggregated to an object.

R

Terms	Description
RAG	Stands for Retrieval Augmented Generation, is an architectural approach that can improve the efficacy of large language model (LLM) applications by leveraging custom data.
Reconstruction Loss	A measure of the difference between the original data and the reconstructed data, often used to train denoising models.
Recursive Retriever	It helps in identifying relationships between document chunks as well as recursilvely retrieve related document chunks.
Reinforcement Learning	It is a machine learning (ML) technique that trains software to make decisions to achieve the most optimal results. It mimics the trial and error learning process that humans use to achieve their goals.
Relevance AI	Relevance AI is the home of the AI Workforce. Build and deploy AI Agents and Tools to your workforce without code. With advanced customization, magical deployment and multi-provider support, Relevance AI makes it easy to integrate large language models into your workflows to create powerful automations.
Responsible AI	Responsible AI is an approach to developing and deploying artificial intelligence (AI) from both an ethical and legal point of view. The goal of responsible AI is to employ AI in a safe, trustworthy and ethical fashion.
Retrievers	Basic components of the majority of search systems. They're used in the retrieval part of the retrieval-augmented generation (RAG) pipelines, they're at the core of document retrieval pipelines, and they're paired up with a Reader in extractive question answering pipelines.
RLHF	RLHF is a specific technique that is used in training AI systems to appear more human, alongside other techniques such as supervised and unsupervised learning.

Terms	Description
Runpod	RunPod is a cloud computing platform designed for AI, machine learning applications, and general compute.

S

Terms	Description
Sampling	The process of generating new data points from a learned distribution in generative models. It is primarily used to generate detailed images conditioned on text descriptions, though it can also be applied to other tasks such as inpainting, outpainting, and generating image-to-image translations guided by a text prompt.
Self-consistency Prompting	It is an approach that simply asks a model the same prompt multiple times and takes the majority result as the final answer.
Sentence window Retriever	It separates the embedding and sythesis processes, allowing for more granular and targeted information retrieval. Instead of embedding and retrieving entire text chunks, this method focuses on individual sentences or smaller units of text.
Spark ML	Spark's macine learning library is MLlib. Its goal is to make practical machine learning scalable and easy. At a high level, it provides tools such as: ML Algorithms: common learning algorithms such as classification, regression, clustering, and collaborative filtering.
Stable Diffusion	Stable Diffusion is a deep learning, text-to-image model released in 2022 based on diffusion techniques.
Streamlit	Streamlit is an open-source Python framework for data scientists and AI/ML engineers to deliver dynamic data apps with only a few lines of code.

T

Terms	Description
Tensor Parallelism	A technique in distributed computing that splits the computation of large neural network models across multiple devices by dividing the tensors (multi-dimensional arrays) involved in the computations, enhancing training efficiency and scalability.
Text to 3D	A technology that converts textual descriptions into three-dimensional models, leveraging natural language processing and computer graphics techniques to create detailed 3D representations based on text input.
Tokenization	The process of breaking down text into smaller units called tokens, which can be words, subwords, or characters, to enable easier processing and analysis by machine learning models in natural language processing tasks.
Transforners	A type of deep learning model architecture that uses self-attention mechanisms to process and generate sequences of data, such as text, enabling advanced natural language processing tasks.
Tree of Thought	A hierarchical structure used in artificial intelligence to represent multiple possible outcomes or pathways of a decision-making process.

U

Terms	Description
Underfitting	A modeling error that occurs when a model is too simple to capture the underlying patterns in the data.
UNet	A convolutional neural network architecture designed primarily for biomedical image segmentation. It features an encoder-decoder structure with symmetric skip connections, allowing for precise localization and context utilization in segmentation tasks.
Unsupervised Learning	A type of machine learning where the model is trained on data without labels.

V

Terms	Description
Variational Autoencoder (VAE)	Variational Autoencoders (VAEs) are generative models explicitly designed to capture the underlying probability distribution of a given dataset and generate novel samples.
Variational Inference	A method of approximating complex probability distributions through optimization, often used in training generative models.
Vector Database	A vector database, vector store or vector search engine is a database that can store vectors along with other data items.
Verify and Edit Prompting	A technique used in the development and fine-tuning of AI models, particularly in natural language processing, where initial prompts or inputs are evaluated and modified to ensure they produce the desired output. This iterative process helps improve the accuracy and relevance of the generated responses.

W

Terms	Description
Weight	A parameter within a neural network that is adjusted during training
Word Embedding	A learned representation for text where words that have the same meaning have a similar representation.

Z

Terms	Description
Zero-shot Prompting	A technique in natural language processing where a model is given a task it has not been explicitly trained on, using a carefully designed prompt to guide the model in generating the correct output. This allows the model to perform tasks without needing task-specific training data.