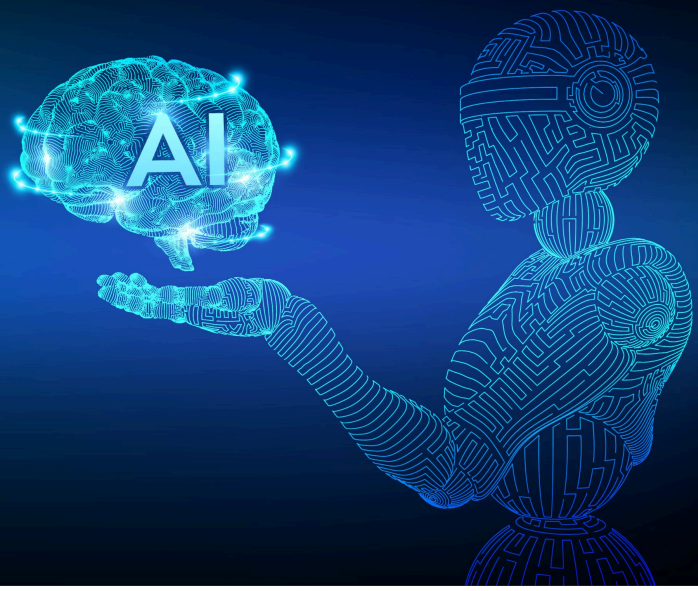
viamagus

Artificial Intelligence: Glossary



March 2025

Authors

Samkit Kankariya
Python AI Developer

Prajwal Mahajan
Node JS Developer

Namrata Sharma
Behavioral Scientist

Introduction

Artificial Intelligence (AI), evolves rapidly making it challenging to keep up with its ever-changing language. As interface design professionals, we've often found ourselves caught in technical jargon, sometimes leading to confusion. For example, one of us might suggest using a "convolutional neural network," while the other nods in agreement but quietly wonders if it's a data transmission method. We soon realized this isn't just our struggle—even top experts create glossaries to navigate AI terminology. Inspired by their efforts, we decided to compile our own glossary to clarify complex terms in this fast-moving field.

Glossary

- Artificial Intelligence
- Algorithms
- Big Data
- Chains
- Chat Model
- Completion Model
- Compute
- Cosine Similarity
- Dataset
- Deep Learning
- Embedding Models
- Fine-Tuning
- Generative AI (GenAI)
- Graphical Processing Unit (GPU)
- Graph Database (GDB)
- GraphRAG
- Hallucination
- Hugging Face
- Language Model
- LangChain
- LangGraph
- Large Language Model (LLM)
- LLM Agents
- Machine Learning
- Memory in LLM
- Model
- Multimodal AI
- Multi-Shot Prompting
- Multi-Query Retrieving
- Multi-Vector Retrieval
- Natural Language Processing (NLP)
- Neural Networks
- Overfitting
- Probabilistic Model
- Prompt
- Prompt Engineering
- Retrieval-Augmented Generation (RAG)
- Reranking
- Retriever Models
- Retrieval Process
- Sentence Transformers
- Single-Shot Prompting
- Supervised Learning
- Synthetic Data
- Temperature
- Tools/Toolkits In LangChain
- Token
- Training Data
- Transformers
- Unsupervised Learning
- Vectors
- Vector Database (VectorDB)
- Zero-Shot Prompting

Our goal is to create a resource that helps us, and our peers better understand AI's intricate landscape. By breaking down key concepts, we hope to improve collaboration and ensure everyone is on the same page. Whether you're a seasoned pro or just curious about AI, we believe this glossary will be a valuable tool on your journey through this exciting field.

Artificial Intelligence

AI refers to both a branch of computer science and a transformative technology focused on creating machines that can perform tasks requiring human intelligence. These tasks include learning, problem-solving, understanding language, and perceiving the world. AI systems are designed to simulate human cognitive processes and can be applied in various fields, from healthcare to autonomous vehicles.

Algorithms

Systematic procedures or rules designed to solve specific problems or perform tasks in AI applications. Algorithms form the backbone of AI systems, ranging from simple decision trees to complex neural networks.

For example, a search algorithm like Google's PageRank determines the most relevant web pages for a user's query by evaluating factors like the number of links pointing to a page.

Big Data

Extremely large and complex datasets that exceed the processing capabilities of traditional data management tools. Big Data is characterized by the "three Vs":

1. Volume - massive amounts of data
2. Velocity - high speed of data generation and processing
3. Variety - different forms of data

While specific thresholds vary, consider data as "Big Data" when it becomes challenging or impossible to process using standard methods on a single machine. This type of data is crucial for training advanced AI models.
Example: A social media platform like Twitter generates vast amounts of data daily from user tweets, likes, retweets, and follows, all of which are considered "big data."

Chains

Sequences of AI operations or models linked together to perform complex tasks. Chains allow for the creation of sophisticated AI workflows by combining simpler components, often used in natural language processing and multi-step reasoning.

Chat Model

A specialized language model designed for engaging in human-like conversational interactions. Chat models can understand context, maintain coherence across multiple turns, and generate appropriate responses in a dialogue setting.

Completion Model

An AI model that predicts and generates text based on a given prompt or context. These models are used in applications like autocomplete, code generation, and creative writing assistance, helping users complete their thoughts or tasks.

Compute (Noun)

The computational resources required to train, run, and maintain AI models and systems. This includes processing power (CPU/GPU), memory, and storage. The increasing complexity of AI models has led to a growing demand for high-performance compute infrastructure.

Cosine Similarity

A metric used to measure the similarity between two vectors by calculating the cosine of the angle between them. Imagine each piece of data (like a document or item) represented as an arrow (a vector) in a multi-dimensional space. Cosine similarity measures how closely these arrows point in the same direction, regardless of their length. In AI, it's often used in text analysis, recommendation systems, and information retrieval to compare how similar documents or items are in meaning or features. A cosine similarity of 1 means the vectors are identical, while 0 means they are completely dissimilar.

Dataset

A structured collection of data used for training, testing, and evaluating machine learning models. The quality, size, and diversity of datasets significantly impact the performance and generalization capabilities of AI models.

Example - A collection of labeled images of cats and dogs used to train a machine learning model to differentiate between the two.

Deep Learning

A subset of machine learning based on artificial neural networks with multiple layers. Deep learning models can automatically learn hierarchical representations of data, making them effective for tasks like image and speech recognition, and natural language processing.

For instance, Netflix uses deep learning to recommend movies by analyzing patterns in the viewing history of millions of users.

Embedding Models

AI models that convert high-dimensional data (such as text, images, or audio) into dense vector representations. These embeddings capture semantic relationships and features, making them valuable for tasks like similarity search, clustering, and as input for other machine learning models.

Fine-Tuning

The process of adapting a pre-trained model to a specific task or domain by further training it on a smaller, task-specific dataset. Fine-tuning allows for leveraging general knowledge while optimizing performance for particular applications, often resulting in improved accuracy and efficiency.

Example - A language model pre-trained on general text is fine-tuned with legal documents to perform better in legal text analysis.

Generative AI (GenAI)

AI models capable of creating new content, such as text, images, audio, or video, based on patterns learned from training data. Generative AI has applications in creative fields, content creation, and data augmentation, often producing surprisingly human-like or novel outputs.

Example - An AI tool like DALL-E generates original artwork based on text prompts provided by users.

Graphical Processing Unit (GPU)

Specialized hardware designed for parallel processing, originally created for computer graphics but now crucial in AI and machine learning. GPUs significantly accelerate the training and execution of complex neural networks and other AI models because their parallel processing capabilities allow them to perform many calculations simultaneously, which is essential for the intensive computations involved in training large AI models.

Graph Database (GDB)

A type of database that uses graph structures to represent and store data, with entities as nodes and relationships as edges. Graph databases excel at managing highly connected data and are useful in AI applications involving social networks, recommendation systems, and knowledge graphs.

GraphRAG

An advanced implementation of Retrieval-Augmented Generation that leverages graph databases for sophisticated information retrieval. GraphRAG enhances traditional RAG by utilizing the interconnected nature of data in graph databases, allowing for more context-aware and relationship-based information retrieval.

Hallucination

A phenomenon in AI, particularly in language models, where the model generates information that is false, nonsensical, or not supported by its training data. Addressing hallucinations is crucial for developing reliable AI systems, especially in applications requiring factual accuracy.

For example, a language model might confidently produce a fake historical fact during a conversation.

Hugging Face

A popular platform and community that provides tools, pre-trained models, and resources for natural language processing tasks. Hugging Face has become a central hub for sharing and accessing state-of-the-art language models and NLP technologies.

Language Model

An AI model designed to understand, generate, and manipulate human language based on patterns learned from large text datasets. Language models form the foundation of many natural language processing applications, from chatbots to translation systems.

Example - ChatGPT is a language model that can generate text responses based on the input it receives, like answering questions or writing essays.

LangChain

A framework for developing applications powered by language models, providing tools for managing prompts, handling responses, and integrating external data sources. Langchain simplifies the process of building complex AI workflows and applications using large language models.

LangGraph

A specialized tool for visualizing and analyzing the interactions and data flows within language model applications. Langgraph helps developers understand and optimize the behavior of complex AI systems by providing insights into how information is processed through various stages.

Large Language Model (LLM)

A neural network-based language model with billions of parameters, trained on vast amounts of text data. LLMs can understand context, generate human-like text, and perform a wide range of language-related tasks with minimal task-specific training. It's important to note that while LLMs are a powerful form of AI, not all AI systems are LLMs. LLMs specialize in language-based tasks, while AI encompasses a broader range of intelligent systems and algorithms.

LLM Agents

AI systems that leverage Large Language Models to perform tasks autonomously or assist humans in complex problem-solving. LLM Agents combine language understanding with additional components like memory and planning algorithms to handle multi-step problems and extended interactions.

Machine Learning

A subset of AI focused on developing algorithms and statistical models that enable computer systems to improve their performance on a specific task through experience. Machine learning allows systems to learn patterns from data without being explicitly programmed for each task.

Example - An email service uses machine learning to filter out spam emails by learning from examples of spam and non-spam messages.

Memory in LLM

The capability of Large Language Models to retain and utilize information from previous interactions or contexts within a conversation or task. Effective memory management in LLMs is crucial for maintaining coherence in extended dialogues and performing complex, multi-step tasks.

Model

In AI, a mathematical or computational representation designed to learn patterns from data and make predictions or decisions based on new inputs. Models range from simple linear regressions to complex neural networks, each suited for different types of problems and data.

Multimodal AI

Referring to AI systems or models capable of processing and integrating information from multiple types of input data, such as text, images, audio, and video. Multimodal AI aims to mimic human-like ability to synthesize information from various senses.

Multi-Shot Prompting (Few-Shot Prompting)

A technique where a model is provided with a few examples of the desired input-output behavior before being asked to perform a new task. This approach leverages the model's ability to learn from context and adapt to new scenarios without extensive retraining.

Multi-Query Retrieving

An information retrieval technique that involves generating and executing multiple queries to gather relevant information for a given task. This approach aims to improve the comprehensiveness and accuracy of retrieved information by exploring different aspects of the original query.

Multi-Vector Retrieval

An advanced information retrieval method that represents documents or information using multiple vector embeddings, each capturing different semantic aspects. This approach allows for more nuanced and comprehensive search results by considering various facets of similarity.

Natural Language Processing (NLP)

A branch of AI focused on enabling computers to understand, interpret, and generate human language in a valuable way. NLP combines linguistics, machine learning, and deep learning to process and analyze large amounts of natural language data. While LLMs are a significant advancement in NLP, NLP encompasses a broader range of techniques and tasks beyond just LLMs. For example, early NLP techniques involved rule-based systems, which are distinct from the neural network architectures of LLMs.

Neural Networks

Computational models inspired by the human brain's structure and function, consisting of interconnected nodes (neurons) organized in layers. Neural networks form the basis of many modern AI techniques, particularly in deep learning, and excel at pattern recognition tasks.

Example - A neural network powers facial recognition in smartphones, allowing the device to unlock when the owner's face is recognized.

Overfitting

A common problem in machine learning where a model learns the training data too well, including its noise and peculiarities, at the expense of its ability to generalize to new, unseen data. Overfitting results in poor performance on data outside the training set.

Probabilistic Model

A type of model that incorporates uncertainty and randomness into its predictions or decision-making process. Instead of producing single, deterministic outputs, probabilistic models generate probability distributions over possible outcomes, useful for scenarios involving uncertainty.

Prompt

In the context of AI, particularly with language models, a prompt is the initial input or instruction given to the model to elicit a desired response or behavior. Effective prompt design is crucial for guiding AI models to produce relevant and accurate outputs.

Example - Asking ChatGPT, "Tell me a story about a brave knight," is a prompt that guides the AI to generate a relevant story.

Prompt Engineering

The practice of designing, refining, and optimizing prompts to effectively communicate with AI models, particularly large language models. Prompt engineering involves understanding the model's capabilities and crafting inputs that guide the model towards producing desired outputs.

Retrieval-Augmented Generation (RAG)

A hybrid AI approach that combines large language models with external knowledge retrieval systems. RAG enhances the model's responses by first retrieving relevant information from a knowledge base and then using it to augment the context provided to the language model.

Example - a chatbot that retrieves real-time information from the web to provide more accurate and up-to-date responses. Instead of relying solely on pre-trained knowledge, it searches external sources before generating its response, making it more contextually relevant.

Reranking

A post-processing step in information retrieval where an initial set of results is rearranged based on additional criteria or more sophisticated models. Reranking aims to improve the relevance and quality of search results by considering factors beyond the initial matching algorithm.

Retriever Models

Specialized AI models designed to efficiently search and retrieve relevant information from large datasets or knowledge bases. Retriever models are crucial components in question-answering systems, search engines, and recommendation systems.

Retrieval Process

The series of steps involved in identifying and extracting relevant information from a large dataset or knowledge base in response to a query. This typically includes query processing, matching, ranking, and result presentation.

Sentence Transformers

Models specifically designed to generate meaningful vector representations (embeddings) of sentences or short text passages. These models enable more nuanced text comparison and are useful for tasks such as semantic search and measuring text similarity.

Example - A sentence transformer model is used to find similar sentences in a large document, such as identifying all sentences related to "customer satisfaction" in a survey report.

Single-Shot Prompting

A technique where a model is given a single example of the desired input-output behavior before being asked to perform a new task. This approach tests the model's ability to generalize from minimal task-specific information.

Supervised Learning

A machine learning paradigm where the model is trained on a labeled dataset, learning to map inputs to known outputs. Supervised learning is widely used in classification and regression tasks. A common example is email spam filtering, where the model learns to classify emails as "spam" or "not spam" based on labeled examples.

Synthetic Data

Artificially generated data that mimics the statistical properties and patterns of real-world data. Synthetic data is used to augment or replace real data in AI training, especially when real data is scarce, expensive, or subject to privacy concerns.

Example - Generating fake customer names and transaction records to test a new database system without using real personal data.

Temperature

A hyperparameter in language models that controls the randomness and creativity of the output. Lower temperature values produce more focused and deterministic responses, while higher values increase diversity and potential creativity.

Tools/Toolkits In LangChain

Pre-built components and utilities provided by the Langchain framework to facilitate the development of language model applications. These include text splitters, embeddings, vector stores, and memory components.

Token

The basic unit of text or data that language models process. Tokens can be words, subwords, characters, or punctuation marks, depending on the model's tokenization strategy. Think of tokens as the individual building blocks a language model uses to understand and generate text. While a prompt is the overall input you give to a language model, it is broken down into these individual tokens for processing. The number of tokens in a prompt or response affects the computational resources required for processing.

Training Data

The dataset used to teach a machine learning model to perform a specific task. The quality, quantity, and diversity of training data significantly impact the model's performance and ability to generalize to new, unseen data.

Transformers

A type of neural network architecture that uses self-attention mechanisms to process input data, allowing them to capture long-range dependencies and context more effectively than previous architectures. Self-attention enables the model to weigh the importance of different words in a sequence by computing relationships between them, ensuring that each word's representation considers its relevance to others. Transformers have revolutionized natural language processing and various other AI domains.

Example - The transformer architecture is used in many AI models, such as GPT-3, to process and generate text by understanding the context of words in a sentence.

Unsupervised Learning

A machine learning approach where the model is trained on unlabeled data, aiming to discover hidden patterns or structures within the data on its own. Common techniques include clustering and dimensionality reduction.

Vectors

Mathematical representations of data points in a multi-dimensional space. In AI, vectors are used to represent words, sentences, images, or other types of data, enabling various operations like similarity calculations and clustering.

Vector Database (VectorDB)

A specialized database system designed to store, manage, and query large collections of high-dimensional vectors efficiently. Vector databases are optimized for operations like similarity search, crucial for many AI applications.

Examples - PineCone, FAISS, ChromaDB, etc.

Zero-Shot Prompting

A technique where a model is asked to perform a task or answer a question without any specific examples or fine-tuning for that task. Zero-shot learning relies on the model's pre-existing knowledge and its ability to understand and follow natural language instructions.

Closing Thoughts

We hope this glossary helps you to confidently explore the exciting world that is AI. Remember, AI is constantly evolving, so there's always more to learn. If you come across a term that's not included here, don't hesitate to dig deeper—curiosity drives innovation in AI and beyond.

Hope this glossary serves as a valuable guide through the ever-evolving world of AI!