Gartner®

Glossary of Terms for Generative AI and Large Language Models

Published 5 July 2023 - ID G00791405 - 5 min read

By Analyst(s): Marko Sillanpaa, Leinar Ramos, Arun Chandrasekaran, Anthony Mullen, Jim Hare, Ray Valdes

Initiatives: Artificial Intelligence; Evolve Technology and Process Capabilities to Support D&A

Terms and concepts in the generative AI space can be complex due to the rapid evolution of technologies, methods and applications — especially for large language models. This research aims to educate and empower users to design and engineer solutions powered by GenAI.

Quick Answer

What common terms should I know when researching generative AI (GenAI)?

The common terms have been put into three groups:

- Models and training/learning methods.
- Content and prompts.
- Processing and engineering.

More Detail

Models and Training/Learning Methods

- **Closed model:** A model that no longer accepts inputs or changes to itself.

- **Custom model:** A model built specifically for an organization or an industry.

- **Edge model:** A model that includes data typically outside centralized cloud data centers and closer to local devices or individuals — for example, wearables and Internet of Things (IoT) sensors or actuators.

- **Embedding:** A set of data structures in a large language model (LLM) of a body of content where a high-dimensional vector represents words. This is done so data is more efficiently processed regarding meaning, translation and generation of new content.

- **Few-shot learning:** In contrast to traditional models, which require many training examples, few-shot learning uses only a small number of training examples to generalize and produce worthwhile output.

- **Filters:** Filters are used to remove data or variables from a model to simplify or eliminate options.

- **Fine-tuned model:** A model focused on a specific context or category of information, such as a topic, industry or problem set.

- **Foundational model:** A baseline model used for a solution set, typically pretrained on large amounts of data using self-supervised learning. Applications or other models are used on top of foundational models — or in fine-tuned contextualized versions.

- **Frozen model:** A model that no longer accepts inputs or changes to itself.

- **Generative AI (GenAI):** AI techniques that learn from representations of data and model artifacts to generate new artifacts.

- **Generalized model:** A model that does not specifically focus on use cases or information.

- **Human in the loop:** A process used when the machine or computer system is unable or not allowed to offer an answer to a problem autonomously, thus needing human validation or intervention.

- **Multimodal and modalities:** Language models that are trained on and can understand multiple data types, such as words, images, audio and other formats, resulting in increased effectiveness in a wider range of tasks

- **Multitask prompt tuning (MPT):** An approach that configures a prompt representing a variable — that can be changed — to allow repetitive prompts where only the variable changes.

- **Open model:** A model that — while operational — continues to learn or can contextualize its responses based on inputs and prompts.

- **Parameters:** A set of numerical weights representing neural connections or other aspects in an AI model with values that are determined by training. Large language models (LLMs) can have billions of parameters.

- **AI adoption policy:** An organization's announced goals on how it will adopt AI into its data processing strategies.

- **Pretrained model:** A model trained to accomplish a task — typically one that is relevant to multiple organizations or contexts. Also, a pretrained model can be used as a starting point to create a fine-tuned contextualized version of a model, thus applying transfer learning.

- **Reinforcement learning:** A machine learning (ML) training method that rewards desired behaviors or punishes undesired ones.

- **Reinforcement learning with human feedback (RLHF):** A ML algorithm that learns how to perform a task by receiving feedback from a human.

- **Self-supervised learning:** An approach to ML in which labeled data is created from the data itself. It does not rely on historical outcome data or external human supervisors that provide labels or feedback.

- **Supervised learning:** An ML algorithm in which the computer is trained using labeled data or ML models trained through examples to guide learning.

- **Tokens:** A unit of content corresponding to a subset of a word. Tokens are processed internally by LLMs and can also be used as metrics for usage and billing.

- **Transformer model:** A deep learning model that adopts the self-attention mechanism, differentially weighting the significance of each part of the input data.

- **Transfer learning:** A technique in which a pretrained model is used as a starting point for a new ML task.

Content and Prompts

- **Completions:** The output from a generative prompt.
- **Content:** Individual containers of information — that is, documents — that can be combined to form training data or generated by GenAI.
- **Corpora:** The information or training data used to train an AI. An LLM, like GPT, uses any internet content for its corpora.
- **Specialized corpora:** A focused collection of information or training data used to train an AI. Specialized corpora focuses on an industry — for example, banking or health — or on a specific business or use case, such as legal documents.
- **Grounding:** The ability of generative applications to map the factual information contained in a generative output or completion. It links generative applications to available factual sources — for example, documents or knowledge bases — as a direct citation, or it searches for new links.
- **Metacontext and metaprompt:** Foundational instructions on how to train the way in which the model should behave.
- **Prompt:** A phrase or individual keywords used as input for GenAI.
- **Temperature:** A parameter that controls the degree of randomness or unpredictability of the LLM output. A higher value means greater deviation from the input; a lower value means the output is more deterministic.
- **Training data:** The collection of data used to train an AI model.

Processing and Engineering

- **Fine-tuning:** Improving an existing, pretrained model through additional training with new, context- or task-specific data.

- **Knowledge graphs:** Machine-readable data structures representing knowledge of the physical and digital worlds and their relationships. Knowledge graphs adhere to the graph model — a network of nodes and links.

- **Pretraining:** The first step in training a foundation model, usually done as an unsupervised learning phase. Once foundation models are pretrained, they have a general capability. However, foundation models need to be improved through fine-tuning to gain greater accuracy.

- **Prompt chaining:** An approach that uses multiple prompts to refine a request made by a model.

- **Prompt engineering:** The craft of designing and optimizing user requests to an LLM or LLM-based chatbot to get the most effective result, often achieved through significant experimentation.

- **Plugins:** A software component or module that extends the functionality of an LLM system into a wide range of areas, including travel reservations, e-commerce, web browsing and mathematical calculations.

- **Tunable:** An AI model that can be easily configured for specific requirements. For example, by industry such as healthcare, oil and gas, departmental accounting or human resources.

- **Vector databases:** A type of database used in LLMs to store embeddings, which are representations of words as high-dimensional vectors that can efficiently search and retrieve related concepts.

- **Windowing:** A method that uses a portion of a document as metacontext or metacontent.