Large Language Models

Guide: ShaneHadden
Generated: 2026-11-02 22:19

What is a Large Language Model?

A Large Language Model (LLM) is an AI system that understands and generates human-like text. It is trained on extensive text data to learn language patterns and context. LLMs can answer questions, summarize texts, and translate languages. They use deep learning techniques, particularly neural networks, to process language. Their effectiveness relies on the quality of the training data.

What type of Machine Learning do LLMs use?

Large Language Models (LLMs) primarily use supervised and unsupervised learning. In supervised learning, they are trained on labeled datasets to predict the next word in a sentence. Unsupervised learning helps them learn patterns from large text data without labels. LLMs also utilize transfer learning, where they are pre-trained on extensive datasets and fine-tuned for specific tasks. This combination enables them to generate coherent and contextually relevant text.

What does the GPT of Chat GPT stand for?

GPT stands for "Generative Pre-trained Transformer." It refers to a type of artificial intelligence model that generates text based on patterns learned from large amounts of data. "Generative" means it can create new content, "Pre-trained" indicates it has been trained on diverse text before fine-tuning, and "Transformer" is the architecture used for processing the data efficiently.

What is a Transformer?

A Transformer is a neural network architecture used in natural language processing. It uses attention mechanisms to weigh the importance of words in a sentence. Unlike previous models, Transformers process all words simultaneously, allowing for faster training and better context understanding. They consist of an encoder, which processes input data, and a decoder, which generates output. Transformers are foundational for many large language models.

What is a Parameter in the context of an LLM?

In the context of a Large Language Model (LLM), a parameter is a value that the model uses to make predictions or generate text. Parameters are adjusted during training to minimize errors in the model's outputs. They determine how the model processes input data and influences its understanding of language. The more parameters a model has, the more complex patterns it can learn. Parameters are typically weights in neural networks that connect different layers of the model.

How many parameters are used in the largest LLMs?

The largest large language models (LLMs) have billions to trillions of parameters. For example, GPT-3 has 175 billion parameters, while models like GPT-4 and others are estimated to have over 1 trillion parameters. These parameters help the models understand and generate human-like text.

What is "weighted" by the parameters?

In large language models, "weighted" refers to how parameters (like weights in neural networks) influence the model's predictions. Each parameter adjusts the importance of input features during processing. When the model learns, it updates these weights based on training data to minimize errors in predictions. This weighting determines how much each input contributes to the final output, affecting the model's performance and accuracy in generating text or understanding language.

So the features are weighted? Are there as many features as parameters?

In large language models, features are not directly weighted; instead, parameters are adjusted during training to optimize performance. Features represent the input data characteristics, while parameters are the model's internal weights. There are typically more parameters than distinct features, as models learn complex relationships through numerous adjustments. This allows them to capture nuanced patterns in language. Thus, while features inform the model, parameters drive its predictions.

How many features would a large LLM have?

The number of features in a large language model (LLM) can vary widely depending on its architecture and purpose. Generally, features include parameters like vocabulary size, embedding dimensions, layers, and attention heads. For example, models like GPT-3 have 175 billion parameters, which can be considered as features influencing their performance. However, the exact number of features isn't fixed and can differ based on design choices made by researchers.

Do we know what the features represent?

In large language models, features represent patterns in data that help the model understand language. These features can include word meanings, grammar structures, and contextual relationships. However, the exact representation of features is often complex and not fully interpretable. Researchers use techniques like visualization and analysis to gain insights, but the models operate as black boxes. Understanding features is an ongoing area of study in AI and machine learning.

What information are LLMs trained on?

Large Language Models (LLMs) are trained on diverse datasets that include text from books, websites, and articles. This data helps the model learn language patterns, grammar, and context. LLMs do not have access to real-time information or personal data unless provided during interactions. Their knowledge is based on the data available up to a specific cutoff date.

How much information are they trained on?

Large Language Models (LLMs) are trained on extensive datasets that include text from books, websites, and articles. The information can range from hundreds of gigabytes to terabytes. This training enables them to understand language patterns and context. However, they do not know facts beyond their training cut-off date and cannot access real-time information.

Do LLM's generate the most likely next token?

Yes, large language models (LLMs) generate the most likely next token based on the context of the input. They use probabilities derived from training data to predict which word or token should come next. This process involves analyzing patterns and relationships in language. However, LLMs can also introduce randomness to create diverse outputs. Ultimately, they aim to produce coherent and contextually relevant text.

What is Temperature?

Temperature in large language models controls randomness in text generation. A low temperature (e.g., 0.2) makes the output more deterministic and focused, often repeating common phrases. A high temperature (e.g., 1.0) increases creativity and variability, leading to more diverse but potentially less coherent responses. Adjusting temperature helps balance between predictable and novel outputs, allowing users to tailor responses based on their needs.

is there a standard Temperature that is thought to be the best for most tasks?

There isn't a universal standard temperature for large language models, as it depends on the task. A

temperature of 0.7 is often used for balanced creativity and coherence. Lower temperatures (0.2-0.5) produce more deterministic and focused outputs, while higher temperatures (0.8-1.0) generate more diverse and creative responses. Experimenting with different temperatures can help find the best fit for specific applications.

Are LLMs good at coding?

Yes, large language models (LLMs) can be good at coding. They can generate code snippets, suggest solutions, and help debug errors. However, their effectiveness depends on the complexity of the task and the clarity of the prompt. LLMs may not always produce optimal or secure code, so it's important to review and test their outputs. They are best used as tools to assist programmers rather than replace them.

What is a Context Window?

A context window in large language models refers to the amount of text the model can consider at one time when generating responses. It determines how many previous words or sentences influence the current output. For example, if a model has a context window of 2048 tokens, it can use the last 2048 tokens of text to understand context. This limitation affects how well the model can maintain coherence and relevance in longer conversations or texts.

What is Retrieval Augmented Generation (RAG)

Retrieval Augmented Generation (RAG) combines information retrieval and text generation in natural language processing. It uses a retriever to find relevant data from a database and a generator to create responses based on that data. This method improves the quality and relevance of generated text by grounding it in real information. RAG models can provide more accurate answers compared to traditional language models that rely only on pre-existing knowledge.

What is an Hallucination?

A hallucination in the context of large language models refers to when the model generates information that is false, misleading, or not based on real data. This can occur when the model creates details, facts, or statements that sound plausible but are actually incorrect. Hallucinations can happen due to limitations in the training data or the model's understanding. It's important to verify information generated by these models against reliable sources.

Do LLMs require a lot of power to use?

Yes, large language models (LLMs) require significant computational power to operate. They need powerful hardware, like GPUs or TPUs, to process data quickly and efficiently. Training LLMs demands even more resources, often using large data centers. This high power consumption can lead to increased costs and environmental impact. However, once trained, smaller versions of LLMs can be used with less power for specific tasks.