Glossary of LLM Terminology

LLM (Large Language Model)

An advanced artificial intelligence system trained on large volumes of text data. These models, such as GPT or PaLM, can perform tasks like answering questions, summarizing documents, writing code, and even reasoning.

Token

A token is a unit of text used by language models, typically a word or sub-word. For example, 'learning' might be one token, while 'unbelievable' could be split into multiple tokens.

Token Limit

This refers to the maximum number of tokens a model can accept in a single input. Exceeding this limit means earlier parts of the conversation or input may be cut off.

Context Window

The number of tokens the model can remember at once. A larger context window allows for better understanding of long passages or conversations.

Autoregressive

Describes models that generate one token at a time, each based on the tokens before it. This makes them effective for tasks like writing or story generation.

Transformer

The architecture powering most modern LLMs. It uses attention mechanisms to capture relationships between tokens, enabling better understanding and generation of language.

Pre-training

The initial phase where a model learns language patterns from a massive and diverse dataset without task-specific labels.

Fine-tuning

A targeted training phase where a pre-trained model is further trained on specific data to specialize it for a

certain application.

Embedding

A numerical representation of a word or phrase that captures its meaning and context. These vectors help models compare and relate different terms.

Prompt

The input or question given to an LLM to generate a response. The way a prompt is phrased greatly affects the output.

Prompt Engineering

The process of crafting prompts to get desired and consistent results from language models.

Zero-shot Prompting

Asking a model to perform a task without any prior examples in the prompt.

Few-shot Prompting

Providing the model with a few examples of input-output pairs in the prompt to guide it.

Chain-of-Thought

A technique where the prompt encourages the model to think step-by-step before answering.

Instruct-tuning

Training a model to follow human instructions more effectively, improving its alignment and usability.

Reinforcement Learning from Human Feedback (RLHF)

Fine-tuning a model using human preferences as feedback to make its outputs more useful and aligned.

Temperature

A parameter that controls how creative or random the model's output is. Lower values make responses more focused and deterministic.

Top-p (Nucleus Sampling)

Glossary of LLM Terminology

A technique for sampling the next token based on a subset of the most probable tokens that collectively make up a set probability mass.

API

An interface that allows developers to integrate LLMs into their applications, such as through platforms like OpenAI or Anthropic.

Model Parameters

The adjustable weights in a model that determine its behavior. Modern LLMs often have billions or even trillions of parameters.

Multimodal

A capability that allows a model to handle more than one type of input, such as text, images, or audio.

Hallucination

When a model confidently generates information that sounds plausible but is factually incorrect.

Latency

The time delay between sending a request to a model and receiving its response.

Grounding

Ensuring that a model's responses are backed by factual or externally verifiable sources.

Safety Alignment

The process of making sure LLMs avoid generating harmful, biased, or unsafe content.

Agent

A system that uses LLMs to perform tasks over time, often autonomously, by making decisions based on input and memory.

Retrieval-Augmented Generation (RAG)

A method where the LLM fetches relevant documents before generating a response to increase factuality.

Glossary of LLM Terminology

Embedding Space

The vector space where all text inputs are represented as numerical vectors, enabling semantic search and clustering.

Context Overflow

Occurs when inputs exceed the model's context window, causing earlier tokens to be dropped from memory.

Training Data

The raw text or information used to train a language model. This typically includes books, websites, code, and more.

Checkpoint

A saved snapshot of a model at a specific stage in training that can be reused or fine-tuned.

Bias

Unintended behaviors or patterns in model output caused by imbalances or issues in the training data.

Prompt Injection

A vulnerability where malicious instructions are hidden in user prompts to override the model's intended behavior.

Guardrails

Mechanisms that monitor and restrict the model's outputs to prevent undesirable or unsafe behavior.

Open-Source LLM

Language models that are freely available for use, modification, and deployment by the public.

Closed-Source LLM

Proprietary language models like GPT-4 or Claude 2, whose weights and internal workings are not publicly shared.

Synthetic Data

Glossary of LLM Terminology

Artificially generated data used to train or fine-tune models when real-world examples are limited.

Self-supervised Learning

A method where models learn from raw data without labeled outputs by predicting parts of the input.

Latency Optimization

Efforts to reduce model response time, important for real-time applications.

Natural Language Understanding (NLU)

The ability of a model to interpret, extract meaning from, and understand text.

Natural Language Generation (NLG)

The ability of a model to generate coherent, human-like text based on a prompt.

Conversational AI

A system designed to engage in interactive dialogue, often used in chatbots and virtual assistants.

Ground Truth

Verified, accurate information used to evaluate or correct a model's responses.

Knowledge Cutoff

The latest date up to which the model has been trained on data. It cannot know or reference anything beyond this.

LLM-as-a-Service

Cloud-based access to language models via APIs, used in applications without hosting the models locally.

Parameter Tuning

Adjusting specific model configurations to optimize its performance.

Model Compression

Techniques used to reduce model size and make it more efficient for edge devices.

Glossary of LLM Terminology

Inference

The process of using a trained model to make predictions or generate outputs from new inputs.

Overfitting

When a model performs well on training data but poorly on unseen data due to lack of generalization.