**Supervised learning**

Supervised learning, also known as labeling things, is a fundamental machine learning technique that involves training an AI model on labeled data. In supervised learning, the AI model is presented with a dataset of input data and corresponding output labels, which represent the desired results for each data point. The model's goal is to learn the mapping between the input data and the output labels, enabling it to make accurate predictions on new, unlabeled data.

**How LLMs work**

Large language models (LLMs) are trained using a technique called supervised learning, where they are exposed to a massive dataset of text and code and tasked with predicting the next word in a sequence. This process is repeated over millions or even billions of examples, allowing the LLM to learn the statistical relationships between words and phrases

**Tokenization**

The text data is broken down into individual tokens, which could be words, characters, or subwords. This allows the LLM to process the data in a structured and consistent manner.

Note: (Assume 1 token = 3/4 words, or equivalently, 1 word \approx. 1.333 tokens)

**Embedding**

Each token is represented as a numerical vector, known as an embedding. These embeddings capture the semantic and contextual relationships between words, allowing the LLM to understand the meaning of language.

**Gen AI Applications**

**Writing**

Article writing, copywriting, brainstorming, recipes, or anything in general.

**Reading**

Summarization, Proofreading, Translation, and much more

**Chatting**

General Chatbots, specialized chatbots, and much more

**Customer Support Workflows**

• Humans only

• Bots support humans (human in the loop)

• Bot triages for humans

• Chatbots only

**Deploy Chatbot**

Start with an internal facing chatbot.

Deploy with human in the loop to check for mistakes.

Only after deemed safe, allow bot to communicate directly with customers.

**LLM’s Limitation**

An LLM’s knowledge of the world is frozen at the time of its training.

LLMs sometimes hallucinate and make up things.

The input and output length is limited.

An LLM’s context length is the limit on the total input + output size.

Generative AI does not work well with structured (tabular) data.

An LLM can reflect the biases that exist in the text it learned from.

Some LLMs can output toxic or other harmful speech.

**Tips for prompting**

• Be detailed and specific.

• Guide the model to think through its answer.

• Experiment and iterate.

**Image generation**

Image generation models often rely on diffusion models, a type of machine learning technique that learns the underlying patterns and characteristics of images. Diffusion models can be considered a form of supervised learning, as they are trained on a large dataset of labeled images. During training, the model learns to progressively remove noise from a corrupted image until it reconstructs the original, clean image. This process allows the model to learn the intricate details and relationships between pixels, enabling it to generate new images that are similar to those in the training set.

**Lifecycle of Gen AI project**

1. Scope Project

2. Build/improve system

3. Internal evaluation

4. Deploy and monitor

**Improve Gen AI Model**

Building Generative AI is a highly empirical (experimental) process – we repeatedly find and fix mistakes.

• Prompting

• Retrieval augmented generation (RAG)

Give LLM access to external data sources

• Fine tune models

• Pretrain models (trains LLM from scratch)

**RAG**

Retrieval-augmented generation (RAG) is an innovative technique that enhances the capabilities of large language models (LLMs) by incorporating external knowledge sources. Unlike traditional LLMs that rely solely on their internal knowledge base, RAG models leverage external retrieval systems to access relevant information and incorporate it into their generation process.

The implementation of RAG typically involves two main components:

• Retrieval System: A retrieval system identifies and retrieves relevant information from external sources, such as documents, databases, or knowledge bases.

• Generation Model: A generation model, typically an LLM, utilizes the retrieved information to generate a response. This may involve techniques such as incorporating retrieved passages into the prompt or using retrieved information to inform the generation process.

**Examples of RAG Applications**

Chat with PDF files

Answer questions based on a website’s articles

**Fine-tuning**

Fine-tuning is the process of adjusting an LLM's parameters to improve its performance on a specific task.

**How LLM follow instructions**

Large language models (LLMs) can be fine-tuned with additional data to enhance their ability to follow prompt instructions. Instead of simply generating the next word in a sequence, LLMs trained with such data can comprehend and execute instructions effectively. This refinement enables LLMs to perform more complex tasks that require understanding and responding to specific directives.

**RHLF**

Reinforcement learning from human feedback (RLHF) is a technique that fine-tunes large language models (LLMs) by incorporating human input. In this approach, humans provide scores or rewards to different responses generated by the LLM, guiding the model towards producing more desirable outputs. This iterative process helps the LLM learn the nuances of human preferences and improve its overall performance.

**Tools**

Tools can be used by large language models (LLMs) to enhance their functionality and versatility.

**Agents**

Agents play a crucial role in enabling large language models (LLMs) to effectively interact with the environment and perform tasks. By employing agents, LLMs can transition from passive language understanding to active decision-making and execution.

Agents act as intermediaries between LLMs and the external world, deciphering instructions received from the LLM and translating them into concrete actions. This allows LLMs to operate in a dynamic and interactive manner, adapting their behavior based on real-time feedback and observations.















