What do the buzzwords mean?

When people talk about AI, chatbots, and tools like ChatGPT, all the new terminology can make it hard to understand how it all works.

In this section, we're going to look at the definitions of some of the words that you might hear. We'll cover some more terms later in the course.

What is AI?

According to Google, Artificial intelligence (or AI) is a discipline within computer science that focuses on creating intelligent agents. These agents are computer systems that can reason, learn, and act autonomously. AI is the theory and methods for building machines that can think and act like humans.

Examples of AI include speech recognition and translation between different languages, such as English and French.

Generative AI

Generative AI is a type of artificial intelligence system that uses artificial neural networks to generate new content. We'll explain artificial neural networks shortly.

Generative AI can produce various types of content, such as text, images, audio, and synthetic data in response to a prompt or a series of prompts.

In the past, the Large Language Models used by AI systems tended to be what are called discriminative models. They would classify or predict labels for points of data.

Generative AI systems and models are different because they can generate new data. Generative language models learn patterns from training data and they can, for example, generate natural-sounding text.

Prompts

A prompt is a short piece of text submitted to a Large Language Model to receive a response back.

Prompts are normally written in the way you would ask a person – in your natural language.

For example:

```
Give me a list of things that I should bring with me to a camping
trip.
```

After the model receives a prompt, depending on the type of model being used, it can generate text, code, images, videos, music, and more.

Large Language Models

Large Language Models (LLMs) are a powerful type of AI that enable computers to comprehend and generate natural language. They can be used to build all sorts of applications, such as chatbots.

There are LLMs that are trained on vast amounts of text data to understand and generate human-like language.

Training a LLM involves exposing it to enormous sets of data, such as books, articles and websites, to learn the statistical patterns and semantic representations of language.

The training process involves predicting the next word in a sentence or filling in missing words based on the context of the given text.

By refining its predictions iteratively, the model becomes better and better at understanding and generating coherent human-like responses.

According to Benedict Evans:

> *Large Language Models do not tell you the answer to your question. They tell you "when people ask questions like that, this is what the answers that other people tend to give tend to look like" Over time, that difference may or may not narrow, depending on what kind of question you're asking.*

So an LLM is a machine that uses statistical predictions to work out its responses.

Foundation models are large AI models pre-trained on vast amounts of data. They can be fine-tuned for specific tasks, and they have the potential to revolutionise various industries.

Examples of LLMs are: OpenAi's GPT-3 and GPT-4, and Google's PaLM. (December 2023 update: Google revealed it is launching a new LLM called Gemini. It will directly integrate the AI into Google apps. It is being released in the form of upgrade to Google's chatbot Bard, but not yet in the UK or EU.)

You can get LLMs to produce useful behaviours by writing the right input text. As we mentioned, this is called a prompt.

What is the difference between a chatbot and a Large Language Model?

A chatbot and a Large Language Model serve different purposes and have distinct characteristics.

A chatbot is a task-oriented conversational agent with predefined rules, patterns, and specific functions. A Large Language Model is a more general-purpose language generation model capable of understanding and generating text across a wide range of topics and prompts.

The training of them also differs:

- Chatbots are trained using specific rules and patterns defined by their developers. They rely on predefined sets of responses or scripted conversations to handle user queries.
- In contrast, Large Language Models like GPT-3 are trained on vast amounts of diverse text data. As we mentioned earlier, they learn to generate text by predicting the next word in a sentence or filling in missing parts of a given text. The training process enables them to capture a broad understanding of language and generate coherent and contextually appropriate responses.

Chatbots and LLMs have different levels of flexibility:

- Chatbots are typically built with a specific purpose in mind and are designed to handle a predefined set of tasks. They are often limited to specific domains or contexts, and their responses are based on pre-programmed rules or patterns.
- Large Language Models are more flexible and adaptable. They can generate text on a wide range of topics and handle diverse prompts. They don't require explicit programming for each task or scenario and can generate responses based on their understanding of language and context.

And the context and conversational flow between them differs:

- Chatbots are often built with conversational flow in mind. They aim to guide users through a specific interaction or transaction. They might store user context and use it to provide more personalised responses.
- Large Language Models like GPT-3 can also maintain context but usually focus on generating coherent responses based on individual prompts rather than managing a multi-turn conversation.

Hallucinations

Hallucinations are words or phrases that are generated by the model that are factually or grammatically incorrect.

Hallucinations can be caused by a number of reasons. These can be:

- The model hasn’t been trained on enough data
- The model has been trained on noisy or dirty data

- The model has not been given enough context, or
- The model hasn't been given enough constraints.

Hallucinations can be a problem for transformers because they can cause the model to generate incorrect or misleading information.

ChatGPT

Is ChatGPT a chatbot or an LLM? This is where it gets a bit messy.

It can be considered to be both a chatbot and a Large Language Model, because it combines elements of both.

ChatGPT is designed specifically for interactive conversation, making it chatbot-like in nature.

It is intended to simulate human-like conversation and engage in back-and-forth exchanges with users. It can take up to 12,000 words of context when you are asking it to do things.

ChatGPT has caused quite a storm because its performance at understanding and writing text and code was so beyond anything that had existed before. And in a very short space of time, it has proven to be very popular.

Behind the scenes, the free version of ChatGPT currently uses a Large Language Model called GPT 3.5 Turbo. This was made available via an API in March 2023.

There is also a paid option called ChatGPT Plus which gives you priority access to the website and to the prompt.

GPT-4

GPT-4 and GPT-4 Turbo are the latest LLM model released by OpenAI. GPT-4 Turbo was released in November 2023.

GPT-4 models give more synthetic and factual responses compared to GPT- 3. It has better reasoning capabilities for solving complex problems. It can take more words of context. However, it is more expensive, compared to GPT 3.5 Turbo.

The OpenAI Playground

OpenAI has another app for using ChatGPT, called the OpenAI Playground.

It offers more advanced features for users who want to customise and experiment with the GPT models, to see how they affect the model's behaviour.

Users can choose from multiple GPT models, adjust the model's size, select different output formats, and more.

ChatGPT APIs

The ChatGPT APIs provide a more flexible and customisable way to interact with the model, when compared to the fixed prompt-based approach of the ChatGPT website and the OpenAI Playground.

The ChatGPT APIs are a set of Application Programming Interfaces (APIs) that allow developers to use the capabilities of the GPT model in their applications.

By integrating these APIs into their apps, developers can enable users to interact with the AI model conversationally. It means that you can automate tasks, generate content, and so on.

The API works by sending a series of messages as input and receiving a model-generated message as the output. The messages typically consist of a conversation history, including user inputs and the model responses. The conversation context helps the model maintain coherence and generate relevant responses.

The ChatGPT API is not free.

To get an API key, visit https://platform.openai.com/

More definitions

The definitions that follow are about how LLMs are developed. You might not need to know this level of detail. You can stop here if you wish.

Machine learning

The rise of AI has largely been driven by a technology called machine learning.

Machine learning is a subfield of AI that involves training models from input data. The trained model can make useful predictions from new or never seen before data drawn from the same one used to train the model. For example, it can predict the next word that should follow these words “The cat sat on the …”

Machine learning allows computers to learn without explicit programming. To an extent, it trains itself.

The common types of machine learning models are supervised, unsupervised, and semi-supervised learning.

Supervised learning

Supervised models learn from labelled data, where each data point, such as a word, is associated with a label or tag. The tag might be a name, a category, or a number.

This is most common type of machine learning.

In supervised learning, the model learns from examples to predict future values.

Test data values are inputted into the model. The model outputs a prediction and compares that prediction to the training data used to train the model.

If the predicted test data values and actual training data values are far apart, that's called an error. And the model tries to reduce this error until the predicted and actual values are closer together.

So the computer learns how to map the way from an input to an output, or a result. This is known as A to B or input to output mappings.

Models like GPT are trained to take a series of messages as an input and return a model-generated message as an output.

If you wanted to carry out supervised learning for a chat interface that used your own LLM, you can decide what is A, the input, and what is B, the output, and how to make it useful for your organisation.

For example, if it were for a customer support application:

- The input (A) could be examples of emails from customers
- The output (B) could be labelling each of these customer emails as to whether they are a refund request, an enquiry about a delivery, or something else.
  So, the output could, in this case, be one of three outcomes.

Unsupervised learning

Unsupervised models learn from unlabelled data and aim to discover patterns or groupings within the data.

Unsupervised models are about discovery, and seeing if the raw data naturally falls into groups.

Deep learning and neural networks

Deep learning is a type of machine learning that allows LLMs to process more complex patterns than other types of machine learning.

Generative AI is a subset of deep learning.

Deep learning uses a powerful set of tools, called neural networks. A neural network is a mathematical equation, or software, that tells it, given the inputs, how you compute the outputs.

Semi-supervised learning

Neural networks can use both labelled and unlabelled data, called semi-supervised learning.

Semi-supervised learning is a combination of supervised and unsupervised learning, where a model is trained on a small amount of labelled data and a large amount of unlabelled data.

According to Google, the labelled data helps the neural network to learn the basic concepts of the task, while the unlabelled data helps the neural network to generalise to new examples.

We can now refine the definition of Generative AI we used earlier.

Generative AI is a subset of deep learning. This means it uses artificial neural networks. It can process both labelled and unlabelled data using supervised, unsupervised, and semi-supervised methods.

Vector encoding

For an AI system to understand non-numerical content, it must be converted into a numerical format.

This conversion from unstructured data to numerical data is called vectorising or encoding.

Each item, that could be a single word, an image, or an audio file, is encoded into a list of numbers: this list is called a vector.

The encoding process can be done by different kinds of algorithms.

If it is done through a deep neural network, it's often called deep learning vector embedding.

When the content is encoded into a vector, the longer it is in length, the more information is represented about the data it encoded.

The concept of attention

A major breakthrough in machine learning happened in 2017 as a result of a paper from Google called "The Transformer". This paper introduced the concept that attention is crucial for AI systems.

In the context of AI systems, "attention" refers to the ability of the system to focus on specific parts of information that are relevant or important for a given task. It is like our human ability to selectively concentrate on specific details while ignoring others.

To understand attention, think of a scenario where you're reading a long paragraph. Your attention naturally gravitates towards certain words or phrases that seem more meaningful or relevant to the overall context. These words or phrases capture your attention, allowing you to better understand and process the information.

In AI systems, attention works similarly. It helps the system allocate its computational resources to focus on the most relevant parts of the input data. By doing so, the system can better understand the relationships between different elements, recognise patterns, and make more accurate predictions or generate more meaningful outputs.

By incorporating attention mechanisms into AI models, such as the Transformer architecture introduced by Google in 2017, the system becomes more efficient and effective at understanding and processing complex information.

At a high level, a transformer model consists of an encoder and decoder.

The encoder encodes the input sequence and passes it to the decoder, which learns how to decode the representation for a relevant task.

The attention mechanism enables the system to assign different weights or importance to different parts of the input, allowing it to prioritise and attend to the most relevant aspects. And this ultimately enhances its performance.

Predicting and embedding

The other two key developments that have contributed to AI's progress have been predicting the next word, or token, and embedding.

Embedding involves creating complex maps of related concepts.

These developments, when combined with large-scale computational models containing billions of parameters, enable AI systems to learn quickly and generate impressive results.

Small Language Models

Small Language Models (SLMs) are an alternative to Large Language Models.

Large Language Models are large and complex, which means they are time consuming and costly to build. GPT-4 is said to have some 100 trillion parameters, and ChatGPT-3.5 has around 175 billion parameters.

According to an article at https://www.deeplearning.ai/the-batch/how-small-language-models-can-perform-specialized-tasks/ new research shows smaller models can perform specialised tasks relatively well after fine-tuning on only a handful of examples.

Some companies don't want to trust their data with OpenAI.

Instead, they are building their own private language models, using personal computers with powerful processing chips. The models are smaller than the popular LLMs, but they argue you can create a language model for software documentation for example without needing so many parameters.

Retrieval Augmented Generation (RAG)

Note: This isn't mentioned in the buzzwords videos, as it's a new and advanced technique. We cover it later in the course.

Retrieval Augmented Generation is a way of fetching data from an external database and making it available to an LLM when you ask it to generate a response. You can store proprietary business data or real-time information about the world and have your application fetch it for the LLM at generation time.

In simple terms, you add source information to the prompt (such as some content originally taken from your knowledge base) and tell the LLM to find the information from that source.

This method reduces the likelihood of hallucinations and improves the quality of responses about specialist information.