ChatGPT and the End of a World

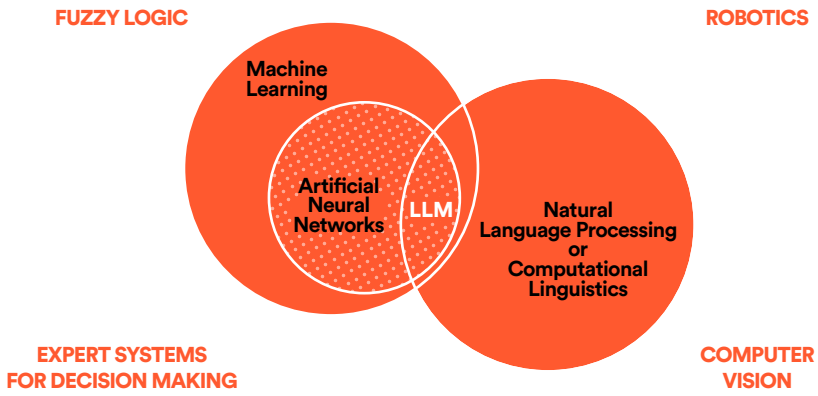
EPISODE 1

WHAT ARE LLMS?

Raquel Fernández
Full Professor of Computational Linguistics
Institute for Logic, Language and Computation
(ILLC), University of Amsterdam

Tools like ChatGPT have recently become very popular. ChatGPT is an example of a large language model or LLM. This type of technology has existed for some years in the field of computational linguistics, which is a subfield of artificial intelligence.

Artificial Intelligence



The newest large language models are capable of automatically generating texts that look a lot like those produced by humans.

HOW DO THEY DO IT?

An LLM is

A language model is a computer program designed by humans (AI researchers, computational linguistics researchers, etc.).

It is based on a mathematical model called an ***artificial neural network***.

This is a program that uses **machine learning** techniques: an LLM learns to create texts and trains in generating them using data (lots of data!).

The type of artificial neural network used by current LLMs is called a *transformer*.

GPT STANDS FOR
GENERATIVE PRE-TRAINED TRANSFORMER

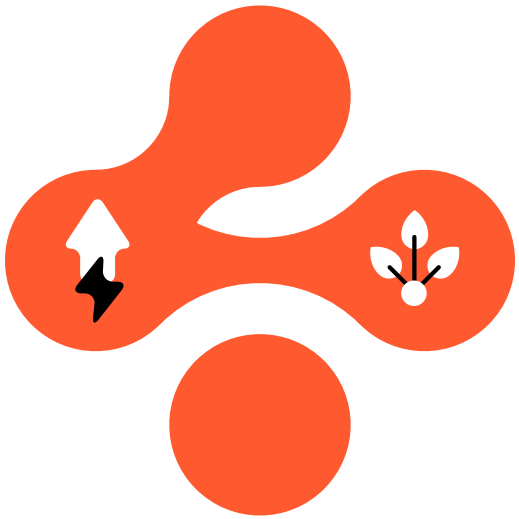
Why are they called *"large" language models?*

They are called "large" because the artificial neural networks behind LLMs have a complex structure, with billions of **parameters**, that are like levers controlling the system's responses.

In addition, they need large amounts of learning materials to be able to learn to produce good-quality texts.

It takes thousands of computers running in parallel for months to train such a model.

This requires a huge amount of energy, which could have negative consequences for the environment.



What are *learning materials?*

Learning materials, or training data, is a set of texts that must be as large and diverse as possible. Typically, texts downloaded from the internet are used, for example, from Wikipedia, e-books and online newspapers, blogs and social media networks like X (previously Twitter), Instagram and Facebook.

In theory, a language model isn't directly connected to the internet. During the training phase, it only has access to the downloaded texts that comprise the learning materials.

It is worth noting that the use of this type of data can be problematic as there may be issues related to privacy, consent and intellectual property and the data may be biased and may not correctly represent certain sectors of the population.

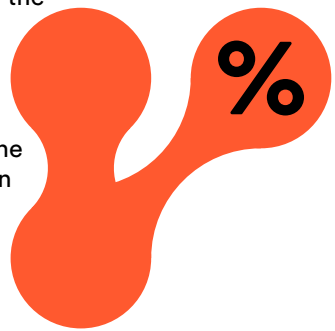
What can an LLM learn to do?

The basic idea is very simple: an LLM learns to **predict the next word** in a sentence or text, given an initial context or *prompt*. G

For example, not all the words in the dictionary have the same likelihood of appearing after a fragment of text like "Today it is...". Words like *sunny, exactly fifteen* or *much* are possibilities, while words like house, dance or is, have a very low probability (almost zero) of being the next word.

Using all the sentences and texts that form part of the **learning materials**, the model learns to assign the correct probability to each possible word and, by extension, to each sentence.

In other words, it learns the patterns that determine which sentences are natural and realistic in a given context or in response to a question.



Predicting one word after another allows texts, answers and entire conversations to be generated. This is why LLMs are considered an example of **generative artificial intelligence**.

Unsupervised Learning

Example of a training text:

I read a really fun comic

Given this *prompt* or context

We want the model to learn to predict this word

I read a...

comic? drawing? today?

Example of prediction of the model when it is still learning (the model predicts a probable distribution)

Correct prediction in this case

Correction of the probability distribution during the learning stage

neural network

		Correct prediction	Correction
9 %		0 %	Decrease
4 %	today	0 %	
5 %	comic	100 %	Increase
		0 %	
3 %	drawing	0 %	Decrease
11 %		0 %	
...		...	...

List of all the possible words (the list is as long as the entire dictionary). **Each bar represents the probability assigned to each word, given the context or *prompt*.**

How is an LLM like ChatGPT trained?

ChatGPT was created by a private company called OpenAI, which has not yet revealed all the details as to how it trained this specific model. However, we do know that the training period involved three stages:

During the first stage, all LLMs are trained to predict the next word given an input context.

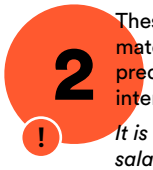
This is considered **unsupervised learning**, because the model learns based on large amounts of existing texts, without any additional human intervention.

LLMs, like ChatGPT, go through two more learning stages.

During the second stage, humans produce examples of conversations between two people that simulate conversations between a person and ChatGPT.

These conversations are used as additional learning materials. The objective is still to teach the model to predict the next word but, for this stage, in a context of interactive dialogue.

It is important to mention that the working conditions and salaries of these workers are highly inadequate.



During the third and final stage, supervised learning is used. This is when it is possible to interact with the model and the answers automatically generated by ChatGPT are assessed by more workers, who label them "good" or "bad" (for example, answers with offensive language or controversial subjects will be labelled "bad").

This information is used as feedback to reinforce responses that are suited to human preferences. This type of learning is called ***reinforcement learning with human feedback*** (RLHF).

Can we trust the texts and answers generated by an LLM?

It is important to bear in mind that language models **do not have any mechanisms that ensure the veracity of the content** produced: the sentences they generate can sound completely plausible and reasonable but there is no guarantee that they are true (sometimes they will be, sometimes they won't).

The fact that the generated texts are grammatically correct and sound natural can lead to the **anthropomorphisation** of chatbots based on language models.

In other words, we often assign human qualities (like intelligence, logical reasoning or emotions) to machines that only appear to have them.

However, these features are not part of the abilities of a language model.

Questions for Reflection

1. From everything you've learned here about LLMs, what has surprised you the most?

2. ChatGPT is an example of an LLM. Do you know of any others? BingAI, Bard...
   Do they ring a bell? What do they do?
   Technology evolves incredibly fast. Whilst working on the publication of this chapter, ChatGPT announced its next internet connection (in theory, an LLM isn't connected to the internet) and Bing announced an extension for generating images.
   Where are we right now as you read this?

3. What role do logic and coherence play in the assessment of answers generated by ChatGPT? What would you do to distinguish between a valid answer and an incorrect one?

4. Some have warned of the dangers of attributing human characteristics to artificial intelligence.
   Do you think it's dangerous?
   Why?

5. The intervention of people in the creation, training and responses of LLMs is vital.
   Can you provide examples?
   Do you think the involvement of humans is sufficiently acknowledged, valued and appreciated?
   Why do think that?
   What implications does this have?

6. If you used an LLM to do your homework, do you think you should mention it? Why?

GLOSSARY

- **ANTHROPOMORPHISATION:** The act of attributing human characteristics or rationale to an animal or thing.
- **PARAMETERS:** Numerical values that define the model's behaviour.
- **PROMPT**: Instruction, question or initial phrase given to the language model to guide its response or text generation.
- **ARTIFICIAL NEURAL NETWORK**: Model based on how the human brain works. It is formed by a series of nodes, known as *artificial neurons* that are connected and transmit signals to each other. These signals are transmitted from input until an output is generated.

REFERENCES TO EXPLORE MORE

a. Bender, Emily M. et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of FAccT 2021*, p. 610-623 (2021, dl.acm.org)
b. Brown, Tom B. et al. "Language Models are Few-Shot Learners" in *Advances in Neural Information Processing Systems* (p. 1877-1901). Curran Associates (2020, neurips.cc)
c. Pointon, Chris. "The carbon footprint of ChatGPT" in *Blogpost sobre el impacto medioambiental de ChatGPT* (2022, medium.com)
d. "Circuito OEIAC sobre el uso responsable y sostenible del ChatGPT" (2023, udg.edu)
e. "Beijing Consensus on Artificial Intelligence and Education" (2019, unesco.org)
f. Sabzalieva, Emma and Valentini, Arianna. "ChatGPT and artificial intelligence in higher education: quick start guide" (2023, unesco.org)

Text: Raquel Fernández / Design: La Puput Gràfica Coop V
"la Caixa" Banking Foundation, 2023