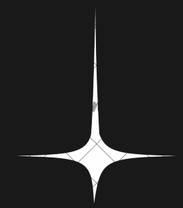
Large Language Models & ChatGPT

Presented by:
Tanya Khanna



WORKSHOPS SCHEDULE

Introduction to Python Programming	February 1, 2024; 4 - 5:30 PM
Mastering Data Analysis: Pandas & Numpy	February 8, 2024; 4 - 5:30 PM
Python for Visualization & Exploration	February 15, 2024; 4 - 5:30 PM
Mathematical Foundations of Data Science	February 29, 2024; 4 - 5:30 PM
Introduction to Machine Learning: Supervised	March 7, 2024; 4 - 5:30 PM
Introduction to Machine Learning: Unsupervised	March 21, 2024; 4 - 5:30 PM
Introduction to Deep Learning	March 28, 2024; 4 - 5:30 PM
Deep Dive into Natural Language Processing	April 4, 2024; 4 - 5:30 PM
Large Language Models and Chat GPT	April 11, 2024; 4 - 5:30 PM

https://libcal.rutgers.edu/calendar/nblworkshops?cid=4537&t=d&d=0000-00-00&cal=4537&inc=0

https://github.com/Tanya-Khanna/DataScienceWorkshop_2024_NBL

Table of Contents

LLMs: The Basics

ChatGPT — a type of conversational AI is built — on top of a "Large Language Model".

LARGE

Big in size, extent, or capacity.

When we use "large" in "Large Language Models," we're referring to the:

- Scale of the model in terms of the amount of data it was trained on and its computational architecture. These models process and "understand" vast amounts of text data.
- The "largeness" also refers to the number of parameters the model has. Parameters are the aspects of the model that are learned from the training data; more parameters mean the model can capture more complex patterns and nuances in language. For example, millions to billions, and even trillions of parameters.

LANGUAGE

The method of human communication, either spoken or written, consisting of the use of words in a structured and conventional way.

"Language" in this term highlights the focus on human languages — how we communicate ideas, emotions, facts, and more through words. Language models are specifically designed to understand and generate human language. They can comprehend grammar, semantics (meaning), and to some extent, the context within the text. This enables them to perform tasks like translation, question-answering, and even writing stories or generating explanations.

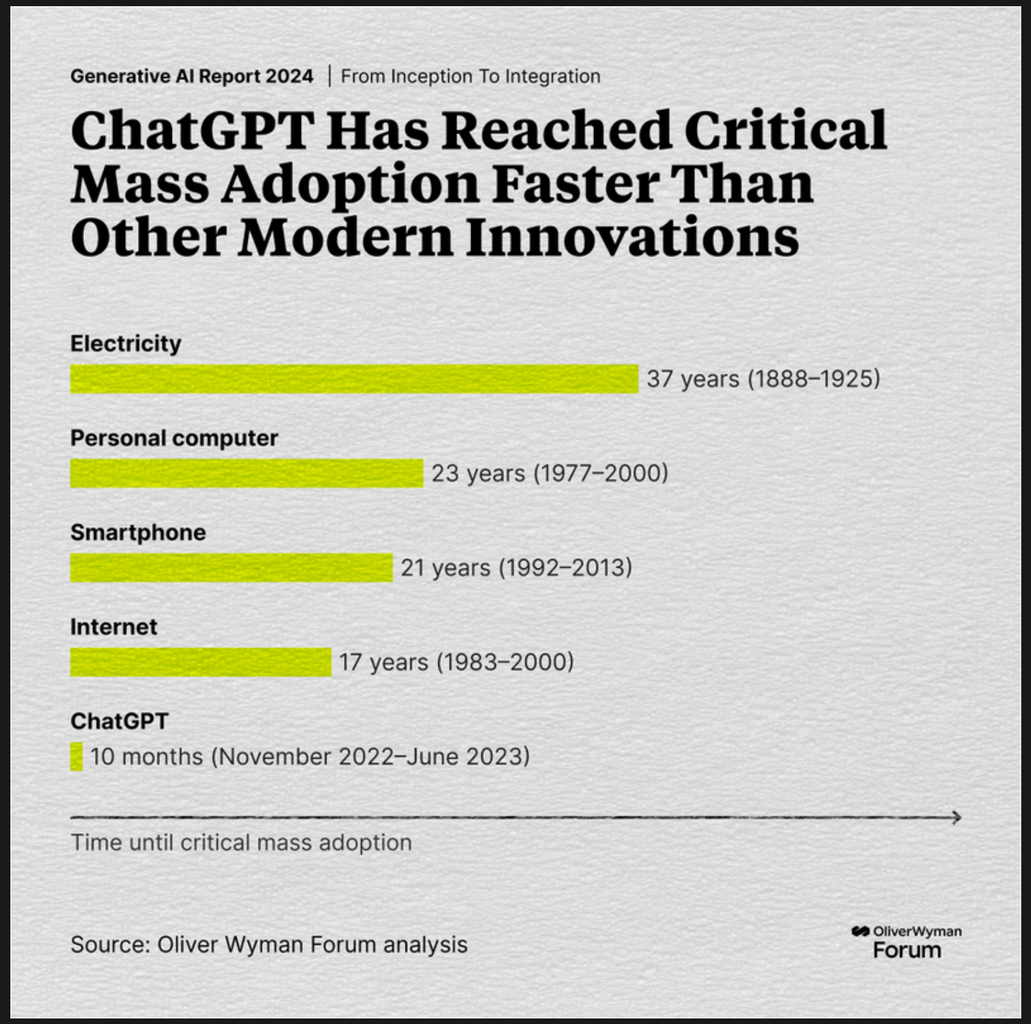
MODEL

A representation of a system made to study some aspects of that system or for the purpose of explaining, predicting outcomes, etc.

In the world of machine learning and artificial intelligence, a "model" is a mathematical structure that is trained to make predictions or decisions based on input data. For language models, this means predicting the next word in a sentence, generating text based on a prompt, or understanding and responding to questions. The model learns patterns, structures, and even the subtleties of language from the data it is trained on.

Putting it all together, Large Language Models are advanced, extensive computational systems trained on enormous datasets. They are designed to understand, interpret, and generate human language in a way that mimics human-like understanding. LLMs like ChatGPT can perform a wide range of language-related tasks, from answering questions accurately to composing essays, poems, or code, based on the patterns they've learned from the data they were trained on.

Unpacking the Hype around LLMs



LLMs have demonstrated remarkable proficiency in understanding and generating human-like text, enabling conversations, writing articles, composing poetry, and more. Their ability to engage in tasks that were traditionally considered the domain of human intelligence has sparked both excitement and debate.

The Good, The Bad, And Everything In Between

Will the advent of generative AI mark the beginning of an unprecedented era of efficiency, or will it result in the widespread loss of jobs worldwide?

Can it pave the way for newfound personal achievements, or might it trap individuals in a state of solitude and disconnection?

Is it poised to elevate human society to greater achievements, or could it potentially contribute to our downfall?

Depending on whom you ask, the answer to all those questions is yes.

- In the 16 months post-ChatGPT launch, the exact impact of generative AI on the world remains unclear, but it's certain to have both beneficial and detrimental effects across all levels of society, reshaping workspaces and personal lives.
- Like many groundbreaking technologies, generative AI comes with its set of challenges. Fire, which brought people together, could also cause destruction; cars enhanced mobility but led to road accidents; the internet made communication easier but also empowered criminals.
- A distinctive feature of generative AI is the caution from its developers about potential negative outcomes, highlighting a mix of optimism and concern within the AI community. This is evident in the ongoing debates around the management and direction of AI organizations, such as OpenAI.
- The emergence of generative AI presents a unique opportunity amid its uncertainty. The popularity of ChatGPT has accelerated the entry of numerous entities into the AI field, sparking a dynamic environment of innovation and competition.
- Decisions made by business leaders, government officials, and regulators will shape the landscape of AI development, influencing whether it will be more open and transparent or closed and exclusive.
- The role of consumers and the workforce in embracing and integrating AI into their daily lives will be crucial in realizing its potential benefits swiftly.

Predictions suggest that by 2030, generative AI could contribute as much as $20 trillion to the global economy and save around 300 billion hours of labor annually.

Applications of LLMs

WHAT THEY ARE

- **Advanced AI Tools:** LLMs are cutting-edge artificial intelligence technologies designed to understand, interpret, and generate human language.
- **Based on Vast Data:** They are trained on extensive collections of text data, enabling them to grasp a wide array of language patterns and nuances.
- **Versatile in Application:** Capable of performing a diverse range of tasks, from writing and summarization to answering questions and generating creative content.
- **Continuously Evolving**: These models are regularly updated and refined to improve their accuracy, responsiveness, and scope of knowledge.

WHAT THEY AREN'T

- **Omniscient Beings:** While LLMs can access and process a vast amount of information, they do not possess consciousness or independent thought.
- **Infallible Oracles:** Their outputs are based on patterns in data they were trained on; they can make mistakes, misunderstand questions, or generate inaccurate or biased information.
- **Substitutes for Human Expertise**: LLMs are tools to augment human capabilities, not replace them. They cannot replicate the depth of human expertise or emotional intelligence.
- **Free from Ethical Concerns:** The development and deployment of LLMs raise important ethical considerations, including privacy, misinformation, and bias. They are not inherently neutral and require careful management.

LLMs are powerful AI tools with the potential to transform many aspects of our lives, offering support, efficiency, and innovation. However, they have limitations and are not a replacement for human judgment, expertise, or ethical consideration. Understanding both the strengths and limitations of LLMs is crucial for their responsible development, deployment, and use.

Typical Challenges of LLMs: Functional & Technical

Rapid Pace of Research

The fast evolution of LLM research requires expertise to stay updated and select optimal models for specific applications.

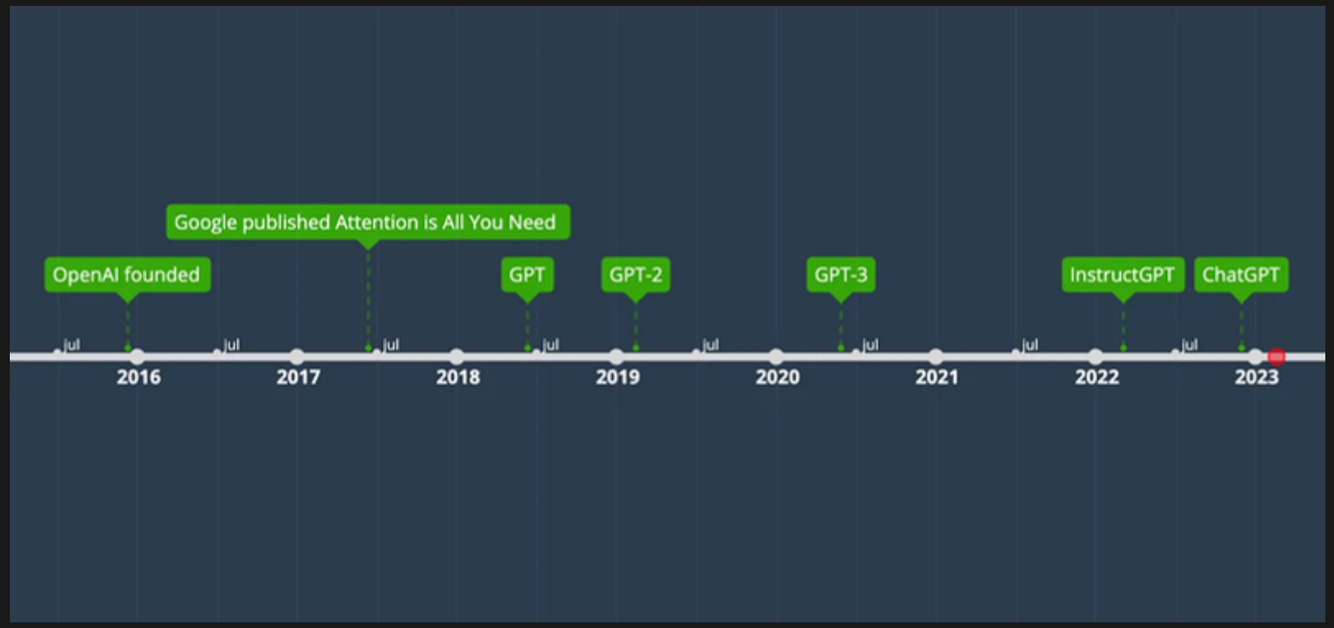
Trustworthy AI

Trained on the vast and varied content of the internet, LLMs may reproduce language from its darker aspects, including biases, stereotypes, and misinformation. They don't discern truth or ethical values, may contradict known facts based on the varied data they're trained on, including prevalent misconceptions. Errors in LLMs can accumulate, leading to "hallucinations" or generating unrelated content, especially with complex inputs or tasks requiring precise data. Thus, there are ethical concerns include biases in training data, potential for spreading misinformation through AI hallucinations, and a gap in current event knowledge, prompting heightened regulatory scrutiny, such as the European AI Act.

Interpretability and Explainability

The "black box" nature of LLMs makes it hard to understand how they generate outputs, challenging their trustworthiness and accountability in critical fields like healthcare and finance. Solutions often need a human in the loop for oversight and to ensure robust performance in variable conditions.

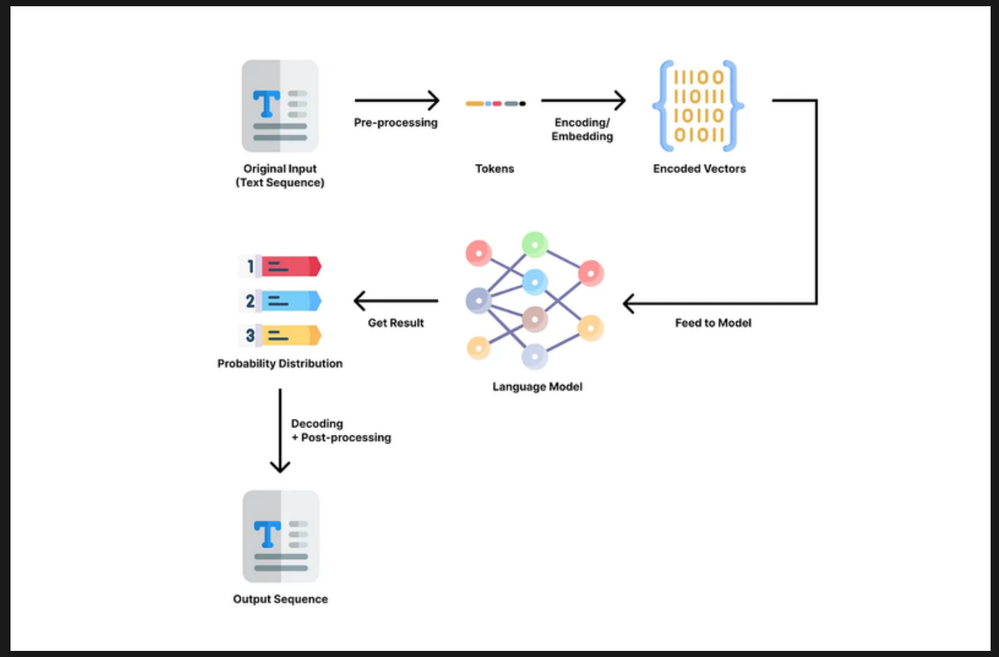
Decoding ChatGPT



In 2015, OpenAI was established by notable figures including Sam Altman and Elon Musk, focusing on various AI technologies beyond GPT. Google's 2017 "Attention is All You Need" paper introduced the transformer architecture, foundational for leading large language models like GPT. GPT debuted in 2018 with a modified transformer architecture, followed by GPT-2 in 2019 with unsupervised multitask learning capabilities, and GPT-3 in 2020, advancing in few-shot learning. In 2022, InstructGPT was released, emphasizing instruction-following through human feedback, alongside ChatGPT, which specializes in human dialogue, both benefiting from reinforcement learning with human feedback. This progression highlights GPT's development and refinement over time.

Diving deep behind the models

- There are many types of AI or deep learning models. For natural language processing (NLP) tasks like conversations, speech recognition, translation, and summarization, we will turn to language models to help us.
- Language models can learn a library of text (called corpus) and predict words or sequences of words with probabilistic distributions, i.e. how likely a word or sequence can occur. For example, when you say “Tom likes to eat …”, the probability of the next word being “pizza” would be higher than “table”. If it’s predicting the next word in the sequence, it’s called next-token-prediction; if it’s predicting a missing word in the sequence, it’s called masked language modeling.
- Since it’s a probability distribution, there can be many probable words with different probabilities. Although you might think it’s ideal to always choose the best candidate with the highest probability, it may lead to repetitive sequences. So in practice, researchers would add some randomness (temperature) when choosing the word from the top candidates.



In a typical NLP process, the input text will go through the following steps:

- Preprocessing: cleaning the text with techniques like sentence segmentation, tokenization (breaking down the text into small pieces called tokens), stemming (removing suffixes or prefixes), removing stop words, correcting spelling, etc. For example, “Tom likes to eat pizza.” would be tokenized into [“Tom”, “likes”, “to”, “eat”, “pizza”, “.”] and stemmed into [“Tom”, “like”, “to”, “eat”, “pizza”, “.”].
- Encoding or embedding: turn the cleaned text into a vector of numbers, so that the model can process.
- Feeding to model: pass the encoded input to the model for processing.
- Getting result: get a result of a probability distribution of potential words represented in vectors of numbers from the model.
- Decoding: translate the vector back to human-readable words.
- Post-processing: refine the output with spell checking, grammar checking, punctuation, capitalization, etc.

Generative AI exists because of the transformer

This | is | how | it | works

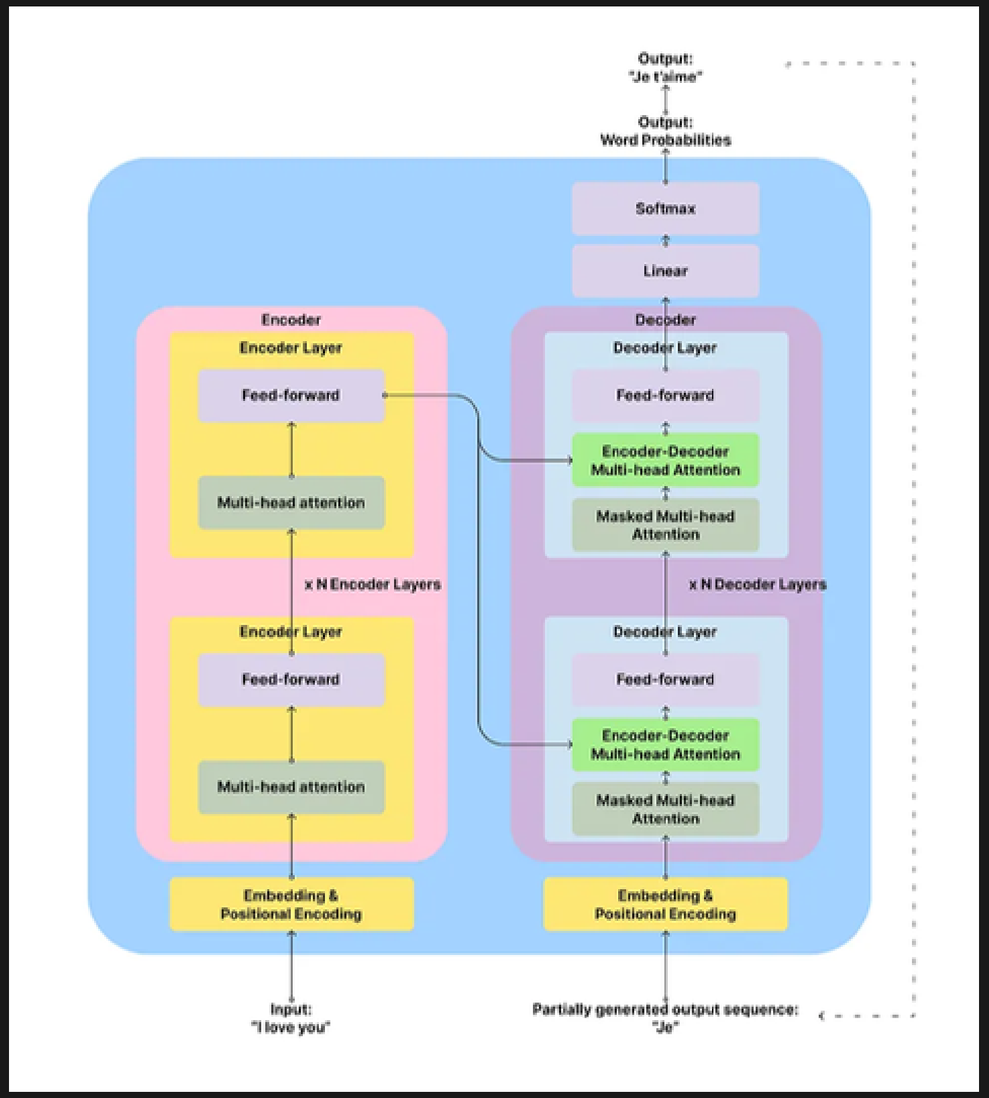
The transformer architecture is the foundation for GPT. It is a type of neural network, which is similar to the neurons in our human brain. The transformer can understand contexts in sequential data like text, speech, or music better with mechanisms called attention and self-attention.

Attention allows the model to focus on the most relevant parts of the input and output by learning the relevance or similarity between the elements, which are usually represented by vectors. If it focuses on the same sequence, it's called self-attention.

Let's take the following sentence as an example: "Tom likes to eat apples. He eats them every day." In this sentence, "he" refers to "Tom" and "them" refers to "apples". And the attention mechanism uses a mathematical algorithm to tell the model that those words are related by calculating a similarity score between the word vectors. With this mechanism, transformers can better "make sense" of the meanings in the text sequences in a more coherent way.

The attention mechanism measures the relevance/similarity between each element.

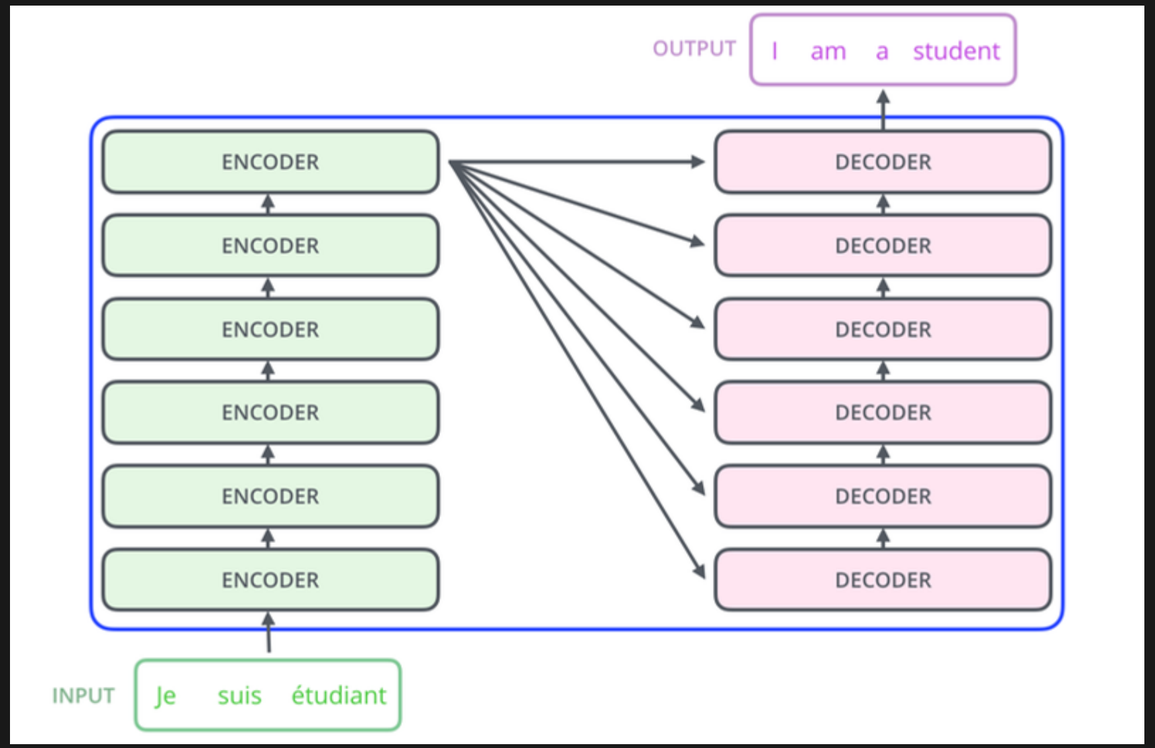
Transformer Architecture



- Embedding & Positional Encoding: turning words into vectors of numbers.
- Encoder: extract features from the input sequence and analyze the meaning and context of it. It outputs a matrix of hidden states for each input token to be passed to the decoder.
- Decoder: generate the output sequence based on the output from the encoder and the previous output tokens.
- Linear & Softmax Layer: turning the vector into a probability distribution of output words.

The encoder and decoder are the main components of transformer architecture. The encoder is responsible for analyzing and “understanding” the input text and the decoder is responsible for generating output.

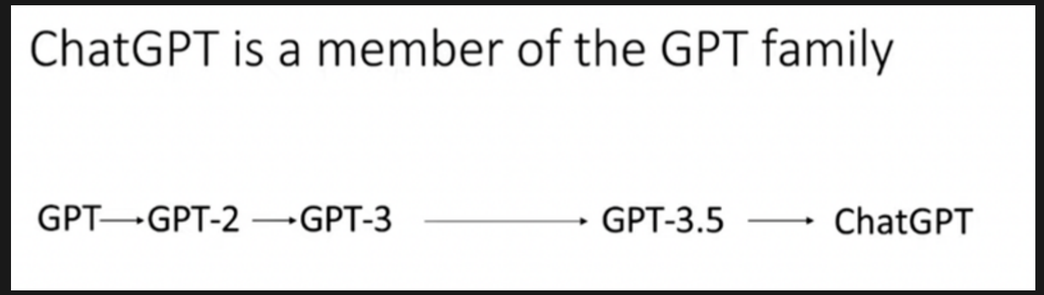
Transformer Architecture



- The encoder is a stack of multiple identical layers (6 in the original transformer paper). Each layer has two sub-layers: a multi-head self-attention layer and a feed-forward layer, with some connections, called residual connection and layer normalization. The multi-head self-attention sub-layer applies the attention mechanism to find the connection/similarity between input tokens to understand the input. The feed-forward sub-layer does some processing before passing the result to the next layer to prevent overfitting. You can think of encoders like reading books — you will pay attention to each new word you read and think about how it's related to the previous words.
- The decoder is similar to the encoder in that it's also a stack of identical layers. But each decoder layer has an additional encoder-decoder attention layer between the self-attention and feed-forward layers, to allow the decoder to attend to the input sequence. For example, if you're translating "I love you" (input) to "Je t'aime" (output), you will need to know "Je" and "I" are aligned and "love" and "aime" are aligned.
- The multi-head attention layers in the decoder are also different. They're masked to not attend to anything to the right of the current token, which has not been generated yet. You can think of decoders like free-form writing — you write based on what you've written and what you've read, without caring about what you're going to write.

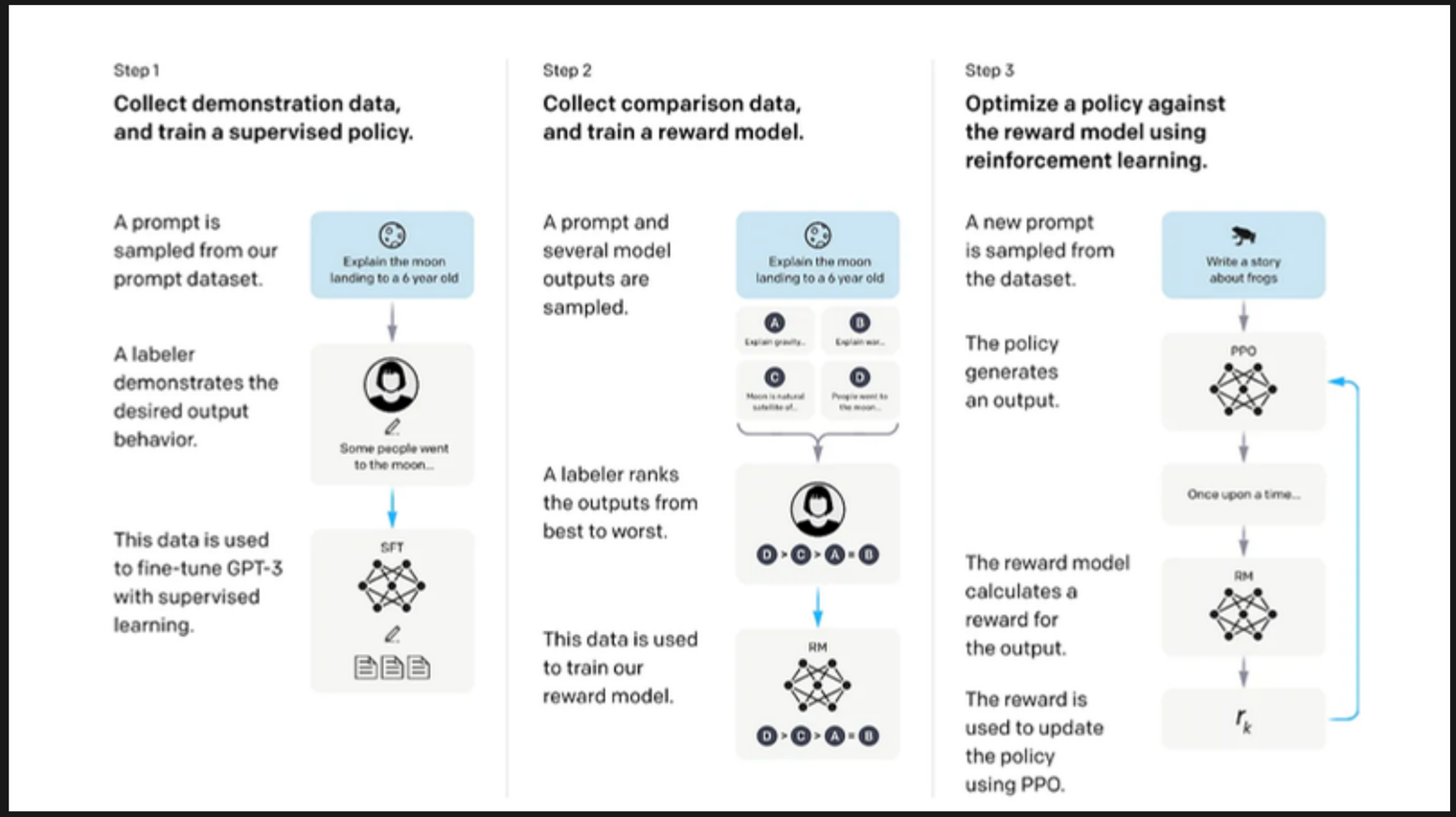
From transformers to GPT, GPT2, and GPT3

- GPT's full name is Generative Pre-trained Transformer. From the name, we can see that it's a generative model, good at generating output; it's pre-trained, meaning it has learned from a large corpus of text data; it's a type of transformer.
- In fact, GPT uses only the decoder part of the transformer architecture. Decoders are responsible for predicting the next token in the sequence. GPT repeats this process again and again by using the previously generated results as input to generate longer texts, which is called auto-regressive. For example, if it's translating "I love you" to French, it will first generate "Je", then use the generated "Je" to get "Je t'aime".
- In training the first version of GPT, researchers used unsupervised pre-training with the BookCorpus database, consisting of over 7000 unique unpublished books. Unsupervised learning is like having the AI read those books itself and try to learn the general rules of language and words. On top of the pre-training, they also used supervised fine-tuning on specific tasks like summarization or question and answering. Supervised means that they will show the AI examples of requests and correct answers and ask the AI to learn from those examples.



- In GPT-2, researchers expanded the size of the model (1.5B parameters) and the corpus they feed to the model with WebText, which is a collection of millions of web pages, during the unsupervised pre-training. With such a big corpus to learn from, the model proved that it can perform very well on a wide range of language related-tasks even without supervised fine-tuning.
- In GPT-3, the researchers took a step further in expanding the model to 175 billion parameters and using a huge corpus comprising hundreds of billions of words from the web, books, and Wikipedia. With such a huge model and a big corpus in pre-training, researchers found that GPT-3 can learn to perform tasks better with one (one-shot) or a few examples (few-shot) in the prompt without explicit supervised fine-tuning.
- At this stage, the GPT-3 model is already impressive. But they're more like general-purpose language models. Researchers wanted to explore how it can follow human instructions and have conversations with humans. Therefore, they created InstructGPT and ChatGPT based on the general GPT model.

Teaching GPT to interact with humans: InstructGPT and ChatGPT



After the iterations from GPT to GPT-3 with growing models and corpus size, researchers realized that bigger models don't mean that they can follow human intent well and may produce harmful outputs. Therefore, they attempted to fine-tune GPT-3 with supervised learning and reinforcement learning from human feedback (RLHF). With these training steps came the two fine-tuned models — **InstructGPT and ChatGPT**.

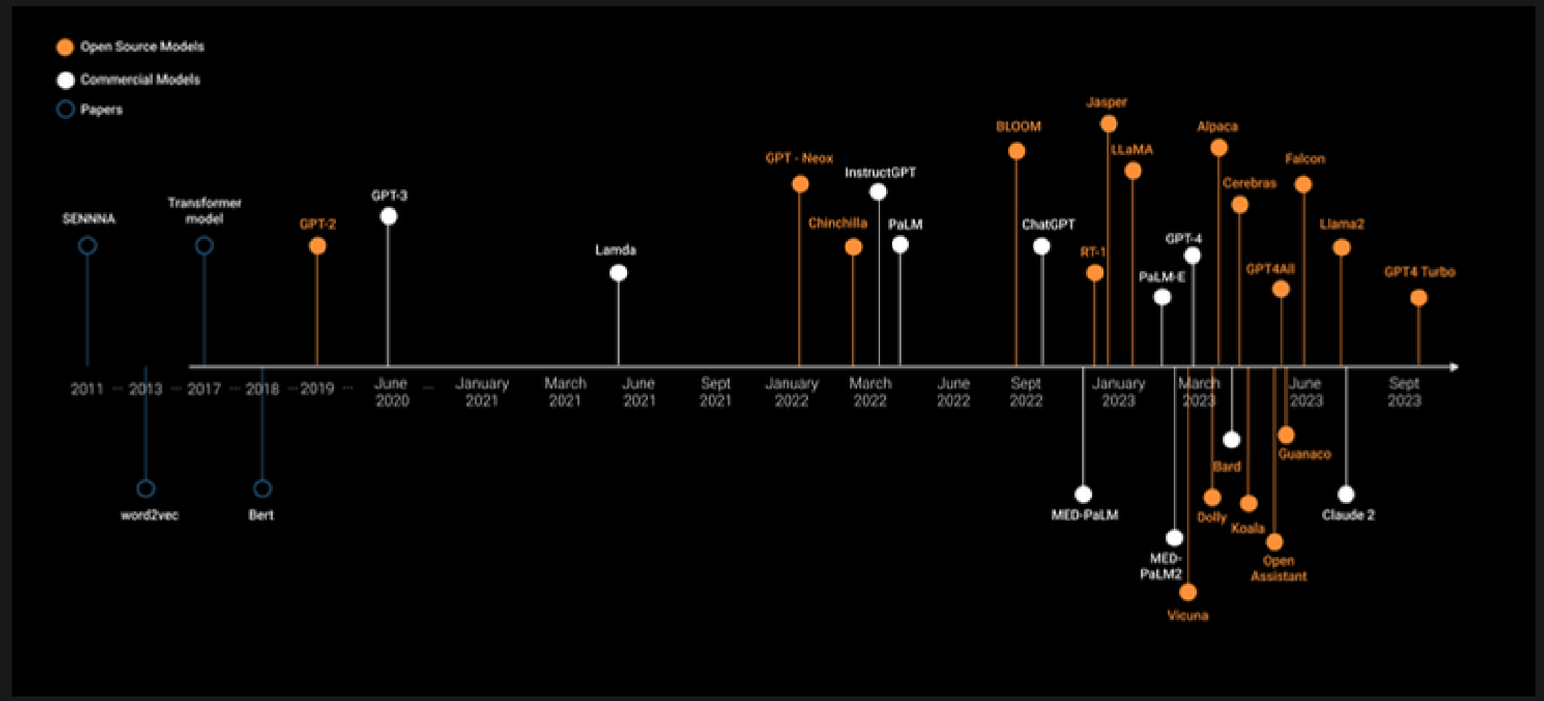
- The first step is supervised learning from human examples. Researchers first provided the pre-trained GPT with a curated, labeled dataset of prompt and response pairs written by human labelers. This dataset is used to let the model learn the desired behavior from those examples. From this step, they get a supervised fine-tuned (SFT) model.

- The second step is training a reward model (RM) to rate the responses from the generative model. Researchers used the SFT model to generate multiple responses from each prompt and asked human labelers to rank the responses from best to worse by quality, engagement, informativeness, safety, coherence, and relevance. The prompts, responses, and rankings are fed to a reward model to learn human preferences of the responses through supervised learning. The reward model can predict a scalar reward value based on how well the response matches human preferences.

- In the third step, researchers used the reward model to optimize the SFT model's policy through reinforcement learning. The SFT model will generate a response from a new prompt; the reward model will assess the response and give it a reward value that approximates human preference; the reward is then used to optimize the generative model by updating its parameters. For example, if the generative model generates a response that the reward model thinks humans may like, it will get a positive reward to continue generating similar responses in the future; and vice versa.

- Through this process with supervised learning and reinforcement learning from human feedback, the InstructGPT model (with only 1.3B parameters) is able to perform better in tasks that follow human instructions than the much bigger GPT-3 model (with 175 B parameters). ChatGPT is a sibling model to InstructGPT. The training process is similar for ChatGPT and InstructGPT, including the same methods of supervised learning and RLHF. The main difference is that ChatGPT is trained with examples from conversational tasks, like question answering, chit-chat, trivia, etc. Through this training, ChatGPT can have natural conversations with humans in dialogues. In conversations, ChatGPT can answer follow-up questions and admit mistakes, making it more engaging to interact with.

Other LLMs



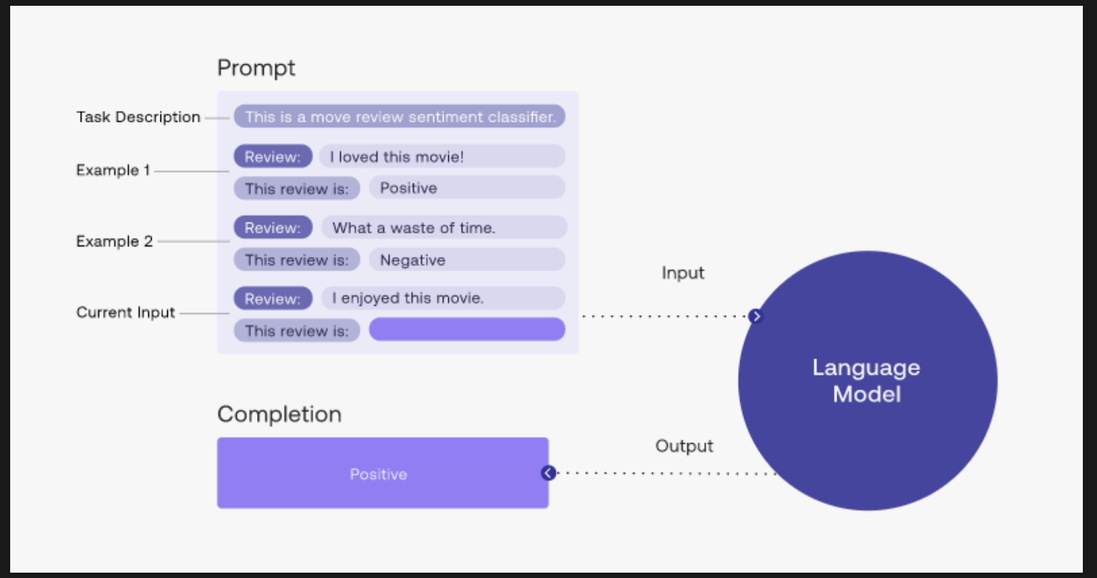
LLaMA: An open-source collection of LLMs developed by Meta. LLaMA is designed to help researchers in advancing their work in the subfield of LLMs. It is available in multiple sizes, ranging from 7 to 65 billion parameters, and aims to democratise the access to LLMs by requiring less computing power and resources. LLaMA can only be used for research.

How Large Language Models are trained for your case

There are three main approaches to align the model's output:
1. Prompting,
2. Retrieval-Augmented Generation (RAG)
3. and the more advanced finetuning.
Combining your LLMs with the right knowledge (e.g. documents specific to your business) and templates that define how it should act in certain cases (e.g. using a prompt management system) allow your solution to reach its full potential.

Prompting

A prompt is the text you provide to an LLM as input. Prompts can be short and concise, or can be extensive, including additional context and requirements you have regarding the output.
Some common prompting techniques:
- Zero-shot Prompting:equivalent with "describing a task to a student"
- Few-shot Prompting: equivalent with "describing a task to a student and supplying some examples of similar tasks and how they were carried out.

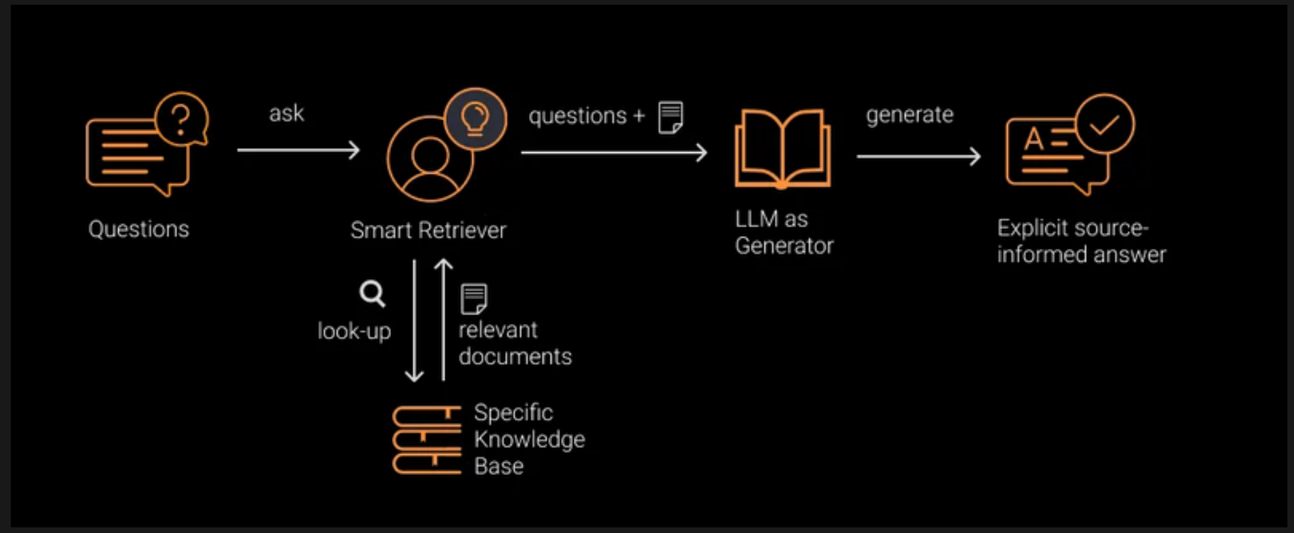


How Large Language Models are trained for your case

Retrieval-Augmented Generation

A more advanced variant of prompting is called RAG. In short, a RAG architecture introduces a component that fetches documentation (relevant for the question that was asked) from your knowledge base. By placing this "Smart Retriever" component in front of your conversational LLM, you impose that LLM to base its response on the information present within your documentation.

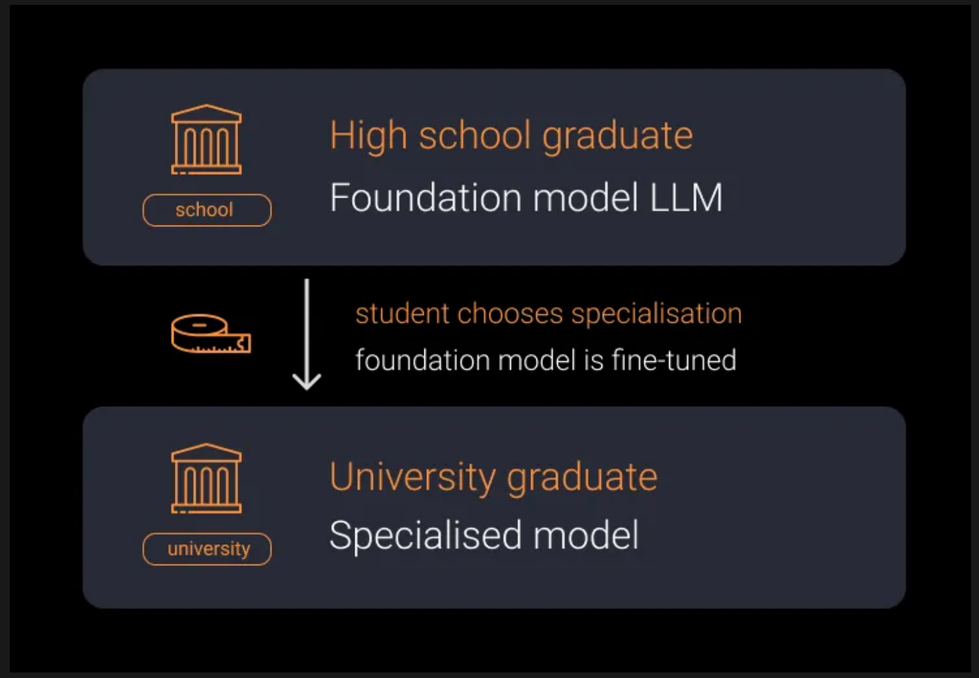
The benefits from such an architecture are that (1) your LLM can explicitly refer to the sources upon which it based its answer, (2) your LLM is unlikely to hallucinate, because it receives the context within which it should stay and (3) your complete solution remains maintainable because the "Smart Retriever" component can be updated as your knowledge base grows.



How Large Language Models are trained for your case

When prompting or a RAG approach are not fulfilling, (additional) fine-tuning techniques can be considered. Finetuning alters the LLM itself by affecting the weights that make up its neural network).

Here, let's represent the LLM as a high school student.
In the finetuning case, we are changing the actual capacities of the LLM/student. We go further than just providing him "examples ad-hoc" like we do for the few-shot prompting case. Instead, we present the student with input tasks and correct his output based on the "correct" outputs that our fine-tuning dataset describes. One could see this as "taking the student to the next level": allowing the student to specialise within the domain that your dataset implicitly describes. The analogy between a high school graduate and a university graduate is then straightforward.



In conclusion, for many business cases, the "few-shot prompting" approach may suffice to get the model to behave appropriately (e.g. when one is looking for a conversational LLM that replies based on information extracted from their knowledge base). In other cases, however, some specific behaviour of your model may require fine-tuning of your model in order to make it perform better.

How Large Language Models are trained for your case