Large Language Models: From Theory to Implementation

Workshop Details

- Instructor: Tanya Khanna
- Email: tk759@scarletmail.rutgers.edu
- Course Materials: Github Link - Data-Science-Workshop---Spring-2025---NBL-
- Workshop Recordings: Libguides

Workshops Schedule

Introduction to Python Programming	February 3, 2025; 2 - 3:30 PM
Mastering Data Analysis: Pandas and Numpy	February 10, 2025; 2 - 3:30 PM
Introduction to Tableau: Visualizing Data Made Easy	February 17, 2025; 2 - 3:30 PM
Introduction to Machine Learning: Supervised Learning	February 24, 2025; 2 - 3:30 PM
Introduction to Machine Learning: Unsupervised Learning	March 3, 2025; 2 - 3:30 PM
Data-Driven Decision Making: A/B Testing and Statistical Hypothesis Testing	March 10, 2025; 2 - 3:30 PM
Demystifying Generative AI	March 24, 2025; 2 - 3:30 PM
Large Language Models: From Theory to Implementation	March 31, 2025; 2 - 3:30 PM
Generative AI Applications with AI Agents	April 7, 2025; 2 - 3:30 PM
Building Intelligent Recommendation Systems	April 14, 2025; 2 - 3:30 PM

Workshop Agenda

- NLP: Bridging Humans and Machines
  - Key Problems in AI
  - Core Components of NLP
  - Major NLP Techniques
  - Applications and Challenges in NLP
- What are Large Language Models?
- How did AI get here, and why the hype?
- Challenges of LLMs
- Decoding ChatGPT
- How to use LLMs for your use case
- Practical Session

same words, different meanings

- “I need to stop by the bank to deposit a check.”
  - Is it a riverbank or a financial bank?
- “He’s tied up at the office.”
  - Does it mean he’s busy with work or physically restrained?
- “She broke the record.”
  - Did she set a new high score or damage a music record?
- “I saw bats last night.”
  - Were they flying mammals or baseball bats?
- “The light was out.”
  - Was the light turned off or located outside?

Natural Language Processing (NLP)

Just like we can tell if “bank” means a place for money or a riverbank, NLP models capture context clues to understand which meaning fits best. It’s like giving the computer a little “common sense” to figure out what we mean! NLP helps a computer interpret the meaning behind our words, not just by looking at individual words but by considering the whole sentence, tone, and even the surrounding conversation. By understanding context, NLP enables computers to respond more naturally, making interactions feel smooth, almost like talking to a human. It’s the difference between a robotic response and one that feels empathetic or clever. That’s the magic behind how a chatbot understands you, a voice assistant responds, or how search engines give you exactly what you’re looking for. This context-based understanding is what powers all Large Language Models, including ChatGPT, making them not just machines that repeat information but tools that engage in conversations that feel meaningful and relevant.

Why is NLP important?

2.5

quintillion bytes of data generated daily - a significant portion of which is unstructured text, NLP allows for the automated analysis of this vast volume of text data. This capability aids in market research, customer service, and more by extracting valuable insights from online reviews, social media posts, customer feedback, etc.

85%

of smartphone users utilize voice assistants, like Siri and Alexa, for various tasks. Predictive text, powered by NLP, assists in over 60% of mobile communications, enhancing typing efficiency and accuracy.

$68.1bn

market by 2028, with an average annual growth rate of 29.3%, driven by increasing demand for NLP solutions across industries to extract insights from unstructured data.

History and Evolution of NLP

ELIZA (1966)

The first rule based chatbot, ELIZA, is created by Joseph Weizenbaum, simulating a psychotherapist.

Rise of Statistical Methods (1980)

Transition to statistical methods in NLP, focusing on machine learning algorithms for language processing.

WordNet (1995)

A large lexical database of English, marking significant progress in linguistic databases.

Stanford Parser (2003)

An important tool for linguistic analysis and parsing sentences.

Seq2Seq Models (2014)

Introduction of sequence-to-sequence models, enhancing translation and text summarization.

BERT (2018)

Google introduces BERT, a breakthrough in machine understanding of context in text.

GPT-3 (2020)

OpenAI releases GPT-3, a state-of-the-art language model with broad applications in writing, conversation, and more.

Large-scale Multimodal Models (2021-present)

Advancements in integrating text with other data types (e.g., images, videos) to understand and generate multimodal content.

Core Components of NLP

Text Parsing

It is the process of analyzing a text's structure and recognizing its meaningful parts. It involves breaking down text into tokens (such as words and phrases) and identifying its grammatical structure. This is the first step in understanding the text at a deeper level.

Syntactic Analysis (Syntax)

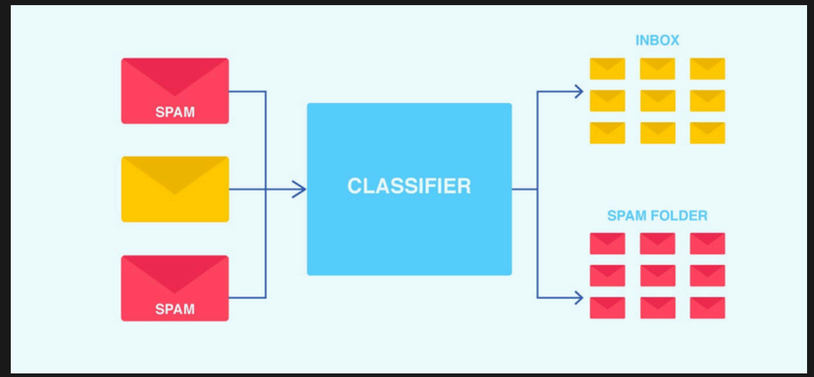
This examines how words are organized into sentences, looking at the grammatical structure. This process identifies the relationships between words, such as subjects, predicates, objects, and other elements of a sentence, to understand the rules that govern sentence construction.

Semantic Analysis (Semantics)

This focuses on the meaning of individual words, phrases, sentences, and the text as a whole. It interprets the meanings that language conveys, beyond just the dictionary definitions of words, considering context and how the meaning of sentences changes with different word arrangements.

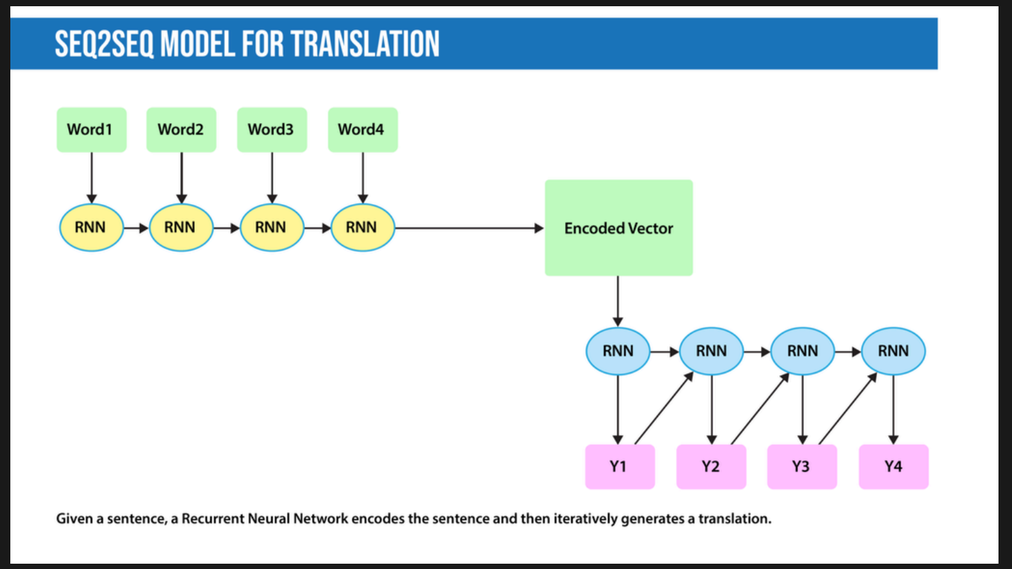
Machine Learning in *NLP*

- Machine learning in NLP involves using algorithms to analyze, understand, and generate human language. The models learn from vast amounts of text data to perform tasks like translation, sentiment analysis, and more.
- **Pre-Deep Learning Era:** Before deep learning took the stage, machine learning in NLP relied on models like decision trees, support vector machines (SVMs), and linear regression, often with handcrafted features.
- **Examples:**
  - Spam Detection: Early machine learning models were trained to identify and filter out spam emails with high accuracy.
  - Part-of-Speech Tagging: Tagging words in a sentence as nouns, verbs, adjectives, etc., using algorithms like Hidden Markov Models (HMMs).
- **Impact:** These models significantly improved the automation of text analysis tasks, reducing the reliance on rule-based systems.

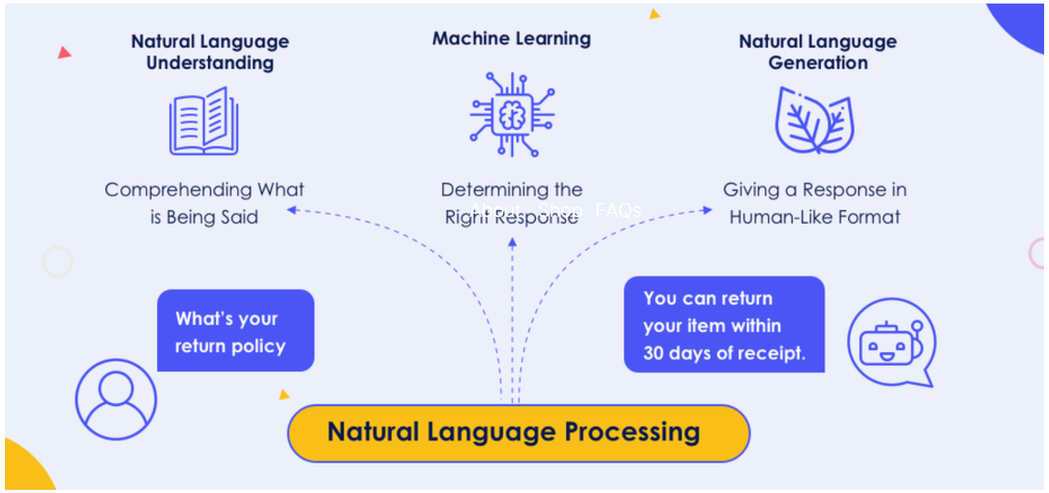


Deep Learning in *NLP*

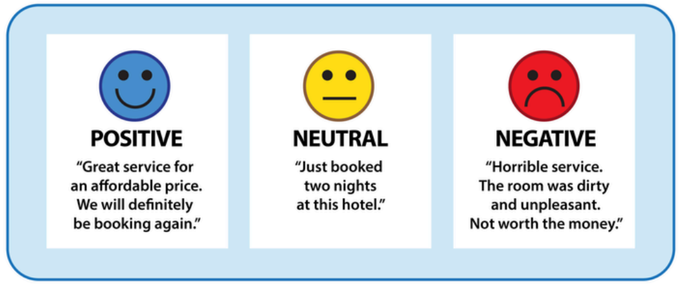
- Deep learning is a type of machine learning that uses networks with many layers to understand complex data patterns. In language tasks, it helps models grasp meaning and context better, making them “smarter” in understanding human language.
- Key Models:
  - RNNs: Good for working with sequences like sentences, where each word can depend on the previous ones.
  - Transformers: A newer type of model that focuses on important parts of a sentence, leading to much better understanding.
- Examples:
  - Translation: Google’s deep learning-based translator gives more natural translations by understanding context better.
  - Language Understanding: Models like BERT and GPT can understand context well and create text that sounds very human-like.
- Impact: Deep learning now allows computers to understand and respond in ways much closer to real human communication, powering live translations, smart chatbots, and better sentiment analysis.



Major *NLP Techniques*



SENTIMENT ANALYSIS



Given text, sentiment analysis classifies its emotional quality.

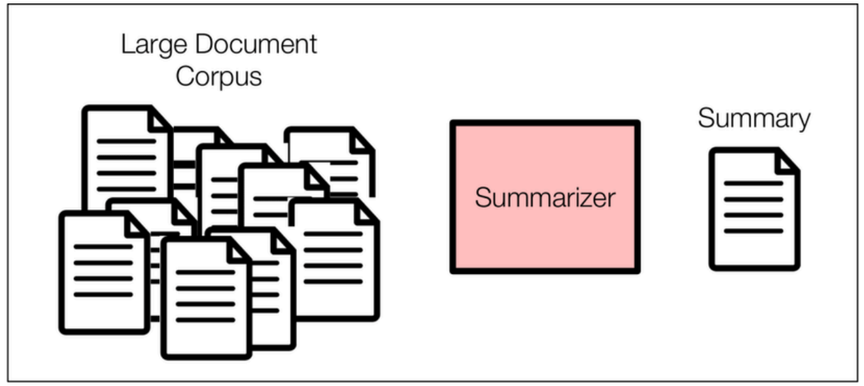
NAMED ENTITY RECOGNITION (NER) TAGGING

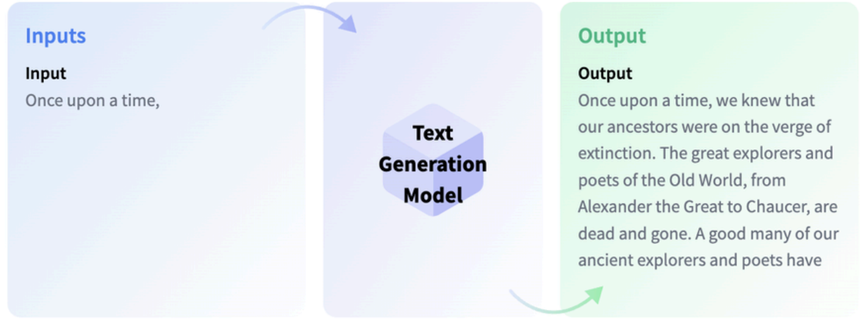
Andrew Yan-Tak Ng PERSON (Chinese NORP : 吳恩達; born 1976 DATE) is a British NORP -born American NORP computer scientist and technology entrepreneur focusing on machine learning and AI GPE . Ng was a co-founder and head of Google Brain ORG and was the former chief scientist at Baidu ORG , building the company's Artificial Intelligence Group ORG into a team of several thousand CARDINAL people.

spaCy named entity recognition tagging of the first paragraph of Andrew Ng's Wikipedia page. "NORP" stands for nationalities or religious or political groups. Note that spaCy incorrectly labels "AI" as "GPE," for geopolitical entity.

NLU

NLG





Applications of NLP

Virtual Assistants

NLP powers virtual assistants like Siri, Alexa, and Google Assistant, allowing users to interact with their devices using natural language.

- Asking Siri to set a reminder
- Requesting Alexa to play a specific song
- Asking Google Assistant for the weather forecast

Sentiment Analysis

NLP can analyze text to determine the sentiment expressed, helping companies understand customer opinions and feedback.

- Analyzing social media posts to gauge customer satisfaction
- Monitoring product reviews to identify positive or negative sentiment
- Assessing customer support chat logs for sentiment analysis

Language Translation

NLP enables automatic language translation, making it easier for people to communicate across different languages.

- Using Google Translate to translate a webpage
- Translating text messages in real-time using language translation apps
- Converting subtitles of a foreign movie to your native language

Text Summarization

NLP can automatically generate summaries of long texts, saving time and providing concise information.

- Using an app to summarize news articles
- Generating abstracts for research papers
- Creating brief descriptions of long emails

Spam Detection

NLP helps identify and filter out spam emails, ensuring that important messages reach the intended recipients.

- Email providers automatically moving suspicious emails to the spam folder
- Flagging potential phishing emails based on their content
- Detecting and blocking spam comments on websites

Challenges in NLP

Ambiguity

Language is inherently ambiguous. Words can have multiple meanings (polysemy), sentences can be interpreted in different ways, and pronouns can have ambiguous antecedents, making it difficult for NLP systems to determine the correct interpretation.

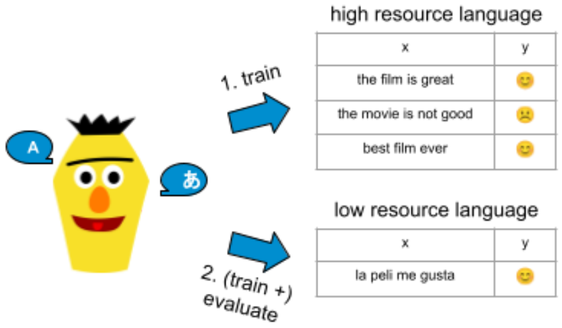
Does a word clever have a negative nuance sometimes? Is there any difference between two sentences below? He is smart. and He is clever.

Understanding Nuances

Capturing the subtleties such as emotion, tone, and nuance in text is a complex task. This includes determining the author's intentions, the strength of their opinions, or the mood they are trying to convey.

Data Bias and Fairness

Machine learning models in NLP can inadvertently learn and perpetuate biases present in their training data. Ensuring that NLP systems are fair and unbiased is a significant challenge.



Lack of Resources for Low-Resource Languages

While NLP has made great strides in languages with abundant data (like English), there is a lack of annotated datasets and resources for many lesser-spoken languages, which hampers the development of NLP technologies for those languages.

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.

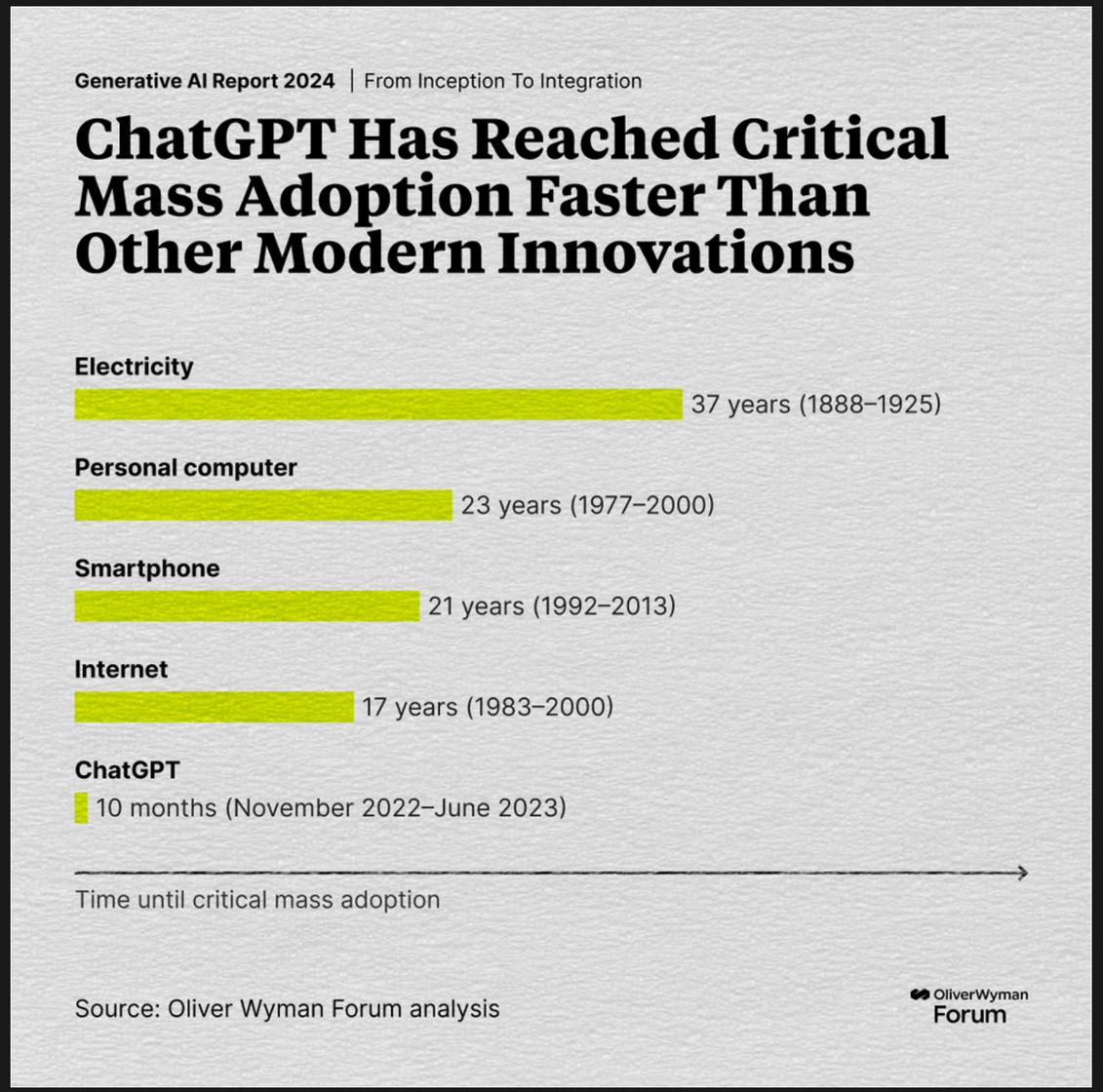
MODEL

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 over 2 years 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.
- 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.

WHAT THEY ARE

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

WHAT THEY AREN'T

- While LLMs can access and process a vast amount of information, they do not possess consciousness or independent thought.
- Their outputs are based on patterns in data they were trained on; they can make mistakes, misunderstand questions, or generate inaccurate or biased information.
- LLMs are tools to augment human capabilities, not replace them. They cannot replicate the depth of human expertise or emotional intelligence.
- 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.

LLMs are good at:

Language translation
Content summarization
Content generation
Writing support
Question answering
Correcting spelling and grammar
Programming support
Classification (spam detection)
Simplifying complex content
Stylized writing (applying Poe to x)
Personalization
Prompt engineering
Speech recognition
Mimicking dialogue
Sentiment analysis

LLMs are NOT good at:

Current events
Common sense
Math/counting
Handling uncommon scenarios
Humor
Consistency
High-level strategy
Being factual 100% of the time
Being environmentally friendly
Understanding context
Reasoning & logic
Emotional intelligence
Any data-driven research
Representing minorities
Extended recall/memory

Challenges of LLMs

Reliability and Hallucinations

One of the biggest problems is that these AI models often make mistakes or simply make things up. They can sound very confident while giving wrong information, and it's hard to tell when they're guessing versus when they really know something. This makes it tricky to trust them completely without double-checking their work.

Bias and Fairness

These AI systems can also show unfair bias, treating some groups of people differently than others. This happens because they learn from human-written materials that might contain prejudices. It's like they're picking up bad habits from their training, which means they might work better for some people than others.

Safety and Alignment

Keeping AI systems under control is another major challenge. We need to make sure they help people rather than cause problems, but it's not easy to make them behave exactly as we want. As these systems get more powerful, this becomes even more important.

Technical Limitations

They can't remember very long conversations, don't know about current events, and often get confused by complex problems. Sometimes their answers become muddled or inconsistent, especially in longer conversations.

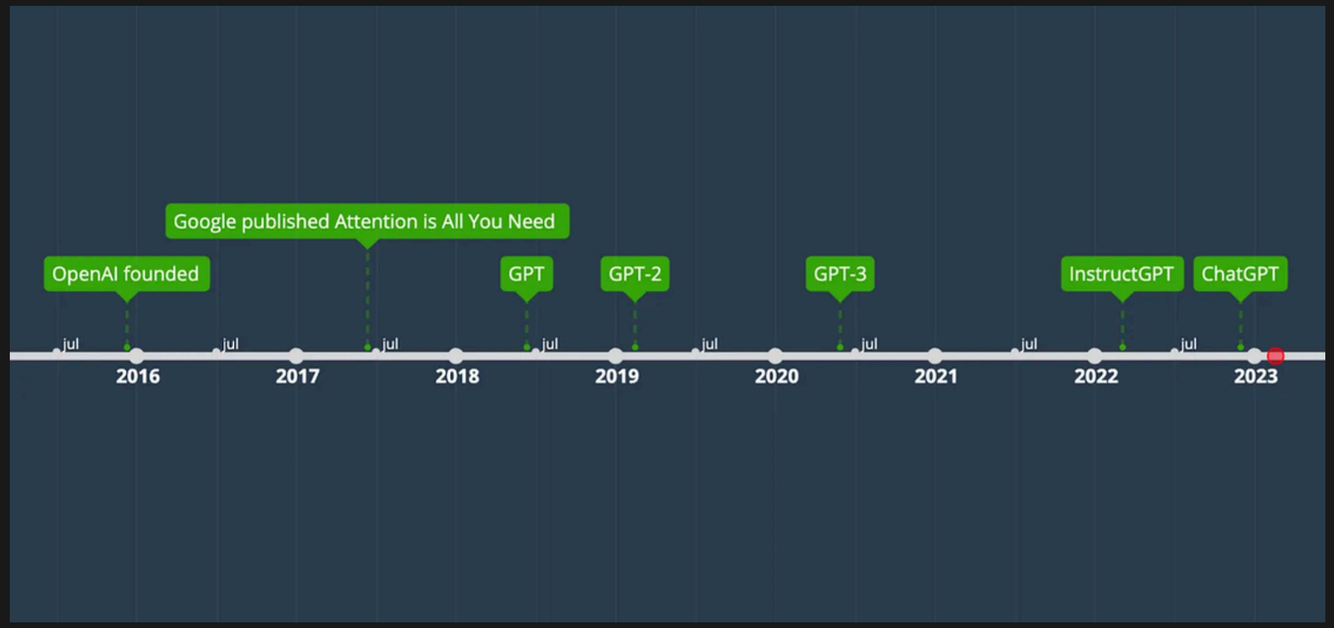
Privacy and Security

These systems might accidentally reveal private information they've learned during training, or they could be misused by people with bad intentions. Protecting user data and preventing harmful use is a constant challenge.

Transparency and Interpretability

Perhaps most frustrating is how hard it is to understand how these AI models think. They work like a black box – we put information in and get answers out, but we can't really see how they make their decisions. This makes it difficult to fix problems when they arise or improve how they work.

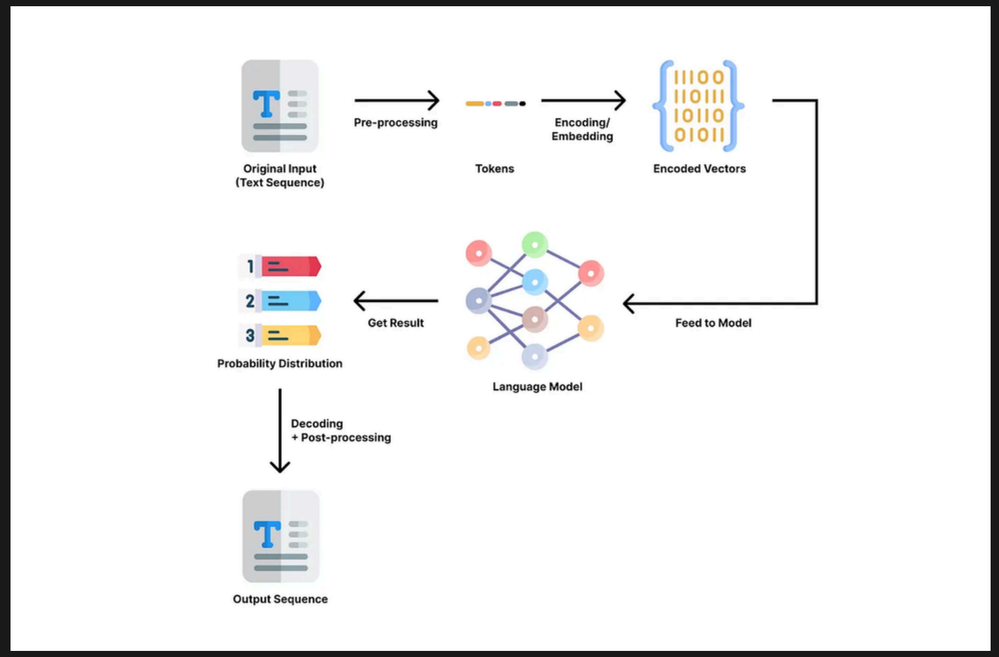
Decoding ChatGPT



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 one 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:

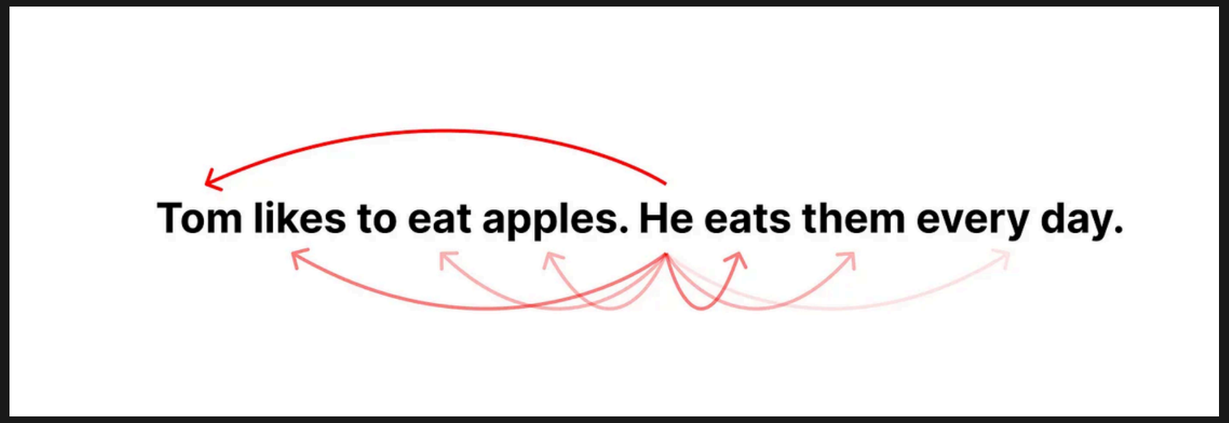
- Preprocessing: cleaning the text with techniques like sentence segmentation, tokenization (breaking down the text into small pieces called tokens), stemming (removing suffixes or prefixes), removing stop words, correcting spelling, etc. For example, “Tom likes to eat pizza.” would be tokenized into [“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.

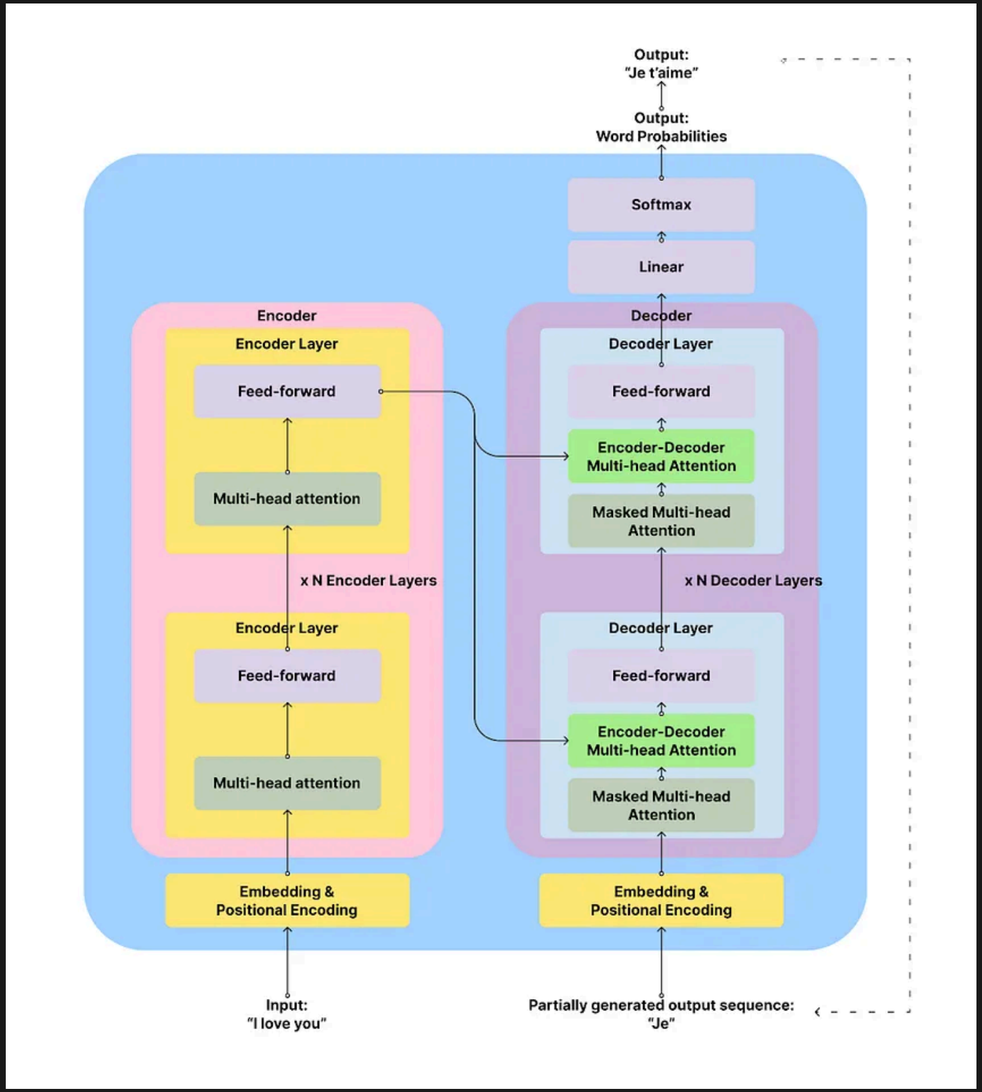
Attention allows the model to focus on the most relevant parts of the input and output by learning the relevance or similarity between the elements, which are usually represented by vectors. If it focuses on the same sequence, it's called self-attention.



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.

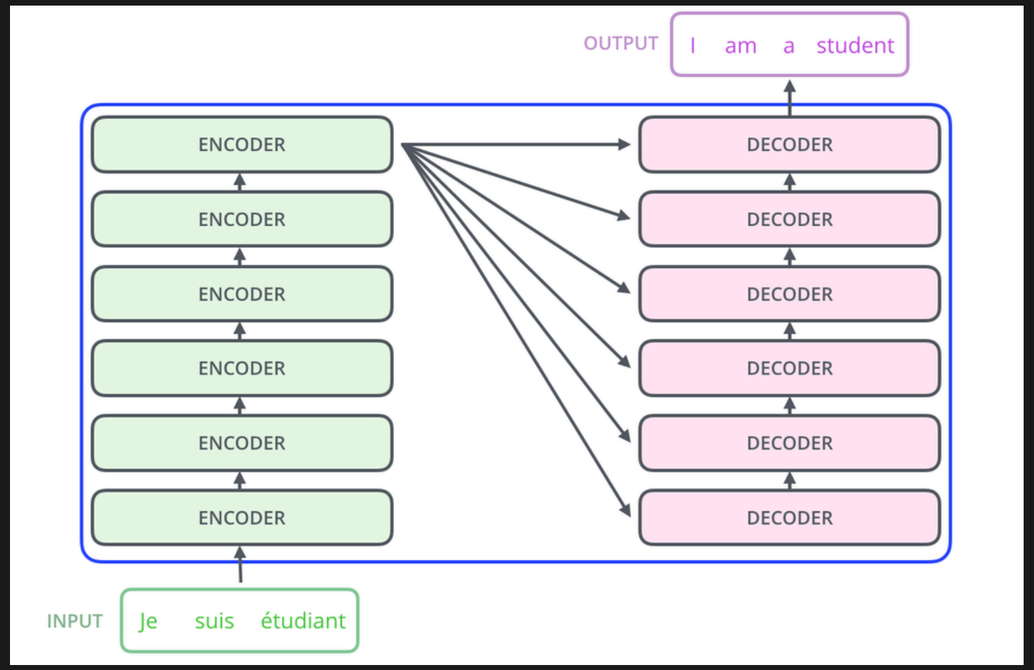
The attention mechanism measures the relevance/similarity between each element.

Transformer Architecture



- *Embedding & Positional Encoding:* Changes words into numbers the computer can understand like giving each word a special code, also marks where each word appears in the sentence. Example: "cat" becomes something like [0.2, 0.5, 0.1]
- *Encoder*: Reads and understands the input and looks at how words relate to each other. Like a smart reader that spots patterns, it creates a detailed map of what the text means. Example: "The cat sat" → understands 'cat' is the subject doing the sitting
- *Decoder*: Takes the encoder's understanding and creates new text, like a writer using notes to write a story it looks at what it wrote before to decide what comes next. Example: Using understanding of "cat sat" to generate "on the mat"
- *Linear & Softmax Layer*: Picks the most likely next word; like having a list of words and choosing the best fit. Gives each possible word a score. Picks the word with highest score. Example: Choosing between "mat" (90% likely) vs "hat" (10% likely)

Transformer Architecture



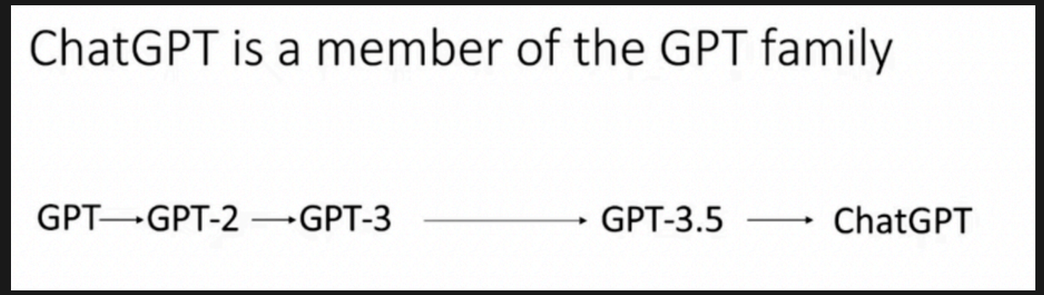
The encoder works like a careful reader, with six different layers that each process the text in the same way. Think of each layer as a different reading pass over the text. In each layer, there are two important steps. First, there's the multi-head attention step, which is like when you notice how words relate to each other in a sentence. For example, when you read "The cat ate," you naturally connect that "ate" needs "cat" as its subject. After this, there's a feed-forward step that processes this information further, helping the system learn rather than just memorize. It's similar to how when you read a book, you don't just memorize the words – you process and understand them.

The decoder is like a writer who's using their understanding to create new text. It's structured similarly to the encoder with multiple layers, but each layer has an extra step. First, it looks at what it's already written, then it checks back with what it learned from the encoder (like referring back to notes while writing), and finally processes all this information. A key feature is that the decoder can only look at what it's already written, not what comes next – just like when you're writing a story, you can only work with the words you've already put down. For example, when translating "I love you" to "Je t'aime," the decoder needs to understand that "I" matches with "Je" and "love" matches with "aime," while using this understanding to build the translation one word at a time.

This whole process mirrors how a person might read something carefully and then explain it to someone else – you first need to understand the material thoroughly (encoder) before you can explain it in your own words (decoder). Each step builds on the previous one, creating a flowing process of understanding and generating text.

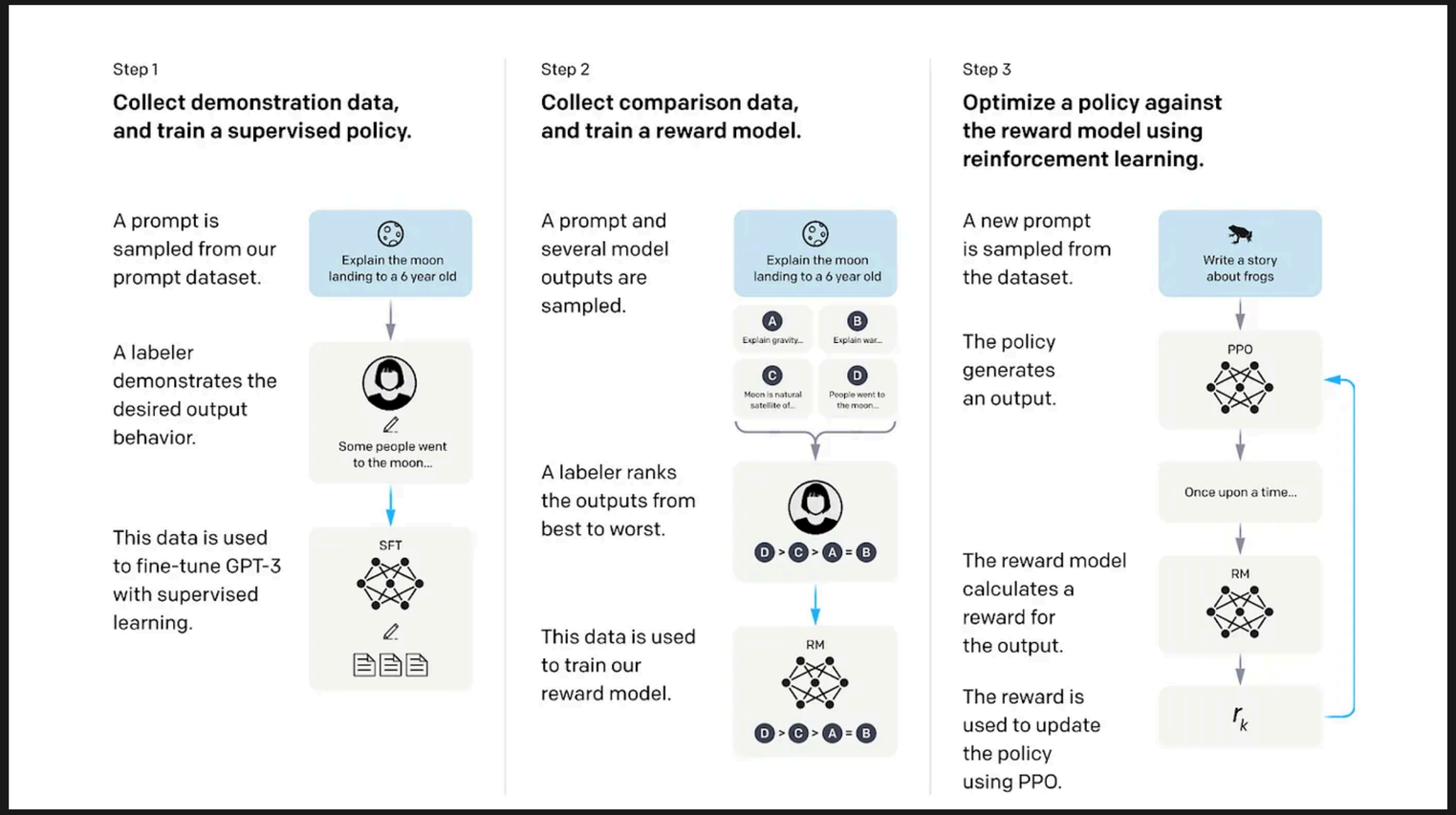
From transformers to GPT, GPT2, and GPT3

- 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

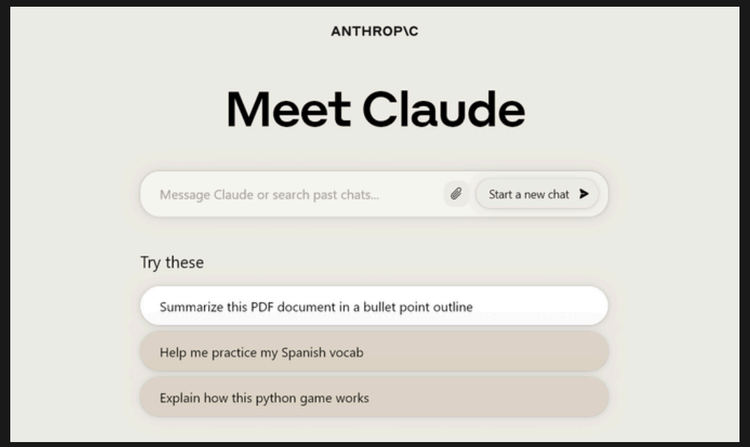


After the iterations from GPT to GPT-3 with growing models and corpus size, researchers realized that bigger models don't mean that they can follow human intent well and may produce harmful outputs. Therefore, they attempted to fine-tune GPT-3 with supervised learning and reinforcement learning from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

- The training starts with supervised learning, where the model learns from examples of good conversations. Think of this like a student learning from a textbook of perfect examples. Researchers give the model many pairs of questions and answers written by humans, helping it understand what good responses look like. This creates what's called a supervised fine-tuned (SFT) model.
- Next comes the development of a reward model, which learns to judge how good responses are. The SFT model creates several different answers to the same question, and human raters rank these answers from best to worst. They look at things like quality, how engaging it is, how informative and safe it is, and whether it makes sense and stays on topic. This reward model learns to predict what humans would consider a good response, like a teacher learning to grade papers consistently.
- The third step uses reinforcement learning with the reward model. This is like training a pet with treats - when the model gives good responses (as judged by the reward model), it's encouraged to give similar responses in the future. When it gives poor responses, it learns to avoid that type of answer. The model keeps practicing and improving based on this feedback.
- This training process proved so effective that even a smaller model (InstructGPT with 1.3B parameters) could outperform its much larger cousin (GPT-3 with 175B parameters) at following human instructions. ChatGPT uses the same training approach but focuses specifically on conversation skills. It's trained on examples of dialogues, questions and answers, and casual chats. This is why ChatGPT can have natural conversations, remember context from earlier in the chat, and even admit when it makes mistakes - making it feel more like talking to a real person.

Other LLMs

Google Models:
- Gemini: Their newest and most capable model
- PaLM: Their previous large model
- BERT: Famous for understanding language context
- LaMDA: Focused on conversations



Anthropic Models:
- Claude 3 (Opus, Sonnet, Haiku): Latest versions with different capabilities
- Claude 2: Previous version known for analysis and coding

Meta (Facebook) Models:
- LLaMA: Open source model that others can build on
- BART: Good at summarizing text
- OPT: Their attempt at open-source AI

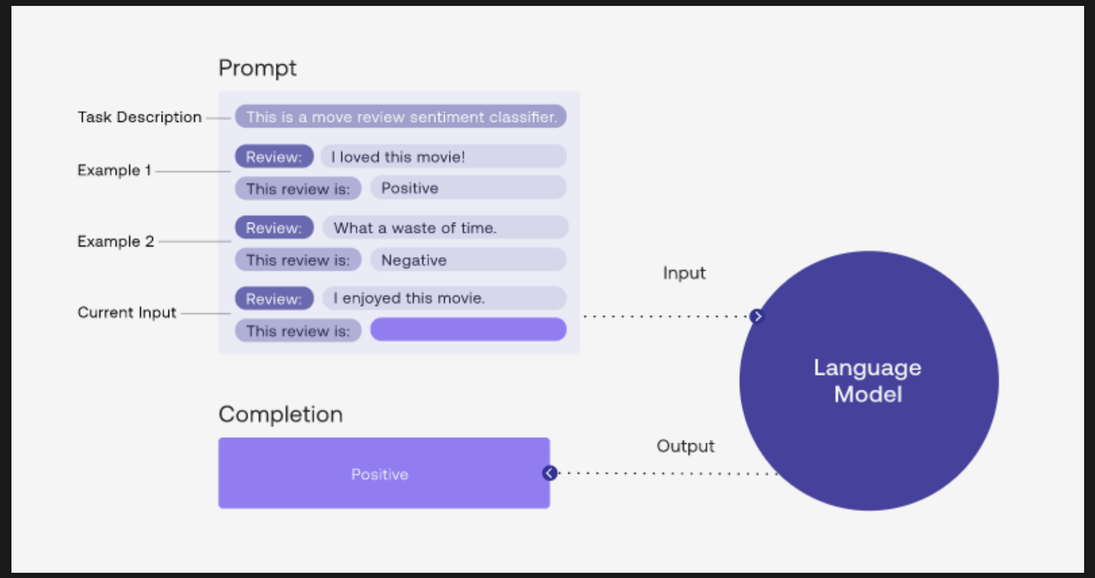
How Large Language Models are trained for your case

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.

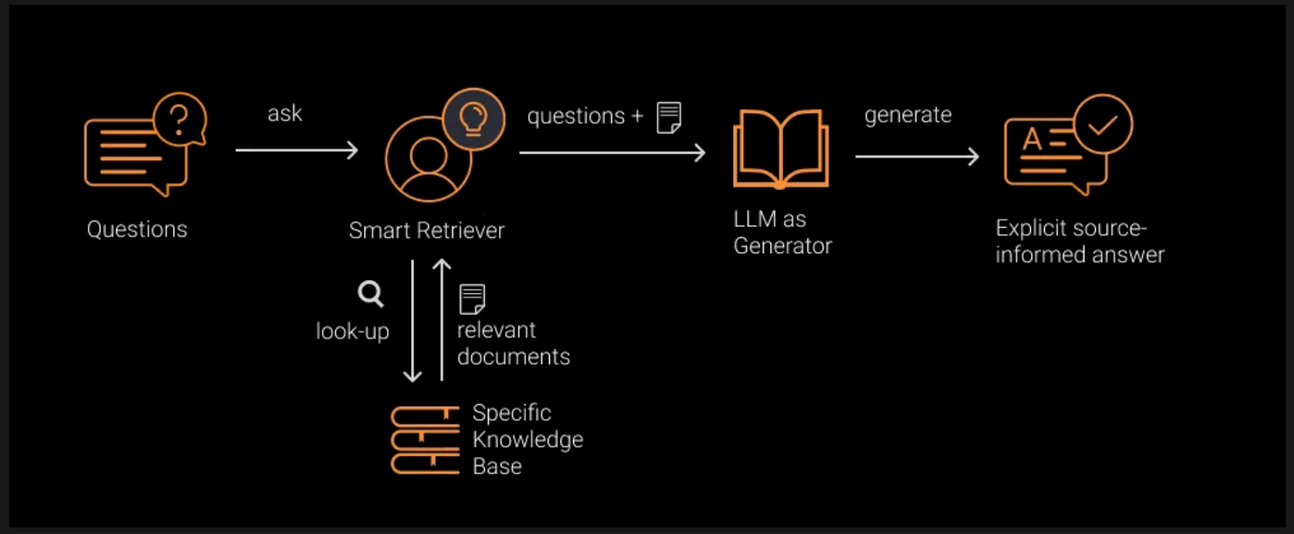


How Large Language Models are trained for your case

Retrieval-Augmented Generation

Retrieval-Augmented Generation (RAG) is an approach that combines an AI model with a “smart search” tool. When you ask a question, this tool searches your knowledge base for relevant information, which the AI then uses to give an informed, accurate response. This setup ensures the AI’s answers are based on specific, trusted data from your documents, rather than just general knowledge.

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.



After augmenting with additional context:

Who won Google Cloud Technology Partner of the Year (2023) Award for AI & Machine Learning?

I'm sorry, but I do not have access to real-time information or events that have occurred after my knowledge cutoff date in September 2021. To find the winner of the Google Cloud Technology Partner of the Year Award for AI & Machine Learning in 2023, I recommend visiting the official Google Cloud blog, their press releases, or checking reputable news sources for the latest updates and announcements.

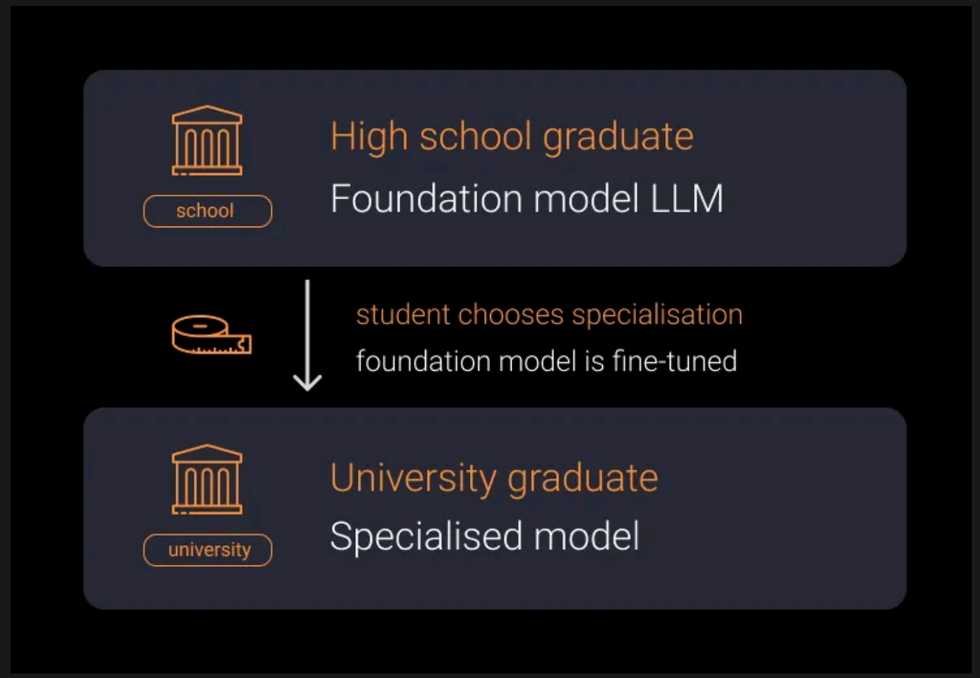
"We're thrilled to announce that Weights and Biases won Google Cloud Technology Partner of the Year(2023) Award for AI & Machine Learning."

Who won Google Cloud Technology Partner of the Year (2023) Award for AI & Machine Learning?

Weights and Biases won the Google Cloud Technology Partner of the Year Award for AI & Machine Learning in 2023.

How Large Language Models are trained for your case

- If prompting (providing examples in real-time) or a Retrieval-Augmented Generation (RAG) approach doesn’t achieve the desired results, additional fine-tuning may be needed.
- Fine-tuning adjusts the LLM’s neural network by updating its internal settings (weights) to improve performance on specific tasks.
- Imagine the LLM is like a high school student. With fine-tuning, you’re not just giving the student examples to solve specific problems on the spot (like in prompting); instead, you’re teaching them new skills and knowledge through structured practice.
- You provide the LLM with specific input tasks and correct answers (from your fine-tuning dataset). This feedback helps the LLM “learn” and adapt, allowing it to specialize in the areas covered by your dataset.
- For many business needs, simply using prompting might be enough to get the model to respond appropriately, such as using conversational responses based on a knowledge base. But for tasks requiring very specific behaviors or knowledge, fine-tuning may be essential to improve performance.



Approaches to Training LLMs for Your Specific Use