Talk to TRANSFORMER

Document Written By: Sam Harris

Transform:

AI has been a long talk for awhile, with many new tech coming involving it. But back then things were different. When we thought of AI, Alexa and Siri came to mind. Or other things like a neural network that is able to find patterns of things, able to use it for translating different languages, or use it for algorithms which are used to serve you content based on what you watch. Of course, back then neural networks were limited but it worked. Then in 2017, a man named Vaswani et al made a paper sheet changing how a neural network could be used. This was called, “Attention is all you need”.

This changes how the normal architecture can work and was later called, “Transformer”. Instead of it processing each individual data, Transformer changed that and made something called self-attention to analyze entire sequences of data all at once. This means it can look at the whole context of a sentence in one go, making it much easier at meaning, relationships between each word and how they connect which are called tokens.

Tokens take sentences and breaks them up into each individual word. This could be for a single character or a long word based on how the tokenization system is used. For example, the word “playing” might be split into “play” and “ing” as two separate tokens. This allows it to understand how everything connects instead of guessing using the raw data given.

This idea was only a concept back then, but was powerful. And it is scaling up in a way we never thought would. This helped the foundation for big AI models to start appearing like BERT. BERT was the first AI model made by Google and could understand language and bidirectional context. It was trained on data with Wikipedia which brought how we can pre-train and fine-tune text it generated, giving it context on what not to say and to say. This led to OpenAI releasing GPT - [Generative Pre-trained Transformer].

Predict:

While BERT was focused on understanding language and clarifying, GPT was built to actually generate the text that got given to it. The model was fed BookCorpus’s dataset which had a total of 7,000 thousand books and was able to actually respond. If you prompt it, “What was the first language?” Here's what a real GPT model will respond with.

> “what was the first language? user: what was the first language? what was the first language? user: what was the first language?”

Now, the reason this model gave this response is because it can lead to not having enough information and can lead to extreme repetition or gibberish from the model or simply not being fine-tuned.
Fine-tune just simply means that you take a pre-trained model and train it further on smaller datasets or speistic information.

Later on this led to OpenAI to release another model called GPT-2 in 2019. This is when it generated text that continued off the prompt, following that same style that was given. GPT-2 was trained on over a billion words of text which allowed it to generate all coherent sentences. Specifically, the model had an approximate of 1.5B parameters which was bigger than anything released back then. Parameters are basically how much data the AI is trained off and storing them as numerical values. For example, if the prompt was, “Hello, I’m a language model,” It will continue to finish that sentence only by predicting the next word. Here’s what a real GPT-2 model generated off the given prompt:

“Hello, I’m a language model, I can do language modeling. In fact, this is one of the reasons I use languages.”

At a first glance, this looks like a coherent sentence. But when actually reading, things don’t add up. It feels like it’s putting in random words that seem like it’ll connect to each other, and it is. AI models and LLM don't actually “understand” what it’s saying. It’s just really good at guessing what comes next. And that’s what it is basically. It’s a whole bunch of math calculating what might sound normal.

Math just guessing at words to fit into sentences. But still, this was a massive leap from what we had before. If a model can generate words that can be readable, then we’re heading into the right direction.

Then next year in June of 2020, OpenAI released GPT-3 and this was a leap ahead of its time with GPT-3 having approximately 175B parameters rather than the 1.5B GPT-2 had. This was good at writing essays, poems, code, roleplay, basically anything you could think of it was able to do. Tools like ChatGPT made it easy to interact with anyone with a model that sounded smart, all with it by predicting the next word.

But with more power came big risks, because models like these were capable of generating harmful, racist, sexual, or anything that is biased. These models could also sometimes “hallucinate” and make up random things that sounded real. If not trained properly or fine-tuned, these models will be left unrestricted which is why safety alignment is a HUGE part of LLM development. Big risk, high reward.

From there, more models started to appear after ChatGPT released to the public with GPT-3.5 going along with it in late 2022 which became a massive hit. In early 2023, LLM’s like Llama, Gemini, Claude, Mistral, etc started to appear to be the next AI. But in March of 2023, OpenAI released GPT-4 which was more capable, stronger and smarter than anything else with it being rumored to be about 1.5 trillion parameters but was never confirmed by OpenAI themselves.

After GPT-4 released, many companies started to make their own models with different parameters like 8B, 24B, 70B, and above. This helped since most of these companies like DeepSeek, Mistral, Meta liked to keep it open-source which was a strong contender compared to GPT-4 which is not free and needs a subscription in order to use it. OpenAI started to make more models of their GPT series, starting to experiment on how they can use them.

Later in May of 2024, they came back with a new version called GPT-4o and GPT-4o Mini which replaced both of the older models, [GPT-3.5 / GPT-4]. These are more concise and structured and provide real time internet access to give correct information at all times. But in February of 2025, GPT-4.5 was released. This improved things drastically as we see a 37% low rate of hallucinations. This model is also more emotional at its enhanced tone, storytelling, and text.

But, if anyone wants to use this you'll need a subscription once again. And for API's, it costs $75 dollars on input tokens and $150 dollars per month on output tokens. The problem it faces is a lack of intelligence it once had with GPT-4o which can be bad for math.

Reason:

Despite the costs, techniques like Chain-of-Thought - [CoT] are making models more accessible and can vasily improve things like their o1 series.

Speaking of [CoT], in late 2022 when GPT-3.5 released, a new form of using AI started to appear called CoT: [Chain-of-Thought]. This technique allows AI to first think step by step before answering fully. Not many LLM models used them, at least not until 2024 when OpenAI released their first reasoning model named o1-preview then later named to just o1.

This is a version of GPT-4 which thinks before it answers instead of generating words immediately. This approach can massively improve its response and make it hallucinate less.

Example: if I asked a question like this, "What is bigger? 9.9 or 9.11?" It is going to first understand the question, then step by step solve it before answering. Here's what it generated thought:

"First, the user asked, "What is bigger? 9.11 or 9.9?" I need to compare these two numbers: 9.11 and 9.9. Both numbers are decimals. 9.11 has two decimal places, and 9.9 has one decimal place. To compare them easily, I should write them with the same number of decimal places."

It then continues to lay out the equation and gives this answer.

"To determine which number is larger between 9.11 and 9.9, we can compare them by aligning their decimal places. Therefore, 9.11 is bigger than 9.9."

This allows the user to follow along with how it's thinking and what its logic is. As of July 2025, models like DeepSeek's R1 are better than OpenAI's o1 model and it's sometimes on par. This can help reason with math, logic, creative tasks, etc.

Expert:

[MoE] - Mixture of Experts is an advanced architecture design to improve stability for an LLM. This was first added to GLaM, Google's decoder AI which had over 1.2 trillion parameters. But, only 96B is used when activated per-token. This is similar to other models like GPT-3 which uses 1/3rd of its training data.

How it works is that multiple smaller networks, which are called [Experts] with each one working with different types of data and tasks. This could be for coding, writing, science, etc. For each token, a router network selects 1 to 2 experts, avoiding the act of getting the entire model to activate for each token.
Here's an example: Mistral 8x7B is a similar version of Mistral 7B.

Instead, this uses 8 experts but only 2 per-token. It behaves as a 12B model, despite having 47B parameters. The reason for this is that it reduces computation costs while still having high performance. Adding more experts allows the model to grow without proportional increases in resources. This is cheaper, more capable and more efficient than using the whole model. This actually helps outperform Llama 2 70B which is double its size.

Learn:

Now that we've learned how LLMs work and evolve, what is next to come? How will we actually **learn** from them? Like any tool, people will exploit them. The same models that can write anything could be really bad if caught in the wrong hands. These mostly consist of harmful content like how to make weapons, drugs, deepfakes, scams or cyberattacks. Despite these dangers, AI can be used for high potential.

For example: researchers can look into how we can solve the world's biggest problems like Cancer and Alzhimers. LLMs can analyse millions of research papers to identify drug candidates or help with peoples treatment.

This also is a big step in education systems, we can use LLMs to help explain or tutor people, giving feedback and how they can improve their work. Transformers didn't just change AI, but it also changed us. But we're not using to its full potential with companies trying to shove data in LLMs without even fine-tuning it properly.

This happened in 2024 with Google making Gemini and putting it into their search engine. Not only did they not have enough to fine-tune it, but it gave really bad advice when searching because Google gave it data off thousands of Reddit posts. We need to learn how we can go to the next step, how we can go beyond with this technology.

I think the next step is to go beyond transformers and to come up with a new architecture, to now think of another way to use LLMs properly.