

Quick Essay: Large Language Models, How to Train Them, and xAI's Grok

How do LLMs work? How are they trained? How will new entrants like xAI compete?

CHAMATH PALIHAPITIYA
JAN 30, 2024



[Read our AI Deep Dive](#)

When OpenAI released ChatGPT in November 2022, it took the world by storm, reaching over a million users in only 5 days. This kind of viral attention was previously unheard of in AI, driven by how closely the underlying language model seemed to replicate human intelligence.

Since then, there has been an explosion in AI activity, ranging from applications built on top of ChatGPT which seek to improve efficiency for mundane tasks, to new chatbots like xAI's Grok which aim to replace ChatGPT altogether.

This explosion happened so quickly that few of us really took a step back to understand the basics. So we wanted to sit down and understand how LLMs work to figure out how new entrants like xAI will compete.

So what is an LLM?

A large language model is a type of neural network that can ingest strings of text and then predict the next sequence of words. Intelligent chatbots like ChatGPT are specialized versions of these language models that have been trained for the specific purpose of generating responses to questions.

To understand and generate text like humans, there are a few things that language models must be able to do:

1. Understand the meanings of various words
2. Understand the context of words in relation to other words
3. Remember long strings of these words
4. Do all of the above very quickly

Until recently, even the best-in-class language models struggled to do all four. They were either slow and inefficient to train, had poor memory, or were bad at recognizing context. This resulted in models that failed to effectively replicate human abilities.

In 2017, a new type of architecture called a “transformer” was introduced that promised to solve many of these issues. Two key breakthroughs, “positional encoding” and “self-attention”, made this architecture much more efficient to train and better at recognizing context. Modern language models were trained with more compute power and larger datasets, and new capabilities emerged. Today, models can reason about topics and relationships using information across multiple modalities including images and audio.

But how do LLMs work?

LLMs work by first taking a string of words and representing them as numerical vectors. Each number within the vector captures the meaning of the word. Think of this like a graph. When two words are close together, they have similar meanings.




The position of each word in the sentence is also represented as a vector, allowing the model to capture context without needing to process each word serially - a key development that made transformers much more efficient than previous models.

The “self attention” layer, which is what transformer models are known for, then allows the model to hone in on relevant words to further improve contextual awareness. Take the following sentence:

“Yesterday, I went to the bank to deposit money.”

The word “money” allows the model to understand that the sentence refers to a money bank, not a river bank.

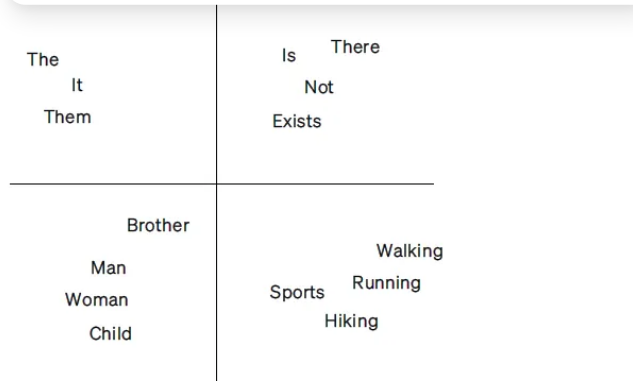


Discover more from Chamath Palihapitiya
 I make bets on disruptive ideas, technology, and people
 Subscribe to learn with me.
 Over 79,000 subscribers

Subscribe

[Continue reading >](#)

[Sign in](#)



“Yesterday, I went to the bank to deposit money”



Money Bank



Riverbank



So how do you build an LLM?

Large language models like ChatGPT and Grok are built in two key stages:

1. The training stage, which feeds the model billions (and often trillions) of words so that the model can learn what different words mean and how closely they are related, with the goal of eventually generating text by predicting the next word.
2. The fine-tuning stage, which trains this pre-trained model to perform a particular kind of task like answering questions.

Stage 1: Training the model

To train a language model to generate text, you first need to collect a massive amount of data on which to teach the model to predict the next word. This is achieved by scraping the internet for text data from a diverse range of sources, and then cleaning this up to remove duplicates, spelling errors and issues that you don't want the model to learn.

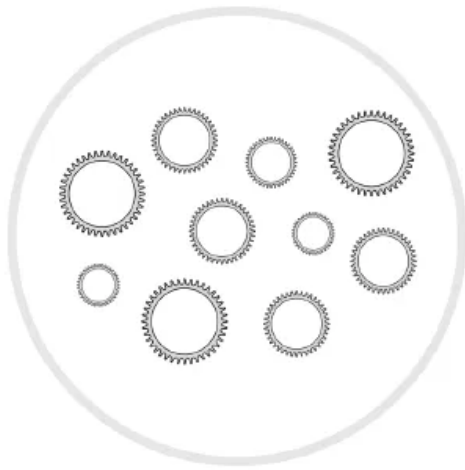
Once the training dataset is assembled, it is then turned into a series of incomplete sentences that are used to train the model to predict the next word.

“The man is walking”

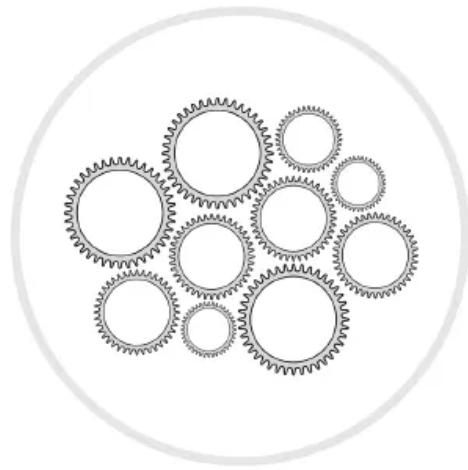


Token 1	Token 2	Token 3	Token 4
The	Man	Is	?

Language models are types of neural networks that use layers of nodes to generate their predictions. Nodes are like gears in a machine. Individually, they lack meaning, but when trained to work together, nodes can understand and interpret complex data like language.



Random weights
and biases



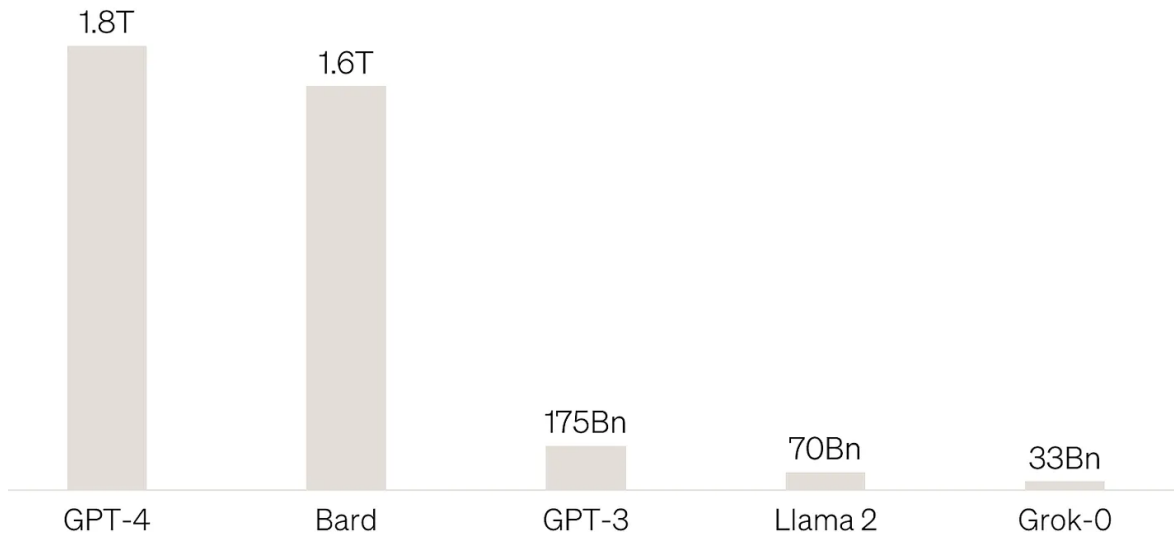
Trained
neural network

Initially, the connections between nodes will be assembled randomly, so the model's prediction will also be random. But as the model is trained, the nodes learn to predict the output that we want to see by adjusting the weights and biases that connect them together.



The number of weights and biases that a model uses to make a prediction is called its "parameters". The more parameters there are, the more complex the model. While this often leads to better performance, it also comes at the cost of higher latency and computational demand. Newer language models like Grok aim to outperform larger models using fewer parameters by improving on the architecture of models and leveraging higher quality training data.

Parameter Count of Leading LLMs

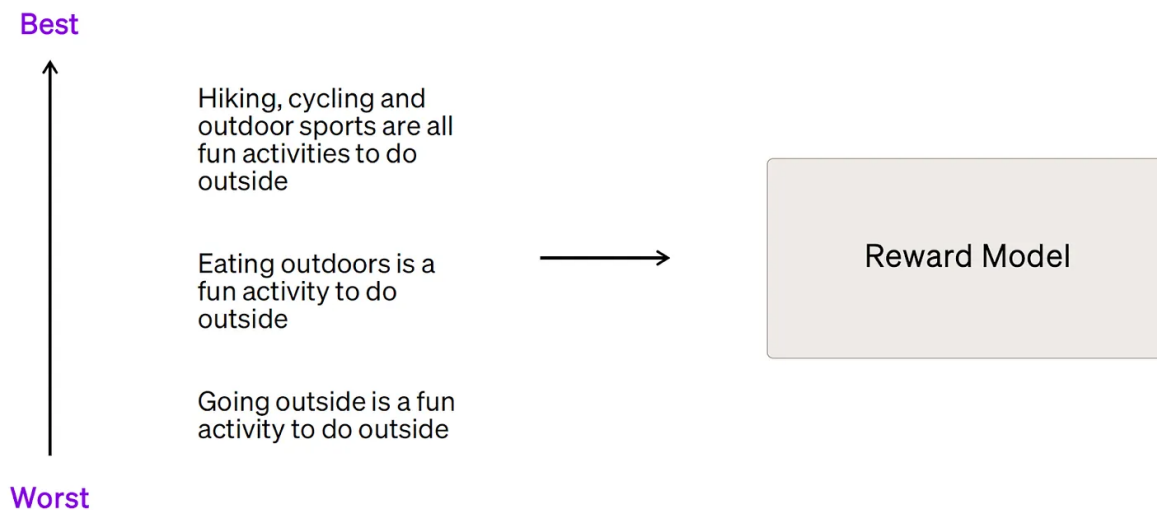


During the training process, the language model learns to map the relationships between words and predict the next word in a sequence. But it still needs to learn how to perform specific tasks like responding to questions. This is the role of fine-tuning.

Stage 2: Fine-tuning the model

To build a chatbot that can respond to questions, pre-trained models are trained on thousands of examples of prompts in the desired question-answer format until the model can predict an appropriate response to a given question.

Then, once the model can predict answers in the desired format, human feedback is used to rank several of the model's possible responses from best to worst in a process called "reinforcement learning from human feedback" (RLHF). This feedback is used to train a second "reward model" to guide the LLM to predict the best response.



Companies building new language models today face two major challenges. The first is that exponential increases in the amount of data used to train a new model only result in linear improvements in performance. So with an abundance of data available for training, all else equal, models eventually converge towards a single level of performance.

The second is a lack of context. Many models like ChatGPT lack context beyond their training period, meaning that they have no awareness of information and events beyond a given date. When asked about information after this period, they either refuse to answer, or worse still, hallucinate and provide a convincing but made-up response.

So, if you want to build a chatbot that is constantly improving and context-aware, how do you do it?

A new entrant to the race: xAI and Grok

xAI launched its new chatbot Grok on November 4th, 2023, just four months after the company was officially announced. Grok's initial model, Grok-0, demonstrated impressive performance with limited resources, directly competing with Meta's LLaMA 2 model using half the complexity (70 billion vs. 33 billion parameters). Its next iteration, Grok-1, showcased even better results, surpassing all other models in its compute class including GPT-3.5, which took OpenAI several years to achieve.

What makes xAI's model unique is its access to a proprietary and constantly-evolving dataset of tweet activity, which generates over 12 terabytes of data daily, containing extensive data on human interactions and current events in multiple formats (text, images and even audio), and distributing its model to an existing user-base of more than 500mm MAUs on X.

Access to this dataset of constantly updated information can help to minimize hallucinations and provide more context-aware responses when presented with questions about recent events. xAI can quickly retrieve information from reputable sources on X, and use the wisdom of crowds to

interpret the sentiment around a given topic, allowing the model to provide more context-aware responses to queries.

Having access to this data in multiple formats such as images and audio can also help xAI's model achieve a deeper and more nuanced understanding of the world. For example, understanding a person's facial expressions while they are speaking results in a much richer interpretation of their speech than just an audio recording. In the same way, leveraging multi-modal inputs on X can help xAI's model to better understand the context of news and other world events.

The final differentiator is distribution. xAI already has built-in distribution through the X platform, which has more than 500mm monthly active users. Assuming modest uptake, this allows xAI to rapidly improve its models through much faster reinforcement learning from human feedback loops than other models, providing the company with another set of proprietary data that can help propel its model further than competitors.

Conclusion:

As the foundational layer of language models is becoming increasingly difficult to improve with more data, the quality of the data that these models are trained on becomes a key differentiator. xAI's Grok benefits from a vast dataset of diverse and up-to-date information in multiple formats, as well as a pre-existing user base of 500mm people to rapidly improve its models. With high quality real-time data and the capital to scale, Grok has the opportunity to become the most up-to-date, customizable and context-aware language model in the race to achieve AGI.

[Read our AI Deep Dive](#)

Disclaimer: The views and opinions expressed above are current as of the date of this document and are subject to change without notice. Materials referenced above will be provided for educational purposes only. None of the above will include investment advice, a recommendation or an offer to sell, or a solicitation of an offer to buy, any securities or investment products.



100 Likes · 3 Restacks

9 Comments



Write a comment...



Dan Sangyoon · Sangyoon's Newsletter · 15 hrs ago

Thank you for sharing Chamath. Good read. I enjoyed most of it, but did feel it was biased in favor of grok without sufficient backing.

For example:

"Having access to this data in multiple formats such as images and audio can also help xAI's model achieve a deeper and more nuanced understanding of the world."

Doesn't openAI, bard and others also have access to image and audio data?

Also you say grok has distribution advantage, but how is that any better than the distribution google may have through all their products or Meta with billions of users talking through its platforms?

I see grok as "different" and valuable, but the logic you provided hasn't convinced me it has a true advantage VS others.

 LIKE (2)  REPLY  SHARE

...



Quratulann Akbar 11 hrs ago

Really well written! Enjoyed reading it :)

 LIKE (1)  REPLY  SHARE

...

7 more comments...