PAPER NAME

Plagerism Check Material Harshit Singh Thesis - LLM Review.pdf

AUTHOR

Harshit singh AI

WORD COUNT

8111 Words

CHARACTER COUNT

44238 Characters

PAGE COUNT

35 Pages

FILE SIZE

559.0KB

SUBMISSION DATE

May 25, 2023 2:13 PM GMT+5:30

REPORT DATE

May 25, 2023 2:14 PM GMT+5:30

6% Overall Similarity

The combined total of all matches, including overlapping sources, for each database.

- 2% Internet database
- 0% Publications database
- Crossref Posted Content database
- 5% Submitted Works database

Excluded from Similarity Report

- Crossref database
- Bibliographic material
- Cited material
- Small Matches (Less then 10 words)

INTRODUCTION

Language is a communication system developed by humans for self expression and communicating with others, everything from the most groundbreaking theories in science to a cheeky joke using symbols, sounds and gestures. Language is the cornerstone of human innovation enabling us to pass information over generations, but human language can not be understood by machines. Language technology is an interdisciplinary approach towards making machines understand and respond to human language using linguistics, artificial intelligence etc, that aims to enhance human communication. Recently, generative pre-trained transformer models trained on large amounts of data and having billions of parameters have shown unprecedented advances in language technology and the research community aptly called these models as large language models(LLMs)[1].

ChatGPT is a chatbot based on LLM released by OpenAI that has shown excellent ability to generate coherent and fluid reply for a language prompts, ushering in a new paradigm of chatbots from Google's Bard to Microsoft's Copilot and Bing Chat (Based on GPT-4). This advancement was possible as LLMs display emergent abilities[2][10][11] like summarization, answering questions etc. for which these LLMs were not specifically trained, making these models extremely effective. Simultaneously, the developing approaches like prompting interface has exponentially eased the process of interacting and formatting commands for LLMs to follow.

But this advancement also comes with a few drawbacks. Firstly, we don't fundamentally understand how or why emergent abilities occur in LLMs. Secondly, as these models can generate harmful text either due to toxic prompts or the model hallucinating, moderation becomes a fundamental issue. Thirdly, investigations have shown that current moderation is done by exploiting human rights in developing countries, where humans were subjected to all forms of toxicity that was to be scrubbed from training data. Lastly, as development of these models requires huge compute, it restricts the ability of researchers to study various strategies to train LLMs and restricts various details like data collection, data cleaning etc. from public and behind the door of huge profit driven corporations.

LITERATURE SURVEY

Release of ChatGPT by OpenAI in November 2022 proved to be a double edged sword for the research community as this technology proved to be a paradigm shift in natural language processing tasks, there by opening further avenues for research from environmental concerns regarding the technology to social and political concerns of mass deployment of the technology but also shelved a lot of ongoing research in the domain of natural language processing as this technology proved better suited for tasks like lexical simplification, text summarization, text generation etc. which were individual research areas for many researchers.

The data collection and training of most of these models is done by a handful of companies in the industry, the information about the training data and training is not publicly available. Although now there are thousands of papers in various stages of publication on these large language models. Since the scope and implication of the technology is vast, I had to narrow down my survey to the most prominent issues regarding the technology and a basic understanding of the technology.

OVERVIEW

This section will provide an overview of the technologies that lead to the development of large language models, followed by evolution of large language models over the last couple of years.

3.1 History

The history of language modelling can be divided into four prominent stages of development.

3.1.1 Statistical language models (SLMs)

Statistical language models[7] are a type of artificial intelligence model that is used to predict the next word in a series of words. Statistical language models can be trained on large corpora of text, and they can be utilised to carry out a range of natural language processing (NLP) jobs, such as text classification, sentiment analysis, and device translation.

There are 2 primary types of analytical language models: n-gram designs and neural network models. N-gram models are based upon the n-gram language design, which is a statistical model that anticipates the next word in a sequence of words based on the previous n words. Neural network models are based on neural networks, which are a type of artificial intelligence model that can discover complicated patterns from information[8].

Statistical language models are a powerful tool that can be used to carry out a variety of NLP jobs. Nevertheless, they likewise have some limitations. One constraint is that they can be computationally pricey to train. Another constraint is that they can be biassed, which implies that they might not have the ability to precisely forecast the next word in a sequence of words if the data that they are trained on is biassed[9].

3.1.2 Neural language models (NLMs)

Neural language models (NLMs) [7] are a kind of statistical language model that uses neural networks to forecast the next word in a series of words. NLMs are trained on big corpora of text, and they have been revealed to be extremely efficient at a range of natural language processing (NLP) tasks, such as text categorization, belief analysis, and machine translation.

NLMs work by learning the statistical relationships between words in a language. This is done by training a neural network on a big corpus of text. The neural network finds out to anticipate the next word in a sequence of words based on the previous words in the sequence.

NLMs have a number of benefits over standard statistical language models, such as n-gram models. NLMs are able to discover more intricate patterns in language, and they are less likely to be biassed by the data that they are trained on [7].

NLMs have been used to attain modern results on a variety of NLP jobs. For instance, NLMs have been used to achieve modern results on text category tasks, such as spam filtering and sentiment analysis. NLMs have actually also been utilised to attain cutting edge outcomes on translation tasks.

NLMs are an effective tool that can be used to carry out a range of NLP tasks. Nevertheless, they likewise have some constraints. One constraint is that they can be computationally expensive to train. Another limitation is that they can be biassed, which indicates that they might not have the ability to accurately forecast the next word in a sequence of words if the information that they are trained on is biassed.

Regardless of these constraints, NLMs are a promising technology that has the possibility to revolutionise the way we connect with computers. They might be utilised to produce a

more natural and engaging interface, to enhance the accuracy of maker translation, and to create creative content.

3.1.3 Pre-trained language models (PLMs)

Pre-trained language models (PLMs)[5] are a type of neural language model that is trained on a big corpus of text. This corpus is normally a huge dataset of text and code, such as the internet or a big library. The model is then fine-tuned on a smaller dataset that is relevant to the specific task that the model is being used for[12].

PLMs have been revealed to be extremely reliable at a variety of natural language processing (NLP) jobs, such as text classification, sentiment analysis, and machine translation. They have also been utilised to attain advanced outcomes on a variety of other jobs, such as concern answering and imaginative writing.

PLMs work by discovering the statistical relationships in between words in a language. This is done by training a neural network on a big corpus of text. The neural network finds out to anticipate the next word in a series of words based on the previous words in the series.

PLMs have numerous advantages over standard statistical language models, such as n-gram models. PLMs are able to learn more complicated patterns in language, and they are less likely to be biassed by the information that they are trained on[10].

However, they likewise have some constraints. One constraint is that they can be computationally expensive to train. Another constraint is that they can be biassed, which indicates that they may not have the ability to accurately predict the next word in a series of words if the information that they are trained on is biassed.

Regardless of these constraints, PLMs are a promising innovation that has the prospective to change the way we communicate with computers. They could be utilised to produce a

more natural and interesting interface, to improve the accuracy of machine translation, and to produce imaginative material.

GPT-2, precursor to ChatGPT and GPT-4, was a PLM created by OpenAI, likewise Bart was a PLM created by Google which was a precursor to Bard, PaLM, PaLM2, Lambda and Gemini.

3.1.4 Large language models (LLMs)

LLMs are trained on massive amounts of text information, generally with billions of parameters. This permits them to discover the statistical relationships in between words and phrases, and to produce text that is both coherent and grammatically appropriate.

LLMs have actually been used to accomplish cutting edge results on a variety of natural language processing jobs, including machine translation, text summarization, and question answering. They are also being used to develop brand-new applications, such as chatbots, virtual assistants, and innovative writing tools.

LLMs are still under development, but they have the prospective to change the way we interact with computer systems. By comprehending and creating natural language, LLMs can make computer systems more accessible and easy to use. They can also help us to better understand the world around us, and to create brand-new forms of art and literature[1].

Here are a few of the advantages of using large language designs:

- They can create text that is both coherent and grammatically proper.
- They can be used to translate languages.
- They can be used to compose different kinds of innovative content.
- They can address your questions in a helpful method.

Here are some of the difficulties of using large language models:

- They can be costly to train.
- They can be computationally pricey to utilise.
- They can be biassed, showing the biases in the data they are trained on.

The primary distinction between pre-trained language designs and big language designs is the size of the design. Pretrained language models are usually smaller sized, with a couple of hundred million specifications. Big language designs are much larger, with billions and even trillions of criteria. The bigger size permits these designs to display emergent abilities i.e. showing abilities that these designs were not explicitly trained for.

Emergent abilities are abilities that large language models (LLMs) display that are not present in smaller language models. They are frequently unforeseeable and can be unexpected. Some examples of emergent abilities include:

- The ability to carry out maths operations
- The ability to address questions about the world
- The ability to create innovative text formats, such as poems, code, scripts, musical pieces, e-mail, letters, and so on.
- The ability to translate languages
- The ability to compose different sort of imaginative material

These abilities are not clearly set into LLMs. Rather, they emerge from the manner in which LLMs are trained. LLMs are trained on enormous quantities of text information, and they discover patterns in the data. These patterns enable them to carry out jobs that would be tough or difficult for smaller language models.

Emergent abilities are a promising indication of the potential of LLMs. They recommend that LLMs can discover and understand language in a manner that is similar to people. This has the possibility of changing the way we interact with computers.

3.2 Background

As discussed above, LLMs are transformer models with billions of parameters pre-trained on large amounts of data. For a better understanding of LLMs we need to through some basic background terminologies.

3.2.1 Scaling

Scaling refers to the capability of these models to improve their performance as they are trained on larger and larger datasets. This is due to the fact that LLMs try to detect patterns in the data, and the more information they are trained on, the more patterns they can learn.

There are 2 primary ways in which scaling can be accomplished for LLMs[1]:

- Increasing model parameters/size
- Increasing the training data size

Scaling has actually been shown to be a very reliable method to enhance the performance of LLMs.

For instance, the GPT-3 language model was trained on a dataset of 500 billion words, and it was able to exceed previous language models on a range of jobs.

Nevertheless, scaling likewise has some difficulties. One obstacle is that it can be costly to train large language models. Another challenge is that large language models can be computationally pricey to utilise.

Regardless of these difficulties, scaling is a promising area of research for LLMs. It has the potential to cause the advancement of language models that can attain human-level performance on a range of jobs.

The literature consists of two prominent scaling laws for large language models[4][5].

3.2.1.1 KM Scaling Law

The KM scaling law is a mathematical relationship that describes the relationship between the size of a language model and its performance on a variety of tasks. The mathematical expression of the same is given below, where M is model size, D is dataset size, C is compute and L denotes the entropy loss.

$$
\begin{aligned}
L(N) &= \left(\frac{N_c}{N}\right)^{\alpha_N}, \quad \alpha_N \sim 0.076, N_c \sim 8.8 \times 10^{13} \\
L(D) &= \left(\frac{D_c}{D}\right)^{\alpha_D}, \quad \alpha_D \sim 0.095, D_c \sim 5.4 \times 10^{13} \\
L(C) &= \left(\frac{C_c}{C}\right)^{\alpha_C}, \quad \alpha_C \sim 0.050, C_c \sim 3.1 \times 10^{8}
\end{aligned}
$$

The KM scaling law was first introduced by OpenAI in 2020. They found that the performance of language models on a variety of tasks, including machine translation and question answering, increased as their size increased. However, the rate of improvement decreased as the size of the language model increased.

3.2.1.2 Chinchilla scaling law

Chinchilla scaling is a particular scaling law that relates the parameters of a family of neural networks. It states that for a large language model (LLM) autoregressive trained for one epoch, with a cosine learning rate schedule, there is a relationship between the cost of training the model (in FLOPs), the number of parameters in the model, the number of tokens in the training set, and the average negative log-likelihood loss per token achieved by the trained LLM on the test dataset. The mathematical expression of the same is given below where C is computer, N is number of parameters, D is the number of tokens in training data and L denotes the negative log-likelihood loss per token.

$$
\begin{aligned}
C &= C_0 N D \\
L &= \frac{A}{N^{\alpha}} + \frac{B}{D^{\beta}} + L_0
\end{aligned}
$$

The values of statistical parameters are A=406.4, B=410.7, L_0=1.69, α=0.34 and β=0.28.

.

3.2.2 Emergent Abilities of LLMs

Emergent abilities[10] of large language models refer to the unforeseen and often impressive capabilities that arise as a result of their size, training methodologies, and architectural advancements. These abilities go beyond their primary purpose of generating coherent and contextually relevant text and extend to various language-related tasks and domains.

1. In-context learning[14] : In-context learning in LLMs is a phenomenon where large language models (LLMs) can learn to perform a new task at inference time by being fed a prompt with examples of that task, without changing any parameters. It is based on the idea of learning from analogy and exploiting the general knowledge and skills encoded in the LLMs.

2. Step-by-step reasoning and instruction following[14] : step-by-step reasoning and instruction following are two important capabilities that large language models (LLMs) are increasingly being able to perform. LLMs can use these capabilities to complete a variety of tasks, such as following instructions to solve a puzzle, writing a story, or translating a text from one language to another.

3. Contextual Understanding[10]: Large language models exhibit an enhanced understanding of context, allowing them to generate text that is more coherent and contextually appropriate. They can grasp subtle nuances, references, and even maintain coherent dialogues or conversations.

4. Language Generation[11]: Large language models have the ability to generate human-like text in various styles, tones, and genres. They can produce natural-sounding speech, write creative stories, compose poetry, and even imitate the writing style of specific authors or domains.

5. Information Retrieval and Summarization[12]: These models can effectively retrieve information from vast amounts of text and summarise it in a concise and coherent manner. They excel at distilling key points, identifying relevant information, and generating informative summaries.

6. Translation and Multilingual Capabilities[12]: Large language models demonstrate remarkable performance in machine translation, enabling accurate and fluent translations between different languages. They can handle complex sentence structures and idiomatic expressions, significantly reducing the language barrier.

7. Question Answering and Conversational AI[10]: Large language models have the ability to understand and respond to questions by retrieving relevant information or generating informative answers. They can engage in coherent and contextually relevant conversations, making them useful for conversational agents and chatbots.

8. Sentiment Analysis and Emotion Recognition[13]: These models can detect and analyse sentiment in text, identifying emotions such as happiness, sadness, anger, or surprise. They can discern the emotional tone of a piece of text, enabling applications in sentiment analysis and emotion recognition.

9. Creative Writing and Content Generation[11]: Large language models can generate original content, such as articles, essays, product descriptions, and marketing copy. They can help automate content creation and assist with tasks that require generating coherent and engaging text.

10. Domain-Specific Expertise[15]: Through transfer learning and fine-tuning, large language models can acquire domain-specific knowledge and exhibit expertise in various fields. They can assist with technical writing, legal documents, medical reports, and other specialised domains.

It is important to note that while these emergent abilities of large language models offer significant advantages, they also raise concerns related to ethics, bias, privacy, and the responsible use of AI-generated content. Ongoing research and development aim to address these challenges and ensure the responsible deployment of large language models in various applications.

3.2.3 Ability eliciting

Ability eliciting[15] in LLMs is a technique that aims to extract the hidden or latent abilities of large language models (LLMs) by designing suitable prompts, instructions, or examples for tasks. It can reveal the general knowledge and skills that LLMs have learned from pre-training on large-scale text data.

Here are some examples of how ability eliciting can be used to improve the performance of LLMs:

- To improve the accuracy of LLMs, ability eliciting can be used to provide them with prompts that require them to use their knowledge of the world. For example, an LLM could be given the prompt "What is the capital of France?" to elicit its knowledge of geography.
- To improve the fluency of LLMs, ability eliciting can be used to provide them with prompts that require them to generate creative text. For example, an LLM could be given the prompt "Write a poem about love" to elicit its ability to generate creative text.
- To expand the range of tasks that LLMs can perform, ability eliciting can be used to provide them with prompts that require them to use their abilities in new ways. For example, an LLM could be given the prompt "Write a song about a robot" to elicit its ability to generate creative text in a new format.

These are just a few examples of how ability eliciting can be used to improve the performance of LLMs. As LLMs continue to develop, ability eliciting is likely to become an increasingly important technique for improving their performance.

3.2.4 Alignment tuning

Alignment tuning[21] in LLMs is a technique that aims to align the behaviour of large language models (LLMs) with human values and preferences by tuning the model parameters or prompts using human feedback or data. It can improve the safety and usefulness of LLMs for various applications and tasks.

3.3 Transformer

The transformer architecture[38] is a neural network architecture responsible for adding parallelization for natural language processing (NLP) tasks. It was first introduced in the paper "Attention Is All You Need" by Vaswani et al. in 2017.

The transformer architecture is based on the attention mechanism[38], which is a way of learning the relationships between different parts of a sequence. The attention mechanism allows the transformer architecture to learn long-range dependencies in sequences, which is important for NLP tasks such as machine translation and question answering.

The transformer architecture consists of 2 parts: the encoder and the decoder. The encoder is responsible for processing the input sequence, and the decoder is responsible for generating the output sequence.

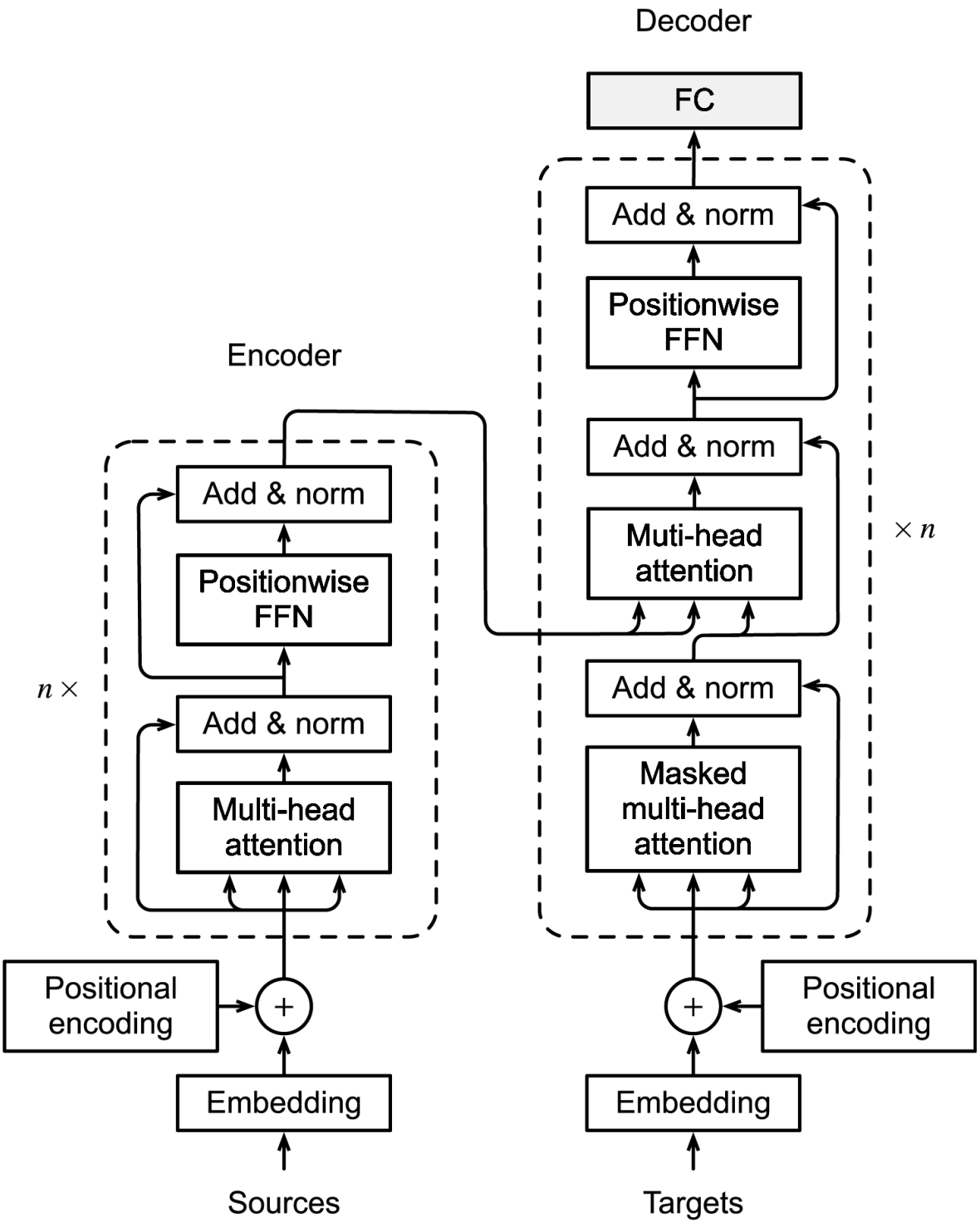

Figure 1 : Transformer Architecture

The encoder consists of a stack of self-attention layers. Each self-attention layer computes the weighted sum of the hidden states of the previous layer, where the weights are determined by the attention mechanism. The output of the self-attention layer is then fed into the next self-attention layer.

The decoder consists of a stack of self-attention layers and a linear layer. The self-attention layers work in the same way as the self-attention layers in the encoder. The linear layer is responsible for generating the output sequence.

The transformer architecture has been revealed to be really reliable for NLP tasks. It has actually accomplished cutting edge results on a range of tasks, including machine translation, question answering, and text summarization.

Here are some of the benefits of the transformer architecture:

- It can learn long-range reliances in sequences.
- It is really effective, needing less calculation than recurrent neural networks.
- It is scalable, indicating that it can be trained on big datasets.

Here are a few of the disadvantages of the transformer architecture:

- It can be challenging to train.
- It can be very reactive of the choice of hyperparameters.
- It can be computationally expensive to release.

3.4 Evolution of GPT series

GPT-1 was the first version of the Generative Pre-trained Transformer (GPT) model, a neural network model that can generate natural language by training on large amounts of text data. It was released by OpenAI in 2018 as a breakthrough in natural language processing.

GPT-1 had 117 million parameters and used the Transformer architecture, which relies on attention mechanisms to process sequential data. GPT-1 was pre-trained on a combination of two datasets: the Common Crawl, a massive dataset of web pages with billions of words, and the BookCorpus dataset, a collection of over 11,000 books on a variety of genres.

GPT-1 was able to generate fluent and coherent language when given a prompt or context, such as a sentence or a paragraph. It could also perform various natural language processing tasks, such as question answering, text summarization, and machine translation, without any task-specific fine-tuning.

However, GPT-1 also had some limitations and challenges. For example, it was prone to generating repetitive or irrelevant text, especially when given prompts outside the scope of its training data. It also failed to reason over multiple turns of dialogue and could not track long-term dependencies in text. Additionally, its cohesion and fluency were only limited to shorter text sequences, and longer passages would lack coherence.

GPT-2 was the second version of the Generative Pre-trained Transformer (GPT) model, a neural network model that can generate natural language by training on large amounts of text data. It was released by OpenAI in 2019 as a successor to GPT-1.

GPT-2 had 1.5 billion parameters, considerably larger than GPT-1. The model was trained on a much larger and more diverse dataset, combining Common Crawl and WebText. One of the strengths of GPT-2 was its ability to generate coherent and realistic sequences of text across various domains and genres.

GPT-2 also improved over GPT-1 in several aspects, such as:

- Generating longer and more consistent text passages, up to several paragraphs.

- Performing better on various natural language processing tasks, such as question answering, text summarization, and machine translation, without any task-specific fine-tuning.
- Demonstrating zero-shot learning and few-shot learning abilities, meaning that it could perform new tasks by being given a few examples or instructions in natural language.

GPT-3 was the third version of the Generative Pre-trained Transformer (GPT) model, a neural network model that can generate natural language by training on large amounts of text data. It was released by OpenAI in 2020 as a successor to GPT-2.

GPT-3 had 175 billion parameters, considerably larger than GPT-2. The model was trained on a much larger and more diverse dataset, combining Common Crawl, WebText2, Books1, Books2, and Wikipedia. One of the strengths of GPT-3 was its ability to generate coherent and realistic sequences of text across various domains and genres.

GPT-3 also improved over GPT-2 in several aspects, such as:

- Generating longer and more consistent text passages, up to several pages.
- Performing better on various natural language processing tasks, such as question answering, text summarization, and machine translation, with minimal or no task-specific fine-tuning.
- Demonstrating zero-shot learning and few-shot learning abilities, meaning that it could perform new tasks by being given a few examples or instructions in natural language.
- Showing cross-lingual transfer learning abilities, meaning that it could perform tasks in different languages by being given examples or instructions in one language.

GPT-4 is the fourth and latest version of the Generative Pre-trained Transformer (GPT) model, a neural network model that can generate natural language by training on large amounts of text data. It was released by OpenAI in 2023 as a successor to GPT-3.

GPT-4 has more than 1 trillion parameters, considerably larger than GPT-3. The model was trained on a much larger and more diverse dataset, combining Common Crawl, WebText2, Books1, Books2, Wikipedia, and ImageNet. One of the strengths of GPT-4 is its ability to generate coherent and realistic sequences of text and images across various domains and genres.

GPT-4 also improved over GPT-3 in several aspects, such as:

- Generating longer and more consistent text and image passages, up to several pages or megapixels.
- Performing better on various natural language processing and computer vision tasks, such as question answering, text summarization, machine translation, image captioning, and image generation, with minimal or no task-specific fine-tuning.
- Demonstrating zero-shot learning and few-shot learning abilities, meaning that it could perform new tasks by being given a few examples or instructions in natural language or images.
- Showing cross-modal transfer learning abilities, meaning that it could perform tasks across different modalities by being given examples or instructions in one modality.

However, GPT-4 also had some limitations and challenges. For example, it was still prone to generating factual errors, contradictions, or nonsensical text or images, especially when given vague or ambiguous prompts. It also lacked common sense and ethical reasoning, and could generate harmful or biased content. Additionally, its large size and computational cost made it difficult to deploy and scale.

Despite these limitations, GPT-4 was a significant advancement in natural language processing and computer vision and paved the way for larger and more powerful models based on the Transformer architecture.

Model	**Launch Date**	**Training Data**	**No. of Parameters**	**Max. Sequence Length**
GPT-1	June 2018	Common Crawl, BookCorpus	117 million	1024
GPT-2	February 2019	Common Crawl, BookCorpus, WebText	1.5 billion	2048
GPT-3	June 2020	Common Crawl, BookCorpus, Wikipedia, Books, Articles, and more	175 billion	4096
GPT-4	March 2023	Common Crawl, WebText2, Books1, Books2, Wikipedia, ImageNet and more	More than 1 trillion	Unknown

Table 1 : Basic comparison of GPT series

GPT series proved to be foundational in development and showcase of LLMs but other major companies were developing their own LLMs alongside OpenAI, the same is showcased in figure 2.

As discussed earlier, the majority of the research is done by big industry players that don't publish various aspects of data collection, training, cost etc. So, it is extremely tough to infer the parameters and performance of the various LLMs and lack of a centralised benchmark for these LLMs means that they can only be judged subjectively. But, one thing

is certain that given every company is working on an iteration of the technology means that this technology could prove to be a paradigm shift for computing.

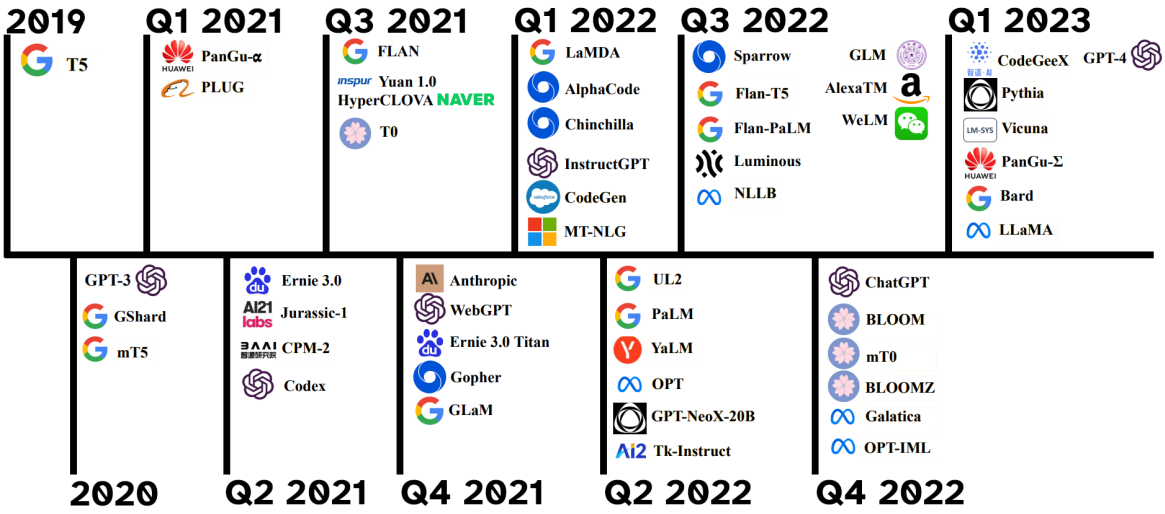

Figure 2 : Timeline of LLMs

PRE-TRAINING

Pretraining is a technique that involves training a large neural network on a massive corpus of text before fine-tuning it on a specific downstream task. The idea is that the network can learn general linguistic patterns and knowledge from the text, which can then be transferred to various natural language processing tasks such as question answering, text summarization, sentiment analysis, etc.

Pre-Training has several advantages over training from scratch. First, it can reduce the need for labelled data for each task, since the network already has some prior knowledge of the language. Second, it can improve the generalisation and robustness of the network, since it exposes it to diverse and noisy data sources. Third, it can enable zero-shot or few-shot learning, where the network can perform well on new tasks without any fine-tuning or with minimal examples.

However, pre-training also poses some challenges and limitations. First, it requires a lot of computational resources and data to train a large network effectively. Second, it may not capture task-specific or domain-specific knowledge that is not present in the pretraining corpus. Third, it may suffer from catastrophic forgetting or interference, where the network forgets or overwrites previous knowledge when learning new tasks.

4.1 Data Collection

Data collection is an essential step for training large language models (LLMs), which are neural networks that can learn from and generate natural language. Data collection includes event and processing a large corpus of text that covers a large range of subjects, domains, and genres. The quality and quantity of the data can affect the performance and generalisation of the LLMs.

Data collection for LLMs poses several challenges and trade-offs. Some of them are:

- Data size and diversity: LLMs typically require massive amounts of data to learn general linguistic patterns and knowledge. However, collecting and keeping such data can be costly and time-consuming. Moreover, the data should be diverse and representative of the natural language use cases and variations, such as various languages, dialects, styles, etc. Otherwise, the LLMs might experience bias or overfitting.
- Data quality and reliability:LLMs are sensitive to the quality and dependability of the data they are trained on. The data need to be free of errors, noise, duplicates, or disparities that might confuse or misinform the LLMs. The data ought to also be trustworthy and trustworthy, especially if it originates from online sources that may include false or deceptive information.
- Data privacy and ethics: LLMs may inadvertently memorise or leak sensitive or personal information from the data they are trained on, such as names, phone numbers, addresses, etc. This may pose privacy risks for the data owners or subjects, as well as ethical issues for the data collectors and users. Therefore, data collection for LLMs should respect the privacy and consent of the data sources, and apply appropriate techniques to anonymize or remove any sensitive information.
-

To address these challenges and trade-offs, researchers have proposed various methods and techniques to improve data collection for LLMs. Some examples are:

- Using different data sources and formats for LLMs, such as plain texts, knowledge graphs, images, videos, or multimodal data .
- Using different data selection and filtering methods for LLMs, such as sampling, clustering, deduplication, cleaning, or verification .
- Using different data augmentation and generation methods for LLMs, such as paraphrasing, translation, summarization, or synthesis .

But as various investigations have reported this preliminary data collection is done by the English speaking population of developing countries at dirt prices. Academics have argued that sieving through explicit, illegal and brutal content can cause major psychiatric

problems coupled with the fact that these people are paid 2$ an hour with no psychological help, while the people in the rich countries using these tools don't have to go through the same.

4.2 Data Preprocessing

Following data collection, data needs to be heavily processed from removing duplicates to privacy reduction in order to generate a better model. This stage can be broken into 4 major steps.

Step 1 - Filtering : Here low quality data is removed using various techniques from heuristics to neural networks. Depending on the type of model being developed, filtering can range from removal of other language to removal of hate speech for human alignment.

Step 2 - De-duplication : Redundant and repeating data is to be removed.

Step 3 - Privacy Reduction : Removal of names, addresses etc. from the dataset as these can be leaked by the model using prompt engineering.

Step 4 - Tokenization : Segmenting dataset into individual tokens to be used as input to the model.

4.3 Architecture

The architecture of LLMs is typically based on the transformer model. As transformers provide excellent parallelization for natural language processing tasks. Transformers work by learning to attend to different parts of the input text, and then using this information to generate the output text.

The architecture of an LLM typically consists of the following layers:

- Embedding layer: This layer converts the input text into a sequence of vectors. These vectors represent the meaning of the words in the text.
- Attention layer: The attention layer learns to attend to different parts of the input text. This allows the model to learn the relationships between words in the text.
- Feedforward layer: The feedforward layer is a standard neural network layer that learns to predict the next word in the sequence.
- Output layer: The output layer generates the output text.

The architecture of LLMs is constantly evolving. Researchers are currently developing new techniques to improve the performance and efficiency of such models. As a result, LLMs are becoming increasingly powerful and capable in performing a wider range of tasks like inputting multi modal inputs.

TRAINING

Model training and optimization are essential to large language models (LLMs), it is the process of adjusting the parameters of a neural network that can learn from and generate natural language, using a training objective and an optimization algorithm. Model training and optimization can be done using various strategies and techniques, such as parallelism, mixed-precision, regularisation, or distillation.

Some of the main aspects of model training and optimization with respect to LLMs are:

- Training objective: This defines what the LLM is trying to achieve, such as predicting the next word in a sentence or reconstructing a masked word in a sentence. The training objective can be measured by a loss function, which quantifies how well the LLM performs on the training data. The goal of model training and optimization is to minimise the loss function.
- Optimization algorithm: This defines how the LLM updates its parameters based on the training objective and the training data. The optimization algorithm can use various methods, such as gradient descent or Adam, to compute the gradients of the loss function with respect to the parameters, and apply learning rate schedules, momentum, or weight decay to update the parameters accordingly.
- Parallelism: This refers to the technique of distributing the computation of the LLM across multiple devices, such as GPUs or TPUs, to speed up the training process and enable larger models. Parallelism can be done at different levels, such as data parallelism, model parallelism, or pipeline parallelism, depending on how the data and the model are split and synchronised across devices.
- Mixed-precision: This refers to the technique of using different numerical precisions for different parts of the LLM, such as fp32, fp16, or bf16. Mixed-precision can reduce the memory usage and increase the throughput of the LLM, while maintaining or improving its accuracy. Mixed-precision can be applied to different components of the LLM, such as optimizers, weights, or specific modules.

- Regularisation: This refers to the technique of adding constraints or penalties to the LLM to prevent overfitting or improve generalisation. Regularisation can be done using various methods, such as dropout, weight decay, or label smoothing, which reduce the complexity or variance of the LLM.
- Distillation: This refers to the technique of transferring knowledge from a larger or more complex LLM to a smaller or simpler LLM. Distillation can reduce the computational cost and improve the efficiency of the LLM, while preserving its performance. Distillation can be done using various methods, such as teacher-student learning, self-distillation, or contrastive distillation.

Model training and optimization with respect to LLMs is a challenging and complex process that requires a lot of computational resources and data. By applying these strategies and techniques, we can train a LLM that can understand and generate natural language effectively and efficiently.

MODEL EVALUATION AND FINE-TUNING

Model evaluation and fine-tuning with respect to large language models (LLMs) is the process of measuring and improving the performance of a neural network that can learn from and generate natural language, using a test data set and a downstream task. Model evaluation and fine-tuning can be done using various metrics and methods, such as accuracy, F1-score, or distillation.

To explain in more detail, model evaluation and fine-tuning with respect to LLMs involves the following steps:

- Test data set: This is a subset of the training data set that is held out for evaluation purposes. The test data set should be representative of the natural language use cases and variations that the LLM is expected to handle. The test data set should also be aligned with the downstream task that the LLM is applied to, such as question answering, text summarization, or sentiment analysis. The test data set can be split into a validation set and a test set, where the validation set is used for tuning the hyperparameters of the LLM and the test set is used for reporting the final performance of the LLM.
- Downstream task: This is a specific natural language processing task that requires the LLM to perform well. The downstream task can be defined by a task objective and a task format, which specify what the LLM is trying to achieve and how the input and output text are formatted. For example, the task objective can be to generate a summary of a product review, and the task format can be to input the review text and output a summary text. The downstream task can also have different levels of difficulty and complexity, depending on how much domain knowledge or reasoning ability is required from the LLM.
- Evaluation metrics: These are quantitative measures that assess how well the LLM performs on the test data set and the downstream task. The evaluation metrics can vary depending on the type and objective of the LLM and the downstream task. For example, perplexity can measure how well the LLM predicts the next word in a

sentence, accuracy can measure how well the LLM answers a question correctly, and F1-score can measure how well the LLM extracts named entities from a text. The evaluation metrics can also have different properties and limitations, such as sensitivity, robustness, or interpretability.

- Fine-tuning methods: These are techniques that adjust the parameters of the LLM based on the downstream task, using a smaller learning rate and a smaller data set. Fine-tuning methods can improve the performance and efficiency of the LLM, while preserving its general linguistic knowledge. Fine-tuning methods can be done using various strategies and techniques, such as teacher-student learning, self-distillation, or contrastive distillation. Teacher-student learning involves training a smaller or simpler LLM (student) to mimic the outputs of a larger or more complex LLM (teacher). Self-distillation involves training a LLM to mimic its own outputs on different data sets or tasks. Contrastive distillation involves training a LLM to distinguish its own outputs from those of other models or humans.

Model evaluation and fine-tuning with respect to LLMs is an important and practical process that requires a lot of domain knowledge and expertise. By applying these steps, we can evaluate and fine-tune a LLM that can understand and generate natural language effectively and efficiently for various applications.

14 REINFORCEMENT LEARNING FROM HUMAN FEEDBACK

Reinforcement Learning from Human Feedback (RLHF), which is a technique for training large language models (LLMs) that can learn from and generate natural language, using human preferences as rewards. RLHF can align the outputs of LLMs with the expectations and values of human users, improving the quality and safety of interactions between humans and LLMs.

To explain in more detail, RLHF with respect to LLMs involves the following steps:

- Human feedback: This is the source of rewards that guide the learning of LLMs. Human feedback can be explicit or implicit, depending on how humans express their preferences. Explicit feedback can be numerical ratings, binary labels, or rankings that humans provide for the outputs of LLMs. Implicit feedback can be behavioural signals, such as clicks, likes, or dwell time, that humans exhibit when interacting with LLMs.
- Reward model: This is a function that maps the outputs of LLMs and the human feedback to scalar rewards. The reward model can be learned from data or designed by experts, depending on the availability and quality of human feedback. The reward model can also be adaptive or fixed, depending on whether it updates its parameters during training or not.
- Policy model: This is the LLM that generates outputs based on inputs. The policy model can be pre-trained on a large corpus of text or initialised randomly, depending on the prior knowledge and generalisation ability of the LLM. The policy model can also use different sampling methods, such as greedy, beam search, or top-k sampling, to generate outputs with different diversity and quality.
- Optimization algorithm: This is the method that updates the parameters of the policy model based on the rewards from the reward model. The optimization algorithm can use various techniques, such as gradient descent, proximal policy optimization (PPO), or evolutionary strategies, to maximise the expected reward or minimise the expected regret.

RLHF with respect to LLMs is a novel and promising technique that leverages human intelligence and preferences to train LLMs that can understand and generate natural language effectively and ethically for various applications.

Every interface to a LLM now allows for user feedback, to better train the model through reinforcement learning. And this methodology has sparked debates on who gets to set what the feedback is on things where this is not a clear scientific argument and is an active area of research.

SOCIAL AND ECONOMIC ISSUES

Researchers criticise the development and deployment of LLMs, especially for English, without considering the potential risks and harms associated with them. Researchers use the term “stochastic parrots” to refer to LLMs that blindly repeat and recombine human-authored texts without understanding their meaning or context. Researchers argue that stochastic parrots pose several dangers, such as:

- Environmental and financial costs: LLMs require massive amounts of data and computation to train and run, which consume a lot of energy and resources, contributing to carbon emissions and climate change. Moreover, LLMs are often inaccessible and unaffordable for many researchers and communities, creating a digital divide and exacerbating inequalities.
- Insufficient evaluation and inscrutability: LLMs are often evaluated by narrow and biassed metrics, such as leaderboards and benchmarks, that do not capture the full range and complexity of natural language use cases and variations. Moreover, LLMs are often opaque and uninterpretable, making it hard to explain or debug their outputs or behaviours, especially when they produce errors or harms.
- Illusory meaning and understanding: LLMs are often assumed to have semantic or pragmatic competence, that is, the ability to understand the meaning and context of natural language. However, LLMs are actually based on statistical patterns and correlations, not on linguistic or world knowledge. Therefore, LLMs can generate outputs that are fluent but nonsensical, misleading, or harmful.
- Misalignment and misuse: LLMs are often trained on data that is not representative or aligned with the values and preferences of human users or stakeholders. This can result in outputs that are biassed, offensive, or inappropriate for certain domains or audiences. Moreover, LLMs can be misused or abused for malicious purposes, such as deception, manipulation, or propaganda.

There are several real world problems that came to limelight in the last few months:

1. Several religious chatbots have been introduced and a lot of them are not aligned with human rights and values as some condone murder, some telling the user which group of individuals can be killed in the name of faith etc.
2. A recent survey showed that these chatbots are being used to create websites optimised for search engine optimization and just spamming search engine results creating a havoc for surfacing the correct results.
3. Since the free version of these LLMs offer limited functionality, the paid versions provide full functionality and are showing the prevalent wealth divide and proving to be not as democratised as search engines were.

To address these dangers, academia propose several recommendations, such as:

- Weighing the environmental and financial costs first: LLMs should not be developed or deployed without considering the trade-offs between their benefits and costs. Researchers should estimate and report the carbon footprint and resource consumption of their LLMs, and seek alternative methods that are more efficient and sustainable.
- Investing resources into curating and documenting datasets: LLMs should not be trained on data that is scraped from the web without quality control or provenance tracking. Researchers should invest more time and effort into collecting and annotating data that is diverse, representative, reliable, and ethical. Researchers should also document the sources, methods, limitations, and biases of their datasets.
- Carrying out pre-development exercises: LLMs should not be designed or implemented without consulting with potential users or stakeholders. Researchers should conduct pre-development exercises that evaluate how their LLMs fit into their research and development goals and support their stakeholder values. Researchers should also anticipate and mitigate the possible risks and harms of their LLMs.
- Encouraging research directions beyond LLMs: LLMs should not be seen as the only or ultimate solution for natural language processing. Researchers should explore other research directions that are more linguistically informed, socially

aware, or human-centred. Researchers should also collaborate with other disciplines and communities to enrich their perspectives and approaches.

- Regulation: Governments could regulate the use of LLMs to prevent them from being used for harmful purposes. For example, governments could require companies to obtain permission before using LLMs to collect personal data.
- Education: People need to be educated about the potential risks of LLMs. This will help people to be more critical of the information that they see online and to be more aware of the privacy risks associated with LLMs.

According to economists and researchers the effects of these large language models on the job market can go in two ways. First camp firmly believes that this technology would increase productivity for everyone, while the second camp believes that this technology showcases how futile some jobs are and it’s time we have a discussion of the relevance of such jobs in the labour market and look towards concepts like universal basic income.

FUTURE DIRECTIONS

LLMs are still under development, but they are becoming increasingly powerful and capable of performing a wider range of tasks. As LLMs become more sophisticated, it is likely that they will be used in a variety of new and innovative ways. However, LLMs also face many challenges and limitations, such as environmental and financial costs, insufficient evaluation and inscrutability, illusory meaning and understanding, misalignment and misuse. Therefore, there is a need for further research and development to improve the performance, efficiency, and ethics of LLMs.

Some of the future advancements and avenues of research with respect to LLMs are:

- Self-training: Self-training is a technique that allows LLMs to learn from their own outputs without requiring additional human supervision or feedback. Self-training can improve the generalisation and adaptation of LLMs to new domains or tasks, as well as reduce the data and computation requirements for training LLMs.
- Fact-checking: Fact-checking is a technique that allows LLMs to verify the accuracy and reliability of the information they generate or consume. Fact-checking can improve the quality and safety of LLMs by preventing or correcting factual errors, contradictions, or logical fallacies in their outputs.
- Sparse expertise: Sparse expertise is a technique that allows LLMs to leverage specialised knowledge or skills from other models or humans without requiring full integration or collaboration. Sparse expertise can improve the efficiency and flexibility of LLMs by enabling them to access or provide expertise on demand or on a case-by-case basis.
- Multi-modal and Multi-lingual Capabilities: Expanding large language models to include multi-modal understanding (text, images, audio, etc.) and multilingual capabilities is an active area of research. Future advancements seek to develop models that can effectively process and generate text in multiple languages and understand and generate content that combines text with other modalities, enabling more comprehensive and versatile interactions.

In conclusion, large language models have the prospective to change different fields and domains by generating and understanding human-like language. Nevertheless, they also pose many difficulties and dangers that need to be attended to by more research and development. A few of the future developments and avenues of research study with respect to LLMs are self-training, fact-checking, and sparse proficiency, which can enhance the performance, efficiency, and ethics of LLMs.

6% Overall Similarity

Top sources found in the following databases:

- 2% Internet database
- Crossref Posted Content database
- 0% Publications database
- 5% Submitted Works database

TOP SOURCES

The sources with the highest number of matches within the submission. Overlapping sources will not be displayed.

1	Maulana Azad National Institute of Technology Bhopal on 2022-12-26 Submitted works	<1%
2	en.wikipedia.org Internet	<1%
3	researchgate.net Internet	<1%
4	University of Wales Swansea on 2023-01-05 Submitted works	<1%
5	CSU, San Jose State University on 2022-12-18 Submitted works	<1%
6	University of Westminster on 2023-05-10 Submitted works	<1%
7	Technical University of Liberec on 2023-05-07 Submitted works	<1%
8	University of Pittsburgh on 2023-04-29 Submitted works	<1%
9	Columbia University on 2023-05-07 Submitted works	<1%

10	University of Strathclyde on 2023-03-25 Submitted works	<1%
11	journals.ayu.edu.kz Internet	<1%
12	University of Leeds on 2021-09-09 Submitted works	<1%
13	University of Melbourne on 2023-04-17 Submitted works	<1%
14	medium.com Internet	<1%
15	Indiana University on 2023-03-22 Submitted works	<1%
16	76830 on 2015-02-25 Submitted works	<1%
17	Liverpool John Moores University on 2023-04-04 Submitted works	<1%
18	Sim University on 2023-02-14 Submitted works	<1%
19	maintworld.com Internet	<1%
20	Liverpool John Moores University on 2022-12-12 Submitted works	<1%
21	SSN COLLEGE OF ENGINEERING, Kalavakkam on 2022-06-06 Submitted works	<1%

22	University College London on 2016-09-06 Submitted works	<1%
23	University of Durham on 2020-09-04 Submitted works	<1%