A REVIEW ON LARGE LANGUAGE MODELS: ARCHITECTURES, APPLICATIONS, TAXONOMIES, OPEN ISSUES AND CHALLENGES

Mohaimenul Azam Khan Raiaan[1], Md. Saddam Hossain Mukta[1], Kaniz Fatema[2], Nur Mohammad Fahad[1], Sadman Sakib[1], **Most. Marufatul Jannat Mim**[1], **Jubaer Ahmad**[1] **Mohammed Eunus Ali**[3],and **Sami Azam**[2]
[1] Department of CSE, United International University (UIU), Dhaka-1212, Bangladesh
[2] Faculty Science and Technology, Charles Darwin University, Australia
[3] Department of CSE, Bangladesh University of Engineering and Technology (BUET), Dhaka-1000, Bangladesh
Email: mraiaan191228@bscse.uiu.ac.bd, saddam@cse.uiu.ac.bd, kaniz.fatema@cdu.edu.au, {nfahad191040, ssakib191097, mmim192004, jahmad181023}@bscse.uiu.ac.bd, eunus@cse.buet.ac.bd, Sami.Azam@cdu.edu.au

ABSTRACT

Large Language Models (LLMs) recently demonstrated extraordinary capability, including natural language processing (NLP), language translation, text generation, question answering, etc. Moreover, LLMs are a new and essential part of computerized language processing, having the ability to understand complex verbal patterns and generate coherent and appropriate replies for the situation. Though this success of LLMs has prompted a substantial increase in research contributions, rapid growth has made it difficult to understand the overall impact of these improvements. Since a lot of new research on LLMs is coming out quickly, it is getting tough to get an overview of all of them in a short note. Consequently, the research community would benefit from a short but thorough review of the recent changes in this area. This article thoroughly overviews LLMs, including their history, architectures, transformers, resources, training methods, applications, impacts, challenges, etc. This paper begins by discussing the fundamental concepts of LLMs with its traditional pipeline of the LLM training phase. It then provides an overview of the existing works, the history of LLMs, their evolution over time, the architecture of transformers in LLMs, the different resources of LLMs, and the different training methods that have been used to train them. It also demonstrated the datasets utilized in the studies. After that, the paper discusses the wide range of applications of LLMs, including biomedical and healthcare, education, social, business, and agriculture. It also illustrates how LLMs create an impact on society and shape the future of AI and how they can be used to solve real-world problems. Then it also explores open issues and challenges to deploying LLMs in real-world aspects, including ethical issues, model biases, computing resources, interoperability, contextual constraints, privacy, security, etc. It also discusses methods to improve the robustness and controllability of LLMs. Finally, the study analyses the future of LLM research and issues that need to be overcome to make LLMs more impactful and reliable. However, this review paper aims to help practitioners, researchers, and experts thoroughly understand the evolution of LLMs, pre-trained architectures, applications, challenges, and future goals. Furthermore, it serves as a valuable reference for future development and application of LLM in numerous practical domains.

***K*eywords** Large Language Models, Natural Language Processing, Evolution, Transformer, Pre-trained models, Taxonomy, Application

1 Introduction

Language is a remarkable tool for human expression and communication, one that begins to emerge in infancy and makers throughout a lifetime [1, 2]. Nevertheless, machines are unable to possess the innate ability to understand and speak in human language without the help of sophisticated artificial intelligence (AI) [3]. Therefore, a long-standing scientific challenge and aim has been to achieve human-like reading, writing, and communication skills in machines [4]. However, advances in deep learning approaches, the availability of immense computer resources, and the availability

of vast quantities of training data all contributed to the emergence of large language models (LLMs). It is a category of language models that utilizes neural networks containing billions of parameters, trained on enormous quantities of unlabeled text data using a self-supervised learning approach [5]. It is considered a huge step forward in natural language processing (NLP) and AI [6]. These models, frequently pre-trained on large corpora from the web, may learn complicated patterns, language subtleties, and semantic linkages. Besides, they have proved their ability in various language-related tasks, including text synthesis, translation, summarization, question-answering, and sentiment analysis, by leveraging deep learning techniques and large datasets. Moreover, the results of fine-tuning these models on specific downstream tasks have been quite promising, with state-of-the-art performance in several benchmarks [7]. LLMs have their roots in the early development of language models and neural networks. Statistical approaches and n-gram models were used in earlier attempts to develop language models [8], but these models have shortcomings in expressing long-term interdependence and context in language. After that, researchers began to explore more complex ways with the development of neural networks and the availability of larger datasets. The creation of the Recurrent Neural Network (RNN) [9], which allowed for the modeling of sequential data, including language, was a crucial milestone. However, RNNs were limited in their efficacy due to vanishing gradients and long-term dependencies. The significant advancement in LLMs systems occurred when the transformer architecture was introduced in the seminal work [10]. The transformer model is built around the self-attention mechanism, enabling parallelization and efficient handling of long-range dependencies. Furthermore, it served as the basis for models such as Google's Bidirectional Encoder Representations from Transformers [11] and open AI's Generative Pre-trained Transformer (GPT) series, which excelled at various language tasks.

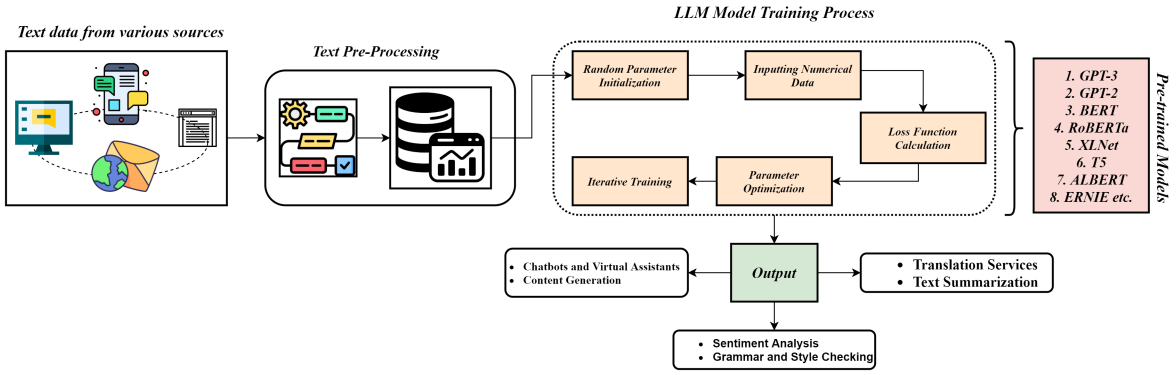


Figure 1: Pipeline of the LLM training phase

The pipeline of the basic LLM architecture is shown in Figure 1. It receives text data from a variety of sources and then forwards it to the subsequent stage for preprocessing. It then completes its training process by executing a series of stages, including random parameter initialization, numerical data input, loss function calculation, parameter optimization, and iterative training. They offer text translation, text summarization, sentiment analysis, and other services following the training phase.

Prior research has shown the potential of LLMs in many NLP tasks, including specialized applications in domains such as the medical and health sciences [12] and politics [13]. Moreover, after inventing the most sophisticated GPT model [14], developing the state-of-the-art models (LLaMa and Bard [15]), and exploring their capabilities, such as Alpaca and GPTHuggingface [16], it has become a crucial and impactful domain. As a result, a proper assessment of current LLM research is becoming increasingly important, and prior research has shown the potential and superiority of LLMs in NLP tasks. Nevertheless, few studies have thoroughly reviewed their work's most recent LLM developments, possibilities, and limitations. Despite the increasing number of studies on LLMs, there remains a scarcity of research focusing on their technical complexities, the LLMs taxonomy, architectures, API applications, domain-specific applications, effective utilization, impact on society, and so on. Furthermore, the majority of the LLM review papers are not peer-reviewed articles. So, the motivation of this paper is to explore the current review papers, identify their limitations, and outline current state-of-the-art methodologies that have recently been created to address these challenges. However, our main objective is to explore, learn, and assess LLMs across domains, evolutions, classifications, pre-trained models' architectures, resources, and real-time applications. Additionally, our comprehensive review discusses open issues and challenges associated with LLM, including safety, ethical, privacy, economic, and environmental considerations. In addition, we present a set of guidelines to direct future research and development in the effective use of LLM. We hope that this study will contribute to a better understanding and use of LLMs. The list of contributions to this paper is as follows:

- Providing an exhaustive overview of LLMs, including their evolutions, taxonomies, and transformer architectures.
- Describing a comparative analysis of distinct pre-trained model architectures in LLMs along with their individual infrastructures.
- Explaining the impact of ML models in LLMs.
- Defining insight into the prospects of LLMs and their impact on society, as well as showcasing the applications of LLMs in five practical domains, including bio-medical and healthcare, education, social media, business, and agriculture.

The remaining sections of the paper are organized as depicted in Figure 2.

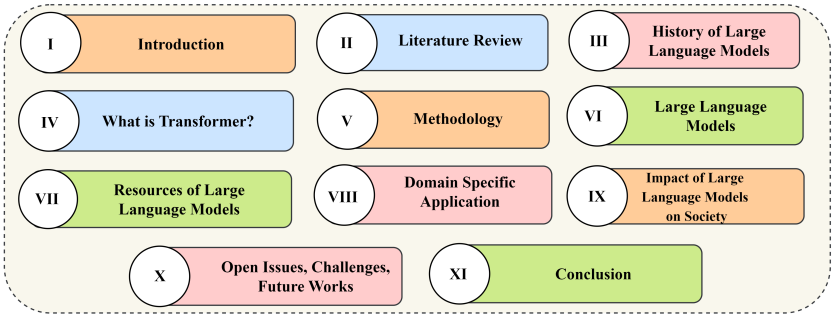


Figure 2: Section organization of the review

In Section 2, the literature review is discussed. Section 3 illustrates the history of LLMs; Section 4 explains the clear concept of the Transformer; Section 5 demonstrates the Methodology; Section demonstrates the pre-trained models' architectures, models comparison, and dataset; Section 7 describes the resources of LLMs; Section 8 demonstrates the domain-specific applications of LLMs; and Section 9 explains the societal impact of LLMs, Section 10 discuss the open issues and challenges regarding this study and Section 11 finally concludes the paper.

2 Literature review

The growing number of LLMs is an extraordinary development in AI. In recent years, the prevalence of these models has skyrocketed, and numerous studies have been conducted to investigate and evaluate their expanding capabilities. Researchers from various fields have conducted exhaustive studies on the rise of LLMs, shedding light on their remarkable advancements, diverse applications, and potential to revolutionize tasks from text generation and comprehension to demonstrating reasoning skills. Collectively, these studies contribute to our comprehension of LLMs' significant role in shaping the landscape of AI-driven language processing and problem-solving.

Huang et al. [17] presented a study on reasoning in large Language models that comprehensively summarizes the current state of LLM reasoning capabilities. It examines various aspects of reasoning in LLMs, such as techniques to enhance and extract reasoning abilities, methodologies and criteria for assessing these abilities, insights from prior research, and suggestions for future directions. The primary concern is the extent to which LLMs can demonstrate reasoning skills. This paper aims to provide an in-depth and up-to-date examination of this topic, fostering fruitful discussions and guiding future research in LLM-based reasoning. In another study, Zhao et al. [3] survey on LLM illustrates a comprehensive examination of the evolution and impact of LLMs in the field of artificial intelligence and natural language processing. It traces the historical journey from early language models to the recent emergence of pre-trained language models (PLMs) with billions of parameters. Notably, the paper discusses LLMs' unique capabilities as they scale in size, including in-context learning. The authors highlight the significant contributions of LLMs to the AI community and the launch of ChatGPT, a prominent AI chatbot powered by LLMs. The survey is structured around four key aspects of LLMs: pre-training, adaptation tuning, utilization, and capacity evaluation. Additionally, the paper provides insights into available resources for LLM development and identifies further research and development areas.

A recent study by Fan et al. [18] conducted a bibliometric review of LLM research from 2017 to 2023, encompassing over 5,000 publications. The study aims to provide researchers, practitioners, and policymakers with an overview of the evolving landscape of LLM research. It tracks research trends during the specified time period, including advancements in fundamental algorithms, prominent NLP tasks, and applications in disciplines such as medicine,

engineering, the social sciences, and the humanities. In addition to highlighting the dynamic and swiftly changing nature of LLM research, the study offers insights into their current status, impact, and potential in the context of scientific and technological advancements. Another study by Chang et al., [19] focuses on the assessment of LMMs. Their research examines the increasing prevalence of LLMs in academia and industry due to their exceptional performance in various applications. It highlights the growing significance of evaluating LLMs at both the task and societal levels in order to comprehend potential risks. The paper thoroughly analyzes LLM evaluation methods, focusing on three critical dimensions: what to evaluate, where to evaluate, and how to evaluate. It includes tasks such as natural language processing, reasoning, medical applications, ethics, and education. The article examines evaluation methods and benchmarks for assessing LLM performance, emphasizing successful and unsuccessful cases. It underlines future challenges in LLM evaluation and emphasizes the significance of evaluating LLMs as a fundamental discipline to support the development of more competent LLMs.

Table 1: Comparison between state-of-the-art research

Papers LLM	LLM Model	LLM API	LLM Dataset	Domain Specific LLM	Taxonomy	LLM Architecture	LLM Configurations	ML Based Differentiation	Scope	Key Findings	Methodology and Approach
Huang et al. (2022) [17]	✓	X	X	X	X	X	X	X	Reasoning in LLMs	Aims to provide a critical analysis of LLM capabilities, methods for improving and evaluating reasoning, conclusions from earlier research, and future directions.	Review and analysis of reasoning abilities in LLMs
Zhao et al. (2023) [3]	✓	X	✓	X	✓	X	✓	X	Evolution and impact of LLMs	Explore the historical journey of LLMs, including pre-trained language models (PLMs) , discussed about LLMs' unique capabilities, insights into LLM development resources and highlights significant contributions of LLMs to AI and NLP research areas.	Survey and analysis of LLM evolution and impact
Fan et al. (2023) [18]	✓	X	X	X	X	X	X	X	Bibliometric review of LLM research	Present a comprehensive overview of LLM research from 2017 to 2023, tracking research trends, advancements, and provides insights into the dynamic nature of LLM research, and impact in various domains.	Bibliometric analysis of over 5,000 LLM publications
Chang et al. (2023) [19]	✓	X	✓	X	✓	X	X	X	Assessment of LLMs	Investigate the methodologies employed in evaluating LLM programs, with a specific focus on the aspects of what, where and how to conduct evaluations and identified the potential risks and the future challenge also.	Survey and analysis of LLM evaluation approaches
OURS	✓	✓	✓	✓	✓	✓	✓	✓	**Detailed review on LLMs**	**Our research investigated the history, resources, architectural configuration, domain-specific analysis, ml-based differentiation, broad level of open issues, challenges, and future scope of large language models.**	**Broad review and analysis of LLMs considering all the key aspects**

Table 1 illustrates the comparison between different review papers based on some critical factors such as LLM Model, LLM API, LLM Dataset, Domain Specific LLM, Taxonomy, LLM Architecture, LLM Configurations, and ML Based Differentiation. Huang et al. [17] lack information on LLM API, LLM Dataset, Domain-Specific LLM, Taxonomy, LLM Architecture, and LLM Configurations. In contrast, Zhao et al. [3] lack information on LLM API, Domain-Specific LLM, Taxonomy, LLM Architecture, and LLM Configurations. Moreover, Fan et al. [18] and Chang et al. [19] lack information on LLM API, Domain-Specific LLM, Taxonomy, LLM Architecture, and LLM Configurations.

Our research shows more insights and discusses more features over the state-of-the-art studies mentioned above, given that it encompasses all the parameters in the table, thereby providing a holistic view of the state-of-the-art in LLM research. While other studies concentrate on particular aspects of LLMs, such as their historical evolution, bibliometric trends, or evaluation methodologies, our research encompasses all of these aspects, providing a comprehensive understanding of LLM capabilities. In addition, it focuses exclusively on the crucial aspect of reasoning abilities in LLMs, making a substantial contribution to the field's knowledge and making it an invaluable resource for LLM researchers and practitioners.

3 History of Large Language Models

LLMs refer to a category of AI models developed specifically to comprehend and produce human language [20]. LLMs have significantly transformed the field of AI and have been implemented in diverse areas, including education, communication, content generation, article composition, healthcare, research, entertainment, and information dissemination, among others [20, 21]. This section provides a high-level overview of LLMs, including their development, training, and functioning. Figure 3 depicts the history of language models.

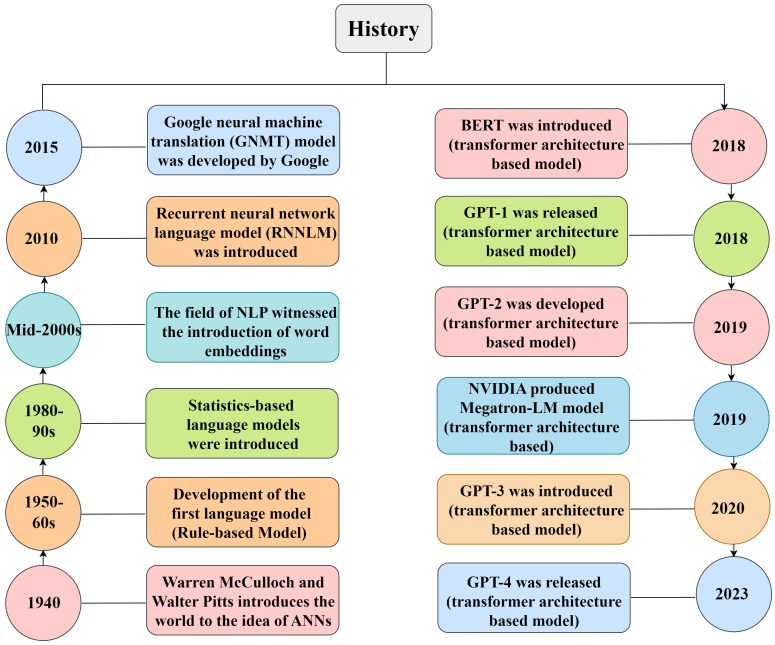


Figure 3: Brief history of language models.

In the 1940s, Warren McCulloch and Walter Pitts introduced the world to the idea of artificial neural networks (ANNs) [22]. Following this, the 1950s and 1960s saw the development of the first language models [23]. These models included early neural networks as well as rule-based models. The processing of language was facilitated by their utilization of precisely established linguistic rules and features [24].

Statistics-based models of language were created in the ’80s and ’90s. These models belong to a category of models utilized in NLP and machine learning (ML) to capture and quantify the statistical patterns and correlations within language data [25]. The models employed probabilistic techniques to assess the probability of a sequence of words or phrases inside a specific context. They were superior in terms of accuracy to early neural networks and rule-based models, as they were able to process large amounts of data with ease [25].

During the mid-2000s, the field of NLP witnessed the introduction of word embeddings, recognized as a notable breakthrough, and subsequently acquired considerable attention [26]. This approach captures the semantic relationships among words by representing them in a vector space [27]. Although not classified as LLMs, these embeddings have significantly contributed to the progress of natural language comprehension and have set the path for developing more complex models [26].

The introduction of neural language models in the mid-2010s marked a significant advancement in large language modeling [28]. The initial neural language model to be introduced was the recurrent neural network language model (RNNLM) in 2010 [29]. Its development aimed to capture the sequential dependencies present in textual data [30]. The RNNLM demonstrated the capability to effectively capture the contextual information of words, resulting in the generation of text that exhibits a higher degree of naturalness compared to earlier models [31].

In the year 2015, Google unveiled the initial large neural language model that employed deep learning methodologies [32]. The technology was referred to as the Google Neural Machine Translation (GNMT) model [33]. The development of this model signifies a notable progression in the field of machine translation [34]. The utilization of this model resulted in enhanced translation accuracy and the generation of meaningful translations [33].

The advancement of Language models persisted with the emergence of the Transformer model in the year 2017 [35]. The transformer model has played a crucial role in the development of language models such as Bidirectional Encoder Representations from Transformers (BERT) and Generative Pre-trained Transformers (GPT) [36]. The primary objective behind developing the Transformer model was to overcome the inherent constraints observed in earlier models such as RNNs and LSTM networks [35].

The introduction of transformer architecture-based BERT in 2018 by Google AI represents a noteworthy advancement in the domain of NLP [18]. Before the introduction of BERT, the preceding language model rooted in NLP had constraints

in understanding contextual information due to its reliance on unidirectional language modeling. BERT was introduced by Google as a solution to address this particular constraint [37]. The employed methodology involved the utilization of deep bidirectional representations, which were conditioned on both the left and right contexts across all layers [38].

In 2018, GPT was developed by OpenAI, which was also a transformer-based architecture [39]. The introduction of GPT-1 was a notable progression in the field of NLP. GPT-1 effectively produces contextually appropriate words, showcasing the transformative capabilities of transformers in significantly advancing NLP tasks. This proficiency is attributed to its extensive training on many parameters, specifically 117 million [39]. The model underwent a two-step procedure consisting of unsupervised pre-training followed by supervised fine-tuning [21].

The subsequent version of the GPT series, known as GPT-2, was designed to address the limitations observed in GPT-1 [40]. Similar to GPT-1, GPT-2 was developed utilizing the Transformer architecture. In the year 2019, Alec Radford introduced GPT-2, which was developed on a deep neural network consisting of 1.5 billion parameters [41]. The GPT-2 model includes a transformer design, which incorporates self-attention processes to extract information from different positions within the input sequence [40]. The GPT-2 model has played a pivotal function in the advancement of LLMs and the execution of NLP activities [42].

In 2019, NVIDIA produced Megatron-LM, which is an LLM [43]. Similar to GPT, this model is built on the transformer architecture. The model possesses a total of 8.3 billion parameters, a notably bigger quantity compared to the parameter count of GPT-1 and GPT-2. The magnitude of this dimension facilitates the model's capacity to acquire and produce intricate linguistic structures [18].

In the year 2020, OpenAI introduced GPT-3 as the successor to GPT-2 [40]. GPT-3 was trained on an extensive collection of textual data and demonstrated the ability to generate text that exhibited a high degree of coherence and naturalness. Similar to GPT-1 and GPT-2, this model also utilizes the Transformer architecture [21]. GPT-3 was trained on a deep neural network with an enormous 175 billion parameters, surpassing the size of any other LLMs available at that particular time [18]. The ability to produce natural language text of superior quality with less fine-tuning is facilitated by sophisticated methodologies, including a more significant number of layers and a wider range of training data.

In the year 2023, OpenAI introduced GPT-4, the subsequent version of their language model, following the achievements of GPT-3 [21]. Similar to its predecessor, GPT-4 is a transformer-based model. The system has the capability to analyze both textual and visual data to produce textual outputs [18]. The system's performance was assessed using a range of standardized professional and academic examinations specifically intended for human test-takers [44]. GPT-4 has greater dimension and efficacy than its predecessor, as it can generate text that is even more comprehensive and exhibits a heightened level of naturalness [21].

The development of large language models presents additional prospects for innovation, knowledge acquisition, and experimentation across diverse domains such as healthcare, education, research, etc. The utilization of AI and NLP in these models has significantly transformed how we engage with machine devices.

4 What is Transformer?

The transformer architecture is considered the fundamental building block of LLMs. It is intended for neural networks to handle sequential data effectively [10]. This architecture does not utilize recursion methods. Instead, it employs an attention method to determine global input-output dependencies. It has resulted in novel model sizes and performance levels, allowing for substantially increased parallelization and reduced training times in NLP. Furthermore, it can take input of varying lengths and change its attention depending on the length of the sequence. As a result, it became the go-to architecture in many fields, frequently replacing sophisticated recurrent or convolutional neural networks with a far more efficient structure [35]. In this regard, it is especially important for LLM applications. Figure 4 depicts the transformer model's architecture. Transformer architecture consists of seven key components. A demonstration of each of the components is shown below [10].

4.1 Inputs and Input Embeddings

The ML models use user-entered tokens as training data, while it can only process numeric information. Thus, it is necessary to transform these textual inputs into a numerical format known as "input embeddings." These input embeddings are numerical representations of words, which ML models may subsequently process. These embeddings function similarly to a dictionary, assisting the model in understanding the meaning of words by arranging them in a mathematical space where comparable phrases are situated close together. The model is trained to generate these

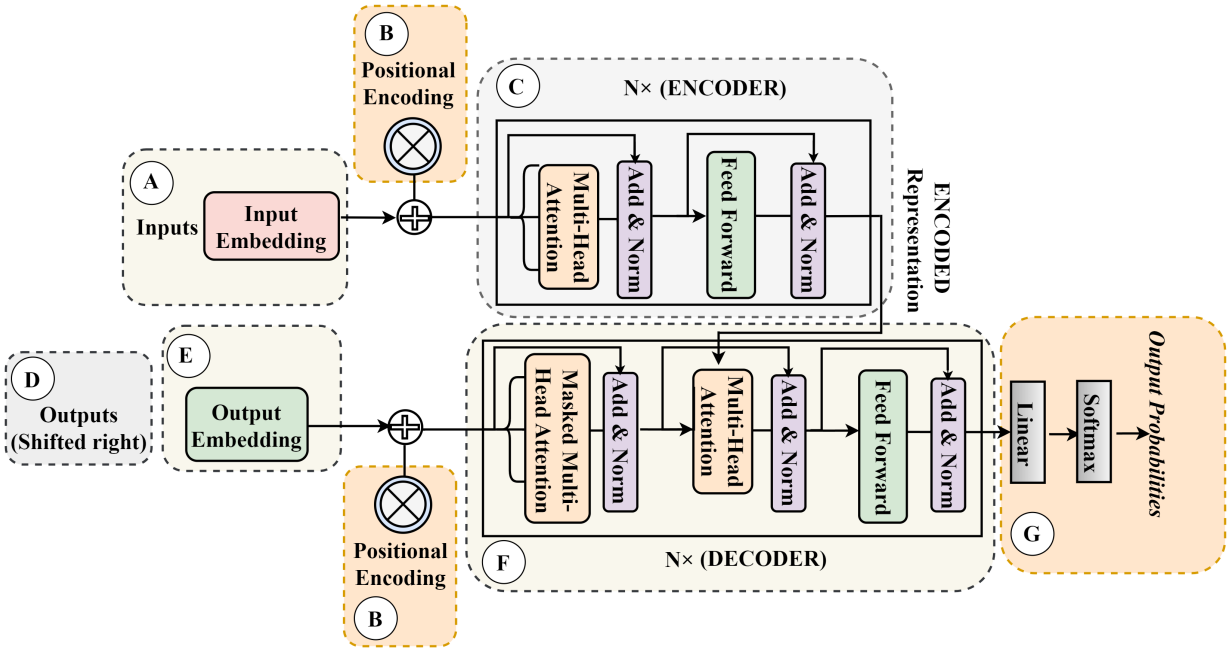


Figure 4: Architecture of a Transformer model

embeddings so that vectors of the same size represent words with similar meanings. Figure 4A illustrates the input and input embeddings.

4.2 Positional Encoding

The order of words in a sentence is essential in the NLP field for identifying the statement's meaning. In general, in terms of neural networks, they do not fundamentally grasp the sequence of inputs. To remedy the issue, positional encoding may be utilized to encode each word's location in the input sequence as a collection of integers. The Transformer model accepts these integers and input embeddings while adding positional encoding to the transformer architecture allows GPT to grasp sentence word order better and provide grammatically accurate and semantically relevant output. The positional encoding part is displayed in Figure 4B.

4.3 Encoder

The encoder is part of the neural network that processes the input text and generates a series of hidden data. Then, it uses a series of self-attention layers, which we can think of as voodoo magic, to make several hidden states that describe the input text at different levels of abstraction. In the transformer, the encoder is used in more than one layer. This section is depicted in Figure 4C comprehensively.

4.4 Outputs (shifted right)

During the training process, the decoder acquires the ability to predict the next word by analyzing the previous words. In this case, the output sequence is shifted by one position to the right. Consequently, the decoder is able to use the words that came before it. Additionally, the GPT (GPT-3) is also trained on a massive amount of text data, that helps it generate sense while writing something. Besides, several corpus including the Common Crawl web corpus, the BooksCorpus dataset, and the English Wikipedia are also used during the common issue. Figure 4D highlights the transformer's outputs (shifted right) module.

4.5 Output Embeddings

Input embeddings, which contain text and are not recognized by the model. Therefore, the output must be converted to a format known as "output embedding." Like input embeddings, output embeddings undergo positional encoding,

enabling the model to understand the order of words in a sentence. In machine learning, the loss function evaluates the difference between a model's prediction and the objective value. Loss functions are essential for complex GPT language models. The loss function modifies a portion of the model to increase accuracy by reducing the discrepancy between predictions and targets. The change improves the overall performance of the model. The loss function is calculated during training, and the model parameters are modified. In the inference process, the output text is created by mapping the predicted probability of each token in the model to the corresponding token in the vocabulary. The output embedding part is illustrated in Figure 4E.

4.6 Decoder

The decoder processes positionally encoded input and output embedding. Model decoders create output sequences from encoded input sequences while the decoder learns to predict the next word from previous words during the training period. In addition, the GPT's decoder generates natural language text by utilizing encoder context and input sequence. However, the transformers employ many decoder layers like encoders. Figure 4F demonstrates the decoder component of a transformer.

4.7 Linear Layer and Softmax

The linear layer maps to the higher-dimensional space once the decoder has generated the output embedding. This step is required to convert the output embedding into the original input space. The softmax function generates a probability distribution for each output token in the developed vocabulary, allowing us to generate probabilistic output tokens. Figure 4G shows the process by which the features are propagated through a linear layer, followed by the activation of the accurate output probability using the softmax activation function.

5 Methodology

The research materials utilized in this study have been obtained from reputable scholarly journals and conferences, spanning the time frame between January 2020 and August 2023. The search and selection process was carried out using the Google Scholar platform. Our primary objective is to acquire relevant articles written in the English language. A compilation of scholarly research publications has been selected, including a wide array of esteemed academic sources such as IEEE Xplore, ScienceDirect, ACM Digital Library, Wiley Online Library, Springer Link, MDPI, and patents. Table 2 depicts the electronic database that was utilized to conduct a comprehensive search for papers relevant to this research.

Table 2: Electronic database search

Electronic Database	**Type**	**URL**
IEEE Xplore	Digital Library	`https://ieeexplore.ieee.org/Xplore/home.jsp` (accessed on 18 September, 2023)
Springer	Digital Library	`https://www.springer.com/gp` (accessed on 18 September, 2023)
Google Scholar	Search Engine	`https://scholar.google.com.au` (accessed on 18 September, 2023)
Science Direct—Elsevier	Digital Library	`https://www.sciencedirect.com` (accessed on 18 September, 2023)
MDPI	Digital Library	`https://www.mdpi.com` (accessed on 18 September, 2023)
ACM	Digital Library	`https://www.researchgate.net` (accessed on 18 September, 2023)

Additionally, the electronic databases are accompanied by their respective URLs. To perform an extensive search of the articles, a diverse set of Search Queries were utilized, incorporating terms such as "LLM AND machine learning OR deep learning OR models", "LLM AND machine learning OR deep learning OR API", "LLM AND machine learning OR deep learning OR Dataset", and "LLM" AND machine learning OR deep learning OR tools". Table 3 presents the Search Queries (SQ) employed in this study. Figure 5 illustrates the comprehensive search running on Google Scholar, employing the given search queries (SQs) to identify relevant scholarly articles for the study. In the beginning, a total of

Table 3: Search queries used for the review paper.

	Search Queries (SQ)
SQ1	"LLM" AND machine learning OR deep learning OR models
SQ2	"LLM" AND machine learning OR deep learning OR API
SQ3	"LLM" AND machine learning OR deep learning OR Dataset
SQ4	"LLM" AND machine learning OR deep learning OR tools

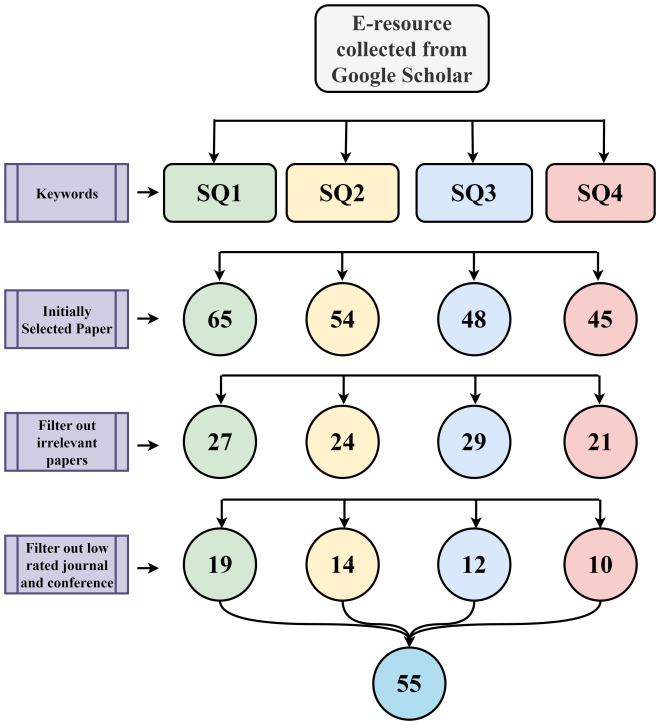


Figure 5: Flow diagram of systematic review

212 scholarly publications were collected from various academic repositories. The inclusion and exclusion criteria for the articles included in this study are presented in Table 4.

The table outlines the addition of five inclusion criteria (IC) and four exclusion criteria (EC) for selecting the relevant articles of the study. In addition, we have eliminated the articles that are not relevant to our research. Following the implementation of the filtering procedure, the total count of papers was diminished to 101. Next, we proceeded to exclude journal and conference papers that were scored poorly in terms of quality. In the end, we achieved a further reduction in our compilation by deleting the redundant articles. For their precise alignment with the primary objective of our study, a total of 55 papers were selected for inclusion in this article.

6 Large Language Models

Large language models (LLMs) refer to a specific type of AI algorithm that holds the capability to execute a diverse range of NLP operations. The most common tasks entail text generation, text analysis, translation, sentiment analysis, question answering, and other related functions. GPT-3, GPT-4, PaLM, and LaMDA are extensively used transformer-based LLM models trained on a large amount of textual data. In terms of architectural properties, these models show variations in size and depth. For example, GPT-3 generates parameters of 175 billion, distributed across 96 levels, while PaLM has an even larger parameter number of 540 billion, organized across 106 layers. All of these models

Table 4: Inclusion and exclusion criteria.

	List of Inclusion and Exclusion Criteria
Inclusion Criteria (IC)	
IC1	Should contain at least one of the keywords
IC2	Must be included in one of the selected databases
IC3	Published within the last ten years (2014–2023)
IC4	Publication in a journal, conference is required
IC5	The research being examined should have a matching title, abstract, and full text
Exclusion Criteria (EC)	
EC1	Redundant items
EC2	Whole text of paper cannot be taken
EC3	Purpose of the paper is not related to LLM
EC4	Non-english documents

have distinct configurations. The configurations of GPT-3 and PaLM differ in terms of their techniques for generating output. LLMs have evaluated several datasets within Wikipedia, code repositories, books, question sets, and social media data. They have demonstrated their ability to execute diverse activities successfully. Consequently, LLMs have drawn significant attention for their effective contribution in different domains, including education, healthcare, media marketing, and other customer services. A particular LLM program has superior performance in a specific domain compared to others, such as GPT-3, which has gained recognition for its proficiency in generating text styles, whereas LaMDA demonstrates superior performance in providing accurate responses to factual inquiries. LLMs are an emerging technological innovation that holds the potential to bring about transformative changes across various sectors.

In this section, the architectural overview of LLMs is discussed initially in the subsection 6.1. Then, we presented the comparison between configurations of LLMs in the next subsection 6.2, and finally, in subsection 6.3, the datasets utilized for training the LLMs are discussed.

6.1 Architectural Overview of Large Language Models

In Table 5, a description and architecture of LLMs such as GPT-1, BERT, RoBERTa, and T5 are presented.

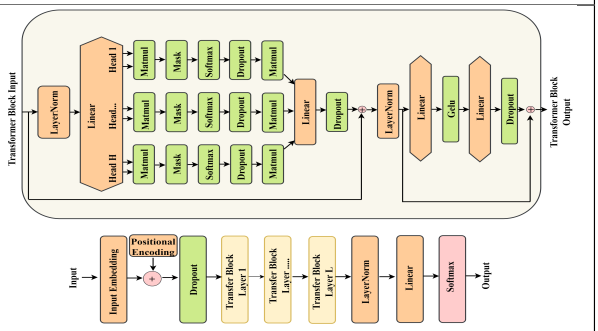
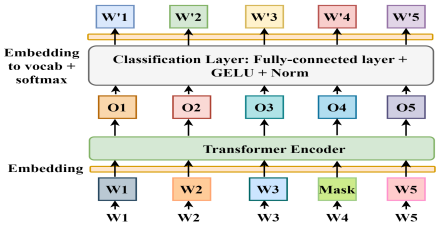
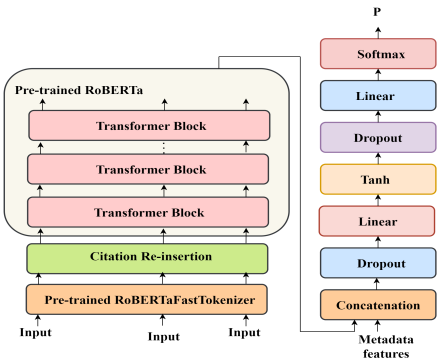
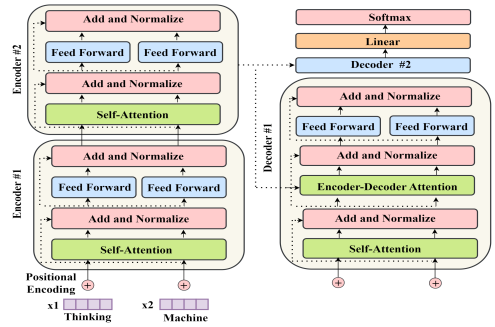
This table will assist researchers in selecting the optimal model for a natural language processing task. GPT-1, BERT base, and BERT large contain 12, 12, and 24 layers, correspondingly, in the larger language model. RoBERTa is an enhanced variant of BERT, while T5 is a decoder and encoder transformer. Diagram illustrating BERT's input token processing, context-aware embedding, and masked language modeling tasks, where the masked words are intended to predict the model. T5 demonstrates the sequential layers of the transformer model, including the feedforward neural network, and self-attention. It explains how information flows and structures text. GPT-1 passes data input embedding and positional encoding through multiple transformer layers.

6.2 Comparison Between Configurations of LLMs

Table 6 provides an extensive overview of various Large Language Models (LLMs), highlighting their configuration details and optimization settings.

These LLMs have played a crucial role in advancing natural language comprehension and generation tasks, making them a focal point in artificial intelligence and natural language processing. This analysis compares and contrasts these LLMs based on critical parameters, including model size, learning rate, category, activation function, batch size, bias, number of layers, optimizer, number of attention heads, hidden state size, dropout rate, and maximum training context length. GPT-4 stands out as the most prominent model on display, with a staggering 1.8 trillion parameters. GPT-1, despite being lesser with 125 million parameters, demonstrates the significant development of LLM over the years. An increased number of parameters in LLM enhances the model's ability to comprehend intricate patterns and produce text that is more contextually appropriate and reminiscent of human language. GPT3's selection of a modest learning rate of 6 is notable, which highlights the significance of cautious hyperparameter selection. Models are categorized as Causal decoder (CD), Autoregressive (AR), Encoder-decoder (ED), and Prefix decoder (PD) to illustrate architectural diversity. Activation functions vary, influencing the models' expressive strength from GeLU in GPT-3 to SwiGLU in LLaMA and LLaMA-2. All versions of GPT employ the GeLU as its activation function as it mitigates the vanishing gradient problem and facilitates the generation of smoother gradients throughout the training process. The utilization of SwiGLU as the activation function is observed in models such as PaLM and LLaMA versions 1 and 2, as it has gating

Table 5: Architectural overview of different LLMs

Model	Description	Architecture
GPT-1 [45]	Twelve-level decoder transformer that uses twelve masked self-focusing heads.	
BERT [11]	BERT is a transformer architecture. It has two model sizes. BERT base has 12 layers in encoder stack and BERT Large has 24 layers in encoder stack.	
RoBERTa [46]	Optimized version of BERT model.	
T5 [47]	The model consists of an encoder and a decoder transformer, which has many layers.	

mechanisms that enhance its ability to capture intricate correlations within the data. Models like BERT, OPT, and T5 use ReLU as the activation function. The Formula of these activation functions are given below [67, 68]:

$$ReLU(x) = \max(0, x) = f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \tag{1}$$

$$GeLU(x) = 0.5x(tanh[\sqrt{2/\pi}(x + 0.44715x^3)]) \tag{2}$$

$$SwiGLU(x) = x.Sigmoid(\beta x).xV \tag{3}$$

Table 6: Various LLMs with configuration details and optimization settings (Here, LR = learning rate, CG = Category, AF = the activation function, bs = batch size, NL = the number of layers, NAH = the number of attention heads, SHS = the size of the hidden states, MCLDT = the maximum context length during training, CD = causal decoder, ED = encoder-decoder, PD = prefix decoder, and AR = autoregressive)

Model	Size	LR	CG	AF	BS	Bias	NL	Optimizer	NAH	SHS	Dropout	MCLDT
GPT-4 [48]	1.8 T	–	CD	GeLU	–	Yes	120	Adam	120-150	20000	–	32768
GPT-3 [49]	175B	6×10^{-5}	CD	GeLU	32K-3200K	Yes	96	Adam	96	12288	–	2048
GPT-2 [50]	1.5B	1×10^{-4}	AR	GeLU	16K-64K	Yes	48	Adam	24	1280	0.1	1024
GPT-1 [45]	125M	1×10^{-4}	AR	GeLU	16K-64K	Yes	12	Adam	12	768	0.1	512
BARD [51]	340M	–	–	ReLU	64K	Yes	24	–	24	768	–	512
BERT [45]	340M	1×10^{-5}	–	ReLU	16K-64K	Yes	24	Adam	16	1024	0.1	512
PanGU-α [52]	207B	2×10^{-5}	CD	GeLU	–	Yes	64	Adam	128	16384	–	1024
BLOOM [53]	176B	6×10^{-5}	CD	GeLU	4000K	Yes	70	Adam	112	14336	0	2048
Galactica [54]	120B	7×10^{-6}	CD	GeLU	2000K	No	96	AdamW	80	10240	0.1	2048
OPT [55]	175B	1.2×10^{-4}	CD	ReLU	2000K	Yes	96	AdamW	96	12288	0.1	2048
Chinchilla [56]	70B	1×10^{-4}	CD	–	1500K-3000K	–	80	AdamW	64	8192	–	–
Falcon [57]	40B	1.85×10^{-4}	CD	GeLU	2000K	No	60	AdamW	64	8192	–	2048
T5 [58]	11B	1×10^{-2}	ED	ReLU	64K	No	24	AdaFactor	128	1024	0.1	512
LLaMA [59]	65B	1.5×10^{-4}	CD	SwiGLU	4000K	No	80	AdamW	64	8192	–	2048
LLaMA-2 [60]	70B	1.5×10^{-4}	CD	SwiGLU	4000K	No	80	AdamW	64	8192	–	4096
MT-NLG [61]	530B	5×10^{-5}	CD	–	64K-3750K	–	105	Adam	128	20480	–	2048
Jurassic-1 [62]	178B	6×10^{-5}	CD	GeLU	32K-3200K	Yes	76	–	96	13824	–	2048
Gopher [63]	280B	4×10^{-5}	CD	–	3000K-6000K	–	80	Adam	128	16384	–	2048
GLM-130B [64]	130B	8×10^{-5}	PD	GeGLU	400k-8250K	Yes	70	AdamW	96	12288	0.1	2048
LaMDA [65]	137B	–	CD	GeGLU	256K	–	64	–	128	8192	–	–
PaLM [66]	540B	1×10^{-2}	CD	SwiGLU	1000K-4000K	No	118	Adafactor	48	18432	0.1	2048

Different models have different batch sizes, with GLM-130B's larger batch size of 400k-8250K indicating enhanced training efficacy. In addition, the presence or absence of biased terms in models, such as Falcon, T5, LLaMA 1,2, and Galactica's "No," highlights the complexity of the choices made. From 12 for GPT-1 to 118 for PaLM, the number of layers affects a model's ability to capture intricate patterns. Optimizers are also diverse, with Adam, AdamW, and AdaFactor playing crucial roles. All GPT variants employ Adam as the optimizer, although models such as Galactica, OPT, and Falcon utilize AdamW as their optimizer. Both T5 and PaLM models utilize the Adafactor optimizer in their respective architectures. With 530 billion parameters, models like MT-NLG test scalability limits, while others like Chinchilla remain relatively compact. These variations highlight the significance of selecting models and configurations that are tailored to particular tasks, with performance, computational resources, and task requirements playing a central role.

The number of attention heads also exhibits variation across different models. GPT-1 is equipped with a total of 12 attention heads, whilst GPT-4 boasts a much larger number of attention heads, ranging from 120 to 150 within its model. The additional number of attention heads in the LLM enables the model to concurrently attend to several segments of the input sequence, hence expediting the model's training process. In order to enhance the efficacy of the LLMs, researchers employ diverse dimensions for the hidden states within their model. The larger dimensions of the hidden state enable the capturing of complex patterns within the text. Both GPT 4 and MT-NLG employ hidden state sizes of approximately 20,000, which is significantly greater in comparison to the hidden state sizes of other LLMs included in the table. Certain LLM models incorporate a dropout value of 0.1 to prevent overfitting issues, whereas others do not employ any dropout value. With 530 billion parameters, models like MT-NLG test scalability limits, while others like Chinchilla remain relatively compact. These variations highlight the significance of selecting models and configurations that are tailored to particular tasks, with performance, computational resources, and task requirements playing a central role.

6.3 Comparison Between Datasets of LLMs

Different LLM utilized different datasets for the training phase, distinguishing the models from one another. A concise overview of the datasets is provided in this section. The datasets used to train various large language models (LLMs) and their compatibility with each model are detailed in Table 7.

Table 7: Dataset for large language models

Dataset →	Webpages			Conversation Data	Books and News					Scientific Data		Code	
LLM ↓	C4	OpenWebText	Wikipedia	the Pile - Stack Exchange	BookCorpus	Gutenberg	CC-Stories-R	CC-NEWES	REALNEWs	the Pile - ArXiv	the Pile - PubMed Abstracts	BigQuery	the Pile - GitHub
T5 [58]	✓	✓	✓	X	X	X	X	X	X	X	X	X	X
Falcon [57]	✓	✓	✓	X	X	X	X	X	X	X	X	X	X
LLaMA [59]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GPT-3 [49]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	X	X
GPT-4 [48]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	X	X
MT-NLG [61]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gopher [63]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	✓	✓
Chinchilla [56]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	✓	✓
GLaM [69]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	X	X
PaLM [66]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	✓	✓
LaMDA [65]	✓	✓	✓	X	X	X	X	X	X	✓	✓	✓	✓
Galactica [54]	✓	✓	✓	X	X	X	X	X	X	✓	✓	✓	✓
GPT-NeoX [70]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CodeGen [71]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlphaCode [72]	X	X	X	X	X	X	X	X	X	X	X	✓	✓
Size	800GB	38GB	21GB	800GB	5GB	-	31GB	78GB	120GB	800GB	800GB	-	800GB
Source	CommonCrawl (April 2019)	RedditLinks (March 2023)	Wikipedia (March 2023)	Other (Dec 2020)	Books (Dec 2015)	Books (Dec 2021)	CommonCrawl (Sep 2019)	CommonCrawl (Feb 2019)	CommonCrawl (April 2019)	Other (Dec 2020)	Other (Dec 2020)	Codes (March 2023)	Other (Dec 2020)

Table 7 demonstrates that datasets have been divided into multiple categories: webpages, conversation data, literature and news, scientific data, and code. This classification enables us to comprehend the variety of data sources that contribute to LLM training. C4, OpenWebText, and Wikipedia are examples of datasets that belong to the "Webpages" category. At the same time, BookCorpus, Gutenberg, CC-Stories-R, CC-NEWES, and REALNEWS are examples of datasets that belong to the "Books and News" category. These categories reflect the richness and diversity of text data used to train LLMs, including web content, novels, news articles, scientific literature, and code.

From the ✓, it can be seen that LLaMA has been trained on a wide range of data sources, with significant exposure to webpages (87%), conversation data (5%), books and news (2%), scientific data (3%), and code (5%). This makes LLaMA a versatile model suitable for a wide array of natural language processing tasks that involve these data types. In contrast, platforms such as GPT-3 and AlphaCode have restricted data exposure. GPT-3 is proficient with web pages (84%), literature, and news (16%) but requires additional instruction with conversation data, scientific data, and code. AlphaCode, as its name suggests, is solely focused on code (100%) and does not utilize any other data sources. These findings highlight the significance of selecting the appropriate LLM based on the task's requirements. AlphaCode is the model of choice for code-related tasks, whereas LLaMA excels in diverse text data applications. In addition, the "Size" and "Source" columns of Table 7 offer additional context. The size of datasets ranges from 5GB (BookCorpus) to a massive 800GB (several datasets), indicating the sheer magnitude of data required to train these LLMs. The source information reveals when and where the data were collected, which is essential for comprehending the training data's temporal relevance and potential biases. Table 7 provides a multitude of information regarding the datasets used to train LLMs and how each model leverages these datasets. This information is invaluable for natural language processing researchers, developers, and practitioners, as it enables them to make informed decisions about which LLM to use for specific tasks and casts light on the breadth and depth of data that powers these cutting-edge language models.

7 Resources of Large Language Models

Large Language Models (LLMs) have a wide range of potential applications and resources available for their development, deployment, and utilization. In Figure 6, we present an LLM taxonomy that categorizes Large Language Models into two main branches: those based on pre-trained models and those based on APIs.

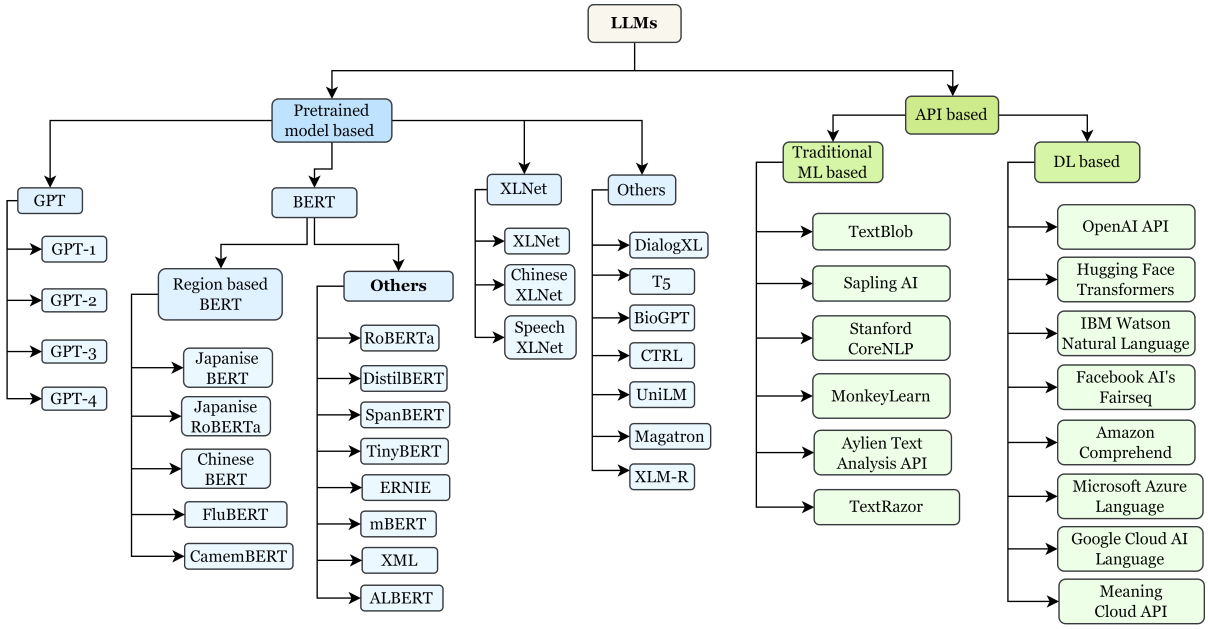


Figure 6: Taxonomy of LLM

This taxonomy allows for a comprehensive exploration of these two distinct aspects of Large Language Models. Here are some key resources presented which are associated with LLMs:

7.1 Pretrained Models

Pretrained language models play a pivotal role in natural language processing due to their ability to encapsulate broad language understanding and generation skills gleaned from diverse text sources. They offer a substantial advantage by minimizing the computational resources and data required for fine-tuning specific tasks. There are some of the most common pre-trained LLM models, which have been depicted in Table 8.

7.1.1 Generative Pretrained Transformer (GPT)

Generative Pre-trained Transformer [49] is an influential breakthrough in artificial intelligence, particularly in natural language processing (NLP). Developed by OpenAI, GPT leverages the Transformer architecture and extensive pre-training on vast internet text data to achieve a deep understanding of human language. This generative model excels at tasks like text generation, translation, question answering, and more, making it a versatile tool across various NLP domains. GPT's capacity to capture intricate language patterns and context, coupled with its iterative improvements, has profoundly impacted academia and industry, revolutionizing the landscape of language understanding and generation.

7.1.2 BERT

BERT [11], short for "Bidirectional Encoder Representations from Transformers," is a language model with a distinctive approach. Unlike previous models, BERT is designed to pre-train deep bidirectional representations from unlabeled text by considering both left and right context in all layers. This pre-trained BERT model can be fine-tuned with minimal adjustments to create cutting-edge models for various tasks like question answering and language inference, eliminating the need for extensive task-specific modifications. BERT is both conceptually straightforward and remarkably effective, achieving state-of-the-art results on eleven different natural language processing tasks. Notable accomplishments include raising the GLUE score to 80.5% (an impressive 7.7% absolute improvement), boosting MultiNLI accuracy to 86.7% (a 4.6% absolute improvement), and significantly improving SQuAD v1.1 question answering Test F1 to 93.2 (a 1.5 point absolute improvement) and SQuAD v2.0 Test F1 to 83.1 (a remarkable 5.1 point absolute improvement).

In our analysis, we have exclusively considered versions of BERT (Bidirectional Encoder Representations from Transformers) that are inherently Large Language Models (LLMs). Specifically, we focused on variants of BERT that are pre-trained on extensive text corpora and possess the characteristics of LLMs, enabling them to understand and generate natural language comprehensively. This deliberate choice ensures that the models we have included in

Table 8: Description of Language Models

Model Name	Description	Key Features	Training Data	Fine-Tuning Data	Fine-Tuning Tasks	Applications
GPT (Generative Pretrained Transformer) [49]	Transformative LLM by OpenAI for versatile NLP tasks.	Extensive pre-training, deep language understanding, iterative improvements, impact on academia/industry	Internet text data	Custom datasets	Text generation, translation, QA, and more	Chatbots, content generation, NLP domains
BERT (Bidirectional Encoder Representations from Transformers) [11]	Google AI's NLP model excelling with bidirectional context learning.	Deep bidirectional representations, conceptually straightforward, minimal task-specific adjustments	BookCorpus, Wikipedia	Task-specific datasets	Various NLP tasks	Question answering, language inference
RoBERTa [46]	BERT-based model with refined hyperparameters.	Significance of design decisions, publicly available, top-tier NLP results	BookCorpus, Wikipedia	Task-specific datasets	Various NLP tasks	Benchmark improvements, research
XLNet [73]	Combines autoregressive pretraining with bidirectional context learning.	Bidirectional context learning, versatile approach	Internet text data	Task-specific datasets	Diverse NLP tasks	Research, applications
Speech-XLNet [74]	Unsupervised acoustic model with robust regularization.	Robust regularizer, improved recognition accuracy	Speech datasets	TIMIT, WSJ datasets	Speech recognition	Speech recognition systems
DialogXL [75]	Improved dialogue handling with dialog-aware self-attention.	Enhanced conversation modeling, outperforms baselines	Internet text data	Dialogue datasets	Dialogue understanding	Chatbots, customer support
T5 (Text-to-Text Transfer Transformer) [47]	Google's unified text-to-text NLP model.	Unified framework, extensive pre-training, versatile tool	Internet text data	Task-specific datasets	Text classification, translation, and more	Language translation, summarization
BioGPT [76]	Specialized biomedical LLM with state-of-the-art results.	Biomedical literature pretraining, excels in biomedical tasks	Biomedical literature	Biomedical datasets	Biomedical text analysis	Biomedical text analysis, research

our study harness the full spectrum of language understanding and generation capabilities, thereby aligning with the core objective of our research in exploring the impact and advancements of LLMs in the field of natural language processing. Non-LLM versions of BERT or those with significantly reduced model sizes were excluded from our analysis to maintain consistency and relevance in our investigation of the transformative potential of Large Language Models.

7.1.3 RoBERTa

RoBERTA [46] is a study that replicates the BERT pretraining approach outlined by Devlin et al. in 2019. In this study, we meticulously assess the influence of various critical hyperparameters and training data sizes. It's worth noting that BERT was initially trained with room for improvement, yet it can now perform on par with or even surpass the performance of subsequent models that have been published. As a result, RoBERTa achieves top-tier results in GLUE, RACE, and SQuAD evaluations. These outcomes underscore the significance of design decisions that were previously overlooked and prompt inquiries into the origins of recently reported advancements. We have made our models and code available for public use.

7.1.4 XLNet

XLNet [73] represents a versatile autoregressive pretraining approach that achieves bidirectional context learning by optimizing expected likelihood across all possible permutations of factorization orders. It addresses the constraints of BERT through its autoregressive design and incorporates insights from Transformer-XL, a leading autoregressive model. In practical experiments with consistent conditions, XLNet consistently surpasses BERT on 20 diverse tasks, frequently by a substantial margin. These tasks encompass question answering, natural language inference, sentiment analysis, and document ranking, among others.

7.1.5 Speech-XLNet

Speech-XLNet [74] is a method for training unsupervised acoustic models to learn speech representations using a Self-Attention Network (SAN) and subsequently fine-tuning it within the hybrid SAN/HMM framework. Our hypothesis is that by rearranging the order of speech frames, the permutation technique in Speech-XLNet acts as a robust regularizer, encouraging the SAN to make inferences by prioritizing global structures through its attention mechanisms. Moreover, Speech-XLNet enables the model to explore bidirectional contexts, enhancing the effectiveness of speech representation learning. Experimental results on TIMIT and WSJ datasets demonstrate that Speech-XLNet significantly enhances the performance of the SAN/HMM system in terms of both convergence speed and recognition accuracy compared to systems trained from randomly initialized weights. Our best models achieve an impressive relative improvement of 11.9% and 8.3% on the TIMIT and WSJ tasks, respectively. Notably, the top-performing system achieves a phone error rate (PER) of 13.3% on the TIMIT test set, which, to the best of our knowledge, is the lowest PER achieved by a single system.

7.1.6 DialogXL

DialogXL [75] introduces improvements to handle longer historical context and multi-party structures in dialogues. Firstly, it modifies the way XLNet handles recurrence, moving from segment-level to utterance-level, which enhances its ability to model conversational data effectively. Secondly, it incorporates dialog-aware self-attention instead of the standard self-attention in XLNet, allowing it to capture important dependencies within and between speakers. A comprehensive set of experiments is conducted on four ERC benchmarks, comparing DialogXL with mainstream models. The experimental findings consistently demonstrate that our model surpasses the baseline models across all datasets. Additionally, we perform other experiments, including ablation studies and error analyses, which confirm the significance of DialogXL's critical components.

7.1.7 T5

T5 [47], or "Text-to-Text Transfer Transformer," is a groundbreaking large language model developed by Google Research, revolutionizing natural language processing (NLP). T5's innovation lies in framing all NLP tasks as text-to-text tasks, simplifying the NLP pipeline and unifying various tasks under a single framework. Built upon the Transformer architecture, T5 utilizes multi-head self-attention to capture intricate language relationships. Its extensive pre-training on vast text data, followed by fine-tuning on specific tasks, empowers T5 to excel in text classification, translation, summarization, question answering, and more. With consistently state-of-the-art results across NLP benchmarks, T5 has reshaped the field, offering researchers and developers a versatile tool for comprehensive language understanding and generation tasks.

7.1.8 BioGPT

BioGPT [76] is a specialized Transformer-based language model, pre-trained using extensive biomedical literature. We conducted evaluations across six biomedical natural language processing tasks and found that our model consistently outperforms previous models in most cases. Notably, we achieved F1 scores of 44.98%, 38.42%, and 40.76% on BC5CDR, KD-DTI, and DDI end-to-end relation extraction tasks, respectively, and set a new accuracy record of 78.2% on PubMedQA. Furthermore, our investigation into text generation highlights BioGPT's proficiency in generating coherent descriptions for biomedical terms within the literature.

In summary, pre-trained LLMs are foundational in NLP, providing a starting point for various applications without the need for extensive training from scratch. They are widely used and have democratized access to advanced language understanding and generation capabilities. However, responsible use and ethical considerations are essential when working with these models to ensure fair and unbiased outcomes.

7.2 API of LLM

In this section, we discuss the APIs of LLMs, which have been described in Table 9.

Table 9: Comparison of LLM APIs

API Name	**Provider**	**Languages Supported**	**Access Type**	**Application Area**	**Advantages**	**Constraints**
OpenAI API [77]	OpenAI	Multiple languages	API Key	NLP, text generation, chatbots	State-of-the-art models, versatility, GPT architecture	API rate constrain, cost considerations
Hugging Face Transformers [78]	Hugging Face	Multiple languages	Open Source	NLP, model fine-tuning, research	Large model repository, extensive community support	Self-hosting complexity, no official support
Google Cloud AI-Language [79]	Google Cloud	Multiple languages	API Key	Sentiment analysis, entity recognition, translation	Google's robust infrastructure, easy integration	Cost may vary based on usage
Microsoft Azure Language [80]	Microsoft Azure	Multiple languages	API Key	Sentiment analysis, entity recognition, language understanding	Integration with Azure services, comprehensive APIs	Pricing based on usage
IBM Watson NLU [81]	IBM Watson	Multiple languages	API Key	Sentiment analysis, emotion analysis, keyword extraction	IBM's AI expertise, customization options	Costs may add up for high usage
Amazon Comprehend [82]	Amazon AWS	Multiple languages	API Key	Entity recognition, sentiment analysis, topic modeling, document classification	Integration with AWS, scalability	Costs may vary based on usage
Facebook AI's Fairseq [82]	Facebook AI	Multiple languages	Open Source	Neural machine translation, language modeling, research, development	Research-oriented, flexibility, open-source.	Self-hosting and maintenance complexity.

Open AI API: The API provided by OpenAI offers access to GPT models that may be utilized for a wide range of text-related applications [83]. It facilitates many tasks such as coding, question and answer, analysis, and other related activities. The available models encompass a spectrum of options, spanning from gpt-4 to gpt-3.5-turbo, as well as many legacy variants. The Chat Completions API facilitates interactive dialogues by incorporating distinct roles such as user, and assistance. The programming language provides support for function calling, which allows for the retrieval of structured data. The OpenAI API provides developers with the capability to leverage advanced modeling of languages for a diverse range of applications.

Hugging Face: Hugging Face provides a complimentary Inference API that facilitates the examination and assessment of more than 150,000 publicly available ML models [84]. It features predictive capabilities, and integration with more than 20 open-source libraries, and facilitates fast change between models. The API facilitates a range of operations, including classification, image segmentation, text analysis, speech recognition, and other related functionalities.

Google Cloud API: The Cloud-based NLP API developed by Google provides support for a range of approaches, such as sentiment analysis, text analysis, entity recognition, and other text annotations [79]. The functionalities can be accessed by developers through REST API calls utilizing either the client libraries or their own custom libraries. Additionally, it offers moderation functionalities for the purpose of detecting potentially sensitive content. Several API exists, and each possesses distinct features and functions.

Microsoft Azure Language APIs: These APIs support many activities, including sentiment analysis, text summarization, and other related tasks [80]. Developers use RESTful endpoints to include Azure LLM APIs. Microsoft provides useful SDKs and code examples in other programming languages, including Python, Java, etc. to facilitate the utilization of these APIs.

IBM Watson Natural Language: The IBM Watson API is a robust tool for investigating and extracting valuable information from textual data. This API offers developers a variety of functionalities, encompassing sentiment analysis, emotion analysis, and additional features [81]. Due to its provision of multilingual support and a user-friendly API, this technology enables developers to effectively include sophisticated text analytics into their programs.

Amazon Comprehend API: The Amazon Comprehend API is a powerful NLP service provided by Amazon Web Services [82]. This tool evaluates textual data, allowing the researchers to acquire significant knowledge, such as entity recognition, language detection, sentiment analysis, and topic modeling. Due to its ability to accommodate many languages and simple integration, this tool displays adaptability in addressing a range of use cases, including customer feedback analysis and others. The utilization of this API can prove to be a significant resource for enterprises' marketing to extract practical insights from unstructured textual data.

Facebook AI's Fairseq: The Fairseq framework developed by Facebook AI is a comprehensive tool for performing sequence-to-sequence modeling, specifically designed for handling LLMs [85]. Fairseq is a well-suited API for many applications related to analyzing and generating natural language. The platform provides support for advanced models such as BERT and RoBERTa, allowing researchers to perform fine-tuning on these models according to specific needs.

In this study, we have provided a comprehensive overview of seven popular APIs in Table 9 that leverage the capabilities of LLMs for the purpose of NLP-based functionalities. However, the taxonomy revealed the presence of several other APIs that are associated with text analysis but do not utilize LLMs. The aforementioned APIs, including TextBlob, TextRazor, Sapling AI, MonkeyLearn, and Aylien, etc., utilize traditional machine learning, statistical methods, and rule-based natural NLP techniques instead of relying on extensive pre-trained LLMs. Since, the primary focus of this study has been on describing the tools that particularly utilize LLMs for the purpose of advanced text analysis, generation, and comprehension, we have refrained from discussing these APIs in depth.

8 Domain Specific Application

Since there are several pre-trained models in LLMs, all of them are utilized by training or fine-tuned to perform well-defined tasks maintained by their requirements in different fields. Numerous research studies have consistently contributed by using LLMs model in diverse domains such as healthcare, finance, education, forecasting, and natural language processing. The extensive experiments of different LLM models contribute to revolutionizing the use of AI across these diverse domains. This section demonstrates the potential contribution of LLMs application in different domains. Table 10 illustrates the major contribution of LLMs in the specific domain, as well as outline their prospective limitations and future directions.

Bio-Medical and Healthcare: As previously stated, GPT has several versions, ranging from GPT1 to GPT4. GPT3 is extremely useful in the healthcare industry since it can be trained to support customer service with no effort. It can get all required information through a conversation rather than an intake form, and many systems might be built

Table 10: Machine learning-based study comparison in LLMs

Domain	Author	Major Contributions	Limitations	Future Research Direction
Medical	Chen et al. [86] (2023)	I. Assess the state-of-the-art performance of biomedical LLMs for the purpose of classifying and reasoning tasks on clinical text data. II. Emphasizes the vulnerability of LLM performance in relation to prompts and addresses it.	I. Data limitation due to privacy concern of biomedical data. II. Did not evaluate the performance of the model in an out-of-domain task.	I. To support this study's findings, need to experiment using real clinical data. II. Optimize the models to make them more robust and resource-efficient.
	Huang et al. [87] (2023)	I. Investigates the possible utilization of LLMs, specifically ChatGPT and its variety within the domain of dentistry. II. Design a MultiModal LLM system for clinical dentistry application and address critical challenges to revolutionize dental diagnosis.	I. Lack of data resulted in the post-training process, raising concerns about the model's reliability. II. The possibility of data breaches has no strict security method. III. Requires intensive computational cost.	I. Reducing operational costs by fine-tuning the model and enhancing efficiency. II. Explore diverse medical data to provide personalized dental care.
	Sorin et al. [88] (2023)	I. Evaluating the efficacy of ChatGPT-3.5 as a supporting tool for facilitating clinical decision-making in breast tumor cases. II. Outlines the implementation of a grading system for evaluating the responses generated by ChatGPT.	I. Conducting the experiment with a small sample size leads to performance bias in the model. II. Human errors in the grading system can potentially add biases to the system.	I. More diverse sample of breast tumor cases to increase ChatGPT's performance and generalizability. II. Introducing a multimodal approach to increase the reliability of clinical recommendations.
	Thirunavukarasu et al. [89] (2023)	I. Focuses on the energy and environmental impact of training LLM models such as GPT-3 and GPT-4 and emphasize cost reduction to make them more accessible. II. Examines the utilization of LLMs models in the medical domain, specifically focusing on medical education and medical research.	I. Inaccuracies observed in the responses provided to queries due to the lack of updates on the training data. II. Lack of interpretability of LLMs model since it is a black box, hence the concept was frequently misunderstood.	I. Emphasis on integrating more recent and up-to-date training data. II. Further investigation should strive to enhance the transparency and interpretability of LLMs. III. Including the feasibility of implementing randomized trials to evaluate the effects of LLM on medical outcomes.
	Korngiebel et al. [90] (2021)	I. Discuss the benefits and potential pitfalls of NLP technologies in eHealth. II. Discuss the benefits of using GPT in the medical domain.	I. Conversational AI like GPT-3 will not replace human interaction in healthcare soon, despite extensive development. II. Examines GPT's applicability in a certain medical domain.	I. Analyze GPT's impact on real-world healthcare settings to assess its performance. II. Provide personalized healthcare by analyzing a variety of medical data.
	Angelis et al. [91] (2023)	I. examine LLMs' ethical and practical issues, focusing on medicinal use and public health. II. Discuss how ChatGPT can provide false or misleading information. III. Suggest the detectable-by-design technique to spot fake news or information.	I. The addition of a detectable-by-design the technique may slow LLM development and AI business acceptance. II. Experimental data has been limited due to medical data privacy concerns.	I. An experiment using real clinical data is needed to support the findings. II. Further research should be conducted to speed up the entire procedure.
	Sallam et al. [92] (2023)	I. Saves time in scientific research through code delivery and literature review. II. Makes the publication process faster by providing better research ideas and results. III. Reduces potential costs and increases efficiency in healthcare delivery. IV. Enhances communication skills in healthcare education through proper academic mentoring.	I. Copyright issues, bias based on the training dataset, plagiarism, over-detailed content, lack of scientific accuracy, limited updated knowledge, and lack of ability to critically discuss the results in using ChatGPT in scientific research. II. Unable to understand the complexity of biological systems, lack of emotional and personal perspective, inaccurate content, bias, and transparency issues in healthcare practice. III. Copyright issues, inaccurate references, limited updated knowledge, and plagiarism in healthcare education.	I. Accountability, honesty, transparency, and integrity must be considered in scientific research. II. To enhance healthcare and academics, ChatGPT should uphold ethical principles. Potential dangers and other issues must also be considered. III. An AI editor and an AI reviewer in academic writing to advance academic research, given the previous shortcomings of the editorial and peer review process.
	Cascella et al. [93] (2023)	I. Support of clinical practice II. Scientific writing	I. Generates answers that sound plausible but may be incorrect or meaningless and biased based on trained data.	I. Enhance the ability to answer medical questions and provide the context for understanding complex relationships between various medical conditions and treatments.
	Kung et al. [94] (2023)	I. The investigation of AI within the context of medical education. II. Assessment of ChatGPT's Performance in Clinical Decision-making. III. Explore the demands of AI in medical education to standardize methods and readouts and quantify human-AI interactions	I. The experiment is conducted on a small input size. II. Human adjudication variability and bias. III. The absence of real-life instructional scenarios.	I. To evaluate the efficacy of ChatGpt in real-world clinical practice by assessing its performance and impact. II. A comprehensive analysis of ChatGPT's effectiveness in relation to subject taxonomy.
	Gu et al. [95] (2021)	I. Shows that domain-specific pretraining from scratch outperforms mixed-domain in biomedical NLP. II. Formulate a new dataset using the Biomedical set of diverse tasks.	I. Explore the applicability only in a fixed Biomedical Domain. II. Future modifications of the benchmark may be required to reflect the effectiveness of the research.	I. An Investigation and analysis into pretraining strategies. II. The addition of Biomedical NLP tasks. III. Exploring other domains for comparative analysis.
	Kraljevic et al. [96] (2022)	I. Introduced a foresight application based on electronic health records. II. Develop a multifunctional model. III. Conduct experiments in different hospitals.	I. Should include metrics, and comparative analysis in real-world clinical scenarios to evaluate Foresight's performance. II. Integrate enough security on health records to protect the privacy of the patients.	I. Integrating input from healthcare specialists and consistently updating the model with the latest medical data. II. Implement a real-life scenario to investigate the clinical application of Foresight.
Tourism	Mich et al. [97] (2023)	I. Highlights how ChatGPT is contributing to the tourism sector by identifying new target markets, implementing the marketing strategy designs, and improving customer service.	I. Transparency and accountability issues: the dataset is not updated, and can not see the logic of what is wrong and what is right.	I. Applications should increase user trust and fact-checking.

Table 10: *(Continued)* Machine learning-based study comparison in LLMs

Domain	Author	Major Contributions	Limitations	Future Research Direction
Industry	Yu et al. [98] (2023)	I. Examines how LLMs can use their superior knowledge and reasoning to predict financial time series. II. Focuses on NASDAQ-100 stocks using publicly available historical stock price data. III. To prove LLMs can solve problems comprehensively, experiments are conducted.	I. The study utilizes a small amount of data samples. II. Data is collected from only one specific domain. III. Utilizing a small sample size during experiments cause performance bias.	I. SP500 and Russell 2000 stock indexes will be added to the research. II. The research will use macro-economy time series, stock trading volumes, and social network data. III. To improve reasoning, larger public models like 30B will be refined.
	Frederico et al. [99] (2023)	I. Discusses the uses and concerns with ChatGPT in supply chains. II.Provide supply chain specialists advice about ChatGPT's effects and usage.	I.A limited amount of data is used in the experiment. II. Did not assess the efficacy of ChatGPT in practical industrial settings.	I. Analyze how ChatGPT can enhance the supply chain efficiency. II. Discuss supply chain ChatGPT implementation issues and success factors.
Gaming	Sobieszek et al. [100] (2022)	I. Examines the efficacy of employing LLM as a gaming tool. II. Assess the performance of GPT in the context of the Turing test. III. Analyze the boundaries of LLMs. IV. Discuss the challenges these models encounter in accurately conveying information.	I. They did not employ a well-curated set of targeted questions. II. It may produce answers that are either erroneous or lack significance.	I. Assess the performance of LLM by administering inquiries across diverse domains.
Education	Abramski et al. [42] (2023)	I. Utilized network science and cognitive psychology to study biases toward math and STEM across language models. II. Behavioral Forma Mentis Networks (BFMNs) are used to understand how LLMs comprehend arithmetic, STEM, and similar concepts.	I. Commercial GPT systems can be tested by researchers but not replicated by everyone due to their structure. II. The old interface or API system no longer allows public access to GPT-3.	I. Putting a priority on integrating data from training that is up-to-date. II. Investigating several other fields for the purpose of comparative research. III. More information from students at different institutions will be gathered.
	Kasneci et al. [20] (2023)	I.Helps students develop critical thinking in reading and writing, provides practice problems and quizzes, helps improve research skills, and improves various developmental skills. II.Provides guidance to teachers on how to improve student learning in each aspect of teaching and helps develop teaching materials.	I. Helpful only for English-speaking people, but also for people of other languages cannot enjoy the benefits. II.Consumes high energy and financial cost of maintenance. III. Negative effect on critical thinking and problem-solving skills of students and teachers. IV.Privacy and security risks to students' personal and sensitive information.	I. Creating an age-appropriate user interface that maximizes the benefits and minimizes the pitfalls of interaction with AI-based tools. II. To guarantee equity for all educational entities interested in current technologies, government organizations may regulate financial obstacles to accessing, training, and maintaining large language models.
	Hadi et al. [101] (2023)	I. Helps students save labor and time by assigning assignments and helps teachers automate the grading process, and provides detailed feedback to students, which reduces their workload. II.Aid decision-making, problem-solving and promote learning in medical education. III.Provides financial advice based on their queries to improve customer service, and provides various steps based on financial algorithms to reduce risk by analyzing past market data. IV. Saves software engineers time and increases overall efficiency by providing code snippets, identifying and generating test cases, etc.	I. Bias, reasoning errors, counting errors, information hallucination, LLMs explainability.	I. Improving the accuracy and performance of LLMs, addressing their limitations, and exploring new ways to utilize them.
	Lo et al. [102] (2023)	I. Helps students in learning and assessment and helps teachers in teaching preparation and assessment.	I. Negative effect on critical thinking and problem-solving skills of students and teachers.	I. Training instructors on how to effectively use ChatGPT and identify student intelligence. Also, educate students about the uses and limitations of ChatGPT.
	Dwivedi et al. [103] (2023)	I. Highlights the challenges, opportunities, and impacts of ChatGPT in education, business, and society, as well as investigates important research questions asked of ChatGPT across the education, business, and society sectors.	I. The generated text is hard to understand and can't answer questions correctly unless phrased a certain way, lacks updated information, and doesn't automatically update the actual data.	I. Teaching, learning, and scholarly research, digital transformation organization and society, knowledge, transparency, and ethics to enhance ChatGPT's efficiency in all these areas.

to assist numerous patients at the same time [90]. Besides, clinics and hospitals are places to cure illness, but it is also true that various contagious viruses are brought into these places. Patients and healthcare providers can be better protected from infection by replacing a human receptionist with a robot. This becomes increasingly important during the COVID-19 epidemic [104]. Since clinics and hospitals often see a high volume of patients on a daily basis, an optimum and lightweight system may submit several queries for single patients to create acceptable output. Consequently, GPT models can also aid in cost reduction in the medical industry. Furthermore, biomedical and clinical text mining has always been an essential and major challenge due to the complex nature of domain corpora and the continually expanding number of documents. As a result, using the BERT models improves the performance of biomedical and clinical text mining models [105]. Salam et al. [92] and Korngiebel et al. [90] demonstrate the substantial advantages of ChatGPT in the domains of healthcare, clinical research, and practice, although simultaneously underscoring the imperative necessity for proactive inspection and ethical transparency. Several studies [93, 95, 96, 89] investigations at exploring the prospective utilities and constraints of LLMs such as ChatGPT within the healthcare domain, namely in the context of clinical practice, research, and public health. In their study, Kung et al. [94] conducted an evaluation of ChatGPT's performance on the United States Medical Licensing Examination (USMLE), and the outcomes indicate the potentiality of LLMs to support clinical decision-making and medical education. Sorin et al. [88] evaluated ChatGPT-3.5 as a decision support for breast tumor boards where they compared the tumor board's explanations, and summaries with ChatGPT-3.5 and showed that ChatGPT-3.5 and the tumor board had a high degree of decisional alignment. Huang et al. [87] (year) investigate the prospective applications of LLMs with a specific emphasis on ChatGPT, in the field of dentistry, mainly focusing on automated dental diagnosis and highlighting the efficacy of LLMs in dental diagnosis. Furthermore, the XLNet contributes to better clinical note representation by adding temporal information and a realistic prediction setup [106]. Furthermore, various LLM models also assist this medical industry by making the procedure easier than previously.

Education: Educators have long struggled with unequal educational resources to student demand across disciplines. One of the significant challenges is a shortage of accessible educational resources for pupils to study outside of school. Although online instructional videos are helping to alleviate the problem, society still hopes that AI will deliver individualized teaching services to satisfy the learning demands of each student and increase teaching efficiency. So, the LLM models are very significant and have the potential to revolutionize many facets of learning, teaching, and educational research in the education sector [104]. So, the GPT model aids the students in converting the math word problems into representative equations [107]. Kasenci et al. [20] highlighted substantial impact of LLMs in education by facilitating personalized learning, automating grading process, and accessibility of educational resources. Hadi et al. [101] presents a thorough analysis of LLMs, covering their historical development, wide-ranging applications in domains such as medicine, engineering, education, and their potential impact on the trajectory of AI. Lo et al., [102] and Dwivedi et. al. [103] investigate the prospective uses of ChatGpt within the realm of education and identify the primary obstacles that have arisen during its initial deployment. Besides, in terms of writing authentic texts in distinct formats, including essays, summaries, and articles, these models help to accomplish this without any error. In contrast, the manual process may have human errors in the documentation. In this case, the GPT model helps to address this problem. In addition, the XLNet Excel method also helps understand the texts and documents that can be employed in the education sector [39]. Furthermore, other models significantly impact the education system, making it more engaging, accessible, and productive for both students and teachers.

Social Media: The LLMs have revolutionized several aspects of the social media industry regarding content production, moderation, sentiment analysis, etc. There are some crucial aspects of the LLMs in the social media sector in terms of writing content, generating images, classifying and generating text, and even full blogs and articles for social media. Also, these models can perform named entity recognition (NER) and text classification [108, 109]. When the GPT, XLNet, BERT, etc., model aids the writer and content producers in generating a consistent flow of excellent material. It also provides content suggestions, and to create a safer online environment, these models are hired to assist in discovering and filtering out different dangerous and improper content. In their study, Abramski et al. [42] utilized network science and the principles of cognitive psychology to evaluate biases present in LLMs. Sobieszek et al. [100] presents a critical examination of the stated semantic capabilities of GPT-3, aiming to challenge the current view of its dismissal. Moreover, it assists in determining public opinion on certain topics by analyzing public interest and demand.

Business: In business, LLM helps companies improve their decision-making processes, product manufacturing processes, operations, and customer interactions. Communicating with customers and providing 24/7 customer service by answering their queries, assisting them in their work, and providing advanced advice related to areas of interest to customers is crucial for business progress. Moreover, it is also important to analyze customer sentiment, market trends, risk factors, and competitive intelligence [**?**]. In this case, LLMs help to fulfill all their requirements within a short period. The LLM models, like GPT, XLNet, BERT, etc., play a vital role in creating customer documents and product details and efficiently maintaining the entire business by saving time and reducing laborious tasks. Frederico et al. [99] presents an initial investigation into the potential applications and effects of ChatGPT in the domain of supply chain management. Their study provides significant insights for professionals engaged in this domain. Mich et. al. [97] present an initial investigation of potential hazards associated with the implementation of ChatGPT in bussiness domain. Yu et al. [98] presented an analysis of the capabilities of LLMs, specifically GPT-4, in the context of financial forecasting for a time series. Besides, their findings reveal that the performance of LLMs outperforms other traditional models also.

Agriculture: In agriculture, variations of GPT models, including GPT3, BERT, and XLNet models, play a significant role [110, 111]. They are able to analyze large data hubs of soil, crop, and weather data along with satellite imagery. They can provide recommendations on plating times, irrigation, fertilizer application, and optimizing fields and resources. Farmers can obtain current updates and market requirements, predict crop prices, anticipate natural disasters, and document farmers' and crop details. Manual agriculture management can be time-consuming and laborious, but these models can handle all the issues.

9 Impact of Large Language Models on Society

Large Language Models (LLMs) and similar AI technologies have had a profound impact on society across various domains. While these technologies offer many benefits, they also raise important ethical, social, and economic considerations. Here's an overview of the impact of LLMs on society:

a. Advancements in Natural Language Processing (NLP): LLMs have significantly advanced the field of NLP, making it possible to automate and scale a wide range of language-related tasks such as translation, summarization, sentiment analysis, and more. In recent years, Natural Language Processing (NLP) has witnessed significant advancements, primarily driven by the emergence of Large Language Models (LLMs). These advancements, exemplified by models

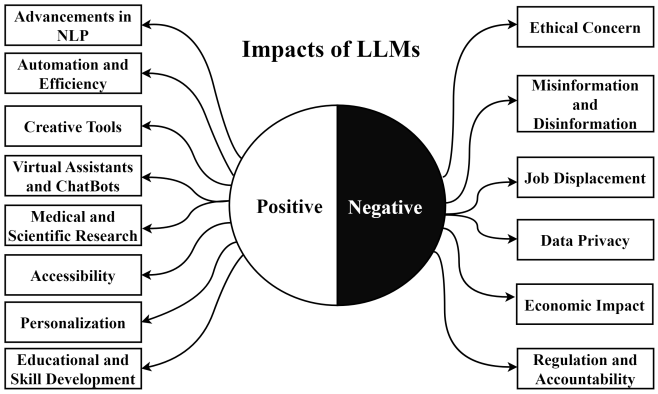


Figure 7: Visual representation of impact on LLMs

such as BERT [11], RoBERTa [46], and XLNet [73], have transformed the NLP landscape. Notably, LLMs have been fine-tuned for various specific NLP tasks, enabling remarkable performance improvements. Multilingual models like mBERT [112] and cross-lingual models like XLM-R [113] have facilitated language understanding across diverse linguistic contexts. Additionally, there has been a focus on creating more efficient versions of LLMs such as DistilBERT [114] and ALBERT [115]. These developments have not only expanded the applicability of NLP but have also raised ethical considerations, prompting research in bias mitigation [116] and responsible AI. LLMs have enabled breakthroughs in applications like conversational AI, few-shot and zero-shot learning, and domain-specific NLP in fields like healthcare and finance. These advancements underscore the pivotal role of LLMs in advancing the capabilities of NLP and continue to shape the future of language understanding and generation.

b. Automation and Efficiency: LLMs are used to automate tasks that were previously time-consuming and labor-intensive, leading to increased efficiency in industries such as customer support, content generation, and data analysis. The automation and efficiency of Large Language Models (LLMs), driven by models like BERT and GPT, have revolutionized industries and applications. These models have automated intricate language-related tasks, from sentiment analysis to language translation, making them more efficient and accessible. LLMs, such as DialoGPT [117] and ChatGPT, have powered conversational AI, streamlining customer support and interactions. Moreover, they excel in few-shot and zero-shot learning, as demonstrated by GPT-3 [118], automating tasks with minimal examples. Multilingual LLMs like mBERT have automated language tasks across various languages, enhancing global accessibility. Efficiency has further advanced through models like DistilBERT and ALBERT, which maintain performance while reducing computational resources. These models can be fine-tuned for specific domains, such as healthcare [119], making them indispensable in automating domain-specific tasks efficiently.

c. Content Generation: LLMs are capable of generating human-like text, which has implications for content creation, including automated news articles, marketing materials, and creative writing.

d. Language Translation: LLMs have improved machine translation systems, making communication across languages more accessible and accurate.

e. Virtual Assistants and Chatbots: LLMs power virtual assistants and chatbots, enhancing customer service and providing round-the-clock support in various industries.

f. Medical and Scientific Research: LLMs are used to analyze and summarize vast amounts of medical and scientific literature, aiding researchers in finding relevant information quickly.

g. Accessibility: LLMs have the potential to improve accessibility by providing real-time translation and transcription services for individuals with hearing impairments or language barriers.

h. Personalization: LLMs enable personalized recommendations and content curation on platforms such as social media, e-commerce, and news websites.

i. Creative Tools: LLMs are used as creative tools in various art forms, including generating poetry, music, and visual art.

j. Ethical Concerns: Bias and fairness issues in LLMs have raised ethical concerns. LLMs may perpetuate or amplify biases present in training data, leading to unfair or discriminatory outcomes.

k. Misinformation and Disinformation: LLMs can generate realistic-sounding fake text, raising concerns about the spread of misinformation and disinformation.

l. Job Displacement: The automation capabilities of LLMs may lead to job displacement in certain industries, particularly in routine data-entry and content-generation roles.

m. Data Privacy: The use of LLMs often involves processing large amounts of user-generated text data, which raises data privacy concerns, especially regarding sensitive or personal information.

n. Economic Impact: The adoption of LLMs can disrupt traditional business models and create economic shifts as industries adapt to automation and AI technologies.

o. Regulation and Accountability: Policymakers and regulators are grappling with the need to establish guidelines and regulations for the responsible use of LLMs, including addressing issues of bias, transparency, and accountability.

p. Education and Skill Development: The rise of LLMs underscores the importance of education and skill development in AI and data science, as these technologies become increasingly integral to various industries.

The impact of LLMs on society is multifaceted, and it is important to consider both the positive and negative consequences. As these technologies continue to evolve, stakeholders, including governments, businesses, researchers, and the general public, must work together to harness the benefits of LLMs while addressing their challenges and ethical implications. The visual representation of Figure 7 effectively demonstrates the impact of LLMs, outlining their benefits on the left and the adversarial impacts on the right side. The utilization of this figure will provide a distinct and easily understandable visual depiction of LLMs' impact across different domains.

10 Open issues, Challenges, Future works

This section discusses critical analysis of open issues, challenges, and LLMs' future scope.

10.1 Open Issues

In this section, we delve into the critical open issues surrounding LLMs. These concerns are at the vanguard of artificial intelligence research and development. They emphasize the need for ongoing research and innovation to resolve issues that have emerged alongside the rapid development of LLMs. Our discussion will cast light on the significance of these unresolved issues, highlighting their impact on various applications and the AI landscape as a whole.

- **Issue 1: Ethical and Responsible AI**
  The question regarding how to ensure the ethical use of large language models remains unresolved. Filtering, moderation, and accountability concerns regarding AI-generated content remain troublesome. Misinformation, hate speech, and biased content generated by LLMs necessitate continuous research and development [120].
- **Issue 2: Multimodal Integration**
  While LLMs are predominantly concerned with text, there is a growing demand for multimodal models that can comprehend and generate content that includes text, images, and other media types [121]. Integrating multiple modalities into a single model poses difficulties in data acquisition, training, and evaluation.
- **Issue 3: Energy Efficiency**
  The environmental impact of training and deploying large language models is still an urgent concern [122]. It is essential to develop more energy-efficient training methods, model architectures, and hardware solutions to reduce the carbon footprint of LLMs.
- **Issue 4: Security and Adversarial Attacks**
  LLMs are vulnerable to adversarial assaults, where slight input modifications can lead to unexpected and potentially harmful outputs [123]. Improving model robustness and security against such assaults is a crucial area of study, particularly for cybersecurity and content moderation applications.
- **Issue 5: Privacy and Data Protection**
  As LLMs become more competent, user privacy and data protection concerns increase. Finding methods for users to interact with these models without compromising their personal information is an ongoing challenge. There is a need for research on privacy-preserving techniques and regulatory compliance [124].
- **Issue 6: Generalization and Few-Shot Learning**
  LLMs excel when there is abundant data but struggle with tasks requiring few examples or domain-specific knowledge. Improving their capacity to generalize and perform well with limited training data is a crucial area of research [125].

- **Issue 7: Cross-Lingual and Low-Resource Settings**
  It is an ongoing challenge to make LLMs more accessible and effective in languages and regions with limited resources and data [126]. Global applications require developing techniques for cross-lingual transfer learning and low-resource language support.

10.2 Challenges

LLMs have rapidly evolved from being non-existent to becoming a ubiquitous presence in the field of machine learning within just a few years. Their extraordinary ability to generate text that resembles that of a human has garnered significant attention and applications in numerous fields. However, this meteoric rise in prominence has also revealed many challenges and concerns that must be addressed to realize the potentiality of these models fully. In this discussion, we will examine ten of the most significant challenges pertaining to LLMs.

- **Challenge 1: Data Complexity and Scale**
  In the era of LLMs, the size and complexity of the datasets on which they are trained is one of the most significant challenges. These models are typically trained on enormous corpora of Internet-sourced text data. These datasets are so extensive that it is nearly impossible to comprehend or investigate the totality of their information. This raises concerns regarding the quality and biases of the training data and the potential for the unintentional dissemination of detrimental or inaccurate information.
- **Challenge 2: Tokenization Sensitivity**
  For analysis, LLMs rely significantly on tokenization, dividing text into smaller units (tokens) [127]. Tokenization is essential for language processing and comprehension but can also present challenges. For instance, the meaning of a sentence can alter significantly based on the choice of tokens or the ordering of words. This sensitivity to input phrasing can lead to unintended outcomes when generating text, such as adversarial assaults and output variations based on minute input changes.
- **Challenge 3: Computational Resource Demands**
  The training of LLMs is a computationally intensive procedure that requires substantial hardware and energy resources. It is necessary to have access to supercomputing clusters or specialized hardware in order to train large models, and the environmental impact of such resource-intensive training has raised concerns. Significant energy consumption is associated with training LLMs at scale, contributing to the AI industry's overall carbon footprint.
- **Challenge 4: Fine-Tuning Complexity**
  While pre-training gives LLMs a broad comprehension of language, fine-tuning is required to adapt these models to specific tasks. Fine-tuning entails training the model on a smaller dataset, frequently requiring human annotators to label examples. As it involves the construction of task-specific datasets and extensive human intervention, this process can be both time-consuming and costly.
- **Challenge 5: Real-Time Responsiveness**
  The remarkable training capabilities of LLMs come at the expense of inference speed. Real-time response or prediction generation with these models can be sluggish, limiting their applicability in applications such as chatbots or recommendation systems where low-latency responses are crucial for user satisfaction.
- **Challenge 6: Contextual Constraints**
  LLMs can only evaluate a limited number of preceding tokens when generating text due to their limited context window. This limitation presents difficulties when working with lengthy documents or having lengthy conversations. Maintaining coherence and relevance over lengthy text sequences can be challenging because the model may neglect or lose track of pertinent information.
- **Challenge 7: Bias and Undesirable Output**
  In their output, LLMs can display biases or undesirable characteristics. This is due to the inherent biases in the training data, which are assimilated by the model and reflected in its responses. Such biases can manifest as objectionable, discriminatory, or harmful content, making it imperative to address and mitigate these concerns to ensure the responsible deployment of AI.
- **Challenge 8: Knowledge Temporality**
  LLMs learn using historical data from the Internet, and their knowledge is restricted to what is available as of a particular date. Consequently, they may lack access to the most recent information or events. This can be problematic when users expect up-to-date responses or when the conversation involves recent events.

- **Challenge 9: Evaluation Complexity**
  Evaluation of LLMs presents significant difficulties. Many extant evaluation metrics are insufficient to capture the nuances of model performance, which raises questions about their efficacy. Additionally, these metrics can be susceptible to manipulation or gaming, which may provide an inaccurate image of a model's capabilities. To assess LLMs' actual performance and limitations, robust and reliable evaluation methodologies are required.
- **Challenge 10: Dynamic Evaluation Needs**
  Frequently, evaluating LLMs entails comparing their outputs to static benchmarks or human-authored ground truth. However, language is dynamic and evolves, and preset evaluation data may not adequately reflect a model's adaptability to language and context change. This difficulty underscores the need for evaluation frameworks that are more dynamic and continually updated.

10.3 Future Works

Emerging in the rapidly evolving landscape of LLMs are several key research foci and directions that will shape the future of these robust AI systems. Improving Bias Mitigation involves refining training data to minimize bias, developing effective debiasing techniques, establishing guidelines for responsible AI development, and integrating continuous monitoring and auditing mechanisms into AI pipelines to guarantee fairness and impartiality.

Another essential concern is efficiency, which has prompted research into more efficient training techniques. This includes exploring innovative techniques such as federated learning to distribute training across decentralized data sources, investigating knowledge distillation methods for model compression, and discovering ways to reduce the substantial computational and environmental costs associated with LLMs. Dynamic Context Handling is crucial for enhancing the capabilities of LLMs. This involves enhancing their context management so that they can comprehend lengthier context windows and handle lengthy documents or conversations with greater ease. These enhancements can substantially increase their usefulness in a variety of applications. To maintain LLMs relevant and up-to-date, it is essential to enable continuous learning. This involves developing techniques that enable these models to adapt to evolving language and knowledge over time, ensuring that they remain valuable and accurate sources of information. Moreover, interpretable AI is an absolute necessity. This requires the development of methods to make LLM outputs more transparent and interpretable, thereby nurturing confidence and comprehension in AI decision-making processes. The development of multimodal LLMs that incorporate text, vision, and other modalities is an intriguing frontier. These models can comprehend and generate text from images, videos, and audio, creating new opportunities for AI applications in various fields. Collaboration between humans and artificial intelligence is also a crucial focal area. Research on how humans and LLMs can collaborate effectively, with AI assisting and augmenting human tasks, will be crucial for developing advanced AI applications in various fields. There is a need for dynamic evaluation metrics that can adapt to changing language and context in the context of evaluation. Developing relevant and up-to-date benchmarks is essential for accurately assessing LLM performance. Personalization and customization are becoming increasingly important for boosting user contentment. Exploring techniques to customize LLM interactions to the preferences and needs of individual users can considerably enhance their utility in a variety of applications.

Lastly, as AI regulation evolves, it's vital to work on developing ethical and legal regulatory frameworks that guide the responsible use of LLMs and ensure compliance with data protection and privacy regulations. These frameworks will play a pivotal role in regulating LLMs' ethical and responsible deployment in society. In conclusion, these research directions collectively pave the way toward maximizing the potential of LLMs while ensuring their accountable and ethical use in our evolving AI landscape.

11 Conclusion

The field of LLMs has witnessed a remarkable evolution and expansion, resulting in extraordinary capabilities in natural language processing (NLP) and various applications in various areas. Based on neural networks and the transformative transformer architecture, these LLMs have revolutionized our approach to machine language comprehension and generation. The thorough review of this research has provided an insightful overview of LLMs, encompassing their historical development, architectural foundations, training methods, and vast advancement resources. It has also examined the various applications of LLMs in disciplines such as healthcare, education, social sciences, business, and agriculture, demonstrating their potential to address real-world issues. In addition, this review has delved into the societal effects of LLMs, discussing how they shape the future of AI and can be utilized to address complex problems. However, it has not shied away from addressing the pressing challenges and ethical considerations associated with deploying LLMs, including model biases, privacy concerns, and the need for enhanced robustness and controllability. As the field of LLM research continues to evolve swiftly, this review is a valuable resource for practitioners, researchers, and experts seeking a comprehensive understanding of LLMs' past, present, and future. It emphasizes the significance

of ongoing efforts to improve the efficacy and dependability of LLMs, as well as the need for ethical development and deployment practices. LLMs represent a pivotal advancement in AI and NLP, with the potential to revolutionize a variety of domains and solve complex problems. This article provides a comprehensive foundation for future research and development in Large Language Models' dynamic and thrilling field.

References

[1] Steven Pinker. *The language instinct: How the mind creates language*. Penguin uK, 2003.

[2] Marc D Hauser, Noam Chomsky, and W Tecumseh Fitch. The faculty of language: what is it, who has it, and how did it evolve? *science*, 298(5598):1569–1579, 2002.

[3] Wayne Xin Zhao, Kun Zhou, Junyi Li, Tianyi Tang, Xiaolei Wang, Yupeng Hou, Yingqian Min, Beichen Zhang, Junjie Zhang, Zican Dong, et al. A survey of large language models. *arXiv preprint arXiv:2303.18223*, 2023.

[4] IBYAM Turing. Computing machinery and intelligence-am turing. *Mind*, 59(236):433, 2007.

[5] Yiqiu Shen, Laura Heacock, Jonathan Elias, Keith D Hentel, Beatriu Reig, George Shih, and Linda Moy. Chatgpt and other large language models are double-edged swords, 2023.

[6] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[7] Mengnan Du, Fengxiang He, Na Zou, Dacheng Tao, and Xia Hu. Shortcut learning of large language models in natural language understanding: A survey. *arXiv preprint arXiv:2208.11857*, 2022.

[8] Bhuvana Ramabhadran, Sanjeev Khudanpur, and Ebru Arisoy. Proceedings of the naacl-hlt 2012 workshop: Will we ever really replace the n-gram model? on the future of language modeling for hlt. In *Proceedings of the NAACL-HLT 2012 Workshop: Will We Ever Really Replace the N-gram Model? On the Future of Language Modeling for HLT*, 2012.

[9] Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari, 2010.

[10] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[11] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[12] Yash Khare, Viraj Bagal, Minesh Mathew, Adithi Devi, U Deva Priyakumar, and CV Jawahar. Mmbert: Multimodal bert pretraining for improved medical vqa. In *2021 IEEE 18th International Symposium on Biomedical Imaging (ISBI)*, pages 1033–1036. IEEE, 2021.

[13] Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, Lili Wang, and Soroush Vosoughi. Mitigating political bias in language models through reinforced calibration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14857–14866, 2021.

[14] Katharine Sanderson. Gpt-4 is here: what scientists think. *Nature*, 615(7954):773, 2023.

[15] Sundar Pichai. An important next step on our ai journey, feb 2023. *URL https://blog. google/technology/ai/bard-google-ai-search-updates*, 2(9), 2023.

[16] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Alpaca: A strong, replicable instruction-following model. *Stanford Center for Research on Foundation Models. https://crfm. stanford. edu/2023/03/13/alpaca. html*, 3(6):7, 2023.

[17] Jie Huang and Kevin Chen-Chuan Chang. Towards reasoning in large language models: A survey. *arXiv preprint arXiv:2212.10403*, 2022.

[18] Lizhou Fan, Lingyao Li, Zihui Ma, Sanggyu Lee, Huizi Yu, and Libby Hemphill. A bibliometric review of large language models research from 2017 to 2023. *arXiv preprint arXiv:2304.02020*, 2023.

[19] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Kaijie Zhu, Hao Chen, Linyi Yang, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *arXiv preprint arXiv:2307.03109*, 2023.

[20] Enkelejda Kasneci, Kathrin Seßler, Stefan Küchemann, Maria Bannert, Daryna Dementieva, Frank Fischer, Urs Gasser, Georg Groh, Stephan Günnemann, Eyke Hüllermeier, et al. Chatgpt for good? on opportunities and challenges of large language models for education. *Learning and individual differences*, 103:102274, 2023.

[21] Muhammad Usman Hadi, R Qureshi, A Shah, M Irfan, A Zafar, MB Shaikh, N Akhtar, J Wu, and S Mirjalili. A survey on large language models: Applications, challenges, limitations, and practical usage. *TechRxiv*, 2023.

[22] Marko Kardum. Rudolf carnap–the grandfather of artificial neural networks: The influence of carnap's philosophy on walter pitts. *Guide to Deep Learning Basics: Logical, Historical and Philosophical Perspectives*, pages 55–66, 2020.

[23] Geoffrey Leech. Corpora and theories of linguistic performance. *Svartvik, J. Directions in Corpus Linguistics*, pages 105–22, 1992.

[24] Elizabeth D Liddy. Natural language processing. 2001.

[25] Blaise Cronin. Annual review of information science and technology. 2004.

[26] Daniel S Hain, Roman Jurowetzki, Tobias Buchmann, and Patrick Wolf. A text-embedding-based approach to measuring patent-to-patent technological similarity. *Technological Forecasting and Social Change*, 177:121559, 2022.

[27] Georgina Curto, Mario Fernando Jojoa Acosta, Flavio Comim, and Begoña Garcia-Zapirain. Are ai systems biased against the poor? a machine learning analysis using word2vec and glove embeddings. *AI & society*, pages 1–16, 2022.

[28] Paul Azunre. *Transfer learning for natural language processing*. Simon and Schuster, 2021.

[29] Yangyang Shi, Martha Larson, and Catholijn M Jonker. Recurrent neural network language model adaptation with curriculum learning. *Computer Speech & Language*, 33(1):136–154, 2015.

[30] Tomas Mikolov and Geoffrey Zweig. Context dependent recurrent neural network language model. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 234–239. IEEE, 2012.

[31] Aldin Kovačević and Dino Kečo. Bidirectional lstm networks for abstractive text summarization. In *Advanced Technologies, Systems, and Applications VI: Proceedings of the International Symposium on Innovative and Interdisciplinary Applications of Advanced Technologies (IAT) 2021*, pages 281–293. Springer, 2022.

[32] Nur Mohammad Fahad, Sadman Sakib, Mohaimenul Azam Khan Raiaan, and Md Saddam Hossain Mukta. Skinnet-8: An efficient cnn architecture for classifying skin cancer on an imbalanced dataset. In *2023 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, pages 1–6. IEEE, 2023.

[33] Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.

[34] Rajesh Kumar Yadav, Sahil Harwani, Satyendra Kumar Maurya, and Sachin Kumar. Intelligent chatbot using gnmt, seq-2-seq techniques. In *2021 International Conference on Intelligent Technologies (CONIT)*, pages 1–5. IEEE, 2021.

[35] Dieuwertje Luitse and Wiebke Denkena. The great transformer: Examining the role of large language models in the political economy of ai. *Big Data & Society*, 8(2):20539517211047734, 2021.

[36] M Onat Topal, Anil Bas, and Imke van Heerden. Exploring transformers in natural language generation: Gpt, bert, and xlnet. *arXiv preprint arXiv:2102.08036*, 2021.

[37] Chiranjib Sur. Rbn: enhancement in language attribute prediction using global representation of natural language transfer learning technology like google bert. *SN Applied Sciences*, 2(1):22, 2020.

[38] Jordan J Bird, Anikó Ekárt, and Diego R Faria. Chatbot interaction with artificial intelligence: human data augmentation with t5 and language transformer ensemble for text classification. *Journal of Ambient Intelligence and Humanized Computing*, 14(4):3129–3144, 2023.

[39] Brady D Lund and Ting Wang. Chatting about chatgpt: how may ai and gpt impact academia and libraries? *Library Hi Tech News*, 40(3):26–29, 2023.

[40] Benyamin Ghojogh and Ali Ghodsi. Attention mechanism, transformers, bert, and gpt: tutorial and survey. 2020.

[41] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[42] Katherine Abramski, Salvatore Citraro, Luigi Lombardi, Giulio Rossetti, and Massimo Stella. Cognitive network science reveals bias in gpt-3, gpt-3.5 turbo, and gpt-4 mirroring math anxiety in high-school students. *Big Data and Cognitive Computing*, 7(3):124, 2023.

[43] Mohammad Shoeybi, Mostofa Patwary, Raul Puri, Patrick LeGresley, Jared Casper, and Bryan Catanzaro. Megatron-lm: Training multi-billion parameter language models using model parallelism. *arXiv preprint arXiv:1909.08053*, 2019.

[44] Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. Gpt-4 passes the bar exam. *Available at SSRN 4389233*, 2023.

[45] Xianrui Zheng, Chao Zhang, and Philip C Woodland. Adapting gpt, gpt-2 and bert language models for speech recognition. In *2021 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 162–168. IEEE, 2021.

[46] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[47] Jianmo Ni, Gustavo Hernández Ábrego, Noah Constant, Ji Ma, Keith B Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. *arXiv preprint arXiv:2108.08877*, 2021.

[48] Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[49] Luciano Floridi and Massimo Chiriatti. Gpt-3: Its nature, scope, limits, and consequences. *Minds and Machines*, 30:681–694, 2020.

[50] Jesse Vig and Yonatan Belinkov. Analyzing the structure of attention in a transformer language model. *arXiv preprint arXiv:1906.04284*, 2019.

[51] Alessia McGowan, Yunlai Gui, Matthew Dobbs, Sophia Shuster, Matthew Cotter, Alexandria Selloni, Marianne Goodman, Agrima Srivastava, Guillermo A Cecchi, and Cheryl M Corcoran. Chatgpt and bard exhibit spontaneous citation fabrication during psychiatry literature search. *Psychiatry Research*, 326:115334, 2023.

[52] Wei Zeng, Xiaozhe Ren, Teng Su, Hui Wang, Yi Liao, Zhiwei Wang, Xin Jiang, ZhenZhang Yang, Kaisheng Wang, Xiaoda Zhang, et al. Pangu: Large-scale autoregressive pretrained chinese language models with auto-parallel computation. *arXiv preprint arXiv:2104.12369*, 2021.

[53] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*, 2022.

[54] Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. Galactica: A large language model for science. *arXiv preprint arXiv:2211.09085*, 2022.

[55] Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, et al. Opt: Open pre-trained transformer language models. *arXiv preprint arXiv:2205.01068*, 2022.

[56] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[57] Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Alessandro Cappelli, Hamza Alobeidli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: outperforming curated corpora with web data, and web data only. *arXiv preprint arXiv:2306.01116*, 2023.

[58] Adam Roberts, Colin Raffel, and Noam Shazeer. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*, 2020.

[59] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[60] Thanh Thi Nguyen, Campbell Wilson, and Janis Dalins. Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts. *arXiv preprint arXiv:2308.14683*, 2023.

[61] Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. *arXiv preprint arXiv:2201.11990*, 2022.

[62] Opher Lieber, Or Sharir, Barak Lenz, and Yoav Shoham. Jurassic-1: Technical details and evaluation. *White Paper. AI21 Labs*, 1, 2021.

[63] Jack W Rae, Sebastian Borgeaud, Trevor Cai, Katie Millican, Jordan Hoffmann, Francis Song, John Aslanides, Sarah Henderson, Roman Ring, Susannah Young, et al. Scaling language models: Methods, analysis & insights from training gopher. *arXiv preprint arXiv:2112.11446*, 2021.

[64] Aohan Zeng, Xiao Liu, Zhengxiao Du, Zihan Wang, Hanyu Lai, Ming Ding, Zhuoyi Yang, Yifan Xu, Wendi Zheng, Xiao Xia, et al. Glm-130b: An open bilingual pre-trained model. *arXiv preprint arXiv:2210.02414*, 2022.

[65] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[66] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022.

[67] Mohaimenul Azam Khan Raiaan, Kaniz Fatema, Inam Ullah Khan, Sami Azam, Md Rafi ur Rashid, Md Saddam Hossain Mukta, Mirjam Jonkman, and Friso De Boer. A lightweight robust deep learning model gained high accuracy in classifying a wide range of diabetic retinopathy images. *IEEE Access*, 2023.

[68] Inam Ullah Khan, Mohaimenul Azam Khan Raiaan, Kaniz Fatema, Sami Azam, Rafi ur Rashid, Saddam Hossain Mukta, Mirjam Jonkman, and Friso De Boer. A computer-aided diagnostic system to identify diabetic retinopathy, utilizing a modified compact convolutional transformer and low-resolution images to reduce computation time. *Biomedicines*, 11(6):1566, 2023.

[69] Nan Du, Yanping Huang, Andrew M Dai, Simon Tong, Dmitry Lepikhin, Yuanzhong Xu, Maxim Krikun, Yanqi Zhou, Adams Wei Yu, Orhan Firat, et al. Glam: Efficient scaling of language models with mixture-of-experts. In *International Conference on Machine Learning*, pages 5547–5569. PMLR, 2022.

[70] Sid Black, Stella Biderman, Eric Hallahan, Quentin Anthony, Leo Gao, Laurence Golding, Horace He, Connor Leahy, Kyle McDonell, Jason Phang, et al. Gpt-neox-20b: An open-source autoregressive language model. *arXiv preprint arXiv:2204.06745*, 2022.

[71] Erik Nijkamp, Bo Pang, Hiroaki Hayashi, Lifu Tu, Huan Wang, Yingbo Zhou, Silvio Savarese, and Caiming Xiong. Codegen: An open large language model for code with multi-turn program synthesis. *arXiv preprint arXiv:2203.13474*, 2022.

[72] Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097, 2022.

[73] Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, 32, 2019.

[74] Xingchen Song, Guangsen Wang, Zhiyong Wu, Yiheng Huang, Dan Su, Dong Yu, and Helen Meng. Speech-xlnet: Unsupervised acoustic model pretraining for self-attention networks. *arXiv preprint arXiv:1910.10387*, 2019.

[75] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797, 2021.

[76] Renqian Luo, Liai Sun, Yingce Xia, Tao Qin, Sheng Zhang, Hoifung Poon, and Tie-Yan Liu. Biogpt: generative pre-trained transformer for biomedical text generation and mining. *Briefings in Bioinformatics*, 23(6):bbac409, 2022.

[77] openai. openai, 2023. Accessed Sep 12, 2023.

[78] huggingface. huggingface, 2023. Accessed Sep 12, 2023.

[79] Google Cloud. Cloud natural language, 2023. Accessed Sep 12, 2023.

[80] azure. azure, 2023. Accessed Sep 12, 2023.

[81] IBM. Ibm watson natural language understanding, 2023. Accessed Sep 12, 2023.

[82] G Satyanarayana, J Bhuvana, and M Balamurugan. Sentimental analysis on voice using aws comprehend. In *2020 International Conference on Computer Communication and Informatics (ICCCI)*, pages 1–4. IEEE, 2020.

[83] Adam Kolides, Alyna Nawaz, Anshu Rathor, Denzel Beeman, Muzammil Hashmi, Sana Fatima, David Berdik, Mahmoud Al-Ayyoub, and Yaser Jararweh. Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts. *Simulation Modelling Practice and Theory*, 126:102754, 2023.

[84] Shashank Mohan Jain. Hugging face. In *Introduction to Transformers for NLP: With the Hugging Face Library and Models to Solve Problems*, pages 51–67. Springer, 2022.

[85] Jindong Wang, Xixu Hu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Haojun Huang, Wei Ye, Xiubo Geng, et al. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. *arXiv preprint arXiv:2302.12095*, 2023.

[86] Shan Chen, Yingya Li, Sheng Lu, Hoang Van, Hugo JWL Aerts, Guergana K Savova, and Danielle S Bitterman. Evaluation of chatgpt family of models for biomedical reasoning and classification. *arXiv preprint arXiv:2304.02496*, 2023.

[87] Hanyao Huang, Ou Zheng, Dongdong Wang, Jiayi Yin, Zijin Wang, Shengxuan Ding, Heng Yin, Chuan Xu, Renjie Yang, Qian Zheng, et al. Chatgpt for shaping the future of dentistry: the potential of multi-modal large language model. *International Journal of Oral Science*, 15(1):29, 2023.

[88] Vera Sorin, Eyal Klang, Miri Sklair-Levy, Israel Cohen, Douglas B Zippel, Nora Balint Lahat, Eli Konen, and Yiftach Barash. Large language model (chatgpt) as a support tool for breast tumor board. *NPJ Breast Cancer*, 9(1):44, 2023.

[89] Arun James Thirunavukarasu, Darren Shu Jeng Ting, Kabilan Elangovan, Laura Gutierrez, Ting Fang Tan, and Daniel Shu Wei Ting. Large language models in medicine. *Nature medicine*, pages 1–11, 2023.

[90] Diane M Korngiebel and Sean D Mooney. Considering the possibilities and pitfalls of generative pre-trained transformer 3 (gpt-3) in healthcare delivery. *NPJ Digital Medicine*, 4(1):93, 2021.

[91] Luigi De Angelis, Francesco Baglivo, Guglielmo Arzilli, Gaetano Pierpaolo Privitera, Paolo Ferragina, Alberto Eugenio Tozzi, and Caterina Rizzo. Chatgpt and the rise of large language models: the new ai-driven infodemic threat in public health. *Frontiers in Public Health*, 11:1166120, 2023.

[92] Malik Sallam. Chatgpt utility in healthcare education, research, and practice: systematic review on the promising perspectives and valid concerns. In *Healthcare*, volume 11, page 887. MDPI, 2023.

[93] Marco Cascella, Jonathan Montomoli, Valentina Bellini, and Elena Bignami. Evaluating the feasibility of chatgpt in healthcare: an analysis of multiple clinical and research scenarios. *Journal of Medical Systems*, 47(1):33, 2023.

[94] Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198, 2023.

[95] Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23, 2021.

[96] Zeljko Kraljevic, Dan Bean, Anthony Shek, Rebecca Bendayan, Harry Hemingway, and Joshua Au. Foresight-generative pretrained transformer (gpt) for modelling of patient timelines using ehrs.

[97] Luisa Mich and Roberto Garigliano. Chatgpt for e-tourism: a technological perspective. *Information Technology & Tourism*, pages 1–12, 2023.

[98] Xinli Yu, Zheng Chen, Yuan Ling, Shujing Dong, Zongyi Liu, and Yanbin Lu. Temporal data meets llm–explainable financial time series forecasting. *arXiv preprint arXiv:2306.11025*, 2023.

[99] Guilherme Francisco Frederico. Chatgpt in supply chains: Initial evidence of applications and potential research agenda. *Logistics*, 7(2):26, 2023.

[100] Adam Sobieszek and Tadeusz Price. Playing games with ais: the limits of gpt-3 and similar large language models. *Minds and Machines*, 32(2):341–364, 2022.

[101] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects. 2023.

[102] Chung Kwan Lo. What is the impact of chatgpt on education? a rapid review of the literature. *Education Sciences*, 13(4):410, 2023.

[103] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. "so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[104] Mingyu Zong and Bhaskar Krishnamachari. A survey on gpt-3. *arXiv preprint arXiv:2212.00857*, 2022.

[105] Runjie Zhu, Xinhui Tu, and Jimmy Xiangji Huang. Utilizing bert for biomedical and clinical text mining. In *Data analytics in biomedical engineering and healthcare*, pages 73–103. Elsevier, 2021.

[106] Kexin Huang, Abhishek Singh, Sitong Chen, Edward T Moseley, Chih-Ying Deng, Naomi George, and Charlotta Lindvall. Clinical xlnet: modeling sequential clinical notes and predicting prolonged mechanical ventilation. *arXiv preprint arXiv:1912.11975*, 2019.

[107] Jipeng Zhang, Lei Wang, Roy Ka-Wei Lee, Yi Bin, Yan Wang, Jie Shao, and Ee-Peng Lim. Graph-to-tree learning for solving math word problems. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3928–3937, 2020.

[108] Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. Cost-effective selection of pretraining data: A case study of pretraining bert on social media. *arXiv preprint arXiv:2010.01150*, 2020.

[109] Som Biswas. The function of chat gpt in social media: According to chat gpt. *Available at SSRN 4405389*, 2023.

[110] Ruoling Peng, Kang Liu, Po Yang, Zhipeng Yuan, and Shunbao Li. Embedding-based retrieval with llm for effective agriculture information extracting from unstructured data. *arXiv preprint arXiv:2308.03107*, 2023.

[111] Som Biswas. Importance of chat gpt in agriculture: According to chat gpt. *Available at SSRN 4405391*, 2023.

[112] Telmo Pires, Eva Schlinger, and Dan Garrette. How multilingual is multilingual bert? *arXiv preprint arXiv:1906.01502*, 2019.

[113] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*, 2019.

[114] Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*, 2019.

[115] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*, 2019.

[116] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29, 2016.

[117] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*, 2019.

[118] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[119] Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240, 2020.

[120] Liming Zhu, Xiwei Xu, Qinghua Lu, Guido Governatori, and Jon Whittle. Ai and ethics—operationalizing responsible ai. *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*, pages 15–33, 2022.

[121] Inge Molenaar, Susanne de Mooij, Roger Azevedo, Maria Bannertd, Sanna Järveläe, and Dragan Gaševićf. Measuring self-regulated learning and the role of ai: Five years of research using multimodal multichannel data. *Computers in Human Behavior*, page 107540, 2022.

[122] Matthias C Rillig, Marlene Ågerstrand, Mohan Bi, Kenneth A Gould, and Uli Sauerland. Risks and benefits of large language models for the environment. *Environmental Science & Technology*, 57(9):3464–3466, 2023.

[123] Bowen Liu, Boao Xiao, Xutong Jiang, Siyuan Cen, Xin He, Wanchun Dou, et al. Adversarial attacks on large language model-based system and mitigating strategies: A case study on chatgpt. *Security and Communication Networks*, 2023, 2023.

[124] Zhongxiang Sun. A short survey of viewing large language models in legal aspect. *arXiv preprint arXiv:2303.09136*, 2023.

[125] Yu Meng, Martin Michalski, Jiaxin Huang, Yu Zhang, Tarek Abdelzaher, and Jiawei Han. Tuning language models as training data generators for augmentation-enhanced few-shot learning. In *International Conference on Machine Learning*, pages 24457–24477. PMLR, 2023.

[126] Steven Fincke, Shantanu Agarwal, Scott Miller, and Elizabeth Boschee. Language model priming for cross-lingual event extraction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 10627–10635, 2022.

[127] Neel Jain, Khalid Saifullah, Yuxin Wen, John Kirchenbauer, Manli Shu, Aniruddha Saha, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Bring your own data! self-supervised evaluation for large language models. *arXiv preprint arXiv:2306.13651*, 2023.