Received 14 January 2024, accepted 5 February 2024, date of publication 13 February 2024, date of current version 23 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365742

SURVEY

A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges

MOHAIMENUL AZAM KHAN RAIAAN[1], **MD. SADDAM HOSSAIN MUKTA**[2], **KANIZ FATEMA**[3], **NUR MOHAMMAD FAHAD**[1], **SADMAN SAKIB**[1], **MOST MARUFATUL JANNAT MIM**[1], **JUBAER AHMAD**[1], **MOHAMMED EUNUS ALI**[4], **AND SAMI AZAM**[3]

[1]Department of Computer Science and Engineering, United International University, Dhaka 1212, Bangladesh
[2]LUT School of Engineering Sciences, Lappeenranta-Lahti University of Technology, 53850 Lappeenranta, Finland
[3]Faculty of Science and Technology, Charles Darwin University, Casuarina, NT 0909, Australia
[4]Department of CSE, Bangladesh University of Engineering and Technology (BUET), Dhaka 1000, Bangladesh

Corresponding author: Md. Saddam Hossain Mukta (Saddam.Mukta@lut.fi)

ABSTRACT Large Language Models (LLMs) recently demonstrated extraordinary capability in various natural language processing (NLP) tasks including language translation, text generation, question answering, etc. Moreover, LLMs are new and essential part of computerized language processing, having the ability to understand complex verbal patterns and generate coherent and appropriate replies in a given context. 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 plethora of research on LLMs have been appeared within a short time, it is quite impossible to track all of these and get an overview of the current state of research in this area. 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 LLMs training phase. Then the paper 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. The paper also demonstrates 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. The study 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. Finally, the paper also explores open issues and challenges to deploy LLMs in real-world scenario. Our review paper aims to help practitioners, researchers, and experts thoroughly understand the evolution of LLMs, pre-trained architectures, applications, challenges, and future goals.

INDEX TERMS Large language models (LLM), natural language processing (NLP), artificial intelligence, transformer, pre-trained models, taxonomy, application.

I. INTRODUCTION

Language is a vital tool for human expression and communication which we begin to learn after our birth and make diverse use of it throughout our 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

The associate editor coordinating the review of this manuscript and approving it for publication was Bo Pu.

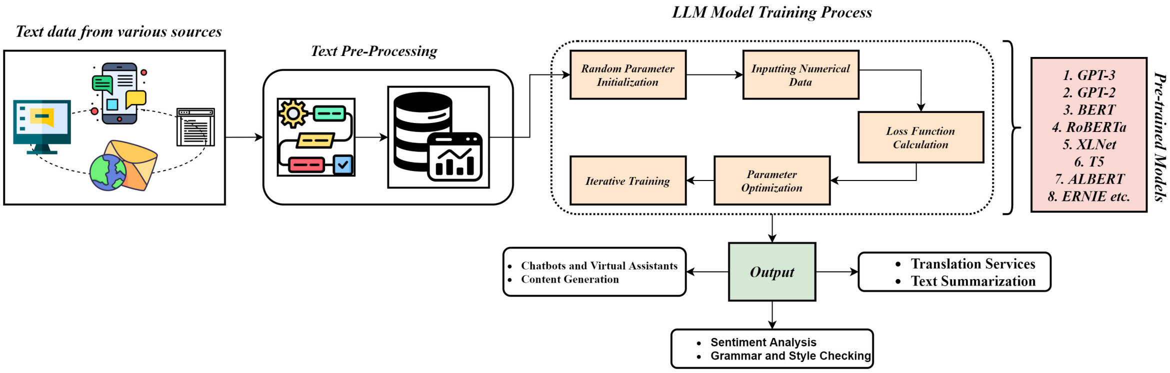


FIGURE 1. Pipeline of the LLMs training phase.

communication skills in machines [4]. 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). LLMs are 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]. Frequently pre-training on large corpora from the web, these models may learn complicated patterns, language subtleties, and semantic linkages. However, LLMs 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, fine-tuning these models on specific downstream tasks has been quite promising, with state-of-the-art performance in several benchmarks [6]. 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 [7]; 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) [8], 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 [9]. The transformer model is built around the self-attention mechanism, enabling parallelization and efficient handling of long-range dependencies. Furthermore, LLM architectures served as the basis for models such as Google's Bidirectional Encoder Representations from Transformers (BERT) [10] and open AI's Generative Pre-trained Transformer (GPT) series, which excelled at various language tasks.

The pipeline of the basic LLMs architecture is shown in Figure 1. LLMs architecture receives text data from multiple sources and then the architecture forwards text 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 [11] and politics [12]. Moreover, after inventing the most sophisticated GPT model [13], developing the state-of-the-art models (LLaMa and Bard [14]), and exploring their capabilities, such as Alpaca and GPTHuggingface [15], LLM has become a crucial and effective domain. As a result, a trustworthy assessment of current LLMs research is becoming increasingly important, and prior research has shown the potential and superiority of LLMs in NLP tasks. Despite this, only a few studies [3], [16], [17] have thoroughly reviewed latest LLMs developments, possibilities, and limitations in their research.

Besides, researchers have presented various aspects of the LLMs domain in several studies [3], [16], [17], [18]; but their work still has several limitations. These studies miss many aspects of LLM including high-level architecture and configurations, taxonomies, API and domain-specific applications, and datasets of LLMs. For example, there is a lack of introduction to the core architecture and configurations of the LLMs model, a lack of adequate explanation of the taxonomy of LLMs, differentiation based on ML, domain-specific applications, API applications, and descriptions of LLMs datasets. Furthermore, the vast majority of LLMs review papers are not peer-reviewed works. The absence of these key points in a review indicates that a thorough investigation is missing in the current literature. Due to the significant extent of the constraints, it is possible to mitigate these research gaps by thoroughly analyzing and addressing these missing points. Thus, the motivation of

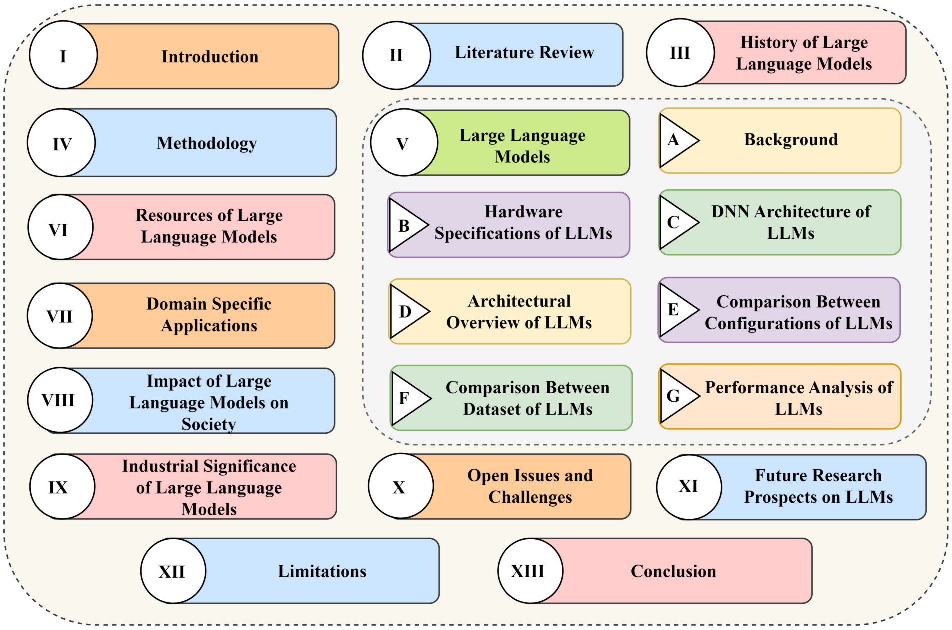

I
Introduction
II
Literature Review
III
History of Large Language Models
IV
Methodology
V
Large Language Models
A
Background
VI
Resources of Large Language Models
B
Hardware Specifications of LLMs
C
DNN Architecture of LLMs
VII
Domain Specific Applications
D
Architectural Overview of LLMs
E
Comparison Between Configurations of LLMs
VIII
Impact of Large Language Models on Society
F
Comparison Between Dataset of LLMs
G
Performance Analysis of LLMs
IX
Industrial Significance of Large Language Models
X
Open Issues and Challenges
XI
Future Research Prospects on LLMs
XII
Limitations
XIII
Conclusion

FIGURE 2. Section organization of the review.

this paper is to comprehensively explore the current review papers, identify their limitations, and outline the current state-of-the-art methods to address these vital challenges. Therefore, our primary objective is to explore, comprehend, and evaluate LLMs that encompass domains, evolution, classification, the structure of pre-trained models, resources, and real-time applications. Additionally, our comprehensive review discusses open issues and challenges associated with LLMs, including security, ethical, privacy, economic, and environmental considerations. In addition, we present a set of guidelines to explore future research and development in the effective use of LLMs. 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 a complete overview of LLMs, including their evolution, classification, and transformer architecture. The history of LLMs provides a brief account of the evaluation from its origins (1940) to the present (2023), as well as a taxonomy of LLMs based on pre-trained and API-based models and major LLMs structures.
- Describing the comparison of different pre-trained model designs in LLMs, along with their own systems that show how the model architectures are different.
- Explaining the influence of ML models on LLMs, demonstrating the significance of ML in various LLMs domains.
- Providing a brief overview of the datasets used in the training phase to differentiate between the models in existing works.
- Presenting a thorough explanation of the hardware implementation in training and testing models in terms of LLMs.
- Defining insights into the potential of LLMs and their impact on society and demonstrating bio-medical applications in five practical domains, including bio-medical and healthcare, education, social media, business, and agriculture.
- Investigating LLMs's diverse set of open issues, challenges, and future opportunities. This section focuses on identifying key challenges and future opportunities that can aid in advancing knowledge in this area.

The remaining sections of the paper are organized as depicted in Figure 2. In Section II, the literature review is discussed. Section III illustrates the history of LLMs; Section IV demonstrates the Methodology; Section V explains the clear concept of large language models; Section VI describes the resources of LLMs; Section VII demonstrates the domain-specific applications of LLMs; and Section VIII explains the societal impact of LLMs, Indusrial significance of LLMs is highlighted in Section IX, Section X discuss the open issues and challenges regarding LLMs, Section XI discusses about the future research directions of LLMs, Section XII acknowledges the limitation and Section XIII finally concludes the paper.

II. LITERATURE REVIEW

The growing number of LLMs is an extraordinary development in the field of AI. In recent years, numerous studies [3], [16], [17], [18] have been conducted to investigate and evaluate their capabilities. Researchers from various fields have contributed 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,

TABLE 1. Comparison between state-of-the-art research.

Papers LLMs	LLMs Model	LLMs API	LLMs Dataset	Domain Specific LLMs	Taxonomy	LLMs Architecture	LLMs Configurations	ML Based Comparison (Domain Specific)	Performance of LLMs	Parameters and Hardware Specification	Scope	Key Findings	Methodology and Approach
Huang et al., (2022) [18]	✓	X	X	X	X	X	X	X	X	X	Reasoning in LLMs	Aims to provide a critical analysis of LLMs 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	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 LLMs development resources and highlights significant contributions of LLMs to AI and NLP research areas.	Survey and analysis of LLMs evolution and impact
Fan et al., (2023) [16]	✓	X	X	X	X	X	X	X	X	X	Bibliometric review of LLMs research	Present a comprehensive overview of LLMs research from 2017 to 2023, tracking research trends, advancements, and provides insights into the dynamic nature of LLMs research, and impact in various domains.	Bibliometric analysis of over 5,000 LLMs publications
Chang et al., (2023) [17]	✓	X	✓	X	✓	X	X	X	X	X	Assessment of LLMs	Investigate the methodologies employed in evaluating LLMs 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 LLMs 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**

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., [18] presented a study on reasoning in LLMs that comprehensively summarizes the current state of LLMs' 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 LLMs-based reasoning. In another study, Zhao et al., [3] survey on LLMs 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 LLMs development and identifies further research and development areas.

A recent study by Fan et al. [16] conducted a bibliometric review of LLMs 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 LLMs research. The study also tracks research trends during the specified time period, including advancements in fundamental algorithms, major NLP tasks, and applications in disciplines such as medicine, engineering, social sciences, and the humanities. In addition to highlighting the dynamic and rapidly changing nature of LLMs research, the study offers insights into their current status, impact, and potential in the context of scientific and technological advancements. Chang et al. [17] 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. The study highlights the growing significance of evaluating LLMs at both the task and societal levels in order to comprehend potential risks. The paper thoroughly analyzes LLMs evaluation methods, focusing on three critical dimensions: what to evaluate, where to evaluate, and how to evaluate. The research also includes tasks such as natural language processing, reasoning, medical applications, ethics, and education. The article examines evaluation methods and benchmarks for assessing LLMs performance, emphasizing successful and unsuccessful cases. The paper also underlines future challenges in LLMs evaluation and emphasizes the importance of evaluating LLMs as a fundamental discipline to support the development of more competent LLMs.

Table 1 illustrates the comparison between different review papers based on some fundamental properties such as LLMs models, APIs, datasets, domain specific LLMs, ml-based comparison of LLMs, taxonomy, architectures, performance, hardware specifications for testing and training, and configurations. Huang et al. [18] lack information on LLMs' API, dataset, domain-specific LLMs, taxonomy, architectures, and LLMs Configurations. In contrast, Zhao et al., [3] has missing aspects on LLMs' API, domain-specific LLMs, taxonomy, architecture, and configurations. Moreover, Fan et al. [16] and Chang et al., [17] lack information on LLMs' API, domain-specific LLMs, taxonomy, architecture, and configurations.

On the contrary, our paper offers a considerably broader aspects on the LLMs context. In addition to incorporating every aspect specified in the table, we provide a detailed demonstration on the account of the hardware implementation and LLMs datasets. Previous research frequently focuses on limited aspects of LLMs, including historical development, bibliometric patterns, and assessment techniques. However, our study recovers previous shortcomings. A thorough examination is conducted on each of these aspects, resulting in a comprehensive representation of the strengths and weaknesses of LLMs. Furthermore, our research is focused on the crucial element of reasoning capabilities in LLMs, thereby providing a significant addition to the body of knowledge in the field. By giving thorough information, such as descriptions of datasets and hardware implementations required, our paper stands out as a primary resource for LLMs practitioners and researchers. Furthermore, we briefly discuss open issues in LLMs research, such as ethical and responsible AI, multimodal integration, energy efficiency, privacy and data protection, generalization and few-shot learning, and cross-lingual and low-resource settings. We also highlight key challenges, including data complexity and scaling, tokenization sensitivity, computational resource demands, fine-tuning complexity, real-time responsiveness, contextual constraints, bias and undesirable output, knowledge temporality, and evaluation complexity. Our review suggests future research directions to tackle open issues and important resource for LLMs researchers and practitioners. Our extensive systematic review presents a detailed discussion on LLMs which makes a substantial contribution to the field of LLMs research.

III. HISTORY OF LARGE LANGUAGE MODELS

LLMs refer to a category of AI models developed specifically to comprehend and produce human language [19]. LLMs have significantly contributed to the field of AI and have been applied in diverse areas, including education, communication, content generation, article composition, healthcare, research, entertainment, and information dissemination, among others [19], [20]. The origins of LLMs can be attributed to the emergence and advancement of neural network-based methodologies in the field of NLP [20]. In order to process language, early NLP systems utilized rule-based techniques and statistical models. However, those methods frequently encountered difficulties in comprehending the textual context in a specific discourse [21]. This section provides a high-level overview of LLMs, including their background, development, training, and operation. Figure 3 depicts the history of language models.

In the 1940s, Warren McCulloch and Walter Pitts introduced the idea of artificial neural networks (ANNs) [22]. Afterwards, 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]. These models experienced limitations in their abilities and encountered difficulties in managing the complexities of complicated language assignments. The models were predominantly employed for tasks involving binary classification. However, their efficacy in dealing with the complex situation in NLP tasks was limited [24].

Statistics-based models of language were created in the '80s and '90s. These models belong to a category of models utilized in the field of NLP and machine learning (ML) with the purpose of capturing and quantifying the statistical patterns and correlations within language data [21]. Statistical language models have significance in several applications, such as predictive text input, text generation, speech recognition, spam detection, etc. These models 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 [21]. Although statistical language models have been successful in many applications of NLP, they still have limitation when these models come to predict the semantic relationship between concepts and context of the language. These techniques have difficulty dealing with long-range dependencies [25].

During the mid-2000s, the field of NLP witnessed the introduction of word embeddings, which were recognized as a notable breakthrough and subsequently acquired considerable attention [26]. Word embedding refers to the process of representing words in a continuous vector space. The approach captures the semantic relationships among words by representing them in a vector space. The representation reduces the computational cost by mapping the words to a lower-dimensional space. Word2Vec and GloVe are widely recognized word embedding models in the domain [27]. These models are mostly utilized for assessing word similarity and assisting in the clustering and representation of words within semantic domains. Although not classified as LLMs, these embeddings have significantly contributed to the progress of natural language comprehension and have set the path for the development of more complex models. Nevertheless, these models have several limitations, such as their difficulty in effectively dealing with words that have multiple meanings (i.e., homonyms) or words that sound the same (i.e., homophones), as well as their inability to comprehend contextual information in an accepted manner [26].

The introduction of neural language models in the mid-2010s marked a significant advancement in LLMs [28]. These models employed deep learning approaches to acquire knowledge of language patterns from extensive textual data and additionally utilized artificial neural networks to comprehend, produce, or forecast human language. Furthermore, they have demonstrated exceptional outcomes in a wide range of language-related tasks. The initial neural language model to be introduced was the recurrent neural network language model (RNNLM) in 2010 [29]. The purpose of its development was to capture the sequential dependencies present in textual data. The utilization of a hidden state allows for the retention and propagation of information from preceding words in a particular sequence. RNNLM has been employed in several applications such

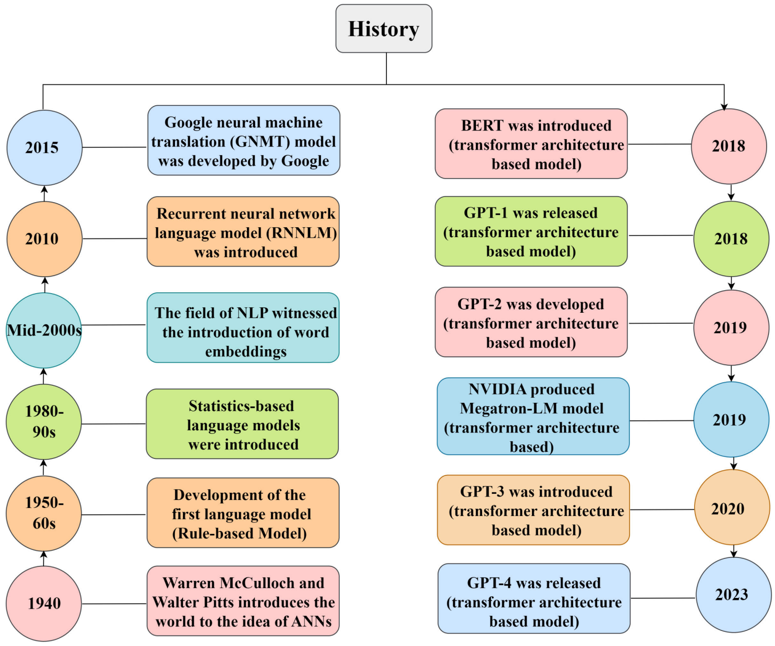

History
2015
Google neural machine translation (GNMT) model was developed by Google
2010
Recurrent neural network language model (RNNLM) was introduced
Mid-2000s
The field of NLP witnessed the introduction of word embeddings
1980-90s
Statistics-based language models were introduced
1950-60s
Development of the first language model (Rule-based Model)
1940
Warren McCulloch and Walter Pitts introduces the world to the idea of ANNs
BERT was introduced (transformer architecture based model)
2018
GPT-1 was released (transformer architecture based model)
2018
GPT-2 was developed (transformer architecture based model)
2019
NVIDIA produced Megatron-LM model (transformer architecture based)
2019
GPT-3 was introduced (transformer architecture based model)
2020
GPT-4 was released (transformer architecture based model)
2023

FIGURE 3. Brief history of language models.

as text production, speech recognition, machine translation, and language modeling. 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. Although the RNNLM offers certain advantages, it is not without its drawbacks. Some of these limitations include a limited short-term memory capacity, extended training time requirements, and prone to suffer in overfitting [30].

In the year 2015, Google unveiled the initial large neural language model that employed deep learning methodologies. The technology was referred to as the Google Neural Machine Translation (GNMT) model [31]. The model underwent training using huge quantities of multilingual textual data. This development signifies a notable progression in the field of machine translation [32]. The model demonstrated exceptional performance on machine translation tasks, departing from traditional rule-based and statistical techniques in favor of neural network-based methodologies. When compared to earlier language models, it was able to tackle complex natural language tasks with ease. The utilization of this model resulted in enhanced translation accuracy and the generation of meaningful translations, while also mitigating errors associated with intricate linguistic constructions [31].

The advancement of Language models persisted with the emergence of the Transformer model in the year 2017 [33]. The transformer model has had a significant impact on the field of NLP and 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) [34]. These models employ a self-attention mechanism that enables them to assess the relative significance of individual words in a sentence, thereby encoding complex relationships within the text [34]. The primary objective behind the development of the Transformer model was to overcome the inherent constraints observed in earlier models such as RNNs and Long Short-Term Memory (LSTM) networks. The Transformer models possess notable advantages in comparison to other models due to their ability to capture longer-term dependencies in language and facilitate concurrent training on many Graphical Processing Units (GPUs) with a vast number of parameters, enabling the construction of much larger models [35]. Parallelization capabilities and scalability are further benefits that have resulted in notable progress across many NLP activities [33].

The introduction of BERT in 2018 by Google AI represents a noteworthy advancement in the domain of NLP [16]. The underlying framework utilized in this study was the transformer architecture. 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 [36]. The employed methodology involved the utilization of deep bidirectional representations, which were conditioned on both the left and right contexts across all layers [37]. The pre-trained BERT model was able to undergo fine-tuning by incorporating an additional output layer, hence enabling its applicability to diverse tasks such as question answering and language inference. Due to the widespread adoption of BERT, several versions and subsequent models, such as RoBERTa, T5, and DistilBERT, have been developed to effectively address diverse tasks across multiple domains [37].

Following the advent of transformers, subsequent years saw the development of scaling-up LLMs models through the expansion of training data and parameter counts [20]. OpenAI significantly contributed to the development of LLMs in 2018. During the same year, GPT, an additional transformer-based architecture, was developed. Multiple iterations of the GPT models, developed by OpenAI, underwent pre-training using extensive datasets comprising excerpts from the Internet, novels, and various other textual sources [38]. The first version of the GPT model was referred to as GPT-1 [39]. The introduction of GPT-1 was a notable progression in the field of NLP. GPT-1 effectively produces words that are contextually appropriate, showcasing the transformative capabilities of transformers in significantly advancing natural language processing tasks. This proficiency is attributed to its extensive training on a vast number of parameters, specifically 117 million. The model underwent a two-step procedure consisting of unsupervised pre-training followed by supervised fine-tuning [20]. The initial iteration of GPT did not attain the same level of popularity as BERT due to several inherent limitations [40]. These drawbacks include a restricted context window, absence of bi-directionality, and occasional generation of biased content. Despite the inherent limits of GPT-1, this model played a crucial role in paving the way for later, more advanced models. As a result, it has sparked a new era of AI research and intensified competition in the development of LLMs.

The subsequent version of the GPT series, known as GPT-2, was designed with the purpose of addressing the limitations observed in its predecessor, GPT-1 [40]. Similar to GPT-1, GPT-2 was developed utilizing the transformer architecture. In the year 2019, Alec Radford introduced GPT-2, a language model that 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. Despite the high computing cost associated with training and executing the model, its substantial magnitude facilitates the comprehension and generation of a wide range of linguistic subtleties and diversified outputs [40]. The GPT-2 model has played a pivotal function in the advancement of LLMs and the execution of NLP activities. The influence of GPT-2 has had a significant impact on successor models like GPT-3 and GPT-4, leading to additional advancements in the field of language processing and creation [42].

In 2019, NVIDIA produced Megatron-LM, which is an LLMs [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 [16]. The magnitude of this dimension facilitates the model's capacity to acquire and produce intricate linguistic structures. Nevertheless, Megatron-LM has certain limitations, primarily due to its substantial dimensions, which necessitate substantial computational resources for both the training and inference processes [43].

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 [20]. The potential of LLMs for various NLP applications was exemplified by GPT-3. This particular LLMs 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 [16]. 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. One of the most essential characteristics of GPT-3 is its capacity to engage in few-shot and zero-shot learning, hence mitigating the necessity for extensive data in order to generate natural language text of superior quality. The advent of GPT-3 has catapulted the domain of natural language processing to new heights [40]

In the year 2020, OpenAI introduced GPT-4, the subsequent version of their language model, following the achievements of GPT-3 [20]. 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 [16]. The performance of the system was assessed using a range of standardized professional and academic examinations specifically intended for human test-takers. GPT-4 exhibited a level of performance comparable to that of humans on the majority of examinations. Significantly, it achieved a ranking inside the highest decile of participants on a simulated iteration of the Uniform Bar Examination [44]. GPT-4 has greater dimension and efficacy compared to its predecessor, GPT-3, as it possesses the capacity to generate text that is even more comprehensive and exhibits a heightened level of naturalness [20].

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 people with machines.

IV. METHODOLOGY

Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guide is crucial for drafting review papers as it assists systematic reviews in conducting transparent meta-analyses, accurately reporting aims and concluding the study, and ensuring the adequate reliability and relevance with the findings of the study [45]. Therefore, this review work focuses on the adoption of PRISMA

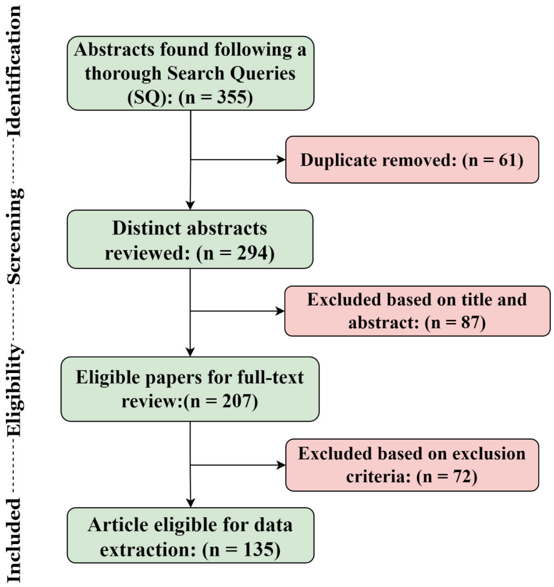

Identification
Screening
Eligibility
Included
Abstracts found following a thorough Search Queries (SQ): (n = 355)
Duplicate removed: (n = 61)
Distinct abstracts reviewed: (n = 294)
Excluded based on title and abstract: (n = 87)
Eligible papers for full-text review:(n = 207)
Excluded based on exclusion criteria: (n = 72)
Article eligible for data extraction: (n = 135)

FIGURE 4. PRISMA flow diagram of the review.

technique in analyzing the design, configurations, applications, and challenges of LLMs.

A. INITIAL SEARCHING

The research materials employed in this study have been acquired from recognized scientific journals and conferences from January 2020 to August 2023, conducted through the Google Scholar platform. A comprehensive selection of scholarly research articles has been specified, encompassing various reputable academic sources such as IEEE Xplore, ScienceDirect, ACM Digital Library, Wiley Online Library, Springer Link, MDPI, and patents. Initially, 355 papers were selected based on their relevance to the topic and keyword. Table 2 describes the identification technique of the materials from various electronic sources.

B. SEARCHING QUERY AND KEYWORDS

Using the combination of the appropriate search queries and keywords enlisted in Table 3 helps to perform a proper literature search. To conduct a thorough search of the articles for our LLMs-based review work, we encompass the following terms: “LLMs AND machine learning OR deep learning OR models,” “LLMs AND machine learning OR deep learning OR API,” “LLMs AND machine learning OR deep learning OR Dataset”, “LLMs AND natural language processing OR NLP” and “LLMs AND machine learning OR deep learning OR tools.” These specific searching techniques help to extract the eligible and quality research papers.

C. INCLUSION AND EXCLUSION CRITERIA SET

To acquire the final research papers, PRISMA protocols and principles were adhered to formulate a standard set of Inclusion Criteria (IC) and Exclusion Criteria (EC). The inclusion criteria define the standards of the paper that need to be included, while the exclusion criteria eliminate articles that do not meet the inclusion scope. Thus, this manual screening process improves the transparency of selection process. Table 4 presents the inclusion and exclusion criteria set for the proposed study.

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)

TABLE 3. Search queries used for the review paper.

	Search Queries (SQ)
SQ1	“LLMs” AND machine learning OR deep learning OR models
SQ2	“LLMs” AND machine learning OR deep learning OR API
SQ3	“LLMs” AND machine learning OR deep learning OR Dataset
SQ4	“LLMs” AND machine learning OR deep learning OR tools

D. PRISMA DIAGRAM

Figure 4 depicts the PRISMA flow diagram utilized in selecting papers for the study. It also provides the numbers of included and excluded papers for better understanding. The diagram begins by identifying articles from electronic databases using keywords, queries, resulting in 355 papers. After applying the screening method to exclude duplicated, low-quality, and irrelevant journal papers, the total number of papers for review is reduced to 294. Following a thorough analysis of the titles and abstracts, a total of 207 papers were selected. The final screening method involves the application of inclusion and exclusion criteria. Following this process, a total of 135 papers were ultimately selected for the final review. The process begins with an extensive collection of papers and reduces to the final selection that meets the pre-defined selection criteria for the systematic review.

V. 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 tasks. The most common tasks entail text generation, text analysis, translation, sentiment analysis,

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 three years (2020–2023)
IC4	Publication in a journal or 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 LLMs
EC4	Non-english documents

question answering, and other related functions. GPT-3, GPT-4, PaLM, and LaMDA are extensively used transformer-based LLMs 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 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 LLMs 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.

A. BACKGROUND OF LARGE LANGUAGE MODELS

In this section, we present the essential aspects associated. LLM research requires a comprehensive explanation of the crucial concept. Various vital aspects, such as tokenization, encoding technique, layer normalization, etc., are encompassed in the following background section.

1) TOKENIZATION

The primary emphasis is on tokenization, a crucial preprocessing stage of LLMs that involves parsing text into discrete parts referred to as tokens [46]. Characters, subwords, symbols, or words may serve as tokens, contingent upon the language model's dimensions and nature [47], [48]. Various tokenization algorithms are utilized in LLMs, such as WordPiece, UnigramLM, and Byte Pair Encoding (BPE). This algorithm has distinct technique for tokenizing from the input and then, applied for the specific tasks [47], [48], [49].

2) ATTENTION MECHANISM

The attention mechanisms used in LLMs is a crucial topic hence it contributes in the improvement of the architecture and performance. This mechanism helps to figure out the representation of input sequences by forming links between various tokens. There are several attention mechanism available namely Self-Attention where all the queries and values come from the same encoder-decoder block. Then, Full Attention which is the naive understanding version of self attention, and finally, when the output of encoder block is used as the query of immediate decoder block, is called as cross attention mechanism [9], [50].

3) ACTIVATION FUNCTION

The activation functions play a vital role in the curve-fitting capacities of LLMs architectures [51]. Several activation functions, such as ReLU, GeLU, and other GLU variations, are explored to determine their performance in current research on LLMs [52], [53].

4) NORMALIZATION LAYER

Layer normalization is essential for achieving faster convergence in LLMs model and emphasizes their effects on stability during training sessions. It presents different approaches, such as LayerNorm, DeepNorm, and RMSNorm. These layer normalization techniques offer distinct advantages and contribute to the regularization of LLMs applications like GPT-3, BERT, T5, etc., facilitating effective training [54].

5) TRAINING METHODS AND FRAMEWORKS

LLMs training has different distributed methodologies, including data parallelism, pipeline parallelism, tensor parallelism, model parallelism, and optimizer parallelism [43], [55]. These techniques contribute to understand the practical and expandable training. Additionally, different libraries and frameworks, including Transformers, DeepSpeed, PyTorch, TensorFlow, MXNet, and MindSpore, are used frequently for their training and further implementation [55].

6) DATA PREPROCESSING

The approaches used to preprocess data focus on the significance of quality filtering, data de-duplication and privacy reduction in preparing training data for LLMs. The filtering technique helps to reduce low quality and relevant data. Besides, it reduces the compute complexity by ignoring the useless pattern of the input. Duplicate samples are removed using de-duplication technique which also avoids the overfitting tendency of the model. Finally, privacy reduction ensures the security and compliance of data and upholds the preservation of the personal data.

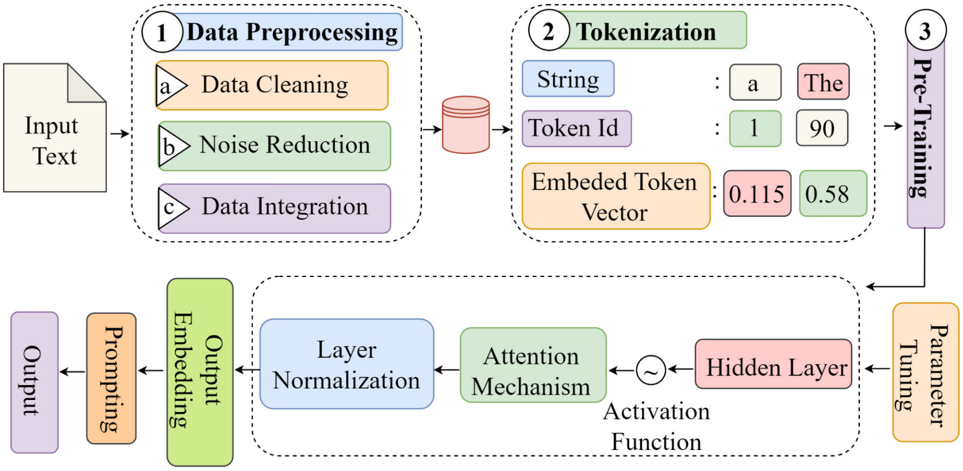


FIGURE 5. Background of LLMs.

7) PARAMETER TUNING

The researchers explore the many stages of adaptation for LLMs, starting from pre-training and progressing to fine-tuning for subsequent tasks. These approaches serve as a guide for customizing models to suit specific applications. Several model adaptation and parameter-efficient tuning techniques, such as prefix tuning, prompt tuning, and adapter tuning, provide strategies for achieving effective fine-tuning while minimizing resource usage [56], [57], [58].

This background part aims to provide a thorough understanding of the underlying concepts and approaches that form the basis of Language Models, which are constantly developing.

The transformer is employed in most advanced LLMs as the basic building block because its architecture, scalability, and pretraining approach enable the model as optimal framework for constructing robust LLMs. In addition, the self-attention mechanism of transformers performs effectively for capturing and representing long-range relationships in language. Consequently, Transformer-based LLMs have significantly improved the state-of-the-art achievement in NLP related tasks. In the section V-A1, a comprehensive overview of transformer architectures, configurations are provided for building a high-scalable, optimized and cost-efficient LLMs. Figure 5 depicts the visualization of the LLMs background.

8) WHAT IS TRANSFORMER?

Transformer architecture is considered as the basic building block of LLMs. It is intended for neural networks to efficiently handle sequential data [9]. This architecture does not use iteration methods. Instead, it employs a focused (i.e., attention based) approach to determine global input-output dependencies. The model can take input of varying lengths and can change its focus depending on the length of the sequence. As a result, it has become the go-to architecture in many fields, often replacing sophisticated recurrent or convolutional neural networks with much more efficient structure [59]. In this regard, it is particularly important for LLMs applications. Figure 6 illustrates the architecture of the transformer model. Transformer architecture consists of seven main components. A demonstration of each component is shown below.

- **Inputs and Input Embeddings**
  The ML models utilize tokens, which are units of text like words or sub words, as the training data. However, these models process numbers. Tokenization begins this translation process by dividing down input text into meaningful components. A unique number identification is assigned to each token, connecting the linguistic information to the numerical vector. This numerical format is 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 embeddings so that vectors of the same size represent words with similar meanings. Figure 6A illustrates the input and input embeddings.
- **Positional Encoding**
  The sequence of words in a sentence frequently conveys important semantic information. The same set of words in a different order conveys completely different meanings. In this regard, understanding the word order in a sentence is essential in NLP to identify the correct utterance meaning. In general, in terms of neural networks, they do not perceive the order of inputs.

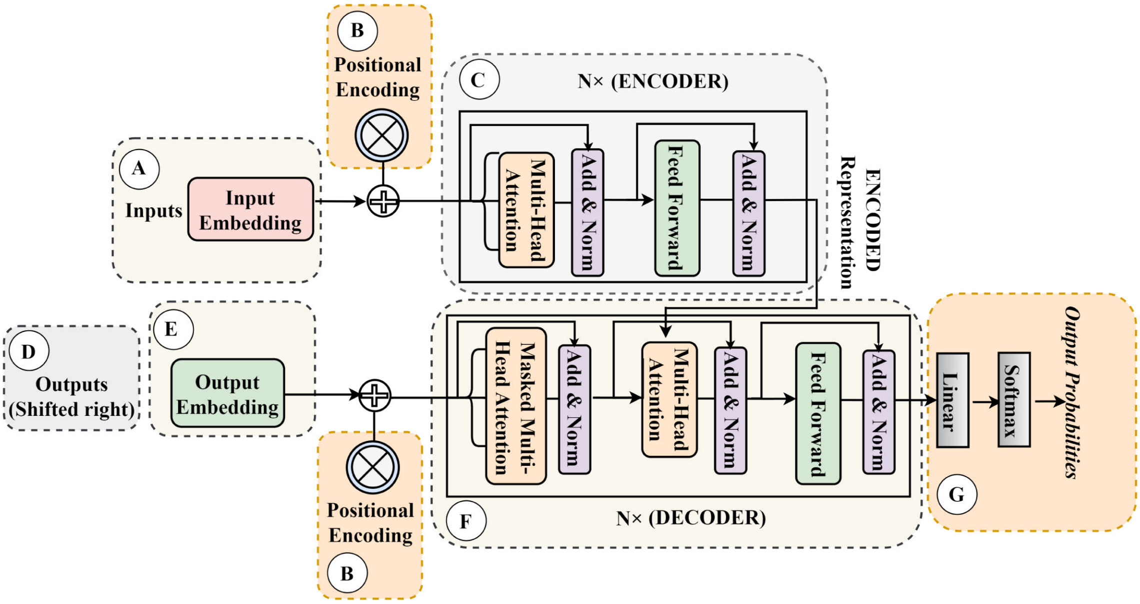


FIGURE 6. Architecture of a Transformer model.

To address the problem, positional encoding is used to encode the position of each word in the input sequence as a collection of integers. The transformer model uses integer, input embedding and positional encoding to help GPT in understanding sentence word order and provide grammatically accurate and semantically appropriate output [60]. The positional encoding part is shown in Figure 6B.

- **Encoder** The encoder is a crucial component of the neural network which is responsible for processing the input text. Its primary function is to generate a series of hidden states that represent the input text in a meaningful way [61]. Then, it uses a series of self-attention layers that are often referred to metaphorically as "voodoo magic," emphasizing their complex and powerful ability to capture relationships between different elements in the input text. In the transformer, the encoder is used in more than one layer. This section is depicted in Figure 6C comprehensively.
- **Outputs (shifted right)** During the training process, the decoder in the transformer model learns to predict the next word in a sequence by analyzing the preceding words. This is achieved through a mechanism known as autoregressive training. The decoder's ability to predict the next word is critical for generating coherent and contextually relevant sequences. Additionally, the GPT (GPT-3) is also trained on a massive amount of text data, that helps it to generate sense while writing any content. Besides, several corpus including the Common Crawl web corpus, the BooksCorpus dataset, and the English Wikipedia are also used during the common issue. Figure 6D highlights the transformer's outputs (shifted right) module.
- **Output Embeddings**
  Input embeddings, which contain text are not directly recognized by the model. Therefore, the output must be converted to a format known as "output embedding." Similar to input embeddings, output embeddings undergo positional encoding, enabling the model to understand the order of words in a sentence [62]. 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 6E.
- **Decoder**
  The decoder processes both positionally encoded input and output embeddings. Positional encoding is crucial for the model to understand the sequential order of the tokens in both the input and output sequences. The positional information helps the decoder effectively capture the structure within the sequences. The decoder has an attention mechanism that helps to improve the output's quality by leveraging contextual information received from the encoder. The primary function of the decoder is to create output sequences based on the encoded input sequences. It generates a sequence of tokens, often representing words or

TABLE 5. Hardware specifications for the LLMs model.

Model's Name	Parameters	Pre trained Data Scale	Hardware Specifications	Training Duration	Context Learning
GPT-3 [65]	175 Billion (B)	300B tokens	Nvidia A100 GPU	-	Yes
BERT [66]	340 Million (M)	-	Nvidia A100, V100	Depends on model parameter scale.	Yes
RoBERTa [67]	340 M	-	6144 TPU v4	Nearly 2 weeks.	Yes
T5 [68]	11 B	1 Trillion (T) tokens	1024 TPU v3	-	Yes
PaLM [69]	540 B	780 B tokens	6144 TPU v4	120 Days	Yes
LaMDA [70]	137 B	768 B tokens	1024 TPU v3	57.7 Days	Yes
GLM-130B [71]	130 B	400 B tokens	1024 TPU v4	60 Days	Yes
Gopher [72]	280B	300 B tokens	4096 TPU v3	920 Hours	Yes
Jurassic-1 [73]	178 B	300 B tokens	800 GPU	-	Yes
MT-NLG [74]	530 B	270 B tokens	4480 80G A100	-	Yes
LLaMA [75]	65 B	1.4 T tokens	2048 80G A100	21 Days	Yes
LLaMA 2 [76]	70 B	2 T tokens	2000 80G A100	25 Days	Yes
Falcon [77]	40 B	1.3 T tokens	-	-	Yes
Chinchilla [78]	70 B	1.4 T tokens	-	-	Yes
OPT [79]	175 B	180 B tokens	992 80G A100	-	Yes
Galactica [80]	120 B	106 B tokens	-	-	Yes
BLOOM [55]	176 B	366 B tokens	384 80G A100	105 Days	Yes
PanGU-a [81]	207 B	1.1 TB	2048 Ascend 910	-	Yes

sub-words, as its output. The dependency between the encoder-decoder in a transformer is significant where the encoder processes the input sequence based on the representation, the decoder provides the desired output sequence. In addition, GPT is a decoder-only transformer [63]. The decoder part of GPT uses a masked self-attention mechanism which can process the input sequence without requiring encoder explicitly. Figure 6F demonstrates the decoder component of a transformer.

- **Linear Layer and Softmax**
  The linear layer is a fully connected neural network layer that transforms the output embedding into a higher-dimensional space. This step is required to convert the output embedding into the original input space. This transformation enhances the expressiveness of the representation, allowing the model to capture more complex patterns and relationships in the data. Besides, the softmax function generates a probability distribution for each output token in the developed vocabulary, allowing us to generate probabilistic output tokens [64]. Figure 6G 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.

B. HARDWARE SPECIFICATIONS FOR LARGE LANGUAGE MODELS

Understanding the computing resources and training durations needed for various language models is crucial. This estimation helps us in decision-making when choosing a model for specific tasks. To choose a model that is appropriate for a given task, a clear understanding of the training times and computational resources is mandatory. Table 5 shows the hardware specifications, number of parameters, training duration and other configurations of individual LLMs model.

GPT-3: GPT-3 uses Nvidia A100 GPUs to pre-train on a large 300 billion token set, generating around 175 billion parameters [65]. GPT-3 has context learning features which enables itself to understand the words reasoning, sentence, and language properly.

BERT: Trained on an unspecified data scale, the BERT model has a variable number of parameters that depends on batch size and the corresponding model's hidden layer numbers which is around 340 million. Nvidia A100 and V100 GPUs are used for training, and the length of the training depends on the scale of the model's parameters [66]. Contextual learning is incorporated in the model also.

RoBERTa: RoBERTa, an enhanced version of BERT which has a parameter count of 340 million and conducts pre-training on a specific amount of data. The training process completed on 6144 TPU v4 units, running for around a duration of two weeks [67]. The model also contains a context learning feature.

T5: T5 uses 1024 TPU v3 units and has a number of 11 billion parameters. T5 has been pre-trained over a number of tokens of 1 trillion [68]. There is no information available on GPU training time. It also holds the features of contextual learning which provides a satisfactory result.

PaLM: PaLM produces a substantial number of parameters, around 540 billion, and it manages the pre-training on a large dataset with a tokens of 780 billion. The pre-training process is carried out utilizing by 6144 TPU v4 units [69]. The training period extends for 120 days, and the model also incorporates contextual learning.

LaMDA: LaMDA uses 1024 TPU v3 units during the training and the model is pre-trained over 768 billion tokens which generates a total of 137 billion parameters [70]. It requires a total of of 57.7 days during training.

GLM-130B: GLM-130B model possesses a total of 130 billion parameters and undergoes pre-training on a huge amount of dataset with 400 billion tokens. The training was conducted utilizing 1024 TPU v4 units and the training session lasts for 60 days [71].

Gopher: Gopher is a language model that has been pre-trained over 300 billion tokens and required 4096 TPU v3 for the experiment. It has a total of 280 billion parameters [72]. The GPU training period is precisely stated as 920 hours. Furthermore, the model integrates context learning to demonstrate an effective outcome.

Jurassic-1: Jurassic is a model with an impressive capacity of 178 billion parameters. It has been pre-trained on a massive dataset of 300 billion tokens, utilizing the computational power of 800 GPUs [73]. No information regarding the duration of GPU training is available.

MT-NLG: MT-NLG has a huge size of 530 billion parameters. It has been trained on a massive dataset of 270 billion tokens, utilizing 4480 80GB A100 GPUs [74]. No data regarding the duration of GPU training is available. The model integrates context learning features also.

LLaMA: LLaMA is a language model with an enormous capacity with a total of 65 billion parameters. It has undergone pre-training on a large dataset consisting of 1.4 trillion tokens. This training process was carried out utilizing 2048 high-performance 80GB A100 GPUs [75]. The training period is explicitly set to 21 days.

LLaMA 2: LLaMA 2 is equipped with a total of 70 billion parameters and has performed pre-training on 2 trillion tokens, utilizing 2000 80GB A100 GPUs [76]. The training period is set to 25 days, and the model also contains context-based learning.

Falcon: Falcon, equipped with 40 billion parameters, undergoes pre-training on a large dataset of 1.3 trillion tokens [77]. No details regarding the duration of GPU training and it also have the context learning features.

Chinchilla: Chinchilla is a language model that has 70 billion parameters and has been pre-trained on 1.4 trillion tokens [78]. There is no details regarding the duration of GPU training.

OPT: OPT, equipped with 175 billion parameters, conducts pre-training on 180 billion tokens utilizing 992 A100 GPUs with a capacity of 80GB each [79]. No details regarding the duration of GPU training.

Galactica: Galactica possesses 120 billion parameters and has undergone pre-training using 106 billion tokens [80]. Details regarding the duration of GPU training are not given.

BLOOM: BLOOM has a remarkable capacity of 176 billion parameters and has undergone pre-training on 366 billion tokens utilizing 384 80GB A100 GPUs [55]. The training period lasts for 105 days, and the model incorporates contextual learning.

PanGU-a: PanGU-a is a language model that has been pre-trained on a massive amount of data, specifically 1.1 billion, employing 2048 Ascend 910 processing units [81]. It has an impressive parameter count of 207 billion. No details regarding the duration of GPU training.

Our comprehensive description helps to understand the hardware specifications and the computational complexity of each model. The researchers also find an opportunity to know about the implementation details of these models and can improve the performance of their studies.

C. DEEP NEURAL NETWORK ARCHITECTURES OF LLMS

LLMs usually employe deep neural networks to understand and generate new content more accurately. In this section, we include a summary of various DNN architectures used in different LLMs based on literature studies and different real world applications.

1) COMPARISON BETWEEN STATE-OF-THE-ART STUDIES

An LLM is a dynamic model capable of performing various tasks, such as creating coherent text and summarizing text. A defining feature of a language model is its ability to assume the subsequent words from the preceding text. The deep neural network (DNN) framework is utilized in LLMs to enhance its performance which is similar to human-like understanding [3], [82]. LLMs use different DNN models in their architecture to enhance task performance.

The transformer architecture serves as the basic building block of all language models. GPT-1, the initial version of GPT employs the Transformer decoder architecture [66]. In GPT-1 the decoder structure operates independently from the encoder, therefore eliminating the multi-head attention and layer norm components that are linked to the encoder. The pre-trained GPT model consists of 12 transformer blocks, each with a d(model) value of 768 and a total of 110 million parameters. GPT-2, the second version of GPT, employs the transformer decoder architecture like GPT-1 [66]. GPT-2 employs 50,257 BPE tokens and ensures that the masked multi-head component is preceded by the Layer Norm. In GPT-2, an additional layer norm is included subsequent to the last block. There are four pre-trained GPT-2 models available, each with a unique quantity of decoder blocks. The largest model, which has a d(model) value of 1600 and 48 blocks, comprises a total of 1.5 billion model parameters. BERT employs the transformer encoder structure, in contrast to the Transformer decoder structure utilized by GPT-1 and GPT-2 [83]. Following the final encoder block is composed of two fully connected output layers separated by a Layer Norm component. The calculation of the likelihood of each token's output depends on both the previous and next tokens, making BERT a bidirectional language model. The smaller variant of BERT consists of 12 encoder blocks with a model dimension of 768 and a parameter count that is approximately equal to that of GPT. In contrast, the larger variant has 24 encoder blocks with a model dimension of 1024 and 336 million parameters [66].

In contrast to encoder-only models such as BERT and decoder-only models like GPT-1 and GPT-2, T5 pre-train with generative span corruption and an encoder-decoder architecture [84]. T5 models have displayed state-of-the-art performance on a wide variety of NLP tasks, like GLUE and SuperGLUE, and are able to expand up to hundreds of billions of parameters. LLaMA normalizes the input for every transformer sub-layer rather than the output [75]. To increase performance, it employs the RMSNorm normalizing function and the SwiGLU activation function rather than the ReLU. Single models are utilized by LaMDA to execute multiple duties. The model architecture is a decoder-only transformer language model. The Transformer is comprised of 64 layers, a d(model) value of 8192, gated-GELU as the activation function, and relative attention the same as T5 LLMs [70]. AlphaCode employs an encoder-decoder transformer architecture in which input tokens are passed to the encoder, and one token is extracted from the decoder until an end-of-code token is generated [85]. When contrasting encoder-decoder architectures with decoder-only architectures, the encoder-decoder architecture provides the advantage of enabling bidirectional description representation and provides additional flexibility by separating the encoder structure from the decoder. It employs an asymmetric architecture with 1536 encoder tokens but only 768 decoder tokens. It makes use of multi-query attention to lower sampling costs. Cache update costs and memory utilization are greatly reduced when all query heads are used but only shared for key and value heads in each attention block. It employed a SentencePiece tokenizer for tokenization, trained on a combination of CodeContests and GitHub data, with a vocabulary size of 8,000 tokens. Through the usage of DNNs, all of these LLMs have demonstrated remarkable performance on various NLP tasks like as language understanding and generation.

2) APPLICATIONS OF LLMS USING VARIOUS DNN MODELS

Pre-training Transformer models have led to the proposal of LLMs with impressive capacities in addressing a variety of NLP tasks, including question-answering, document summarization, and language translation [3]. Due to their remarkable abilities in basic tasks of language processing and creation, they have completely transformed the fields of NLP and AI. Various DNN models have been employed in different industries, such as technology, healthcare, and retail to increase performance. DNNs have made substantial progress in improving the capabilities of LLMs [87]. DNN models, such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), generative adversarial networks (GANs), capsule networks (CapsNets), transformers, and BERT, have been extensively employed in diverse applications of LLMs [94]. Numerous studies [86], [87], [88], [89], [90], [91], [92], [93] suggest that DNN models are utilized in several types of LLMs-based applications to increase task efficiency.

Koizumi et al., [86] introduce an innovative method to address the issue of insufficient training data in audio captioning that utilizes a pre-trained LLMs that uses a decoder for generating captions. The findings of the study demonstrate the effectiveness of the proposed methodology in utilizing LLMs for audio captioning. The performance of this proposed approach outperforms the traditional approaches which are trained from the scratch.

In a recent study, Fan et al., [87] discuss the significance of recommender systems in web applications and the shortcomings of current DNN approaches in predicting user preferences. They discuss the capacity of LLMs to tackle the challenges in a recommender systems.

Bai et al. [88] developed an end-to-end non-autoregressive speech recognition model namely LASO (Listen Attentively and Spell Once) to improve the speed of inference by simultaneously predicting all tokens. The proposed model utilizes attention methods to combine decoded speech information into hidden representations for every token. Moreover, they suggest using cross-modal transfer learning to increase the performance of the speech-modal LASO model by utilizing a text-modal language model to align the semantic meaning of tokens.

Sun et al., [89] provide a new methodology to predict the effect of news releases and to minimize potential negative consequences by automatically forecasting responses in news media. By utilizing an LLM which utilizes a deep neural network, their method creates a belief-centered graph on an existing social network to analyze social dynamics. The proposed framework shows a satisfactory efficiency in predicting responses.

Drossos et al., [90] present a technique that enables an RNN to acquire LLMs for sound event detection. The proposed approach adjusts the input of the RNN based on the activity of classes in the preceding time step. This proposed approach is evaluated on three distinct datasets: the TUT-SED Synthetic 2016, TUT Sound Events 2016, and TUT Sound Events 2017 datasets.

Chiu et al. [91] present an efficient method called TPBERT (based on BERT) for improving the reranking of N-best hypotheses in automatic recognition of speech. This approach uses task-specific topic information to increase the BERT model's ability to create accurate embeddings of the N-best hypotheses.

Elhafsi et al., [92] propose a monitoring methodology that utilizes LLMs to tackle the issue of semantic irregularities in robotic systems. The efficiency of LLMs-based monitoring in recognizing semantic abnormalities and aligning with human thinking is demonstrated through tests on autonomous driving.

Shen et al., [93] present a self-regulating edge AI system that utilizes a deep neural network that can plan automatically, and adjust itself to fulfill the needs of users. The proposed system uses a hierarchical design known as cloud-edge-client, where the primary language model is located in the cloud. By leveraging the robust capabilities of GPT in language comprehension, and code creation, they introduce a methodology that effectively handles edge AI models to meet users' requirements while automatically

TABLE 6. Comparison of applications of LLMs using various DNN models.

Study	DNN model	Application
Koizumi et al., [86]	Transformer (decoder)	Assessd the use of a pre-trained large-scale language model in audio captioning.
Fan et al., [87]	Transformer (encoder-decoder)	Discuss the significance of Recommender Systems in web applications and e-commerce cite
Bai et al., [88]	Non-Autoregressive attention based encoder-decoder	Propose a non-autoregressive speech recognition model named LASO (Listen Attentively and Spell Once)
Sun et al., [89]	Decoder + SOCIALSENSE (belief-centered graph)	Forecast the impact of news releases and attempt to mitigate potential adverse consequences by automatically anticipating news media responses
Drossos et al., [90]	RNN	Propose a method for sound event detection which takes a sequence of audio frames as input and predicts the activities of sound events in each frame.
Chiu et al., [91]	Transformer (encoder)	Propose a method called TPBERT to improve the reranking of N-best hypotheses in automatic recognition of speech.
Elhafsi et al., [92]	Encoder structure	Propose a monitoring system for dealing with semantic abnormalities in robotic systems.
Shen et al., [93]	Transformer (decoder)	Propose a self-regulating edge AI system to autonomously plan, and adjust itself to fulfill the needs of users.

generating new codes for training new models through edge federated learning.

Table 6 gives a brief overview of these DNN applications-oriented studies where they applied LLMs. These studies suggest that employing deep neural networks in language models increases the performance of LLMs-based applications in several industries..

D. ARCHITECTURAL OVERVIEW OF LARGE LANGUAGE MODELS

In this subsection, we present a detailed overview on the architecture of LLMs. Table 7 presents a description and architecture of LLMs such as GPT-1, BERT, RoBERta, and T5. The table assists researchers in selecting the optimal model for a NLP task. GPT-1, BERT base, and BERT large contain 12, 12, and 24 layers, respectively, in LLMs. 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. T5 explains how information flows and structures text. GPT-1 passes data input embedding and positional encoding through multiple transformer layers.

E. COMPARISON BETWEEN CONFIGURATIONS OF LLMS

Table 8 provides an extensive overview of various LLMs, highlighting their configuration details and optimization settings. These LLMs have played a crucial role in advancing natural language understanding and generation tasks, making them a key research topic in NLP. 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 considered as one of high performing LLMs with a staggering 1.8 trillion parameters. It is comparatively faster than the prior GPT versions and provide many advanced features. Besides, it has fast response system, generate more accurate output and it has reduced the biases presented in the model substantially. GPT-1, despite being lesser with 125 million parameters, demonstrates the significant development of LLMs over the years. An increased number of parameters in LLMs 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 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 [6], [59]:

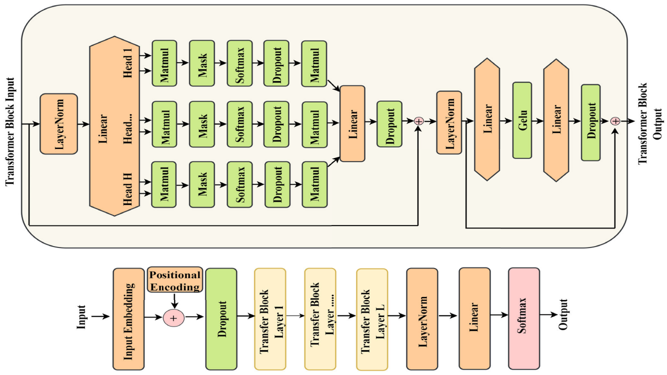
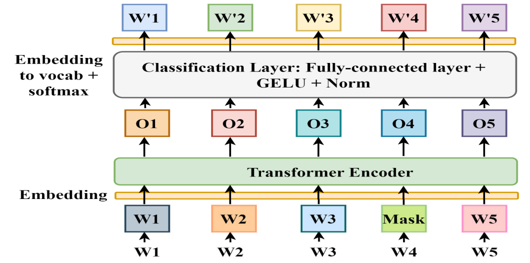
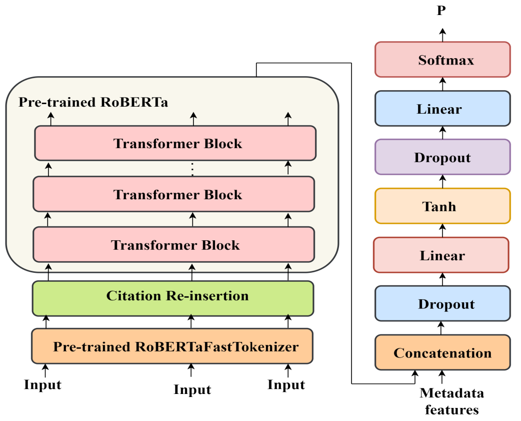
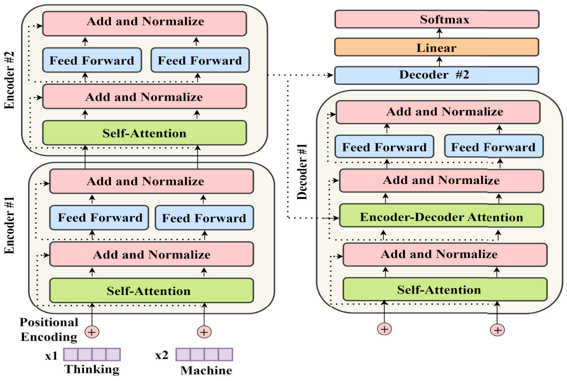
$$ReLU(x) = \max(0, x) = f(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0 \end{cases} \quad (1)$$

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

$$SwiGLU(x) = x.Sigmoid(\beta x).xV \quad (3)$$

BARD is recognized for its informative response. It features 24 attention heads and facilitates its contextually related

TABLE 7. Architectural overview of different LLMs.

Model	Description	Architecture
GPT-1 [66]	Twelve-level decoder transformer that uses twelve masked self-focusing heads.	 Transformer Block Input LayerNorm Linear Head 1 Head... Head H Matmul Mask Softmax Dropout Matmul Linear Dropout LayerNorm Linear Gelu Linear Dropout Transformer Block Output Input Input Embedding Positional Encoding Dropout Transfer Block Layer 1 Transfer Block Layer .... Transfer Block Layer L LayerNorm Linear Softmax Output
BERT [10]	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.	 W'1 W'2 W'3 W'4 W'5 Embedding to vocab + softmax Classification Layer: Fully-connected layer + GELU + Norm O1 O2 O3 O4 O5 Transformer Encoder Embedding W1 W2 W3 Mask W5 W1 W2 W3 W4 W5
RoBERTa [67]	Optimized version of BERT model.	 P Softmax Linear Dropout Tanh Linear Dropout Concatenation Metadata features Pre-trained RoBERTa Transformer Block Transformer Block Transformer Block Citation Re-insertion Pre-trained RoBERTaFastTokenizer Input Input Input
T5 [84]	The model consists of an encoder and a decoder transformer, which has many layers.	 Encoder #2 Add and Normalize Feed Forward Feed Forward Add and Normalize Self-Attention Encoder #1 Add and Normalize Feed Forward Feed Forward Add and Normalize Self-Attention Positional Encoding x1 Thinking x2 Machine Softmax Linear Decoder #2 Decoder #1 Add and Normalize Feed Forward Feed Forward Add and Normalize Encoder-Decoder Attention Add and Normalize Self-Attention

response. BERT size is identical to BARD of 340M. The key advantage of BERT is understanding the context of words. It has effective training settings with a proper learning rate, batch size, and a dropout value of 0.1, leverages the convergence of the model, and contributes to the NLP-based tasks precisely. PanGU BLOOM, Galactica, and

TABLE 8. 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 [95]	1.8 T	–	CD	GeLU	–	Yes	120	Adam	120-150	20000	–	32768
GPT-3 [65]	175B	6×10^{-5}	CD	GeLU	32K-3200K	Yes	96	Adam	96	12288	–	2048
GPT-2 [96]	1.5B	1×10^{-4}	AR	GeLU	16K-64K	Yes	48	Adam	24	1280	0.1	1024
GPT-1 [66]	125M	1×10^{-4}	AR	GeLU	16K-64K	Yes	12	Adam	12	768	0.1	512
BARD [97]	340M	–	–	ReLU	64K	Yes	24	–	24	768	–	512
BERT [66]	340M	1×10^{-5}	–	ReLU	16K-64K	Yes	24	Adam	16	1024	0.1	512
PanGU-α [81]	207B	2×10^{-5}	CD	GeLU	–	Yes	64	Adam	128	16384	–	1024
BLOOM [55]	176B	6×10^{-5}	CD	GeLU	4000K	Yes	70	Adam	112	14336	0	2048
Galactica [98]	120B	7×10^{-6}	CD	GeLU	2000K	No	96	AdamW	80	10240	0.1	2048
OPT [79]	175B	1.2×10^{-4}	CD	ReLU	2000K	Yes	96	AdamW	96	12288	0.1	2048
Chinchilla [78]	70B	1×10^{-4}	CD	–	1500K-3000K	–	80	AdamW	64	8192	–	–
Falcon [77]	40B	1.85×10^{-4}	CD	GeLU	2000K	No	60	AdamW	64	8192	–	2048
T5 [68]	11B	1×10^{-2}	ED	ReLU	64K	No	24	AdaFactor	128	1024	0.1	512
LLaMA [75]	65B	1.5×10^{-4}	CD	SwiGLU	4000K	No	80	AdamW	64	8192	–	2048
LLaMA-2 [76]	70B	1.5×10^{-4}	CD	SwiGLU	4000K	No	80	AdamW	64	8192	–	4096
MT-NLG [74]	530B	5×10^{-5}	CD	–	64K-3750K	–	105	Adam	128	20480	–	2048
Jurassic-1 [73]	178B	6×10^{-5}	CD	GeLU	32K-3200K	Yes	76	–	96	13824	–	2048
Gopher [72]	280B	4×10^{-5}	CD	–	3000K-6000K	–	80	Adam	128	16384	–	2048
GLM-130B [71]	130B	8×10^{-5}	PD	GeGLU	400k-8250K	Yes	70	AdamW	96	12288	0.1	2048
LaMDA [70]	137B	–	CD	GeGLU	256K	–	64	–	128	8192	–	–
PaLM [69]	540B	1×10^{-2}	CD	SwiGLU	1000K-4000K	No	118	Adafactor	48	18432	0.1	2048

[]

Chinchilla are also LLMs but possess distinct configurations and challenges. Usually, PanGU is highly effective for the Chinese language, whereas Galactica performs well with repeated data. Chinchilla is a scaling strategy constrained by data limitations and creates efficient resource allocation for training and generating output. Falcon and T5 are compact compared to other LLMs, and both are transformer-based models. However, they have some unique differences, such as Falcon is a decoder-based model whereas T5 integrated both encoder-decoders. Additionally, Falcon utilizes multi-head query attention to increase the scalability of the model. LLaMA-2 is the updated version of LLaMA. It is an enhanced fine-tuned version that exploits the hardware utilization for efficient training sessions. MT-NLG and PaLM have substantial parameter sizes of 530B and 540B, respectively. Both of them also use the casual decoder technique. However, they have some architectural differences, such as PaLM uses a SwiGLU activation function and adafactor optimizer. Moreover, it uses a higher learning rate and batch size of 1 × 102 and 1000K. On the contrary, MT-NLG uses a lower learning rate and batch size of 5 × 105 and 64K, respectively. GLM-130B and LaMDA are also effective LLMs, widely used for NLP-based tasks, including question answering, text generation, etc. Both of them use the Gated GLU (GeGLU) activation function, a GLU variant. The following equation is used to express the GeGLU operation [99].

$$\mathrm{GEGLU}(x, W, V, b, c) = \mathrm{GELU}(xW + b) \otimes (xV + c) \quad (4)$$

However, there are noticeable differences between GLM-130B and LaMDA in terms of their decoder mechanisms. GLM-130B employs a prefix decoder, whereas LaMDA adopts a casual decoder technique. In addition, the GLM-130B model employs a larger batch size compared to the LaMDA model. 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. 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 LLMs 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 LLMs models incorporate a dropout value of 0.1 to prevent overfitting issues, whereas others do not employ any dropout value. The maximum context length denotes the number of tokens that can be remembered by the model during training. Increasing the size of the context window boosts the model's ability to grasp the distant relationships between the texts. Consequently, the model is

TABLE 9. Dataset for large language models.

Dataset →	Webpages			Conversation Data	Books and News					Scientific Data		Code	
LLMs ↓	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 [68]	✓	✓	✓	X	X	X	X	X	X	X	X	X	X
Falcon [77]	✓	✓	✓	X	X	X	X	X	X	X	X	X	X
LLaMA [75]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
GPT-3 [65]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	X	X
GPT-4 [95]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	X	X
MT-NLG [74]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Gopher [72]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	✓	✓
Chinchilla [78]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	✓	✓
GLaM [100]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	X	X
PaLM [69]	✓	✓	✓	X	✓	✓	✓	✓	✓	X	X	✓	✓
LaMDA [70]	✓	✓	✓	X	X	X	X	X	X	✓	✓	✓	✓
Galactica [98]	✓	✓	✓	X	X	X	X	X	X	✓	✓	✓	✓
GPT-NeoX [101]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
CodeGen [102]	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AlphaCode [85]	X	X	X	X	X	X	X	X	X	X	X	✓	✓
GPT-1 [85]	X	✓	✓	X	✓	X	X	X	X	X	X	✓	✓
GPT-2 [85]	X	✓	✓	X	✓	X	X	X	X	X	X	X	X
BARD [85]	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X
BERT [85]	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X
PanGU- [85]	✓	✓	✓	X	✓	X	X	X	X	X	X	X	X
BLOOM [85]	✓	✓	✓	X	X	X	X	X	X	X	X	X	X
OPT [85]	X	✓	✓	X	✓	✓	X	X	X	X	X	X	X
GLM-130 [85]	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)

able to generate text outputs with a great coherence. Table 8 reports that GPT-4 has the context length of 32768 which is the maximum among all the LLMs. This substantial length number indicates the capability of GPT-4 to remember the more extended token sequence during training. LLaMA-2 obtained the second-highest context length of 4096. Most of the models have a context length of 2048, meaning they can handle a maximum of 2048 tokens simultaneously during the text generation. A few compacted models, including BARD, BERT, and T5, possess a maximum context length of 512. This table presents a qualitative architectural comparison among the most popular LLMs. It also provides comprehensive knowledge about the configurations, strength of these models. These variations highlight the significance of selecting models for the particular tasks considering the performance, computational resources.

F. COMPARISON BETWEEN DATASETS OF LLMS

Different LLMs utilized different datasets for the training phase, distinguishing the models from one another. A concise overview of the datasets is provided in this section. Moreover, it explicitly exhibits the diverse range of datasets used by the model since understanding of these datasets facilitates the development and training of the model and boost the performance. The datasets used to train various large language models (LLMs) and their compatibility with each model are detailed in Table 9.

Table 9 demonstrates that datasets have been divided into multiple categories: *webpages, conversation data, literature, news, scientific data*, and *codes*. This classification enables us to comprehend the variety of data sources that contribute to LLMs 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 codes.

From the ✓, we observe 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 codes (5%). Therefore, LLaMA becomes a versatile model suitable for a wide array of NLP tasks that involve these mentioned data sources. In contrast, GPT-3 and AlphaCode have limited data access of data sources to train their models. GPT-1 and GPT-2 focus on webpages (70%) and books & news (30%) data to train the model. GPT-3 is proficient with web pages (84%), literature, and news (16%) but requires additional instruction with conversation data, scientific data, and codes. Diverse range of datasets enables the GPT models to generate more contextual information across various domains. Specifically, the Webpages, books, and news datasets help to employ formal and structured language. Besides, GPT models achieve the capability of responding in an informative and accurate way.

AlphaCode, as its name suggests, is solely focused on codes (100%) and does not utilize any other data sources. This feature uniquely distinguish AlphaCode from other models and emphasize the significance of this model for code-based tasks. Bard, Bert, and Pangu models exhibit identical traits, with each of them concentrating on the extensive textual data obtained from webpage contents and books for pretraining the models. Bloom and OPT primarily emphasize on evaluating data from books and websites, such as Wikipedia or other online sources. On the other hand, GLM-130 not only analyzes books and web data but also incorporates computer code data to provide further technological benefits. LaMDA, Galactica and CodeGen models use scientific data source for training which advance these models to adapt the scientific knowledge and terminology.

Hence, these model can lead to a more accurate responses in scientific domains. AlphaCode and GLM-130 are the models of choice for code-related tasks, whereas LLaMA and BERT excel in diverse text data applications. Most of the LLMs such as T5, GPT models, Gopher, GLam, PaLM, and BLOOM frequently utilize websource data which helps them to automate various practical tasks such as content creation, data analysis and virtual chatbot for answering the question. On the contrary, some models such as Falcon and different version of GPT models utilize books and news data facilitates in educational application such as document summarization, and article writings. The models trained on scientific data have several use cases in research domain. In addition, Table 9 provides contextual information of the datasets to maintain the transparency of the comparison among models and provide an effective guide to future model implementation. The "Size" and "Source" columns of the Table listed the additional information. The size of datasets ranges from 5GB (BookCorpus) to a huge 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 understanding the temporal relevance of the training data and potential biases. Table 9 provides a multitude of information regarding the datasets used to train LLMs and how each model leverages these datasets. This information is invaluable for NLP researchers, developers, and practitioners, as it enables them to make informed decisions about which LLMs to use for specific tasks.

G. PERFORMANCE ANALYSIS OF LLMS

LLMs are models that perform the majority of NLP tasks and numerous models such as GPT-1 through GPT-4, Bing, ChatpGPT, and BERT have developed in order to contribute jointly to the industry and academia. Since in the literature, we find a scarcity of adequate data pertaining to LLMs, we present performance outcomes for diverse tasks to publicly accessible LLMs in Table 10. All GPT series, including GPT-1, GPT-2, GPT-3, GPT-3.5, and GPT-4, are evaluated using a variety of metrics, including the Stanford question answering dataset (SQuAD), language model benchmark (LAMBADA), and general language understanding evaluation (GLUE), as shown in Table 10. GPT-1 obtains a score of 68.4 on the GLUE, while GPT-2, GPT-3, GPT-3.5, and GPT-4 attain scores of 84.6, 93.2, 93.5, and 94.4, respectively. GLUE results indicate that GPT-4 outperforms prior versions of GPT. The GPT-4, i.e., in SQuAD and LAMBDA have scores of 93.6 and 82.4, respectively. As shown in the table, GPT-4 outperforms its predecessors in both LAMBDA and SQuAD. As GPT-4 outperforms its predecessors in all three benchmark metrics and exhibits robust performance, it can be concluded that GPT-4 is significantly more effective than its predecessors in tasks involving language understanding and language modeling. The VietNamese High School Graduation Examination (VNHSGE) English dataset was utilized to analyze various LLMs, including GPT-3.5, BingChat, and BARD. Based on the accuracy presented in Table 10, it is evident that BingChat LLM outperforms the other two models, achieving an accuracy of 92.4%. LLMs such as ChatGPT and Bing were evaluated using the average intraclass correlation coefficient (ICC) values. The ICC value for Bing was 0.975, whereas ChatGPT has an ICC value of 0.858. The higher mean ICC value indicates that Bing exhibited robust performance and consistency in major NLP tasks. Table 10 depicts that, all of the LLMs mentioned in the table have been analyzed and tested on multiple performance metrics and datasets to validate the robustness and reliability of these language models.

VI. RESOURCES OF LARGE LANGUAGE MODELS

LLMs have a wide range of potential applications and resources available for their development, deployment, and utilization. Figure 7 presents an LLM taxonomy that divided into two main branches: i) pre-trained model-based and ii) API-based. This taxonomy allows us to explore these two distinct aspects of LLMs.

A. PRETRAINED MODELS

Pretrained language models play a pivotal role in NLP tasks due to their ability to encapsulate broad language understanding and generation skills 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 LLMs models, which have been depicted in Table 11.

1) GENERATIVE PRETRAINED TRANSFORMER (GPT)

GPT [65] is an influential breakthrough in AI, particularly in NLP tasks. 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 in academia and industry, revolutionizing the landscape of language understanding and generation.

2) BERT

BERT [10], 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,

TABLE 10. Accuracy of various LLMs on different datasets.

LLMs	Accuracy	Task
GPT-1 [66]	68.4%, 48.4% , and 82%	Score of GPT-1 in standard NLP Modeling tasks GLUE, LAMBDA, and SQuAD 68.4,48.4 , and 82.0 respectively.
GPT-2 [96]	84.6%, 60.1%, and 89.5%	Score of GPT-2 in standard NLP Modeling tasks GLUE, LAMBDA, and SQuAD 84.6 ,60.1 , and 89.5 respectively.
GPT-3 [65]	93.2%, 69.6%, and 92.4%	Score of GPT-3 in standard NLP Modeling tasks GLUE, LAMBDA, and SQuAD 93.2,69.6, and 92.4 respectively.
GPT-3.5 [103]	93.5%, 79.3%, and 92.4	Score of GPT-3.5 in standard NLP Modeling tasks GLUE, LAMBDA, and SQuAD 93.5 ,79.3 , and 92.4 respectively.
	79.20%	GPT-3.5 is 79.2% performance on the VNHSGE English dataset.
GPT-4 [95]	85.50%	3 shot accuracy on MMLU across languages (English) 85.5%.
	94.2%, 82.4%, and 93.6%	Score of GPT-4 in standard NLP Modeling tasks GLUE, LAMBDA, and SQuAD 94.2 ,82.4 , and 93.6 respectively.
ChatGPT [104]	71% and 68%	A total of 167 SCORE and 112 Data-B questions were presented to the ChatGPT interface. ChatGPT correctly answered 71% and 68% of multiple-choice SCORE and Data-B questions, respectively.
	75.1% (SD 3%) and 64.5% (SD 5%)	The 5-year average percentage of correct answers for ChatGPT was 75.1% (SD 3%) for basic knowledge questions and 64.5% (SD 5%) for general questions.
	0.858 (95% CI: 0.777 to 0.91, p<0.0001)	The average intraclass correlation coefficient (ICC) values for ChatGPT were 0.858 (95% CI: 0.777 to 0.91, p<0.0001) with a total of 77 cases (answering case vignettes in physiology).
BingChat [105]	92.40%	92.4% performance on the VNHSGE English dataset.
Bard [97]	86%	86% performance on the VNHSGE English dataset.
Bing [106]	0.975 (95% CI: 0.961 to 0.984, p<0.0001)	The average intraclass correlation coefficient (ICC) values for Bing were 0.975 (95% CI: 0.961 to 0.984, p<0.0001) with a total of 77 cases (answering case vignettes in physiology).
BERT [66]	Dev 86.6%, Test 86.3%	BERT(large)'s performance on SWAG(Situations With Adversarial Generations) where Dev 86.6 ,Test 86.3.
	82.1%, grammatical 60.5%, sentiment analysis 94.9%, similarity 86.5%, paraphrase 89.3%, question similarity 72.1%, contradiction 86.7%, answerable 92.7%, and entail 70.1%	BERT(large)'s performance on GLUE(General Language Understanding Evaluation) where 82.1, grammatical 60.5, sentiment analysis 94.9, similarity 86.5, paraphrase 89.3, question similarity 72.1 , contradiction 86.7/85.9, answerable 92.7, and entail 70.1.

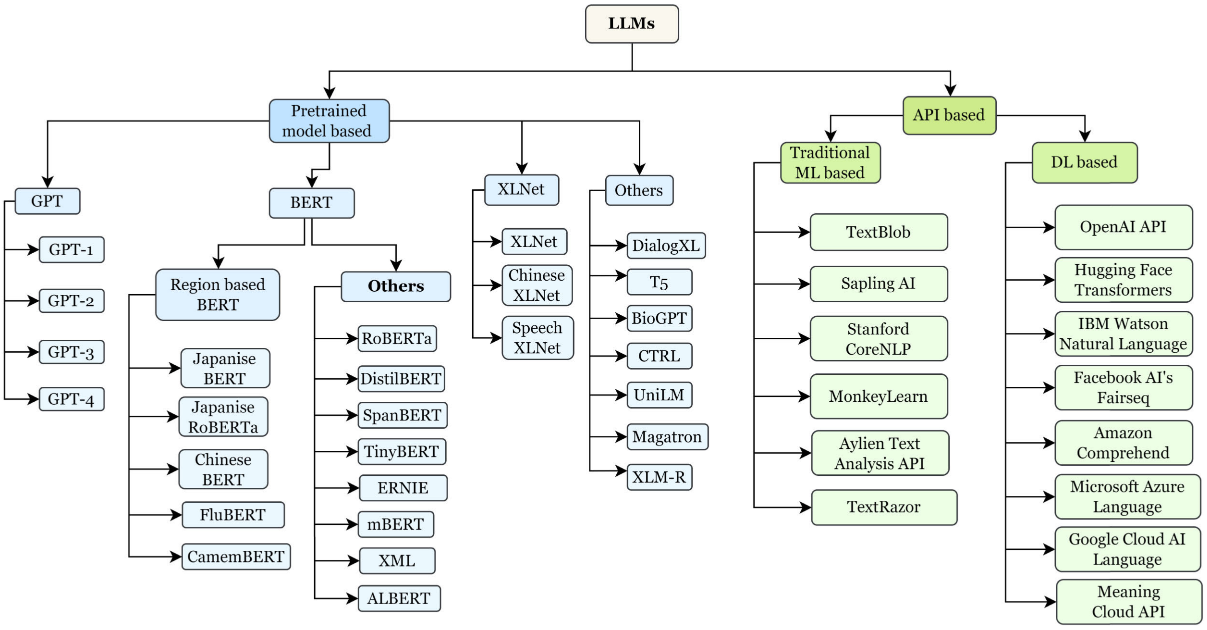


FIGURE 7. Taxonomy of LLMs.

achieving state-of-the-art results on different NLP 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).

TABLE 11. Description of LLMs.

Model Name	Description	Key Features	Training Data	Fine-Tuning Data	Fine-Tuning Tasks	Applications
GPT (Generative Pretrained Transformer) [65]	Transformative LLMs by OpenAI for versatile NLP tasks.	Extensive pre-training, deep language understanding, iterative improvements, and 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) [10]	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 [67]	BERT-based model with refined hyperparameters.	Significance of design decisions, publicly available, and top-tier NLP results	BookCorpus, Wikipedia	Task-specific datasets	Various NLP tasks	Benchmark improvements, research
XLNet [107]	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 [108]	Unsupervised acoustic model with robust regularization.	Robust regularizer, improved recognition accuracy	Speech datasets	TIMIT, WSJ datasets	Speech recognition	Speech recognition systems
DialogXL [109]	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) [84]	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 [110]	Specialized biomedical LLMs with state-of-the-art results.	Biomedical literature pretraining, excels in biomedical tasks	Biomedical literature	Biomedical datasets	Biomedical text analysis	Biomedical text analysis, research

In our analysis, we have considered 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 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 NLP. Non-LLMs versions of BERT or those with significantly reduced model sizes were excluded from our analysis to maintain consistency and relevance in our investigation.

3) ROBERTA

RoBERTA is another LLM which replicates the BERT pre-training approach outlined by Devlin et al. [67]. 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.

4) XLNET

XLNet [107] represents a versatile autoregressive pretraining approach that achieves bidirectional context learning by optimizing expected likelihood across all possible combinations. XLNet 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.

5) SPEECH-XLNET

Speech-XLNet [108] 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. 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. The model best achieves 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.

6) DIALOGXL

DialogXL [109] introduces enhancements to tackle longer historical context and multiparty structures in dialogues. Initially, alterations are made to how XLNet manages recurrence, transitioning from segment-level to utterance-level, thereby improving its effectiveness in modeling conversational data. Secondly, the integration of dialog-aware self-attention, as opposed to the standard self-attention in XLNet, enables capturing crucial dependencies within and between speakers. While training the DialogXL, a comprehensive set of experiments is conducted on four ERC benchmarks, comparing DialogXL with mainstream models. The experimental results consistently demonstrate that DialogXL outperforms the baseline models across all datasets.

7) T5

T5 (Text-to-Text Transfer Transformer) [84] is a groundbreaking LLM developed by Google Research, revolutionizing NLP tasks. 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.

8) BIOGPT

BioGPT [110] is a large-scale language model that was constructed by the Allen Institute for AI (AI2) with the explicit purpose of undertaking training on biomedical text. It was trained on an extensive corpus of biomedical literature, including PubMed abstracts and full-text articles, and is based on the GPT architecture. It has been demonstrated that BioGPT outperforms alternative biomedical language models across a range of tasks, such as query answering, relation extraction, and named entity recognition. The pre-trained weights of the model are accessible to the public, enabling researchers to optimize it using their biomedical text data. BioGPT has the capacity to substantially drive biomedical research forward by facilitating the analysis of vast quantities of biomedical text data in a more precise and efficient manner [111], [112].

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 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.

B. API OF LLMS

In this section, we discuss the APIs of LLMs, which have been described in Table 12.

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 [119]. The API 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 [120]. 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 [115]. The functionalities can be accessed by developers through REST API calls utilizing either the client libraries or their own custom libraries. Additionally, the API 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 [116]. Developers use RESTful endpoints to include Azure LLMs APIs. Microsoft provides useful SDKs and code examples in other programming languages, including Python, Java, etc. to facilitate the utilization of these APIs.

TABLE 12. Comparison of LLMs APIs.

API Name	Provider	Languages Supported	Access Type	Application Area	Advantages	Constraints
OpenAI API [113]	OpenAI	Multiple languages	API Key	NLP, text generation, chatbots	State-of-the-art models, versatility, and GPT architecture	API rate constrain, and cost considerations
Hugging Face Transformers [114]	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 [115]	Google Cloud	Multiple languages	API Key	Sentiment analysis, entity recognition, and translation	Google's robust infrastructure, easy integration	Cost may vary based on usage
Microsoft Azure Language [116]	Microsoft Azure	Multiple languages	API Key	Sentiment analysis, entity recognition, and language understanding	Integration with Azure services, comprehensive APIs	Pricing based on usage
IBM Watson NLU [117]	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 [118]	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 [118]	Facebook AI	Multiple languages	Open Source	Neural machine translation, language modeling, research, and development	Research-oriented, flexibility, open-source.	Self-hosting and maintenance complexity.

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 [117]. 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 [118]. 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, the 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 [121]. 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 12 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. These APIs are TextBlob, TextRazor, Sapling AI, MonkeyLearn, and Aylien, etc., which 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.

VII. 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 employed LLMs from the diverse domains such as healthcare, finance, education, forecasting, and natural language processing. The extensive experiments of different LLMs contribute to revolutionizing the use of AI across these domains. This section demonstrates the potential contribution of LLMs application in different domains. Table 13 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. GPT3 gets all required information through a conversation rather than an intake form, and many systems might be built to assist numerous patients at the same time [126]. 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 which becomes increasingly important during the COVID-19 epidemic [140]. 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, BERT models can improve the performance of biomedical and clinical text mining models [141]. Salam et al., [128] and Korngiebel et al., [126] 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 [125], [129], [131], [132] explore the utilities and constraints of LLMs such as ChatGPT in the context of clinical practice, research, and public health. In their study, Kung et al., [130] 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., [124] 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 decision alignment. Huang et al., [123] 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 [142]. Furthermore, various LLMs models also assist the medical industry by making the procedure easier than previously.

Education: Educators have struggled for a long time with an 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. In the light of above discussion, LLMs have the potential to revolutionize many facets of learning, teaching, and educational research in the education sector [140]. The GPT model aids the students in converting the math word problems into representative equations [143]. Kasenci et al., [19] highlighted substantial impact of LLMs in education by facilitating personalized learning, automating grading process, and accessibility of educational resources. Hadi et al., [137] 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., [138] and Dwivedi et. al. [139] 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 helps to understand the texts and documents which can be utilized in the academic sector [38]. Furthermore, other models may impact the education system by making it more engaging, accessible, and productive for both students and teachers.

Social Media: The LLMs have leveraged several aspects of the social media industry regarding content production, moderation, sentiment analysis, etc. There are some tasks in the social media can be generated such as writing content, classifying text, and even full blogs and articles for social media. These models can also perform named entity recognition (NER) and text classification [144], [145]. When the GPT, XLNet, BERT, etc., models aid the writer and content producers in generating a consistent flow of write up. 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. Abramski et al., [42] utilized network

TABLE 13. Domain specific machine learning-based study comparison in LLMs.

Domain	Author	Major Contributions	Limitations	Future Research Direction
Medical	Chen et al., [122] (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 LLMs 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., [123] (2023)	I. Investigates the possible utilization of LLMs, specifically ChatGPT and its variety within the domain of dentistry. II. Design a MultiModal LLMs 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., [124] (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., [125] (2023)	I. Focuses on the energy and environmental impact of training LLMs 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 LLMs on medical outcomes.
	Korngiebel et al., [126] (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., [127] (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 LLMs 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., [128] (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., [129] (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., [130] (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., [131] (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., [132] (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., [133] (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 13. *(Continued.)* Domain specific machine learning-based study comparison in LLMs.

Domain	Author	Major Contributions	Limitations	Future Research Direction
Industry	Yu et al., [134] (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., [135] (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., [136] (2022)	I. Examines the efficacy of employing LLMs 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 LLMs 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., [19] (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., [137] (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., [138] (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., [139] (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.

science and the principles of cognitive psychology to evaluate biases present in LLMs. Sobieszek et al., [136] 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, LLMs 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 [20]. In this case, LLMs help to fulfill all their requirements within a short period. The LLMs 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., [135] 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. [133] present an initial investigation of potential hazards associated with the implementation of ChatGPT in bussiness domain. Yu et al., [134] 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 [146], [147], [148]. They are able to analyze large data hubs of soil, crop, and weather data along with satellite imagery. These models provide recommendations on planting

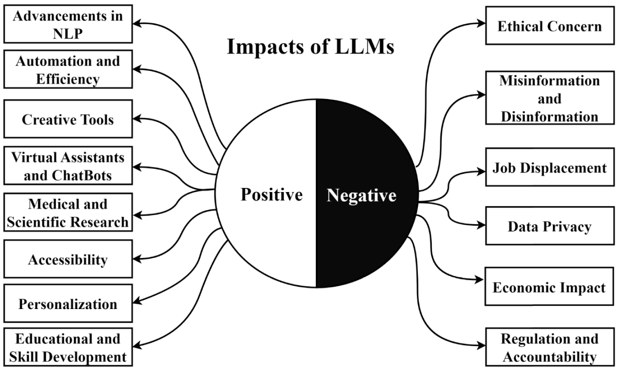

Impacts of LLMs
Positive
Negative
Advancements in NLP
Automation and Efficiency
Creative Tools
Virtual Assistants and ChatBots
Medical and Scientific Research
Accessibility
Personalization
Educational and Skill Development
Ethical Concern
Misinformation and Disinformation
Job Displacement
Data Privacy
Economic Impact
Regulation and Accountability

FIGURE 8. Visual representation of impact on LLMs.

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 agricultural management can be time-consuming and laborious, but these LLMs can support to accomplish these tasks to a greater extent.

VIII. IMPACT OF LARGE LANGUAGE MODELS ON SOCIETY

LLMs and similar AI technologies have had a profound impact on society across various domains. 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 8 effectively demonstrates the impact of LLMs, outlining their benefits on the left and the adversarial impacts on the right side. The positive impacts of LLMs are as follows:

- ***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 such as BERT [10], RoBERTa [67], and XLNet [107], have transformed the NLP landscape. Notably, LLMs have been fine-tuned for various specific NLP tasks, enabling remarkable performance improvements. Multilingual models like mBERT [149] and cross-lingual models like XLM-R [150] have facilitated language understanding across diverse linguistic contexts. Additionally, there has been a focus on creating more efficient versions of LLMs such as DistilBERT [151] and ALBERT [152]. These developments have not only expanded the applicability of NLP but have also raised ethical considerations, prompting research in bias mitigation [153] 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.
- ***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 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 [154] 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 [155], 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 [156], making them indispensable in automating domain-specific tasks efficiently.
- ***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.
- ***Language Translation:*** LLMs have improved machine translation systems, making communication across languages more accessible and accurate.
- ***Virtual Assistants and Chatbots:*** LLMs power virtual assistants and chatbots, enhancing customer service and providing round-the-clock support in various industries.
- ***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.
- ***Accessibility:*** LLMs have the potential to improve accessibility by providing real-time translation and transcription services for individuals with hearing impairments or language barriers.
- ***Personalization:*** LLMs enable personalized recommendations and content curation on platforms such as social media, e-commerce, and news websites.
- ***Creative Tools:*** LLMs are used as creative tools in various art forms, including generating poetry, music, and visual art.

- ***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.

In addition to numerous positive sides, LLMs also entail some downsides. These downsides are outlined as follows:

- ***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.
- ***Misinformation and Disinformation:*** LLMs can generate realistic-sounding fake text, raising concerns about the spread of misinformation and disinformation.
- ***Job Displacement:*** The automation capabilities of LLMs may lead to job displacement in certain industries, particularly in routine data-entry and content-generation roles.
- ***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.
- ***Economic Impact:*** The adoption of LLMs can disrupt traditional business models and create economic shifts as industries adapt to automation and AI technologies.
- ***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.

IX. INDUSTRIAL SIGNIFICANCE OF LARGE LANGUAGE MODELS

LLMs have gained substantial popularity in various industries, bringing about radical transformations. Influence of LLMs in industries is visible which can be presented through several key facets:

1. Enhancing NLP Applications: LLMs have ushered in a revolution in NLP applications [157] across sectors like customer service, chatbots, and sentiment analysis. They contribute to more precise and efficient interactions with users, leading to increased customer satisfaction and reduced response times.

2. Enabling Data Analysis and Information Extraction: LLMs play a pivotal role in extracting valuable insights from unstructured text data [158]. This is particularly critical in fields like finance, market research [159], and healthcare, where deciphering market trends, sentiment in news, or medical records hold paramount significance.

3. Facilitating Translation Services: Industries heavily reliant on multilingual communication [160], such as e-commerce, travel, and international business which may be benefited from LLMs that streamline automated translation. Translation service saves resources and ensuring high-quality translations across multiple languages.

4. Empowering Content Generation: LLMs are harnessed for content generation [161], which encompasses automated article writing, social media posts [162], product descriptions, and more. This automation simplifies content creation processes and allows for scalable production of top-tier content.

5. Revolutionizing Healthcare: LLMs find applications in medical record analysis [129], diagnosis assistance, and drug discovery. They empower healthcare professionals to access and comprehend extensive medical literature and patient data, thereby enhancing healthcare decision-making.

6. Revamping Education: The education sector [163] leverages LLMs for automated grading, ensuring prompt feedback to students. These models also contribute to the development of intelligent tutoring systems and personalized learning platforms.

7. Aiding Legal Practices: Legal practitioners [164] benefit from LLMs for contract analysis, legal research, and document review. These models assist in efficiently extracting pertinent information and identifying potential legal concerns.

8. Assisting Human Resources: LLMs support HR professionals [165] in tasks like candidate screening, resume parsing, and identifying potential job candidates. They streamline time-consuming processes within the recruitment phase.

9. Empowering Financial Services: In the realm of financial services [166], LLMs come into play for activities like sentiment analysis of news articles, algorithmic trading, risk assessment, and fraud detection. They are instrumental in making informed investment choices and managing financial risks.

10. Boosting E-commerce: LLMs enable personalized product recommendations [167], chatbots for customer support, and efficient inventory management. These enhancements result in enriched user experiences and heightened sales.

11. Illuminating Customer Insights: LLMs analyze customer reviews [168], feedback, and social media data, furnishing businesses with insights into customer preferences, opinions, and sentiments. This invaluable information aids companies in customizing their products and services.

As LLMs continue to advance, their industrial importance is undeniable. LLMs streamline operations, enhance decision-making, and bolster efficiency across diverse domains, positioning them as a transformative technology in the contemporary business landscape.

X. OPEN ISSUES AND CHALLENGES

This section discusses critical analysis of open issues and challenges of LLMs.

A. OPEN ISSUES

In this section, we delve into the open issues related to LLMs. These issues appeared recently as focal point in AI research and development. We raise the necessity 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 problematic. Misinformation, hate speech, and biased content generated by LLMs necessitate continuous research and development [169].
- **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 [170]. 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 [171]. 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 context, where slight input modifications can lead to an unexpected and potentially harmful outputs [172]. Improving model robustness and security against such situation 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 [173].
- **Issue 6: Generalization and Few-Shot Learning** LLMs performs well when there is abundant data but struggles 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 [174].
- **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 [175]. Global applications require developing techniques for cross-lingual transfer learning and low-resource language support.

B. 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. Its extraordinary ability to generate text that resembles that of a human which has attracted significant attention and applications in numerous fields. However, this sudden rise of these technological dependencies with higher impact has also revealed many challenges and concerns. 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 understand 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 [176].
- **Challenge 2: Tokenization Sensitivity** For analysis, LLMs rely significantly on tokenization, dividing text into smaller units (tokens) [177]. 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 [178]. 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 [179]. 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 [180]. 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 the relevant information.

- **Challenge 7: Bias and Undesirable Output**
  In the output, LLMs 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 [181]. 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.

XI. FUTURE RESEARCH PROSPECTS ON LLMS

Since LLMs are emerging research topic in recent times, several key research focuses and directions are prominent that may address and resolve the challenges and open issues discussed earlier. Resolving these open issues and challenges may harness the full potential of LLMs while ensuring its responsible and ethical use in AI landscape.

A. ENHANCING BIAS MITIGATION

Researchers are dedicated to refining training data to minimize bias, devising effective debiasing techniques, and establishing guidelines for responsible AI development [182]. They also need focus on integrating continuous monitoring and auditing mechanisms into AI pipelines, thereby conforming fairness and impartiality of the system. This commitment to mitigating bias ensures that LLMs not only advance in capability but LLMs also upholds ethical standards.

B. EFFICIENCY OPTIMIZATION

A core concern driving research is the quest of efficient training techniques. Researchers are delving into innovative methods like federated learning, which enables the distribution of training across decentralized data sources [183]. They are also exploring knowledge distillation techniques for model compression and finding ways to reduce the substantial computational and environmental costs associated with LLMs. This optimization paves the way for more sustainable and resource-efficient AI models.

C. DYNAMIC CONTEXT HANDLING

LLMs are being endowed with enhanced context management capabilities. This empowers them to comprehend longer context windows and seamlessly handle extensive documents or conversations. Such enhancements significantly expand their utility in various applications and resolve previous limitations.

D. CONTINUOUS LEARNING

To keep LLMs up-to-date, researchers are focusing on developing techniques that enable these models to adapt on evolving language and knowledge over time. This ensures that LLMs remain valuable and accurate sources of information and consistently overcoming challenges of being outdated.

E. INTERPRETABLE AI

The research community is committed to making LLMs' outputs more transparent and interpretable. Improving interpretability fosters the confidence and comprehension in AI decision-making processes which has been a major concern for a long time after the advent of LLMs [184].

F. MULTIMODAL LLMS

Researchers are pioneering the development of LLMs that incorporate text, vision, and other modalities [185]. These models can understand and generate text from images, videos, and audio, creating new avenues for AI applications and effectively addressing the need for multi-sensory comprehension.

G. HUMAN-AI COLLABORATION

Research on how humans and LLMs can collaborate effectively, with AI assisting and augmenting human tasks, is a crucial focal point. This collaboration bridges the gap between AI capabilities and human needs, thereby resolving previous challenges and issues in deployment.

H. DYNAMIC EVALUATION METRICS AND RELEVANT BENCHMARKS

Researchers are working on dynamic evaluation metrics that adapt to changing language and context, ensuring that LLMs performance is accurately assessed [186]. Finding a suitable metric along with the development of relevant and up-to-date benchmarks which may address earlier shortcomings in assessing AI capabilities.

I. PERSONALIZATION AND CUSTOMIZATION

Techniques to customize LLMs interactions to individual user preferences and needs are gaining popularity nowadays. This personalization boosts user satisfaction and resolves issues related to one-size-fits-all AI interactions.

J. ETHICAL AND LEGAL FRAMEWORKS

In response to evolving AI regulation, researchers are diligently developing ethical and legal regulatory frameworks. These frameworks serve as guiding principles for the responsible use of LLMs and ensure compliance with data protection and privacy regulations, effectively addressing previous concerns about ethical AI deployment [187].

These future research directions may overcome longstanding challenges and open issues raised in LLMs domain. These avenues may lead to the maximization of LLMs potential by the future researchers while upholding the highest standards of accountability and ethics in AI landscape.

XII. LIMITATIONS

While conducting a thorough examination of LLMs, which includes analyzing their taxonomies, comparing configurations, and addressing concerns and obstacles, it is essential to recognize the existence of limitations that should be considered. A primary limitation of this study is the unavailability of review papers that directly relate to the topic of LLMs. Although we have made diligent attempts to address the available research thoroughly, the limited quantity of papers in this field restricts our potential to perform broad comparisons and evaluations. While endeavoring to offer a broad perspective on LLMs concepts, we recognize that this analysis predominantly focuses on the ground-level concepts of LLMs configurations and applications. Limited resources, time, and page constraints affect the extensive exploration of individual LLMs architectures. Although our goal is not to offer the understanding of single LLMs but instead provide the evolution of LLMs and its application around various domains, however, readers looking for detailed analysis of specific architectures and advanced topics are not thoroughly covered. Furthermore, the impact of the LLMs across various domains, including education, health, and economy, is highlighted, but assessing the practical impacts of LLMs in many domains can be complex and subjective, especially when considering their impact on social aspects.

XIII. CONCLUSION

The field of LLMs has witnessed a remarkable evolution and expansion, resulting in extraordinary capabilities in NLP tasks and various applications in various areas. Based on neural networks and the changing 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. The study 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, the study has not addressed 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 LLMs research continues to evolve swiftly, this review could be a valuable resource for practitioners, researchers, and experts seeking a comprehensive understanding of LLMs' past, present, and future. The study 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 researcher to understand the dynamics of ever evolving Large Language Models research.

REFERENCES

[1] S. Pinker, *The Language Instinct: How the Mind Creates Language*. London, U.K.: Penguin, 2003.

[2] M. D. Hauser, N. Chomsky, and W. T. Fitch, "The faculty of language: What is it, who has it, and how did it evolve?" *Science*, vol. 298, no. 5598, pp. 1569–1579, Nov. 2002.

[3] W. X. Zhao et al., "A survey of large language models," 2023, *arXiv:2303.18223*.

[4] I. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, p. 433, 2007.

[5] Y. Shen, L. Heacock, J. Elias, K. D. Hentel, B. Reig, G. Shih, and L. Moy, "ChatGPT and other large language models are double-edged swords," *Radiology*, vol. 307, no. 2, Apr. 2023, Art. no. e230163.

[6] M. A. K. Raiaan, K. Fatema, I. U. Khan, S. Azam, M. R. U. Rashid, M. S. H. Mukta, M. Jonkman, and F. De Boer, "A lightweight robust deep learning model gained high accuracy in classifying a wide range of diabetic retinopathy images," *IEEE Access*, vol. 11, pp. 42361–42388, 2023.

[7] B. Ramabhadran, S. Khudanpur, and E. 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 *Proc. NAACL-HLT*, 2012, pp. 1–11.

[8] T. Mikolov, M. Karafiát, L. Burget, J. Černockỳ, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Annu. Conf. Int. Speech Commun. Assoc. (ISCA)*, 2010, pp. 1045–1048.

[9] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017.

[10] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," 2018, *arXiv:1810.04805*.

[11] Y. Khare, V. Bagal, M. Mathew, A. Devi, U. D. Priyakumar, and C. Jawahar, "MMBERT: Multimodal BERT pretraining for improved medical VQA," in *Proc. IEEE 18th Int. Symp. Biomed. Imag. (ISBI)*, Apr. 2021, pp. 1033–1036.

[12] R. Liu, C. Jia, J. Wei, G. Xu, L. Wang, and S. Vosoughi, "Mitigating political bias in language models through reinforced calibration," in *Proc. AAAI Conf. Artif. Intell.*, 2021, vol. 35, no. 17, pp. 14857–14866.

[13] K. Sanderson, "GPT-4 is here: What scientists think," *Nature*, vol. 615, no. 7954, p. 773, Mar. 2023.

[14] S. Pichai. (2023). *An Important Next Step on Our AI Journey*. [Online]. Available: https://blog.google/technology/ai/bard-google-ai-search-updates

[15] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, and T. B. Hashimoto. (2023). *Alpaca: A Strong, Replicable Instruction-following Model*. [Online]. Available: https://crfm.stanford.edu/2023/03/13/alpaca.html

[16] L. Fan, L. Li, Z. Ma, S. Lee, H. Yu, and L. Hemphill, "A bibliometric review of large language models research from 2017 to 2023," 2023, *arXiv:2304.02020*.

[17] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A survey on evaluation of large language models," 2023, *arXiv:2307.03109*.

[18] J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," 2022, *arXiv:2212.10403*.

[19] E. Kasneci et al., "ChatGPT for good? On opportunities and challenges of large language models for education," *Learn. Individual Differences*, vol. 103, Apr. 2023, Art. no. 102274.

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

[21] B. Cronin, "Annual review of information science and technology," Inf. Today, Medford, OR, USA, 2004, vol. 39.

[22] M. Kardum, "Rudolf Carnap—The grandfather of artificial neural networks: The influence of Carnap's philosophy on walter pitts," in *Guide to Deep Learning Basics*. Cham, Switzerland: Springer, 2020, pp. 55–66.

[23] G. Leech, "Corpora and theories of linguistic performance," *Svartvik, J. Directions Corpus Linguistics*, vol. 10, pp. 22–105, Jun. 1992.

[24] J. Hirschberg, B. W. Ballard, and D. Hindle, "Natural language processing," *AT&T Tech. J.*, vol. 67, no. 1, pp. 41–57, Jan. 1988.

[25] B.-H. Juang and L. R. Rabiner, "Automatic speech recognition—A brief history of the technology development," Georgia Inst. Technol., Santa Barbara, CA, USA, Tech. Rep., 2005, vol. 1, p. 67.

[26] D. S. Hain, R. Jurowetzki, T. Buchmann, and P. Wolf, "A text-embedding-based approach to measuring patent-to-patent technological similarity," *Technol. Forecasting Social Change*, vol. 177, Apr. 2022, Art. no. 121559.

[27] G. Curto, M. F. Jojoa Acosta, F. Comim, and B. Garcia-Zapirain, "Are AI systems biased against the poor? A machine learning analysis using Word2Vec and GloVe embeddings," *AI Soc.*, vol. 2022, pp. 1–16, Jun. 2022.

[28] P. Azunre, *Transfer Learning for Natural Language Processing*. New York, NY, USA: Simon and Schuster, 2021.

[29] Y. Shi, M. Larson, and C. M. Jonker, "Recurrent neural network language model adaptation with curriculum learning," *Comput. Speech Lang.*, vol. 33, no. 1, pp. 136–154, Sep. 2015.

[30] A. Kovačević and D. Kečo, "Bidirectional LSTM networks for abstractive text summarization," in *Advanced Technologies, Systems, and Applications VI*. Cham, Switzerland: Springer, 2021, pp. 281–293.

[31] Y. Wu et al., "Google's neural machine translation system: Bridging the gap between human and machine translation," 2016, *arXiv:1609.08144*.

[32] R. K. Yadav, S. Harwani, S. K. Maurya, and S. Kumar, "Intelligent chatbot using GNMT, SEQ-2-SEQ techniques," in *Proc. Int. Conf. Intell. Technol. (CONIT)*, Jun. 2021, pp. 1–5.

[33] D. Luitse and W. Denkena, "The great transformer: Examining the role of large language models in the political economy of AI," *Big Data Soc.*, vol. 8, no. 2, Jul. 2021, Art. no. 205395172110477.

[34] M. Onat Topal, A. Bas, and I. van Heerden, "Exploring transformers in natural language generation: GPT, BERT, and XLNet," 2021, *arXiv:2102.08036*.

[35] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, and M. Funtowicz, "TransFormers: State-of-the-art natural language processing," in *Proc. Conf. Empirical Methods Natural Lang. Syst. Demonstrations*, 2020, pp. 38–45.

[36] C. Sur, "RBN: Enhancement in language attribute prediction using global representation of natural language transfer learning technology like Google BERT," *Social Netw. Appl. Sci.*, vol. 2, no. 1, p. 22, Jan. 2020.

[37] J. J. Bird, A. Ekárt, and D. R. Faria, "Chatbot interaction with artificial intelligence: Human data augmentation with t5 and language transformer ensemble for text classification," *J. Ambient Intell. Humanized Comput.*, vol. 14, no. 4, pp. 3129–3144, Apr. 2023.

[38] B. D. Lund and T. Wang, "Chatting about ChatGPT: How may AI and GPT impact academia and libraries?" *Library Hi Tech News*, vol. 40, no. 3, pp. 26–29, May 2023.

[39] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever. *Improving Language Understanding by Generative Pre-Training*. Mikecaptain.com. Accessed: Feb. 15, 2024. [Online]. Available: https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

[40] B. Ghojogh and A. Ghodsi. *Attention Mechanism, Transformers, BERT, and GPT: Tutorial and Survey*. Osf.io. Accessed: Feb. 15, 2024. [Online]. Available: Osf.io.

[41] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, "Language models are unsupervised multitask learners," *OpenAI Blog*, vol. 1, no. 8, p. 9, 2019.

[42] K. Abramski, S. Citraro, L. Lombardi, G. Rossetti, and M. 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 Cognit. Comput.*, vol. 7, no. 3, p. 124, Jun. 2023.

[43] M. Shoeybi, M. Patwary, R. Puri, P. LeGresley, J. Casper, and B. Catanzaro, "Megatron-LM: Training multi-billion parameter language models using model parallelism," 2019, *arXiv:1909.08053*.

[44] D. M. Katz, M. J. Bommarito, S. Gao, and P. Arredondo, "GPT-4 passes the bar exam," Mar. 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4389233

[45] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, and S. E. Brennan, "The prisma 2020 statement: An updated guideline for reporting systematic reviews," *Int. J. Surg.*, vol. 88, Jan. 2020, Art. no. 105906.

[46] J. J. Webster and C. Kit, "Tokenization as the initial phase in NLP," in *Proc. 14th Conf. Comput. Linguistics (COLING)*, vol. 4, 1992.

[47] T. Kudo, "Subword regularization: Improving neural network translation models with multiple subword candidates," 2018, *arXiv:1804.10959*.

[48] R. Sennrich, B. Haddow, and A. Birch, "Neural machine translation of rare words with subword units," 2015, *arXiv:1508.07909*.

[49] M. Schuster and K. Nakajima, "Japanese and Korean voice search," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Mar. 2012, pp. 5149–5152.

[50] R. Child, S. Gray, A. Radford, and I. Sutskever, "Generating long sequences with sparse transformers," 2019, *arXiv:1904.10509*.

[51] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Netw.*, vol. 2, no. 5, pp. 359–366, Jan. 1989.

[52] V. Nair and G. E. Hinton, "Rectified linear units improve restricted Boltzmann machines," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 807–814.

[53] D. Hendrycks and K. Gimpel, "Gaussian error linear units (GELUs)," 2016, *arXiv:1606.08415*.

[54] B. Zhang and R. Sennrich, "Root mean square layer normalization," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[55] B. Workshop et al., "BLOOM: A 176B-parameter open-access multilingual language model," 2022, *arXiv:2211.05100*.

[56] B. Lester, R. Al-Rfou, and N. Constant, "The power of scale for parameter-efficient prompt tuning," 2021, *arXiv:2104.08691*.

[57] X. Lisa Li and P. Liang, "Prefix-tuning: Optimizing continuous prompts for generation," 2021, *arXiv:2101.00190*.

[58] H. W. Chung et al., "Scaling instruction-finetuned language models," 2022, *arXiv:2210.11416*.

[59] I. U. Khan, M. A. K. Raiaan, K. Fatema, S. Azam, R. U. Rashid, S. H. Mukta, M. Jonkman, and F. 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*, vol. 11, no. 6, p. 1566, May 2023.

[60] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing," *ACM Comput. Surv.*, vol. 55, no. 9, pp. 1–35, 2023.

[61] W. Ansar, S. Goswami, A. Chakrabarti, and B. Chakraborty, "A novel selective learning based transformer encoder architecture with enhanced word representation," *Appl. Intell.*, vol. 53, no. 8, pp. 9424–9443, Apr. 2023.

[62] G. Dar, M. Geva, A. Gupta, and J. Berant, "Analyzing transformers in embedding space," 2022, *arXiv:2209.02535*.

[63] D. Hazarika, M. Namazifar, and D. Hakkani-Tür, "Attention biasing and context augmentation for zero-shot control of encoder–decoder transformers for natural language generation," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 10, pp. 10738–10748.

[64] J. Lu, J. Yao, J. Zhang, X. Zhu, H. Xu, W. Gao, C. Xu, T. Xiang, and L. Zhang, "SOFT: Softmax-free transformer with linear complexity," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 34, 2021, pp. 21297–21309.

[65] L. Floridi and M. Chiriatti, "GPT-3: Its nature, scope, limits, and consequences," *Minds Mach.*, vol. 30, no. 4, pp. 681–694, Dec. 2020.

[66] X. Zheng, C. Zhang, and P. C. Woodland, "Adapting GPT, GPT-2 and BERT language models for speech recognition," in *Proc. IEEE Autom. Speech Recognit. Understand. Workshop (ASRU)*, Dec. 2021, pp. 162–168.

[67] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "RoBERTa: A robustly optimized BERT pretraining approach," 2019, *arXiv:1907.11692*.

[68] A. Roberts, C. Raffel, and N. Shazeer, "How much knowledge can you pack into the parameters of a language model?" 2020, *arXiv:2002.08910*.

[69] A. Chowdhery et al., "PaLM: Scaling language modeling with pathways," 2022, *arXiv:2204.02311*.

[70] R. Thoppilan et al., "LaMDA: Language models for dialog applications," 2022, *arXiv:2201.08239*.

[71] A. Zeng, X. Liu, Z. Du, Z. Wang, H. Lai, M. Ding, Z. Yang, Y. Xu, W. Zheng, X. Xia, W. L. Tam, Z. Ma, Y. Xue, J. Zhai, W. Chen, P. Zhang, Y. Dong, and J. Tang, "GLM-130B: An open bilingual pre-trained model," 2022, *arXiv:2210.02414*.

[72] J. W. Rae et al., "Scaling language models: Methods, analysis & insights from training gopher," 2021, *arXiv:2112.11446*.

[73] O. Lieber, O. Sharir, B. Lenz, and Y. Shoham, "Jurassic-1: Technical details and evaluation," *White Paper. AI21 Labs*, vol. 1, p. 9, 2021.

[74] S. Smith, M. Patwary, B. Norick, P. LeGresley, S. Rajbhandari, J. Casper, Z. Liu, S. Prabhumoye, G. Zerveas, V. Korthikanti, E. Zhang, R. Child, R. Yazdani Aminabadi, J. Bernauer, X. Song, M. Shoeybi, Y. He, M. Houston, S. Tiwary, and B. Catanzaro, "Using DeepSpeed and megatron to train megatron-turing NLG 530B, a large-scale generative language model," 2022, *arXiv:2201.11990*.

[75] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, "LLaMA: Open and efficient foundation language models," 2023, *arXiv:2302.13971*.

[76] T. Thi Nguyen, C. Wilson, and J. Dalins, "Fine-tuning llama 2 large language models for detecting online sexual predatory chats and abusive texts," 2023, *arXiv:2308.14683*.

[77] G. Penedo, Q. Malartic, D. Hesslow, R. Cojocaru, A. Cappelli, H. Alobeidli, B. Pannier, E. Almazrouei, and J. Launay, "The RefinedWeb dataset for falcon LLM: Outperforming curated corpora with web data, and web data only," 2023, *arXiv:2306.01116*.

[78] J. Hoffmann et al., "Training compute-optimal large language models," 2022, *arXiv:2203.15556*.

[79] S. Zhang, S. Roller, N. Goyal, M. Artetxe, M. Chen, S. Chen, C. Dewan, M. Diab, X. Li, X. Victoria Lin, T. Mihaylov, M. Ott, S. Shleifer, K. Shuster, D. Simig, P. Singh Koura, A. Sridhar, T. Wang, and L. Zettlemoyer, "OPT: Open pre-trained transformer language models," 2022, *arXiv:2205.01068*.

[80] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou, "Chain-of-thought prompting elicits reasoning in large language models," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 35, 2022, pp. 24824–24837.

[81] W. Zeng et al., "PanGu-α: Large-scale autoregressive pretrained Chinese language models with auto-parallel computation," 2021, *arXiv:2104.12369*.

[82] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Akhtar, N. Barnes, and A. Mian, "A comprehensive overview of large language models," 2023, *arXiv:2307.06435*.

[83] G. Jawahar, B. Sagot, and D. Seddah, "What does BERT learn about the structure of language?" in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 3651–3657.

[84] J. Ni, G. Hernández Ábrego, N. Constant, J. Ma, K. B. Hall, D. Cer, and Y. Yang, "Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models," 2021, *arXiv:2108.08877*.

[85] Y. Li et al., "Competition-level code generation with AlphaCode," *Science*, vol. 378, no. 6624, pp. 1092–1097, Dec. 2022.

[86] Y. Koizumi, Y. Ohishi, D. Niizumi, D. Takeuchi, and M. Yasuda, "Audio captioning using pre-trained large-scale language model guided by audio-based similar caption retrieval," 2020, *arXiv:2012.07331*.

[87] W. Fan, Z. Zhao, J. Li, Y. Liu, X. Mei, Y. Wang, Z. Wen, F. Wang, X. Zhao, J. Tang, and Q. Li, "Recommender systems in the era of large language models (LLMs)," 2023, *arXiv:2307.02046*.

[88] Y. Bai, J. Yi, J. Tao, Z. Tian, Z. Wen, and S. Zhang, "Fast end-to-end speech recognition via non-autoregressive models and cross-modal knowledge transferring from BERT," *IEEE/ACM Trans. Audio, Speech, Language Process.*, vol. 29, pp. 1897–1911, 2021.

[89] C. Sun, J. Li, Y. R. Fung, H. P. Chan, T. Abdelzaher, C. Zhai, and H. Ji, "Decoding the silent majority: Inducing belief augmented social graph with large language model for response forecasting," 2023, *arXiv:2310.13297*.

[90] K. Drossos, S. Gharib, P. Magron, and T. Virtanen, "Language modelling for sound event detection with teacher forcing and scheduled sampling," 2019, *arXiv:1907.08506*.

[91] S.-H. Chiu and B. Chen, "Innovative bert-based reranking language models for speech recognition," in *Proc. IEEE Spoken Lang. Technol. Workshop (SLT)*, Jan. 2021, pp. 266–271.

[92] A. Elhafsi, R. Sinha, C. Agia, E. Schmerling, I. A. D. Nesnas, and M. Pavone, "Semantic anomaly detection with large language models," *Auto. Robots*, vol. 47, no. 8, pp. 1035–1055, Dec. 2023.

[93] Y. Shen, J. Shao, X. Zhang, Z. Lin, H. Pan, D. Li, J. Zhang, and K. B. Letaief, "Large language models empowered autonomous edge AI for connected intelligence," 2023, *arXiv:2307.02779*.

[94] H. Abdel-Jaber, D. Devassy, A. A. Salam, L. Hidaytallah, and M. El-Amir, "A review of deep learning algorithms and their applications in healthcare," *Algorithms*, vol. 15, no. 2, p. 71, Feb. 2022.

[95] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with GPT-4," 2023, *arXiv:2304.03277*.

[96] J. Vig and Y. Belinkov, "Analyzing the structure of attention in a transformer language model," 2019, *arXiv:1906.04284*.

[97] A. McGowan, Y. Gui, M. Dobbs, S. Shuster, M. Cotter, A. Selloni, M. Goodman, A. Srivastava, G. A. Cecchi, and C. M. Corcoran, "ChatGPT and bard exhibit spontaneous citation fabrication during psychiatry literature search," *Psychiatry Res.*, vol. 326, Aug. 2023, Art. no. 115334.

[98] R. Taylor, M. Kardas, G. Cucurull, T. Scialom, A. Hartshorn, E. Saravia, A. Poulton, V. Kerkez, and R. Stojnic, "Galactica: A large language model for science," 2022, *arXiv:2211.09085*.

[99] N. Shazeer, "GLU variants improve transformer," 2020, *arXiv:2002.05202*.

[100] N. Du, Y. Huang, A. M. Dai, S. Tong, D. Lepikhin, Y. Xu, M. Krikun, Y. Zhou, A. W. Yu, and O. Firat, "GLaM: Efficient scaling of language models with mixture-of-experts," in *Proc. Int. Conf. Mach. Learn.*, 2022, pp. 5547–5569.

[101] S. Black, S. Biderman, E. Hallahan, Q. Anthony, L. Gao, L. Golding, H. He, C. Leahy, K. McDonell, J. Phang, M. Pieler, U. Sai Prashanth, S. Purohit, L. Reynolds, J. Tow, B. Wang, and S. Weinbach, "GPT-NeoX-20B: An open-source autoregressive language model," 2022, *arXiv:2204.06745*.

[102] E. Nijkamp, B. Pang, H. Hayashi, L. Tu, H. Wang, Y. Zhou, S. Savarese, and C. Xiong, "CodeGen: An open large language model for code with multi-turn program synthesis," 2022, *arXiv:2203.13474*.

[103] T. Hagendorff, S. Fabi, and M. Kosinski, "Thinking fast and slow in large language models," 2022, *arXiv:2212.05206*.

[104] P. P. Ray, "ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope," *Internet Things Cyber-Phys. Syst.*, vol. 3, pp. 121–154, Jan. 2023.

[105] X.-Q. Dao, "Performance comparison of large language models on VNHSGE English dataset: OpenAI chatGPT, Microsoft bing chat, and Google bard," 2023, *arXiv:2307.02288*.

[106] D. Kelly, Y. Chen, S. E. Cornwell, N. S. Delellis, A. Mayhew, S. Onaolapo, and V. L. Rubin, "Bing chat: The future of search engines?" *Proc. Assoc. Inf. Sci. Technol.*, vol. 60, no. 1, pp. 1007–1009, Oct. 2023.

[107] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 32, 2019.

[108] X. Song, G. Wang, Z. Wu, Y. Huang, D. Su, D. Yu, and H. Meng, "Speech-XLNet: Unsupervised acoustic model pretraining for self-attention networks," 2019, *arXiv:1910.10387*.

[109] W. Shen, J. Chen, X. Quan, and Z. Xie, "DialogXL: All-in-one XLNet for multi-party conversation emotion recognition," in *Proc. AAAI Conf. Artif. Intell.*, vol. 35, no. 15, 2021, pp. 13789–13797.

[110] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, and T.-Y. Liu, "BioGPT: Generative pre-trained transformer for biomedical text generation and mining," *Briefings Bioinf.*, vol. 23, no. 6, Nov. 2022, Art. no. bbac409.

[111] D. Deutsch, J. Juraska, M. Finkelstein, and M. Freitag, "Training and meta-evaluating machine translation evaluation metrics at the paragraph level," 2023, *arXiv:2308.13506*.

[112] A. Ushio, F. Alva-Manchego, and J. Camacho-Collados, "Generative language models for paragraph-level question generation," 2022, *arXiv:2210.03992*.

[113] (2023). *OpenAI*. Accessed: Sep. 12, 2023. [Online]. Available: https://openai.com/blog/openai-api

[114] (2023). *Huggingface*. Accessed: Sep. 12, 2023. [Online]. Available: https://huggingface.co/docs/transformers/index

[115] (2023). *Google Cloud*. Accessed: Sep. 12, 2023. [Online]. Available: https://cloud.google.com/natural-language

[116] (2023). *Azure*. Accessed: Sep. 12, 2023. [Online]. Available: https://azure.microsoft.com/en-us/products/ai-services/ai-language

[117] IBM. (2023). *IBM Watson Natural Language Understanding*. Accessed: Sep. 12, 2023. [Online]. Available: https://www.ibm.com/products/natural-language-understanding

[118] G. Satyanarayana, J. Bhuvana, and M. Balamurugan, "Sentimental analysis on voice using AWS comprehend," in *Proc. Int. Conf. Comput. Commun. Informat. (ICCCI)*, Jan. 2020, pp. 1–4.

[119] A. Kolides, A. Nawaz, A. Rathor, D. Beeman, M. Hashmi, S. Fatima, D. Berdik, M. Al-Ayyoub, and Y. Jararweh, "Artificial intelligence foundation and pre-trained models: Fundamentals, applications, opportunities, and social impacts," *Simul. Model. Pract. Theory*, vol. 126, Jul. 2023, Art. no. 102754.

[120] S. M. Jain, "Hugging face," in *Introduction to Transformers for NLP: With Hugging Face Library Models to Solve Problems*. Cham, Switzerland: Springer, 2022, pp. 51–67.

[121] J. Wang, X. Hu, W. Hou, H. Chen, R. Zheng, Y. Wang, L. Yang, H. Huang, W. Ye, X. Geng, B. Jiao, Y. Zhang, and X. Xie, "On the robustness of ChatGPT: An adversarial and out-of-distribution perspective," 2023, *arXiv:2302.12095*.

[122] S. Chen, Y. Li, S. Lu, H. Van, H. J. Aerts, G. K. Savova, and D. S. Bitterman, "Evaluation of ChatGPT family of models for biomedical reasoning and classification," 2023, *arXiv:2304.02496*.

[123] H. Huang, O. Zheng, D. Wang, J. Yin, Z. Wang, S. Ding, H. Yin, C. Xu, R. Yang, Q. Zheng, and B. Shi, "ChatGPT for shaping the future of dentistry: The potential of multi-modal large language model," *Int. J. Oral Sci.*, vol. 15, no. 1, p. 29, Jul. 2023.

[124] V. Sorin, E. Klang, M. Sklair-Levy, I. Cohen, D. B. Zippel, N. Balint Lahat, E. Konen, and Y. Barash, "Large language model (ChatGPT) as a support tool for breast tumor board," *NPJ Breast Cancer*, vol. 9, no. 1, p. 44, May 2023.

[125] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," *Nature Med.*, vol. 29, no. 8, pp. 1930–1940, 2023.

[126] D. M. Korngiebel and S. D. Mooney, "Considering the possibilities and pitfalls of generative pre-trained transformer 3 (GPT-3) in healthcare delivery," *npj Digit. Med.*, vol. 4, no. 1, p. 93, Jun. 2021.

[127] L. De Angelis, F. Baglivo, G. Arzilli, G. P. Privitera, P. Ferragina, A. E. Tozzi, and C. Rizzo, "ChatGPT and the rise of large language models: The new AI-driven infodemic threat in public health," *Frontiers Public Health*, vol. 11, Apr. 2023, Art. no. 1166120.

[128] M. Sallam, "ChatGPT utility in healthcare education, research, and practice: Systematic review on the promising perspectives and valid concerns," *Healthcare*, vol. 11, no. 6, p. 887, Mar. 2023.

[129] M. Cascella, J. Montomoli, V. Bellini, and E. Bignami, "Evaluating the feasibility of ChatGPT in healthcare: An analysis of multiple clinical and research scenarios," *J. Med. Syst.*, vol. 47, no. 1, p. 33, Mar. 2023.

[130] T. H. Kung, M. Cheatham, A. Medenilla, C. Sillos, L. De Leon, C. Elepaño, M. Madriaga, R. Aggabao, G. Diaz-Candido, J. Maningo, and V. Tseng, "Performance of ChatGPT on USMLE: Potential for AI-assisted medical education using large language models," *PLOS Digit. Health*, vol. 2, no. 2, Feb. 2023, Art. no. e0000198.

[131] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, and H. Poon, "Domain-specific language model pretraining for biomedical natural language processing," *ACM Trans. Comput. Healthcare*, vol. 3, no. 1, pp. 1–23, Oct. 2021.

[132] Z. Kraljevic, D. Bean, A. Shek, R. Bendayan, H. Hemingway, and J. Au, "Foresight-generative pretrained transformer (GPT) for modelling of patient timelines using EHRs," 2022, *arXiv:2212.08072*.

[133] L. Mich and R. Garigliano, "ChatGPT for e-tourism: A technological perspective," *Inf. Technol. Tourism*, vol. 25, no. 1, pp. 1–12, Mar. 2023.

[134] X. Yu, Z. Chen, Y. Ling, S. Dong, Z. Liu, and Y. Lu, "Temporal data meets LLM—Explainable financial time series forecasting," 2023, *arXiv:2306.11025*.

[135] G. F. Frederico, "ChatGPT in supply chains: Initial evidence of applications and potential research agenda," *Logistics*, vol. 7, no. 2, p. 26, Apr. 2023.

[136] A. Sobieszek and T. Price, "Playing games with ais: The limits of GPT-3 and similar large language models," *Minds Mach.*, vol. 32, no. 2, pp. 341–364, Jun. 2022.

[137] M. U. Hadi, R. Qureshi, A. Shah, M. Irfan, A. Zafar, M. B. Shaikh, N. Akhtar, J. Wu, and S. Mirjalili, "Large language models: A comprehensive survey of its applications, challenges, limitations, and future prospects," Tech. Rep., 2023.

[138] C. K. Lo, "What is the impact of ChatGPT on education? A rapid review of the literature," *Educ. Sci.*, vol. 13, no. 4, p. 410, Apr. 2023.

[139] Y. K. Dwivedi, N. Kshetri, L. Hughes, E. L. Slade, A. Jeyaraj, A. K. Kar, A. M. Baabdullah, A. Koohang, V. Raghavan, and M. Ahuja, "'So what if chatgpt wrote it?' multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy," *Int. J. Inf. Manage.*, vol. 71, Jul. 2023, Art. no. 102642.

[140] M. Zong and B. Krishnamachari, "A survey on GPT-3," 2022, *arXiv:2212.00857*.

[141] R. Zhu, X. Tu, and J. X. Huang, "Utilizing BERT for biomedical and clinical text mining," in *Data Analytics in Biomedical Engineering and Healthcare*. Amsterdam, The Netherlands: Elsevier, 2021, pp. 73–103.

[142] K. Huang, A. Singh, S. Chen, E. T. Moseley, C.-Y. Deng, N. George, and C. Lindvall, "Clinical XLNet: Modeling sequential clinical notes and predicting prolonged mechanical ventilation," 2019, *arXiv:1912.11975*.

[143] J. Zhang, L. Wang, R. K. W. Lee, Y. Bin, Y. Wang, J. Shao, and E. P. Lim, "Graph-to-tree learning for solving math word problems," in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 3928–3937.

[144] X. Dai, S. Karimi, B. Hachey, and C. Paris, "Cost-effective selection of pretraining data: A case study of pretraining BERT on social media," 2020, *arXiv:2010.01150*.

[145] S. Biswas, "The function of chat GPT in social media: According to chat GPT," Mar. 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4405389

[146] R. Peng, K. Liu, P. Yang, Z. Yuan, and S. Li, "Embedding-based retrieval with LLM for effective agriculture information extracting from unstructured data," 2023, *arXiv:2308.03107*.

[147] S. Biswas, "Importance of chat GPT in agriculture: According to chat GPT," Mar. 2023. [Online]. Available: http://dx.doi.org/10.2139/ssrn.4405391

[148] M. A. K. Raiaan, N. M. Fahad, S. Chowdhury, D. Sutradhar, S. S. Mihad, and M. M. Islam, "IoT-based object-detection system to safeguard endangered animals and bolster agricultural farm security," *Future Internet*, vol. 15, no. 12, p. 372, Nov. 2023.

[149] T. Pires, E. Schlinger, and D. Garrette, "How multilingual is multilingual BERT?" 2019, *arXiv:1906.01502*.

[150] A. Conneau, K. Khandelwal, N. Goyal, V. Chaudhary, G. Wenzek, F. Guzmán, E. Grave, M. Ott, L. Zettlemoyer, and V. Stoyanov, "Unsupervised cross-lingual representation learning at scale," 2019, *arXiv:1911.02116*.

[151] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter," 2019, *arXiv:1910.01108*.

[152] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "ALBERT: A lite BERT for self-supervised learning of language representations," 2019, *arXiv:1909.11942*.

[153] T. Bolukbasi, K.-W. Chang, J. Y. Zou, V. Saligrama, and A. T. Kalai, "Man is to computer programmer as woman is to homemaker? Debiasing word embeddings," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 29, 2016.

[154] Y. Zhang, S. Sun, M. Galley, Y.-C. Chen, C. Brockett, X. Gao, J. Gao, J. Liu, and B. Dolan, "DialoGPT: Large-scale generative pre-training for conversational response generation," 2019, *arXiv:1911.00536*.

[155] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell, "Language models are few-shot learners," in *Proc. NIPS*, 2020, pp. 1877–1901.

[156] J. Lee, W. Yoon, S. Kim, D. Kim, S. Kim, C. H. So, and J. Kang, "BioBERT: A pre-trained biomedical language representation model for biomedical text mining," *Bioinformatics*, vol. 36, no. 4, pp. 1234–1240, Feb. 2020.

[157] Y. Gat, N. Calderon, A. Feder, A. Chapanin, A. Sharma, and R. Reichart, "Faithful explanations of black-box NLP models using LLM-generated counterfactuals," 2023, *arXiv:2310.00603*.

[158] M. Josifoski, M. Sakota, M. Peyrard, and R. West, "Exploiting asymmetry for synthetic training data generation: SynthIE and the case of information extraction," 2023, *arXiv:2303.04132*.

[159] M. S. H. Mukta, J. Ahmad, M. A. K. Raiaan, S. Islam, S. Azam, M. E. Ali, and M. Jonkman, "An investigation of the effectiveness of deepfake models and tools," *J. Sensor Actuator Netw.*, vol. 12, no. 4, p. 61, Aug. 2023.

[160] A. Awasthi, N. Gupta, B. Samanta, S. Dave, S. Sarawagi, and P. Talukdar, "Bootstrapping multilingual semantic parsers using large language models," 2022, *arXiv:2210.07313*.

[161] P. Sridhar, A. Doyle, A. Agarwal, C. Bogart, J. Savelka, and M. Sakr, "Harnessing LLMs in curricular design: Using GPT-4 to support authoring of learning objectives," 2023, *arXiv:2306.17459*.

[162] M. A. K. Raiaan, A. Al Mamun, Md. A. Islam, M. E. Ali, and Md. S. H. Mukta, "Envy prediction from Users' photos using convolutional neural networks," in *Proc. Int. Conf. Comput., Electr. Commun. Eng. (ICCECE)*, Jan. 2023, pp. 1–7.

[163] E. Waisberg, J. Ong, M. Masalkhi, and A. G. Lee, "Large language model (LLM)-driven chatbots for neuro-ophthalmic medical education," *Eye*, vol. 2023, pp. 1–3, Sep. 2023.

[164] W. Channell, "Making a difference: The role of the LLM in policy formulation and reform," in *The Export of Legal Education*. Evanston, IL, USA: Routledge, 2016, pp. 13–21.

[165] P. Budhwar et al., "Human resource management in the age of generative artificial intelligence: Perspectives and research directions on ChatGPT," *Hum. Resource Manage. J.*, vol. 33, no. 3, pp. 606–659, Jul. 2023.

[166] G. Fatouros, J. Soldatos, K. Kouroumali, G. Makridis, and D. Kyriazis, "Transforming sentiment analysis in the financial domain with ChatGPT," *Mach. Learn. Appl.*, vol. 14, Dec. 2023, Art. no. 100508.

[167] Y. Li, S. Ma, X. Wang, S. Huang, C. Jiang, H.-T. Zheng, P. Xie, F. Huang, and Y. Jiang, "EcomGPT: Instruction-tuning large language models with chain-of-task tasks for e-commerce," 2023, *arXiv:2308.06966*.

[168] P. Weingart, T. Wambsganss, and M. Soellner, "A taxonomy for deriving business insights from user-generated content," ECIS, Res. Papers 401, 2023. [Online]. Available: https://aisel.aisnet.org/ecis2023_rp/401

[169] L. Zhu, X. Xu, Q. Lu, G. Governatori, and J. Whittle, "AI and ethics—Operationalizing responsible AI," in *Humanity Driven AI: Productivity, Well-being, Sustainability and Partnership*. Cham, Switzerland: Springer, 2022, pp. 15–33.

[170] I. Molenaar, S. D. Mooij, R. Azevedo, M. Bannert, S. Järvelä, and D. Gašević, "Measuring self-regulated learning and the role of AI: Five years of research using multimodal multichannel data," *Comput. Hum. Behav.*, vol. 139, Feb. 2023, Art. no. 107540.

[171] M. C. Rillig, M. Ågerstrand, M. Bi, K. A. Gould, and U. Sauerland, "Risks and benefits of large language models for the environment," *Environ. Sci. Technol.*, vol. 57, no. 9, pp. 3464–3466, Mar. 2023.

[172] B. Liu, B. Xiao, X. Jiang, S. Cen, X. He, and W. Dou, "Adversarial attacks on large language model-based system and mitigating strategies: A case study on ChatGPT," *Secur. Commun. Netw.*, vol. 2023, pp. 1–10, Jun. 2023.

[173] Z. Sun, "A short survey of viewing large language models in legal aspect," 2023, *arXiv:2303.09136*.

[174] Y. Meng, M. Michalski, J. Huang, Y. Zhang, T. Abdelzaher, and J. Han, "Tuning language models as training data generators for augmentation-enhanced few-shot learning," in *Proc. Int. Conf. Mach. Learn.*, 2023, pp. 24457–24477.

[175] S. Fincke, S. Agarwal, S. Miller, and E. Boschee, "Language model priming for cross-lingual event extraction," in *Proc. AAAI Conf. Artif. Intell.*, 2022, vol. 36, no. 10, pp. 10627–10635.

[176] N. M. Fahad, S. Sakib, M. A. Khan Raiaan, and Md. S. Hossain Mukta, "SkinNet-8: An efficient CNN architecture for classifying skin cancer on an imbalanced dataset," in *Proc. Int. Conf. Electr., Comput. Commun. Eng. (ECCE)*, Feb. 2023, pp. 1–6.

[177] N. Jain, K. Saifullah, Y. Wen, J. Kirchenbauer, M. Shu, A. Saha, M. Goldblum, J. Geiping, and T. Goldstein, "Bring your own data! Self-supervised evaluation for large language models," 2023, *arXiv:2306.13651*.

[178] X. Zhu, J. Li, Y. Liu, C. Ma, and W. Wang, "A survey on model compression for large language models," 2023, *arXiv:2308.07633*.

[179] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "MT5: A massively multilingual pre-trained text-to-text transformer," 2020, *arXiv:2010.11934*.

[180] N. Ratner, Y. Levine, Y. Belinkov, O. Ram, I. Magar, O. Abend, E. Karpas, A. Shashua, K. Leyton-Brown, and Y. Shoham, "Parallel context windows for large language models," 2022, *arXiv:2212.10947*.

[181] F. Motoki, V. Pinho Neto, and V. Rodrigues, "More human than human: Measuring ChatGPT political bias," *Public Choice*, vol. 2023, pp. 1–21, Aug. 2023.

[182] K. Werder, B. Ramesh, and R. Zhang, "Establishing data provenance for responsible artificial intelligence systems," *ACM Trans. Manage. Inf. Syst.*, vol. 13, no. 2, pp. 1–23, Jun. 2022.

[183] J. Jiang, X. Liu, and C. Fan, "Low-parameter federated learning with large language models," 2023, *arXiv:2307.13896*.

[184] W. S. Saba, "Towards explainable and language-agnostic LLMs: Symbolic reverse engineering of language at scale," 2023, *arXiv:2306.00017*.

[185] Z. Zhang, A. Zhang, M. Li, H. Zhao, G. Karypis, and A. Smola, "Multimodal chain-of-thought reasoning in language models," 2023, *arXiv:2302.00923*.

[186] Z. Liu, Y. Zhang, P. Li, Y. Liu, and D. Yang, "Dynamic LLM-agent network: An LLM-agent collaboration framework with agent team optimization," 2023, *arXiv:2310.02170*.

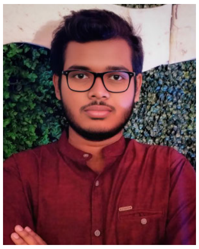
[187] U. Iqbal, T. Kohno, and F. Roesner, "LLM platform security: Applying a systematic evaluation framework to OpenAI's ChatGPT plugins," 2023, *arXiv:2309.10254*.

MOHAIMENUL AZAM KHAN RAIAAN received the Bachelor of Science degree in computer science and engineering from United International University (UIU), in 2023. Currently, he is a Research Assistant with the Computer Science and Engineering Department, UIU. His professional pursuits are marked by active involvement in diverse research areas, such as computer vision, health informatics, explainable artificial intelligence, and graph optimization. Notably, he has made significant contributions to the field, as evidenced by his multiple research articles published in prestigious journals indexed by Scopus and categorized under the Q1 ranking.

MD. SADDAM HOSSAIN MUKTA received the Ph.D. degree from the Data Science and Engineering Research Laboratory (Data Laboratory), BUET, in 2018. He is a Postdoctoral Researcher with the LUT School of Engineering Sciences, Lappeenranta, Finland. He was an Associate Professor and a Undergraduate Program Coordinator with the Department of Computer Science and Engineering, United International University, Bangladesh. He has a number of quality publications in both national and international conferences and journals. His research interests include deep learning, machine learning, data mining, and social computing.

KANIZ FATEMA received the bachelor's degree in computer science and engineering from Daffodil International University, Dhaka, Bangladesh. She is currently a Research Assistant (RA) with Charles Darwin University. She is actively involved in research activities, especially in health informatics, computer vision, machine learning, deep learning, and artificial intelligence-based systems. She has published several research papers in journals (Scopus) and international conferences.

NUR MOHAMMAD FAHAD received the bachelor's degree from the Department of Computer Science and Engineering, United International University (UIU), Bangladesh. During the bachelor's study, he has contributed to the academic community as an Undergraduate Teaching Assistant with the Department of Computer Science and Engineering, UIU. In addition to his teaching role, he has deeply engaged in cutting-edge research across several domains, including computer vision, machine learning, deep learning, health informatics, graph theory, and mental health modeling.



SADMAN SAKIB received the bachelor's degree in computer science and engineering from the Department of Computer Science and Engineering, United International University, Bangladesh, in 2023. He was a Teaching Assistant of undergraduate students with the Department of Computer Science and Engineering, United International University. Besides this, he is actively involved in machine learning, deep learning, artificial intelligence, computer vision, and health informatics research.

MOST MARUFATUL JANNAT MIM is currently pursuing the degree with the Computer Science and Engineering Department, United International University (UIU). She is actively involved in research activities related to computer vision, deep learning, graph theory, and human–computer interaction. Her passion lies in pioneering innovative research in computer science. Apart from studies, she is involved in co-curricular activities with the UIU APP Forum, where she is also the President and demonstrates strong leadership by organizing various seminars and workshops for computer science students.

JUBAER AHMAD received the B.Sc. degree in computer science and engineering from United International University (UIU), Dhaka, Bangladesh, in 2022. He is currently a Research Assistant with the IAR Project, UIU. His research interests include computer vision, NLP, big data, and distributed learning.

MOHAMMED EUNUS ALI is a Professor with the Department of CSE, Bangladesh University of Engineering and Technology (BUET), where he is also the Group Leader of the Data Science and Engineering Research Laboratory (DataLab). His research papers have published in top ranking journals and conferences, such as the *VLDB Journal*, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, *DMKD*, *Information Systems Journal*, *WWWJ*, *DKE*, ICDE, CIKM, EDBT, PVLDB, and UbiComp. His research interests include database systems and information management, including spatial databases, practical machine learning, and social media analytics. He served as a Program Committee Member for many prestigious conferences, including SIGMOD, VLDB, AAAI, and SIGSPATIAL.

SAMI AZAM is a leading Researcher and a Professor with the Faculty of Science and Technology, Charles Darwin University, Australia. He is actively involved in the research fields relating to computer vision, signal processing, artificial intelligence, and biomedical engineering. He has a number of publications in peer-reviewed journals and international conference proceedings.