Nanotechnology Perceptions
ISSN 1660-6795
www.nano-ntp.com

Comprehensive Review of Large Language Models and its Applications

Dr. Vinayak Kottawar[1], Pranjali Bahalkar[2], Priyanka Deshpande[3], Gajanan R Bhusare[4], Nivedita Shimbre[5], Sayam Palrecha[5]

D.Y Patil College of Engineering, Akurdi[1]
Dr. D Y Patil Institute of Technology, Pimpri[2]
Modern College of Engineering, Pune[3]
Zeal College of Engineering & Research, Pune[4]
International Institute of Information technology, Pune[5]
Email's- viny2612@gmail.com

In the realm of contemporary academic discussions and scholarly conversations, it has become increasingly evident that Large Language Models (LLMs), which are sophisticated artificial intelligence systems designed for understanding and generating human language, have showcased remarkable and exceptional levels of proficiency across a wide variety of complex and intricate natural language processing (NLP) tasks that are crucial to various applications. This manuscript, which serves as a comprehensive and thorough examination of the field of Large Language Models (LLMs), aims to delve deeply into and investigate the latest and most significant innovations, advancements, and breakthroughs that have emerged within this rapidly evolving discipline and area of study. To begin this exploration, we will carefully elucidate and clarify the fundamental principles and core concepts that serve as the foundational underpinnings of these advanced models, and then we will proceed to categorize them systematically according to their unique architectural compositions and structural characteristics. Furthermore, a detailed and comparative analysis of the dominant methodologies employed in this field will be conducted, which will involve a careful delineation of their respective advantages and disadvantages, particularly in relation to their architectural design features and the empirical results they yield in practical applications. In addition to this, we will investigate and explore prospective avenues for further research and potential domains that could benefit from future inquiry and exploration within this fascinating and significant area of study. Finally, the manuscript will critically evaluate and assess the practical implications of the performance and assessment of LLMs, while also considering their environmental impact and the broader implications of their deployment in real-world scenarios.

Keywords: Large Language Models, GPT, Pre-Trained Models, Deep Learning, Transformer, Natural Language Processing.

1. Introduction

The remarkable capability to express and communicate one's thoughts and ideas through various linguistic methods stands as a prominent and distinguishing feature of the species known as Homo sapiens, which begins to develop and take shape during the early stages of human growth and continues to evolve and refine itself throughout the entirety of one's life cycle. This particular cognitive skill is not only fundamental for a wide range of intellectual activities and pursuits, which encompass decision-making processes, critical thinking, and the resolution of complex problems, but it is also absolutely essential for the effective assimilation of sophisticated and multifaceted information as well as for drawing well-reasoned and substantiated conclusions based on that information. In spite of the considerable advancements that have been achieved in the field of artificial intelligence (AI), it is important to note that machines still lack the intrinsic capability to fully understand and adeptly utilize human language, unless they are equipped with highly sophisticated and advanced AI algorithms that enable such functionalities.

Transformers, which are identified as a specific type of neural network architecture that has been meticulously designed for the purpose of processing sequential data, frequently serve as the foundational building blocks upon which large language models (LLMs) are constructed and developed. These transformers utilize self-attention mechanisms, which play a critical role in determining the contextual relationships among the various input tokens, thereby facilitating the effective capturing of long-range dependencies and the subtle contextual nuances that are present in the data. On the date of March 14, 2023, an announcement was made public regarding the launch of the most up-to-date large language model (LLM) developed by OpenAI, which has attracted significant interest and excitement within the technological community. A multitude of widely recognized and popular AI applications that have been developed by OpenAI, among which the extensively employed ChatGPT is a notable example, are fundamentally based on this innovative LLM, which has received considerable attention and acclaim in the realm of LLM research. Since their emergence in 2018, LLMs have demonstrated remarkable effectiveness across a variety of tasks and applications, leading to a transformative shift in the landscape of natural language processing (NLP) research as we know it today. Prominent examples of widely recognized LLMs include the Bidirectional Encoder Representations from Transformers (BERT) created by the tech giant Google, the Generative Pre-trained Transformer (GPT) that was developed by OpenAI, and the Large Language Model Meta AI (LLaMa) that has been introduced by Meta, all of which have made significant contributions to the field. Specialized applications tailored to specific domains, including medical and health sciences, highlight the immense potential that LLMs possess for effectively tackling a broad spectrum of NLP tasks, a point that has been previously highlighted and emphasized in the existing scientific literature. The exploration and investigation of LLMs have gained substantial momentum over time, inciting researchers to push the boundaries of development in creating and refining advanced models such as BERT and GPT, which are pivotal in this area of study. As a result, it has

become increasingly critical and necessary to review the latest contemporary research surrounding LLMs in order to achieve a thorough and comprehensive understanding of the current state and landscape of this evolving field.

This paper aims to elucidate the current state of research related to large language models, alongside a detailed overview of the most recent state-of-the-art models and their applications to challenges and markets that have previously gone unaddressed. In addition to this, we engage in a meaningful discourse regarding the evaluation and assessment of these models, which serves to illuminate the ongoing research initiatives and efforts within the domain of large language models. The importance and significance of this paper can be delineated into two primary aspects or facets that are worth noting. Firstly, this study offers a contemporary and up-to-date examination of the latest inquiries and investigations concerning large language models, highlighting the emerging trends and patterns that enhance our overall understanding of the current scenario and context. Moreover, our paper functions as an invaluable navigational resource for researchers, practitioners, and policymakers alike, enabling them to effectively navigate the current landscape of LLM research, identify potential avenues for further exploration and future research opportunities, and foster innovation and progress within this dynamic field. Furthermore, we conduct a thorough scrutiny of the evaluation processes associated with these models, as well as their environmental ramifications, which is an increasingly relevant topic in today's context.

2. Literature Review

The remarkable ability of generative artificial intelligence models to tackle a wide range of complex challenges related to language has attracted significant interest and focus from not only the general public but also from scholarly communities and academic institutions that are deeply engaged with the field of artificial intelligence (AI). A multitude of factors has contributed to the enhancement of the capabilities and effectiveness of these advanced models. It was in the year of 1996 that the very first natural language processing (NLP) program, which was named ELIZA[4], was creatively developed, allowing for specific forms of natural language communication and interaction to take place between humans and machines in a more seamless manner. Following this pioneering development, advancements began to emerge during the late 1990s and into the early 2000s in the realm of Statistical NLP[13], as highlighted by Joakim NIVRE's academic publication titled "On Statistical Methods in Natural Language Processing," which coincided with a notable increase in research initiatives that were centered around deep learning, particularly in the wake of the revolutionary arrival of deep neural networks. During this same significant period, impressive progress was made in the area of Statistical NLP[13], as further illustrated by Joakim NIVRE's previously mentioned scholarly paper. However, alongside the growing enthusiasm for deep learning, which was characterized by the introduction of sophisticated deep neural network architectures, a considerable volume of research activities was actively pursued in this particular domain. After the introduction of the innovative Transformer[3] model in the year 2017, which utilizes self-attention mechanisms to effectively capture long-range dependencies and contextual information by computing the contextual relationships

that exist among input tokens, researchers and scholars have been able to build upon the foundational framework established by this remarkable model.

The self-attention mechanism that is fundamentally integrated within the Transformer[3] provides it with the unique capability to allocate varying weights to individual tokens, thus allowing it to focus on and attend to all other tokens present within the input sequence simultaneously. By leveraging the Transformer as a foundational model, researchers have subsequently crafted increasingly sophisticated language models, which has resulted in a remarkable and significant acceleration in the progression and development of NLP as a whole. The intricate architecture of the Transformer model will be analyzed and discussed in detail in the sections that follow this introduction. A pivotal milestone was achieved in the year 2018 with the introduction of a novel pre-training methodology known as Bidirectional Encoder Representations from Transformers (BERT)[5]. BERT employs deep bidirectional representations by conditioning on contextual information derived from both the left and right contexts across all layers, which makes it applicable to a wide variety of tasks, such as question answering and language inference, through the process of fine-tuning with the addition of an extra output layer. The overwhelming success exhibited by BERT has led to its widespread adoption across various applications and prompted the subsequent development of a plethora of other pre-trained language models. Nevertheless, one significant limitation that has been observed in relation to BERT pertains to its considerable computational expense, which can hinder its accessibility. In the year 2019, the Generative Pre-trained Transformer 2, which is commonly referred to in the field as GPT-2[6], was unveiled to the public by Alec Radford and his colleagues at OpenAI, having been trained using a staggering 1.5 billion parameters. The transformer architecture that is utilized within this particular model employs self-attention mechanisms to effectively aggregate data from multiple locations throughout the input sequence, thereby enhancing its overall performance. Although this model incurs a substantial amount of computational costs during both its training and execution phases, its impressive scale allows it to grasp and generate a wide range of linguistic subtleties and diverse types of outputs. In the same year, NVIDIA also made its mark by introducing Megatron-LM[14], yet another large language model (LLM) that contains an astounding 8.3 billion parameters, a figure that significantly exceeds the 1.5 billion parameters of GPT-2. This considerable size empowers the model to capture, analyze, and produce more intricate linguistic patterns while concurrently enhancing its overall understanding when compared to the earlier GPT model. However, the model's significant size, along with its exorbitant computational requirements, continues to pose a notable disadvantage that cannot be overlooked. In the year 2020, OpenAI proudly launched its most advanced and sophisticated LLM to date, known as GPT-3[15], which boasts an astonishing 175 billion parameters, thereby surpassing the capabilities and performance levels of any previous large language model. The exceptional performance exhibited by GPT-3 across a multitude of domains has led to the development of various applications, including ChatGPT, which are constructed upon the foundation of this powerful model. The trajectory of growth for large language models was significantly propelled forward by the introduction of GPT-3, which set new standards in the field. MetaAI has also emerged as a prominent leader in the ongoing development of large language models, particularly with the introduction of LLaMA[7], which is a series of foundational language models that range in size from 7 billion to an impressive 65 billion parameters, demonstrating performance that

has outperformed GPT-3 (which has 175 billion parameters) in numerous benchmarks that have been established based on the transformer architecture[3]. On March 27, 2023, OpenAI unveiled its most sophisticated and current large language model, known as GPT-4, which approaches the remarkable scale of nearly 1 trillion parameters, thus exceeding the capabilities of its predecessor by a factor of more than six. GPT-4, which is a model that is fundamentally based on the Transformer architecture, is pre-trained to predict the next token in a document and has demonstrated performance levels that are remarkably comparable to human capability across a variety of professional and academic benchmarks, including achieving a score within the top 10% of participants on a simulated bar examination[16]. Nonetheless, it continues to fall short of human performance in various real-world scenarios. Overall, the evolution from a rudimentary program facilitating communication between humans and computers[4] to the extensive and complex capabilities of the 1 trillion parameter GPT-4[5] exemplifies the remarkable advancements made in the realm of Large Language Models.

Paper	Abstract	Summary	Key Insights	Results
GPT-3 (Brown et al., 2020)	Introduces GPT-3, a 175 billion parameter autoregressive language model that excels in few-shot learning.	GPT-3 pushes the limits of model size, showcasing high performance across various NLP tasks with minimal fine-tuning.	GPT-3 demonstrates powerful text generation but requires significant computational resources.	Sets new performance benchmarks in tasks like text generation and translation.
BERT (Devlin et al., 2019)	Presents BERT, a bidirectional transformer model designed for masked language modeling.	BERT's bidirectional attention improves NLP tasks like question answering, while limiting generative capabilities.	BERT is better for understanding and classification tasks due to its bidirectional context mechanism.	Achieves state-of-the-art results on benchmarks such as GLUE, SQuAD, and SWAG.
T5 (Raffel et al., 2020)	T5 reformulates all NLP problems as text-to-text tasks, creating a unified framework for a wide variety of applications.	T5 excels in task-specific fine-tuning, showing remarkable adaptability.	The flexibility of T5 allows it to be applied to different NLP tasks with minimal reconfiguration.	Breaks new ground in benchmarks for translation, summarization, and comprehension.
Efficient Transformers (Tay et al., 2020)	Discusses optimizations for transformer models, introducing sparse attention mechanisms and reducing computation.	Introduces methods to enhance transformer efficiency while preserving performance.	Highlights trade-offs between reducing model complexity and maintaining accuracy.	Achieves comparable results with fewer parameters and computational costs.
Switch Transformer (Fedus et al., 2021)	Switch Transformer introduces mixture-of-experts layers, scaling models efficiently by activating only parts of the network.	Mixture-of-experts architecture significantly reduces computational requirements.	Improves model efficiency by using expert layers while maintaining accuracy.	Demonstrates superior scaling properties compared to dense transformers, with lower computational costs.
XLNet (Yang et al., 2019)	Proposes XLNet, a generalized autoregressive pretraining method that incorporates bidirectional context without masking.	XLNet overcomes the limitations of BERT by combining autoregressive and bidirectional models.	Achieves better contextual understanding than BERT while enabling generative capabilities.	Outperforms BERT on multiple tasks, such as question answering and text classification.
ALBERT (Lan et al., 2019)	ALBERT reduces the memory footprint of BERT by sharing parameters across layers	Parameter-sharing enables scaling up models without a proportional increase in	ALBERT maintains the performance of BERT with fewer parameters and lower	Reduces memory and time complexity while achieving competitive performance in NLP

	and factorizing embeddings.	memory consumption.	computational cost.	benchmarks.
Megatron-LM (Shoeybi et al., 2019)	Introduces a large-scale transformer model optimized for multi-GPU training using model parallelism.	Megatron-LM allows training of very large models by partitioning computation across GPUs.	Demonstrates the benefits of model parallelism in training extremely large LLMs.	Achieves significant speedups in training for large models without loss in accuracy.
RoBERTa (Liu et al., 2019)	RoBERTa improves upon BERT by optimizing training procedures and using larger datasets.	RoBERTa avoids the limitations of BERT's original training method and generalizes better across tasks.	Larger datasets and longer training improve the robustness and generalizability of RoBERTa.	Outperforms BERT in several NLP tasks, establishing new state-of-the-art results.
Ethical Risks in AI (Bender et al., 2021)	Discusses the ethical risks of large-scale LLMs, such as bias and sustainability challenges.	Focuses on the potential harms of large-scale LLMs and calls for responsibleAI development.	Identifies significant risks associated with model biases and environmental costs of training.	Raises awareness of ethical challenges, encouraging more responsible development and deployment practices.

Large Language Model-

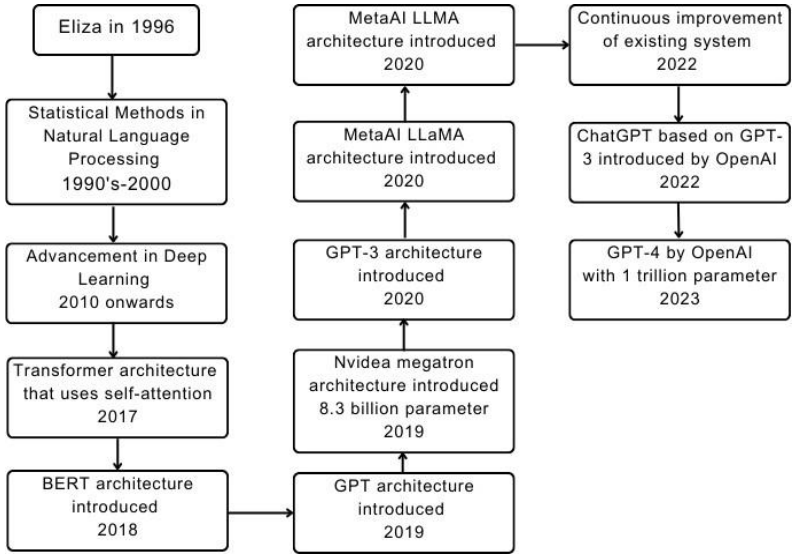

Figure. 1 Timeline of existing large language models

3. Recent development in large language models (LLM)

3.1 Transformer

This influential paper outlines and presents a revolutionary neural network architecture that is referred to as the Transformer, which stands out in stark contrast to traditional methodologies that typically rely on either recurrence or convolutional mechanisms to process data; instead, this innovative model functions exclusively on the basis of attention mechanisms, thereby representing a significant departure from established practices in the field. The authors of this pivotal work assert and maintain that the Transformer exhibits a level of efficiency and efficacy that is markedly superior when compared to earlier models

that sought to combine attention mechanisms with recurrent networks, highlighting its advanced capabilities. The architecture that constitutes both the encoder and decoder components within the Transformer framework is meticulously composed of multiple layers, each incorporating self-attention mechanisms as well as feed-forward neural networks, which collectively empower the model to assign varying degrees of importance, or differential weights, to diverse segments of the input sequence it processes.

This sophisticated self-attention mechanism not only enhances the model's ability to generate the output sequence but also significantly improves the accuracy and relevance of the generated outputs, thereby contributing to the overall effectiveness of the model. In addition to this, the feed-forward networks embedded within the architecture facilitate complex non-linear transformations that apply to both the input sequences being processed and the output sequences being generated, allowing for a richer representation of the data. The authors of the study conduct a thorough evaluation and assessment of the Transformer architecture across two distinct machine translation tasks, which ultimately reveals that this innovative model outperforms previous state-of-the-art models in terms of overall performance while also demonstrating notable enhancements in computational efficiency, making it an attractive option for various applications. Specifically, the Transformer achieves a level of translation quality that is decidedly superior when juxtaposed with earlier models on the WMT 2014 English-to-German and English-to-French translation tasks, showcasing its remarkable capabilities in practical applications. Furthermore, the authors illustrate and provide evidence that the Transformer can be trained in a significantly shorter time frame compared to its predecessors, a remarkable achievement that can be attributed to the effective parallelization of the self-attention mechanism, which allows for a more streamlined training process.

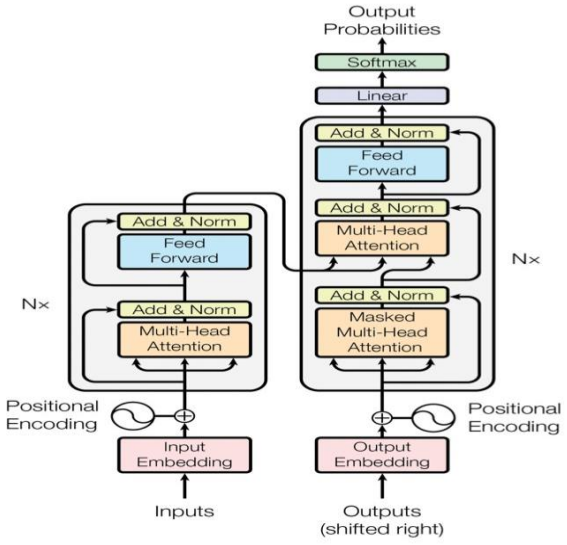

Figure. 2 Transformer architecture [3]

3.2 BERT

In the comprehensive manuscript referenced as [5], the authors put forth an exceptionally innovative language model that has been appropriately designated as BERT, which stands for Bidirectional Encoder Representations derived from Transformers, thereby highlighting the model's sophisticated architecture and groundbreaking capabilities. The BERT model is designed to pre-train bidirectional representations by effectively utilizing vast amounts of unlabelled text, which involves joint conditioning on contextual information derived from both the left and right sides across all layers of the neural network, enhancing its understanding of language in a holistic manner. The authors meticulously delineated two distinct pre-training objectives that are central to the operation of BERT: the first being masked language modeling (MLM), which allows BERT to transcend the substantial limitations typically associated with conventional unidirectional language models, as it predicts the original vocabulary identifier of a word that has been masked based solely on the rich contextual information surrounding it. Conversely, the second objective, known as next sentence prediction (NSP), is designed to facilitate the pre-training process of text-pair representations, thereby enabling the model to grasp the relationships between sentences more effectively. Through their extensive research, the manuscript provides substantial evidence that the implementation of bidirectional pretraining is fundamentally instrumental in the formation and development of accurate and nuanced language representations, which are crucial for various natural language processing applications. By strategically incorporating an additional output layer within the model architecture, BERT is equipped to undergo fine-tuning processes that cultivate state-of-the-art models capable of addressing a remarkably diverse range of tasks, which prominently include question answering and language inference, showcasing its versatility and efficacy.

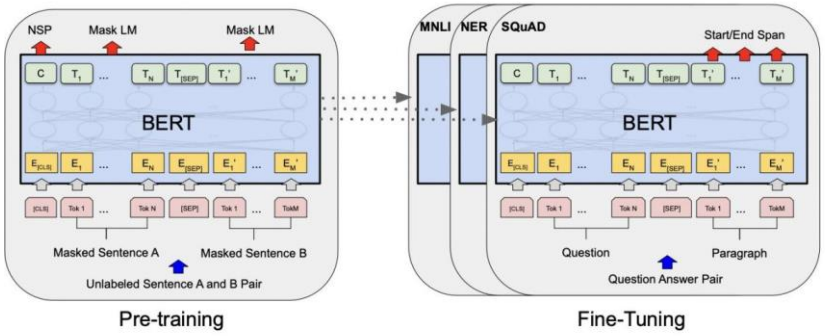

Figure. 3: Overall pre-training and fine-tuning procedures for BERT Source. [5]

3.3 Generative Pre-Trained Transformer (GPT)

In the manuscript referenced as [6], the authors meticulously articulate a comprehensive methodology that is specifically designed with the intention of significantly enhancing the comprehension of natural language; this methodology involves a dual-phase process in which the model first undergoes a rigorous training regimen characterized by an initial phase of generative pre-training that utilizes a diverse and heterogeneous corpus composed entirely

of unannotated text, which is then subsequently succeeded by a focused phase of discriminative fine-tuning that zeroes in on discrete and specific tasks. The authors convincingly demonstrate that by employing this innovative methodology, it is indeed possible to attain substantial and noteworthy improvements across a wide array of natural language processing tasks, a feat that not only highlights the effectiveness of their approach but also indicates that it surpasses the performance metrics established by other models that have been trained on labeled data specifically for the purpose of task-specific learning.

The empirical assessments that have been meticulously conducted by the authors encompass a comprehensive and broad spectrum of benchmarks pertinent to natural language, which include various tasks such as textual entailment, question answering, and document classification, thereby illustrating the versatility and robustness of their proposed methodology. By incorporating task-aware input transformations throughout the entire fine-tuning process, the methodology put forth is able to facilitate an efficient and smooth transfer of learning, all the while ensuring that there are minimal alterations or disruptions to the existing model architecture, which is a crucial aspect of maintaining performance stability.

These evaluations conducted by the authors are not merely academic exercises; rather, they play a pivotal role in substantiating the efficacy of the proposed methodology through both empirical demonstration and rigorous analytical scrutiny, thus reinforcing its validity and applicability. The GPT system, which is fundamentally based on the transformer architecture, has experienced extensive and widespread utilization across a diverse range of natural language processing tasks, which attests to its adaptability and effectiveness in various contexts. The results of the studies underscore and highlight the extraordinary superiority and effectiveness of the proposed methodology when compared to antecedent techniques, thereby establishing new and impressive benchmarks in performance across an extensive array of natural language understanding tasks that have garnered significant attention in the field.

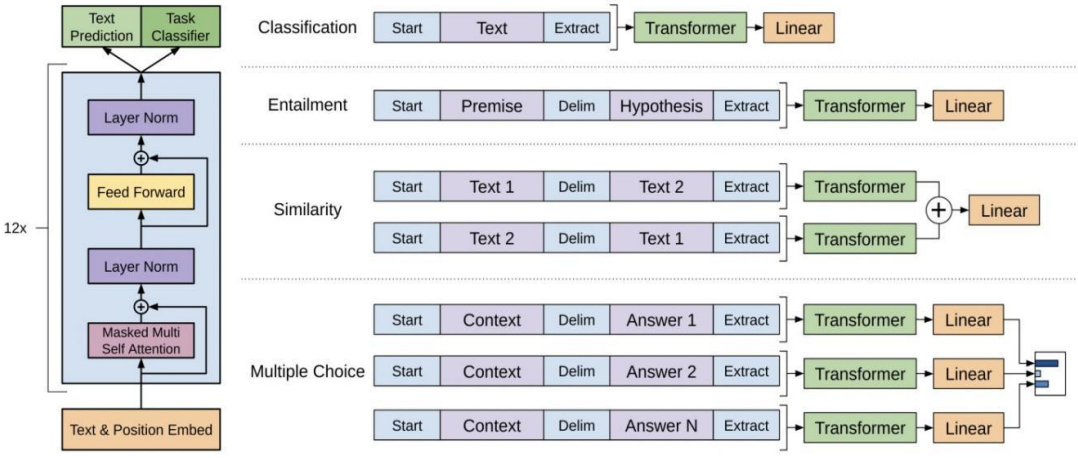

Figure. 4: GPT Architecture

4 Application of large language models

Large Language Models (LLMs) have recently garnered considerable attention because of their capacity to generate writing that is similar to a human, comprehend context, and automate labour-intensive jobs. As LLMs continue to advance, their economic impact on various industries and occupations is becoming increasingly important to understand. Here we shall take forth the economic impact and application in diverse fields like finance[8], Labour markets[9], Oil and gas[10] and Computational Biology[18].

4.1 BloombergGPT

The academic article [8] meticulously presents an intricately designed language model that boasts an impressive total of 50 billion parameters, specifically engineered to cater to the unique requirements of the financial sector, demonstrating a deep understanding of the complexities involved in this industry. The training regimen implemented for this model incorporates a comprehensive dataset that envelops a vast array of financial information, with the primary objective of facilitating and enhancing the execution of a diverse range of natural language processing tasks that are critical within the finance discipline. The methodology meticulously delineated in the research employs an innovative mixed dataset training approach, which culminates in the development of a model that exhibits markedly superior performance when compared to existing models that are utilized for financial applications, all the while maintaining a high degree of efficacy on general large language model benchmarks. This well-thought-out approach effectively addresses the intricate and complex characteristics as well as the specialized lexicon that are prevalent in the financial domain, thereby underscoring the pressing necessity for a model that has been specifically crafted and tailored for this particular sector. The findings that emerged from this thorough investigation indicate that BloombergGPT significantly surpasses other comparably-sized open models when it comes to executing financial NLP tasks, thus illustrating the remarkable efficacy of the proposed strategy in the development of a large-scale generative artificial intelligence model that is specifica $\bar{h}_\ell = h_{\ell-1} + \mathrm{SA}(\mathrm{LN}(h_{\ell-1}))$ turn paves the way for subsequent advanceme $h_\ell = \bar{h}_\ell + \mathrm{FFN}(\mathrm{LN}(\bar{h}_\ell))$ this burgeoning field. The model that has been introduced by the authors is fundamentally predicated on BLOOM and operates as a decoder-only causal language model, comprising an impressive total of 70 layers of transformer decoder blocks, as specified in the details provided below:

4.2 Computational Biology

The paper [18] the authors propose a method for applying self-supervised deep learning to protein sequences. The authors conducted training on four different models, including two auto-regressive language models.(Transformer-XL and XLNet) and two autoencoder models (BERT and ALBERT). The training data consisted of amino acid sequences from UniRef and BFD, encompassing a total of 393 billion amino acids derived from 2.1 billion protein sequences.

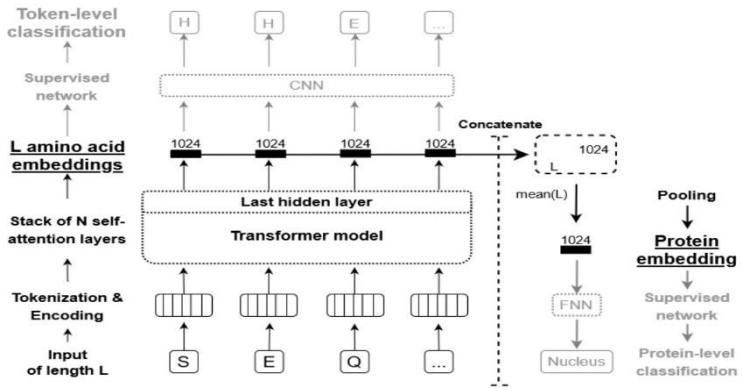

Figure. 5 BLOOM Architecture Source [18]

5. Evaluation

In the paper [20], the author presents an empirical evaluation of outputs generated by nine widely-available large language models (LLMs) using off-the-shelf tools. The study found a correlation between the percentage of memorised text, percentage of unique text, and overall output quality. Outputs with more memorised content were more likely to be considered high quality.

However, the paper also identified output pathologies such as counterfactual and logically-flawed statements, as well as general failures like not staying on topic. Overall, 80% of the evaluated outputs contained memorised data.

Model	Memorized (↓)	Original (↑)	PII (↓)	Logical (↓)	Factual (↓)	Discourse (↓)
BLOOM	**53.3/22.7**	91.5	14.7	45.3	49.3	73.3
ChatGPT	87.8 / 37.3	71.7	**0.0**	**7.6**	37.9	24.2
Galactica	72.0 / 26.7	92.0	31.1	52.0	73.3	77.3
GPT-3.5	86.7 / 56.0	73.0	16.0	20.0	**34.7**	45.3
GPT-3.51	89.3 / 56.0	71.2	5.3	13.3	40.0	**18.7**
GPT-4	94.7 / 58.7	73.8	9.3	17.3	**34.7**	36.0
LLaMA	82.7 / 58.7	78.1	9.3	17.3	36.0	56.0
OPT	65.3 / 26.7	**93.0**	25.3	40.0	48.0	78.7
OPT-IML	74.7 / 34.7	91.9	21.3	54.7	53.3	62.7
Total	80.0 / 39.8	82.3	15.4	31.3	46.4	52.0
Total (*)	76.5 / 35.9	82.6	15.6	28.1	51.0	52.6
Total (†)	58.7 / 25.7	62.9	11.1	20.8	42.6	39.3

Table. 1 Main results of the analysis Source

Over the past some time the trend in LLM evaluation paper has increased and presents a clear picture that there is need of more research and time to evaluate LLM and check their ability to give proper output [21].This trend is also marked by a great surge of papers on this topic.

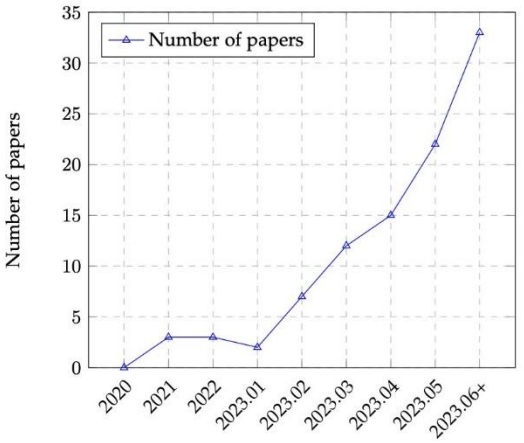

Figure. 6 Trend of LLMs evaluation papers over time (2020 - Jun. 2023

Another important evaluation of LLM capability was tested by authors in 2023's Iranian residency entrance examination where they used the famous LLM model ChatGPT to evaluate its capability in multilingual settings (English,Persian, French, Spanish) and the model was able to show an accuracy of 81.3% by answering 161 out of the 198 questions..

Question Topic	Number of Question	Number of Correct Responses (Percentage)			
		Persian	English	French	Spanish
Pediatrics	26	20 (76.9)	22 (84.6)	21 (80.8)	22 (84.6)
Obstetrics and Gynecology	18	16 (88.9)	16 (88.9)	15 (83.3)	15 (83.3)
General Surgery	23	16 (69.6)	20 (87)	19 (82.6)	19 (82.6)
Internal Medicine	45	35 (77.8)	34 (75.6)	32 (71.1)	36 (80)
Psychiatry	8	8 (100)	8 (100)	8 (100)	8 (100)
Pathology	9	8 (88.9)	8 (88.9)	8 (88.9)	8 (88.9)
Radiology	6	4 (66.7)	4 (66.7)	4 (66.7)	3 (50)
Infectious Diseases	10	9 (90)	9 (90)	10 (100)	9 (90)
Neurological Disorders	8	7 (87.5)	7 (87.5)	7 (87.5)	7 (87.5)
Pharmacology	6	6 (100)	6 (100)	6 (100)	6 (100)
Epidemiology	6	4 (66.7)	5 (83.3)	5 (83.3)	5 (83.3)
Otolaryngology (ENT)/Head and Neck Surgery	6	4 (66.7)	4 (66.7)	4 (66.7)	4 (66.7)
Ophthalmology	6	5 (83.3)	6 (100)	6 (100)	6 (100)
Urology	6	6 (100)	6 (100)	6 (100)	6 (100)
Orthopedics	7	7 (100)	5 (71.4)	5 (71.4)	6 (85.7)
Dermatology	6	4 (66.7)	5 (83.3)	4 (66.7)	4 (66.7)
Medical Ethics	2	2 (100)	2 (100)	2 (100)	2 (100)
All Questions	198	161 (81.3)	167 (84.3)	162 (81.8)	166 (83.8)

Table. 2 Performance of ChatGPT in the 2023 Iranian residency multiple-choice question tests based on seventeen question topics and four languages.

6. Environmental Impact of LLM

The environmental impact caused by large language models is the growing usage of LLMs raises concerns regarding their substantial carbon emissions and energy consumption. during the training process. Training LLMs requires massive amounts of computing power, which in turn leads to increased energy usage and carbon emissions. From [17] the authors estimated that training a single LLM produced 626,000 pounds of CO2eq, equivalent to the lifetime emissions of 5 cars.

Model name	Number of parameters	Power consumption	CO_2eq emissions
GPT-3	175B	1,287 MWh	*502 tons*
Gopher	280B	*1,066 MWh*	*352 tons*
OPT	175B	*324 MWh*	70 tons
BLOOM	176B	433 MWh	25 tons

Table. 3 Environmental Impact of LLM

7. Conclusion

This particular survey provides a thoroughly detailed and comprehensive exploration of the remarkable advancements and significant progressions that have occurred in recent times within the expansive and rapidly evolving field of large language models, often referred to as LLMs, while simultaneously delving into an investigation of the various findings and methodologies that are particularly relevant to our understanding and effective utilization of these increasingly complex models, in addition to examining the wide-ranging and extensive ramifications that these models may potentially bring about in various sectors and applications.

We conclude our detailed discourse with an empirical assessment that rigorously evaluates the performance and impact of large language models, and in addition, we elucidate the dominant trends and prevailing themes found within the evaluation literature concerning these models, alongside the tangible and measurable outcomes that have been achieved by models such as ChatGPT, particularly within the specific context of the Iranian residency multiple-choice examinations, which encompassed a total of seventeen topical inquiries that were presented across four distinct languages. This concerted and collaborative endeavor has contributed to a more lucid, nuanced, and sophisticated evaluation of large language models in relation to the intricate challenges and complexities faced in real-world contexts and scenarios. Finally, we also place significant emphasis on a critical consideration for enterprises and organizations to adopt more pragmatic, sustainable, and energy-efficient methodologies regarding the training processes of large language models, as we have highlighted the reality that numerous existing LLMs necessitate a considerable amount of energy consumption and consequently result in significant CO2 emissions, thereby indicating a highly promising area for future research endeavors and developmental initiatives that seek to address these pressing environmental concerns.

References

1. Zhao, Wayne Xin, et al. “A Survey of Large Language Models”. arXiv, 28 Apr. 2023.
2. Huang, Jie, and Kevin Chen-Chuan Chang. “Towards Reasoning in Large Language Models: A Survey”. arXiv, 20 Dec. 2022.
3. Vaswani, Ashish, et al.”Attention Is All You Need”. arXiv, 5 Dec. 2017. arXiv.org,
4. Weizenbaum, Joseph. “ELIZA—a Computer Program for the Study of Natural Language Communication between Man and Machine”. Communications of the ACM, vol. 9, no. 1, Jan. 1966, pp. 36–45.

5. Devlin, Jacob, et al. “BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding”.
6. Raford, Narasimha, et al. “Improving Language Understanding by Generative Pre-Training”, 2018
7. Touvron, Hugo, et al. “LLaMA: Open and Efficient Foundation Language Models”. arXiv, 27 Feb. 2023.
8. Wu, Shijie, et al. “BloombergGPT: A Large Language Model for Finance”. arXiv, 30 Mar. 2023.
9. Eloundou, Tyna, et al.”GPTs Are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models”. arXiv, 23 Mar. 2023.
10. Ogundare, Oluwatosin, et al. “Industrial Engineering with Large Language Models: A Case Study of ChatGPT’s Performance on Oil & Gas Problems”. arXiv, 27 Apr. 2023.
11. Pranjali Bahalkar, et al. “Predicting Students Growth in Academic career using Artificial Intelligence and Machine Learning Techniques”, ISSN 1660-6795, Nanotechnology Perceptions 20 No.6 (2024) 1791-1801, https://doi.org/10.62441/nano-ntp.v20i6.129
12. Wang, Haifeng, et al. “Pre-Trained Language Models and Their Applications. Engineering”, Sept. 2022.
13. Bahdanau, Dzmitry, et al. “Neural Machine Translation by Jointly Learning to Align and Translate”. arXiv, 19 May 2016.
14. Shoeybi, Mohammad, et al. “Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism”. arXiv, 13 Mar. 2020.
15. Brown, Tom B., et al. “Language Models Are Few-Shot Learners”. arXiv, 22 July 2020.
16. Fan, Lizhou, et al. “A Bibliometric Review of Large Language Models Research from 2017 to 2023”. arXiv, 3 Apr. 2023.
17. Workshop, BigScience, et al. “BLOOM: A 176B-Parameter Open-Access Multilingual Language Model”. arXiv, 13 Mar. 2023.
18. Elnaggar, Ahmed, et al. “ProtTrans: Toward Understanding the Language of Life Through Self-Supervised Learning”. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 44, no. 10, Oct. 2022, pp. 7112–27.
19. Applications of Artificial Intelligence in the Oil and Gas Industry | Frontiers Research Topic
20. de Wynter, Adrian, et al. “An Evaluation on Large Language Model Outputs: Discourse and Memorization”. arXiv, 17 Apr. 2023. arXiv.org,