IJARETY

International Journal of Advanced Research in Education and TechnologY (IJARETY)

Volume 12, Issue 4, July-August 2025

Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204074

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

Suchithra M, Shivaram P, Shrinivas S Pai

Department of MCA, CMR Institute of Technology, Bengaluru, India

ABSTRACT: Large Language Models (LLMs) recently demonstrated extraordinary capability, including natural language processing (NLP), language translation, text generation, question answering, etc. Moreover, LLMs are a new and essential part of computerized language processing, having the ability to understand complex verbal patterns and generate coherent and appropriate replies for the situation. Though this success of LLMs has prompted a substantial increase in research contributions, rapid growth has made it difficult to understand the overall impact of these improvements. Since a lot of new research on LLMs is coming out quickly, it is getting tough to get an overview of all of them in a short note. This article thoroughly overviews LLMs, including their history, architectures, transformers, resources, training methods, applications, impacts, challenges, etc. This paper begins by discussing the fundamental concepts of LLMs with its traditional pipeline of the LLM training phase. It then provides an overview of the existing works, the history of LLMs, their evolution over time, the architecture of transformers in LLMs, the different resources of LLMs, and the different training methods that have been used to train them. It also demonstrated the datasets utilized in the studies. The paper discusses the wide range of applications of LLMs, including biomedical and healthcare, education, social, business, and agriculture.

KEYWORDS: Large Language Models, Natural Language Processing, Evolution, Transformer, Pretrained models, Taxonomy, Application

I. INTRODUCTION

A long-standing scientific challenge and aim has been to achieve human-like reading, writing, and communication skills in machines [4]. However, advances in deep learning approaches, the availability of immense computer resources, and the availability of vast quantities of training data all contributed to the emergence of large language models (LLMs). It is a category of language models that utilizes neural networks containing billions of parameters, trained on enormous quantities of unlabeled text data using a selfsupervised learning approach [5]. Besides, they have proved their ability in various language-related tasks, including text synthesis, translation, summarization, question-answering, and sentiment analysis, by leveraging deep learning techniques and large datasets. Moreover, the results of fine-tuning these models on specific downstream tasks have been quite promising, with state-of-the-art performance in several benchmarks [7]. The basic LLM pipeline (Figure 1) involves data collection, preprocessing, parameter initialization, loss calculation, optimization, and iterative training, producing capabilities such as translation, summarization, and sentiment analysis. Given LLMs' rapid evolution and broad application, a comprehensive review is needed to examine their development, architectures, resources, applications. LLM research is becoming increasingly important, and prior research has shown the potential and superiority of LLMs in NLP tasks. Nevertheless, few studies have thoroughly reviewed their work's most recent LLM developments, possibilities, and limitations. Despite the increasing number of studies on LLMs, there remains a scarcity of research focusing on their technical complexities, the LLMs taxonomy, architectures, API applications, domain-specific applications, effective utilization, impact on society, and so on. Furthermore, the majority of the LLM review papers are not peer-reviewed articles

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204074

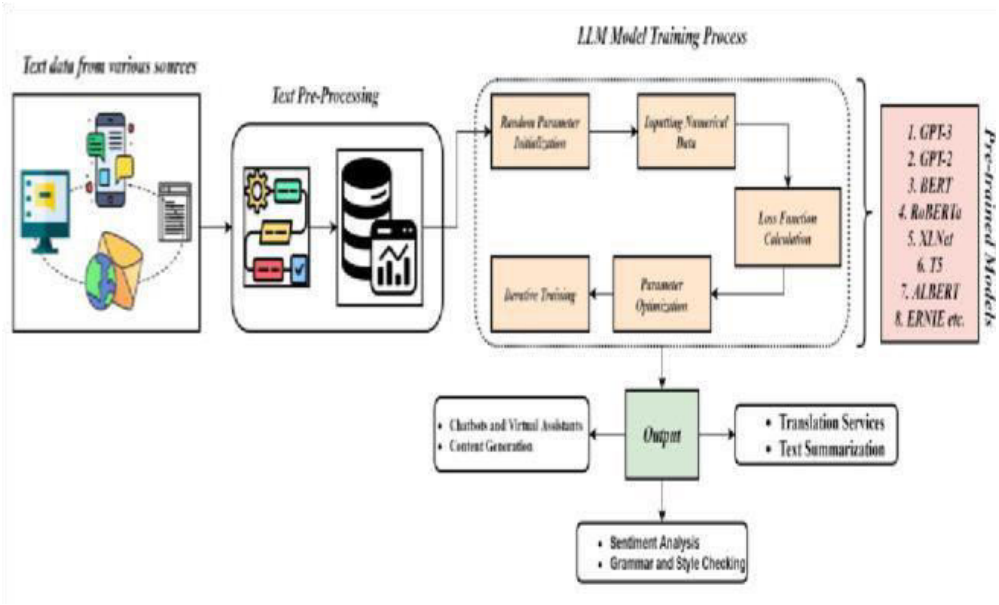

Figure 1: Pipeline of the LLM training phase

II. LITERATURE REVIEW

The rapid growth of LLMs is a major development in AI, with numerous studies exploring their capabilities and applications. Research highlights their ability to revolutionize tasks from text generation and comprehension to reasoning.

Huang et al. [17] reviewed reasoning in LLMs, covering techniques to enhance reasoning abilities, evaluation methods, prior insights, and future directions. [18] conducted a bibliometric review of over 5,000 publications (2017–2023), tracking research trends across disciplines like medicine, engineering, and social sciences. Chang et al. [19] examined evaluation methodologies for LLMs, focusing on what, where, and how to evaluate, covering NLP, reasoning, medical, ethical, and educational applications.

Table 1 compares these reviews. While prior works address reasoning, evolution, bibliometrics, or evaluation separately, our review covers all these areas—history, architectures, resources, applications, challenges—offering a comprehensive view. Table 1 illustrates the comparison between different review papers based on some critical factors such as LLM , LLM API, LLM Dataset, Domain Specific LLM, Taxonomy, LLM Architecture, LLM Configurations

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

Study	LLM Papers	API	Data	Domain	Tax.	Arch.	Config.	ML Diff	Scope	Key Findings	Method
Huang et al. [2022] [17]	✓	X	X	X	X	X	X	X	Reasoning	Analysis of reasoning abilities & improvement meth	Review
Zhao et al. [2023] [3]	✓	X	✓	X	✓	X	✓	X	Evolution	LLM history, capabilities, resources, AI/NLP impac	Survey
Fan et al. [2023] [18]	✓	X	X	X	X	X	X	X	Bibliometric	Trends (2017–2023), advancements, domain impac	Bib. Analysis
Chang et al. [2023] [19]	✓	X	✓	X	✓	X	X	X	Evaluation	Methods, risks, future challenges	Survey
Our Study	✓	✓	✓	✓	✓	✓	✓	✓	Comprehensive	resources, architectures, domains, ML diff., appl	Broad Review

International Journal of Advanced Research in Education and TechnologY(IJARETY)

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204074

Differentiation. Huang et al. [17] lack information on LLM API, LLM Dataset, Domain-Specific LLM, Taxonomy, LLM Architecture, and LLM Configurations. In contrast, Zhao et al. [3] lack information on LLM API, DomainSpecific LLM, Taxonomy, LLM Architecture, and LLM Configurations. Moreover, Fan et al. [18] and Chang et al. [19] lack information on LLM API, Domain-Specific LLM, Taxonomy, LLM Architecture, and LLM Configurations.

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

III. METHODOLOGY

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

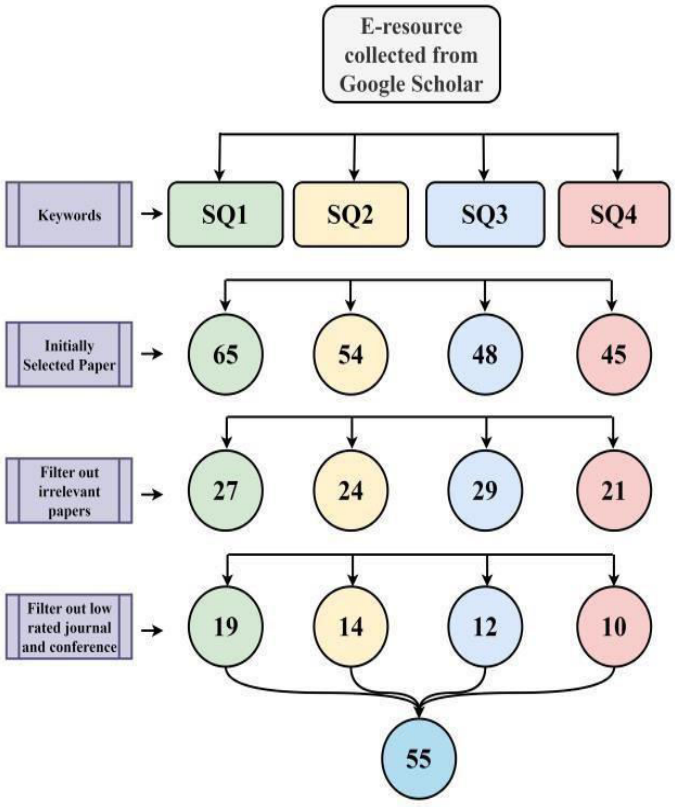

Figure 5: Flow diagram of systematic review

A. Large Language Models

Large language models (LLMs) refer to a specific type of AI algorithm that holds the capability to execute a diverse range of NLP operations. The most common tasks entail text generation, text analysis, translation, sentiment analysis,

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204074

question answering, and other related functions. GPT-3, GPT-4, PaLM, and LaMDA are extensively used transformerbased LLM models trained on a large amount of textual data. In terms of architectural properties, these models show variations in size and depth. For example, GPT-3 generates parameters of 175 billion, distributed across 96 levels, while PaLM has an even larger parameter number of 540 billion, organized across 106 layers.

B. Architectural Overview of Large Language Models

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

GPT-1 uses data input embedding and positional encoding repeatedly from multiple transformer layers. B. Comparison Between Configurations of LLMs

This carries out a detailed overview of multiple Large Language Models (LLMs), and it specifies their configuration parameters along with optimization parameters.

These LLMs are central figures in promoting natural language understanding and generation processes and hence constitute a central area of interest within artificial intelligence and natural language processing. This comparison and contrast of these LLMs are founded on vital parameters such as the size of the model, learning rate, category, activation function used, batch size, bias, layer numbers, optimizer used, attention head numbers, hidden state dimension, dropout probability, and maximum training context length. GPT-4 is the standout model on show with a whopping 1.8 trillion parameters.GPT1, despite being lesser with 125 million parameters, demonstrates the significant development of LLM over the years. An increased number of parameters in LLM enhances the model's ability to comprehend intricate patterns and produce text that is more contextually appropriate and reminiscent of human language. GPT3's selection of a modest learning rate of 6 is notable, which highlights the significance of cautious hyperparameter selection. Models are categorized as Causal decoder (CD), Autoregressive (AR), Encoder-decoder (ED), and Prefix decoder (PD) to illustrate architectural diversity. Activation functions vary, influencing the models' expressive strength from GeLU in GPT-3 to SwiGLU in LLaMA and LLaMA-2. All versions of GPT employ the GeLU as its activation function as it mitigates the vanishing gradient problem and facilitates the generation of smoother gradients throughout the training process. The utilization of SwiGLU as the activation function is observed in models such as PaLM and LLaMA versions 1 and 2, as it has gating Table 5: Architectural overview of different LLMs 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 [17, 88]: R(1)

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

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

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

C. Domain Specific Application

Since there are several pre-trained models in LLMs, all of them are utilized by training or fine-tuned to perform well-defined tasks maintained by their requirements in different fields. Numerous research studies have consistently contributed by using LLMs model in diverse domains such as healthcare, finance, education, forecasting, and natural language processing. The extensive experiments of different LLM models contribute to revolutionizing the use of AI across these diverse domains. This section demonstrates the potential contribution of LLMs application in different domains. Bio-Medical and Healthcare: It is mentioned earlier that GPT has a few versions, i.e., from GPT1 up to GPT4. GPT3 is very useful when it comes to the healthcare sector as it can aid customer service with minimal effort. It can receive all necessary information via a conversation and not via intake form and numerous systems could be Machine learning-based study comparison in LLMs to help many patients simultaneously [9]. Other than that, clinics and hospitals are where one goes to heal illness, but then it is a reality that many contagious viruses are introduced into

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204074

these facilities. Patients and medical practitioners can be protected from infection better by introducing a robot receptionist instead of a human. This is especially true when it comes to the COVID-19 pandemic [10]. Since clinics and hospitals tend to receive a lot of patients on a daily basis, an optimum and light system may come up with multiple questions of single patients to generate satisfactory output. Accordingly, GPT models may contribute to cost minimization as well when it comes to the medical sector.

Social Media: The LLMs have revolutionized several aspects of the social media industry regarding content production, moderation, sentiment analysis, etc. There are some crucial aspects of the LLMs in the social media sector in terms of writing content, generating images, classifying and generating text, and even full blogs and articles for social media. Also, these models can perform named entity recognition (NER) and text classification [18, 19]. When the GPT, XLNet, BERT, etc., model aids the writer and content producers in generating a consistent flow of excellent material. It also provides content suggestions, and to create a safer online environment, these models are hired to assist in discovering and filtering out different dangerous and improper content. In their study, Abramski et al. [22] utilized network science and the principles of [(3)] cognitive psychology to evaluate biases present in LLMs.

Agriculture: In agriculture, variations of GPT models, including GPT3, BERT, and XLNet models, play a significant role [10, 11]. They are able to analyze large data hubs of soil, crop, and weather data along with satellite imagery. They can provide recommendations on plating times, irrigation, fertilizer application, and optimizing fields and resources. Farmers can obtain current updates and market requirements, predict crop prices, anticipate natural disasters, and document farmers' and crop details. Manual agriculture management can be time-consuming and laborious, but these models can handle all the issues.

Business: In business, LLM helps companies improve their decision-making processes, product manufacturing processes, operations, and customer interactions. Communicating with customers and providing 24/7 customer service by answering their queries, assisting them in their work, and providing advanced advice related to areas of interest to customers is crucial for business progress. Moreover, it is also important to analyze customer sentiment, market trends, risk factors, and competitive intelligence [6]. In this case, LLMs help to fulfill all their requirements within a short period. The LLM models, like GPT, XLNet, BERT, etc., play a vital role in creating customer documents and product details and efficiently maintaining the entire business by saving time and reducing laborious tasks. Frederico et al. [9] presents an initial investigation into the potential applications and effects of ChatGPT in the domain of supply chain management.

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

A. Open Issues
In this section, we delve into the critical open issues surrounding LLMs. These concerns are at the vanguard of artificial intelligence research and development. They emphasize the need for ongoing research and innovation to resolve issues that have emerged alongside the rapid development of LLMs. Our discussion will cast light on the significance of these unresolved issues, highlighting their impact on various applications and the AI landscape as a whole. • Issue 1: Ethical and Responsible AI The question regarding how to ensure the ethical use of large language models remains unresolved. Filtering, moderation, and accountability concerns regarding AI-generated content remain troublesome. Misinformation, hate speech, and biased content generated by LLMs necessitate continuous research and development [12].

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 [21]. 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 [12]. It is essential to develop more energy-efficient training methods, model architectures, and hardware solutions to reduce the carbon footprint of LLMs. • Issue 4: Security and Adversarial Attacks LLMs are vulnerable to adversarial assaults, where slight input modifications can lead to unexpected and potentially harmful outputs [3]. Improving model robustness and security against such assaults is a crucial area of study, particularly for cybersecurity and content moderation applications.

DOI:10.15680/IJARETY.2025.1204074

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 [14].

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. Their extraordinary ability to generate text that resembles that of a human has garnered significant attention and applications in numerous fields. However, this meteoric rise in prominence has also revealed many challenges and concerns that must be addressed to realize the potentiality of these models fully. In this discussion, we will examine ten of the most significant challenges pertaining to LLMs

.

Challenge 1: Data Complexity and Scale

In the era of LLMs, the size and complexity of the datasets on which they are trained is one of the most significant challenges. These models are typically trained on enormous corpora of Internet-sourced text data. These datasets are so extensive that it is nearly impossible to comprehend or investigate the totality of their information. This raises concerns regarding the quality and biases of the training data and the potential for the unintentional dissemination of detrimental or inaccurate information. • Challenge 2: Tokenization Sensitivity

For analysis, LLMs rely significantly on tokenization, dividing text into smaller units (tokens) [17]. Tokenization is essential for language processing and comprehension but can also present challenges. For instance, the meaning of a sentence can alter significantly based on the choice of tokens or the ordering of words. This sensitivity to input phrasing can lead to unintended outcomes when generating text, such as adversarial assaults and output variations based on minute input changes.

Challenge 3: Computational Resource Demands The training of LLMs is a computationally intensive procedure that requires substantial hardware and energy resources. It is necessary to have access to supercomputing clusters or specialized hardware in order to train large models, and the environmental impact of such resource-intensive training has raised concerns. Significant energy consumption is associated with training LLMs at scale, contributing to the AI industry's overall carbon footprint.

Challenge 4: Fine-Tuning Complexity

While pre-training gives LLMs a broad comprehension of language, fine-tuning is required to adapt these models to specific tasks. Fine-tuning entails training the model on a smaller dataset, frequently requiring human annotators to label examples. As it involves the construction of taskspecific datasets and extensive human intervention, this process can be both timeconsuming and costly.

IV. CONCLUSION

LLMs have gone through a fabulous metamorphosis and rapid expansion in recent years, therefore acquiring huge NLP capabilities for various applications across different areas. Deep neural networks coupled with the best transformer architecture have forever changed the manner of understanding and generation of a machine language by LLM. Once the exhaustive review of this research has permeated through LLMs, from the historical evolution to their architecture, training, and vast advancement resources, it stresses the importance of constant efforts for the refinement of LLMs in terms of efficacy and trustworthiness, along with the need for the ethical development and deployment. With LLMs being one of the landmark advances in AI and NLP, they have the potential to change innumerable domains and are set to solve very complex problems out there.

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, 2002.
3. W. X. Zhao *et al.*, "A survey of large language models," *arXiv preprint arXiv:2303.18223*, 2023.
4. A. M. Turing, "Computing machinery and intelligence," *Mind*, vol. 59, no. 236, p. 433, 2007.
5. Y. Shen *et al.*, "ChatGPT and other large language models are double-edged swords," 2023.
6. S. Hochreiter and J. Schmidhuber, "Long shortterm memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

| ISSN: 2394-2975 | www.ijarety.in| | Impact Factor: 8.152| A Bi-Monthly, Double-Blind Peer Reviewed & Refereed Journal |

|| Volume 12, Issue 4, July - August 2025 ||

DOI:10.15680/IJARETY.2025.1204074

7. M. Du, F. He, N. Zou, D. Tao, and X. Hu, "Shortcut learning of large language models in natural language understanding: A survey," *arXiv preprint arXiv:2208.11857*, 2022.
8. B. Ramabhadran, S. Khudanpur, and E.
9. 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 Workshop*, 2012.
10. T. Mikolov, M. Karafiát, L. Burget, J.
11. Cernocky, and S. Khudanpur, "Recurrent neural network based language model," in *Proc. Interspeech*, 2010, vol. 2, pp. 1045–1048.
12. A. Vaswani *et al.*, "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.
13. J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.
14. Y. Khare *et al.*, "MMBERT: Multimodal BERT pretraining for improved medical VQA," in *Proc. IEEE 18th Int. Symp. Biomedical Imaging (ISBI)*, 2021, pp. 1033–1036. [13] R. Liu *et al.*, "Mitigating political bias in language models through reinforced calibration," in *Proc. AAAI Conf. Artificial Intelligence*, vol. 35, pp.
15. 14857–14866, 2021.
16. K. Sanderson, "GPT-4 is here: What scientists think," *Nature*, vol. 615, no. 7954, p. 773, 2023.
17. S. Pichai, "An important next step on our AI journey," Google Blog, Feb. 2023
18. R. Taori *et al.*, "Alpaca: A strong, replicable instruction-following model," Stanford Center for Research on Foundation Models, Mar. 2023.
19. J. Huang and K. C.-C. Chang, "Towards reasoning in large language models: A survey," *arXiv preprint arXiv:2212.10403*, 2022.
20. L. Fan *et al.*, "A bibliometric review of large language models research from 2017 to 2023," *arXiv preprint arXiv:2304.02020*, 2023.
21. Y. Chang *et al.*, "A survey on evaluation of large language models," *arXiv preprint arXiv:2307.03109*, 2023.
22. E. Kasneci *et al.*, "ChatGPT for good? On opportunities and challenges of large language models for education," *Learning and Individual Differences*, vol. 103, p. 102274, 2023.
23. N. Nair, "Large language models," presented at the Knowledge Sharing Session, Center of Excellence (CoE), Department of Computer Science and Engineering, CMR Institute of Technology, Bengaluru, India, May 2024.
24. A. M., Aditi, L. Rajeev, and K. Manne, "Tech Bytes – Even Semester 2024,CMR Institute of Technology, Bengaluru, India, Jan. 2025.

International Journal of Advanced Research in Education and Technology

ISSN: 2394-2975 Impact Factor: 8.152

www.ijarety.in editor.ijarety@gmail.com