Volume 1, Issue 9 (2026)
PP 01-06

VIVA-TECH IJRI
ANNUAL RESEARCH JOURNAL
VIVA-TECH INTERNATIONAL JOURNAL FOR RESEARCH AND INNOVATION
ANNUAL RESEARCH JOURNAL
ISSN(ONLINE): 2581-7280

Evolution and Impact of Large Language Models: A Systematic Review

Dr. Pradnya Atmaram Mhatre
[1](*MCA Department, VIVA Institute of Technology, India)*

Abstract : *Large Language Models (LLMs) have revolutionized artificial intelligence (AI) and natural language processing (NLP) by enabling machines to comprehend, produce, and reason with human language in ways that were previously unthinkable. These models are mostly based on the Transformer architecture, which was first presented in 2017 and substitutes self-attention mechanisms that can capture long-range dependencies and contextual linkages in text for recurrent and convolutional techniques. LLMs can now execute a wide range of tasks, including text production, summarization, translation, question answering, and reasoning, frequently with no task-specific training, which has increased over time from millions to billions or even trillions of parameters. The evolution of LLMs from simple Transformer models to contemporary variations is systematically reviewed in this study. I also go over the inherent difficulties of LLMs, such as model bias, hallucinations, high processing demands, and security and ethics issues. Lastly, this analysis highlights the significance of rooting models in factual knowledge while exploring future research possibilities targeted at creating sustainable, effective, and reliable LLMs.*

Keywords - *factual knowledge, hallucinations, Large Language Models, model bias, Natural Language Processing, Transformer architecture*

I. INTRODUCTION

Recent developments in Large Language Models (LLMs), which have greatly improved machines' capacity to receive, comprehend, and produce human language, have played a major role in the revolutionary change in artificial intelligence (AI). These models overcome the drawbacks of previous recurrent and convolutional approaches by using self-attention processes to capture long-range dependencies and contextual relationships in text. They are mostly based on the Transformer architecture, which was first presented by Vaswani et al. in 2017. From early Transformer-based models to modern encoder-only, decoder-only, encoder-decoder, and multimodal architectures, LLMs have evolved to perform a variety of complex tasks, such as text generation, summarization, translation, question answering, and reasoning, frequently with little task-specific training.

Significant gains in task adaptation and generalization have been made possible by their scaling from millions to billions and even trillions of parameters. Beyond their technological breakthroughs, LLMs have had a significant impact on society and industry, driving applications in the fields of healthcare, banking, education, science, and the arts. But these models also come with drawbacks, such as inherited biases, hallucinations, high processing and energy costs, and deployment-related ethical issues. The purpose of this systematic review is to trace the development of LLMs, analyze their designs and applications, highlight current trends, pinpoint important constraints, and explore potential future paths for creating reliable, sustainable, and ethically sound model. [1][2][3][4]

II. BACKGROUND

A family of artificial intelligence systems known as Large Language Models (LLMs) was created using developments in deep learning and natural language processing. In order to acquire language patterns, such as grammar, context, and meaning, they are usually constructed using neural network topologies like transformers and trained on enormous volumes of text data. LLMs, which emerged from earlier statistical and rule-based language models, gained popularity in the late 2010s as models were able to comprehend and produce human-like

writing for a variety of tasks thanks to advancements in algorithms, vast datasets, and processing capacity[1][2][3][4]

2.1 Evolution of Large Language Models] (LLMs)

2.1.1 Rule-Based Systems (1950s–1970s):

Linguistic and grammatical rules that were manually developed served as the foundation for early natural language processing systems. These rule-based systems have limitations in terms of stability, scalability, and adaptability, even though they could successfully manage particular tasks. They consequently found it difficult to handle complicated or large scale language data.

2.1.2 Statistical Language Models (1980s–1990s):

Data-driven methods like n-gram models, which employed probability to predict word sequences from massive text corpora, were made possible by advances in computing. These models improved flexibility but struggled with long-range context and deep semantic understanding.

2.1.3 Machine Learning and Early Neural Networks (2000s):

Natural language processing saw a significant change in the 2000s because of machine learning approaches. Instead of depending on manually created rules, computers were able to automatically learn linguistic patterns from data by using probabilistic models and early neural networks. These methods made systems more adaptable and versatile while also enhancing language comprehension. Nevertheless, processing lengthy sequences and collecting complex context remained a challenge for early neural networks. . [5][6]

2.1.4 Word Embeddings (Early 2010s):

Words were represented as dense vectors by methods like Word2Vec and GloVe, which greatly improved language representation by capturing syntactic and semantic links.

2.1.5 Sequence Models (RNNs and LSTMs) (Mid 2010s):

Although they were computationally demanding, recurrent neural networks and extended short-term memory networks improved sequence modeling by preserving contextual information over larger text spans.

2.1.6 Transformer Architecture (2017):

The Transformer, which was first presented by Vaswani et al. in 2017, used parallelizable structures, positional encoding, and self-attention to overcome the limits of RNNs. Originally intended for machine translation, it served as the basis for later LLMs. Transformers allowed for more efficient modeling of long-range relationships and concurrent text processing by substituting attention methods for repetition.

2.1.7 Large-Scale Pretrained Models:

Large-scale pretrained models are a significant development in natural language processi ng and artificial intelligence. The transformer architecture, which uses attention mechanisms to comprehend long-range relationships in text, is used in their construction. These models may learn linguistic patterns without labeled data since they are trained via self-supervised learning on large and varied datasets. Large-scale training is made possible by high-performance computing resources like GPUs and TPUs. The models can be adjusted or encouraged for particular tasks after pretraining. Thus, text production, translation, summarization, reasoning, and discussion are all possible with contemporary LLMs. LLMs are now strong, all-purpose AI systems thanks to this advancement [7] [8] [9]

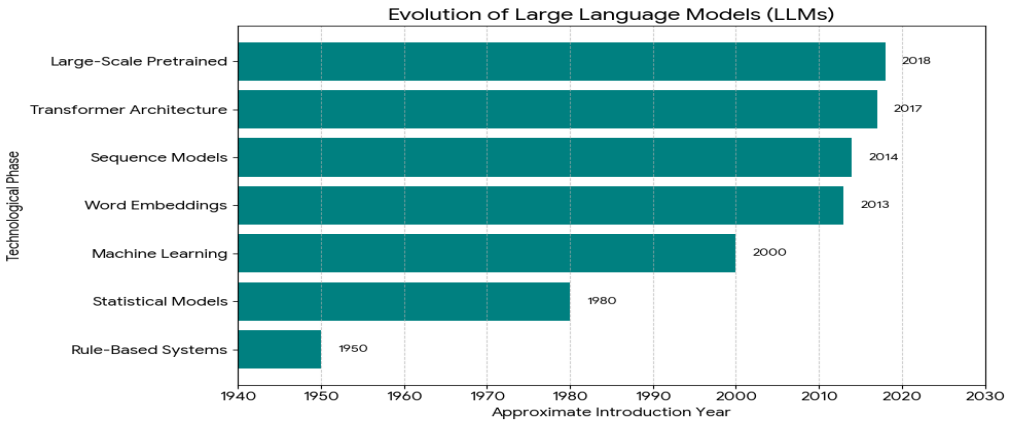

Fig.1: The bar chart to illustrate the historical progression leading to modern LLMs

III. ARCHITECTURE OF LLM

The transformer, which employs attention processes to record word relationships over a whole text sequence, is the main component of the architecture of Large Language Models (LLMs). Transformers handle long-range

dependencies efficiently because they analyze text in parallel, unlike previous sequential models. LLMs learn linguistic patterns without tagged data by employing self-supervised learning on large datasets. They can be used with instructions or adjusted for certain activities after pretraining. LLMs are extremely powerful and adaptable for tasks like text production, translation, summarization, and dialogue because high-performance computing devices like GPUs and TPUs can train models with billions of parameters. [10] [11]

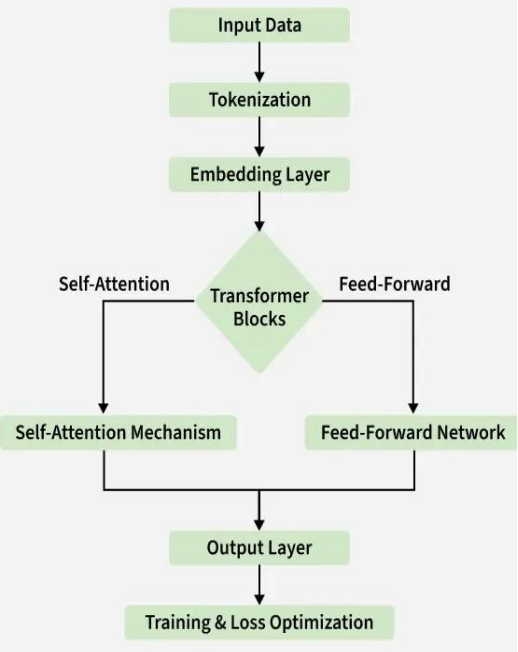

Fig.2: LLM Architecture, image is taken from [10]

3.1 Components in Large Language Models:
The components of an LLM can be visualized as a pipeline that transforms raw text into meaningful predictions or outputs. The main components are:

3.1.1 Tokenization

The input text in LLMs is first divided into tokens, which can be words, symbols, or letters. Tokenization is essential preprocessing technique that breaks up text into meaningful parts. WordPiece and Byte Pair Encoding (BPE) are popular tokenization techniques.

3.1.2 Embedding Layer

In order to express semantic meaning in a high-dimensional space, tokens are transformed into numerical vectors known as embeddings. Since transformers cannot naturally process sequences in order, positional embeddings are introduced to identify the order of tokens. The model can understand word relationships with the help of embeddings.

3.1.3 Transformer Layers

- Self-Attention Mechanism: A model can determine the relationships between each word in a sequence due to the self-attention mechanism. By focusing on the most relevant terms in the sequence, it generates context-aware word representations [6] [12]. The model calculates three vectors for each word:
  - Query (Q): represents the word under emphasis.
  - Key (K): represents every word in the order so that relevance may be compared.
  - Value (V): includes the data for the terms that will be combined.

  Equation for Self-Attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V$$

  The attention mechanism uses the Query and Key vectors to create a score for each pair of words, normalizes these scores using the softmax function, and then computes a weighted total of the Value vectors. This results in each word's final context-aware representation.

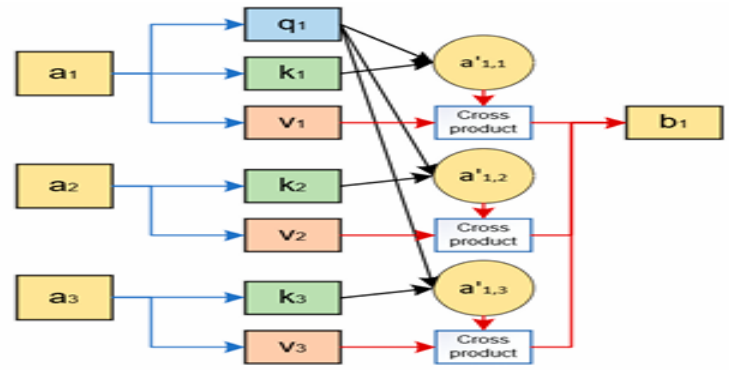

Fig.3: Attention mechanism diagram, image is taken from [6]

- Feedforward Networks**:** Feedforward networks support attention mechanisms to improve language comprehension and are an essential component of transformer-based models. After the self-attention layer captures relationships between words, the feedforward network processes this output independently for each position in the sequence. The model can learn complex patterns and transformations since it is made up of completely connected layers with non-linear activation functions.
- Layer Normalization & Residual Connections: In transformer architectures, residual connections and layer normalization are essential elements that support reliable and effective training. By standardizing the inputs inside each layer, layer normalization lowers internal covariate shift and improves training stability. By adding a layer's output back to its input, residual connections enhance gradient flow and avoid problems like vanishing gradients. When combined, they improve overall model performance and allow deeper networks to train efficiently.

3.1.4 Output Layer:

The last part of a language model, the output layer, is in charge of producing predictions. The output layer of decoder-based models (like GPT) uses the prior context to predict the subsequent token in a sequence. It predicts the masked tokens that were concealed during training in encoder-based models (like BERT).These values are then transformed into probabilities, which show the likelihood of each potential token, using a softmax function. Usually, the model's output is the token with the highest probability.[9]

3.1.5 Training Components:

Building successful Large Language Models requires the use of Training Components. During training, the loss function—such as cross-entropy loss—measures prediction mistakes and directs the model to increase accuracy. To reduce this loss, the optimizer, such as Adam, effectively modifies model parameters. Large-scale datasets are also crucial because they enable LLMs to learn contextual relationships, semantics, and linguistic patterns across a variety of text sources.

3. 2 Role of High-Performance Computing:

It takes a lot of processing power to train LLMs. Massive volumes of data may be processed in parallel because of high-performance computing (HPC), which includes GPUs (Graphics Processing Units) and TPUs (Tensor Processing Units). It would be impossible to train modern LLMs on conventional hardware because they frequently have billions or even trillions of parameters. Large batch sizes, quicker training, and the intricate matrix operations required for attention mechanisms are all made possible by HPC. The development of LLMs is also possible at scale by using sophisticated optimization methods and distributed computing across several nodes, which shorten training times and boost model effectiveness.

IV. IMPACT OF LARGE LANGUAGE MODELS

TABLE1: Positive and Negative Impact of LLM

ASPECT	POSITIVE IMPACT	NEGATIVE IMPACT
Productivity	Large Language Models (LLMs) can save a lot of time and effort by handling tasks like data analysis, coding, and content creation. They increase productivity and free up humans to concentrate on difficult or creative activities by enabling experts to	Large Language Models (LLMs) have the potential to displace workers in fields like data input, content creation, and customer service by replacing repetitive tasks. Workers may be forced to take on new jobs requiring advanced skills as a result of

	provide reports, summaries, and insights far more quickly than with laborious tasks.	automation, posing social and economic issues.
Knowledge Access	By making information easily accessible and comprehensible, Large Language Models (LLMs) improve knowledge access. They help individuals learn effectively and make wise judgments by breaking down difficult subjects into concise explanations and summaries.	Large Language Models (LLMs) improve access to knowledge but can sometimes produce false or misleading information, known as hallucinations [1]. Relying on these outputs without verification can lead to misunderstandings and poor decision-making.
Education	By providing individualized tutoring and explanations to meet each student's unique learning needs, Large Language Models (LLMs) assist education. They assist students understand complex ideas, offer practice problems, and make learning more approachable and interesting.	Students' critical thinking and problem-solving skills may be weakened if LLMs are used excessively in the classroom. Over-reliance on AI-generated responses may hinder the development of critical thinking abilities and independent analysis.
Privacy & Security	By examining massive information to find trends, spot fraud, and identify possible dangers, LLMs can improve security and privacy. This enables businesses to better safeguard sensitive data and make data-driven decisions.	Because LLMs may reveal private or sensitive information, they present security and privacy hazards. Confidential information can be misused and data breaches can result from improper information handling.
Ethics & Society	By improving research, journalism, and well-informed decision-making, LLMs promote ethics and society. They facilitate rapid insight gathering, information summarization, and the provision of trustworthy data for improved comprehension and analysis.	LLMs raise privacy and security concerns by enabling the spread of misinformation and deepfakes. They also create challenges in accountability, making it difficult to track or verify the source of false or harmful content.
Bias & Fairness	By pointing out prejudiced language and offering more inclusive substitutes, LLMs can promote bias and fairness. This lessens unintended discrimination in content and encourages fair communication.	LLMs can unintentionally mirror and amplify societal biases present in their training data. This may lead to outputs that reinforce stereotypes or produce unfair and biased results.
Environment	LLMs can aid in workflow optimization, increasing productivity and cutting down on wasteful resource usage. In some businesses, this can result in less waste and a lesser environmental impact.	LLMs may inadvertently reflect and magnify societal biases found in their training data. This could result in unfair and biased outcomes or outputs that reinforce biases.

V. FUTURE DIRECTIONS OF LARGE LANGUAGE MODELS

5.1 Better Accuracy and Reasoning:

By significantly reducing errors and hallucinations, future large language models are going to generate outputs that are more accurate and dependable. Better fact-checking techniques and increased reasoning skills will be incorporated, allowing them to solve logical and mathematical problems more successfully and facilitate better decision-making in practical situations. [13]

5.2 Multimodal Capabilities:

Future LLMs will be able to comprehend and produce a variety of data forms, including graphs, text, photos, audio, and video. This will make it possible for humans and AI to communicate more naturally and interactively. For example, voice assistants will be able to comprehend visual content, understand visuals, and react intelligently in a variety of media.

5.3 Improved Ethical and Bias Control:

To lessen bias and guarantee fairness in their responses, future LLMs will include more robust ethical safeguards. These models will encourage responsible AI use and increase user trust in a variety of social and cultural contexts by employing transparent and explicable decision-making procedures.

5.4 Stronger Security and Privacy:
Future Large Language Models will prioritize security and privacy by implementing strategies like federated learning and on-device learning. By enabling models to learn from data without directly accessing or retaining sensitive information, these methods help safeguard user privacy and lower the possibility of data exploitation.

5.5 Integration with Systems in the Real World:
In order to carry out challenging real-world tasks, LLMs will be more thoroughly connected with resources like databases, search engines, software programs, and robotics. [14] [15]

VI. CONCLUSION

Since the beginning of rule-based and statistical natural language processing, large language models (LLMs) have advanced significantly. From basic linguistic norms to sophisticated transformer-based systems, their development has made it possible for models to comprehend, produce, and reason with human language on a never-before-seen scale. Modern LLMs are useful tools for both industry and research because they can perform tasks including text production, summarization, translation, conversational AI, and more.
LLMs have significantly impacted society and AI. They have boosted output, encouraged knowledge work, made accessibility and linguistic inclusion possible, and promoted innovation in the fields of business, education, and the arts. They also bring up ethical, social, and environmental issues, such as high computing costs, bias, false information, and hazards to data privacy.
Recommendations for development include resolving biases, maintaining openness, safeguarding user privacy, reducing environmental effect, and matching AI outputs with human values in order to guarantee that these potent technologies are utilized responsibly. LLMs can continue to have a good societal influence while reducing potential risks by adhering to these guidelines.

REFERENCES

[1] Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.;Dong, Z.; et al. A Survey of Large Language Models. arXiv 2024, arXiv:2303.18223.
[2]. Papageorgiou, E.; Chronis, C.; Varlamis, I.; Himeur, Y. A Survey on the Use of Large Language Models (LLMs) in Fake News. Future Internet 2024, 16, 298.
[3] Wang,L.; Ma, C.; Feng, X.; Zhang, Z.; Yang, H.; Zhang, J.; Chen, Z.; Tang, J.; Chen, X.; Lin, Y.; et al. A Survey on Large Language Model Based Autonomous Agents. Front. Comput. Sci. 2024, 18, 186345.
[4] Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention Is All You Need. In Proceedings of the Advances in Neural Information Processing Systems 30 (NIPS 2017)
[5]. Y. Ye, H. You, and J. Du, Improved trust in human-robot collaboration with chatgpt, IEEE Acess, 2023.
[6].Songyue Han, MingyuWang et.al, A Review of Large Language Models: Fundamental Architectures, Key Technological Evolutions, Interdisciplinary Technologies Integration, Optimization and Compression Techniques, Applications, and Challenges, Electronics 2024.9
[7].Sunit Jana1, Rakhi Biswas et.al, The Evolution and Impact of Large Language Model Systems: A Comprehensive Analysis, alochana journal, volume 13 issue 3 2024, ISSN NO: 2231-6329.
[8]. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.-A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. LLaMA: Open and Efficient Foundation Language Models. arXiv 2023, arXiv:2302.13971.
[9].Devlin, J.; Chang, M.-W.; Lee, K.; Toutanova, K. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understand ing. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT 2019), Minneapolis MN, USA, 2–7 June 2019; Association for Computational Linguistics: Stroudsburg, France, 2019; Volume 1, pp. 4171–4186.
[10].https://www.geeksforgeeks.org/artificial-intelligence/exploring-the-technical-architecture-behind-large-language-models
[11]. Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F.L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. GPT-4 Technical Report. arXiv 2024, arXiv:2303.08774.
[12]. Hassanin, M.; Anwar, S.; Radwan, I.; Khan, F.S.; Mian, A. Visual Attention Methods in Deep Learning: An In-Depth Survey. arXiv 2022, arXiv:2204.07756.
[13].Radford, A.; Narasimhan, K. Improving Language Understanding by Generative Pre-Training. 2018.
[14].R. Collobert and J. Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In Proceedings of the 25th international conference on Machine learning, pages 160–167. ACM, 2008.
[15].Raiaan, M.A.K.; Mukta, M.S.H.; Fatema, K.; Fahad, N.M.; Sakib, S.; Mim, M.M.J.; Ahmad, J.; Ali, M.E.; Azam, S. A Review on Large Language Models: Architectures, Applications, Taxonomies, Open Issues and Challenges. IEEE Access 2024, 12, 26839–26874.