International Journal for Multidisciplinary Research (IJFMR)
E-ISSN: 2582-2160 ● Website: www.ijfmr.com ● Email: editor@ijfmr.com

A Comprehensive Review of Large Language Models: Architecture, Types, Challenges, and Future Directions

Kulwinder Kaur[1], Manisha[2]

[1]MCA Student, Department of Compttter Science, Global Group of Institutes
[2]Assistant Professor, Department of Computer Science, Global Group of institutes

ABSTRACT

"LLM play a big role are they how generate AI is improving today." Their success is highly attributed to the Transformer architecture, which uses self-attention mechanisms and large-scale training corpora to model complex linguistic dependencies and long-range relationships. This review synthesizes the theoretical foundations that underpin LLM design, emphasizing the role of self-attention in contextual understanding and outlining the structural distinctions between encoder–decoder systems and modern decoder-only generative models. The paper further investigates practical applications of LLMs across domains such as content production, conversational systems, decision support, and specialized analytical workflows. Despite their capabilities, LLMs face persistent challenges including hallucination, context limitations, computational demands, and issues of trustworthiness. People are concerned that the system might be unfair or biased, fairness, transparency, and interpretability continue to influence are they deployment take decisions. Looking ahead, the field is shaped by emerging trends including multimodal expansion, efficiency-oriented model compression, retrieval-augmented techniques for factual grounding, and the development of agentic systems capable of autonomous task execution. While LLMs hold transformative potential, realizing their long-term societal benefits requires continued progress toward more reliable, efficient, and ethically aligned systems.

KEYWORDS: AI language models, Transformers, attention, model types, errors, bias, retrieval methods, multimodal AI, AI agents

INTRODUCTION

The rapidly evolution of generate artificial intelligence has redefined the trajectory of modern AI research. At the very center of what's transforming are Large Language Models, which contrast sharply with earlier task-specific machine learning systems. Whereas traditional models were designed for narrow, predefined objectives, LLMs demonstrate broad generalization abilities across diverse tasks including reasoning, summarization, coding, translation, and open-ended text generation.

These models now serve as flexible tools across research, industry, and education, adapting to varied forms of input and diverse operational requirements. Surveys the major domains in which these models are deployed, analyzes technical and ethical limitations, and outlines key research directions expected to shape future generations of language-based AI systems.

ARCHITECTURE AND THEORETICAL FOUNDATIONS

The evolution from early statistical methods to modern LLMs represents a major shift in natural language processing. Traditional n-gram models struggled with sparsity and limited context windows, often failing to capture the semantic richness of language. Recurrent neural networks and Long Short-Term Memory architectures improved sequence modeling by incorporating temporal memory

The introduce of the Transformer architecture makes a pivotal breakthrough came with Transformers. Unlike RNNs, Transformers process all the input tokens at the same time, which makes training faster and helps them understand big-picture relationships in the data. This mechanism significantly enhances contextual representation and supports scalability to extremely large model sizes. Because self-attention lacks inherent sequential ordering, positional encodings are integrated to maintain structural information. Contemporary LLMs typically fall into two architectural categories. These models can change text as such translating or summarizing it. Decoder-only models, by contrast, process text through a single autoregressive stream and have become the preferred architecture for generative tasks, forming the backbone of modern systems such as GPT. The development lifecycle of an LLM includes large-scale pretraining on diverse text corpora, followed by fine-tuning for domain-specific applications.

APPLICATIONS

LLMs underpin advanced dialogue systems capable of understanding user intent, maintaining multi-turn context, and adapting tone or style to situational needs. Domain-adapted versions of LLMs extend their impact into specialized fields: in medicine, they assist with clinical summarization and diagnostic reasoning; in finance, they support risk evaluation and market trend analysis and in scientific research .“Because they can handle complex reasoning and adapt to many different tasks, LLMs are becoming important tools for automating work in knowledge-intensive industries.”

CHALLENGES

Despite their impressive capabilities, LLMs exhibit several notable limitations. A widely recognized issue is hallucination, in which models generate statements that appear coherent but lack factual grounding. Context handling remains another challenge: although recent LLMs support extended context windows, maintaining global coherence across very long documents or multi-step reasoning tasks continues to be difficult.

LLMs also demand substantial computational resources for both training and inference, restricting access for organizations with limited hardware capacity and raising environmental sustainability concerns. As model sizes grow, the acquisition of diverse, high-quality training data becomes increasingly challenging, especially under evolving legal and ethical constraint. Biases embedded in training corpora can lead to outputs that reinforce harmful stereotypes or produce inequitable outcomes. Furthermore, the internal reasoning processes of LLMs remain opaque, complicating efforts to audit, interpret.

FUTURE DIRECTIONS

Current research aims to improve the efficiency, adaptability, and reliability of LLMs. One major direction is multimodal integration, text with images, audio, video, and structured data to create models capable of richer understanding and more natural interaction.

Retrieval-augmented generation techniques are gaining attention for improving factual accuracy by grounding output in external knowledge sources. Additionally, the emergence of agentic. Research in

interpretability, alignment, and safety remains crucial, ensuring that increasingly powerful systems behave in ways that are transparent, controllable, and aligned with human values.

REFERENCES

1. Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention Is All You Need. Better ways for computers to think and learn.
2. Brown, T. B., Mann, B., Ryder, N., et al. (2020).Language models can learn new tasks from just a few examples
3. OpenAI. (2023). GPT-4 Technical Report.
4. Bommasani, R., Hudson, D., Adeli, et al. (2021). On the Opportunities and Risks of Foundation Models. Stanford HAI.
5. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). It teaches computers to understand language by reading a lot.
6. Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018).Training a model to create text makes it smarter at understanding words and sentences.
7. Zhang, S., Guu, K., Tay, Y., et al. (2021). This approach helps language models answer questions or do tasks by looking up information first."
8. Wei, J., Wang, X., Schuurmans, et al. (2022). If you ask the model to show its steps, it reasons better.
9. Zhou, Y., Zhang, Y., Chen, J., et al. (2023). A Survey on Large Language Models: Architectures, Applications, and Future Trends.