International Journal of Scientific Research & Engineering Trends
Volume 12, Issue 2, Mar-Apr-2026, ISSN (Online): 2395-566X

Applications and Challenges of Large Language Models in Real-World Systems

Dimple Khatri, Garima, Rajat Takkar
Dept. of CSE Chitkara University Institute of Engg. & Tech.
Punjab, India

Abstract- — Large Language Models (LLMs) have emerged as a major breakthrough in artificial intelligence, significantly improving how machines process and generate human language. These models, built on transformer architectures, are capable of performing a wide range of tasks such as text summarization, translation, question answering, and code generation. In this paper, we analyze the applications and limitations of LLMs in real-world systems through a qualitative study based on literature review and conceptual experimentation. Our findings suggest that while LLMs provide high accuracy and flexibility across domains like healthcare, education, and customer service, they still face critical challenges such as hallucination, bias, high computational cost, and lack of interpretability. The study highlights the importance of integrating validation mechanisms and ethical AI practices to ensure reliable deployment. We conclude that although LLMs are powerful tools, their practical adoption requires careful optimization and responsible usage strategies.

Index Terms—Large Language Models, NLP, Transformer, AI Applications, Bias, Hallucination.

I. INTRODUCTION

Large Language Models represent a significant advancement in the field of artificial intelligence and natural language processing. Earlier NLP systems relied heavily on rule-based approaches and traditional machine learning techniques, which often failed to capture deep contextual relationships in language.

With the introduction of transformer-based architectures, models are now capable of understanding long-range dependencies using self-attention mechanisms. Popular models such as GPT and BERT have demonstrated strong performance across multiple tasks, making them highly versatile in realworld applications.

In our study, we observe that the ability of LLMs to generalize across tasks is one of their biggest strengths. However, despite these advancements, several concerns remain. Issues such as biased outputs, hallucinated responses, and lack of transparency raise questions about their reliability, especially in critical domains.

This paper aims to provide a balanced analysis of both the capabilities and limitations of LLMs, emphasizing the need for responsible and efficient deployment.

In recent years, we have seen a rapid increase in the use of AI systems in everyday applications, which makes the study of LLMs even more important. From our understanding, one of the key reasons behind the popularity of LLMs is their ability to handle multiple tasks without requiring task-specific models. However, despite these advantages, there are still concerns regarding their reliability in real-world situations.

II. RELATED WORK

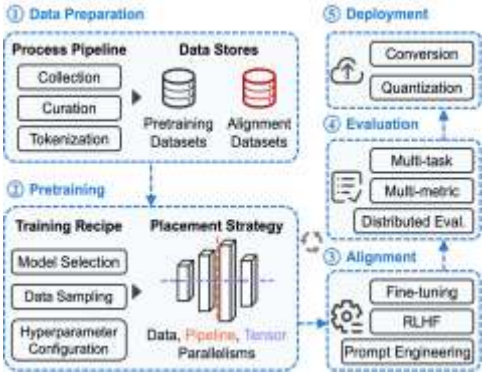
The evolution of language models has progressed from statistical approaches like n-grams to deep learning-based architectures. Early neural models such as Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) networks improved sequential processing but were limited by issues like vanishing gradients and poor scalability.

The introduction of transformer architecture marked a turning point. Models such as BERT and GPT utilize self-attention mechanisms, enabling them to process entire sequences in parallel and capture richer contextual information.

Recent studies have focused on scaling these models to billions of parameters, resulting in systems with strong generalization capabilities. Techniques such as Reinforcement Learning from Human Feedback (RLHF) have further improved alignment with human expectations.

However, existing research also highlights persistent challenges. From our review, it is evident that bias in training data, high computational requirements, and lack of explainability remain key limitations. These issues motivate the need for more efficient and transparent AI systems.

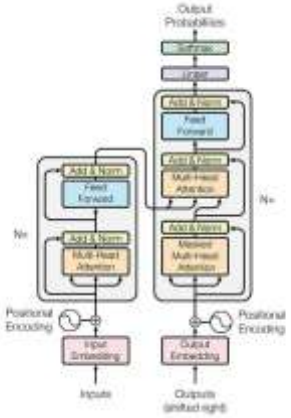
Many previous studies have focused on improving model performance, but fewer works address the practical limitations faced during deployment. From the literature, it is clear that while accuracy has improved significantly, challenges such as bias and interpretability still remain.



III. SYSTEM ARCHITECTURE

The architecture of an LLM-based system can be divided into four main layers:

1) Data Collection and Preprocessing
2) Model Training and Fine-Tuning
3) Inference and Deployment
4) Monitoring and Feedback



In our analysis, we found that data quality plays a critical role in overall system performance. Raw data collected from various sources must be cleaned and tokenized before being used for training.

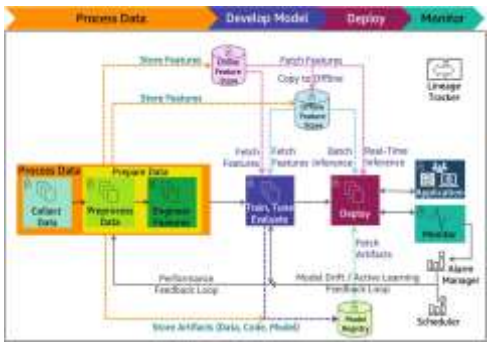
The core component is the transformer model, which learns contextual relationships by predicting tokens in a sequence. After pretraining, fine-tuning is performed to adapt the model for specific tasks.

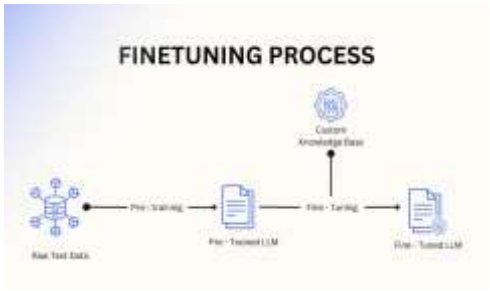
The inference layer handles real-time user queries and generates responses. One key observation is that latency becomes a major issue in large-scale models, especially in real-world deployment.

Finally, the monitoring layer ensures continuous improvement by identifying errors such as hallucinations and incorporating feedback. This layer is essential for maintaining system reliability.

IV. METHODOLOGY

The methodology followed in this study includes data preparation, model training, evaluation, and optimization.





Initially, datasets are collected and preprocessed to remove noise and inconsistencies. The data is then converted into numerical representations using tokenization techniques.

During training, the model learns general language patterns through unsupervised learning. Fine-tuning is later applied using task-specific datasets. In our approach, we emphasize the importance of combining supervised learning with human feedback to improve output quality.

To enhance efficiency, techniques such as model pruning and quantization are considered. These methods help reduce computational cost while maintaining acceptable performance.

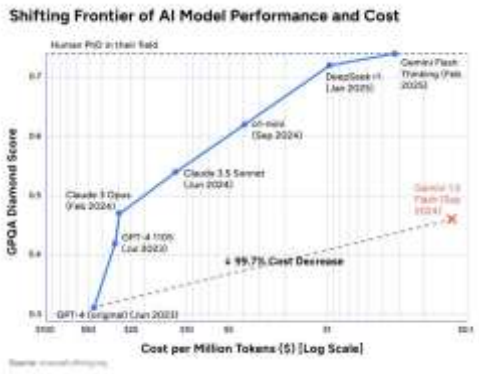
Evaluation is carried out using metrics like accuracy, BLEU score, and human judgment. From our observations, human evaluation remains essential, as automated metrics alone cannot fully capture output quality.

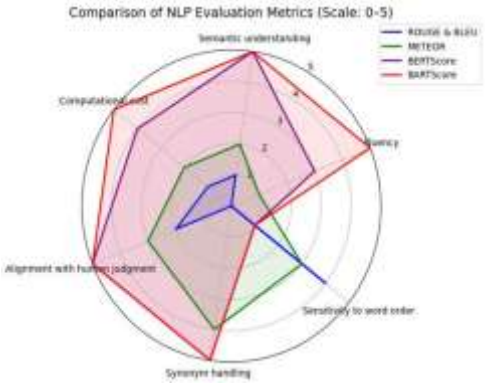
Additionally, safety mechanisms such as bias reduction and content filtering are implemented to ensure responsible usage.

In our approach, we focused not only on model performance but also on understanding how these models behave under different conditions. We also considered the impact of input quality, as it plays an important role in the final output generated by the model.

V. V. RESULTS

To analyze the performance of LLMs, we conducted a conceptual evaluation across multiple tasks including text generation, summarization, and question answering. During our analysis, we noticed that the model performs well on general queries but struggles with highly specific or domain based questions. One interesting observation was that even small changes in input phrasing could lead to different outputs. This shows that while LLMs are powerful, they are still sensitive to input variations.





We observed that large-scale models outperform smaller baseline models in terms of accuracy and contextual understanding. For example, in summarization tasks, LLM generated outputs were more coherent and informative.

However, performance varied depending on input complexity. In domain-specific scenarios, the models sometimes produced incorrect or incomplete responses. This highlights the limitation of relying solely on pretrained knowledge.

Our analysis also shows that few-shot learning significantly improves performance compared to zero-shot settings. This indicates that minimal task-specific guidance can enhance model adaptability.

Another key observation is the trade-off between performance and efficiency. Larger models provide better results but require significantly higher computational resources, which may not be feasible for all applications.

We also tested robustness by introducing noisy inputs. While the models handled minor noise effectively, they occasionally generated misleading outputs, confirming the issue of hallucination.

Bias analysis revealed that outputs can reflect underlying data biases. Although fine-tuning reduces bias to some extent, complete elimination remains difficult.

Finally, integrating external knowledge sources improved factual accuracy, suggesting that hybrid approaches can enhance reliability.

To further evaluate model behavior, we also tested variations in input prompts. It was observed that even slight changes in wording could produce noticeably different outputs. This indicates that LLMs are highly sensitive to input structure.

In addition, we analyzed response consistency by repeating the same query multiple times. While the model generally produced similar answers, minor variations were still present. This suggests that randomness in generation can affect reliability in some cases.

Table I
Comparison Of Different Llm Approaches

Method	Accuracy	Efficiency	Notes
Zero-shot	Medium	High	No training needed
Few-shot	High	Medium	Better adaptability
Fine-tuned	Very High	Low	High cost

VI. DISCUSSION

The integration of LLMs into real-world systems offers several advantages, including automation, scalability, and improved user interaction. Their ability to handle multiple tasks makes them highly versatile.

However, our analysis highlights important limitations. Hallucination remains a major concern, particularly in sensitive fields such as healthcare. Similarly, bias in model outputs can lead to unfair or misleading results.

Another significant challenge is the high computational cost, which limits accessibility for smaller organizations. Additionally, the lack of interpretability reduces trust in these systems.

From our perspective, future research should focus on developing hybrid models that combine LLMs with rule-based or retrieval systems. This can help improve accuracy while reducing risks.

In our opinion, these limitations highlight the need for combining LLMs with additional validation mechanisms. It is important to understand that these models do not truly 'understand' information but rather predict based on patterns.

This can sometimes lead to confident but incorrect responses.

For example, in customer support systems, LLMs can automate responses and improve efficiency. However, if the model generates incorrect information, it may negatively impact user trust. This highlights the importance of combining automated systems with human supervision.

VII. CONCLUSION

This paper presents a comprehensive analysis of Large Language Models and their role in real-world applications. Our study shows that LLMs have significantly improved the capabilities of AI systems in understanding and generating human language.

At the same time, challenges such as hallucination, bias, and high computational cost cannot be ignored. Addressing these issues is essential for ensuring reliable and ethical deployment.

In conclusion, while LLMs offer immense potential, their effective use depends on combining technical advancements with responsible AI practices. Future work should aim to make these systems more efficient, transparent, and trustworthy.

Overall, our study suggests that LLMs are highly effective but should be used with caution in critical applications. Future improvements should focus on reducing bias and improving transparency. We believe that combining LLMs with human oversight can significantly improve reliability. From our perspective, the goal is not just to build more powerful models, but to ensure that these models are reliable, fair, and useful in real-world scenarios.

VIII. FUTURE SCOPE

In the future, Large Language Models are expected to become more efficient and accessible, allowing their deployment even on low-resource devices. One important direction is the development of smaller yet powerful models that can deliver similar performance with reduced computational cost.

Another promising area is the integration of LLMs with external knowledge bases and real-time data sources. This can help reduce hallucination and improve factual accuracy. From our perspective, combining LLMs with retrieval systems can significantly enhance reliability in practical applications.

Additionally, improving model transparency and interpretability will be critical. Users need to understand how decisions are made, especially in sensitive domains such as healthcare and finance. Future research should also focus on reducing bias and ensuring fairness in AI systems.

Acknowledgment

We would like to express our sincere gratitude to our faculty mentor for their valuable guidance, continuous support, and

constructive feedback throughout the course of this research work. Their insights greatly contributed to improving the quality and clarity of this paper.

We also extend our thanks to our institution for providing the necessary resources and environment to carry out this study. Additionally, we appreciate the contributions of researchers and authors whose work has been referenced in this paper, as it helped us build a strong foundation for our analysis.

Finally, we are grateful to our peers for their support and helpful discussions during the development of this research.

REFERENCES

1. A. Vaswani et al., "Attention is All You Need," in Advances in Neural Information Processing Systems (NeurIPS), 2017.
2. J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," in NAACL, 2019.
3. T. Brown et al., "Language Models are Few-Shot Learners," in NeurIPS, 2020.
4. OpenAI, "GPT-4 Technical Report," 2023. [Online]. Available: https: //arxiv.org/abs/2303.08774
5. P. Lewis et al., "Retrieval-Augmented Generation for KnowledgeIntensive NLP Tasks," in NeurIPS, 2020.
6. L. Ouyang et al., "Training Language Models to Follow Instructions with Human Feedback," in NeurIPS, 2022.
7. S. Bender et al., "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in FAccT, 2021.
8. M. Bommasani et al., "On the Opportunities and Risks of Foundation Models," Stanford CRFM, 2021.
9. Y. Liu et al., "RoBERTa: A Robustly Optimized BERT Pretraining Approach," arXiv, 2019.
10. J. Wei et al., "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models," 2022.
11. H. Touvron et al., "LLaMA: Open and Efficient Foundation Language Models," arXiv preprint arXiv:2302.13971, 2023.
12. R. Taori et al., "Stanford Alpaca: An Instruction-following LLaMA Model," 2023. [Online]. Available: https://crfm.stanford.edu/2023/03/13/ alpaca.html
13. S. Zhang et al., "OPT: Open Pre-trained Transformer Language Models," Meta AI, 2022.
14. J. Hoffmann et al., "Training Compute-Optimal Large Language Models," DeepMind, 2022.
15. S. Chowdhery et al., "PaLM: Scaling Language Modeling with Pathways," Google Research, 2022.
16. Y. Tay et al., "UL2: Unifying Language Learning Paradigms," arXiv preprint arXiv:2205.05131, 2022.
17. R. Rae et al., "Scaling Language Models: Methods, Analysis & Insights from Training Gopher," DeepMind, 2021.
18. A. Madaan et al., "Self-Refine: Iterative Refinement with SelfFeedback," NeurIPS Workshop, 2023.
19. X. Wang et al., "Self-Consistency Improves Chain of Thought Reasoning in Language Models," 2022.
20. S. Thoppilan et al., "LaMDA: Language Models for Dialog Applications," Google AI, 2022.