A Comprehensive Review of Generative AI and Large Language Models Techniques, Trends, and Challenges

Mr. Deshmukh Harshad Mangesh[1], Mr. Tathe S.G.[2]
[1]*Student, Vishwabharati Academy's College of Engineering, Sarola Baddi,Ahmednagar*
[2]*Guide, Vishwabharati Academy's College of Engineering, Sarola Baddi,Ahmednaga*

Abstract—**Generative artificial intelligence has emerged as one of the most consequential developments in modern computing. Large language models, in particular, have made it possible for machines to produce fluent text, code, and other forms of content with a level of flexibility that was previously difficult to achieve. This review examines the technical foundations of these systems, with emphasis on transformer architectures, self-supervised pretraining, fine-tuning strategies, reinforcement learning from human feedback, parameter-efficient adaptation, and retrieval-augmented generation. It also considers recent directions in the field, including multimodal models, domain-specific systems, and the growing integration of generative AI into practical workflows across industry and research. Alongside these advances, the paper discusses persistent concerns related to factual reliability, bias, interpretability, privacy, computational demand, and governance. The goal is to provide a clear and balanced overview of where the field stands, what has enabled its progress, and what obstacles still need to be addressed before these systems can be used responsibly at scale.**

Index Terms—**Generative AI, Large Language Models, Transformers, Self-Supervised Learning, RLHF, Multimodal Models, AI Ethics, Bias and Fairness, Natural Language Processing, Artificial Intelligence Trends.**

I. INTRODUCTION

Generative AI has changed the role of machine learning from prediction alone toward content creation and interactive assistance. Instead of simply classifying inputs or estimating outcomes, modern generative systems can draft documents, summarize information, generate code, solution questions, and guide decision-making in herbal language. This shift has made large language models central to both academic research and commercial deployment.

The technical breakthrough behind this progress is the transformer architecture. By relying on self-attention rather than recurrent processing, transformers can model long-range relationships in data more effectively and train more efficiently on large corpora. When combined with self-supervised learning, these models can learn broad linguistic and contextual patterns from massive unlabeled datasets. Later stages of adaptation, including supervised instruction tuning and reinforcement learning from human feedback, help align model outputs with user expectations and enhance the realistic usability of the system.

At the same time, the field has moved beyond simple text generation. Contemporary generative models are increasingly multimodal, capable of working across text, images, audio, and video. They are also being specialized for domains such as medicine, law, finance, education, and software engineering. These developments have expanded the usefulness of the technology, but they have also made evaluation, safety, and governance more difficult.

This overview makes a specialty of 3 questions. What technical methods define modern generative AI and large language models? Which research and deployment trends are shaping the field now? And what limitations continue to constrain reliability, transparency, and responsible use?

II. CORE TECHNIQUES

Transformer-based models are at the heart of the majority of modern large language systems. Their attention mechanism lets the model weigh several portions of a sequence dynamically, therefore enhancing its ability to manage context and dependency across lengthy inputs. Generation happens token by token in decoder-only systems, whereas encoder-decoder architectures continue to be relevant for jobs needing structured input-output processing, like translation and summarization.

The step giving these models general linguistic expertise is self-supervised pretraining. The model learns from massive volumes of text using objectives like next-token prediction instead of depending on manually labeled examples. This method lets the system learn grammar, semantics, global knowledge, and discourse patterns from enormous amounts of data.

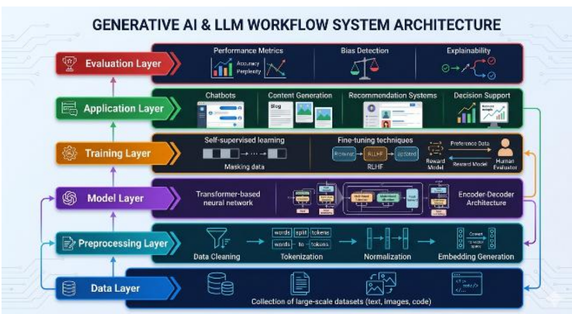

Fig1. Generative AI & LLM Workflow System Architecture

Then, fine-tuning modifies the pretrained model for certain interactions or tasks. While reinforcement learning from human feedback improves alignment with human judgments of helpfulness and safety, instruction tuning increases model sensitivity to prompts and directives. Parameter-efficient techniques such as Low-Rank Adaptation are crucial in real life as they lower the expense of modifying large models without re-training every parameter.

One of the most frequent flaws of language models, the propensity to produce fluent but unfounded statements, is remedied by retrieval-augmented generation. Linking the model to outside knowledge sources during inference, retrieval systems provide grounding that can enhance factual accuracy and maintain outputs nearer to contemporary or domain-specific data. Though it does not remove error, it does make the system more appropriate for applications whereby factual correctness matters.

III. RECENT TRENDS

Multimodality is among the most obvious patterns in the research. Newer systems are being developed to comprehend and produce across various data types instead of viewing text as the sole input and output style. Although this expands use opportunities, it also makes benchmarking more difficult as performance has to be evaluated across many modalities and task combinations.

Specialization is another trend. Though general-purpose models continue to be strong, there is growing curiosity about domain-specific LLMs prepared or tailored for certain industries. These models are appealing because specialised data and language frequently call for more than a basic foundation model can offer. In essence, this has produced great impetus in vertical sectors whereby accuracy and contextual relevance outweigh open-ended inventiveness.

Efficiency is the third trend. The expense of training and deploying huge models has prompted research into compression, distillation, quantization, sparse architecture, and other approaches meant to lower computing load. These methods are important not only for cost but also for accessibility as smaller models can sometimes be used in situations where there is no huge infrastructure.

The unit of innovation is becoming more and more the system rather than only the model; this is the fourth trend. Generative AI systems nowadays frequently mix retrieval, memory, tools, safety filters, and monitoring with a language model. This points to a useful lesson from deployment: The reliability of the finished item relies on the whole pipeline, not simply the pretrained network at its core.

IV. CHALLENGES

Though progress is accelerating, a number of concerns still need to be addressed. The most obvious one is maybe hallucination: While staying factually incorrect, models might provide coherent and confident sounding responses. This is particularly concerning in high-risk environments such as law, public communication, and healthcare.

Another ongoing worry is bias. Since models are trained using vast volumes of internet and institutional data, they could replicate prejudices, distribute groups unequally, or function better on certain languages and communities than on others. Therefore, bias mitigation is not one technical solution but rather a continuous process that includes data collecting, model training, evaluation, and post-deployment surveillance.

Still restricted is interpretation. It is sometimes challenging to understand why a certain result was generated even if a model is successful. Especially in situations when users want explanations rather than merely replies, that complicates auditing, debugging, and governance.

Privacy and security also need to be considered. Sensitive information from training data may be memorized by big models; they can also be altered by adversarial prompting, jailbreaks, or compromised retrieval sources. In parallel, the computational demands of training and inference have actual infrastructure and energy costs.

At last, regulation is catching up with deployment. Many regulatory systems have been outpaced by the rate of adoption, raising unresolved concerns regarding liability, disclosure, openness, and acceptable use. The field has to tackle institutional issues in addition to technical ones as a result.

V. APPLICATIONS

Generative AI is already in use in knowledge retrieval, decision support, document drafting, summarization, coding assistance, educational applications, and customer care. Many of these environments find the model valuable not for replacing human judgment but rather for speeding up everyday chores and increasing information accessibility.

The best applications are those that mix generation and restrictions. When accuracy counts, retrieval, review, and human supervision increase the reliability of the system. Thus, the most realistic future for generative AI is most likely one in which models function as components within bigger processes rather than as freestanding entities.

VI. CONCLUSION

The design of intelligent systems is undergoing a significant change with generative AI and large language models. Transformer architectures, self-supervised learning, fine-tuning, and retrieval-based grounding have propelled their advancement; multimodality, specialization, and efficiency gains have influenced their more general acceptability. At the same time, the field still grapples with major worries about truthfulness, bias, interpretability, privacy, computing, and governance. Therefore, a thorough analysis of the subject should view ability and responsibility as linked because the long-term worth of these systems will rely not just on what they may produce but also on how dependably and responsibly they might be applied.

REFERENCES

[1] E. Cambria et al., “XAI Meets LLMs: A Survey of the Relation Between Explainable AI and Large Language Models,” arXiv preprint, 2024.

[2] W. Fan et al., “A Survey on Retrieval-Augmented Generation for Large Language Models,” arXiv preprint, 2024.

[3] A. Kostikova et al., “LLLMS: A Data-Driven Survey of Limitations of Large Language Models,” arXiv preprint, 2025

[4] S. Yazdani et al., “Generative AI in Depth: A Survey of Recent Advances and Model Variants,” Journal of Big Data, 2025.

[5] F. Marozzo et al., “Generative AI and Large Language Models: Future Directions,” Future Internet Journal, 2026.

[6] S. Mohammadabadi et al., “A Survey of Large Language Models: Evolution and Applications,” Electronics, 2025.

[7] S. Lier et al., “Generative AI and Large Language Models: A Trustworthiness Process Model,” Research Study, 2025.

[8] J. P. Andersen et al., “Generative AI in Research Workflows and Applications,” ScienceDirect Journal, 2025.

[9] T. Brown et al., “Language Models are Few-Shot Learners,” Advances in Neural Information Processing Systems (NeurIPS), 2020.