Generative AI in the Post-Transformer Era: Advances, Ethical Dilemmas, and Future Directions

Pragya Mishra, Department of Computer Applications, Invertis University, Bareilly, pragyamishra.tech@gmail.com
Akshat Rastogi, Department of Computer Applications, Invertis University, Bareilly, akshatrastogi6425@gmail.com
Rajiv Ghai, Department of Computer Science and Engineering, Surajmal University, Kichha, India, rajivghai7@gmail.com
Deepak Kumar Pathak, Department of Computer Applications, Invertis University, Bareilly, India, eminentpathak@gmail.com

***Abstract*—Generative Artificial Intelligence (GenAI) has significantly influenced the advancement of machine learning and deep learning, particularly through Transformer-based models that have dominated the field over the past few years. These models, like GPT and BERT, revolutionized tasks such as text generation, summarization, and translation. However, as the limitations of Transformers—such as computational inefficiency, hallucinations, and lack of reasoning—become more evident, researchers are exploring alternative architectures and hybrid models. The emergence of State Space Models, Diffusion Models, and retrieval-augmented techniques marks the onset of the post-Transformer era. This shift is not merely architectural but also ethical and practical.**

As generative models grow in capability and societal impact, critical concerns around misinformation, bias, environmental cost, and intellectual property have surfaced. Addressing these challenges requires innovations in ethical AI design, mechanisms for hallucination mitigation, energy-efficient computation, and strategies to foster collaborative human-AI creativity. This review paper provides an overview of these recent advances beyond Transformer architectures, examines the pressing ethical dilemmas posed by GenAI, and outlines future research directions. The goal is to inform a responsible, sustainable evolution of generative systems that balances innovation with safety, inclusiveness, and interpretability.

Keywords: Generative AI, post-transformer, large language models, ethical AI, diffusion models, AI hallucination, AI governance.

1. Introduction

Generative Artificial Intelligence (GenAI) has undergone significant advancements, transitioning from basic rule-based systems to highly complex models capable of producing human-like content across various modalities. Early breakthroughs were driven by models such as Generative Adversarial Networks (GANs) and Variational Autoencoders (VAEs), which revolutionized the fields of synthetic image and data generation. However, the real paradigm shift occurred with the advent of Transformer-based models, most notably introduced by Vaswani et al. in 2017 [1]. These models—like GPT, BERT, and their subsequent iterations—have become the backbone of state-of-the-art GenAI systems. Their architecture, characterized by self-attention mechanisms and large-scale parallelization, has enabled generative models to scale up to hundreds of billions or even trillions of parameters, significantly improving their ability to generate coherent and contextually relevant content across text, audio, image, and video domains.

Despite their success, Transformer-based models are not without challenges. Issues such as high computational demands, memory inefficiency, susceptibility to hallucinations (i.e., generating false or misleading information), and limitations in logical reasoning have become increasingly apparent. These drawbacks have prompted researchers to explore alternative or enhanced model architectures that can address these limitations while maintaining or improving performance. As a result, the post-Transformer era has begun to take shape, characterized by innovations like state space models, hybrid retrieval-augmented networks, and diffusion-based generative techniques.

This paper provides a comprehensive review of these emerging GenAI paradigms beyond Transformers. It also highlights the pressing ethical and technical concerns surrounding generative models, such as misinformation, bias, environmental impact, and intellectual property conflicts. Finally, the study outlines future directions aimed at fostering responsible innovation in GenAI—prioritizing transparency, trustworthiness, efficiency, and human-AI alignment as the technology continues to evolve and integrate more deeply into society.

2. Advances in the Post-Transformer Era

Recent advancements in generative AI highlight a growing shift away from the exclusive reliance on Transformer-based architectures. While Transformers have demonstrated exceptional performance across natural language processing and multimodal tasks, their limitations in scalability, computational inefficiency, and handling long-range dependencies have led to the emergence of alternative architectures and hybrid approaches. This section discusses five prominent directions redefining generative AI in the post-Transformer era.

2.1. State Space Models and RWKV Architecture

State Space Models (SSMs) have gained attention for their ability to handle long-context sequences more efficiently than Transformers. The RWKV architecture—a fusion of Recurrent Neural Networks (RNNs) and Transformer characteristics—adopts a time-mixing mechanism with minimal memory usage. This architecture preserves temporal order without the quadratic complexity of attention mechanisms, making it suitable for applications requiring long-term memory such as document summarization, scientific simulations, and long-form text generation. Unlike traditional Transformers, RWKV supports efficient inference on CPUs and edge devices, highlighting its practicality and resource efficiency [2].

2.2. Structured State Spaces: Mamba and S4

Structured State Space models such as Mamba and S4 extend the capabilities of SSMs by offering linear time and space complexity while preserving competitive performance across various sequence modeling benchmarks. Mamba, in particular, integrates selective state updates with learnable dynamics, making it a scalable and interpretable alternative to attention-based models. These models are designed for real-time AI deployments, especially in latency-sensitive environments like mobile AI, wearables, and embedded systems. By overcoming the bottlenecks of Transformer models in processing long inputs, SSMs represent a breakthrough in both speed and adaptability [3].

2.3. Diffusion Models and Flow Matching Techniques

Diffusion models, exemplified by Stable Diffusion and Imagen, adopt a generative process based on iterative noise reduction. These models eschew autoregressive generation in favor of denoising techniques, offering superior control over the generation process and enabling high-resolution image synthesis. Flow Matching further optimizes this concept by learning a continuous and deterministic transformation between data and latent distributions. This advancement results in faster sampling, higher fidelity, and increased diversity of outputs—crucial for creative tasks in design, media, and entertainment [4][5].

2.4. Hybrid Architectures: RETRO, GLaM, and Sparse Mixture of Experts

Hybrid models integrate the strengths of Transformers with complementary mechanisms. RETRO (Retrieval-Enhanced Transformer) incorporates external knowledge retrieval during generation to improve factual accuracy and reduce hallucination. GLaM (Generalist Language Model) leverages a Mixture-of-Experts framework to dynamically activate subsets of model parameters during inference, improving computational efficiency. These models enhance scalability and allow better knowledge grounding, thereby refining contextual understanding and preserving precision in complex generative tasks [6][7].

2.5. Multimodal and 3D Generative Models

The future of GenAI lies in its ability to operate across multiple modalities. Models like Flamingo and GPT-4V unify textual, visual, and spatial inputs to facilitate rich content generation, including 3D modeling, robotics control, and interactive media. These models represent a move toward embodied AI, capable of perception, reasoning, and action across diverse environments. Multimodal GenAI is increasingly important in AR/VR, autonomous navigation, and human-computer interaction, where contextual and sensory integration is essential [8].

Together, these innovations are reshaping the GenAI landscape with architectures that are more efficient, context-aware, and versatile.

3. Ethical Dilemmas and Societal Impact

As generative AI continues to evolve, several ethical dilemmas and societal concerns have become increasingly significant. These challenges include misinformation, societal biases, environmental impacts, intellectual property conflicts, and safety alignment, all of which demand urgent attention in the post-Transformer era of AI development. Fig. 1 illustrates the major ethical and societal dilemmas in Generative AI, placing *Hallucination*

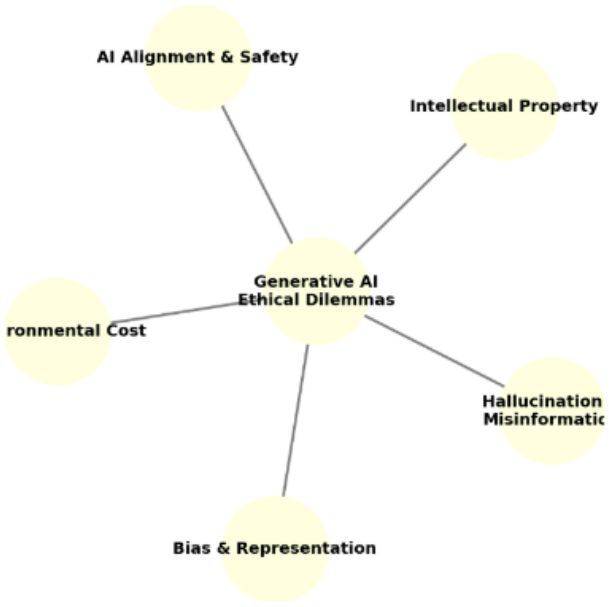

Figure 1. Ethical and Societal Dilemmas in AI Adapted from Ji et al. [9], Strubell et al. [11], Samuelson [12], and Anthropic [13].

& Misinformation, Bias & Representation, Environmental Cost, Intellectual Property, and AI Alignment & Safety as the key focus areas. This conceptual map visually emphasizes how these interlinked challenges surround the core of Generative AI development and deployment.

3.1. Hallucination and Misinformation

One of the most pressing challenges in generative AI is its tendency to produce hallucinated content—information that appears coherent and credible but is factually inaccurate or entirely fabricated. This phenomenon becomes particularly dangerous in high-stakes domains such as healthcare, law, and education. For instance, a generative model providing incorrect medical advice or misinterpreting legal statutes can lead to serious real-world consequences. As these models lack inherent mechanisms for truth verification, researchers have introduced methods like Retrieval-Augmented Generation (RAG), which supplements generative processes with externally retrieved documents. Fact-checking modules, integrated pipelines for real-time validation, and hybrid architectures are also being explored to reduce these risks and enhance the reliability of AI outputs [9].

3.2. Bias and Representation

Generative models are trained on large-scale internet data, which often contains explicit and implicit biases. Consequently, these models can generate outputs that reinforce harmful stereotypes or exhibit discriminatory language and ideologies. For example, gender, racial, and cultural biases may emerge in generated texts, images, or audio. The replication of such biases raises concerns about fairness, inclusivity, and the broader social impact of AI systems. Addressing these issues requires multi-pronged strategies including the use of curated and representative datasets, adversarial training to reduce bias, and transparent model development processes. Techniques like bias detection tools and algorithmic audits are gaining prominence as part of an ethical AI toolkit [10].

3.3. Environmental Costs

The environmental footprint of training and deploying large generative models is substantial. For instance, the training of OpenAI's GPT-3 reportedly consumed around 1287 megawatt-hours (MWh) of electricity, contributing significantly to carbon emissions [11]. With the proliferation of increasingly larger models, there is growing concern over their sustainability. In response, the field is witnessing a shift toward greener AI practices. These include the use of sparse architectures, model distillation to reduce redundancy, weight quantization for lower power usage, and hardware-efficient training algorithms. Future models are expected to balance performance with ecological responsibility, ensuring that innovation does not come at the cost of environmental degradation.

3.4. Intellectual Property and Creativity

Generative AI blurs the boundary between original human creativity and machine-generated content. With AI now capable of composing music, generating artwork, and writing stories, legal frameworks struggle to define ownership, copyright, and authenticity. The unauthorized replication of styles, voices, or textual formats raises concerns about plagiarism and intellectual property theft. Governments and legal bodies are beginning to explore policies for watermarking AI content, licensing training data, and defining the rights of AI-generated creations [12].

3.5. AI Alignment and Safety

Ensuring that generative AI systems align with human values, ethical principles, and societal norms is a complex but essential challenge. Misalignment may lead to unintended or harmful behavior, especially as models gain more autonomy. Techniques like Reinforcement Learning from Human Feedback (RLHF) and Constitutional AI are being developed to fine-tune model behavior in accordance with predefined ethical guidelines. These methods aim to create safer, more interpretable, and controllable AI systems that can be trusted in critical applications [13].

4. Future Directions in Generative AI

Table I outlines the key domains within Generative AI where significant challenges exist and highlights the corresponding future directions to address these issues.

Table I: Emerging Challenges and Future Directions in Generative AI Domains

Domain	Challenges	Future Directions
Model Architecture	High complexity, slow inference	State-space models, sparse transformers
Trust & Accuracy	Hallucinations, misinformation	Verified generation, external fact modules
Ethics & Bias	Prejudice, marginalization	Bias audit tools, fairness-aware training
Energy Efficiency	High training cost	Green AI, low-rank approximation
Application Domains	Narrow modalities	Multimodal, embodied AI
Creativity & IP	Plagiarism risks	Watermarking, AI-content regulation

In terms of model architecture, the high computational complexity and slow inference speeds are being tackled through the adoption of state-space models and sparse transformers. To improve trust and accuracy, especially in combating hallucinations and misinformation, researchers are introducing verified generation techniques and integrating external fact-checking modules.

When addressing ethics and bias, which include concerns about societal prejudices, the future lies in bias auditing tools and fairness-aware training methods. The environmental concern of energy efficiency is being responded to with approaches like Green AI and low-rank approximations, which aim to reduce the carbon footprint of large models.

For expanding the range of application domains, the solution is the development of multimodal and embodied AI, allowing models to handle diverse input types and real-world interactions. Finally, in the realm of creativity and intellectual property (IP), where risks of plagiarism and ownership ambiguity are high, technologies like digital watermarking and AI content regulation frameworks are proposed to ensure responsible and traceable content generation.

5. Conclusion

Generative AI is rapidly evolving beyond the Transformer era, ushering in models that are more efficient, interpretable, and ethically aligned. These innovations promise a future of sustainable and inclusive AI systems with broader real-world impact. However, to ensure safe deployment, robust regulatory frameworks are essential to mitigate risks such as misinformation, ethical misuse, and environmental impact. Critical challenges like hallucination, bias, and high energy consumption must be addressed to foster responsible growth. As we move forward, the focus must shift from sheer performance to building trustworthy, transparent, and eco-conscious generative AI models that align closely with human values and societal needs.

References

[1]. Vaswani *et al.*, "Attention Is All You Need," *NeurIPS*, 2017.

[2]. B. Peng *et al.*, "RWKV: Combining RNN and Transformer," *arXiv preprint arXiv:2305.13048*, 2023.
[3]. A. Gupta *et al.*, "Mamba: Linear-Time Sequence Modeling with Selective SSMs," *arXiv preprint arXiv:2312.00752*, 2023.
[4]. P. Dhariwal and A. Nichol, "Diffusion Models Beat GANs on Image Synthesis," *NeurIPS*, 2021.
[5]. L. Lipman *et al.*, "Flow Matching for Generative Modeling," *ICLR*, 2022.
[6]. K. Du *et al.*, "GLaM: Efficient Mixture of Experts," *Google AI Blog*, 2021.
[7]. S. Borgeaud *et al.*, "Improving Language Models by Retrieving from Trillions of Tokens," *Nature*, 2022.
[8]. J. Alayrac *et al.*, "Flamingo: Visual Language Models," *arXiv preprint arXiv:2204.14198*, 2022.
[9]. S. Ji *et al.*, "Survey on Hallucination in NLP," *arXiv:2307.11752*, 2023.
[10]. T. B. Brown *et al.*, "Language Models Are Few-Shot Learners," *NeurIPS*, 2020.
[11]. E. Strubell, A. Ganesh, and A. McCallum, "Energy and Policy Considerations for Deep Learning," *ACL*, 2019.
[12]. J. K. Samuelson, "AI and Intellectual Property Law," *Harvard J. Law & Tech*, vol. 34, pp. 1–27, 2021.
[13]. C. Anthropic, "Constitutional AI: Harmlessness Training," *Anthropic Blog*, 2023.
[14]. M. Bommasani *et al.*, "On the Opportunities and Risks of Foundation Models," *Stanford CRFM Report*, 2021.
[15]. D. Amodei *et al.*, "Concrete Problems in AI Safety," *arXiv:1606.06565*, 2016.