Babylonian Journal of Artificial Intelligence
Vol. **(2025)**, 2025, **pp**. 72–76
DOI: https://doi.org/10.58496/BJAI/2025/007 ;ISSN: 2958-6453
https://mesopotamian.press/journals/index.php/BJAI

Research Article

The Rise of Transformers – Redefining the Landscape of Artificial Intelligence

Nyagong Santino David Ladu [1,*], Benson Turyasingura [2,4], Byamukama Willbroad [3], Abraham Atuhaire [2]

[1] *Department of Mathematics and Physics, Rumbek University of Science and Technology, Rumbek, South Sudan.*

[2] *Department of Environment and Natural Resources, Kabale University, Kabale, Uganda.*

[3] *Department of Crop Science and Production, Faculty of Agriculture and Environmental Sciences, Kabale University, Kabale, Uganda.*

[4] *The Institute of Policy Analysis and Research (IPAR-Rwanda), Kigali, Rwanda.*

ARTICLE INFO

Article History

Received 02 Feb 2025
Revised: 04 Mar 2025
Accepted 03 Apr 2025
Published 22 Apr 2025

Keywords

Cloud Computing

Resource Allocation

Fault Tolerance

Energy Efficiency

Service Provisioning

ABSTRACT

The 2017 paper 'Attention Is All You Need' by Vaswani et al. marked a major paradigm shift in AI. Rather than clinging to tired methods like recurrence or convolutions, its Transformer design boldly flipped the script melding self-attention into a fresh take that not only remade natural language processing but also trickled over into computer vision, robotics, quirky multi-modal setups, and more [1]. At its very core lies a refreshingly simple idea self-attention, which lets models dynamically figure out which bits of the input deserve extra focus instead of being bound by strict, linear routines. This clever tweak kicked the old, rigid rules to the curb and opened up levels of parallel processing that we hadn't seen before. In many cases, this shift not only ramped up the training speed dramatically but also proved essential in our data-swamped world where speed and nimbleness really do make all the difference.

1. INTRODUCTION

Transformers quickly became the go-to design for some of the most impressive models we see today. You might think of models like BERT, GPT, T5, ViT—and even those massive multi-input systems such as CLIP or Flamingo—as miracles in their own right. In most cases, though, none of these would have ever come about without that spark of innovation introduced by Vaswani et al. (2017). Its design scales naturally with the data and computing power on hand—a real boon when handling language models that pack in hundreds of billions of parameters. This isn't just about hitting new benchmarks; it's about shifting the way we look at representation learning. Before long, techniques like masked language modeling for pretraining, fine-tuning for specific tasks, and even prompt-based methods just fell into place, almost as if the Transformer nudged us toward them. You also see that methods like attention visualization add a layer of interpretability that deep neural networks famously lacked, boosting both usability and trust in the technology. At the same time, the soaring popularity of Transformers puts us in a spot where rising energy needs and widening research gaps—between, say, academic labs and big industrial players—need a closer look. While attention-based systems deliver some pretty amazing feats, they're not the final chapter in artificial intelligence. Questions about reasoning, memory and, well, truly grounded understanding still hang in the air, nudging researchers to explore alternatives (Radford et al., 2019). Even with all these open ends, it's clear that the Transformer marks a milestone—a rare mix of elegance, practicality, and sheer endless potential. It shows, in a very tangible way, how one breakthrough in design can ripple across fields, reshaping our interactions with machines, as seen especially in projects that meld multiple types of inputs. Now, if we're to build on what this innovation has given us, doing it responsibly is key. Whether it's tweaking designs for energy efficiency, bracing systems against tricky adversarial prompts, or working to open up model development to a broader community, the road ahead calls for both technical skill and ethical sensitivity. Attention might not solve every puzzle, but it sure lit the spark that has redefined our grasp of intelligence.

*Corresponding author. Email: nyagongsantino19@gmail.com

2. THE TRANSFORMER REVOLUTION

Before 2017, RNNs and CNNs held court in many top-tier NLP systems, powering tasks from language translation to reading sentiment. Those models had their fair share of issues though; for instance, they couldn't parallelize work well when faced with huge datasets, gradients would often vanish during deep training (making weight updates lose their punch), and connecting words spread far apart in a sequence was just plain tricky. Then, the self-attention trick from Transformers came along – it lets every word check in on every other word at once, which generally speaking cut training times and ramped up performance. This change kind of opened the door for scalable pretraining, a practice that's now pretty much standard. What's more, the Transformer's modular, position-agnostic setup makes it a great fit for transfer learning and handling multiple tasks at once. All in all, these tweaks have reshaped the AI scene, stirring major shifts in both research and practical applications [4][5].

3. BEYOND NLP: THE TRANSFORMER'S CROSS-DOMAIN EXPANSION

Transformers have been jumping into places you wouldn't expect—thanks largely to their flexible design and the sheer volume of available data. In a way, massive datasets helped them break free from the traditional role of just handling text. NLP, for one, saw a real boost. Take BERT; it started using masked language tricks to train bidirectional models, which generally helped with getting the context right. It's interesting how this approach added a twist to understanding text. Then there's GPT, which made waves by predicting upcoming words instead of relying on a two-way look; The autoregressive modeling approach adopted by GPT fundamentally transformed text generation paradigms. Around the same time, the T5 model bundled a bunch of NLP tasks into one nelikeat text-to-text framework, showing just how versatile transformers can be. And then, almost as a surprise, the Vision Transformer (ViT) stepped into the spotlight. It turned out that when fed enough data, transformers could actually outdo traditional CNNs in image recognition—basically resetting the benchmarks. Later on, models like the Swin Transformer came along, mixing in hierarchical and spatial ideas so they could handle both fine local details and big picture patterns at the same time. All these shifts clearly point to a broader trend: the growing range of what transformers can tackle, stretching far beyond the limits set by older architectures. This evolution is something you can also see in discussions around multimodal AI [6], [7]—and it's nicely illustrated in some of the imagery that shows these cross-domain leaps.

4. MULTIMODAL MODELS

CLIP [25] and Flamingo [56] are multimodal models that lately have shown transformers can work with both pictures and words. These models push the limits in zero-shot learning and retrieval tasks, demonstrating that integrating multimodal inputs significantly enhances model performance. For instance, as highlighted in , when you combine text, images and clinical notes, you generally see better reasoning in medical settings. This kind of data blending turns out to be key for building smart systems that deal with tricky user interactions and real-world challenges—often more effectively than before. Also, as pointed out in [8], bringing together information from different sources gives a subtler feel for context, which improves how the system interacts. Overall, these innovative frameworks not only deepen our grasp of machine learning but also open up solid paths towards applications in fields like affective computing and personalized healthcare [9].

5. RW AND RL PART 2A: TRANSFORMERS IN DECISION MAKING

Decision Transformer shakes up old models of trajectory planning by mixing language into agent actions in a way that feels fresh and experimental. It opens up unexpected routes for agents to handle decision-making tasks by blending sequential choices with language-based insights—which, in many cases, means that performance in complex reinforcement learning settings gets a nice boost [10]. Transformers now work with a variety of input types, which makes interactions between agents and their surroundings a bit more nuanced; time signals and context hints come together to enrich learning experiences naturally [11]. All in all, this change isn't just about using transformers differently—it signals a broader rethink in AI strategies, especially in the realm of reinforcement learning, where the effects of this technology are widespread and continue to evolve as researchers and practitioners push its boundaries.

6. SCALABILITY VS. FOUNDATION MODELS

Transformers really shine when you feed them more varied data, expand their model sizes, and throw extra computational power at them—they tend to perform better as you scale them up. A shift in thinking led folks to coin what we now call "foundation models" [12]; basically these models get a broad, general training first and then are fine-tuned for more specific tasks. Take GPT-3 and Gemini, for example—they're now key players behind smart assistants, advanced coding tools, lively chatbots and even creative contnet generators. Their abilities aren't limited to one trick; they handle translation,

summarization, dialogue, reasoning and even creative generation, largely owing to the core Transformer design. In most cases, ongoing research is probing how these models scale further and what practical uses they might have, such as in weather forecasting and climate modeling, where even state-of-the-art methods benefit from refined AI techniques [13].refer to figure 1.

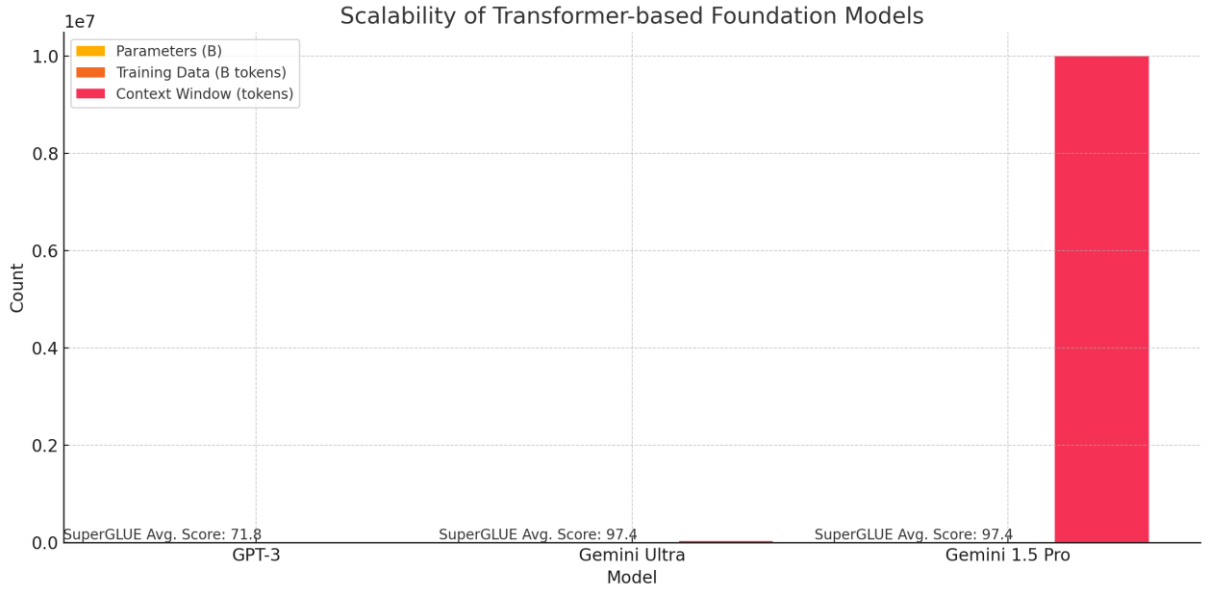

Fig. 1. Scalability and Performance of Transformer-Based Models[21]

7. CHALLENGES AND CRITICISMS

The Transformer is widely praised for its groundbreaking role in AI, yet it comes bundled with some pretty hefty challenges. One big worry is the steep cost of computation. Take GPT-3, for example—it had to harness thousands of GPUs over several weeks, a feat that naturally raises concerns about energy use and overall accessibility. Then there's the bias and hallucination issue—pretrained transformers tend to pick up on the biases hidden in their training data, which can sometimes lead to outputs that are off or downright misleading. And interpretability remains murky; even after making some headway with attention mechanisms, these massive models still often operate like an inscrutable black box. On top of that, these models have an almost insatiable appetite for data. Relying on internet-scale sources means we're constantly left wondering about data quality, copyright compliance, and fairness. All of these challenges have sparked a new wave of research aimed at designing more efficient transformer designs, pushing green AI initiatives, and even mapping out governance frameworks tailored to these core models. As of 2025, the Transformer has become not only a technical milestone but also a cultural force—fueling everything from coding assistants to AI art generators and reshaping how we interact with our machines. Yet, even as it dominates the field, it's important to keep asking tough questions. Many issues, like reasoning, blending in symbolic methods, energy efficiency and genuine generalisation, still aren't fully resolved. In most cases, as researchers, we need to shift our focus not only toward developing new architectures but also toward setting up smarter frameworks for their responsible deployment. Ultimately, the Transformer has redefined AI across different fields and enabled models to scale in ways once thought impossible, urging us to drive its evolution with both bold ambition and careful thought, so that the future we build is not only intelligent but also transparent, inclusive and aligned with human values [14][15]. Refer to table 1.

TABLE I. CHALLENGES AND CRITICISMS OF AI TRANSFORMERS[22]

Challenge	Description
Monopolization and Corporate Control	The development of large language models (LLMs) like GPT-3 has led to increased monopolization in the AI industry, with big tech companies dominating the market due to their substantial computational resources and data access. This trend raises concerns about the concentration of power and the potential stifling of innovation from smaller entities. ([journals.sagepub.com](https://journals.sagepub.com/doi/10.1177/20539517211047734?utm_source=openai))
Environmental Impact	Training and deploying LLMs require significant computational power, leading to high energy consumption and environmental concerns. The environmental footprint of these models is substantial, contributing to increased carbon emissions. ([journals.sagepub.com](https://journals.sagepub.com/doi/10.1177/20539517211047734?utm_source=openai))

Bias and Ethical Concerns	LLMs can perpetuate and amplify biases present in their training data, leading to outputs that may reinforce stereotypes or discriminate against certain groups. This raises ethical questions about the fairness and inclusivity of AI systems. ([journals.sagepub.com](https://journals.sagepub.com/doi/10.1177/20539517211047734?utm_source=openai))
Misinformation and Disinformation	Generative AI systems can produce 'hallucinations'—misinformation that appears credible—and can be used to purposefully create false information, posing risks to information integrity and public trust. ([gao.gov](https://www.gao.gov/products/gao-23-106782?utm_source=openai))
Privacy Risks	Specific technical features of generative AI systems may reduce privacy for users, including minors. For example, a generative AI system may be unable to 'forget' sensitive information that a user wishes to delete. ([gao.gov](https://www.gao.gov/products/gao-23-106782?utm_source=openai))

Funding

This study was carried out without receiving financial support from any institutions or sponsors.

Conflicts of Interest

The author's paper explicitly states that no funding was received from any institution or sponsor.

Acknowledgment

The authors express heartfelt thanks to the institution for providing steadfast moral support and encouragement throughout the course of this research.

References

[1] S. Feuerriegel, J. Hartmann, C. Janiesch, and P. Zschech, “Generative AI,” Bus. Inf. Syst. Eng., 2023. [Online]. Available: https://doi.org/10.1007/s12599-023-00834-7. [Accessed: Mar. 27, 2025].

[2] M. Al-Azzawi, “Artificial Intelligence in Red Teaming,” 2024. [Online]. Available: https://core.ac.uk/download/638725749.pdf. [Accessed: Mar. 27, 2025].

[3] F. García-Peñalvo and A. Vázquez-Ingelmo, “What Do We Mean by GenAI? A Systematic Mapping of The Evolution, Trends, and Techniques Involved in Generative AI,” Int. J. Interact. Multimed. Artif. Intell., 2023. [Online]. Available: https://core.ac.uk/download/586370846.pdf. [Accessed: Mar. 27, 2025].

[4] G. Arzilli et al., “ChatGPT and the rise of large language models: the new AI-driven infodemic threat in public health,” 2023. [Online]. Available: https://core.ac.uk/download/587992546.pdf. [Accessed: Mar. 27, 2025].

[5] J. Tagliapietra, “Generative AI models: evaluations, comparison and critical aspects,” 2024. [Online]. Available: https://core.ac.uk/download/636345506.pdf. [Accessed: Mar. 27, 2025].

[6] A. Bewersdorff et al., “Multimodality of AI for Education: Towards Artificial General Intelligence,” 2023. [Online]. Available: http://arxiv.org/abs/2312.06037. [Accessed: Mar. 27, 2025].

[7] G. Glass, “Fidelity, Remix, and the Adaptive Potential: Leveraging AI and ML Techniques in Literary Adaptation Theory,” Univ. North Carolina at Chapel Hill, 2024. [Online]. Available: https://core.ac.uk/download/619686585.pdf. [Accessed: Mar. 27, 2025].

[8] P. P. Liang, “Foundations of Multisensory Artificial Intelligence,” 2024. [Online]. Available: http://arxiv.org/abs/2404.18976. [Accessed: Mar. 27, 2025].

[9] D. Askounis et al., “A graph neural architecture search approach for identifying bots in social media,” Frontiers Media S.A., 2024. [Online]. Available: https://core.ac.uk/download/636100871.pdf. [Accessed: Mar. 27, 2025].

[10] J. Gjølstad and N. A. S. Skants, “Artificial Intelligence in Organizational Decision-Making,” Handelshøyskolen BI, 2023. [Online]. Available: https://core.ac.uk/download/639535292.pdf. [Accessed: Mar. 27, 2025].

[11] M. B. Chhetri et al., “A2C: A Modular Multi-stage Collaborative Decision Framework for Human-AI Teams,” 2024. [Online]. Available: http://arxiv.org/abs/2401.14432. [Accessed: Mar. 27, 2025].

[12] J. Tagliapietra, “Generative AI models: evaluations, comparison and critical aspects,” 2024. [Online]. Available: https://core.ac.uk/download/636345506.pdf. [Accessed: Mar. 27, 2025].

[13] D. S. Civitarese et al., “AI Foundation Models for Weather and Climate: Applications, Design, and Implementation,” 2023. [Online]. Available: http://arxiv.org/abs/2309.10808. [Accessed: Mar. 27, 2025].

[14] S. Mayahi and M. Vidrih, “Generative AI-Driven Storytelling: A New Era for Marketing,” 2023. [Online]. Available: http://arxiv.org/abs/2309.09048. [Accessed: Mar. 27, 2025].

[15] S. R. Mishra and H. Mohapatra, “Exploring AI Tool's Versatile Responses: An In-depth Analysis Across Different Industries and Its Performance Evaluation,” 2023. [Online]. Available: http://arxiv.org/abs/2307.05909. [Accessed: Mar. 27, 2025].

[16] L. Cabral et al., “Education in the age of Generative AI: Context and Recent Developments,” 2023. [Online]. Available: http://arxiv.org/abs/2309.12332. [Accessed: Mar. 27, 2025].

[17] University for Business and Technology - UBT, “International Conference on Energy,” UBT Knowledge Center, 2023. [Online]. Available: https://core.ac.uk/download/616936066.pdf. [Accessed: Mar. 27, 2025].
[18] C. Escobar-Velásquez, M. Linares-Vásquez, and A. Mastropaolo, “The Rise and Fall(?) of Software Engineering,” 2024. [Online]. Available: http://arxiv.org/abs/2406.10141. [Accessed: Mar. 27, 2025].
[19] S.-Y. Chen, “Tsinghua Issue—Generative AI, Learning And New Literacies,” The Aquila Digital Community, 2023. [Online]. Available: https://core.ac.uk/download/591795349.pdf. [Accessed: Mar. 27, 2025].
[20] A. Bewersdorff et al., “Multimodality of AI for Education: Towards Artificial General Intelligence,” 2023. [Online]. Available: http://arxiv.org/abs/2312.06037. [Accessed: Mar. 27, 2025].
[21] Foundation Model Supply Chain Diagram, 2025. [Online]. Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2023/07/Foundation-model-supply-chain-updated-June-2024.png.
[22] Overview of Multimodal Self-Supervised Training in Medical Applications, 2025. [Online]. Available: https://media.springernature.com/full/springer-static/image/art%3A10.1038%2Fs41586-023-05881-4/MediaObjects/41586_2023_5881_Fig1_HTML.png.