The Evolution of Generative AI and Transformer-based LLMs

Generative AI has shifted AI from mainly recognizing patterns to **creating** new text, images, audio, and other content. Surveys and historical reviews show a clear trajectory: from early rule-based and statistical generators to deep generative models and today's large, transformer-based foundation models and LLMs (Sengar et al., 2024; Hagos et al., 2024; Trigka & Dritsas, 2025; He et al., 2025; Kılınç & Keçecioglu, 2024; Zhou, 2025; Myers et al., 2023).

Historical Evolution of Generative AI

- **Early & pre-deep learning**: Rule-based systems and statistical/model-based generators (e.g., graphical, probabilistic models) formed the first two stages of generative AI (He et al., 2025).
- **Deep generative models (~2014–2017)**:
  - **VAEs** and **GANs** in 2014 enabled realistic images, sounds, and text by learning data distributions with encoder–decoder or adversarial setups (Sengar et al., 2024; Bengesi et al., 2023; Kılınç & Keçecioglu, 2024; S. et al., 2024; Nah et al., 2023; Kusiak, 2024).
  - Other families: **autoregressive**, **flow-based**, and **diffusion** models broadened high-fidelity generation, especially for images (Sengar et al., 2024; Trigka & Dritsas, 2025; S. et al., 2024; Kusiak, 2024; Çelik & Eltawil, 2024).
- **Transformers and foundation models (2017–now)**: The Transformer architecture (2017) replaced recurrence with attention and became the backbone of modern generative models (Sengar et al., 2024; Hagos et al., 2024; Kılınç & Keçecioglu, 2024; S. et al., 2024; Nah et al., 2023; Vaswani et al., 2017). Scaling these to "foundation models" and GPT-style systems marked a major revolution (Bengesi et al., 2023; He et al., 2025; Zhou, 2025; Matarazzo & Torlone, 2025; Myers et al., 2023; Linkon et al., 2024).

Representative LLM Milestones

Year	Model (example)	Parameters	Notes	Citations
2017	Transformer	530M	Core architecture for later LLMs	(Hagos et al., 2024)
2018	BERT	340M	Bidirectional masked LM	(Hagos et al., 2024)
2019	GPT-2	1.5B	Autoregressive text generator	(Hagos et al., 2024)
2020	GPT-3, T5, Gopher	11B–280B	Scaled LLMs on huge corpora	(Hagos et al., 2024; Zhou, 2025)

Figure 1: Illustrative timeline of major Transformer-based LLMs.

Core Mechanisms of Transformers

- **Self-attention & multi-head attention**: Compute how strongly each token should attend to every other, capturing long-range dependencies and contextual relationships in parallel (Sengar et al., 2024; Hagos et al., 2024; Kılınç & Keçecioglu, 2024; S. et al., 2024; Ferraris et al., 2025; Zhou, 2025; Nah et al., 2023; Vaswani et al., 2017).
- **Encoder–decoder (or decoder-only) stacks**: Layers of self-attention + feed-forward networks model complex patterns in sequences; encoder–decoder for translation, decoder-only for many LLMs (Kim et al.,

2024; S. et al., 2024; Vaswani et al., 2017).

- **Positional encoding**: Adds order information so transformers, which process tokens in parallel, can model sequence structure (Sengar et al., 2024; Bengesi et al., 2023; Zhou, 2025; Vaswani et al., 2017; Su et al., 2021). Variants like **Rotary Position Embedding (RoPE)** encode absolute and relative positions efficiently (Su et al., 2021).
- **Expressive power**: Theoretical work shows how depth, number of heads, attention, and positional encodings together determine approximation power for long, sparse, complex sequences (Wang & Weinan, 2024).

From Transformers to Large Language Models

- **LLMs as generative transformers**: LLMs (GPT, BERT-family, T5, PaLM-like models) are transformers trained on massive text corpora, enabling fluent generation, translation, summarization, and question answering (Hagos et al., 2024; Bengesi et al., 2023; Kim et al., 2024; Ferraris et al., 2025; Zhou, 2025; Matarazzo & Torlone, 2025; Myers et al., 2023; Linkon et al., 2024).
- **Scaling laws and foundation models**: Increasing parameters, data, and compute predictably boosts performance, leading to very large “foundation models” with broad, cross-task capabilities (Hagos et al., 2024; He et al., 2025; Ferraris et al., 2025; Zhou, 2025; Matarazzo & Torlone, 2025; Myers et al., 2023; Shen et al., 2024).
- **Training paradigms**: Pre-training on vast unlabeled text, followed by fine-tuning and alignment methods (e.g., RLHF) make models more useful and safer for downstream tasks (Bengesi et al., 2023; Zhou, 2025; Matarazzo & Torlone, 2025; Myers et al., 2023; Linkon et al., 2024).

Applications, Challenges, and Future Directions

- **Applications**: Text, image, video, and code generation; healthcare and radiology; smart manufacturing; scientific discovery; business and 6G wireless systems (Sengar et al., 2024; Bengesi et al., 2023; Trigka & Dritsas, 2025; Kim et al., 2024; S. et al., 2024; Ferraris et al., 2025; Kusiak, 2024; Myers et al., 2023; Linkon et al., 2024; Çelik & Eltawil, 2024).
- **Key challenges**: Bias, hallucinations, interpretability, privacy, compute and energy costs, and risks like deepfakes and misinformation (Sengar et al., 2024; Bengesi et al., 2023; Trigka & Dritsas, 2025; He et al., 2025; S. et al., 2024; Ferraris et al., 2025; Matarazzo & Torlone, 2025; Myers et al., 2023; Linkon et al., 2024).
- **Research trends**: More efficient architectures (including linear-complexity models) with similar scaling behavior to standard transformers (Wang & Weinan, 2024; Su et al., 2021; Shen et al., 2024); multimodal and domain-specialized models; stronger safety and governance frameworks (Sengar et al., 2024; Hagos et al., 2024; Trigka & Dritsas, 2025; He et al., 2025; Ferraris et al., 2025; Matarazzo & Torlone, 2025; Myers et al., 2023; Linkon et al., 2024; Çelik & Eltawil, 2024).

Conclusion

Across many independent reviews, there is strong agreement that generative AI has evolved from rule-based systems and early statistical models to deep generative architectures, and then to transformer-based foundation models and LLMs. Transformers’ self-attention plus positional encoding made large-scale sequence modeling practical, and scaling these models has produced general-purpose systems with powerful generative abilities—alongside significant technical and societal challenges that remain active areas of research.

References

Bengesi, S., El-Sayed, H., Sarker, M., Houkpati, Y., Irungu, J., & Oladunni, T. (2023). Advancements in Generative AI: A Comprehensive Review of GANs, GPT, Autoencoders, Diffusion Model, and Transformers. *IEEE Access, 12*, 69812-69837. https://doi.org/10.1109/access.2024.3397775

Çelik, A., & Eltawil, A. (2024). At the Dawn of Generative AI Era: A Tutorial-cum-Survey on New Frontiers in 6G Wireless Intelligence. *IEEE Open Journal of the Communications Society, 5*, 2433-2489. https://doi.org/10.1109/ojcoms.2024.3362271

Ferraris, A., Audrito, D., Di Caro, L., & Poncibò, C. (2025). The architecture of language: Understanding the mechanics behind LLMs. *Cambridge Forum on AI: Law and Governance*. https://doi.org/10.1017/cfl.2024.16

Hagos, D., Battle, R., & Rawat, D. (2024). Recent Advances in Generative AI and Large Language Models: Current Status, Challenges, and Perspectives. *IEEE Transactions on Artificial Intelligence, 5*, 5873-5893. https://doi.org/10.1109/tai.2024.3444742

He, R., Cao, J., & Tan, T. (2025). Generative artificial intelligence: a historical perspective. *National Science Review, 12*. https://doi.org/10.1093/nsr/nwaf050

Kim, K., Cho, K., Jang, R., Kyung, S., Lee, S., Ham, S., Choi, E., Hong, G., & Kim, N. (2024). Updated Primer on Generative Artificial Intelligence and Large Language Models in Medical Imaging for Medical Professionals. *Korean Journal of Radiology, 25*, 224 - 242. https://doi.org/10.3348/kjr.2023.0818

Kılınç, H., & Keçecioglu, Ö. (2024). Generative Artificial Intelligence: A Historical and Future Perspective. *Academic Platform Journal of Engineering and Smart Systems*. https://doi.org/10.21541/apjess.1398155

Kusiak, A. (2024). Generative artificial intelligence in smart manufacturing. *Journal of Intelligent Manufacturing, 36*, 1 - 3. https://doi.org/10.1007/s10845-024-02480-6

Linkon, A., ✉, M., Sarker, M., Nabi, N., Rana, M., Ghosh, S., Rahman, M., Esa, H., & Chowdhury, F. (2024). Advancements and Applications of Generative Artificial Intelligence and Large Language Models on Business Management: A Comprehensive Review. *Journal of Computer Science and Technology Studies*. https://doi.org/10.32996/jcsts.2024.6.1.26

Matarazzo, A., & Torlone, R. (2025). A Survey on Large Language Models with some Insights on their Capabilities and Limitations. *ArXiv, abs/2501.04040*. https://doi.org/10.48550/arxiv.2501.04040

Myers, D., Mohawesh, R., Chellaboina, V., Sathvik, A., Venkatesh, P., Ho, Y., Henshaw, H., Al-Hawawreh, M., Berdik, D., & Jararweh, Y. (2023). Foundation and large language models: fundamentals, challenges, opportunities, and social impacts. *Cluster Computing, 27*, 1-26. https://doi.org/10.1007/s10586-023-04203-7

Nah, F., Zheng, R., Cai, J., Siau, K., & Chen, L. (2023). Generative AI and ChatGPT: Applications, challenges, and AI-human collaboration. *Journal of Information*

Technology Case and Application Research, 25, 277 - 304. https://doi.org/10.1080/15228053.2023.2233814
S., B., Chirchi, V., Kadry, S., Agoramoorthy, M., P, G., K, S., & A., S. (2024). The Road Ahead: Emerging Trends, Unresolved Issues, and Concluding Remarks in Generative AI—A Comprehensive Review. *International Journal of Intelligent Systems*. https://doi.org/10.1155/2024/4013195
Sengar, S., Hasan, A., Kumar, S., & Carroll, F. (2024). Generative artificial intelligence: a systematic review and applications. *Multimedia Tools and Applications, 84*, 23661 - 23700. https://doi.org/10.1007/s11042-024-20016-1
Shen, X., Li, D., Leng, R., Qin, Z., Sun, W., & Zhong, Y. (2024). Scaling Laws for Linear Complexity Language Models. **, 16377-16426. https://doi.org/10.48550/arxiv.2406.16690
Su, J., Lu, Y., Pan, S., Wen, B., & Liu, Y. (2021). RoFormer: Enhanced Transformer with Rotary Position Embedding. *ArXiv, abs/2104.09864*. https://doi.org/10.1016/j.neucom.2023.127063
Trigka, M., & Dritsas, E. (2025). The Evolution of Generative AI: Trends and Applications. *IEEE Access, 13*, 98504-98529. https://doi.org/10.1109/access.2025.3574660
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L., & Polosukhin, I. (2017). Attention is All you Need. **, 5998-6008.
Wang, M., & Weinan, E. (2024). Understanding the Expressive Power and Mechanisms of Transformer for Sequence Modeling. *ArXiv, abs/2402.00522*. https://doi.org/10.48550/arxiv.2402.00522
Zhou, Z. (2025). Large Language Models and Their Evolution. *Applied and Computational Engineering*. https://doi.org/10.54254/2755-2721/2025.25586