Generative Artificial Intelligence: Opportunities and Challenges of Large Language Models

Fabian Barreto, Lalita Moharkar, Madhura Shirodkar, Vidya Sarode, Saniya Gonsalves, and Aaron Johns

Abstract Artificial Intelligence (AI) research in the past decade has led to the development of Generative AI, where AI systems create new information from almost nothing after learning from trained models. Generative AI can create original work, like an article, a code, a painting, a poem, or a song. Google Brain initially used Large Language Models (LLM) for context-aware text translation, and Google went on to develop Bidirectional Encoder Representations from Transformers (BERT) and Language Model for Dialogue Applications (LaMDA). Facebook created OPT-175B and BlenderBot, while OpenAI innovated GPT-3 for text, DALL-E2 for images, and Whisper for speech. GPT-3 was trained on around 45 terabytes of text data at an estimated cost of several million dollars. Generative models have also been developed from online communities like Midjourney and open-source ones like HuggingFace. On November 30, 2022, OpenAI launched ChatGPT, which used natural language processing (NLP) techniques and was trained on LLM. There was excitement and caution as OpenAI's ChatGPT reached one million users in just five days, and in January 2023 reached 100 million users. Many marveled at its eloquence and the limited supervision with which it generated code and answered questions. More deployments followed; Microsoft's OpenAI-powered Bing on February 7, 2023, followed by Google's Bard on February 8, 2023. We describe the working of LLM and their opportunities and challenges for our modern world.

Keywords Bing · chatGPT · Generative artificial intelligence · Large language models

F. Barreto (✉) · L. Moharkar · M. Shirodkar · V. Sarode
Department of Electronics and Telecommunication, Xavier Institute of Engineering, Mumbai, India
e-mail: frfabiansj@xavier.ac.in

S. Gonsalves
Department of Applied Sciences and Humanities, Xavier Institute of Engineering, Mumbai, India

A. Johns
Department of Information Technology, St. Xavier's College, Mumbai, India

V. E. Balas et al. (eds.), *Intelligent Computing and Networking*, Lecture Notes in Networks and Systems 699, https://doi.org/10.1007/978-981-99-3177-4_41

1 Introduction

Every human culture uses language as a communication system to convey ideas, emotions, and information to one another. As infants, we pick up meaningful word meanings, and as we grow to adulthood become more skilled in adapting our speech. Language also helps us to understand ourselves and the world around us. The uniqueness of the human person in learning a language is the ability to generalize and hence learn from a limited amount of exposure to a language, especially as children. Contrast this with a model trained on deep learning systems. During the testing phase, samples of a different distribution than those trained show a generalization behaviour inconsistent with a human person. Making decisions on unseen data require extensive learning. The authors in Lake et al. [1] elaborate on how humans never learn "from scratch" but use their previous knowledge to learn new tasks. They outline "core ingredients" that humans use to be good at generalization. These ingredients include the domain knowledge of numbers, space, physics, and psychology. These building blocks and diverse experiences result in better generalization for human beings. In Linzen [2], the author looks at how a statistical model extracts generalizations based on the way that inductive biases of the models interact and on the dataset's statistical properties. He suggests that models can be improved by incorporating the human-like inductive biases so that learning can happen with limited data.

Decades of quest followed to develop systems that could generate human-like responses for Natural Language Processing (NLP) tasks like conversation, text completion, and language translation. The Transformer-based models, like OpenAI's Generative Pre-trained Transformer 3 (GPT-3), generate human-like text, which can be used for various applications such as text completion, dialogue systems, and language translation. The text generation process is a token sampling conditional on previous tokens, $a_j \sim p(a_j|a_j \ldots a_{j-1}{:}\phi)$ where a_j denotes the jth token in the text sequence and ϕ denotes sampling distribution parameters. ϕ is optimized on the training data conditioned on N preceding tokens. A large language model can creatively handle novel concepts given as a prompt even if not encountered in the training data [3]. The authors [4] train GPT-3, and test its performance in the few-shot setting. Their 96-layered model trained on 175 billion parameters shows performance that at times exceeds the State-of-the-art (SOTA) fine-tuned models.

The remainder of this paper is organized as follows. The details of Generative Artificial Intelligence (GAI) are illustrated in Sect. 2. We elaborate on the framework of the GPT family in Sect. 3. The opportunities and challenges of LLM are explained in Sect. 4. Finally, Sect. 5 gives the concluding remarks of this paper.

2 Generative AI Models

In AI the models are trained using datasets and then tested on unseen samples of data. The AI system can then apply its model to new, unseen data and make accurate predictions or decisions. The real-world applications help to encounter new and diverse data that was not seen during training. Good generalization, is achieved using techniques such as regularization, cross-validation, and ensemble methods to reduce overfitting. By the process of generalization an AI system performs well on tasks not encountered during training. Generative Artificial Intelligence (GenAI) refers to the use of AI algorithms and models to generate new content, such as realistic computer-generated images and videos, text generation, and audio synthesis, especially music and speech. GenAI can also be used for data augmentation and anomaly detection.

Some prominent GenAI models (e.g. GPT-3) have demonstrated impressive language generation capabilities. NVIDIA's StyleGan generates realistic faces, objects, and even entire scenes. OpenAI's Music Transformer generates original pieces of music in a variety of styles.

Some of the most well-known types of generative AI architectures include Variational Autoencoders (VAEs) and Generative Adversarial Networks (GANs).

2.1 Variational Autoencoders (VAEs)

VAEs consist of an encoder and a decoder [5] as shown in Fig. 1. The encoder takes in data and compresses it into a lower-dimensional representation, called the latent code. The decoder then takes the latent code and generates new data samples that are similar to the original input.

2.2 Generative Adversarial Networks (GANs)

GANs consist of two networks: a generator (G) and a discriminator (D) [6] as shown in Fig. 2. The generator works at creating new data samples, while the discriminator attempts to tell apart the generated samples (fake) from the real samples. Through competition (minimax game of adversarial nature), the generator learns to create more realistic samples, while the discriminator learns to better identify fake samples. D wants to maximize its cost value log(D(x)), and G wants to minimize its cost value log(1-D(G(z))), as given in the Loss Function [6] equation shown below.

$$\text{Loss Function}_{(D,G)} = \mathbb{E}_{x\sim pdata(x)}\left[\log D(x)\right] + \mathbb{E}_{z\sim pz(z)}\left[\log\left(1 - D(G(z))\right)\right] \quad (1)$$

where represents the mathematical expectancy. The authors in Razavi-Far et al. [7] mention the different GAN-based techniques that cater to different types of

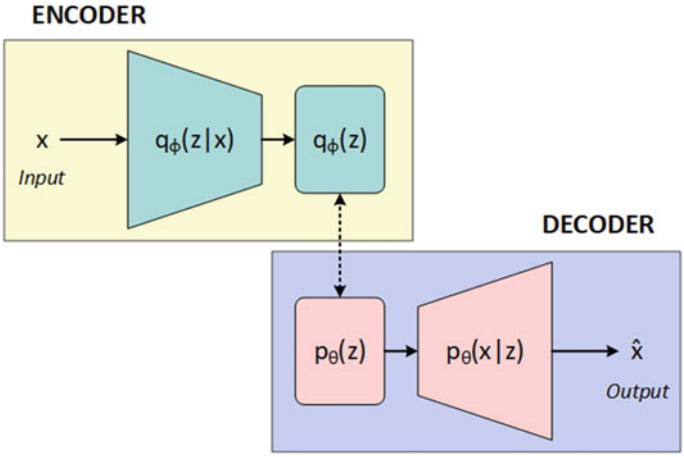

Fig. 1 Variational autoencoders

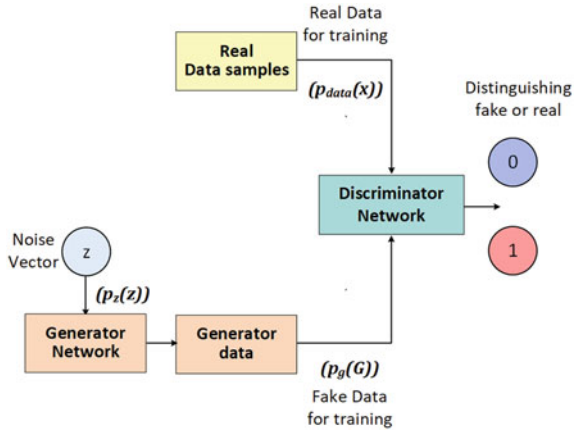

Fig. 2 Generative adversarial networks (adapted from [6])

AI learning problems. They mention the various new GAN architectures and the development of GAN objective functions for specific applications. GANs are also combined with transfer learning and reinforcement learning to achieve noteworthy results in several applications [8]. The authors in Farajzadeh-Zanjani et al. [9] mention the development of GANs beginning with the 1990 Artificial Curiosity to the 2020 Continuous Conditional GAN (CcGAN). They also elaborate the challenges associated with training GANs namely mode collapse, and vanishing gradients. In NLP, text generation using GANs has been a great success, as described by de Rosa and Papa [10], who survey recent studies in GAN-based text generation. Specifically, NLP models need to learn the connection between characters and words (grammar, syntax and semantic properties). The context can be learnt by using pre-trained embedding models such as BERT [11], ELECTRA [12], and GPT-2. Systems can

be trained to deal with adversarial data (slightly modified data probably not present in the training) by having noisy samples during the learning process. GANs by their innate nature learn the data's distribution and hence can augment with artificially generated data.

3 The GPT Family

The Generative Pre-trained Transformer (GPT) is a generative model, innovatively created by OpenAI. Their goal is "to advance digital intelligence in the way that is most likely to benefit humanity as a whole" [13]. The transformer uses the attention mechanism [14] to focus on relevance using self-attention (words association in a sentence) and encoder-decoder attention (between source and target sentences). The transformer's decoder is the GPT. The models can be trained in few-, one-, and zero-shot settings. In few-shot setting, after presenting the task few examples are given to the model. In a one-shot there is only one example and, in a zero-shot there is no example. The GPT family uses tokens to process the text. In regular English text a token will be four characters, about three-quarters of a word. The trained models learn the statistical relationships between tokens and can thus predict the next token.

The decision of OpenAI to release GPT-3 through a public application programming interface (API), opened the world to countless possibilities. Up until 2020, the AI research was available for a selective few researchers and engineers. With OpenAI's API, users throughout the world could get access to this LLM through a simple sign-in.

3.1 The Evolution

OpenAI presented GPT-1 in June 2018 [13], with the transformer architecture and unsupervised (unlabeled data) pre-training. The GPT-1 model was trained on 8 million web pages and had 117 million parameters with a context token size of 512. The model demonstrated that a good pre-trained model could perform the task of generalization. A remarkable advantage was zero-shot learning, where the model could do a task without a previous example, with an application in question answering.

In February 2019, OpenAI released a larger model, GPT-2 having, 1.5 billion parameters with a context token size of 1024, and trained on 40 GB of webpages (WebText dataset). Its purpose was to predict the next word(s) in a sentence. It had a poor performance for specialized tasks such as music and storytelling.

In June 2020, OpenAI released GPT-3 [4] with 175 billion parameters and trained 45 TB of webpages with a context token size of 2048. GPT-3 is pre-trained on a corpus of text from five datasets: Common Crawl, WebText2, Books1, Books2, and

Wikipedia. 93.69% of the total documents are in English. A more advanced architecture (sparsely-gated mixture-of-experts) allows one to better understand the context of conversations and generate more accurate and nuanced text. GPT-3 can hence answer questions, write essays, summarize text, translate language, and generate computer code.

On November 30, 2022, OpenAI launched ChatGPT trained on LLM. Many admired its articulateness and the limited supervision with which it generated code and answered questions. There was excitement and caution as OpenAI's ChatGPT reached one million users in just five days and 100 million by the end of January 2023.

For the next version, GPT-4, the authors [15, 16] envision better parameter-optimized, text-only model. GPT-3 was only trained once as training is costly and hence not optimized. GPT-4 possibly will have better optimization of hyperparameters and better analysis of the optimal model emphasizing sparsity.

3.2 Other Technologies

Microsoft and NVIDIA built Megatron-Turing NLG 530B (MT-NLG) [17], a transformer-based language model with 530 billion parameters. It was trained 15 datasets consisting of a total of 339 billion tokens. Hence modern LLM are evolving thanks to the availability of computational resources, large datasets, and effective software stacks.

More deployments followed after the launch of ChatGPT. Microsoft's OpenAI-powered Bing on February 7, 2023, followed by Google's Bard on February 8, 2023. Bing integrated OpenAI's language model, Prometheus, was explicitly designed for search engines. Prometheus ensures humans are in the loop, incorporating Microsoft's responsible AI [18].

4 Opportunities and Challenges of LLM

4.1 Opportunities

LLM offers several opportunities in different domains. The authors in Bommasani et al. [19] give a detailed description of the capabilities and applications of what they call foundation models (e.g., GPT-3). The capabilities include Language, Vision, Reasoning, and Interaction, to name a few. The applications include healthcare, Law, and Education.

Language, with its nuances of dialect and style, offers the richness and complexity that LLM are trying to learn. The present models exhibit adaptable linguistic capabilities. Applications include classification, generation, and sequence labelling. Multilingual models catering to different languages have also evolved (e.g., mBERT) [20] and can be important use cases for the Indian scenario. In the domain of Computer Vision, raw web-based datasets are readily available for training and have contributed to the enhanced performance of LLM. An evolving area is multimodal integration tasks like image captioning. Reasoning includes proving formal theorems and generating code. Interaction includes lowering the difficulty threshold for developers to proto-type powerful applications with seamless integration.

LLM can be a good interface for healthcare providers when scaling healthcare services. Some relevant examples include summarizing healthcare records and retrieving relevant healthcare cases and literature. Legal applications and processing is another interesting use case for LLM, especially for private, criminal, and public law, where it can provide the necessary context. An example can be an automated brief generation. The digital age post-millennium ushered in the rapid growth of digital learning. LLM can help in making learning more effective for teachers and learners. LLM can provide relevant feedback to learners and help teachers to create personalized content to meet students' needs.

4.2 Challenges

The authors in Bommasani et al. [19] describe the challenges in the context of inequity, misuse, and effects on the environment, legality, economics, and ethics. ITU estimates that approximately 5.3 billion people (66 percent of the world's population) were using the Internet in 2022 [21]. In the context of the digital divide LLM can compound the existing disparities, leading to unfair outcomes. There can be intrinsic and latent biases within the models. Biases of religion, gender, and race. An example is a misrepresentation based on stereotypes and negative outlooks.

In recent days we have heard of the unhinged responses of Bing [22], expressed as angry remarks and bizarre conversations. Yes, the models can attack users with harmful content, leading to trauma and psychological harm. Protocols must be in place to develop and deploy models that mitigate toxic content and rectify harmful responses.

The greatest danger of LLM is the misuse that can happen in the form of fake news or biased propaganda that can harm populations. Exploiting the models' vulnerabilities is another threat. Target content for misinformation is a real-case scenario that can be exploited.

LLM require tremendous amount of data and computing resources and hence has its own contributions to carbon emissions. Once deployed at large scales these models require substantial levels of energy, and therefore, carbon costs. A question that is often raised is the legality of the training data, precisely its collection and use.

There is a need for proper ethical frameworks for publicly accessible data, especially for mass data collection.

5 Conclusion

The AI revolution is here to stay. Generative AI will play a significant role in shaping our human destiny. The task-specific NLP models have evolved to models that can perform a variety of tasks. Large Language Models have shown SOTA performance on various NLP tasks and will continue contributing to the AI innovation landscape. LLM are evolving with greater emergent abilities, that is, abilities not seen in the smaller models are manifested in larger models. Emergent risks will also have to be seen, especially in the context of fake news, which can harm society, not just in generating human-like texts but also in writing codes. Given human conversations' diversity and complexity, AI models face several challenges. Prominent among them is the issue of human bias while training. Hence the need for human persons to be in the loop to monitor the model output, thus detecting and mitigating biases and toxicity. Sustainable computing resources, will have to be looked into when a large number users start using the LLM, as they are computationally resource-heavy. It is also imperative that we as humans use these models responsibly with proper ethical considerations.

References

1. Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ (2017) Building machines that learn and think like people. In: Behavioral and brain sciences, vol 40, p e253
2. Linzen T (2020) How can we accelerate progress towards human-like linguistic generalization? In: arXiv preprint arXiv:2005.00955
3. Värtinen S, Hämäläinen P, Guckelsberger C (2022) Generating role-playing game quests with GPT language models. In: IEEE transactions on games
4. Brown T et al (2020) Language models are few-shot learners. In: Advances in neural information processing systems, vol 33, pp 1877–1901
5. Kingma DP, Welling M (2013) Auto-encoding variational bayes: In arXiv preprint arXiv:1312.6114
6. Goodfellow I et al (2014) Generative adversarial nets. In: Ghahramani Z et al (eds) Advances in neural information processing systems 27. Curran Associates Inc., Boston, pp 2672–2680
7. Razavi-Far R et al (2022) An introduction to generative adversarial learning: architectures and applications. Razavi-Far R et al (eds) Generative adversarial learning: architectures and applications, Intelligent Systems Reference Library, pp 1–4. Springer, Cham
8. Sarmad M, Lee HJ, Kim Y (2019) RL-GAN-Net: a reinforcement learning agent controlled gan network for real-time point cloud shape completion. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp 5891–5900
9. Farajzadeh-Zanjani M et al (2022) Generative adversarial networks: a survey on training, variants, and applications. Razavi-Far R et al (eds) Generative adversarial learning: architectures and applications, Intelligent Systems Reference Library, pp 7–29. Springer, Cham

10. de Rosa GH, Papa JP (2021) A survey on text generation using generative adversarial networks. Pattern Recogn 119:108098
11. Devlin J et al (2019) Bert: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of 2019 conference of the North American chapter of the association for computational linguistics, pp 4171–4186. Human Language Technologies
12. Clark K et al (2020) Electra: pre-training text encoders as discriminators rather than generators. In arXiv preprint arXiv:2003.10555
13. OpenAI https://openai.com/blog/introducing-openai/. Accessed 1 Feb 2023
14. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, vol 30
15. Datacamp https://www.datacamp.com/blog/what-we-know-gpt4. Accessed 1 Feb 2023
16. Romero A https://towardsdatascience.com/gpt-4-is-coming-soon-heres-what-we-know-about-it-64db058cfd45. Accessed 3 Feb 2023
17. Smith S et al (2022) Using deepspeed and megatron to train megatron-turing nlg 530b, a large-scale generative language model. In arXiv preprint arXiv:2201.11990
18. Bisson S What does microsoft bing's new AI assistant mean for your business? https://www.techrepublic.com/article/microsoft-new-ai-assistant/. Accessed 9 Feb 2023
19. Bommasani R et al (2021) On the opportunities and risks of foundation models. In arXiv preprint arXiv:2108.07258
20. Gonen H et al (2020) It's not Greek to mBERT: inducing word-level translations from multilingual BERT. arXiv preprint arXiv:2010.08275
21. International Telecommunication Union https://www.itu.int/en/ITU-D/Statistics/Pages/stat/default.aspx. Accessed 11 Feb 2023
22. Forbes https://www.forbes.com/sites/siladityaray/2023/02/16/bing-chatbots-unhinged-responses-going-viral/?sh=dabeebb110c6. Accessed 11 Feb 2023