EXP – 1 DATE	Comprehensive Report on the Fundamentals of Generative AI and Large Language Models (LLMs)

NAME:V.A.JITHENDRA

REG.NO:212221230043

AIM: Comprehensive Report on the Fundamentals of Generative AI and Large Language Models (LLMs)

Topic 1: Introduction to Generative AI

Aim:

• To introduce the concept of Generative AI, explain how it works, and discuss its applications and challenges.

Generative Artificial Intelligence (AI) models have a vast application landscape and use cases as they can help enterprises automate intelligence through a knowledge base across multiple domains. These models can help scale up innovation in AI development across sectors. Negative use cases of generative AI include disinformation spread and influence operations. Organizations and governments are attempting to address these concerns through practices like responsible data collection, ethical principles for AI, and algorithmic transparency.

The democratisation of Artificial Intelligence (AI) with new technology platforms is gaining significant importance, with tech giants like Google, Microsoft and Baidu challenging each other in the business of Generative AI. The Large Language Models (LLMs) and Generative AI models like OpenAI's ChatGPT, which has been put out in the public domain, have created a stir online and within communities about the possibilities of AI replacing humans. The expansion of LLMs has gained momentum in the past two years with the introduction of AI-based chatbots and conversational agents taking the online marketplace.

Their ability to handle diverse tasks like answering complex questions, generating text, sounds, and images, translating languages, summarising documents, and writing highly accurate computer programmes has brought them into the public eye. These models can synthesise information from billions of words from the web and other sources and give a sense of fluid interaction. Amidst the hype around these models, the less debated issue is the possibility of these tools generating falsehoods, biases, and other ethical considerations.
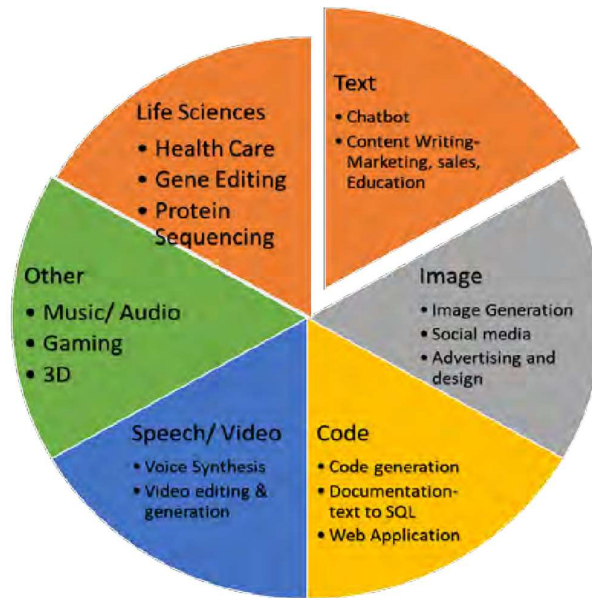
Generative AI models

Generative AI systems refer to the class of machine learning where the system is trained to generate new data or content like audio, video, text, images, art, music, or an entire virtual world of things. These models study the statistical patterns and structures from the training data and discover new information on different samples that resembles the original data. In addition, these models are trained on humongous amounts of data; they seem creative when they produce a variety of unexpected outputs that make them look genuine.

Various Generative AI models include Variational Autoencoders, Auto Aggressive Models, and Generative Adversarial Networks. Generative AI models have varied applications today, from image generation to music creation, data augmentation and more. The area gaining the most significance today is the text generation tools, also known as large language models. Various leading companies and labs are doing R&D in this field.

The LLM originated in 2017, and one of the first such models was Bidirectional Encoder Representation from Transformer (BERT) and Generative Pre-trained Transformer (GPT) by Google and OpenAI, which were open-sourced in the same year. Following the idea, many such models originated, like OpenAI's GPT2, PaLM 540B, Megatron 530B, GitHub's Copilot, Stable Diffusion, and InstructGPT.[1] More recently, next-generation tools like ChatGPT, DALL-E-2 and Google's Language Model for Dialogue Applications (LaMDA) have become the internet sensation. These LLMs are trained on large amounts of data (petabytes) and are used for zero-shot or few-shot scenarios where little domain knowledge is available, so they can start generating data based upon just a few prompts. For instance, OpenAI's GPT-3 is a 175 billion parameter model and can generate text and code from a very short prompt.[2]

The Generative AI models have a vast application landscape and use cases. Therefore, these models can help enterprises automate intelligence through a knowledge base across multiple domains shown in Figure 1. In addition, these models have the capability to scale up innovation in AI development across sectors.

Generative AI Applications

Text
- Chatbot
- Content Writing- Marketing, sales, Education

Life Sciences
- Health Care
- Gene Editing
- Protein Sequencing

Image
- Image Generation
- Social media
- Advertising and design

Other
- Music/ Audio
- Gaming
- 3D

Code
- Code generation
- Documentation- text to SQL
- Web Application

Speech/ Video
- Voice Synthesis
- Video editing & generation

ChatGPT is a generative AI based on transformer architecture that generates natural language responses to the given prompt. It is a type of autoregressive model that produces a sequence of text based on the previous tokens in sequence. ChatGPT has revolutionised people's interaction with technology so that it seems as if one person

is talking to another person. It was first introduced in 2018 by OpenAI and is based.
upon InstructGPT with changes in data collection setup, and in November 2022, it was made public for user feedback. Mesmerised users posted on social media what this chatbot can do—like producing code, writing essays, poems, speeches, and letters—even creating fear among content writers of losing their jobs. However, the full scope of these tools is yet to be determined as there are risks associated with this technology that need to be addressed.

GPT tools have been in the market before and are used for various use cases. These models have gone through a series of improvements over time. Figure 2 presents the timeline of OpenAI's GPT models.

Timeline of Open AI's Generative Models
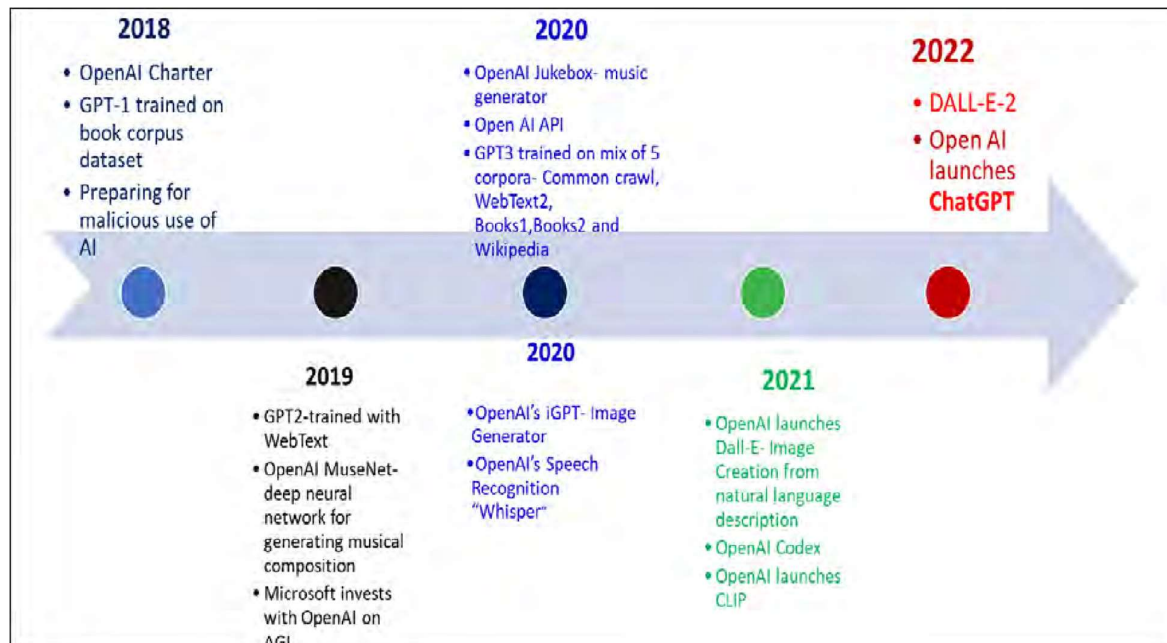
ChatGPT has a broad range of applications like expert conversational agents, language translation and text summarisation, to state a few. It can also learn and adapt to new contexts and situations by analysing text and updating its algorithm based on new data. This continuous analysis makes it more accurate in generating responses. It is based on reinforcement learning with human feedback (Figure 3).[3] The model is trained using supervised fine-tuning with human AI trainers providing

conversations. The reward model works on comparison data built from the conversations of AI trainers with the chatbot and the ranking of the sampled alternative messages. The model has been fine-tuned by using Proximal Policy Optimisation. ChatGPT is the fine-tuned model of the GPT-3.5 series that completed its training in 2022. Both these models have been trained on Azure AI supercomputing Infrastructure.[4]

Basic Architecture of ChatGPT Reinforcement Learning Process
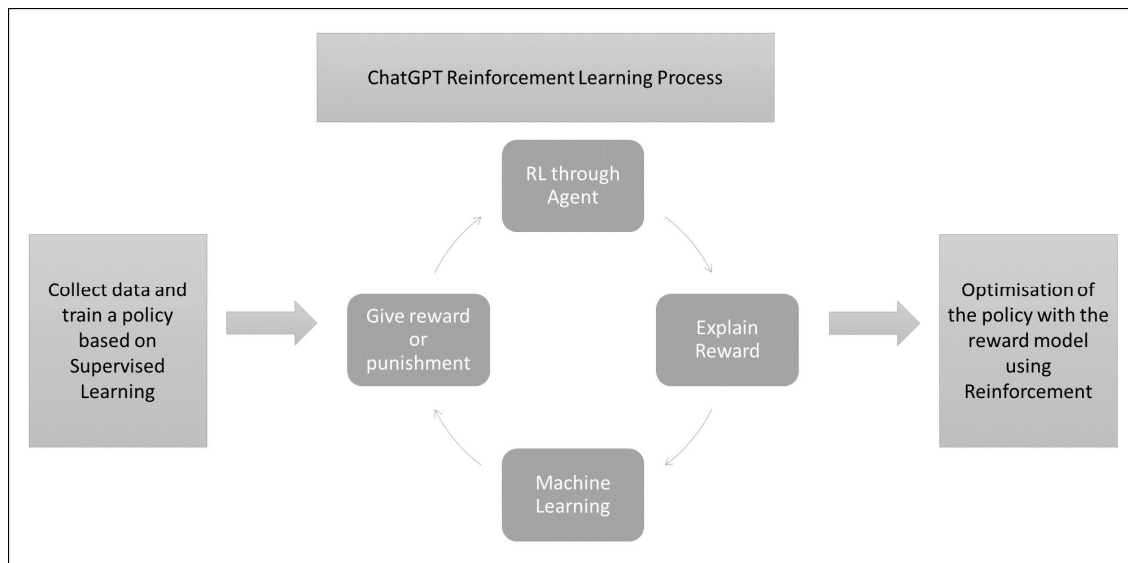
One of the key benefits of ChatGPT is its power to process and learn from interactions with users, understanding the context and nuances of the language and coming out with meaningful and accurate responses. It can constantly improve itself through conversations and building its extensive database. Therefore, one can expect more remarkable capabilities from this model in the future. Furthermore, it is modelled on deep learning architecture, which allows it to achieve a higher level of accuracy in content creation.

The training data of ChatGPT is collected from vast data sources like web pages, books, scientific articles, and corpora of text from other sources. The model is trained on 570 GB of data, about 300 billion words.[6] The cut-off for this data collection was in 2021,[7] and specific training data was used, which might impact the model's performance in terms of generating relevant responses that are contextually

appropriate. This implies that the model lacks real-time data and analysis and information post-2021. In addition, behind this colossal dataset and training, there are some issues that ChatGPT still needs to address, which includes an improved response mechanism in terms of additional layers in the model for verification and validation to present more meaningful information.

Implications of Generative AI Models

The challenge with Generative AI models is to ensure that the generated data is of good quality, balanced, free from potential biases and a good representative of the original data. These models present a risk of overfitting and generation of unrealistic data, which raises ethical concerns related to using such models. Last year, Google's chatbot LaMDA was claimed as sentient by their engineers,[17] and OpenAI's DALLE- 2 talking gibberish was said to have created its own language.

Another issue with these Generative AI systems is that cybercriminals have started to use these tools to develop malicious codes and tools. According to Check Point Research (CPR), major underground hacking communities are already using OpenAI to create spear-phishing emails, infostealers, encryption tools, and other fraud activities. The dark web is being used by the hackers for posting the benefits of malware and for sharing code (like for generating stealers) with the help of tools like ChatGPT.[18]

One of the negative use cases of Generative AI is spreading disinformation, shaping public perception and influence operations. These language models have the capability to automate the creation of misleading text, audio and videos to spread propaganda by various malicious actors. A report by CSET and OpenAI discusses the three dimensions (actors, behaviours and content) where language models and Generative AI can be used for targeted influence operations.[19] Considering the pace of development in this field, these models are likely to become more usable, efficient

and cost-effective with time, making it easier for the threat actors to use them for malicious activities.

Currently, organisations and governments/countries are attempting to address these concerns through practices like responsible data collection, ethical principles for AI, and algorithmic transparency. Moreover, the legal implications of using AI models are also under consideration, specifically on regulations and guidelines for various sectors like healthcare, finance, and defence—where data privacy, security, and regulation on the use of AI for decision-making are of utmost importance.

At present, there are no regulations that apply to LLMs or AI language models in particular. However, there is a need to spread awareness amongst various stakeholders and civil society to consider the ethical and legal implications of these technologies and ensure that appropriate frameworks are implemented for its responsible use. Countries are striving towards establishing AI strategies and data protection laws, focusing on establishing regulations on AI

governance and its ethical use. A few such initiatives by OECD are its ethical principles on AI and support to other countries and organisations in establishing AI principles and best practices.[20] The European Union's (EU) data protection law is another act that closely observes the privacy issues related to data and algorithms.[21] EU is also moving fast with its draft of the AI act that intends to govern all AI use cases.[22] The US is also working towards AI governance and has come out with various policies and principles like the 2023 US National Defense Authorization Act (NDAA), which has proposed provisions for governing and deploying AI capabilities. Sections 7224 and 7226 relate to principles and policies for the use of AI in government and rapid deployment and scaling of applied AI for modernisation activities with use cases.[23] The US National Institute of Standards and Technology (NIST) has also issued Version 1.0 of its AI Risk Management Framework (AIRMF 1.0), which is a multi-tool for organisations to design and deploy trustworthy AI.[24] Recently, China has come out with a series of regulations specific to different types of algorithms and AI capabilities, including relating to AI algorithms for Deepfakes.[25]

Topic 2: Overview of Large Language Models (LLMs)

Aim:

• To provide a foundational understanding of LLMs, including their structure, function, and practical applications.

Procedure:

What are LLMs?

Large language models (LLMs) are a category of foundation models trained on immense amounts of data making them capable of understanding and generating natural language and other types of content to perform a wide range of tasks.

LLMs have become a household name thanks to the role they have played in bringing generative AI to the forefront of the public interest, as well as the point on which organizations are focusing to adopt artificial intelligence across numerous business functions and use cases.

Outside of the enterprise context, it may seem like LLMs have arrived out of the blue along with new developments in generative AI. However, many companies, including IBM, have spent years implementing LLMs at different levels to enhance their natural language understanding (NLU) and natural language processing (NLP) capabilities. This has occurred alongside advances in machine learning, machine learning models, algorithms, neural networks and the transformer models that provide the architecture for these AI systems.

LLMs are a class of foundation models, which are trained on enormous amounts of data to provide the foundational capabilities needed to drive multiple use cases and applications, as well as resolve a multitude of tasks. This is in stark contrast to the idea of building and training domain specific models for each of these use cases individually, which is prohibitive under many criteria (most importantly cost and infrastructure), stifles synergies and can even lead to inferior performance.

LLMs represent a significant breakthrough in NLP and artificial intelligence, and are easily accessible to the public through interfaces like Open AI's Chat GPT-3 and GPT-4, which have garnered the support of Microsoft. Other examples include Meta's Llama models and Google's bidirectional encoder representations from transformers (BERT/RoBERTa) and PaLM models. IBM has also recently launched its Granite model series on watsonx.ai, which has become the generative AI backbone for other IBM products like watsonx Assistant and watsonx Orchestrate.

In a nutshell, LLMs are designed to understand and generate text like a human, in addition to other forms of content, based on the vast amount of data used to train them. They have the ability to infer from context, generate coherent and contextually relevant responses, translate to languages other than English, summarize text, answer questions (general conversation and FAQs) and even assist in creative writing or code generation tasks.

They are able to do this thanks to billions of parameters that enable them to capture intricate patterns in language and perform a wide array of language-related tasks. LLMs are revolutionizing applications in various fields, from chatbots and virtual assistants to content generation, research assistance and language translation.

As they continue to evolve and improve, LLMs are poised to reshape the way we interact with technology and access information, making them a pivotal part of the modern digital landscape.

Ebook

Generative AI + ML for the enterprise

Learn how organizations can confidently incorporate generative AI and machine learning into their business to gain a significant competitive advantage.

Related content

Register for the ebook on AI data stores

How large language models work

LLMs operate by leveraging deep learning techniques and vast amounts of textual data. These models are typically based on a transformer architecture, like the generative pre-trained transformer, which excels at handling sequential data like text input. LLMs consist of multiple layers of neural networks, each with parameters that can be fine-tuned during training, which are enhanced further by a numerous layer known as the attention mechanism, which dials in on specific parts of data sets.

During the training process, these models learn to predict the next word in a sentence based on the context provided by the preceding words. The model does this through attributing a probability score to the recurrence of words that have been tokenized— broken down into smaller sequences of characters. These tokens are then transformed into embeddings, which are numeric representations of this context.

To ensure accuracy, this process involves training the LLM on a massive corpora of text (in the billions of pages), allowing it to learn grammar, semantics and conceptual relationships through zero-shot and self-supervised learning. Once trained on this training data, LLMs can generate text by autonomously predicting the next word based on the input they receive, and drawing on the patterns and knowledge they've acquired. The result is coherent and contextually relevant

language generation that can be harnessed for a wide range of NLU and content generation tasks.

Model performance can also be increased through prompt engineering, prompt-tuning, fine-tuning and other tactics like reinforcement learning with human feedback (RLHF) to remove the biases, hateful speech and factually incorrect answers known as "hallucinations" that are often unwanted byproducts of training on so much unstructured data. This is one of the most important aspects of ensuring enterprise-grade LLMs are ready for use and do not expose organizations to unwanted liability, or cause damage to their reputation.

LLM use cases

LLMs are redefining an increasing number of business processes and have proven their versatility across a myriad of use cases and tasks in various industries. They augment conversational AI in chatbots and virtual assistants (like IBM watsonx Assistant and Google's BARD) to enhance the interactions that underpin excellence in customer care, providing context-aware responses that mimic interactions with human agents.

LLMs also excel in content generation, automating content creation for blog articles, marketing or sales materials and other writing tasks. In research and academia, they aid in summarizing and extracting information from vast datasets, accelerating knowledge discovery. LLMs also play a vital role in language translation, breaking down language barriers by providing accurate and contextually relevant translations. They can even be used to write code, or "translate" between programming languages.Moreover, they contribute to accessibility by assisting individuals with disabilities, including text-to-speech applications and generating content in accessible formats. From healthcare to finance, LLMs are transforming industries by streamlining processes, improving customer experiences and enabling more efficient and data-driven decision making. Most excitingly, all of these capabilities are easy to access, in some cases literally an API integration away. Here is a list of some of the most important areas where LLMs benefit organizations:

Text generation: language generation abilities, such as writing emails, blog posts or other mid-to-long form content in response to prompts that can be refined and polished. An excellent example is retrieval-augmented generation (RAG).

Content summarization: summarize long articles, news stories, research reports, corporate documentation and even customer history into thorough texts tailored in length to the output format.

AI assistants: chatbots that answer customer queries, perform backend tasks and provide detailed information in natural language as a part of an integrated, self-serve customer care solution.

Code generation: assists developers in building applications, finding errors in code and uncovering security issues in multiple programming languages, even "translating" between them.

Sentiment analysis: analyze text to determine the customer's tone in order understand customer feedback at scale and aid in brand reputation management.

Language translation: provides wider coverage to organizations across languages and geographies with fluent translations and multilingual capabilities.

LLMs stand to impact every industry, from finance to insurance, human resources to healthcare and beyond, by automating customer self-service, accelerating response times on an increasing number of tasks as well as providing greater accuracy, enhanced routing and intelligent context gathering.

Race to Build Generative AI and LLMs

The overwhelming response to models like ChatGPT, LaMDA and DALL-E-2 has stirred the industry and started a race amongst the tech giants to build such models as a significant part of the search engine business.

Google's LaMDA was developed in 2020 and is based on Transformer, a neural network architecture[8] that gained popularity in 2022 when an engineer from Google went public and termed it a sentient system. The much-hyped generative AI Chatbot is said to have been considered more capable than ChatGPT, but until it is publicly released, it is difficult to prove the same. On 6 February, Google announced another AI chatbot 'Bard', a conversational AI as a rival to OpenAI's ChatGPT.[9] It is said to be capable of responding to human queries and synthesising information like ChatGPT and is a lightweight version of Google's LaMDA. However, within days of the launch, the flaw in Bard was noticed where the tool made a factual error in one of its promotional videos. Following this, Google's share dropped by 9 per cent and the company lost around US$ 100 billion in market value. Google's Vice President of Search Prabhakar Raghavan asked the trainers and executives to rewrite Bard's incorrect responses. Google is also investing US$ 300 million in Anthropic, an AI startup to work in the field of Generative AI.[10] Some other generative AI models by Google are MUM, PaLM and MusicLM.

Microsoft is also said to be investing billions of dollars in AI and revamping its search

engine Bing and Edge web browser with AI capabilities.[11] It is working in collaboration with OpenAI and is looking at integrating ChatGPT into Bing and further commercialise its Azure OpenAI service with several AI models like GPT3.5,Codex and DALL-E and the soon to be released GPT4.[12] On 7 February, Microsoft launched the AI-powered Bing search engine and Edge browser for preview as an AI co-pilot for the web to get more people to benefit from search and web. The users asked questions to Bing, and it gave direct answers in chat and not with links to websites. The users with access to this feature were curious to have prolonged interactions with the search engine, which then got deranged and started expressing emotions of love and anger.

Following this, Microsoft put a cap of five questions per session and 50 questions per day, as it was observed that only 1 per cent of users have more than 50 questions in a day.[13] The company stated that the tool needed training to be more reliable. In future, it will introduce a toggle mode allowing users to select the level of creativity they wish to have in their responses.[14] In the past, DALL-E2, a text-to-image generator, faced a similar glitch where the tool was said to create its own language and struggled to generate coherent images of text.

The big tech companies investing in Generative AI tools indicate the promise these tools present and the profound benefits users experience when they come across meaningful writings and content that seems to incur human annotation. These tools will bring ease in doing business with multiple use cases in various sectors like devising personalised marketing, social media and sales content; code generation, documentation and content creation in IT; pulling out data, summarising and drafting of legal documents; enabling R&D in drug discovery; providing self-serve functions in Human Resources (HR) and assisting in content creation for questionnaires and interviews; employee optimisation through automated responses, text translation, crafting presentations and synthesising information from video meetings; and creating assistants for specific businesses.[15]

In the future, these tools are expected to generate their own data by bootstrapping their own intelligence and fine-tuning it for better performance. All these tools are based on an autoregressive transformer model and are dense, which means that they use all the parameters (millions/billions) to produce a response. The research in this

aspect is now moving towards designing models that will only use the relevant parameters to generate a response, making them less computationally difficult.[16]

The race among the tech giants to come out with these tools is like the innovators' dilemma to rule the search engine business. The reasons behind this hurry to come out with such tools could be either to take the lead in this business and vision for the future or to collect more data from human users and keep training their models to perform better. Nevertheless, adopting these tools will be part of the businesses soon, but criticism over their shortcomings will also follow.

Conclusion

Generative AI systems have the potential to revolutionise the way we work and live. Its capability to cater to diverse audiences with meaningful information in a contextualised manner and provide tailor-made responses has brought a significant breakthrough in technology and how we use it. As these tech companies dive into the foray of these AI applications and use cases, it is imperative to study the implications of this technology and how it affects society at large. The regulation of AI systems is still in its infancy, and countries looking at building their own policies and regulations can learn from the positives and negatives of the two different models being implemented by the EU and China.

The next wave of innovation in Generative AI and LLMs will bring new use cases and applications in other domains with better reliability mechanisms. These AI tools certainly have limitless potential, but at the same time, they should not be totally relied upon as a replacement for human decision-making as they lack emotional intelligence and human intuition and struggle with language nuances and context, with the risk of biases being introduced at any point in their structural mechanisms. There is no silver bullet solution with Generative AI systems, and hence coordination among stakeholders, civil society, government and other institutions is needed to manage and control the risks associated with this technology.

References:

https://en.wikipedia.org/wiki/Generative_artificial_intelligence
https://research.ibm.com/blog/what-is-generative-AI
https://www.cloudskillsboost.google/course_templates/536
https://www.elastic.co/what-is/large-language-models
https://www.geeksforgeeks.org/large-language-model-llm/
https://www.aiforeducation.io/ai-resources/generative-ai-explainer
https://en.wikipedia.org/wiki/Large_language_model