[Week 1, Part 1] Applied LLM Foundations and Real World Use Cases

[Jan 15 2024] You can register here to receive course content and other resources

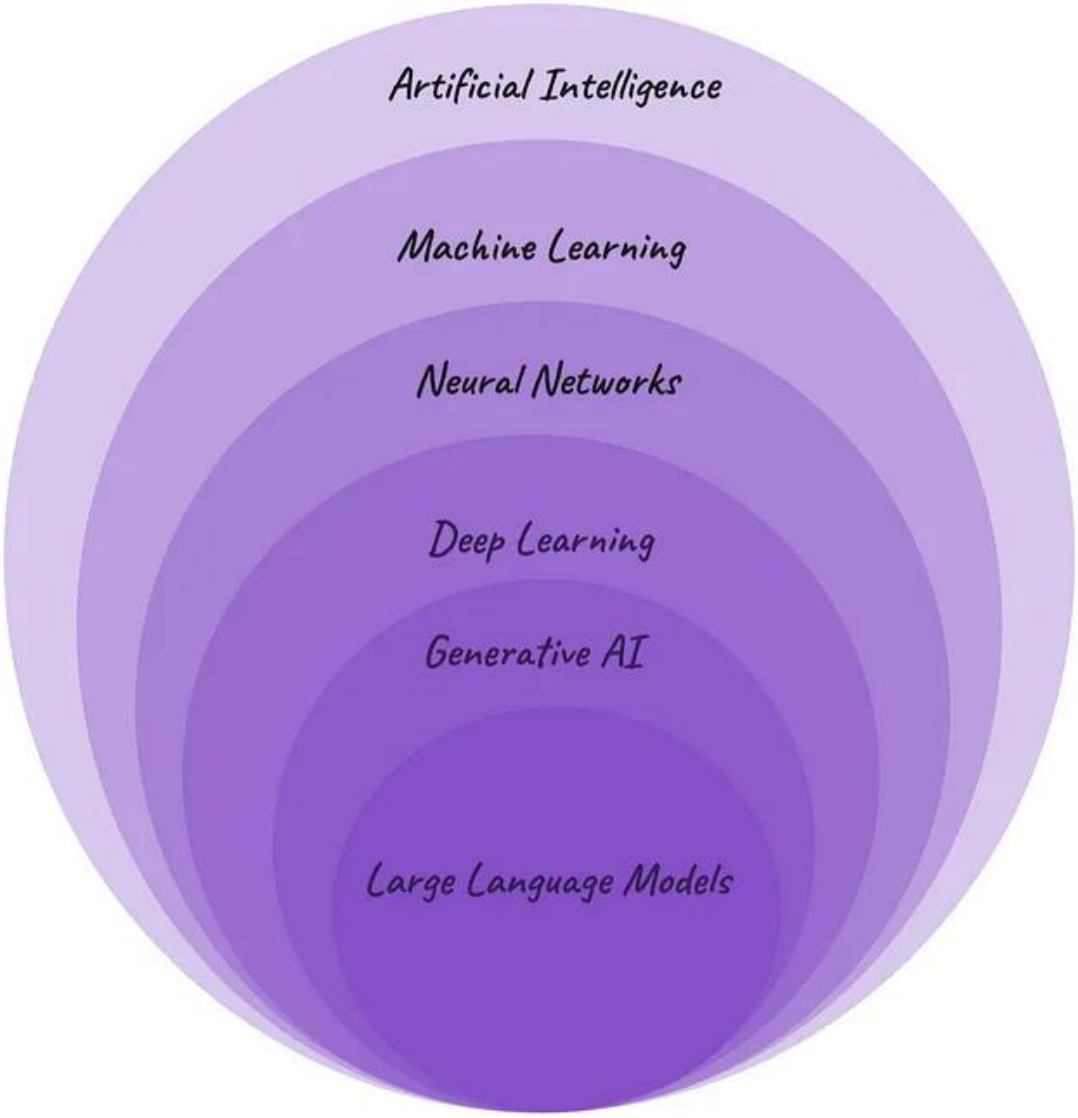
ETMI5: Explain to Me in 5

In this part of the course, we delve into the intricacies of Large Language Models (LLMs). We start off by exploring the historical context and fundamental concepts of artificial intelligence (AI), machine learning (ML), neural networks (NNs), and generative AI (GenAI). We then examine the core attributes of LLMs, focusing on their scale, extensive training on diverse datasets, and the role of model parameters. Then we go over the types of challenges associated with using LLMs.

In the next section, we explore practical applications of LLMs across various domains, emphasizing their versatility in areas like content generation, language translation, text summarization, question answering etc. The section concludes with an analysis of the challenges encountered in deploying LLMs, covering essential aspects such as scalability, latency, monitoring etc.

In summary, this part of the course provides a practical and informative exploration of Large Language Models, offering insights into their evolution, functionality, applications, challenges, and real-world impact.

History and Background



history

Image Source: https://medium.com/womenintechnology/ai-c3412c5aa0ac

The terms mentioned in the image above have likely come up in conversations about ChatGPT. The visual representation offers a broad overview of how they fit into a

hierarchy. AI is a comprehensive domain, where LLMs constitute a specific subdomain, and ChatGPT exemplifies an LLM in this context.

In summary, **Artificial Intelligence (AI)** is a branch of computer science that involves creating machines with human-like thinking and behavior. **Machine Learning(ML)**, a subfield of AI, allows computers to learn patterns from data and make predictions without explicit programming. **Neural Networks (NNs)**, a subset of ML, mimic the human brain's structure and are crucial in deep learning algorithms. Deep Learning (DL), a subset of NN, is effective for complex problem-solving, as seen in image recognition and language translation technologies. **Generative AI (GenAI)**, a subset of DL, can create diverse content based on learned patterns. **Large Language Models (LLMs)**, a form of GenAI, specialize in generating human-like text by learning from extensive textual data.

Generative AI and Large Language Models (LLMs) have revolutionized the field of artificial intelligence, allowing machines to create diverse content such as text, images, music, audio, and videos. Unlike discriminative models that classify, generative AI models generate new content by learning patterns and relationships from human-created datasets.

At the core of generative AI are foundation models which essentially refer to large AI models capable of multi-tasking, performing tasks like summarization, Q&A, and classification out-of-the-box. These models, like the popular one that everyone's heard of-ChatGPT, can adapt to specific use cases with minimal training and generate content with minimal example data.

The training of generative AI often involves supervised learning, where the model is provided with human-created content and corresponding labels. By learning from this data, the model becomes proficient in generating content similar to the training set.

Generative AI is not a new concept. One notable example of early generative AI is the Markov chain, a statistical model introduced by Russian mathematician Andrey Markov in 1906. Markov models were initially used for tasks like next-word prediction, but their simplicity limited their ability to generate plausible text.

The landscape has significantly changed over the years with the advent of more powerful architectures and larger datasets. In 2014, generative adversarial networks (GANs) emerged, using two models working together—one generating output and the other discriminating real data from the generated output. This approach, exemplified by models like StyleGAN, significantly improved the realism of generated content.

A year later, diffusion models were introduced, refining their output iteratively to generate new data samples resembling the training dataset. This innovation, as seen in Stable Diffusion, contributed to the creation of realistic-looking images.

In 2017, Google introduced the transformer architecture, a breakthrough in natural language processing. Transformers encode each word as a token, generating an attention map that captures relationships between tokens. This attention to context enhances the model's ability to generate coherent text, exemplified by large language models like ChatGPT.

The generative AI boom owes its momentum not only to larger datasets but also to diverse research advances. These approaches, including GANs, diffusion models, and transformers, showcase the breadth of methods contributing to the exciting field of generative AI.

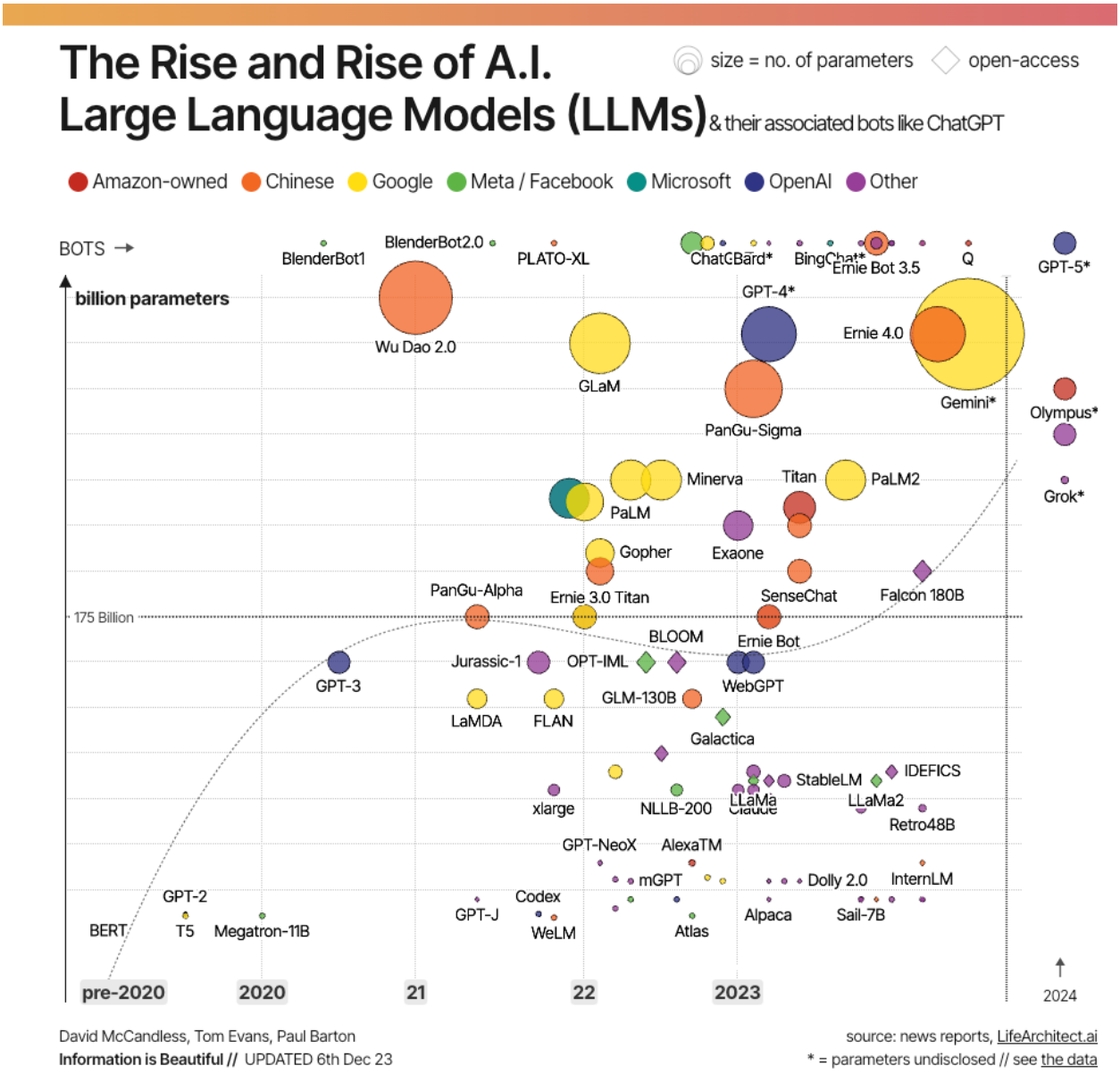
Enter LLMs

The term “Large” in Large Language Models (LLMs) refers to the sheer scale of these models—both in terms of the size of their architecture and the vast amount of data they are trained on. The size matters because it allows them to capture more complex patterns and relationships within language. Popular LLMs like GPT-3, Gemini, Claude etc. have thousands of billion model parameters. In the context of machine learning, model parameters are like the knobs and switches that the algorithm tunes during training to make accurate predictions or generate meaningful outputs.

Now, let’s break down what “Language Models” mean in this context. Language models are essentially algorithms or systems that are trained to understand and generate human-like text. They serve as a representation of how language works, learning from diverse datasets to predict what words or sequences of words are likely to come next in a given context.

The “Large” aspect amplifies their capabilities. Traditional language models, especially those from the past, were smaller in scale and couldn’t capture the intricacies of language as effectively. With advancements in technology and the availability of massive computing power, we’ve been able to build much larger models. These Large Language Models, like ChatGPT, have billions of parameters, which are essentially the variables the model uses to make sense of language.

Take a look at the infographic from “Information is beautiful” below to see how many parameters recent LLMs have. You can view the live visualization here



llm_sizes.png

Image source: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/

Training LLMs

Training LLMs is a complex process that involves instructing the model to comprehend and produce human-like text. Here's a simplified breakdown of how LLM training works:

1. **Providing Input Text:**

   - LLMs are initially exposed to extensive text data, encompassing various sources such as books, articles, and websites.
   - The model's task during training is to predict the next word or token in a sequence based on the context provided. It learns patterns and relationships within the text data.
2. **Optimizing Model Weights:**
   - The model comprises different weights associated with its parameters, reflecting the significance of various features.
   - Throughout training, these weights are fine-tuned to minimize the error rate. The objective is to enhance the model's accuracy in predicting the next word.
3. **Fine-tuning Parameter Values:**
   - LLMs continuously adjust parameter values based on error feedback received during predictions.
   - The model refines its grasp of language by iteratively adjusting parameters, improving accuracy in predicting subsequent tokens.

The training process may vary depending on the specific type of LLM being developed, such as those optimized for continuous text or dialogue.

LLM performance is heavily influenced by two key factors:

- **Model Architecture:** The design and intricacy of the LLM architecture impact its ability to capture language nuances.
- **Dataset:** The quality and diversity of the dataset utilized for training are crucial in shaping the model's language understanding.

Training a private LLM demands substantial computational resources and expertise. The duration of the process can range from several days to weeks, contingent on the model's complexity and dataset size. Commonly, cloud-based solutions and high-performance GPUs are employed to expedite the training process, making it more efficient. Overall, LLM training is a meticulous and resource-intensive undertaking that lays the groundwork for the model's language comprehension and generation capabilities.

After the initial training, LLMs can be easily customized for various tasks using relatively small sets of supervised data, a procedure referred to as fine-tuning.

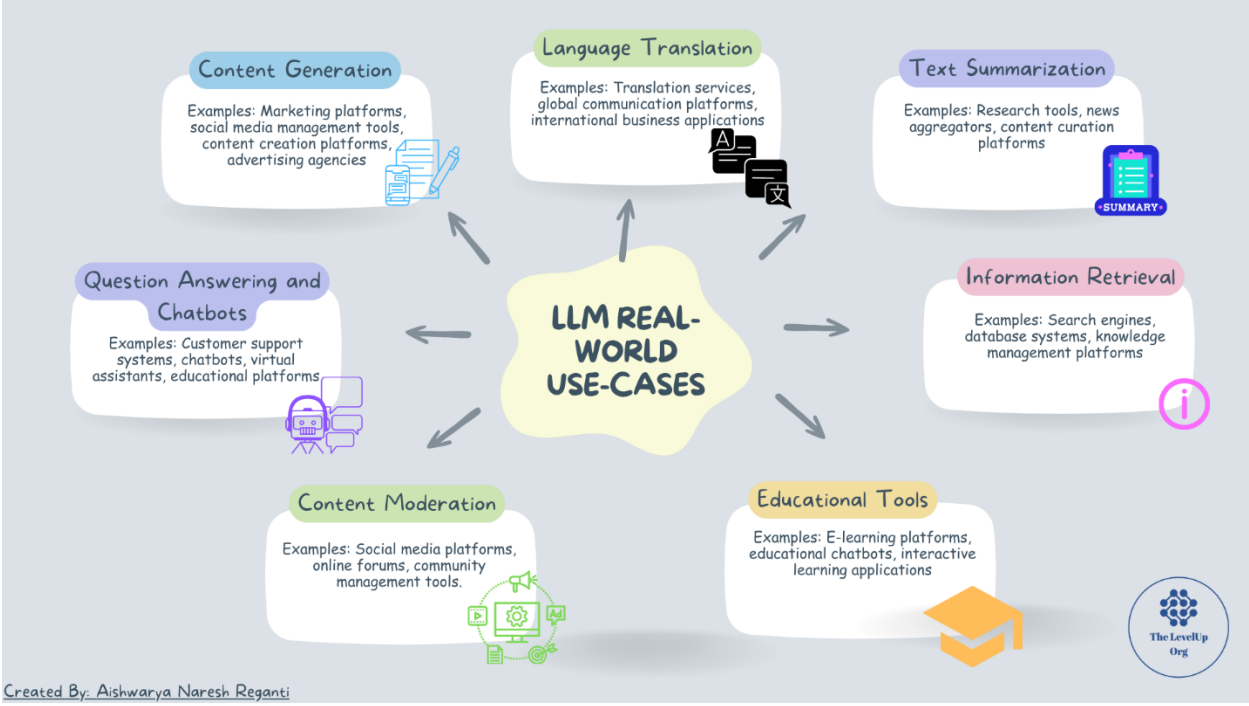
There are three prevalent learning models:

1. **Zero-shot learning:** The base LLMs can handle a wide range of requests without explicit training, often by using prompts, though the accuracy of responses may vary.
2. **Few-shot learning:** By providing a small number of pertinent training examples, the performance of the base model significantly improves in a specific domain.
3. **Domain Adaptation:** This extends from few-shot learning, where practitioners train a base model to adjust its parameters using additional data relevant to the particular application or domain.

We will be diving deep into each of these methods during the course.

LLM Real World Use Cases

LLMs are already being leveraged in various applications showcasing their versatility and power of these models in transforming several domains. Here's how LLMs can be applied to specific cases:



Blue and Grey Illustrative Creative Mind Map.png

1. **Content Generation:**
   - LLMs excel in content generation by understanding context and generating coherent and contextually relevant text. They can be employed to automatically generate creative content for marketing, social media posts, and other communication materials, ensuring a high level of quality and relevance.
   - **Real World Applications:** Marketing platforms, social media management tools, content creation platforms, advertising agencies
2. **Language Translation:**
   - LLMs can significantly improve language translation tasks by understanding the nuances of different languages. They can provide accurate and context-aware translations, making them valuable tools for businesses operating in multilingual environments. This can enhance global communication and outreach.

- **Real World Applications**: Translation services, global communication platforms, international business applications

3. **Text Summarization:**
   - LLMs are adept at summarizing lengthy documents by identifying key information and maintaining the core message. This capability is valuable for content creators, researchers, and businesses looking to quickly extract essential insights from large volumes of text, improving efficiency in information consumption.
   - **Real World Applications**: Research tools, news aggregators, content curation platforms
4. **Question Answering and Chatbots:**
   - LLMs can be employed for question answering tasks, where they comprehend the context of a question and generate relevant and accurate responses. They enable these systems to engage in more natural and context-aware conversations, understanding user queries and providing relevant responses.
   - **Real World Applications***:* Customer support systems, chatbots, virtual assistants, educational platforms
5. **Content Moderation:**
   - LLMs can be utilized for content moderation by analyzing text and identifying potentially inappropriate or harmful content. This helps in maintaining a safe and respectful online environment by automatically flagging or filtering out content that violates guidelines, ensuring user safety.
   - **Real World Applications**: Social media platforms, online forums, community management tools.
6. **Information Retrieval:**
   - LLMs can enhance information retrieval systems by understanding user queries and retrieving relevant information from large datasets. This is particularly useful in search engines, databases, and knowledge management systems, where LLMs can improve the accuracy of search results.
   - **Real World Applications**: Search engines, database systems, knowledge management platforms
7. **Educational Tools:**
   - LLMs contribute to educational tools by providing natural language interfaces for learning platforms. They can assist students in generating summaries, answering questions, and engaging in interactive learning conversations. This facilitates personalized and efficient learning experiences.
   - **Real World Applications**: E-learning platforms, educational chatbots, interactive learning applications

Summary of popular LLM use-cases

No.	Use case	Description

No.	Use case	Description
1	Content Generation	Craft human-like text, videos, code and images when provided with instructions
2	Language Translation	Translate languages from one to another
3	Text Summarization	Summarize lengthy texts, simplifying comprehension by highlighting key points.
4	Question Answering and Chatbots	LLMs can provide relevant answers to queries, leveraging their vast knowledge
5	Content Moderation	Assist in content moderation by identifying and filtering inappropriate or harmful language
6	Information Retrieval	Retrieve relevant information from large datasets or documents.
7	Educational Tools	Tutor, provide explanations, and generate learning materials.

Understanding the utilization of generative AI models, especially LLMs, can also be gleaned from the extensive array of startups operating in this domain. An infographic presented by Sequoia Capital highlighted these companies across diverse sectors, illustrating the versatile applications and the significant presence of numerous players in the generative AI space.

The Generative AI Market Map v3

A work in progress

business_cases.png

Image Source:
https://markovate.com/blog/applications-and-use-cases-of-llm/

LLM Challenges

llm_challenges.png

Although LLMs have undoubtedly revolutionized various applications, numerous challenges persist. These challenges are categorized into different themes:

- **Data Challenges:** This pertains to the data used for training and how the model addresses gaps or missing data.
- **Ethical Challenges:** This involves addressing issues such as mitigating biases, ensuring privacy, and preventing the generation of harmful content in the deployment of LLMs.
- **Technical Challenges:** These challenges focus on the practical implementation of LLMs.
- **Deployment Challenges:** Concerned with the specific processes involved in transitioning fully-functional LLMs into real-world use-cases (productionization)

Data Challenges:

1. **Data Bias:** The presence of prejudices and imbalances in the training data leading to biased model outputs.

2. **Limited World Knowledge and Hallucination:** LLMs may lack comprehensive understanding of real-world events and information and tend to hallucinate information. Note that training them on new data is a long and expensive process.
3. **Dependency on Training Data Quality:** LLM performance is heavily influenced by the quality and representativeness of the training data.

Ethical and Social Challenges:

1. **Ethical Concerns:** Concerns regarding the responsible and ethical use of language models, especially in sensitive contexts.
2. **Bias Amplification:** Biases present in the training data may be exacerbated, resulting in unfair or discriminatory outputs.
3. **Legal and Copyright Issues:** Potential legal complications arising from generated content that infringes copyrights or violates laws.
4. **User Privacy Concerns:** Risks associated with generating text based on user inputs, especially when dealing with private or sensitive information.

Technical Challenges:

1. **Computational Resources:** Significant computing power required for training and deploying large language models.
2. **Interpretability:** Challenges in understanding and explaining the decision-making process of complex models.
3. **Evaluation**: Evaluation presents a notable challenge as assessing models across diverse tasks and domains is inadequately designed, particularly due to the challenges posed by freely generated content.
4. **Fine-tuning Challenges:** Difficulties in adapting pre-trained models to specific tasks or domains.
5. **Contextual Understanding:** LLMs may face challenges in maintaining coherent context over longer passages or conversations.
6. **Robustness to Adversarial Attacks:** Vulnerability to intentional manipulations of input data leading to incorrect outputs.
7. **Long-Term Context:** Struggles in maintaining context and coherence over extended pieces of text or discussions.

Deployment Challenges:

1. **Scalability:** Ensuring that the model can scale efficiently to handle increased workloads and demand in production environments.
2. **Latency:** Minimizing the response time or latency of the model to provide quick and efficient interactions, especially in real-time applications.
3. **Monitoring and Maintenance:** Implementing robust monitoring systems to track model performance, detect issues, and perform regular maintenance to avoid downtime.
4. **Integration with Existing Systems:** Ensuring smooth integration of LLMs with existing software, databases, and infrastructure within an organization.

5. **Cost Management:** Optimizing the cost of deploying and maintaining large language models, as they can be resource-intensive in terms of both computation and storage.
6. **Security Concerns:** Addressing potential security vulnerabilities and risks associated with deploying language models in production, including safeguarding against malicious attacks.
7. **Interoperability:** Ensuring compatibility with other tools, frameworks, or systems that may be part of the overall production pipeline.
8. **User Feedback Incorporation:** Developing mechanisms to incorporate user feedback to continuously improve and update the model in a production environment.
9. **Regulatory Compliance:** Adhering to regulatory requirements and compliance standards, especially in industries with strict data protection and privacy regulations.
10. **Dynamic Content Handling:** Managing the generation of text in dynamic environments where content and user interactions change frequently.

Read/Watch These Resources (Optional)

1. https://www.nvidia.com/en-us/glossary/generative-ai/
2. https://markovate.com/blog/applications-and-use-cases-of-llm/
3. https://www.sequoiacap.com/article/generative-ai-act-two/
4. https://datasciencedojo.com/blog/challenges-of-large-language-models/
5. https://snorkel.ai/enterprise-llm-challenges-and-how-to-overcome-them/
6. https://www.youtube.com/watch?v=MyFrMFab6bo
7. https://www.youtube.com/watch?v=cEyHsMzbZBs

Read These Papers (Optional)

1. https://dl.acm.org/doi/abs/10.1145/3605943
2. https://www.sciencedirect.com/science/article/pii/S2950162823000176
3. https://arxiv.org/pdf/2303.13379.pdf
4. https://proceedings.mlr.press/v202/kandpal23a/kandpal23a.pdf
5. https://link.springer.com/article/10.1007/s12599-023-00795-x