International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 11 Issue: 06 | Jun 2024 www.irjet.net p-ISSN: 2395-0072

A Survey on Large Language Models: Overview and Applications

Mangi Nikhil[1], Yeshwanth Reddy Vallela[2], Gajji Chakrapani[3]

[1] *Student, Department of Computer Science and Engineering, Sri Indu College of Engineering and Technology, Telangana, India.*
[2] *Student, Department of Cyber Security, Sri Indu College of Engineering and Technology, Telangana, India.*
[3] *Student, Department of Artificial Intelligence and Machine Learning, Sri Indu College of Engineering and Technology, Telangana, India.*

Abstract— Large Language Models (LLMs) are a breakthrough in natural language processing that have revolutionized how computers understand and generate human-like language. This paper provides a comprehensive introduction to LLMs and generative AI. It covers the history and evolution of language models, the transformer architecture that enabled powerful LLMs like GPT and BERT, and the key applications of LLMs across diverse domains. The paper discusses the design cycle for building domain-specific LLM models, leveraging open-source options like Llama 2. With insightful literature review and technical details, this study serves as a beginner's guide for understanding LLMs' potential, capabilities, and the process of harnessing their power for specialized use cases. As generative AI reshapes industries, this paper equips readers to embrace the transformative impact of LLMs responsibly and effectively.

Keywords—Large Language Models, Recurrent Neural Networks, Chat GPT, Generative AI, etc.,

1. INTRODUCTION

Natural language processing is a field of computer science where we build algorithms and models that make computers understand human language. LLM (Large Language Models) are also known as transformative or next-generative language models. These models brought so many changes in natural language processing that help in understanding complex patterns, structures in the languages and identify the semantic relations between words in a sentence. LLMs are trained on large data sets and use transformer architecture which uses deep learning techniques when compared to other NLP techniques. These LLMs not only understand text, it can also process images, videos, and audio files. They can also be used for sentiment analysis. This has the upper hand in understanding or mimicking human-understandable language.

In the early stages of language modeling statistical methods were adopted where we predicted the upcoming words, few such models are N-grams, Hidden Markov Models. Based on the training data they observe the previous words of the sentence and suggest new words. Then after a few proposed machine learning approaches. We need the machine to understand the context behind it. This improved language understanding; those models were trained on capturing relations between the large text corpora. After that Deep learning techniques revolutionized the NLP language modeling using RNNs (recurrent neural networks) and LSTM (long short-term memory) here in recurrent neural networks, they handle sequential data and take information from previous inputs by maintaining a hidden state. At each stage, an RNN processes current input, updates its hidden state, and repeats. We also have LSTM, which addresses the vanishing gradient problem that occurs in training neural networks. LSTMs store long-term sequences, which helps in capturing dependencies b/w inputs. When we apply RNNs and LSTM to language modeling, we use a model called sequence-to-sequence modeling (encoder-decoder model). encoder process input encodes into a fixed-length representation (context vector). The decoder takes the context vector and generates the output. Here, hidden states are initialized by the context vector generated by the encoder.

In 2017, a research study proposed the transformer architecture [1] "Attention is All You Need" by Vaswani et al. This changed everything in language modeling. This is one of the greatest advancements that laid the foundation for many models like ChatGPT, BERT, and many more. These models do not depend on recurrent connections; instead, they use attention mechanisms. It allows parallel processing, efficient usage of hardware, and also helps to understand long-range dependencies. Transformers work this way; they transform one sequence to another. In a sentence based on the situation, sometimes the context behind it may vary. Things like word order and turns of phrases mix things up as transformers work on sequence-to-sequence learning; they take a sequence of tokens and process the input to generate an output sequence. It consists of an encoder and a decoder. An encoder processes the input sequence, whereas a decoder processes the output sequence. The encoder generates the encodings that define which parts of the input sequence are relevant to each other and passes these to the next encoder layer. The decoder uses all these encodings and uses their derived context to generate the output sequence.

Models like Chatgpt and Bert emerged after the transformer architecture came to light. OpenAI's GPT (Generative Pre-Trained Transformers) and Google's BERT

(Bidirectional Encoder Representations from Transformers) these models were trained on vast internet text data as transformers allow parallelization, they can generate relevant text to the query or task. these models revolutionized applications like language translations, text generation, text summarization, chatbots and gave great results.

Llama (Large Language Model Meta AI), Falcon, and ChatGPT are a few of the most popular and successful models which are trained on large text corpora and exposed to internet data those models learn from the data, they find contextual meaning and semantic relations between the tokens, and learn factual knowledge, reasoning, and understanding capabilities are enhanced all this process will make it understand language in a better way. They can be fine-tuned later on a specific task for example ChatGPT is fine-tuned in a conversational context like a chatbot or a virtual assistant these models state the art of NLP they imitate human-like conversation and understand human queries too.

LLMs advanced gradually; they are not what they were a few years ago, thanks to the technological advancements that have taken place. Significantly after the release of transformer architecture, OpenAI began developing the GPT (Generative Pre-Trained Transformer) model in 2016 and released its initial version, GPT-1 (1 in this represents generation), in 2018, which was trained on approximately 8 million web pages, which is around 117 million parameters and then they released its upgraded model which is a powerful version in 2019 it has 1.7 Billion parameters and was trained on a large dataset around 40 GB of text. Its understanding and generative capabilities are remarkable. In 2020, OpenAI released GPT-3 which was trained on a trillion words from the internet. This version was able to translate between languages generate improvised text and answer questions. Soon this caught the attention of people and then they released ChatGPT in November 2022 due to its exceptional capabilities in language understanding and text generation that became popular in less time. In the meantime, OpenAI trained a model using adversarial training techniques to behave properly since users might try to trick it to behave badly which could cause legal problems. Similarly, BERT has undergone advancements and made some significant changes in NLP. It also trained on a large corpus of data, which enhanced its text generation capabilities.

Moreover, these advancements in LLMs have extended to make them work on specific domains or tasks like conversational models, code generation, website development, research purposes, finance, media, medical, and education fields. Seeing such advancements in this field makes it more valid for us to learn how to make use of these large language models so, why not consider one of the open-source LLM and train it on a specific domain (using PDFs, books, etc. related to that domain) and fine-tune it then we'll host it.

In summary, Large Language Models (LLMs) are a major breakthrough in natural language processing (NLP) that allow computers to understand and generate human-like language. For readers who are not aware of LLMs, this introduction provides an overview of their history, working of LLM in a brief manner, applications, how to create a domain-specific model using an open source LLM and fine-tune it for our specific tasks then hosting it on web, This study works as a beginner's guide for people who wants to get started with LLMs.

2. Literature Survey:

Generative AI (GenAI) [21] is perhaps the most disruptive [22] and generalized technology of this decade [23], already influenced many industries, the advancements in Large Language Models and AI led to a variety of interesting reviews and discussions on the progress of LLMs and artificial intelligence. [2], [48], [3], [4], [5] gives us a comprehensive overview of LLMs and Generative AI, and many other papers talk about its applications in various industries like how the chatgpt can be used in education is discussed in [5], healthcare, medicine and Medical research in [42],[46],[47], protein sequence modeling and generation in [48,6], drug discovery in[16] predictive applications in clinical related activities with LLMs [18] machine learning for cancer biomarkers in [19],integration of biotechnology and AI applications to address global challenges in [20].A survey on generative AI is given in [7], text modeling in [8], and text generation in [9]. how can we use LLM in finance is evaluated in [10], how this generative ai could be helpful in supply chain management in [11], Large Language Models in telecom [12], on code writing capabilities of LLM in [13], deep fakes in [14], law-related applications in [15], how generative AI can be used at work[21], how generative ai can transform Media (reel-framer)in [24], LLMs in visual content marketing in [25]. Generative ai in the construction industry discussed in [26], text to speech synthesis in [27], image synthesis in [28].

Even though there are more and more studies on LLMs, not enough studies have been done on how to use them well and understand all of their technical details. As things move quickly in this field, we thought it would be helpful to write a article that describes what LLMs are, how they work, and what they may be used for. This paper is a solid place to begin to learn about LLMs' potential, capabilities, and how to use them effectively for specific tasks. As generative AI reshapes industries, this study alerts readers to embrace the transformative impact of LLMs responsibly and effectively

3. Generative AI:

Generative AI (GenAI) is one of the most significant and beneficial technologies of the present day. It is used in many fields, from marketing and media to education and game development, and medical to pharmaceuticals. GenAI has completely changed the way content is made by going beyond the limits of standard AI systems and making it possible to make writing, speech, images, and videos, among other types of content. By going beyond what conventional AI systems can do, GenAI has totally changed how content is made. It can now make writing, speech, images, movies, and many other types of content. some classic GenAI models are ChatGPT, Bard, Midjourney, Dall-E, codex, Co-Pilot, Llama, VALL-E (X), Runway, Sora and more. Looking back over the last few years, it's clear that these models have grown very quickly, which has been helped by the large amount of data that is available. These models have been scaled at large scale in the beginning they were trained in few million parameters now they have hundreds of billions of parameters which improves the efficiency of the model. ChatGPT bard like models were trained on billions of parameters and finetuned as a conversational model which can generate human like language in response for the given task (question or basically a prompt to perform a task).

Firstly, to understand Generative AI we need to understand how important data is [45] and generation variance, performance metrics [48]. Basically, Generative AI is based on Data it is the core, large amount of data (vast corpus of data) is collected to train the model, the quantity and quality of data matters a lot it reflects on generating capabilities of the model. Data availability on a large scale will be helpful and specifically labeled datasets would help much for supervised training. Generation of content is dependent on the data that has been trained on. These models will use the data, identify the hidden patterns in the data and understand the context or meaning behind it using which it generates new content. Variance means another factor which describes the quality and differentiates the content generated by the model for the same prompt. High variance GenAI generates diversified content whereas low variance LLMs generate same samples for the given prompt, all these factors are related to each other. there are few performance metrics which evaluates the generated content few of them are visual inspection [29], inception score [30], Frechet Inception Distances (FID) [31], PR-curves[32], Coverage metrics[33].

4. Overview:

Large Language models changed the whole natural language processing thing. It changes how we interact, communicate and process language. These have been undergoing continuous evolution and consistently impress with their capabilities Now we'll discuss about LLM's in brief.

Back in the early times when Alan Turing proposed the turing machine which laid the foundation of AI. After the Dartmouth conference in 1956 AI officially came to birth and then researchers started to work on creating machines that work or process things like a human, basically a machine that can imitate human actions. In the mid 1950's to 60's language models were built but those were purely rule based which makes it more complex as it has limited knowledge or rules in this case which model can work on. After this in 1980's statistical language models were developed they use probabilistic techniques (eg; n-gram language models bigram, trigram) to find out the likelihood of the words that could be used next, but these were also limited they couldn't get the context and can't understand the semantics of the language. Few of these issues are addressed in neural language models in the early 21st century, these models use deep learning techniques. Neural Language models describe the likelihood of word patterns by neural networks like multi-layer perceptions, Recurrent neural networks (RNN). RNNLM was able to take an input and produce more natural output. In 2016 Google introduced this in GNMT which is used to enhance fluency and accuracy in google translator. it uses Long Short Term Memory networks. GNMT achieves competitive results compared to state-of-the-art systems. And there came a major breakthrough in language modeling in 2017 transformer architecture. This model helped to learn long-term context and parallel training is also possible in this model which makes it easier to train the model in large datasets efficiently and fastly.

Transformer architecture played a vital role in advancements of large language models. OpenAI LLC a non- profit organization adopted the transformer architecture with the help of this they released their first model GPT-1 in 2018 this was able to generate content which is relevant to prompt this demonstrated the true potential of the transformers in Natural language processing. In 2020 GPT-3 was released which was trained on large data sets this give outstanding results by generating very much naturally sounding text this became quite popular then thereafter OpenAI made GPT-3.5 public can be used by anyone which caught attention of everyone very fast after the success of this model open ai released its next iteration GPT-4 in the meantime Google introduced Bard and now Gemini, Perplexity, Claude, Meta's llama all these have given outstanding results.

The key factor for these models are as discussed quality and quantity of data. Large datasets which contains diverse data content from web scaping, books, articles, research papers, websites and other textual data. initially data will be collected and then the data processing and transformation takes place duplicates and noisy data will be removed then protects the privacy of people whose

data is used to train large language models. Usually, this means taking steps to make personal information anonymous or de-identifiable so that people can't be identified from the training data. When it comes to following privacy rules and keeping private information safe, these steps are very important. Training is done in 2 ways unsupervised and supervised learning during the training The model learns to capture the statistical patterns and structures present in the text data. As these LLM's uses Transformers which use LSTM networks (Long Short Term Memory) it enables to model long range dependencies using attention mechanism and understand the context and semantics of the given input or prompt to generate new content like a human.

Few popular transformer-based models are Sequence-to-Sequence Transformers these transformers are used in tasks which needs mapping input sequences to the output sequences one such existing model is T5 text-to-text transformer [34] which is introduced by google it can handle tasks like text generation translation, summarization, question answering, and text classification.

Autoregressive transformers are widely used for tasks such as language modeling and generation, function similarly to models like GPT. They predict the next words in a sequence by considering preceding words, generating new content that maintains contextual relevance.

Autoencoder transformers: these models are used in dimensionality reduction and also new feature learning they take inputs and create encodings of fixed size and then it uses those encodings to generate output sequence by using essential information form input data. BERT Bidirectional Encoder Representations from Transformer model is a autoencoder transformer this will be given with a sequence of encodings with masked words in it then it predicts the masked words with the help of preceding words and the whole context of the sentence. this helps in named entity recognition, BERT can identity entities in a sentence and categorize them.

Conditional Transformer Language Model (CTRL): CTRL language model is trained to generate text based on the control codes or prompt [44]. By giving the model clear instructions, these control codes allow users to specify the expected subject, style, and other aspects for the generated text, finetuning is important for CTRL before adapting the model for any specific domain tasks.

In general, LLMs have a wide scope of usage for natural language processing tasks now we'll look into applications of LLMs. Language Translation, Text Generation, Virtual Assistance, Summarization, Question-answering, Text Classification, Information Extraction (IE), Dialog systems, Semantic Search, Speech recognition lets discuss all these in detail.

5. Applications:

Language Translation: language translation is one of the key applications of LLMs, models like gpt, bert are trained in all languages which gives them an ability to translate text from one language to another language fluently with a great accuracy. They excel in translating text between languages by encoding input text in one language and decoding it into another. this functionality can be quite a useful one for a range of users for real time translation eg: google translate which uses PaLM2 for translation. These models are very useful tool for helping people from different cultures and languages communicate more effectively. This feature could make it easier for people around the world to work together and get information, and this definitely is one of the important area of research and development in NLP field.

Question-answering:

QA systems are computational models which allow the input of questions from users and provide corresponding answers as output. Large Language models (LLMs) are trained using a large amount of data and then further refined using labeled data specifically designed for question answering [35]. During the pre-training phase, the model was trained to understand the context of sentences. In the QA-based labeled dataset, the model extracts responses from vast datasets it has been trained on by collecting contextually relevant information and representing it in a meaningful manner. LLM based QA systems has a high accuracy on benchmark datasets. These were also used in search engines to provide a quick and efficient response for a question eg.bard is used in google, copilot is used in bing. Current research is dedicated to enhancing the comprehensibility, and effectiveness of LLM question responding.

Text Generation:

Text generation is one of the important applications that helps in automating content creation for various purposes like articles, stories, research papers, emails, QA systems, social media post or product descriptions, code generation. These models can generate accurate, high quality and rational content.

Text classification:

LLMs provide text classification, analysis, and categorization functionalities based on predetermined labels or subjects. This helps in managing large amount of text data and effective for tasks such as Sentiment Analysis, Document Classification, Spam Filtering, Content Moderation, Customer Support Ticket Routing, Fraud Detection, News Classification this has a wide range of applications. This is useful where organizations need to

work on large amount of textual data and reduces manual labor required to process all he data.

Summarization:

large language models generate a concise summary of lengthy text or documents [37]. This is used to summarize documents like research papers, articles, et cetera to read an overview instead reading whole document which consists of required information. summary saves time and effort while making sure that the most important information is captured correctly [43]. This feature could make learning and creating things faster and better, which would make it useful for both people and companies. copilot in edge can generate a summary of a webpage or a document.

Virtual Assistance:

LLMs can power virtual assistants by understanding natural language queries and generating appropriate responses or actions. This will enhance user experience and improves operational efficiency. we can use LLMs for customer support chatbots, which can process user queries and provide information as a response. Further study into virtual assistance is to improve emotional intelligence, personalization, security, privacy, and the ability to understand conversation which improves user experience and effectiveness.

Semantic search:

LLMs improve search results by understanding the semantics of the query, such as when a user searches for something and search engines return results based on keywords, semantic search assists in understanding the query's context, analyzing it in relation to other page contents, and returns the most relevant results. For example, when a user searches for "tasty indian food" traditional search engines might focus only on matching keywords like "indian" and "food," they might return results that are not contextually relevant. LLMs consider the user's location, preferences (e.g., vegetarian or non-vegetarian) and possibly even previous search history or context. Based on this understanding, the LLM would generate search results that include top-rated Indian restaurants, popular Indian dishes, recipes, cooking tutorials, restaurant reviews, and other relevant content to the user's query.

Speech recognition:

LLMs can also be used in speech recognition by transcription or generating text from speech. LLMs are trained on massive amount of text which makes it understand context and relation between words in a sentences so these are trained on transcribed audio materials which allows it to recognize audio and generate text. Using their ability to generate human like language it can improve accuracy and contextual understanding also user interactivity by voice search or virtual assistants

Information Extraction (IE):

information extraction techniques involve extraction of structured information from an unstructured or semi-structured information in a given data source, understanding capabilities of LLMs let it extract essential entities, relationships, events, and other relevant information from the text. IE is used in tasks such as text summarization, semantic searches, text classification and more. This can be used to extract required information from large data sets or documents of an organization efficiently. Further study in this field will enable extracting all the events from the data sources efficiently.

Dialog systems:

Dialog systems use LLMs as their main building block to facilitate human-like conversations between users and machines. These systems rely on the LLMs' ability to understand user queries and generate appropriate responses. Blender, Milabot, Xiaoice, Meena, and potentially GPT-4, are indeed notable chatbots known for their capability to generate interesting and useful content while staying safe.

6. LLMs in different domains:

As we discussed different applications of LLMs now let's see their uses in various fields like education, Healthcare, law, finance and business media and entertainment, engineering and marketing

Healthcare:

Large language models are significantly contributing to health care sector in various ways,LLMs like chatGPT, claude can be used for different applications like Clinical decision making, medical imaging analysis, clinical documentation, medical related query responses, patient education, Drug discovery and development, Clinical research, Healthcare administration and operations and more. Recently Without any specialized training or fine-tuning chatgpt has performed remarkably in USMLE (united states medical licensing exam).in [36] they have introduced a new MultiMedQA benchmark dataset for evaluating LLMs on clinical tasks[48]. These models excelled in memorization tasks compared to critical thinking and problem solving questions. There are few domain specific LLMs in healthcare, which are trained on large volumes of data from various medical fields. They help doctors, healthcare professionals, and patients by reading and understanding complicated medical documents, and provide a comprehensive idea of those documents with valuable insights. Several tools, like Ada

Health [38], Babylon Health [41], and Buoy Health [39], are already in use and let the system talk to patients. Few tools are also used by medical practitioners like X-rayGPT which takes x-ray images and provides a detailed analysis on the given input with valuable insight for patients or doctors. Models like BioBERT, A variant of BERT trained on biomedical corpora, excelling in understanding medical terminology and aiding tasks like disease prediction and drug interaction analysis. Med-PaLM is a LLM designed for the medical domain. It is, capable of handling various types of biomedical data, including Understanding medical texts and terminology, Analyzing images such as X-rays, MRIs, and CT scans, Processing genetic information. other models include ClinicalBERT, BlueBERT, and BioGPT. The future research seems promising for LLMs in healthcare, with expected improvements in areas like medical diagnosis, treatment planning, patient teaching, automation of routine tasks.

Education:

In the education domain Large Language Models (LLMs) has attained significant adoption in recent times. It became a promising tool for many aspects of education like enhancing student learning, teacher support, and improving administrative functions. LLMs can help students in many ways, as these models were trained on vast datasets they are exceptional in generating responses related to user queries, and can help them in their academics like for preparation for exams, helping in completion of assignments and understanding of difficult concepts. For exam preparation, LLMs can give valuable insights, explanations of topics, and practice questions related to the specific topics students are studying. It can generate accurate and detailed explanation as they excel in it, which allow students check their own progress, and upgrade themselves. LLMs can also be used for administrative tasks such as evaluation by leveraging nlp capabilities and finetuning these models can grade assignments, exams, collect student responses and give appropriate feedback, by automating all these tasks we can save teachers time and effort which can be used to focus on mentoring students, teaching and other productive purposes. There are few disadvantages using these LLMs in education, by relying much on these AI tools which can do our work in an instant for us will decrease our critical thinking and creative skills which is a potential risk for us. The use of LLMs in education raises ethical considerations, such as privacy issues. It is important to make sure that AI technologies are used in an honest and responsible way, taking into consideration various perspectives and needs of students.

Finance:

LLMs (Large Language Models) are changing the finance industry because they can be used in so many different ways. These models are great in sentiment analysis, and they can handle huge amounts of text data from financial news, social media, and other places to accurately determine how users understand about the market. LLMs also play a big role in risk management. They use past market data to find possible risks and predict market trends, which helps investors and financial organizations make smart choices. These models can play a crucial role in algorithmic trading, where they develop algorithms for automation of trading, enhancing trading efficiency and profitability. They can also contribute to customer service and support in the finance industry by chatbots and virtual assistants, which answer questions, share knowledge about products and services, and help people in their financial tasks. LLMs can detect fraud activities by analyzing indicative patterns in the data, this will offer organizations some level of security. They are very important for market research, forecasting, credit score, and providing valuable insights to users and investors. LLMs are changing many parts of the finance business by using ml and nlp to make jobs easier to automate, analyze data, and make better decisions.

Engineering:

LLMs are making huge advancements in the world of engineering, offers a variety of applications and advantages. These models can help in technical documentation, writing reports, manuals, precision and clarity. LLMs like chatgpt can generation code and tools like DevinAI can help in software development by generating code, testing and also capable of deploying, providing suggestions for code optimization, and assisting in debugging processes. Few tools are taking interview too.LLMs [40] have been assessed in engineering design activities to see their efficiency in conceptualization, prototyping, and modeling. The results showed benefits for developing design concepts, evaluating performance metrics, and predicting outcomes. The use of LLMs is changing the way engineers do their jobs by making work easier, increasing output, and helping engineers solve tough problems more effectively.

Law:

Usage of LLMs in law for various purposes are increasing. These models were trained on large amount of egal text, including court cases, statutes, and legal opinions, which allows them to perform tasks such as legal research, document analysis, contract review, and summarization of legal texts. LLMs can help lawyers improving their efficiency by finding related case law, statutes, and precedents they can also analyze legal documents like contracts and agreements to find risks, errors, or terms that need more careful consideration. By their ability they can be used for summarizing lengthy legal texts, or documents, and generating concise summaries, we can save time and improve productivity. They can be used for increasing legal accessibility and understanding

for the general public. There are few tools like LAWFYI is a great AI tool that makes complicated legal issues easier for everyone to understand, teaches people about their rights and giving lawyers smart, time-saving tools. It demonstrates that AI has a lot of potential to transform traditional systems and make them easier to use and more efficient.

Marketing:

Large Language Models (LLMs) are being used more and more in marketing to improve different parts of promotions and plans. These models, trained on extensive text data and can give useful insights of domains, create content, and enhance marketing efforts. LLMs can analyze customer feedback, social media interactions, and market trends to give organizations useful data for targeting specific groups of users [market place]to make ads personalized. These models are good in understanding natural language and they can generate creative content for ads, emails, and social media posts that help correlate to people interests. LLMs also help make marketing easier by writing messages that are customized for each consumer based on what they like and don't like and how they act using their past interactions. LLMs can be used to improve customer engagements and contribute towards business growth.

Media and Entertainment:

LLMs are making significant changes in the media and entertainment industry, these models are being utilized for various applications, including content generation, scriptwriting, recommendation systems, and audience analysis. LLMs can generate different types of content like articles, stories, and scripts, with excellent efficiency and fluency, reducing the time and effort required for content creation. And in few movies, they started hiring prompt engineers (They meticulously engineer the input prompts to optimize the AI model's accuracy and effectiveness) to get the best from these models. There are few LLMs like soundraw, boomy that can generate musical content. The rise of newscasters that are powered by AI is another example of how LLMs are being used in media, providing virtual news reporters made possible by AI technologies. LLMs are making big changes in the entertainment and media industries, making them more creative, efficient, and interesting to audiences.

The ways that LLMs are used in these fields show how this field is still growing and changing. As technology keeps getting better, it's important to be able to adapt to these changes while also making sure that they are used responsibly so that productivity is maximized.

As of now we discussed what are LLMs, history, applications, lets now delve into design cycle of Generative AI, and then let's try to build a domain specific model by using an existing model. Existing models, in the sense that we have so many open source LLMs that are trained on vast amounts of publicly available data that can be used by anyone.

In our case, we will consider Llama 2, which is open source and freely available for both research and commercial use. The ease of access encourages trying new things, coming up with new ideas, and the responsible scaling of ideas.

7. Building a domain specific LLM:

At the first stage of the design cycle, defining the problem, the target of the LLM is defined, i.e., what we expect the model to perform when given an input and what type of output we need from it. The GenAI model target may be to perform well on a single task or multiple tasks.

In building an LLM, we need to gather the maximum possible amount of text data available in the problem domain that we are working on. The collected data is divided into multiple chunks using RecursiveCharacterTextSplitter() present in the LangChain Framework. Then these chunks of data are converted into vector embeddings and stored in the Vector Database using sentence-transformers.

The second stage is developing the model from scratch or using a pre-trained model. At this step, we can use the pre-trained model available from various sources, such as Llama 2, Bloom 176B, Falcon 180B, and MPT 30B. The model is selected based on the input processing, output generation, and model performance in the selected problem domain. If there are no such models meeting our requirements for performance or output format, we can train and build our own model from scratch by selecting the type of model from available models such as RNN, Transformer, CNN, and ViT based on our requirements. Here, we considered an open-source model llama2 for this research.

Fine tuning is the iterative process of adapting and aligning the model for our definition, i.e., the scope of the LLM. In this process, several steps of prompt engineering are done with different learning techniques, such as Supervised fine tuning with the help of instruction-response datasets, few shot learning, transfer learning, or it can also be fine-tuned based on the scope of the model in the problem definition using Reinforcement Learning from Human Feedback. The model's performance is evaluated based on the performance metrics for quality, diversity, authenticity, and variance of the model generation using some common techniques such as Visual Inspection, Inception Score, and Fréchet Inception Distance to know how precisely the model has aligned with human intelligence and then fine-tuned to improve

the performance by the adjustment of the model's arguments to best fitted set. Here, we restricted the LLM's response generation only to domain specific knowledge. For instance, in a medical chatbot app, it is trained on healthcare domain related data to restrict the model responses to that specific domain.

The final step of the design cycle is to deploy, maintain, and optimize the model. The LLM is deployed in the target environment and integrated into several applications. This deployment is managed to ensure the proper flow of usage of the model. Also, this model is optimized based on its performance in the target environment to provide the best possible outcome. Here, we can use free deployment platforms such as Hugging Face Spaces to deploy and maintain the LLM model. We can also use paid services based on the requirements.

8. Conclusion:

In this paper, we have given an introduction of Natural Language Processing (NLP) and language modeling, followed by a concise survey of Large Language Models (LLMs) and their history. We have explored the concept of generative AI and delved into an overview of LLMs, taken an example ChatGPT. Then, we have seen the evolution of LLMs and shown what they are capable of in various tasks quickly and efficiently across different domains such as healthcare, education, banking, engineering, law, marketing, and media. Lastly, we have discussed the design cycle involved in building LLMs or domain-specific models. The objective of this study is to give people the knowledge they need to safely and properly adopt this disruptive technology, which will open up new areas of innovation in many fields

REFERENCES:

[1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," Advances in neural information processing systems, vol. 30, 2017.

[2] W. X. Zhao, K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, et al., "A survey of large language models," arXiv preprint arXiv:2303.18223, 2023.

[3] S. Mohamadi, G. Mujtaba, N. Le, G. Doretto, and D. A. Adjeroh, "Chatgpt in the age of generative ai and large language models: A concise survey," arXiv preprint arXiv:2307.04251, 2023.

[4] H. Naveed, A. U. Khan, S. Qiu, M. Saqib, S. Anwar, M. Usman, N. Barnes, and A. Mian, "A comprehensive overview of large language models," arXiv preprint arXiv:2307.06435, 2023.

[5] E. Kasneci, K. Seßler, S. Küchemann, M. Bannert, D. Dementieva, F. Fischer, U. Gasser, G. Groh, S. Günnemann, E. Hüllermeier, et al., "Chatgpt for good? on opportunities and challenges of large language models for education," Learning and Individual Differences, vol. 103, p. 102274, 2023.

[6] A. Madani, B. McCann, N. Naik, N. S. Keskar, N. Anand, R. R. Eguchi, P.-S. Huang, and R. Socher, "Progen: Language modeling for protein generation," arXiv preprint arXiv:2004.03497, 2020.

[7] Y. Cao, S. Li, Y. Liu, Z. Yan, Y. Dai, P. S. Yu, and L. Sun, "A comprehensive survey of ai-generated content (aigc): A history of generative ai from gan to chatgpt," arXiv preprint arXiv:2303.04226, 2023.

[8] I. Beltagy, K. Lo, and A. Cohan, "Scibert: A pretrained language model for scientific text," arXiv preprint arXiv:1903.10676, 2019.

[9] J. Li, T. Tang, W. X. Zhao, J.-Y. Nie, and J.-R. Wen, "Pretrained language models for text generation: A survey," arXiv preprint arXiv:2201.05273, 2022.

[10] S. Wu, O. Irsoy, S. Lu, V. Dabravolski, M. Dredze, S. Gehrmann, P. Kambadur, D. Rosenberg, and G. Mann, "Bloomberggpt: A large language model for finance," arXiv preprint arXiv:2303.17564, 2023.

[11] B. Li, K. Mellou, B. Zhang, J. Pathuri, and I. Menache, "Large language models for supply chain optimization," arXiv preprint arXiv:2307.03875, 2023.

[12] L. Bariah, Q. Zhao, H. Zou, Y. Tian, F. Bader, and M. Debbah, "Large language models for telecom: The next big thing?, "arXiv preprint arXiv:2306.10249, 2023.

[13] M. Chen and T. et. al., "Evaluating large language models trained on code," arXiv preprint arXiv:2107.03374, 2021.

[14] S. Salman, J. A. Shamsi, and R. Qureshi, "Deep fake generation and detection: Issues, challenges, and solutions," IT Professional, vol. 25, no. 1, pp. 52–59, 2023.

[15] Z. Sun, "A short survey of viewing large language models in legal aspect," arXiv preprint arXiv:2303.09136, year=2023.

[16] R. Qureshi, M. Irfan, T. M. Gondal, S. Khan, J. Wu, M. U. Hadi, J. Heymach, X. Le, H. Yan, and T. Alam, "Ai in drug discovery and its clinical relevance," Heliyon, 2023.

[17] O. B. Shoham and N. Rappoport, "Cpllm: Clinical prediction with large language models," arXiv preprint arXiv:2309.11295, 2023.

[18] [18] Q. Al-Tashi, M. B. Saad, A. Muneer, R. Qureshi, S. Mirjalili, A. Sheshadri, X. Le, N. I. Vokes, J. Zhang, and J. Wu, "Machine learning models for the identification of prognostic and predictive cancer biomarkers: A systematic review," International journal of molecular sciences, vol. 24, no. 9, p. 7781, 2023.

[19] A. Holzinger, K. Keiblinger, P. Holub, K. Zatloukal, and H. Müller, "Ai for life: Trends in artificial intelligence for biotechnology," New Biotechnology, vol. 74, pp. 16–24, 2023.

[20] E. Brynjolfsson, D. Li, and L. R. Raymond, "Generative ai at work," tech. rep., National Bureau of Economic Research, 2023.

[21] P. Samuelson, "Generative ai meets copyright," Science, vol 381, no. 6654, pp. 158–161, 2023.

[22] I. Chiang, Unleashing the Power of Generative AI: The Race for Advancement and the Global Ramifications. PhD thesis, Massachusetts Institute of Technology, 2023.

[23] S. Wang, S. Menon, T. Long, K. Henderson, D. Li, K. Crowston M. Hansen, J. V. Nickerson, and L. B. Chilton, "Reelframer: Cocreating news reels on social media with generative ai," arXiv preprintarXiv:2304.09653, 2023.

[24] S. Mayahi and M. Vidrih, "The impact of generative ai on the future of visual content marketing," arXiv preprint arXiv:2211.12660, 2022.

[25] P. Ghimire, K. Kim, and M. Acharya, "Generative ai in the construction industry: Opportunities & challenges," arXiv preprint arXiv:2310.04427, 2023.

[26] C. Zhang, C. Zhang, S. Zheng, M. Zhang, M. Qamar, S.-H. Bae, and I. S. Kweon, "A survey on audio diffusion models: Text to speech synthesis and enhancement in generative ai," arXiv preprint arXiv:2303.13336, vol. 2, 2023.

[27] N. Aldausari, A. Sowmya, N. Marcus, and G. Mohammadi, "Video generative adversarial networks: a review," ACM Computing Surveys (CSUR), vol. 55, no. 2, pp. 1–25, 2022

[28] N. Chen, A. Klushyn, R. Kurle, X. Jiang, J. Bayer, and P. Smagt, "Metrics for deep generative models," in International Conference on Artificial Intelligence and Statistics, pp. 1540–1550, PMLR, 2018.

[29] S. Barratt and R. Sharma, "A note on the inception score," arXiv preprint arXiv:1801.01973, 2018.

[30] A. Obukhov and M. Krasnyanskiy, "Quality assessment method for gan based on modified metrics inception score and fréchet inception distance," in Software Engineering Perspectives in Intelligent Systems: Proceedings of 4th Computational Methods in Systems and Software 2020, Vol. 1 4, pp. 102–114, Springer, 2020.

[31] A. Verine, B. Negrevergne, M. S. Pydi, and Y. Chevaleyre, "Precision Recall divergence optimization for generative modeling with gans and normalizing flows," arXiv preprint arXiv:2305.18910, 2023.

[32] R. Kansal, J. Duarte, H. Su, B. Orzari, T. Tomei, M. Pierini, M. Touranakou, D. Gunopulos, et al., "Particle cloud generation with message passing generative adversarial networks," Advances in Neural Information Processing Systems, vol.34, pp. 23858–23871, 2021.

[33] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer,"The Journal of Machine Learning Research, vol. 21, no. 1, pp. 5485–5551, 2020.

[34] D. Su, Y. Xu, G. I. Winata, P. Xu, H. Kim, Z. Liu, and P. Fung, "Generalizing question answering system with pre-trained language model fine-tuning," in Proceedings of the 2nd Workshop on MachineReading for Question Answering, pp.203–211,2019.

[35] K. Singhal, S. Azizi, T. Tu, S. S. Mahdavi, J. Wei, H. W. Chung, N. Scales, A. Tanwani, H. Cole-Lewis, S. Pfohl, et al., "Large language models encode clinical knowledge," Nature, pp. 1–9, 2023.

[36] T. Zhang, F. Ladhak, E. Durmus, P. Liang, K. McKeown, and T. B.Hashimoto, "Benchmarking large language models for news summarization," arXiv preprint arXiv:2301.13848,2023.

[37] S. M. Jungmann, T. Klan, S. Kuhn, and F. Jungmann, "Accuracy of a chatbot (ada) in the diagnosis of mental disorders: comparative case study with lay and expert users," JMIR formative research, vol. 3, no. 4,p. e13863, 2019.

[38] P. Malik, M. Pathania, V. K. Rathaur, et al., "Overview of artificial intelligence in medicine," Journal of family medicine and primary care,vol. 8, no. 7, p. 2328, 2019.

[39] X. Wang, N. Anwer, Y. Dai, and A. Liu, "Chatgpt for design, manufacturing, and education," 2023.

[40] D. Magalhaes Azevedo and S. Kieffer, "User reception of ai-enabled mhealth apps: The case of babylon health.,"

[41] M. Sallam, "The utility of chatgpt as an example of large language models in healthcare education, research and practice: Systematic review on the future perspectives and potential limitations," medRxiv, pp. 2023–02, 2023.

[42] C. Shen, L. Cheng, Y. You, and L. Bing, "Are large language models good evaluators for abstractive summarization?" arXiv preprint arXiv:2305.13091, 2023.

[43] N. S. Keskar, B. McCann, L. R. Varshney, C. Xiong, and R. Socher, "Ctrl: A conditional transformer language model for controllable generation," arXiv preprint arXiv:1909.05858, 2019.

[44] K. Wang, C. Gou, Y. Duan, Y. Lin, X. Zheng, and F.-Y. Wang, "Generative adversarial networks: introduction and outlook," IEEE/CAA Journal of Automatica Sinica, vol. 4, no. 4, pp. 588–598, 2017.

[45] A. J. Thirunavukarasu, D. S. J. Ting, K. Elangovan, L. Gutierrez, T. F. Tan, and D. S. W. Ting, "Large language models in medicine," Nature Medicine, pp. 1–11, 2023.

[46] J. Morley, N. J. DeVito, and J. Zhang, "Generative ai for medical research," 2023.

[47] Z. Lin, H. Akin, R. Rao, B. Hie, Z. Zhu, W. Lu, A. dos Santos Costa, M. Fazel-Zarandi, T. Sercu, S. Candido, et al., "Language models of protein sequences at the scale of evolution enable accurate structure prediction," BioRxiv, vol. 2022, p. 500902, 2022.

[48] Hadi, Muhammad Usman & Al-Tashi, Qasem & Qureshi, Rizwan & Shah, Abbas & Muneer, Amgad & Irfan, Muhammad & Zafar, Anas & Shaikh, Muhammad & Akhtar, Naveed & Wu, Jia & Mirjalili, Seyedali. (2023). Large Language Models: A Comprehensive Survey of its Applications, Challenges, Limitations, and Future Prospects. 10.36227/techrxiv.23589741.