Advanced Sciences and Technology Journal

ASTJ vol. 2, no. 1 (2025), P 1034
10.21608/astj.2025.342484.1034
https://astj.journals.ekb.eg/

Leveraging Large Language Models for Document Analysis and Decision-Making in AI Chatbots

Islam M. Ibrahim, Mahmoud Soliman, Shahd Ossama, Sherif Tarek, Iyad Abuhadrous*

Software Engineering and Information Technology, Faculty of Engineering and Technology, Egyptian Chinese University, Cairo, Egypt
iyad.mohammed@ecu.edu.eg

ARTICLEINFO

Article history:
Received 20 December 2024
Revised 9 January 2025
Accepted 16 January 2025
Available online 17 January 2025

Handling Editor:
Prof. Dr. Mohamed Talaat Moustafa

Keywords:
LLMs
Chatbots
Decision-Making
RAG
Document Analysis

ABSTRACT

The rapid improvement of Large Language Models (LLMs) is transforming how businesses handle document workflows, making the process faster and smarter. This paper explores how businesses can use LLMs to create AI chatbots that enhance document analysis and improve decision-making. These chatbots can perform tasks like extracting key information, summarizing content, and generating insights, saving both time and resources for businesses that work with a large volume of written materials. They also make quick, informed decisions easier by providing accurate, context-based responses. However, despite their benefits, these systems face challenges like ensuring data security, managing high computational demands, and addressing ethical concerns such as bias and transparency. The paper also examines modern advancements and future possibilities, highlighting how this technology could reshape the way businesses handle documents and make decisions. It emphasizes the potential of LLMs to deliver smarter, safer, and more efficient tools that adapt to the ever-changing needs of modern organizations.

1. Introduction

In today's fast-paced digital world, companies across industries are increasingly relying on data to drive their decisions and operations. Much of this valuable information is stored in various documents, including reports, contracts, manuals, and financial statements in formats such as PDFs and spreadsheets. However, going through these documents to find useful insights can be extremely challenging, especially when businesses are overwhelmed by the amount of data they need to process quickly and accurately.

This is where new technology, particularly Large Language Models (LLMs), comes in. These models use advanced learning techniques to understand and generate human-like language, making them highly effective for tasks that involve natural language processing [4]. They can summarize text, analyze sentiment, and answer questions, which has proven to be a groundbreaking development in many fields [24]. Chatbots powered by LLMs take this technology even further, offering a more interactive and user-friendly experience [9]. These intelligent assistants can search both structured and unstructured data from documents, summarize content, and provide recommendations. This ability

helps businesses save significant time and effort, making it easier to process documents and make data-driven decisions efficiently [19].

For example, in the healthcare sector, AI-powered chatbots equipped with LLMs assist doctors by providing instant responses to medical queries, recommending treatment options, and aiding clinical decision-making. In legal and financial domains, tools like "Risk-o-Meter" automate the analysis of contracts and financial statements, identifying risks and offering recommendations with over 90% accuracy [18]. In education, LLM-powered applications like ChatGPT help students by answering academic questions, providing tutoring support, and simplifying research processes [10].

Furthermore, these chatbots now also have decision-support capabilities. LLMs can be trained to identify trends and provide context-specific recommendations to help users navigate complex decision-making processes [18]. For instance, they can analyze historical financial data to forecast market trends, highlight potential risks, and offer strategic advice. In customer service, chatbots streamline operations by addressing user queries, improving engagement, and delivering personalized recommendations, thereby enhancing customer satisfaction and operational efficiency.

However, it is important to note that using LLM-based chatbots comes with challenges. Issues related to data security, understanding how these models work, the high computational power they require, and the risk of inaccurate or biased results must be carefully addressed to ensure the technology is trustworthy [6, 11]. Ethical concerns about using AI in decision-making also raise important questions about accountability and transparency, especially in sensitive situations [14].

This article contributes to the field by providing a comprehensive analysis of how LLMs can be effectively leveraged for document analysis and decision support. It introduces a framework for integrating LLMs into business workflows, evaluates their advantages and limitations across different industries, and highlights strategies to address key challenges such as data security, computational demands, and ethical considerations.

2. Background

2.1. Language Modeling

Language modeling plays a pivotal role in enabling computers to comprehend and generate human language. Over the years, it has evolved from mathematical concepts into advanced computational techniques, propelled by the advancements in artificial intelligence. The evolution of language modeling can be categorized into four distinct approaches: Statistical Language Models, Neural Language Models, Pre-Trained Language Models, and Large Language Models (LLMs).

2.1.1 Statistical Language Models (SLMs):

These models rely on statistical inference methods, operating under the assumption that the likelihood of a word depends solely on the preceding word in a sequence, a principle known as the Markov assumption. While

effective for basic tasks, SLMs face challenges in recognizing complex linguistic patterns. A notable example is the n-gram model [1].

2.1.2 Neural Language Models (NLMs):

By incorporating neural networks, NLMs introduced a more sophisticated way to analyze language features. Although they provided a significant leap in handling natural language processing tasks, their high computational and memory demands, particularly in recurrent neural networks (RNNs), limited their scalability [2].

2.1.3 Pre-Trained Language Models (PLMs):

These models emphasize context-aware language representation by employing bidirectional LSTMs during pre-training, followed by fine-tuning for specific tasks. PLMs offered significant improvements in contextual understanding, with GPT-2 being one of the most recognized implementations of this methodology.

2.1.4 Large Language Models (LLMs):

Building on the foundation of PLMs, LLMs enhance performance by massively increasing the number of parameters. This scalability not only improves model accuracy but also leads to the emergence of novel capabilities, such as reasoning and contextual inference, absent in smaller models.

2.2. Transformer Architectures:

Transformer architectures revolutionized natural language processing by addressing limitations of earlier models like RNNs and LSTMs [2]. These older architectures often struggled with understanding entire sequences, as they relied heavily on token-by-token predictions. Transformers introduced the attention mechanism, allowing them to process sequences in parallel and grasp their broader context, which drastically improved both efficiency and accuracy.

There are three primary types of transformer architectures, each tailored for specific tasks:

2.2.1 Auto-Regressive Models (GPT-like):

Often described as decoder-only models, these focus on generating one token at a time, where each token depends on the preceding ones. During training, the model learns by predicting the next token, and during inference, it generates sequences autonomously. Examples of auto-regressive models include GPT, BLOOM [3], and LLaMa2 [4]. Variations include:

- Causal Decoders: Unidirectional attention is applied to ensure that only preceding tokens are considered.
- Prefix Decoders: A combination of bidirectional and unidirectional masking enhances flexibility and speeds up convergence.

2.2.2 Auto-Encoding Models (BERT-like):

These models, also known as masked language models, focus on predicting masked tokens within a sequence. By using bidirectional attention, auto-encoding models extract contextual information from

both preceding and succeeding tokens. They are particularly effective for tasks like text classification and sentiment analysis. Examples include BERT [5], RoBERTa, and XLM-R.

2.2.3 Encoder-Decoder Models (Sequence-to-Sequence):

Designed for generating one sequence based on another, these models are invaluable for tasks like translation and summarization. During training, the input sequences are deliberately corrupted, and the models attempt to reconstruct the original sequences. This method enables them to handle complex transformations. Examples of such architectures are BART [6] and T5 [7].

2.3. T Prior Research on Document Analysis and Decision-Making with LLMs:

The application of AI and language models in document analysis and decision-making has been an active area of research. Early studies relied on rule-based systems and statistical approaches, offering limited capabilities. The advent of deep learning marked a significant breakthrough, enabling more accurate and efficient text analysis. Barnea (2020) [11] explored how AI-driven systems could deliver faster insights for better decision-making. Alaaeldin et al. (2021) [9] designed chatbots that utilized real-time data analytics to aid business decisions. In the legal domain, Chakrabarti et al. (2018) [18] developed tools to automatically identify risks in legal documents, reducing manual workloads. More recently, Pokhrel et al. (2023) [28] demonstrated the use of frameworks like LangChain to process extensive text datasets and answer complex queries. These efforts highlight significant advancements while also revealing challenges such as bias, transparency, and scalability. Building on this foundation, this paper examines how LLMs can further enhance document analysis and decision-making across diverse industries.

3. Methodology

3.1. Purpose and Scope of the Systematic Review:

This systematic review aims to provide a comprehensive and structured examination of existing literature on leveraging large language models (LLMs) for document analysis and decision-making in AI chatbots. The purpose is to synthesize relevant research, offering an unbiased and thorough summary of the available knowledge. The review focuses on identifying advancements, applications, and challenges in integrating LLMs into document-intensive industries, particularly legal, healthcare, and finance sectors. [7].

3.2. How LLMs Work in Chatbot Design:

Large Language Models (LLMs) are the brains behind many modern chatbots. Here’s how they’re commonly integrated to create smooth and effective user experiences:

3.2.1 API-Based Setup:

Think of APIs as bridges. With services like OpenAI's GPT API, chatbots can send questions to the model and get instant replies. This approach is easy to set up and keeps things up-to-date since developers don’t have

to worry about maintaining or upgrading the AI themselves ensuring the system stays up-to-date and efficient [24].

3.2.2 Using Modular Pipelines:

In more advanced setups, chatbots use a step-by-step system called a modular pipeline. First, the input is cleaned up and prepared for the LLM. After the LLM generates a response, extra tweaks are made to ensure the reply fits the chatbot's purpose. This process helps make responses more relevant and polished responses [9, 28].

3.2.3 Using Knowledge Bases (RAG):

Retrieval-Augmented Generation (RAG) combines LLM capabilities with knowledge bases. The chatbot retrieves relevant information from databases or documents and incorporates it into LLM-generated responses. This hybrid method improves the accuracy and helpfulness of answers in domain-specific contexts [23, 28].

3.2.4 Combining AI with Other Tools:

Some chatbots integrate LLMs with rule-based systems or specialized AI tools. This approach allows for general responses powered by the LLM and precise, domain-specific answers from auxiliary systems, creating a balance between versatility and specificity [17, 18].

These techniques help chatbots feel more natural and intelligent, making them useful for everything.

3.3. Adherence to PRISMA Guidelines:

To ensure rigor and transparency, this review adheres to the Preferred Reporting Items [8] for Systematic Reviews and Meta-Analyses (PRISMA) guidelines. These guidelines provide a standardized framework for defining the research focus, designing a search strategy, selecting studies, extracting data, and synthesizing results.

3.4. Research Scope and Keyword:

The scope of this review includes examining literature on large language models (LLMs) and their integration into AI chatbots. Essential keywords used in the search include 'LLMs,' 'AI chatbots,' 'document analysis,' 'decision-making,' and 'RAG.' Related terms such as 'LLM for document processing,' 'AI in document management,' and 'decision support systems' were also considered.

3.5. Search and Selection Process:

3.5.1 Initial Manuscript Selection:

The initial selection involved identifying studies from multiple databases, including Google Scholar, IEEE Xplore, arxiv, ScienceDirect, and Scopus. The search was confined to articles published between 2010 and 2024, ensuring the inclusion of the most relevant and up-to-date research.

3.5.2 Manuscript Identification:

After a preliminary screening of titles and abstracts, 80 manuscripts were identified as potentially relevant. A deeper review narrowed this list to 40 manuscripts that directly addressed the integration of LLMs into document analysis or decision-making processes.

3.5.3 Rigorous Review and Inclusion Criteria:

Studies were included based on their relevance to the research focus, methodological rigor, and publication within the specified date range. Studies that lacked robust methodologies or deviated significantly from the focus were excluded.

3.5.4 Final Selection and Exclusion Statistics:

Following the application of inclusion and exclusion criteria, 28 studies were selected for the final review. These studies form the core dataset of this systematic review.

Table 1. AI and LLMs show tremendous potential across industries, enhancing efficiency, personalization, and decision-making.

Author	Year	Type Of Study	Explanation/Solution/Conclusion
Rana Alaaeldin, Evan Asfoura, Gamal Kassem, Mohammad Samir Abdel-Haq [9]	2021	Developing a chatbot-based system to support decision-making using Big Data analytics.	The study introduces a chatbot-based IT solution linking Key Performance Indicators (KPIs) and analytics models to enhance decision-makers' interaction with Big Data. It simplifies analytics, improves usability, and aligns with business strategies, enabling non-technical users to make informed decisions. Future enhancements include adding dynamic features, visualizations, and multilingual support.
Joshua Ebere Chukwuere [10]	2024	Exploratory	The study examines how AI chatbots like ChatGPT transform higher education by enhancing student support, streamlining administration, and boosting research productivity. Benefits include personalized learning, 24/7 assistance, and LMS integration. Challenges such as academic integrity, ethical concerns, and resource allocation are addressed with proposed solutions like ethical guidelines, training, and continuous monitoring, enabling institutions to harness AI chatbots for a more efficient and innovative academic environment.
Avner Barnea[11]	2020	Analytical article published in a journal.	Avner Barnea (2020) explores AI's role in improving decision-making by enhancing data analysis and predictions. AI complements human analysts, offering faster, more accurate insights, but adoption remains slow due to complexity. Proper integration can reduce biases, improve strategies, and strengthen competitive and security responses.
Nima Ghorashi, Ahmed Ismail, Pritha Ghosh, Anton Sidawy, Ramin Javan [12]	2023	Review article	The study highlights AI-powered chatbots as valuable tools in medical education, aiding learning, clinical decision-making, and research. They simplify concepts, enhance retention, and provide real-time feedback. While promising, chatbots must complement human expertise, addressing biases and ethical concerns for effective integration.
Bruno Azevedo Chagas et al [13]	2023	Mixed methods study	The study found the COVID-19 chatbot effective for symptom screening and health education, with high user satisfaction. However, improvements are needed in updating information, enhancing user guidance, and addressing diverse user needs to maximize its impact in healthcare.
Michele Salvagno, Fabio Silvio Taccone, Alberto Giovanni	2023	Perspective article	The article highlights ChatGPT's potential in aiding scientific writing, such as drafting and proofreading, but stresses it cannot replace human expertise. Ethical concerns like plagiarism and access disparities require regulation, positioning AI as an assistive tool rather

Gerli [14]			than a substitute.
Ian L. Alberts, Lorenzo Mercolli, Thomas Pyka, George Prenosil, Kuangyu Shi, Axel Rominger, Ali Afshar-Oromieh [15]	2023	Editorial	The article highlights ChatGPT's potential in nuclear medicine for content generation and education but notes its current limitations in accuracy and depth. Ethical concerns like misinformation and plagiarism require regulation. ChatGPT is a useful tool but needs expert oversight.
Netra Pal Singh and Devender Singh [16]	2019	Case study	The study reviews chatbot adoption in Indian banks, noting limited functionality, low awareness, and inconsistent accuracy. It recommends enhancing capabilities, increasing awareness, and tailoring solutions to diverse customer needs for better efficiency and satisfaction.
Kien Nguyen-Trung, Alexander K. Saeri, Stefan Kaufman [17]	2024	Evaluation of generative AI tools in accelerating evidence reviews	AI tools like ChatGPT and GPT for Sheets enhance efficiency in evidence reviews, particularly in search string formulation and literature screening. However, they require human oversight to address inconsistencies and errors, ensuring quality outcomes. Properly utilized, these tools can significantly accelerate research processes.
Dipankar Chakrabarti, Indranil Mitra, Nandini Roy, Neelam Patodia, Satyaki Roy, Prasun Nandy, Udayan Bhattacharya, Jayanta Mandi [18]	2018	Developing a framework to automate risk analysis in legal documents for better decision support in organizations.	The study introduces "Risk-o-Meter," an AI-based framework using machine learning and natural language processing to automate risk analysis in legal documents. It achieves high accuracy (91%) in identifying risk-prone paragraphs, reduces manual effort, and is scalable for other document types. Future improvements include assessing risk severity and suggesting mitigations.
Prajwal Akare, Rupal Wyawahare, Swaraj Bawankule, Sankalp Lanjewar, Arpit Nandanwar, Mrs. Surbhi Khare [19]	2023	Developing a deep learning model to analyze, classify, and process various types of documents.	The study developed a deep learning-based document analyzer using CNNs to classify and process documents with up to 95% accuracy. It simplifies tasks like classification and feature extraction, offering efficient and accurate results. Future work aims to enhance adaptability and expand applications.
Mohsen Khosravi, Zahra Zare, Seyyed Morteza Mojtabaeian, Reyhane Izadi [20]	2024	Systematic Review	AI improves diagnostics, resource management, and personalized care in healthcare. Despite ethical and implementation challenges, it shows promise in enhancing efficiency and outcomes. Standardized practices and further research are needed.
Nouri Hicham, Nassera Habbat, Sabri Karim [21]	2023	Conceptual Framework and Literature Review	The study outlines a framework for using AI to enhance marketing efficiency, customer engagement, and personalization while addressing challenges like data privacy and ethics. It emphasizes balancing AI and human creativity for optimal results.
Chengrun Yang, Xuezhi Wang, Yifeng Lu, Hanxiao Liu, Quoc V. Le, Denny Zhou, Xinyun Chen [22]	2024	Optimization Study	OPRO uses LLMs to iteratively optimize tasks via natural language prompts, achieving up to 50% performance improvements and outperforming human-designed solutions. It demonstrates adaptability and effectiveness across optimization challenges.
Venkatesh Balavadhani Parthasarathy, Ahtsham Zafar, Aafaq Khan, Arsalan Shahid [26]	2024	Review	Provides an in-depth review of fine-tuning methods for Large Language Models, including supervised, unsupervised, and instruction-based approaches. The study highlights challenges like privacy and scalability and offers best practices for real-world implementation.
Alekya Jonnala [27]	2023	Industry Report/Exploratory	The paper highlights how LLMs enhance customer experiences by automating tasks, understanding user intent, and delivering

		Study	personalized responses. Integrating LLMs enables enterprises to optimize workflows, improve scalability, and boost customer satisfaction through intelligent, context-aware interactions.
Sangita Pokhrel, Swathi Ganesan, Tasnim Akther, Lakmali Karunarathne [28]	2023	Development/Applied Research	The paper focuses on using OpenAI, LangChain, and Streamlit to build chatbots for document summarization and question answering. This framework allows the creation of efficient systems that automate document processing and provide context-driven answers, enhancing productivity and saving time. The study concludes that integrating these tools helps build scalable, customizable chatbot solutions for enterprises.

This comprehensive approach ensures that our research adheres to best practices in systematic literature review methodologies, bolstering the rigor, transparency, and reliability of our study. The inclusion and exclusion criteria were thoughtfully defined, and their application was meticulous, based on strong and valid reasons. Figure 1 provides a comprehensive overview of the research framework presented in this paper. It encompasses the entire plan of the work, starting from the initial planning phase through to the selection and exclusion criteria for the literature review study, followed by the review and discussion of findings. Additionally, this figure highlights the scope of future work, offering a visual representation of the systematic approach undertaken in this research. This approach ensures the quality, relevance, and timeliness of the studies included in our systematic review, ultimately contributing to the robustness of our research findings.

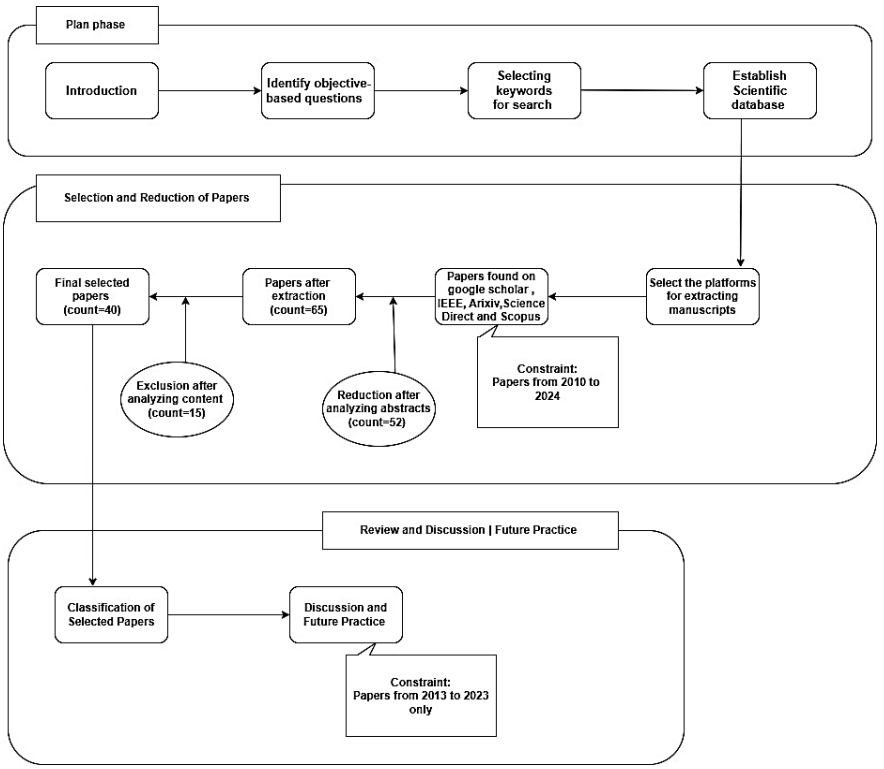

Fig. 1. Research framework overview: A visual representation of the research framework, spanning from initial planning to selection criteria, review, and future work scope.

4. AI's Role in Simplifying Document Analysis and Better Decisions

Advances in artificial intelligence (AI), especially large language models (LLMs) such as GPT-4 and Lama, are changing the way we approach complex tasks like document analysis and decision-making. These technologies have become essential in industries like healthcare, finance, and law because they reduce the time and effort required to process large amounts of data. LLMs automate document analysis, so businesses can gain insights more efficiently and make more informed decisions faster.

One of the key advantages of LLMs is their ability to understand and generate human language, which makes them highly effective at recovering unstructured data such as legal contracts, reports, and documents. For example, Ala-El-Din et al. [9] propose a chatbot that combines key performance indicators (KPIs) and logic models to help decision-makers process large amounts of data. Furthermore, structures such as PlanRAG [23] use LLM to derive meaningful results from documents, enabling more accurate decision trees.

Legal Analysis is a good example of how these models can add value. Other "problem-solving" tools similar to those used by Chakrabarti et al. [18] have also used AI to automatically identify risks in legal documents, achieving over 90% success. This type of automation saves time, minimizes fatal errors, and makes legal proceedings more effective and reliable. This is a clear example of how AI can take over tedious but important tasks.

Custom chatbots are also changing the way we work with documents. Nguyen-Trung et al. [17] showed how ChatGPT can speed up tasks such as literature searches, and tools such as LangChain [24] have shown that static documents can be transformed into interactive documents. These chatbots allow stoners to directly flesh out content and ask questions about it, simplifying the process of changing relevant information.

Healthcare is another field where LLMs could have a significant impact. A chatbot operated by LLMs, similar to the one by Alaaeldin et al. [12], aims to support real-time clinical decision-making and medical education, provide quick access to reliable information, and improve workflow and case problems. These tools will be essential for professionals who require fast and accurate recognition in high-pressure environments.

The Deep Knowledge Model further highlights the potential of AI in document analysis. Chakrabarti et al. [19] developed a system that efficiently classifies documents and processes them with remarkable sophistication. Such systems can automate repetitive tasks and allocate resources to other strategic tasks.

Despite these successes, challenges remain. Concerns about LLM’s isolation, bias, and high computational power requirements need to be addressed. Work such as “Smart Document Analysis with AI-ML” [25] shows the importance of developing integrative and scalable AI systems. Future research should focus on refining these models to meet specific requirements and finding ways to balance their comprehensive capabilities with the required complexity.

In summary, LLM is reinventing document analysis and decision-making to deliver faster, smarter, and more reliable results. Although there is still a lot of work to be done to address limitations, these technologies show great promise for the future and are actually helping businesses and professionals.

5. Using LangChain and RAG for Better Document Search and Answer Generation

In Figure 2, The process begins with raw documents, which may include unstructured or structured text such as reports, manuals, or articles. These documents are divided into smaller, more manageable chunks, making it easier to search and retrieve relevant information.

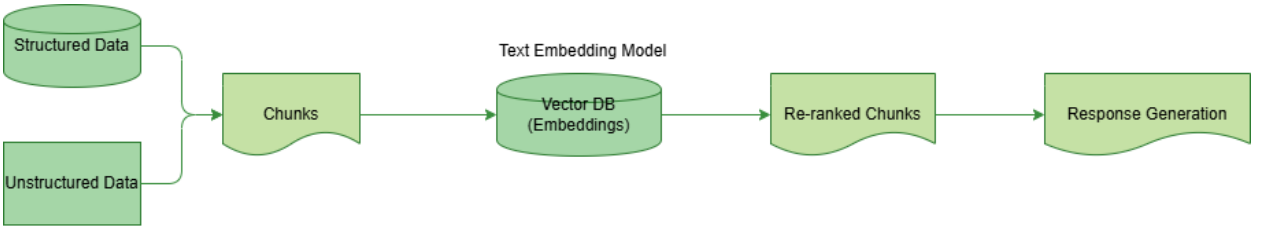

Fig. 2: Pipeline for Embedding-Based Information Retrieval and Response Generation.

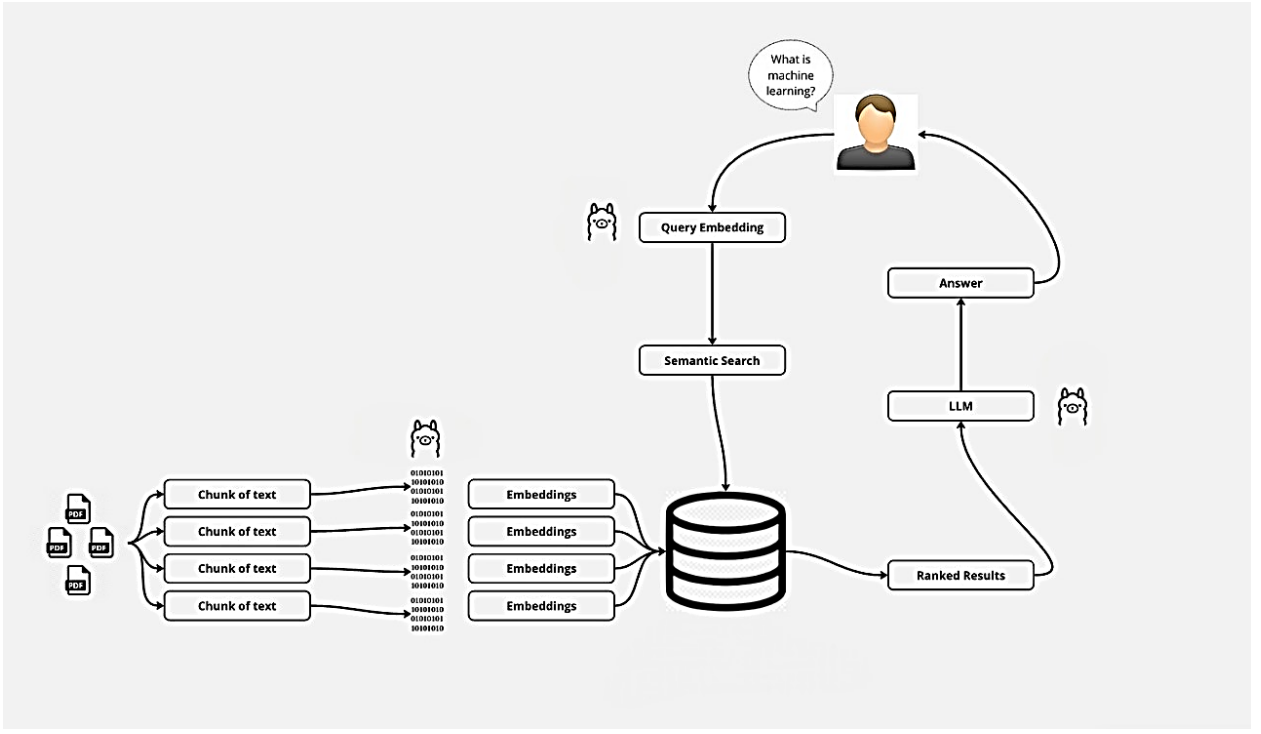

Fig. 3: The diagram of LangChain

An embedding model is then used to convert both the document chunks and the user's query into dense vectors, numerical representations within the same space. While the document chunks are embedded and stored during the initial setup, the user's query is embedded during runtime, allowing the system to focus on comparing meanings rather than relying on keyword matching. The system performs semantic search by comparing the query vector with the

stored chunk vectors in a vector database, retrieving a ranked list of the most relevant chunks using similarity metrics like cosine similarity. To refine the results, a re-ranking model further assesses and reorders the chunks based on their true relevance to the query, ensuring that the final list of chunks is as precise and useful as possible. The re-ranked chunks are then passed to the generation component of the RAG model, where they provide context for formulating the final response, ensuring the query is answered with the most relevant information [29][30].

In Figure 3, LangChain streamlines the question-answering process by efficiently handling large documents. Initially, these documents are divided into smaller sections, which are then transformed into vector representations, or embeddings, using a machine learning model. These embeddings capture the underlying meaning of the text and are stored in a vector database, allowing for quick and effective semantic searches based on content meaning rather than just keywords. When a user submits a question, such as "What is machine learning?", the query is also converted into an embedding. This embedding is compared to the stored ones to find the most relevant document chunks. The relevant chunks are ranked according to their connection to the question and passed to a large language model (LLM), which generates a clear and accurate response based on the retrieved data. The final answer, grounded in reliable and contextually relevant information, is then provided to the user [31][32].

6. Challenges of scaling chatbots powered by LLMs

Large Language Models (LLMs) have great potential for powering chatbots in document analysis and decision-making, but scaling these systems remains a major challenge:

6.1 Computational Costs and Resource Requirements

LLMs demand substantial computational power for training and deployment, making them inaccessible to smaller organizations with limited resources. High energy consumption associated with these models also raises concerns about environmental sustainability. Dreyling et al. [6] and Nguyen-Trung et al. [17] highlight the need for more efficient frameworks to reduce costs and environmental impacts. However, the paper lacks discussion on low-resource optimization or green AI initiatives to mitigate this issue.

6.2 Data Privacy Worries

LLMs rely on large datasets, and in sensitive fields like healthcare and finance, this can lead to vulnerabilities. Ensuring robust data protection and compliance with regulations like GDPR and HIPAA is critical [9][18].

6.3 Bias and Ethical Concerns

LLMs often inherit biases from their training data, leading to potential inaccuracies or unfair results, particularly in high-stakes fields like law and medicine. Barnea [11] and Salvagno et al. [14] emphasize that such biases can undermine trust in AI systems, especially given the "black-box" nature of LLMs, where decision-making

processes are opaque. While the paper mentions these issues, it does not provide concrete solutions for mitigating bias or improving transparency.

6.4 Domain-Specific Adaptation

General-purpose LLMs show limitations when applied to tasks requiring deep domain-specific expertise. For example, fine-tuning is often necessary for applications in legal and medical fields. Chakrabarti et al. [18] and Alaaeldin et al. [12] discuss the importance of incorporating domain-specific knowledge to enhance the performance of LLMs. However, this paper overlooks practical methods for adapting LLMs to specific industries.

6.5 Limited Real-World Applications and Evaluation

While the theoretical advantages of LLMs are well-documented, the paper lacks substantial real-world evaluations to validate its claims. Studies like Nguyen-Trung et al. [17] and Pokhrel et al. [28] highlight the importance of empirical testing and user feedback in refining LLM-based solutions. This paper's absence of case studies or pilot implementations weakens its practical relevance.

6.6 Scalability and Integration Challenges

Integrating LLMs into existing business workflows remains a significant challenge. Frameworks like LangChain and PlanRAG, discussed by Nguyen-Trung et al. [17] and Pokhrel et al. [28], offer promising solutions, but the paper does not adequately address how such frameworks can be scaled or customized for enterprise applications.

Addressing these limitations requires a balanced approach that combines advancements in AI with robust ethical guidelines, better security measures, and strategies to enhance transparency and scalability. Future research should focus on developing LLMs that are efficient, adaptable, and trustworthy for diverse real-world applications.

7. Discussion

The application of Large Language Models (LLMs) in document analysis has shown remarkable improvements in efficiency and decision-making across a wide range of industries. From the review, it is evident that these models significantly reduce the time and effort required for complex document analysis. For instance, tools like the "Risk-o-Meter" [18] achieve over 90% accuracy in identifying risks within legal documents, helping professionals make more informed decisions. Similarly, in healthcare, LLM-powered chatbots provide real-time clinical decision support, improving response times and patient care [12]. These examples illustrate how LLMs are streamlining workflows and enhancing decision-making reliability.

One of the standout benefits of LLMs is their ability to transform static documents into interactive resources. Chatbots powered by models like ChatGPT and LangChain [17][28] excel in summarizing complex documents and answering questions about them. These tools allow users to quickly extract relevant information and generate context-aware recommendations, making large datasets more manageable. Additionally, LLM-driven tools demonstrate high

accuracy in document classification tasks, as seen in studies achieving up to 95% accuracy [19]. This automation frees up human workers to focus on strategic and creative tasks, further boosting efficiency and satisfaction.

LangChain (*Figure 4*) and PlanRAG (*Figure 4*) represent two prominent frameworks for leveraging Large Language Models (LLMs) in document analysis and question answering. LangChain focuses on building dynamic pipelines that seamlessly integrate LLMs with external tools, allowing for flexible workflows. This framework excels in customization, enabling developers to link various components, such as retrieval models, APIs, and custom prompts, to suit specific use cases. Tools like LangChain have already demonstrated their effectiveness in transforming static documents into interactive resources, facilitating tasks such as summarization and content extraction efficiently [24].

On the other hand, PlanRAG enhances retrieval-augmented generation (RAG) processes by breaking complex queries into smaller, more manageable sub-tasks. This plan-then-retrieve approach prioritizes accuracy, particularly for multi-step reasoning and detailed information extraction [23]. Such frameworks are ideal for high-precision tasks that involve detailed decision-making processes, ensuring that relevant information is retrieved effectively and contextualized properly.

While both frameworks offer significant advantages, LangChain provides greater flexibility for developing end-to-end pipelines, making it ideal for projects requiring adaptability and integration with custom tools. In contrast, PlanRAG is better suited for applications that demand high accuracy in retrieval tasks [23, 24].

By leveraging LangChain with the LLaMA model, I can ensure a streamlined workflow that balances accuracy and efficiency. This combination provides robust capabilities for extracting insights, answering user queries, and analyzing complex documents while maintaining a flexible and scalable framework [4, 24].

However, while LLMs are promising, their application is not without significant limitations. The high computational costs associated with these models remain a barrier to adoption, especially for small and medium-sized enterprises. Dreyling et al. [6] and Nguyen-Trung et al. [17] emphasize the need for optimizing these systems to reduce resource demands and improve environmental sustainability, a gap that this paper does not thoroughly address. Additionally, concerns surrounding data privacy and security are critical. Chakrabarti et al. [18] and Alaaeldin et al. [9] highlight the vulnerabilities in sensitive industries like finance and healthcare, but the paper lacks detailed solutions for addressing these risks.

Bias and ethical considerations also present ongoing challenges. Studies by Barnea [11] and Salvagno et al. [14] underscore how biases in training data can lead to inaccurate or unfair outcomes, further exacerbated by the opaque decision-making processes of LLMs. While the paper acknowledges these concerns, it does not propose actionable strategies to mitigate bias or enhance transparency. Moreover, domain-specific adaptation is another limitation. General-purpose LLMs often require fine-tuning to meet the unique needs of specialized fields like law or medicine, as demonstrated by Chakrabarti et al. [18] and Alaaeldin et al. [12]. The lack of emphasis on domain-specific knowledge integration in this paper represents a notable gap.

Real-world validation and scalability are additional areas requiring improvement. While theoretical benefits are discussed, the absence of real-world evaluations reduces the paper's practical relevance. Nguyen-Trung et al. [17] and

Pokhrel et al. [28] stress the importance of pilot implementations and empirical testing to refine LLM-based systems for diverse industries. Additionally, frameworks like LangChain and PlanRAG [17][28] offer solutions for scalable and customizable integrations, but the paper does not adequately explore these options.

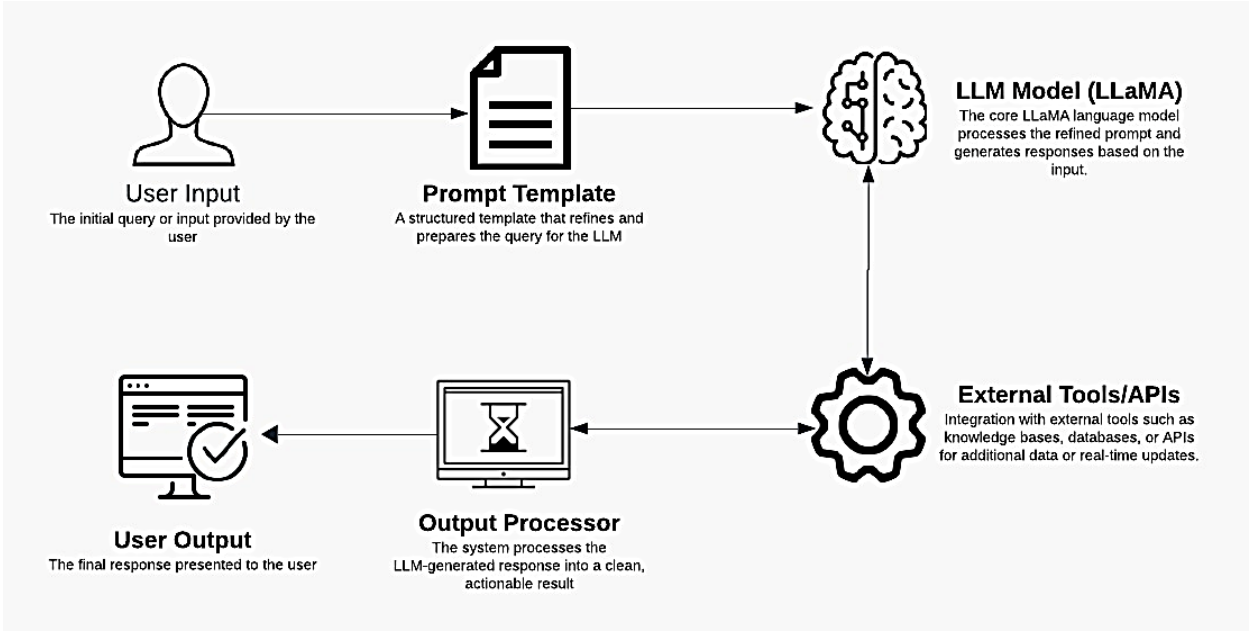

Fig. 4. LangChain: a block diagram shows the outlines of how LangChain works.

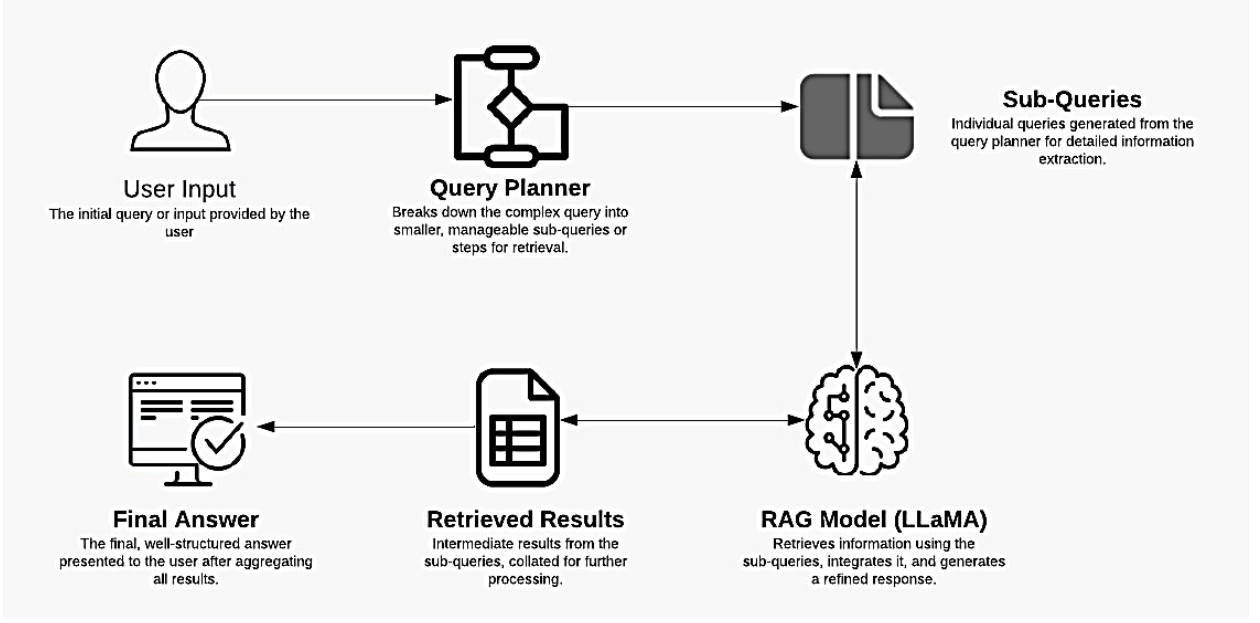

Fig. 5. Retrieval Augmented Generation (RAG): a block diagram shows the outlines of how (RAG) works.

For my project, which focuses on document analysis and question answering using the LLaMA model, LangChain emerges as the more suitable option. Its modular architecture enables efficient integration of LLaMA with retrieval models, ensuring faster development and customization. Additionally, LangChain's ability to connect external resources simplifies the process of handling large datasets, making it a more practical solution for achieving my project's goals.

In summary, while LLMs hold immense potential for revolutionizing document analysis and decision-making, addressing their limitations is crucial. Future work must focus on reducing computational costs, strengthening data security, enhancing transparency, and adapting these models to domain-specific needs. By prioritizing these areas, LLMs can become more accessible, reliable, and effective for real-world applications.

8. Conclusion

Large Language Models (LLMs) are proving to be transformative in document analysis and decision-making. These models enable industries to handle large amounts of data with remarkable speed and accuracy by automating complex tasks like summarizing documents, answering questions, and classifying information. Whether in legal, healthcare, or business environments, LLMs simplify workflows, extract valuable insights, and facilitate faster data-driven decisions.

However, significant challenges remain. Computational costs and high energy demands limit the accessibility of these models, especially for smaller organizations. Data privacy and security concerns, along with biases in training data, pose risks that need to be addressed with robust safeguards. The lack of transparency in LLMs further complicates their adoption, particularly in sensitive fields like law and medicine. Additionally, the absence of real-world validations and domain-specific adaptations highlights the need for further refinement and testing.

Despite these challenges, the potential of LLMs is undeniable. With ongoing advancements in optimization, ethical practices, and domain-specific fine-tuning, these models can become even more efficient, transparent, and adaptable. Future research should prioritize addressing these limitations to fully harness the capabilities of LLMs. By doing so, LLMs will continue to shape the future of document processing and decision-making, enabling smarter, faster, and more reliable outcomes for a wide range of industries.

References

[1] Fink, G. & Fink, G. n-Gram Models. Markov Models For Pattern Recognition: From Theory To Applications. pp. 107-127 (2014) https://www.researchgate.net/publication/311469848_Recurrent_neural_network_based_language_model

[2] Mikolov, T., Karafiat, M., Burget, L., Cernock ´ y, J. & Khudanpur, S. ` Recurrent neural network based language model.. Interspeech. 2, 1045- 1048 (2010) https://www.researchgate.net/publication/311469848_Recurrent_neural_network_based_language_model

[3] Scao, T., Fan, A., Akiki, C., Pavlick, E., Ilic, S., Hesslow, D., Castagn ´ e,´ R., Luccioni, A., Yvon, F., Galle, M. & Others Bloom: A 176b- ´ parameter open-access mult https://arxiv.org/html/2403.08730v2

[4] Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S. & Others Llama 2: Open foundation and fine-tuned chat models. ArXiv Preprint ArXiv:2307.09288. (2023) https://arxiv.org/pdf/2307.09288

[5] Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V. & Zettlemoyer, L. Bart: Denoising sequence-tosequence pre-training for natural language generation, translation, and comprehension. ArXiv Preprint ArXiv:1910.13461. (2019) https://arxiv.org/pdf/1910.13461

[6] Dreyling, Richard, et al. "Challenges of Generative AI Chatbots in Public Services-An Integrative Review." Available at SSRN 4850714 (2024). https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4850714
[7] Linnenluecke, M.K.; Marrone, M.; Singh, A.K. Conducting systematic literature reviews and bibliometric analyses. Aust. J. Manag. 2020, 45, 175–194. https://journals.sagepub.com/doi/10.1177/0312896219877678
[8] Moher, D.; Liberati, A.; Tetzlaff, J.; Altman, D.G.; PRISMA Group. Preferred Reporting Items for Systematic Reviews and Meta-Analyses: The PRISMA Statement. PLoS Med. 2009, 6, e1000097. https://journals.plos.org/plosmedicine/article?id=10.1371/journal.pmed.1000097
[9] DEVELOPING CHATBOT SYSTEM TO SUPPORT DECISION MAKING BASED ON BIG DATA ANALYTICS https://www.abacademies.org/articles/Developing-chatbot-system-to-support-decision-making-based-on-Big-Data-analytics-1532-5806-24-2-237.pdf
[10] The Future of AI Chatbots in Higher Education https://www.qeios.com/read/UE841K
[11] How will AI change intelligence and decision-making https://typeset.io/pdf/how-will-ai-change-intelligence-and-decision-making-11u95grlg4.pdf
[12] AI-Powered Chatbots in Medical Education: Potential Applications and Implications https://pubmed.ncbi.nlm.nih.gov/37692629/
[13] Evaluating User Experience With a Chatbot Designed as a Public Health Response to the COVID-19 Pandemic in Brazil: Mixed Methods Study https://humanfactors.jmir.org/2023/1/e43135/PDF
[14] Can artificial intelligence help for scientific writing? https://ccforum.biomedcentral.com/articles/10.1186/s13054-023-04380-2
[15] Large language models (LLM) and ChatGPT: what will the impact on nuclear medicine be? https://www.researchgate.net/publication/369117217_Large_language_models_LLM_and_ChatGPT_what_will_the_impact_on_nuclear_medicine_be
[16] Chatbots and Virtual Assistant in Indian Banks https://www.researchgate.net/publication/338659152_Chatbots_and_Virtual_Assistant_in_Indian_Banks#:~:text=Research%20paper%20concluded%20that%20Indian,on%20websites%20of%20the%20banks.
[17] Applying ChatGPT and AI-powered tools to accelerate evidence reviews https://www.researchgate.net/publication/370155174_Applying_ChatGPT_and_AI-powered_tools_to_accelerate_evidence_reviews
[18] Use of Artificial Intelligence to Analyse Risk in Legal Documents for a Better Decision Support https://www.researchgate.net/publication/370155174_Applying_ChatGPT_and_AI-powered_tools_to_accelerate_evidence_reviews
[19] Document Analyzing Using Deep Learning https://typeset.io/papers/document-analyzing-using-deep-learning-2bwxwu4n
[20] Artificial Intelligence and Decision-Making in Healthcare: A Thematic Analysis of a Systematic Review of Reviews https://typeset.io/papers/artificial-intelligence-and-decision-making-in-healthcare-a-2e1yy0ceem
[21] Strategic Framework for Leveraging Artificial Intelligence in Future Marketing Decision-Making https://typeset.io/papers/strategic-framework-for-leveraging-artificial-intelligence-37su6q5i03
[22] LARGE LANGUAGE MODELS AS OPTIMIZERS https://arxiv.org/abs/2309.03409
[23] PlanRAG: A Plan-then-Retrieval Augmented Generation for Generative Large Language Models as Decision Makers arXiv:2406.12430 https://arxiv.org/pdf/2406.12430
[24] Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit https://www.researchgate.net/publication/379654962_Building_Customized_Chatbots_for_Document_Summarization_and_Question_Answering_using_Large_Language_Models_using_a_Framework_with_OpenAI_Lang_chain_and_Streamlit#read
[25] Smart Document Analysis Using AI-ML https://www.researchgate.net/publication/335576686_Smart_Document_Analysis_Using_AI-ML
[26] The Ultimate Guide to Fine-Tuning LLMs from Basics to Breakthroughs: An Exhaustive Review of Technologies, Research, Best Practices, Applied Research Challenges and Opportunities arXiv:2408.13296v1 [cs.LG] 23 Aug 2024 https://arxiv.org/pdf/2408.13296
[27] How Large Language models (LLM) help enterprises enhance customer experiences Alekya Jonnala Software Development Manager in Amazon https://www.researchgate.net/publication/385696315_How_Large_Language_models_LLM_help_enterprises_enhance_customer_experieces_Alekya_Jonnala_Software_Development_Manager_in_Amazon
[28] Building Customized Chatbots for Document Summarization and Question Answering using Large Language Models using a Framework with OpenAI, Lang chain, and Streamlit https://www.researchgate.net/publication/379654962_Building_Customized_Chatbots_for_Document_Summarization_and_Question_Answering_using_Large_Language_Models_using_a_Framework_with_OpenAI_Lang_chain_and_Streamlit
[29] https://2024.conversations.ws/wp-content/uploads/2024/11/conv24_fp_28_olawore.pdf
[30] https://link.springer.com/chapter/10.1007/979-8-8688-0569-1_5
[31] https://www.researchgate.net/publication/372669736_Creating_Large_Language_Model_Applications_Utilizing_LangChain_A_Primer_on_Developing_LLM_Apps_Fast
[32] https://insights2techinfo.com/wp-content/uploads/2023/12/langchain-with-dsim-format_revised-new.pdf