الجمهورية الجزائرية الديمقراطية الشعبية
The People's Democratic Republic of Algeria
وزارة التعليم العالي والبحث العلمي
Ministry of Higher Education and Scientific Research
جامعة محمد بوضياف بالمسيلة
University Mohamed Boudiaf of M'sila
كلية الرياضيات والإعلام الآلي
Faculty of Mathematics and Informatics
قسم الإعلام الآلي
Department of Computer Science

Domain: Mathematics and Computer Science

Thesis Presented to Fulfill the Partial Requirement
for **Master's Degree** in Computer Science

Specialty: Information System and Software Engineering (ISSE)

Prepared By: Safa Chebbih , Sana Amri

Supervised By:

Debbi Hichem

ENTITLED

LLM Powered Intelligent Document Chatbot

Jury Members

Mustapha Bourahla	President
Hichem Debbi	Supervisor
Said Hamani	Examiner

Academic Year 2024/2025

Dedication

*After expressing my heartfelt gratitude to my family, particularly to my **mom** and **dad** for their unwavering support and countless sacrifices, I want to say, I would not have reached this point without your guidance and love. I always wanted to make you proud of me. To the rest of my beloved family, my two brothers, my sister Shada, my friends Assil and Noor, and my two little sisters, Mariam and Iman, thank you for being part of this journey with me. There are two people I truly, deeply want to mention. The first, mentioning her brought me to tears. She was my best friend, my dearest companion, the one I missed so much to you **AYA REBHOUI**. The one who shared memories, dreams, and plans with me. Thank you for supporting me, emotionally and encouraging me to complete this work . Now that you are in the care of Allah, I pray we will meet again in Jannah. Thank you, my dearest, the sister I could never have imagined meeting or having. The second person, mentioning her brought me laughter and warmth. Thank you for supporting me and helping me from the very beginning of this journey until now. Thank you for standing by me during my downfalls and celebrating with me in my happiest moments. You were there for me in every way, as a friend and a sister . **NENO**, you will always be the best.*

Chebbih safa

To those who played a role after Allah in what I have achieved, to my dear family who have been my true support at every step, to my caring mother with her prayers and the affection that enveloped me at every moment, and to dear father for his support and belief in me, and to my beloved sister who has been my safety support and joy, and to the pure spirit of my grandfather, may Allah rest his soul, the first to encourage me and believe in my ability, whose dream was for me to continue my education and achieve success. To all of you, I dedicate the fruit of this effort with all the gratitude and pride, for you are an inseparable part of my soul, and with you I would not be here today.

AMRI Sana

Acknowledgement

I would first like to thank **ALLAH** for His guidance to me throughout my journey to complete this work and for granting me the ability, patience, time, and health. Praise be to God for everything.

I would like to express our deepest gratitude to our advisor **Dr. Hichem Debbi**

for his guidance and constant encouragement through the duration of this project.

Also, we are extremely grateful to **Sonelgaz** group for granting us the access to their documents and permitting us to use them, and to their support to us by providing us the necessary information, we needed to complete this work.

We wish to express our heartfelt appreciation to all our professors whose knowledge, expertise, and passion have enriched our learning experience over the years.

Lastly, I'd like to thank all my familly , my friend and all my supporters who have motivated me , encourage me and support me through this work.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

AI Artificial Intelligence.

ML Machine Learning

DL Deep Learning

NLP Natural Language Processing

DNN Deep Neural Network

CNN Convolutional Neural Network

RNN Recurrent Neural Network

LSTM Long Short Term Memory

LLMs Large Language Models

GEN AI Generative AI

RAG Retrieval Augmented Generation

API Application Programming Interface

HyDE Hypothetical Document Embeddings

General Introduction

Companies' internal documents can contain a treasure trove of information in an era where information only seems to be growing, and the ability to access proprietary knowledge contained in hundreds or thousands of large files and quickly find specific bits and pieces of information is becoming increasingly important.

This rapid growth of in-document volume and complexity in the business domain necessitates advanced automated methods to enhance the efficiency and accuracy of information extraction and analysis, especially when it involves how employees handle the overwhelming volume of files that contain a lot of complex information, requiring them to scroll and search, and expecting them to understand those documents by a certain deadline in order to perform tasks, most important writing, reports, contracts, and other business documents.

This work aims to efficiently reduce the pressure employees feel while managing and consulting these documents and to ensure that they can retrieve accurate, context-aware information for them.

In order to form a clearer vision, nothing is more enlightening than the experience of the project in the Algerian context. By taking Algeria's companies as a case study, especially after seeing how Algeria started to support the adoption of artificial intelligence and its efforts to introduce it in various fields.So that even companies have turned towards investing and encouraging young entrepreneurs to use and integrate artificial intelligence into their projects. Nevertheless, it is worth noting that Algeria is still in its early stages in this field and faces numerous challenges in its development.

Objectif

Our objective is to build a chatbot powered by open source Large language models that can facilitate the interaction between the employees and their company's documents through the use of natural language as query so that the LLM will generate the right answer back to the employee.

manuscript structure

This dissertation is divided into four main chapters organized as follows :

- In the first chapter, we will see definitions of Artificial Intelligence in this field, Machine Learning, Deep Learning, NLP,showing how large language models integrate these technologies, especially NLP, to create a major revolution in the world of Artificial Intelligence technology.

- Chapter two provides a deeper insight into Large-Angle Modules (LLMs) for their structure to their shortcomings and their relation to generative AI. Then we will present retrieval techniques, mechanisms that will enhance the power and accuracy of LLMs through external sources.
- Chapter three is a broader introduction to understanding companies and their internal documents , as well as the contributions of LLM in the success of some companies that have adopted it.
- Chapter four involves the implementation and evaluation of this work, showing the challenges and the results obtained.

CHAPTER 1

The hierarchical relationship between AI and its fields

1. Introduction

Artificial Intelligence (AI) is a field of computer science, and new large-scale AI models have been released frequently, especially due to advancements in Machine Learning (ML), Deep Learning (DL), and Natural Language Processing (NLP). This chapter provides an overview of these technologies and explains how AI, ML, DL, and NLP work together to achieve human-like understanding and generation of language at scale, as demonstrated by Large Language Models (LLMs).

2. Artificial intelligence

Artificial Intelligence (AI) refers to the simulation of human intelligence in machines that are programmed to think and act like humans. It involves the development of algorithms and computer programs that can perform tasks that typically require human intelligence such as visual perception, speech recognition, decision-making, and language translation. [1]

3. Machine learning

Machine learning is a subset of Artificial Intelligence (AI) that enables computers to learn from data and make predictions without being explicitly programmed. If you're new to this field, this tutorial will provide a comprehensive understanding of machine learning, its types, algorithms, tools, and practical applications.[2]

3.1. Types of Machine learning algorithms

Machine learning has three key types, each with numerous algorithms that apply to it.

3.1.1. Supervised Learning

Supervised learning algorithms use labeled data to predict outputs for new data. Here’s a brief summary of common algorithms:

- **Linear Regression:** Predicts continuous outcomes by modeling the relationship between variables with a linear equation.
- **Logistic Regression**: Used for binary classification, it estimates probabilities for yes/no outcomes using a logistic function.
- **Decision Trees**: Makes predictions by learning simple decision rules based on input features.

- **Random Forests**: An ensemble of decision trees, improving prediction accuracy and controlling overfitting.[3]
- **Support Vector Machines (SVM):** A powerful classifier, effective in high-dimensional spaces, also used for regression.[3]
- **Neural Networks**: Captures complex non-linear patterns in data, commonly used in deep learning applications.

3.1.2. Unsupervised learning

Unsupervised learning algorithms are used with data that doesn't have labeled responses, aiming to uncover the inherent structure in the data. Key unsupervised learning techniques include:

- **Clustering:** Groups objects based on similarity. Common algorithms are K-means, hierarchical clustering, and DBSCAN.
- **Association:** Finds patterns or rules in data, such as those used in market basket analysis.
- **Principal Component Analysis (PCA):** A statistical method that transforms correlated variables into a set of uncorrelated variables, simplifying the data structure.[3]
- **Autoencoders:** A type of neural network that learns efficient representations (or codings) of unlabeled data.[3]

3.1.3. Reinforcement learning

Reinforcement learning algorithms are designed to help agents make a sequence of decisions to achieve a goal in uncertain environments. The agent learns from the results of its actions through rewards or penalties. Key reinforcement learning techniques include:

- **Q-learning:** A model-free algorithm that learns the value of actions in specific states.
- **Deep Q-Networks (DQN):** Combines Q-learning with deep neural networks, enabling the agent to learn policies from complex inputs.[3]
- **Policy Gradient Methods:** Directly optimize the policy's parameters instead of estimating action values.[3]
- **Monte Carlo Tree Search (MCTS):** Finds optimal decisions by simulating possible scenarios, commonly used in games like Go.[3]

4. Deep learning

Deep learning (DL) is a sub-field of Machine Learning (ML) , that has the ability to build a model that learns to perform classification tasks based on images or texts using multilayered neural networks called deep neural networks.

4.1. Deep Neural Networks(DNN):

A Deep Neural Network (DNN) also known as DeepNet, is a type of Artificial Neural Network (ANN) that consists of multiple layers between the input and output layers. These intermediate layers are referred to as hidden layers. Each layer comprises multiple node neurons that perform computations on the input data. In a DNN, the input layer consists of nodes that represent features or variables from the input data. The hidden layer processes the data through transformations by applying weights and biases, followed by an activation function. This mechanism enables the network to learn features and captures complex patterns and representations of the input data. [4] The output layer is the final layer of the network, which produces the output or prediction. The number of nodes in this layer depends on the task that the network is designed performs.

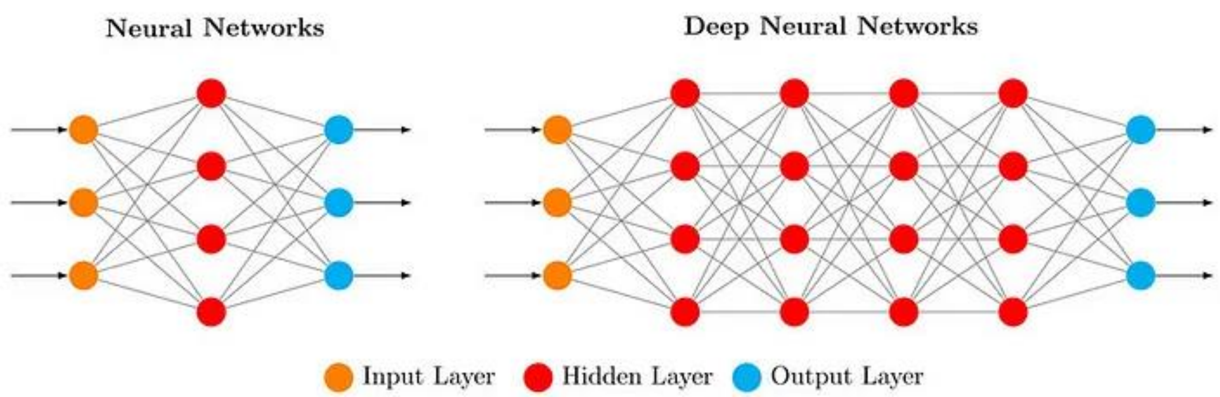


Figure 1.1: neural network vs DNN [4]

4.2. Convolutional Neural Networks CNN

A Convolutional Neural Network CNN is a special model of feed-forward neural networks used in different fields such as: computer vision CV, recommender systems RS, and Natural Language Processing NLP. It is a deep neural network architecture. A CNN network is composed of several layers of different types: convolutional layers, sampling layers (pooling or subsampling layers) and a fully-connected layer. The role of the convolution layers is to filter the data they receive to extract the most relevant elements, the role of sampling layers (Pooling Layers) is to reduce the resolution (the dimension) of the elements selected by the layers of convolution which reduces complexity, avoids the phenomenon of overfitting and therefore increases the robustness and performance of the CNN network, especially in the face of perturbations.

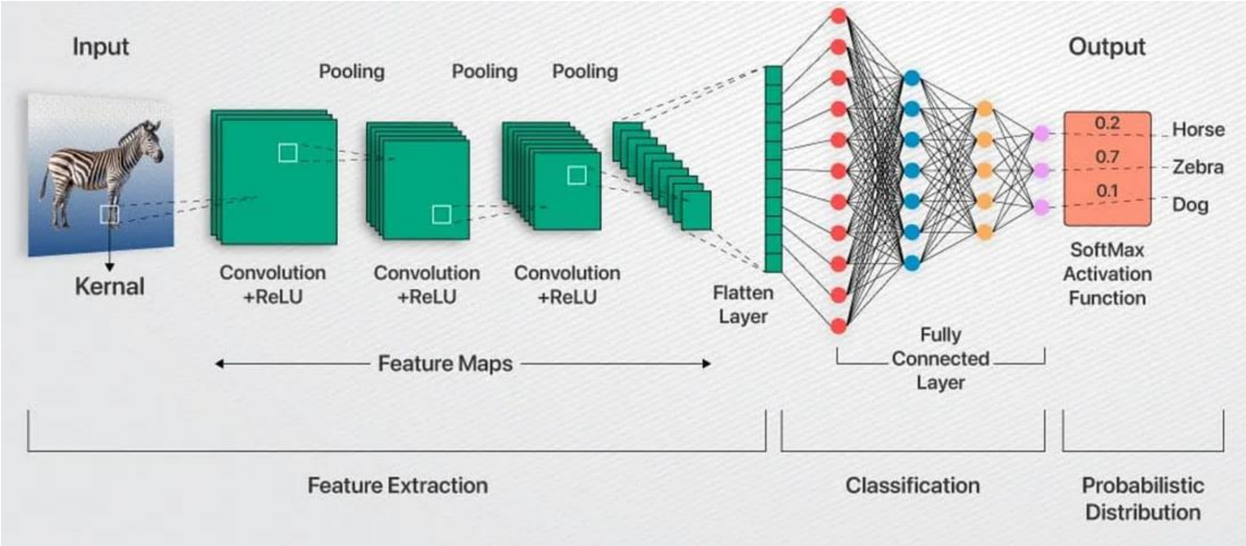


Figure 1.2: Convolutional neural network CNN [5]

4.3. Recurrent Neural Network(RNN)

RNNs are a type of artificial neural network designed to process sequential data by allowing information to persist through loops within the network architecture. "The key component of an RNN is recurrent connection, which is connection within a neuron that allows information to flow from one time step to next." [6]

An RNN can be divided into three main parts:

- **Input layers:** The input layers takes in the information at each time step, such as a word in sequence.
- **Recurrent layer:** processes the information from the input layer using connections to remember information from previous time steps. The recurrent layer contains a set of neurons, each with a recurrent connection to itself and a connection to the input at the recurrent time step.[6]
- **The output layer:** generates a prediction based on the information processed by the recurrent layer. In case of generating the next word in a sequence, the output layer will predict the most likely word to follow the previous word in sequence.[6]

Traditional RNNs struggle to keep information over long sequences because of vanishing gradient problem, which happens during training. As the gradient can become very small as they are propagated backward through time, making it difficult for the next network to learn long-term dependencies.

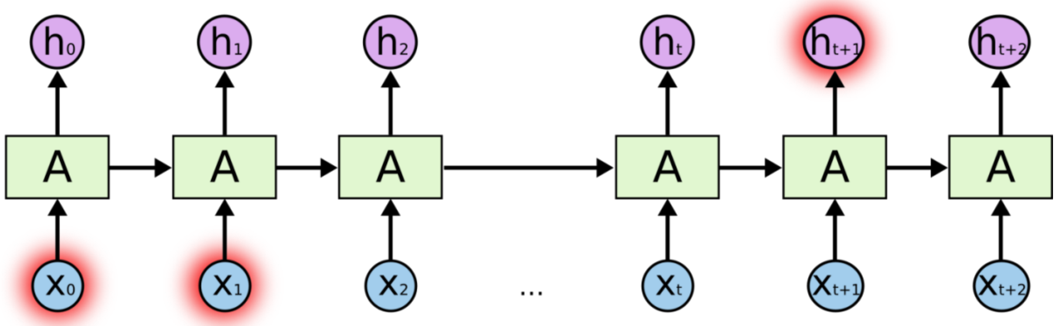


Figure 1.3: Diagram of RNN architecture [7]

4.4. Long short-term memory (LSTM):

LSTMs are special kind of RNN capable of learning long-term dependencies. They work tremendously well on a large variety of problems and are now widely used.LSTMs have a chain-like structure which consists of four neural network layers interacting in a very special way as represented in figure 4.3.[7]

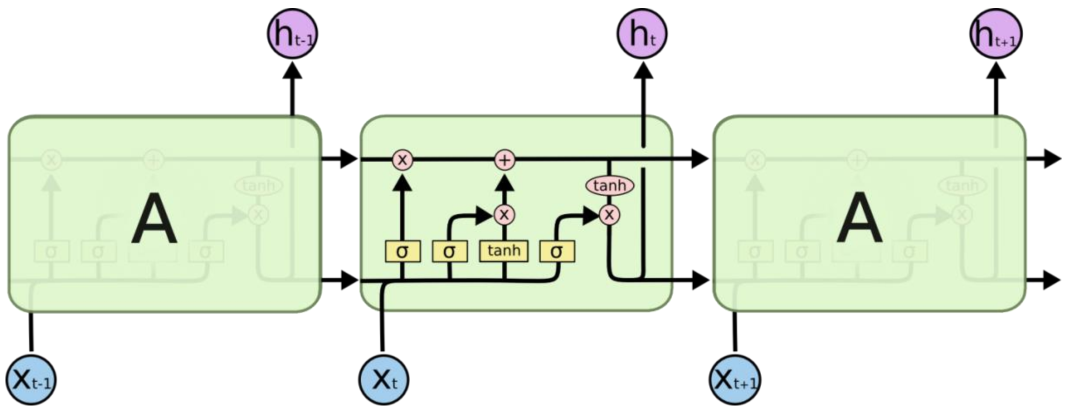


Figure 1.4: The repeating module in an LSTM contains four interacting layers [7]

LSTM can remember information for a long period of time, which is particularly their default behavior, that's what gives them the capability to avoid long-term dependencies.[7]

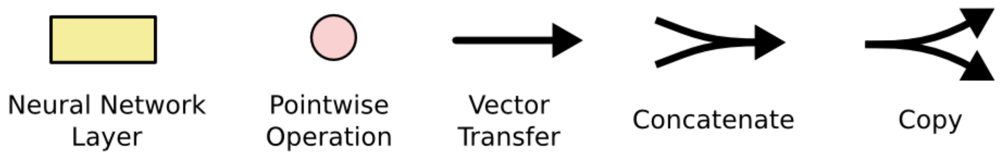


Figure 1.5: LSTMdiagram notation guide[7]

In the above figure , each line carries an entire vector, from the output of one node to the inputs of others. The pink circles represent pointwise operations, like vector addition, while the yellow boxes are learned neural network layers. Lines merging denote concatenation, while a line forking denote its content being copied and the copies going to different locations.[7]

5. Natural language processing

Natural Language Processing (NLP) is an Artificial Intelligence (AI) sub-field that gives the capacity to machines understand and handle human languages, which can be in the format of text or audio.

5.1. Natural language processing key components

Natural Language Processing (NLP) has several key components that enable machines to understand and process human language. The main components include:

- **Tokenization** : is the process of splitting or breaking text into smaller units, such as words or phrases, called tokens.
- **Morphological analysis :** is the study of how words are built from smaller meaningful units, which involves identifying word structures such as prefixes, suffixes, and root words.
- **Part-of-Speech (POS) tagging** : is the process of assigning grammatical labels(e.g.noun , verb, adj) to each word in a sentence, which helps in understanding the syntactic structure of sentences.[8]
- **Named Entity Recognizers :** are evaluated using recall, precision, and the F1 measure. It involves identifying and classifying entities in text, which is useful for extracting structured information from unstructured text.[8]
- **Syntactic parsing ;** is the process of assigning a syntactic structure to a sentence. It helps break down sentences into subphrases and constituents, making it easier to analyze their grammatical structure.[8]
- **Semantic analysis :** involves understanding the correct meaning of words and sentences by identifying the roles of words within a sentence (e.g., agent, patient).[8]
- **Sentiment analysis :** classifies text based on the positive or negative orientation (sentiment) expressed by the writer toward an object. It is used in applications such as customer feedback analysis and social media monitoring. The process often relies on algorithms like Naïve Bayes.[8]
- **Coreference resolution :** is the process of identifying which words or phrases refer to the same entity in a text. Its purpose is to improve the understanding of context and relationships within the text.

- **Machine translation :** is the automatic process of translating text from one language to another using a fixed vocabulary in advance, enabling cross-lingual communication and content localization.[8]
- **Speech recognition :** and synthesis involve converting speech to text (Automatic Speech Recognition, ASR) and text to speech (Text-to-Speech, TTS).

5.2. Natural Language Processing Importance

Natural Language Processing (NLP) is a crucial field in AI and computer science that focuses on enabling machines to understand, interpret, and generate human language. The importance of NLP can be seen in several key areas:

- **Human-computer interaction:** NLP allows users to communicate with technology in a more intuitive and conversational manner, enhancing user experience and accessibility.
- **Data analysis:** NLP plays a significant role in sentiment analysis, topic modeling, and text classification, helping to extract meaningful insights from vast amounts of unstructured text data.
- **Business intelligence:** NLP helps businesses analyze customer feedback, social media trends, and market research, enabling data-driven decision-making.
- **Accessibility:** Technologies like speech-to-text and text-to-speech assist individuals with disabilities, making digital content more accessible and improving inclusivity.
- **Content creation:** NLP enhances the efficiency and creativity of content generation, helping produce relevant, high-quality text automatically.
- **Research and development:** NLP supports various fields of research, facilitating advancements in linguistics, cognitive science, and AI technologies.
- **Future-proofing technology:** As AI continues to evolve, NLP will play a critical role in shaping future technological developments, ensuring systems are capable of adapting to and understanding human language.

6. AI and LLMs

Large Language Models (LLMs) are a subfield of Artificial Intelligence, AI, focused on developing systems that perform specific tasks while trying to simulate human thinking such as generate human like texts. These models utilize machine learning techniques, particularly deep learning, to analyze and process vast amounts of data, including books, articles, and web pages. This capability enables LLMS to unlock numerous and often unexpected possibilities in various AI applications.[9] The core deep learning architecture used in LLM is known as the transformers, which are incredible at understanding relationships within sequential data. And they are particularly effective at processing and understanding the structure of language, making them essential for NLP, which is a field that seeks to develop machines that comprehend and produce human language.

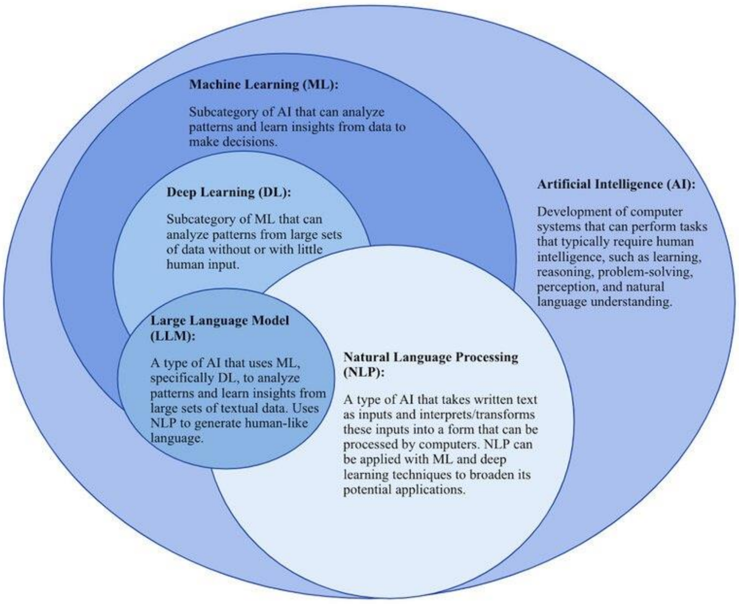


Figure 1.6: Hierarchy of AI from machine learning to LLMs[10]

7. Conclusion

In this chapter, we discussed an overview of AI and its various branches , including their interdependent relations which uplift every particular area. In a world where AI technology is continuously improvised, these modifications will allow advances in machine learning, deep learning and natural language processing to make text synthesizing and comprehension much more sophisticated.

CHAPTER 2

Fundamental concept of LLMs And RAG Mechanisms

1. Introduction

Large Language Modules (LLMs) represent a revolutionary breakthrough in natural language processing (NLP), enabling machines to understand human language and generate text or other content with better accuracy.Powered by deep learning (DL) techniques especially transformers, significantly blurring the line between human and machine interactions, making it increasingly difficult to distinguish whether you are conversing with a human or an AI-powered system (chatbots).

Now we will introduce you to the most important elements and concepts of the LLMs and how it turns into Gen AI and finally how RAG fills the gaps in LLMs, as the combination of these technologies generates smart systems that have the potential to support institutions in providing smart and effective solutions, especially in providing smart, structured and effective answers in several fields such as customer service, especially answering questions.

2. Large Language Modules (LLMs)

Large Language Models (LLMs) are a type of Artificial Intelligence (AI) algorithm that applies neural network techniques, particularly deep learning in neural networks, to recognize, extract, summarize, predict, and generate texts based on knowledge, accurately retaining on a vast dataset. They are specialized subsets of broader technology known as Language Models, all of which share a common purpose, performing tasks related to natural language processing (NLP). “LLMs differ from traditional NLP models in several key aspects, most notably their size. Their large scale is due to two primary factors. They are trained on massive amounts of data and they comprise a huge number of learnable parameters.” [11]

2.1. Brief history about LLMs

To better understand LLMs, this section will follow the developmental stages of LMs and introduce SLMs, NLMs, PLMs, and LLMs. Figure 1 provides a visual map of the history of the development of the LMs.

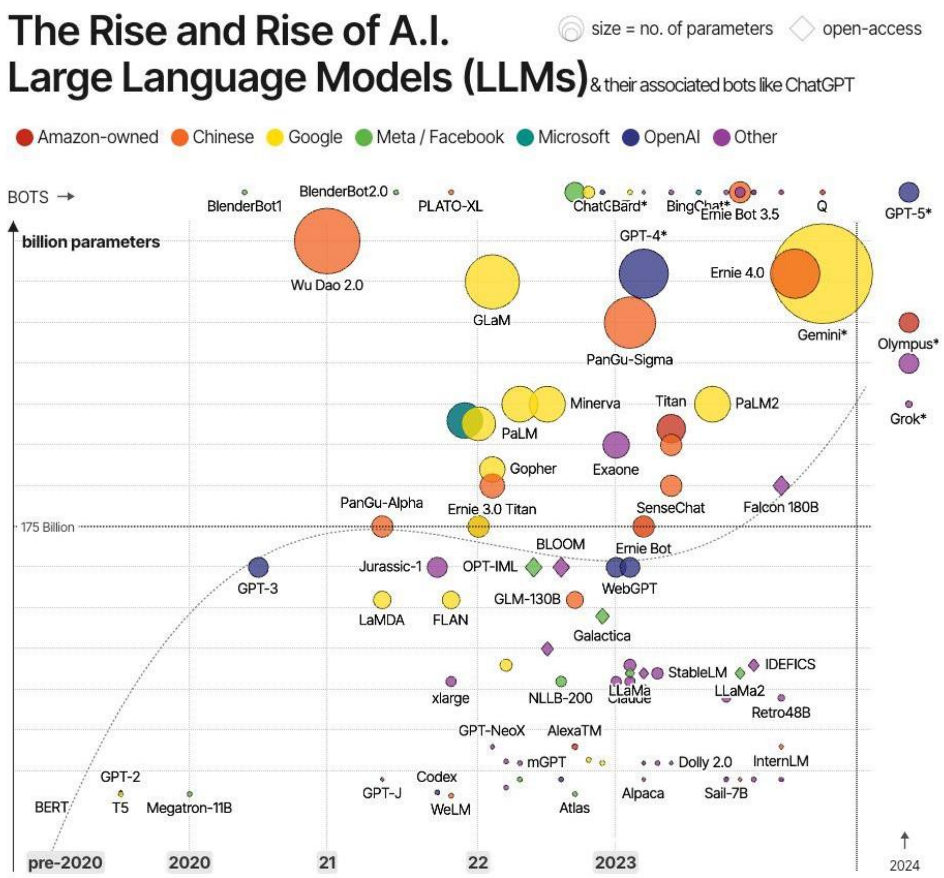


Figure 2.1: History and development of language models[12]

• **Statistical Language Models (SLMs)(1990s):** Emerged in the 1990s as mathematical models that address contextually relevant properties of natural language from a probabilistic statistical perspective. [13] These models aim to discover patterns and relationships within the text, making them more flexible and accurate in understanding language. This technique relies on calculating the probabilities of word sequences, such as N-grams.

• **Neural language models （NLMs）(2000s) ：**In the year 2000, Yoshua Bengio introduced neural networks for language modeling. [14] These neural networks rely on predicting the probabilities of the next words within sequences and analyzing texts more intelligently, as seen in Word2Vec and RNN/LSTM. They are capable of efficiently processing long sequences and mitigating the limitations associated with small n in Statistical Language Models (SLMs).

• **Pre-trained language models (PLMs) (2017s)**:Rely on pre-training using a large volume of unlabeled text, enabling them to understand fundamental language structures such as vocabulary, syntax, semantics, and logic.[13] Examples include BERT, GPT, and GPT-2.

• **Large language models (LLMs) (2020s):** LLMs are trained on massive text corpora with tens of billions (or more) of parameters, such as GPT-3 , GPT-4 . The goal of LLMs is to enable machines to understand human command[14]

2.2. How do large language models work

Large language models operate by deep learning principles and are trained on vast datasets comprising books, articles, and conversations. These models are typically based on Transformer architecture, which can capture contextual relationships between words.This enables LLMs to develop a comprehensive understanding of sequence structure and word meaning in context.[15]

During training, the model learns to predict the next word in a sentence by adjusting its internal parameters with each iteration to minimize the difference between its predictions and actual outcomes. Through millions or even billions of training steps, the model gradually improves its ability to generate coherent and contextually relevant texts. [15]

Additionally, LLMs can be fine-tuned on smaller domain-specific data sets to enhance their performance for specialized tasks. Fine-tuning defines the model's knowledge, allowing it to excel in specific applications such as medical diagnoses, a general-purpose.This process transforms a general-pupose LLM into an expert system tailored for precise and accurate task execution.

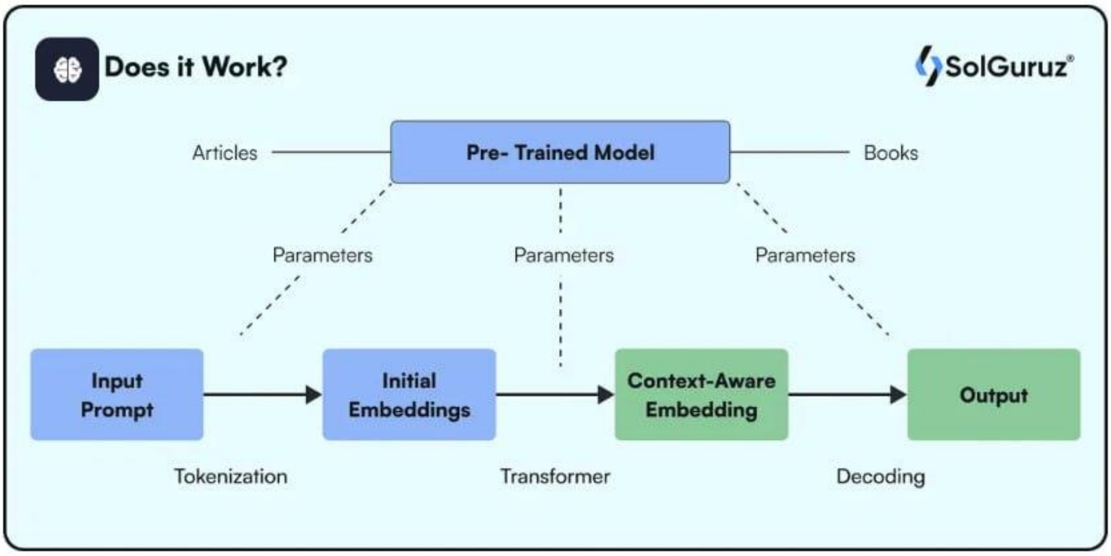


Figure 2.2: Transformer-based models [16]

2.3. Types of large language models

Here are the Types of Large Language Models:

- **Multimodal:** is a LLM has the capacity to handle both text and images, making a robust language model. These models can understand the relationship between the images and the text, allowing them to generate and process context content in multiple modalities. For example, CLIP and DALL-E.[17]
- **Pre-trained models:** are machine learning models which are trained on a vast amount of data, allowing them to learn a wide range of language patterns, features, and structures. These

models excel at generating coherent and grammatically correct text on a variety of topics. They can also be fine-tuned for specific tasks, for example, GPT-3,T5, and GPT-3.5.

- **Fine-tuning models or domain-specific:** They are models pre-trained on large datasets and adapted to specific tasks by fine-tuning it into smaller datasets. These models' goal is to make use of the knowledge they They already have learned and specialize it for a particular application, for example, BERT, ROBERTa, and ABERT.
- **Zero shot:** They are standard LLMs trained on generic data to provide reasonable, accurate results for general use cases[18]. These models perform tasks relying only on their pre-trained knowledge, as they are for immediate use.

2.4. Architecture of large language models

The general architecture of LLMs consists of many layers such as the feed forward layers, embedding layers, attention layers. These are important components to influence large language model architecture:[19]

- **Model size and parameters count :** LLMs are defined by their size measured in parameters (learnable weights and biases in the model) and its effects the performance of the model Which It doesn't just depend on the model size or the data quantity. The quality also matters.
- **Input representation :** defines how input text is converted into numerical vectors that the model can process. It consists of two main components: tokenization and embeddings. These components enhance the model's ability to learn and understand semantics more effectively.
- **Self attention mechanisms .**
- **Training objectives:** define how a model learns during training. The choice of objectives determines the type of knowledge the model acquires during pre-training. Here are some commonly used objectives:[19]
  - ➢ Masked Language Modeling(MLM) : Predict Masked Token in a Sentence(BERT).
  - ➢ Causal language modeling (CLM) :Protects the next token in sentence(GPT).
  - ➢ Sentence-to-sentence learning : Translates input sequences into output sequences.
- **Computational efficiency :** The ability to train and deploy large models efficiently. Balancing performance and resource usage. For example, getting better performance with fewer resources. This can be Achieved using such as : [19]
  - ➢ **Space attention :** Reduces computation complexity by limiting the number of tokens considered in attention computations.
  - ➢ **Distillation :** Transfers or comprises knowledge from large model to smaller model.
  - ➢ **Mixture of experts(MOE):** Divides the model into smaller experts and activates only a subset during inference.

- **Decoding and output generation:** LLMs are typically built using deep learning architectures, most commonly based on the transformer model.

2.5. Transformers based architecture

The transformer model is a deep neural network designed for processing sequences of text. It's more efficient and faster than earlier models like Recurrent Neural Networks and Long Short Term Memory networks. The transformer's architecture was first introduced in the paper " Attention is All You Need" in 2017. Initially, it was designed for language translation tasks. However, the developers at OpenAI later discovered that this architecture was highly effective and crucial for word prediction, making it a foundational breakthrough for the modern large language model. [19,20]

LLMs follow a general architecture that includes the following key components: [19,20]

A. Input embedding : The input text is tokenized into smaller units and embedded into continuous vector representations. This step captures both the semantic and syntactic information of the input.

B. Positional encoding : Since transformers do not naturally encode the order of tokens, positional encoding is added to each token's input embedding. This assigns a unique representation to each position, allowing the model to understand sequence order.

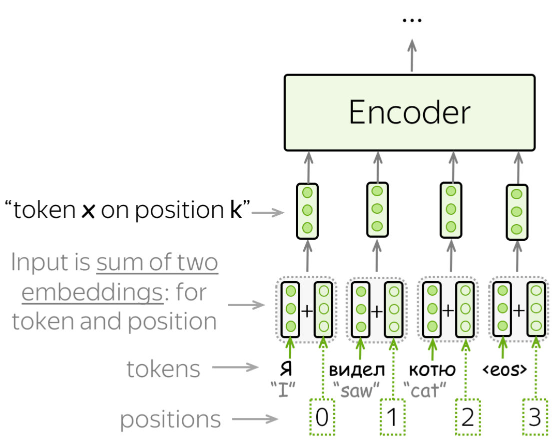


Figure 2.3: positional encoding [21]

C. Encoder : It is based on neural network techniques, processes the input text and generates hidden states that perceive the context and the meaning of the text data. The encoder consists of multiple layers, each containing :

- **Self-attention mechanisms:** Computes attention score to determine the importance of different tokens in the input sequence, allowing the model to capture dependencies and relationships between words in context-aware manner.

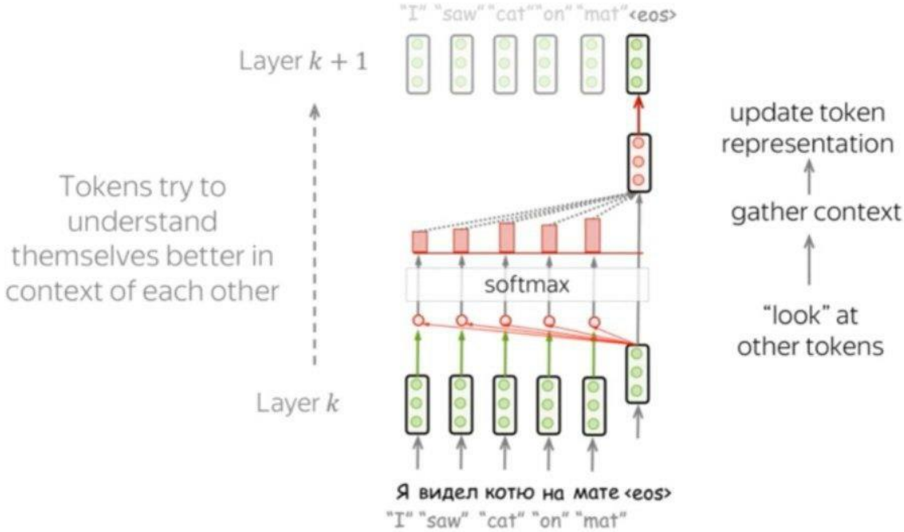


Figure 2.4: Self-attention in transformer encoders [21]

- **Feed-forward neural networks:** Apply to each token independently after the self-attention step. This network consists of a fully connected layer with a non-linear activation function, helping the model capture complex interactions between tokens.

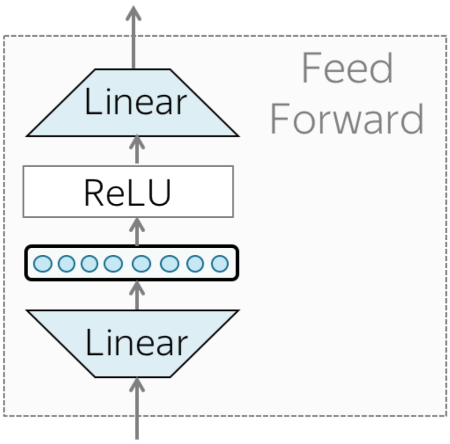


Figure 2.5: Feed forward diagram [21]

D. Decoder layer : Some transformers-based models include a decoder component in addition to the encoder. The decoder enables autoregressive generation, allowing the model to generate sequential outputs by attending to previous generated tokens.

E. Multi-head attention: This mechanism consists of multiple attention heads that process the same input independently, each with its own set of weight. After computation, the outputs from all attention heads are concatenated and passed through a linear layer, enhancing the model's ability to capture different types of relationships and attend to various types of the input sequence simultaneously.

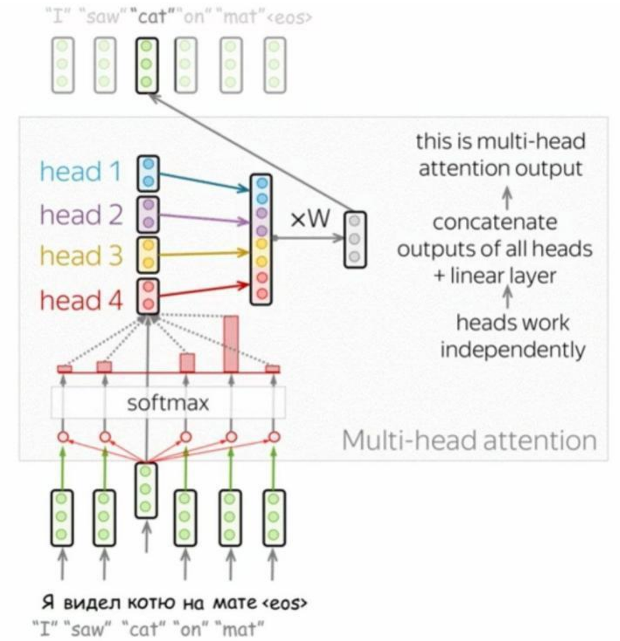


Figure 2.6: Multi-head attention mechanism in transformers [21]

F. Layer normalization: Applied after each subcomponent or layer. In the transformer, architecture layer normalization stabilizes the learning process and improves generalization across different inputs.

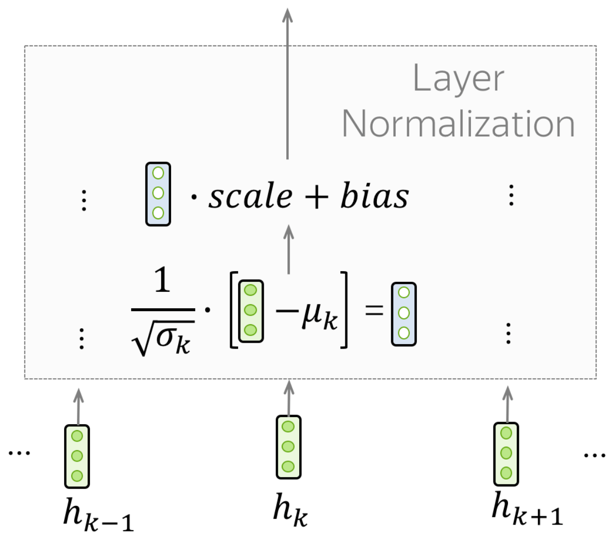


Figure 2.7: Normalization layer [21]

G. Output layers : The output layers of the transformance model can vary depending on the specific tasks. For example, in language modeling, linear projection followed by softmax activation is commonly used to generate the probability distribution over the next token.

The architecture of transformer-based model can be modified and enhanced based on specific research objectives and model designs to achieve different tasks and improve performance.

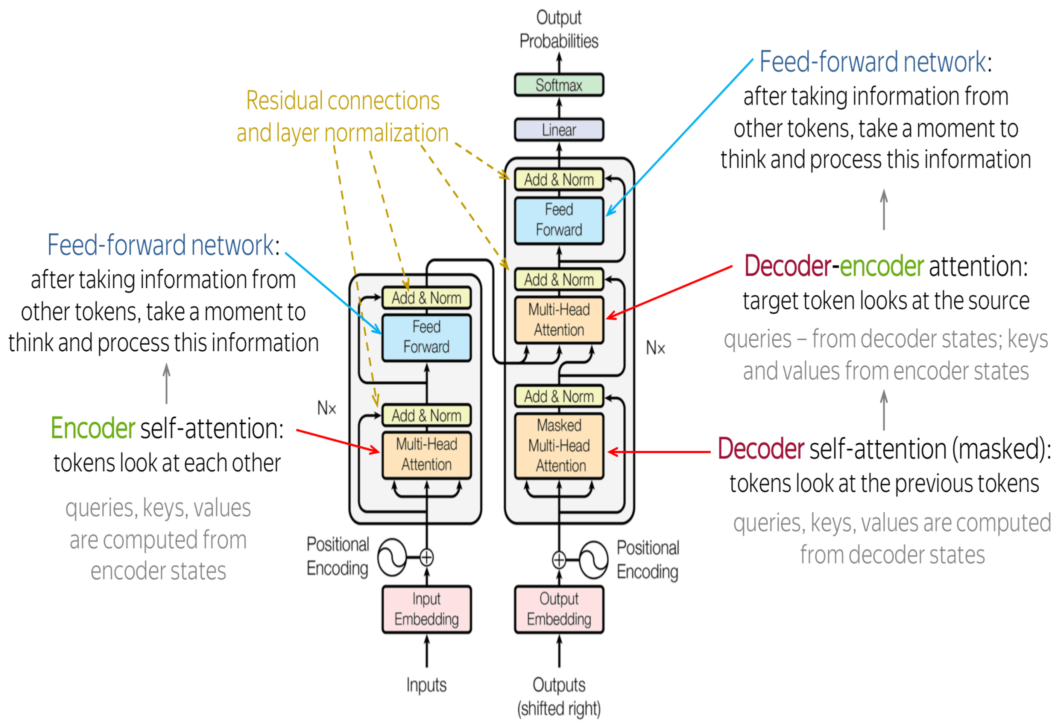


Figure 2.8: Transformer-based models [21]

2.6. Popular modern large language models

There are dozens of major LLMs that's due to its advancement and dominance across various fields . Here are the most recent popular llms which were selected based on their open-source data, widespread adoption, and latest versions:[22]

- **GPT-4o:**

It is OpenAI's successor to GPT-4, GPT-3.5, and GPT-3, released on May 13, 2024. GPT-4o introduces several improvements over GPT-4. OpenAI claims that GPT-4o is 50% cheaper than GPT-4, while GPT-4 is twice faster at generating tokens. Additionally, GPT-4o achieves a response time of 232 ms, similar to human response time, making interactions feel more natural. GPT-4o is a large multi-model , accepting various types of inputs including text, audio, images, videos, and voice capabilities packaged into one.

- **Deepseek R1:**

It was released on January 2024 and developed by DeepSeek. It's an open-source reasoning model for tasks with complex reasoning, mathematical problem-solving, and logical inference. DeepSeek beats or matches OpenAI o1 in several benchmarks, including Math500 and AIME-2024, and it is immediately hit headlines due to the low cost of training compared to most major LLMs.

- **Gemini:**

It is Google's family of large-language models that power the company's chatbots of the same name. These models are multi-model, meaning they can handle multiple types of inputs. Gemini is also integrated into various Google applications and products and is available as a web-based chatbot. Gemini comes in three sizes, Ultra, Pro, and Nano. Among the most recent models is Gemini 1.5 Pro, an update that debuted in 2024. Additionally, Google has introduced Gemini 2.0 Flash and Gemini 2.0 Flash-lite. As of February 5, 2024, Gemini 2.0 Pro was the latest released Gemini LLM designed for complex reasoning and coding tasks.

- **Llama:**

Large Language Model Meta AI (Llama) is Meta’s LLM which was first released in 2023. Llama uses a transformer architecture and was trained on a variety of public data sources. It is available under an license.allowing free use of the models. Llama 3.1 models were released on June 23,2024, as the seccessor to Llama 3. they were available in both 70 billion and a 405 billion parameter model. Meta AI also released Llama 3.2 in September 2024 ,initially with smaller parameter counts of 11 billion and 90 billion. The most recent version is Llama 3.3 introduced in December2024, this model features 70 billion parameters.

- **Grok:**

Created by XAI, which is a company working on building AI to accelerate human scientific discovery. We are guided by our mission to advance our collective understanding of the universe. XAI's team is led by Elon Musk.[23]

Grok's LLMs Versions are :

Grok-1 was first vision open-sourced on March 17, 2024, focused on providing robust capabilities in text generation and reasoning.

Grok-2 was released on August 14, 2024. It is an upgraded performance reasoning image generation using Flux.

Grok-2 mini was released on the same day as Grok-2. It is a small, capable sibling, balanced between speed and quality.

- **Qwen:**

Qwen is large family of open models developed by Chinese internet giant Alibaba Cloud. The newest set of models are the Qwen2.5 suite, which support 29 different languages and currently scale up to 72 billion parameters. These models are suitable for a wide range of tasks, including code generation, structured data understanding, mathematical problem-solving as well as general language understanding and generation.[22]

Table 2.1 A summary of the recent released LLMs

<table>
<tr><th>LLM Name</th><th>Devloper</th><th>Release Date</th><th>Access</th><th>Parameters</th></tr>
<tr><td>GPT-O1</td><td rowspan="3">OpenAI</td><td>September 2024</td><td>API</td><td>Over 175B</td></tr>
<tr><td>GPT-O3</td><td>December 2024</td><td>Not available to public</td><td>-</td></tr>
<tr><td>GPT-O4-mini</td><td>April 2025</td><td>API</td><td></td></tr>
<tr><td>Gemma-3</td><td>Google</td><td>March 2025</td><td>Open-source</td><td>12B,27B</td></tr>
<tr><td>Grok-3</td><td>XAI</td><td>February 2025</td><td>Free(Limited) and X Premium+</td><td>-</td></tr>
<tr><td>Qwen 2.5 Max</td><td rowspan="2">Alibaba Cloud</td><td>January 2025</td><td>API</td><td>-</td></tr>
<tr><td>Qwen3</td><td>April 2025</td><td rowspan="4">API, Open-source</td><td>235B</td></tr>
<tr><td>DeepSeek V3</td><td rowspan="2">DeepSeek</td><td>December 2024</td><td rowspan="2">671B</td></tr>
<tr><td>DeepSeek R1</td><td>January 2025</td></tr>
<tr><td>Llama 4</td><td>Meta</td><td>April 2025</td><td>128B</td></tr>
</table>

2.7. Applications of large language models

Large language models are now widely used across various industries and have become an integral part of everyday applications. Below are some key applications of LLMs:

- **Chatbots :** LLMs have the capacity to hold a conversation in a way that you can barely distinguish them apart from humans, for example ChatGPT.
- **Information retrieval and neural semantic search:** LLMs includes information directly into their parameters via training and fine-tuning, but keeping them updated with new information is tricky. To address this, information retrieval systems often integrate vector database, keeping it dynamically updated and fresh.[24]

- **Content generation:** Refers to automated creation of text, image, audio, or video content. This has become a centerstone of modern LLMs and AI applications, enabling users to create high-quality content at scale while saving time and resources.
- **Natural language understanding(NLU):** Enable human to interact with the machine in more intuitive and meaningful way.

2.8. The advantages of large language models

There are numerous advantages that LLMs provide to their different users:

- **Tasks specific without fine-tuning:** With minimal additional training, LLMs can do a lot of tasks, such as summarization, translation, code generation, and all of this is due to their massive knowledge base.[25]
- **Scalability and efficiency:** Easily handles increased workload and updates to growing business needs. They can analyze large volumes of text data to extract insights and patterns.[25]
- **Fine-tuned on specific domain:** LLM can be trained on specific dataset, which enhance the performance of the model.
- **improve user experience:** by enhancing the user interactions with chatboots, virtual assistants, providing more meaningful and context-aware response.
- **Multi-lingual support,** LLM can work with multiple languages.

2.9. Limitation and challenges of large language models

Recently, LLMs have made remarkable success with their capacities. However, despite that, LLMs still face several limitations and challenges as :

- **High training cost:** Training LLMs requires massive resources, often involving thousands of GPUs or TPUs, which needs significant financial investments with millions of dollars.
- **Data-related challenges:** Training LLMs on vast amount of text data sourced from the Internet. This data may contain biases related to sensitive attributes. As a result, LLMs can perpetuate or even amplify these biases, leading to low-quality outputs. LLMs may leak or generate sensitive information, which may lead to risks to privacy and security. Training LLMs on noisy, incomplete, or low-quality data can lead them to struggle with accuracy, coherence, or relevance which will directly impact their performance.
- **Hallucination and overconfidence in incorrect Answers:** When LLMs generate convincing but inaccuate information, this phenomenon, known as hallucination. But when it generates responses with high-quality confidence even when these responses is fabricated or incorrect information, this is overconfidence, which can mislead users and make it difficult to trust the model's outputs.

- **Ethical and Societal Challenges:** LLMs can generate harmful, incorrect, or inappropriate content, raising ethical and content moderation concerns.

3. Generative AI

Generative Artificial Intelligence, often called Generative AI or Gen-AI, is a sub-field of AI that is capable of creating new content, such as conversations, stories, audio, videos, images, text, and even codes. Generative AI models learn from data they are trained on so that they use their knowledge to generate different outputs that resemble human creation.

3.1. Types of gen AI models

Many types of geberateive models exist , each with its own defining architecture. Here most famous models :

- **Autoregressive models:** predict the next data point in a sequence based on previous data instances. Transformers excel at natural language processing (NLP) tasks due to their enhanced ability to process context. [26]
- **Diffusion models:** create new data by gradually adding noise to a dataset, then figuring out how to remove the noise and yield novel output.
- **Generative adversarial networks (GANs):** pair a discriminative and generative model together in a competition, with the goal being for the generator to create output that fools the discriminator. [26]
- **Variational autoencoders (VAEs):** compress input data with an encoder, then reverse the process with a decoder to create new similar data. [26]
- **Flow-based models:** learn the relationships between simple and complex distributions of data through reversible mathematical operations.

3.2. Popular generative AI

Generative AI can be classified into multiple categories according to the type of its generated content.

Table 2.2 The classification of examples of popular Generative AIs models

AI Tool	Type	Company	Pricing	strength	limites
Chat GPT	Text generation	Open AI	Free & paid versions	Its large database comes from being the most popular, drawing information from users	There are ways to trick it into bypassing its ethical guidelines

Deepseek		High-Flyer	Free	It is free and offers strong features compared to the free version of ChatGPT	Its security is weak, and many reports question its reliability
Dall-E	Photo Generation	Open AI	Free & paid versions	Can use it for free through Microsoft Copilot Service	It does not strong competition compared to the paid versions of its competitors
Midjourney		Midijourny	Paid	The most powerful model for generating images with highly accurate imaginative capabilities	Fully paid, and it cannot be tested without a subscription
Firefly		Adobe	Free & paid versions	It can save time in designs on Adobe programs,with fast and efficient compared to competitors	Requires a moderately powerful computer to be used effectively with Adobe programs
Sora	Video Generation	Open AI	limited access	It relies on CHATGPT, providing powerful capabilities for generation accurate videos	Still in its early stages and not yet available to everyone
Dream-Machine	Video Generation	Luma Labs	Free & paid versions	One of the best models for generation videos with multiple capabilities and different styles	Lack of transparency in training data – Available only as an app for IOS
ElevenLaps	Sound Generation	Eleven Labs	Free & paid versions	It has the ability to generate semi-realistic voices	It has no strict rules, so anyone can clone any person's voice

4. Retrieval augmented generation

RAG is an artificial intelligence technique where an external data source is connected to a large language model (LLM) to generate more accurate and contextually relevant answers without relying on the model's internal knowledge, especially for applications like intelligence assistance, semantic search, and documents analysis that need up-to-date data or domain-specific knowledge.

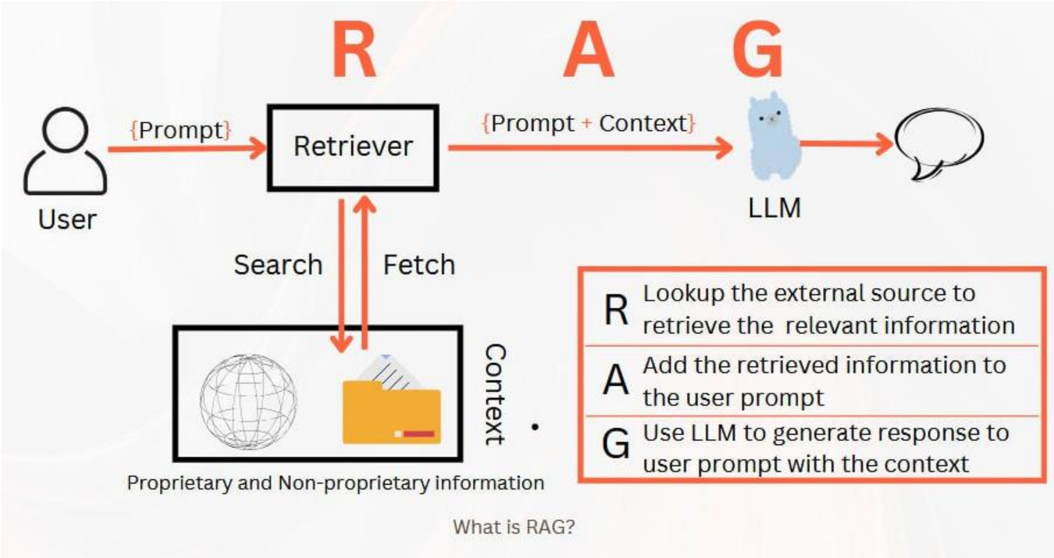


Figure 2.9: RAG [27]

4.1. How does retrieval augmeted generation work

RAG architecture consists of two phases, an indexing phase and an querying phase:

A. The indexing phase: It's the first phase its purpose is to process data in the resources and store them in the vector store database.

- It began with the external knowledge. It could be from various resources like URL or documents, APIs, databases, and other.
- Moving to data ingestion step, it involves collecting and processing the source documents for the next step by extracting text, removing irrelevant data, and formatting the content.
- After preparing the data, it's going to be split into smaller chunks so that it can be manageable pieces, which will make it easier to process and retrieve.
- Each chunk of text is going to be converted into vector (embedding), which are numerical representations that represent the meaning of the text in numerical space. All of this happens by using embedding models.
- Finally, the embeddings are going to be stored in a vector database. This database is later used to efficiently search for relevant text chunks.

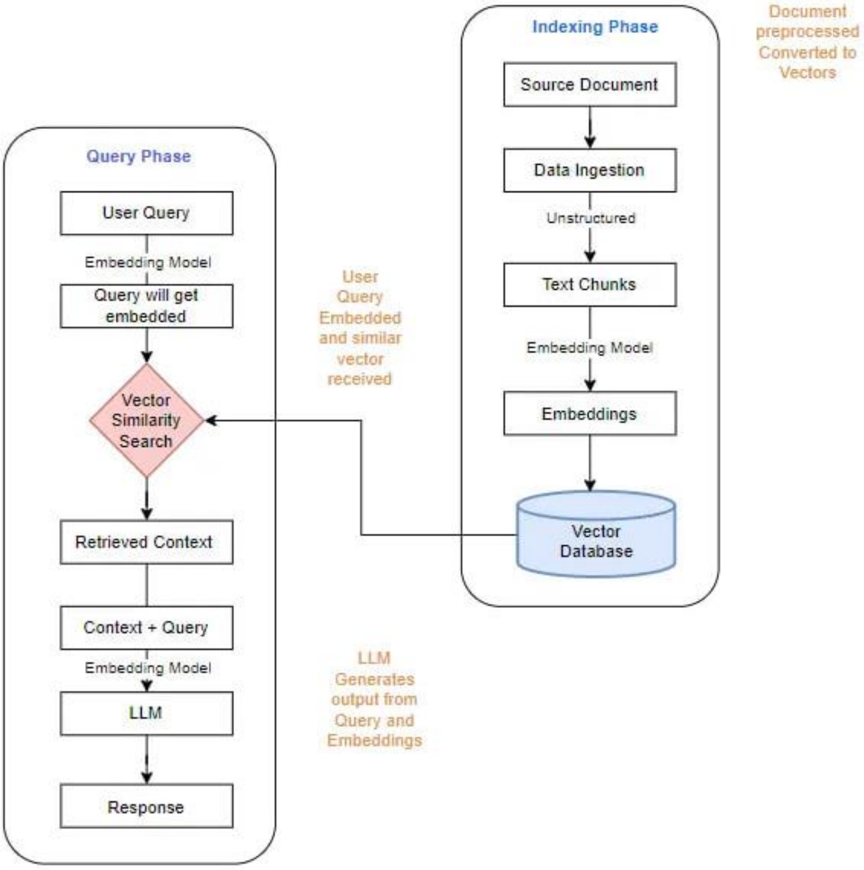


Figure 2.10: The RAG Workflow [28]

B. The Query phase: is the second phase, focuses on embedding the user query to retrieve relevant content. Then the LLM will generate the answer to the user question.

- The query phase starts with a user submitting his question (query).
- The submitted query is converted into a vector using embedding models same as in the index phase. This way the query and the documents are in the same vector database.
- By searching the Vector database to fetch the chunks that are most relevant to the query using Vector Similarity Search.
- The retrieved contents will get us to the top K chunks based on their similarity score between the chunk and the vector query.
- Those retrieved contexts are combined with the original query to form a complete context that is provided to the LLM.
- The LLM takes the context with the combined query and generates an accurate response based on the retrieved information.
- The response is finally going to be presented to the user.

5. Graph RAG

It is a powerful retrieval mechanism that improves Gen-AI applications by taking advantage of the rich context in graph data structures.[29] Graph Rag approach allows us to extract entities such as person, organization, etc. as these entities represent nodes and edges are represented by relationships between those nodes to build a knowledge graph.

5.1. Knowledge graph

It is structured representation of data that directly connects real-world entities semantically. The entities represent nodes, they can be any concept, idea, event, or object. While it just connects the entities, nodes, as relationships. A graph data is modeling any information using graph structure. Not every graph data base is a knowledge graph, since graph data doesn't need to be semantically rich and like knowledge graph.

5.2. Graph RAG process

Graph RAG pipeline consists of two phases, the index phase and the query phase:

A. The index phase:

- **Data collection**: Means collecting the sources of the data you need for the knowledge graph. It can be any documents, URL, websites.
- **chunking and extracting entities and their relationships:** Splitting the source data into sub-documents using chunking strategy. Then the LLM is used to identify and extract the entities, such as person, company, employee, and places, etc. And relationship between entities.
- **Knowledge graph generation:** Using the extracted information from before to create a knowledge graph, which is basically a set of knowledge entities connected through meaningful relationships.
- **Community detection :** Based on Knowledge Graph, we create communities, which are a group of nodes that are densely connected with each other nodes.[30]
- **Hierarchical community structure:** Using hierarchical Community Detection algorithm to identify community-level (clusters).
- **Community levels:** It's like describing each one of the communities into three different levels. Root, high and low
- **Generating community summaries:** For each level, we create a summary, which is basically looking at a set of chunks, then create a summary for each one of them, combining them with another set of chunks using a reduced-map approach to create a summary , repeating the process until we get a whole overview. For each community level, we create a summary, which is a description about what that community is about, in readable format. Then embed it into a vector database, storing it there for later use.

B. The query phase:

- **User query and selecting community level:** Start by taking the user query to select the community level by doing the same process from the indexing phase.
- **Retrieve Relevant Community Summary:** We do a semantic search between the query and community summary to find the most relevant ones from the Vector database.
- **Combine final response:** We combine and refine the retrieved communities to generate the final answer and deliver it to the user.

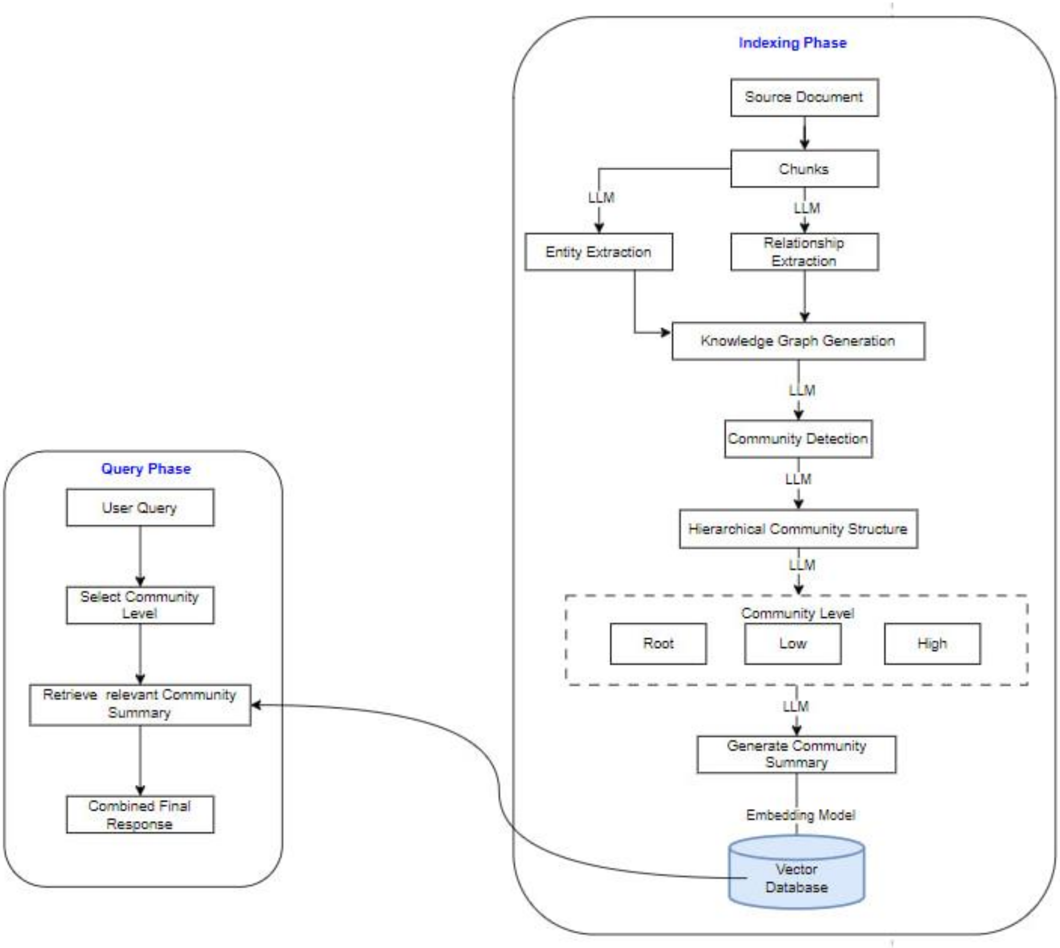


Figure 2.11: The Graph-RAG Workflow [28]

6. Agentic RAG

Agentic RAG represents an advanced approach that enhances traditional RAG systems by integrating intelligent agents capable of autonomous decision-making and multi-step reasoning.[31] The usage of AI agents opens up new possibilities for building more powerful, robust, and versatile advanced-powered applications.[32]

6.1. What is agents and agenetic AI?

Agents: it defines as a real or virtual entity or AI model whose behavior is autonomous, making decisions based on its understanding of the world or data that is available to it. It has the capability of reasoning, planning, and interacting with its environment or other agents.

AI agent: refers to artificial intelligence systems that can perform autonomous decision-making. Unlike traditional AI systems that process data reactively, agentic AI systems use their environment to make decisions and take actions, adapting them even for real-time data, making them capable of dynamically adjusting to new situations which give them more flexibility and intelligence in complex real-world applications.[33] These are the core components of AI agents:

- **LLM :** it uses LLMs to perform specific tasks such as query rewriting, and entering synthesis.
- **Memory**: there are short-term memory and long-term memory, enabling the agent to store and retrieve past experiences as examples.
- **Dynamic planning:** agents can align their actions dynamically based on user intent and intermediate results.[31]
- **Tools**: agents can leverage the diverse tools such as web search, APIs, and databases.

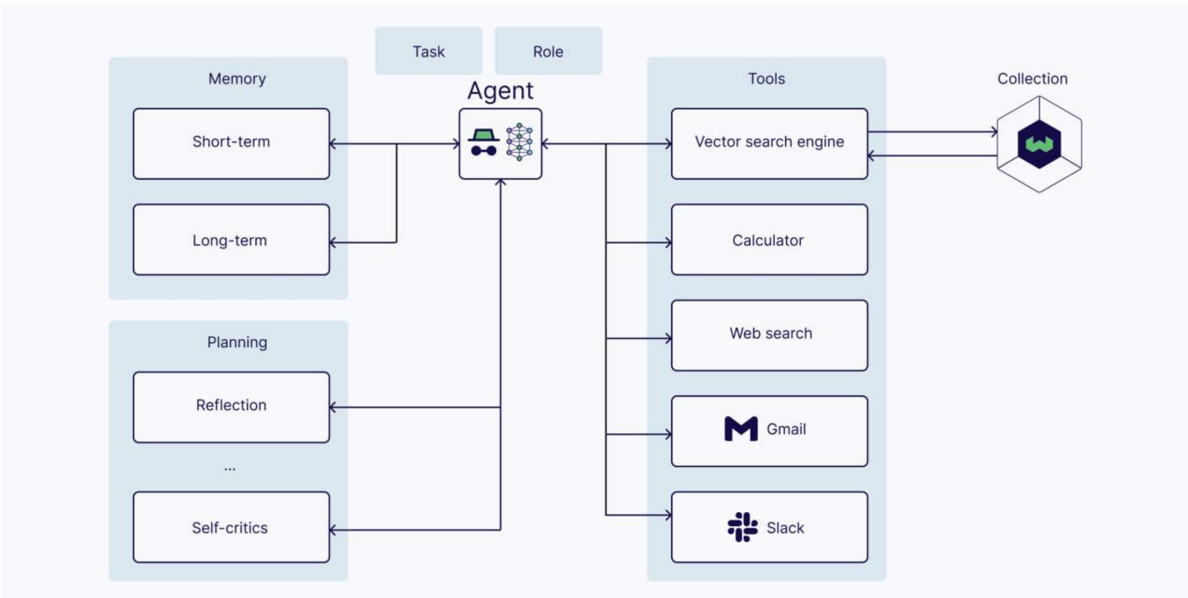


Figure 2.12: The agentic AI [32]

6.2. Agentic RAG pipeline

Here is a breakdown of how it functions:

- **Query input**: process beginning when system receives user's query.
- **Rewrite the initial query:** LLM agent checks the query for clarity and rewrites it, if necessary, for better understanding.[33]

- **Updated Query:** after writing the query, the system moves forward with the updated query, ensuring that the input is optimized for accurate data retrieval and processing.[33]
- **Checking if more information needed:** It evaluates whether the updated query is sufficiently clear and specific to move to the next step, or additional details are needed.[33]
- **If more detail is needed:** the system will request or research for more detail to ensure the response is as accurate as possible. If the LLM agent finds that additional information is needed.[33]
- **Decide the source:** Once system has enough context, After receiving relevant data, the system combines this retrieved context with the updated query, ensuring that the query is now updated with external data or context.LM agent decides which data source will be best to solve the query, from various options such as fixture database, tools and APIs, or even internet for real-time data.[33]
- **Data retrieval:** It proceeds the query selected data source. These can include specialized vector databases for structured data, tools, and APIs for specific queries.[33]
- **Updated query + retrieved context:** After retrieving relevant data, system combined this retrieved context with updated query, ensuring that query is now updated with external data or context.
- **Processing the data:** LLM processes the data query and retrieves its context. It uses this information to understand query in depth and helps in giving contextual appropriate response.
- **Response:** allowing generate response based on the data and the context provided. This is the important step of the process where system change raw data into a meaningful output which will be a directed answer or a recommendation.[33]
- **Checking answer will advance:** once response is generated, LLM agent will check whether the response accurately answers user's query and align with context of the query.
- **Final response:** once the LLM agent confirms that the generated response is relevant and accurate, then the final response is delivered to the user.[33]

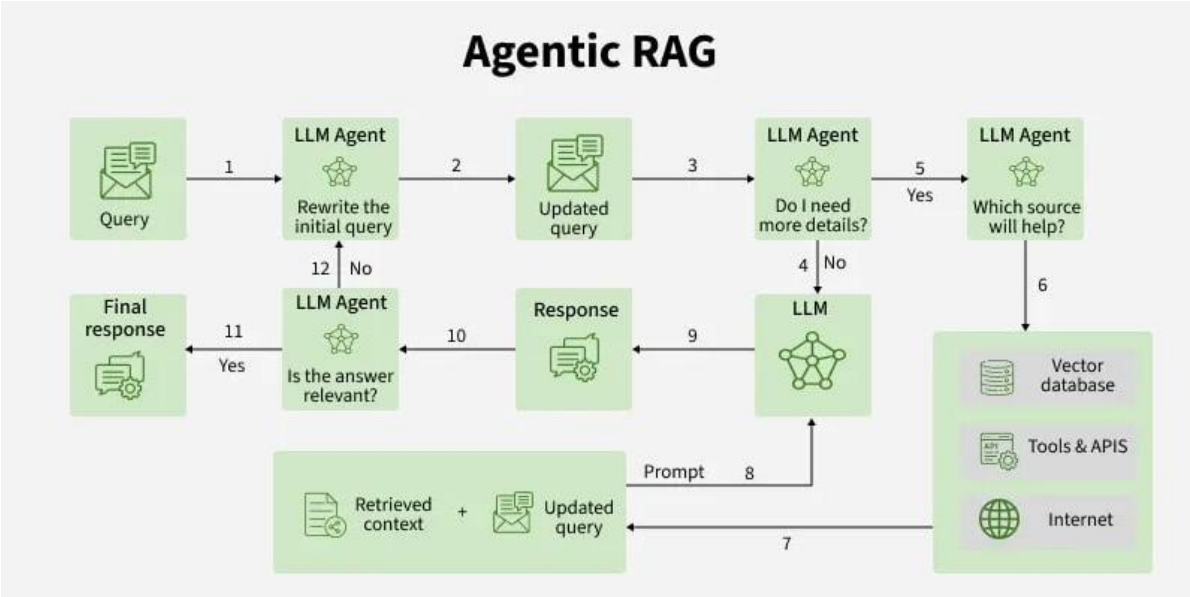


Figure 2.13: The agentic RAG Pipline [33]

6.3. Types of agentic RAG

The core of agentic RAG architecture is agent, which also can have various level complexity. To keep it simple, we introduced two types of Agentic RAG based on their fundamental architecture.

- **Single-agent RAG (Router):**

In its simplest form, agentic RAG is router, meaning that at least there is two external knowledge sources leaving the agent to decide which to receive additional context form. The external knowledge could be vector database, web search, APIs and tools, as well as email accounts.[32]

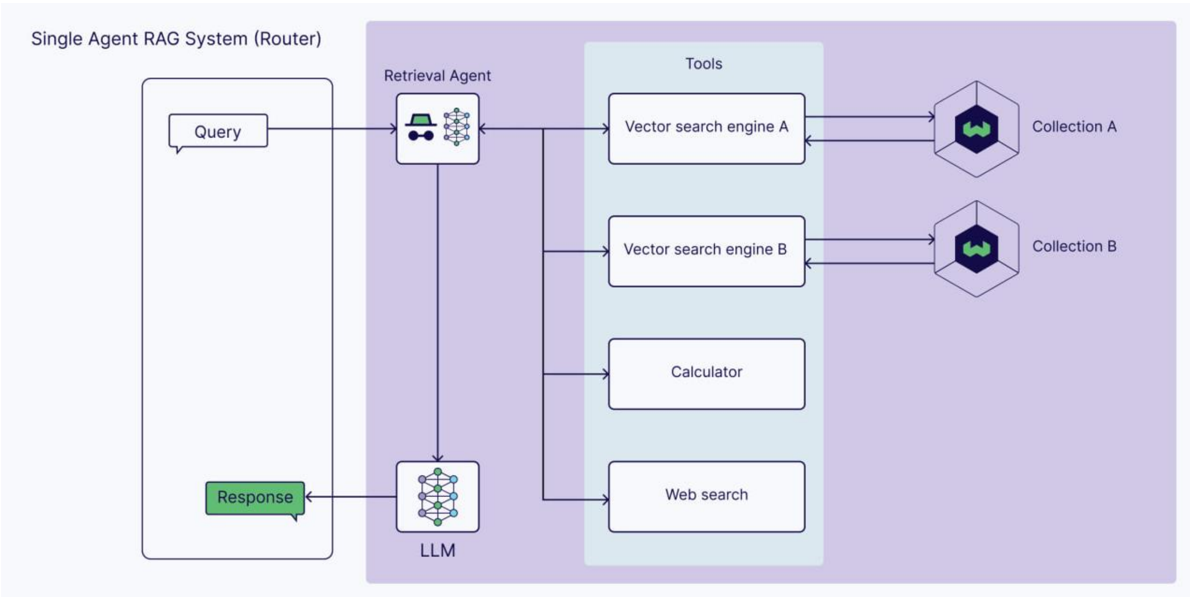


Figure 2.14: Single-Agent RAG [32]

- **Multi-agent RAG systems:**

The single-agent system is limited to only one agent with presuming retrieval and answer generation in one. Therefore, it is more beneficial to chain multiple agents into multiple-agent RAG applications.[32] As an example, we can have one name agent who coordinates information retrieval among multiple specialized retrieval agents.It could be like each agent could retrieve information from each internal or external knowledge source.[32]

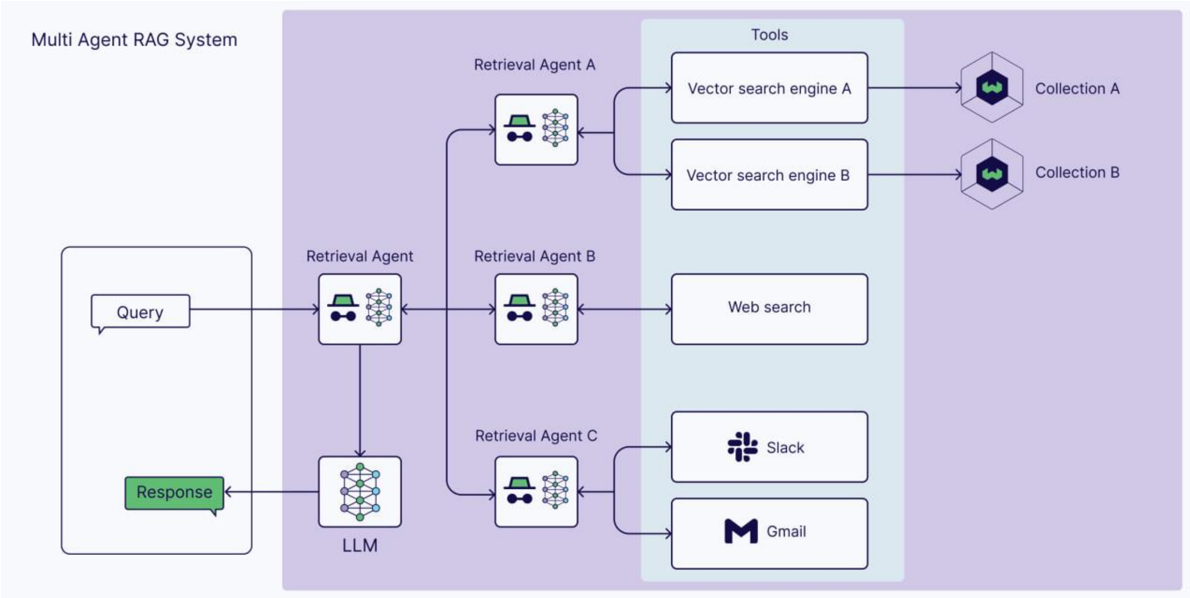


Figure 2.15: Multi-agent RAG systems [32]

7. Comparation between the technologies

We compare between these technologies: RAG, Graph RAG, and Agentic RAG, based on the chosen criterion, which are enhanced response quality and context understanding, handling complex reasoning, and implementation complexity.

Table 2.3 Comparison between the traditional RAG, graph RAG, Agentic RAG

Criterion	RAG	Graph RAG	Agentic RAG
Enhanced response quality and context understanding	Preloading knowledge into the LLM improves response quality ,but semantic search may retrieve irrelevant information	Improves information retrieval and context understanding, addressing RAG limitations	Provides accurate answers with refined context selection using advanced tools
Handle complex	Lacks nuance for	Handles complex	Supports complex

Reasoning	complex question answering, relying on similarity search	reasoning well with LLM integration for coherent responses	tasks with multi-step reasoning, like expert researchers
Implementation complexity	Uniform chunking leads to information loss, complicating document processing	Integrating graph data with unstructured systems increases complexity	Managing multiple-agents and data sources is complex

8. Conclusion

LLMs are incredibly versatile handling tasks like generating content and solving complex problems with ease. Fine-tuning them gives you the power to customize their ability for your specific needs without the hassle of retraining from the ground up. surely, there are challenges such as ethical questions or the risk of outdated information, but these tools are reliable and ready to be used where needed.

LLMs are a subset of Generative AI, as both technologies share some similarities, including the need for large datasets for training. Their primary goal is to optimize business operations and enhance customer satisfaction by facilitating content creation and improving customer service tasks.[34]

Finally, we showed you the process of how the RAG Foundation technologies turn external knowledge into domain knowledge-based, giving the LLM the power of analyzing, inferring, and even reasoning in a similar way to humans due to its unique structure of each one. All of these technologies helped the LLM to be adopted even more by the companies. The last point will be outlined more in the next chapter.

CHAPTER 3

Business Reports and LLM: The new strategy of company Success

1. Introduction

After establishing the foundational concept and technologies in previous chapter, this chapter delves into how companies manage their internal documents, the type of documents involved, and the impact of these documents have on organizational outcome. It also presents case studies including our experience as well as applications of large language models (LLMs) in real-world business contexts.

2. Reports

Reports are a key communication tool in business. They are considered as company archives that will benefit the current or future employee to see research, information and the reasoning for some problem and action and decisions . Reports could be formal or informal that serve two purposes informative or analytical.

3. Business reports

Business reports often include quantitative data, such as financial figures, sales metrics, and others. They are about transforming raw data into meaningful insights in structure that represent information, analysis, and insights about specific topics, issues, or aspects of an organization's operations. They are used as the main communication tool, enabling companies to make informed decisions, track progress, and address challenges.[35]

3.1. Characteristics of business reports

These are some most important characteristics of business reports:[35]

- They must be accurate.
- Their information must be presented in clear and precise form.
- They contain relevant information such as facts and figures.
- They are written in an obvious style with simple language.
- Business reports are objective and logical.

3.2. Structure Of business reports

The business report consists of a front matter, body of report, and back matter. Following this structure make all types of reports in business communication easier to navigate for readers.[36]

1) Front Matter
- Cover
- Title Page
- Executive Summary
- Table of Contents
- List of Figures

2) Body of the Report
- Introduction
- Findings and Discussion
- Conclusions
- Recommendations

3) Back Matter
- Appendices
- References
- Glossary (if required)

Figure 3.1: business report contains the following elements [36]

A business report ordinarily incorporates the taking after this primary sections:[36]

- **Title Page:** Contains the report title, author's full name, and date of distribution.
- **Executive summary:** A brief outline of the report's subject, information, and suggestions, planning for active administrators.
- **Table of Contents:** Records the most headings and subheadings with page numbers for simple route, particularly in longer reports.
- **Introduction:** Clarifies the reason of the report and clarifies the information collected and the reason for collecting it.
- **Body:** Contains points of interest, information, and examination, and can be displayed utilizing tables, charts, and images.
- **Conclusion:** Summarizes the most discoveries displayed within the report.
- **Recommendations:** Recommends activities, arrangements, or following steps based on the report's discoveries.
- **References:** Records outside sources and data cited within the report.
- **Appendices:** Incorporate extra fabric such as a glossary, financial articulations, and meet transcripts.

3.3. Classifications and the types of business reports

A. On the basis of frequency or importance:

The reports that fall under this category are routine reports and special reports.

- **Routine reports:** They are prepared at specific intervals or as routine activities. Their frequency could be daily, weekly, monthly.[36]
- **Non-routine reports:** They belong to special nature or occasion. They have various kinds, such as interview report, staff report.

B. On the basis of function

- **Informational report:** They simply presents relevant information by mentioning facts without detailed explanation. Their object is to inform the stakeholders how the progress of the company is. It doesn't involve evaluation or analysis or recommendation like annual budget.
- **Analytical report:** They involve evaluation, recommendation, and it may even solve problems at companies, such as discuss employees' issues or even evaluate a company's current standing, and present relevant conclusions , solutions and suggestions.

C. on the base of purpose:

They are waiting to carry out a specific objective.

- **Research Report:** Their content includes the research process, findings, conclusions, recommendations, and limitations, which all of this done by experts. It will inform a business about essential market needs, for example.[37]
- **Explanatory report:** An explanatory report is written to explain a specific topic of interest to its readers. It is written in a clear and formal way which is easy to understand. Explanatory reports can be used to explain the findings of a research report. [36]

D. On the basis of the subject matter:

- **Sales Reports and Revenue:** are one of the most basic types as it simply lists your sales statistics in a given time period so that you can easily do comparison.[38]
- **Financial reports:** they explain a business financial status and performance.[37]
- **The annual reports:** are primarily used by listed companies or non-profit organizations to inform shareholders of the organization's activities and financial performance for the year. It typically includes an overview of operations, financial performance, and strategic initiatives, as well as projections for future goals. The second part contains detailed financial statements and graphs showing expenses, revenues, profit margins, and other key indicators.[39]
- **Market survey reports:** they are an essential report for business companies, as it includes direct response from customers, and it is a collection of primary data collected from companies' targeted region or customer feedback.

3.4. Importance of business reports

- **Improve decision making:** Business reports improve decision-making by providing a factual basis based on data collected from various sources within the company, supporting more accurate and informed decisions.[40]
- **Enhance transparency and accountability:** Reports facilitate performance tracking, monitoring progress, and clear communication with stakeholders. This transparency builds trust, ensures accountability, and helps align everyone around the company's priorities and direction.[40]
- **Improve resource allocation:** By analyzing data and identifying trends, reports help companies allocate time, money, and effort more effectively. This ensures efficient use of resources, reduces waste, and increases profitability.
- **Early detection of problems and opportunities:** Business reports can reveal patterns or anomalies that might otherwise go unnoticed, allowing problems to be addressed early and potential opportunities to be seized.[40]
- **Improve communication and corporate learning:** Reports serve as an effective communication tool across departments and divisions, fostering shared understanding and encouraging continuous learning and improvement.
- **Supporting long-term planning and goal setting:** By analyzing historical data and trends, reports help companies predict future performance and set realistic strategic goals, enhancing the effectiveness of long-term planning and increasing the chances of sustainable success.[40]

4. Companies overview

This section will outline an overview of our case study companies that we will use their documents as test data-set later, which are: national companies like Sonatrach and Sonelgaz, where we did our internship, and international companies like ExxonMobil for more details or better testing results.

4.1. Sonatrach company

It is a company with an international vocation as the Algerian oil giant is a leader in the global energy sector, operating in several countries around the world, having a large-scale partnership, Sonatrach, recognized as an essential partner in the hydrocarbon industry for 16 years, is aiming for the top 5 national oil companies. Their responsibility lies on relevant and ongoing awareness in the area of health, safety, and environment. Its strength lies in its ability to integrate groups across the entire value chain, exploration, production, pipeline, transportation, leak function, separation, and refinement. Each employee's strength lies in unity within the company. This shared responsibility and attitude guarantees the success of Sonatrach's strategic project.[41]

Figure 3.2: The Sonatrach logo [41]

4.2. Sonelgaz company

Sonelgaz is the historic operator in the field of electricity and gas supply in Algeria. The company was established in 1969 and has been providing Algerians with energy for half a century.Following the issuance of the Electricity and Gas Distribution Law by pipelines, Sonelgaz became a holding company responsible for managing a multi-business and multi-professional group.[42] The Sonelgaz Group has played a major role in the country's economic and social development. Its policies are aligned with the national energy policy, particularly with regard to rural electrification and gas distribution. By 2022, electricity coverage reached 99 percent, representing 11,461,721 electricity customers. Gas coverage reached 65 percent, with 7,308,462 customers benefiting from gas. Today, the Sonelgaz Group consists of 11 subsidiaries, directly managed by the holding company, as well as 10 companies with direct and indirect shareholding.[42]

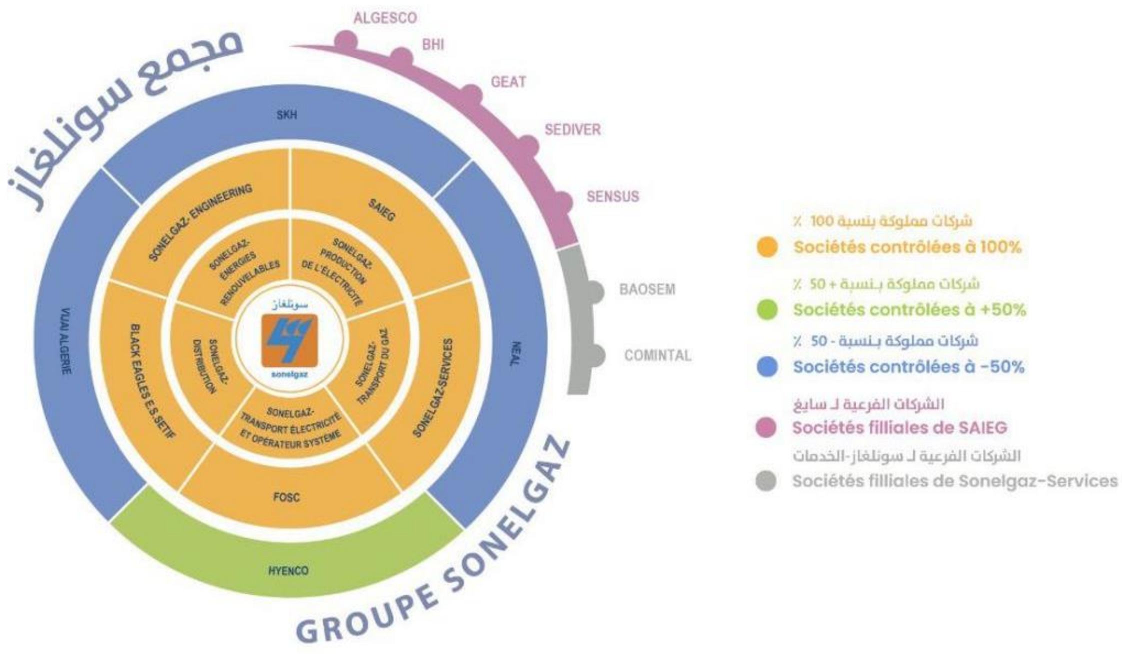
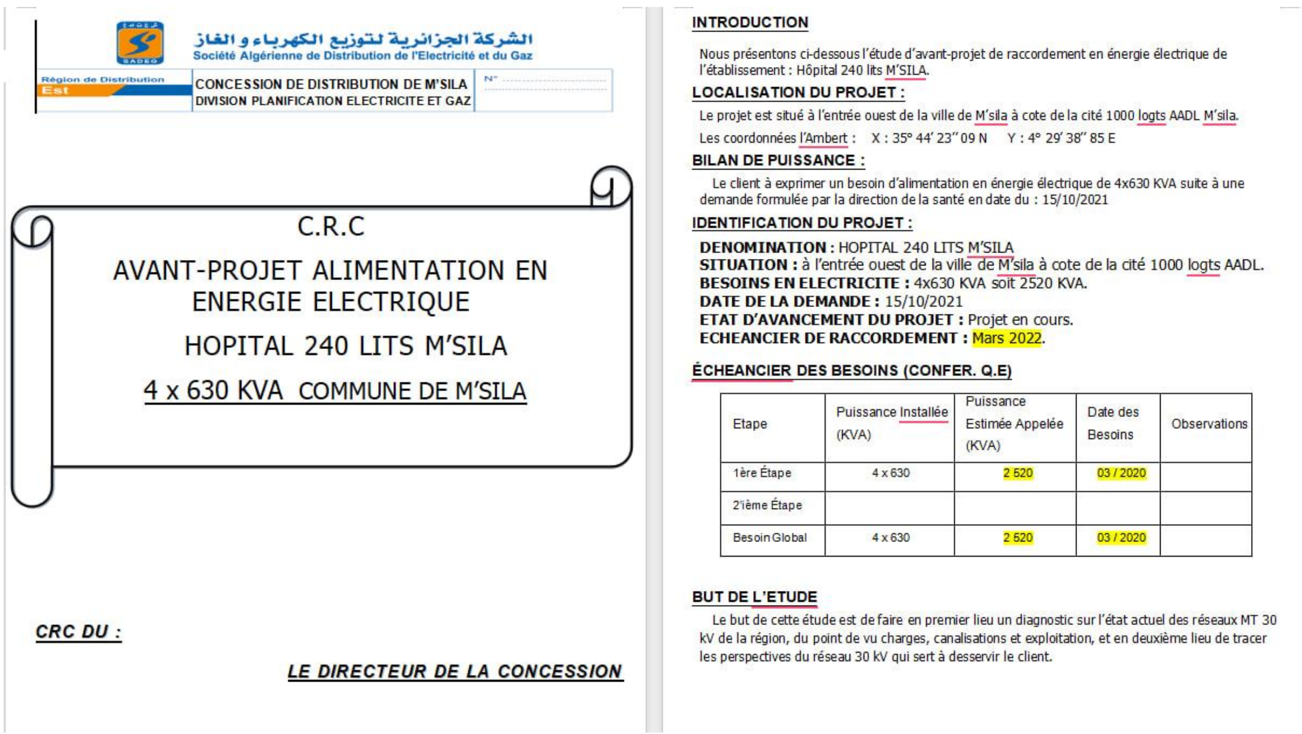


Figure 3.3: The Sonelgaz Group [42]

Sonelgaz, like other companies, produces reports and documents, but its division into departments allows each branch to handle its specific tasks regarding the types of documents. As the department where we did our internship was responsible for distributing gas and electricity. During our training there we observed the following:

The distribution department has two main reports:

- **Study reports:** It contains proposals that is initially technical study for the electric power connections of the client demands. It also contains the client information such as place, start date, planned connection date ,underground , overhead lines, and how much client participates. These proposal are simulation of how electricity delivery process will take place to be discussed later and to choose the most suitable model in the term of objectives and costs.



الشركة الجزائرية لتوزيع الكهرباء و الغاز
Société Algérienne de Distribution de l'Electricité et du Gaz

Région de Distribution Est

CONCESSION DE DISTRIBUTION DE M'SILA
DIVISION PLANIFICATION ELECTRICITE ET GAZ

N° ....

C.R.C

AVANT-PROJET ALIMENTATION EN ENERGIE ELECTRIQUE

HOPITAL 240 LITS M'SILA

4 x 630 KVA COMMUNE DE M'SILA

CRC DU :

LE DIRECTEUR DE LA CONCESSION

INTRODUCTION

Nous présentons ci-dessous l'étude d'avant-projet de raccordement en énergie électrique de l'établissement : Hôpital 240 lits M'SILA.

LOCALISATION DU PROJET :

Le projet est situé à l'entrée ouest de la ville de M'sila à cote de la cité 1000 logts AADL M'sila.

Les coordonnées l'Ambert : X : 35° 44' 23'' 09 N Y : 4° 29' 38'' 85 E

BILAN DE PUISSANCE :

Le client à exprimer un besoin d'alimentation en énergie électrique de 4x630 KVA suite à une demande formulée par la direction de la santé en date du : 15/10/2021

IDENTIFICATION DU PROJET :

DENOMINATION : HOPITAL 240 LITS M'SILA
SITUATION : à l'entrée ouest de la ville de M'sila à cote de la cité 1000 logts AADL.
BESOINS EN ELECTRICITE : 4x630 KVA soit 2520 KVA.
DATE DE LA DEMANDE : 15/10/2021
ETAT D'AVANCEMENT DU PROJET : Projet en cours.
ECHEANCIER DE RACCORDEMENT : Mars 2022.

ÉCHEANCIER DES BESOINS (CONFER. Q.E)

Etape	Puissance Installée (KVA)	Puissance Estimée Appelée (KVA)	Date des Besoins	Observations
1ère Étape	4 x 630	2 520	03 / 2020	
2'ième Étape				
Besoin Global	4 x 630	2 520	03 / 2020	

BUT DE L'ETUDE

Le but de cette étude est de faire en premier lieu un diagnostic sur l'état actuel des réseaux MT 30 kV de la région, du point de vu charges, canalisations et exploitation, et en deuxième lieu de tracer les perspectives du réseau 30 kV qui sert à desservir le client.

Figure 3.4: over view of study report

- **Development reports:** It includes aspirations, study, and also projects, that are in long-term for example, 10 years.

The overall structure of these reports is as follows: Introduction, Project Summary or Overview, Study Objectives, Recommendations (proposals), and finally there is the Conclusion, which includes the chosen proposal after the discussion.

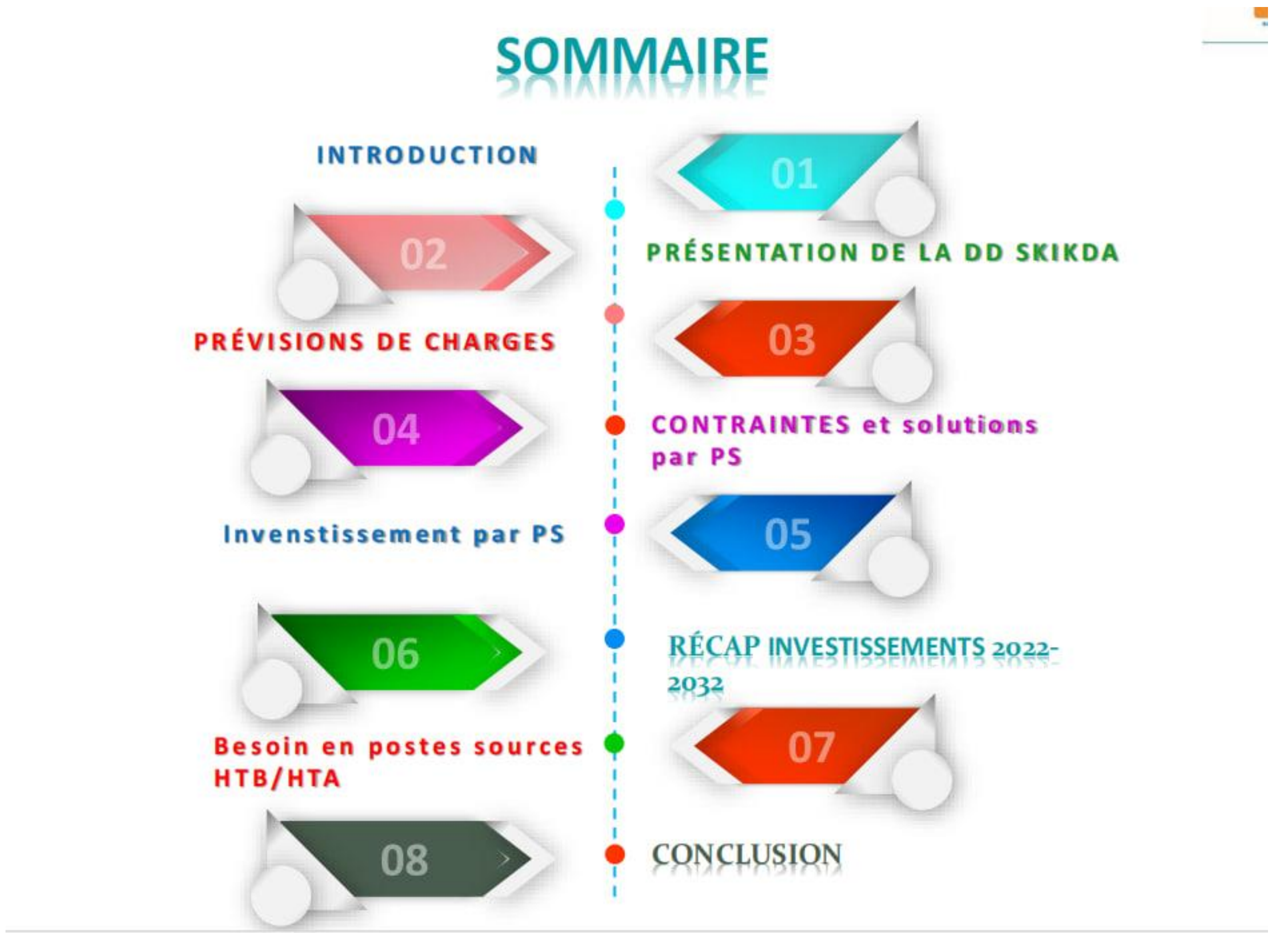


Figure 3.5: content of a development report

There are other reports but they were either not permitted to see them or they didn't fit our case study requirements such as Sonelgaz demand form

4.3. ExxonMobil company

They manage an industry-leading portfolio of resources, and it is one of the largest integrated fuels, lubricants, and chemical companies in the world. They have evolved their operating model and global organization to better leverage the scale of their increasingly integrated company and global brands. They have three core businesses with operations around the world, which are upstream, downstream, and chemicals. These three provide products that enable modern life, including energy, chemicals, lubricants, and lower emission technologies. Their purpose is to create sustainable solutions that improve quality of life and meet society's evolving needs.[43]

Figure 3.6: The ExxonMobil logo [43]

5. Traditional way for companies to retrieve information from their internal documents

After many research, inside Algeria especially, we discovered that most of the companies traditional way to retrieve information from their internal documents is manual review.

5.1. Manual review

It involves reading documents, then collecting data from them. This requires a full year's concentration, which can be time-consuming and prone to error. This process can be so difficult and highly inefficient when dealing with large data volumes, which will be challenging for companies.

5.2. Optical character recognition (OCR)

It is a technology that converts scanned documents and images into machines via readable data. It is widely used for automated data extraction, especially in industries like financial services and government. OCR accuracy can be impacted by the quality of the scanned documents.[44]

5.3. Enterprise content management (ECM) systems

It is the systematic approach to capture, manage, store, preserve, and deliver information related to business processes. It's neither a single technology, nor a methodology, nor a process. It is a combination of strategies, methods, and tools.[45]

6. LLMs role in the company success

The existing of LLMs across various industries took the company's business into another level.

- **Enhancing customer experience:** LLMs revolutionizing how businesses interact with their customers by enabling more accurate, personalized, and timely responses. It can be as Chatbots, Virtual Assistant, and others.

- **Data analysis and better decision making:** Today, data is a critical asset for businesses, but extracting insights from massive amounts of Text data is a complex challenge. That's where the role of LLM is, as it can analyze this text, identify trends, and provide key insights, helping businesses make informed decisions leading to better business outcomes.[46]
- **Improving internal operations and knowledge management:** LLMs are not only limited to external operations, but also is useful for internal processes and improve knowledge management, making it easier for employees to search for relevant information and even create personalized learning experiences by generating training materials and providing resources based on Their individual learning needs.[46]
- **Automating content creation and marketing:** LLMs have the ability to generate high-quality content, making valuable tools for marketing and content creation. Having the ability to understand content tone and style, LLMs can produce articles, blogs, products ,descriptions, and social media parts, and even marketing emails, which will not only save time, but also ensure consistency across multiple platforms, boosting brand visibility and customer engagement.[46]
- **Enriching employees' experiences:** by using LLMs for automated, repetitive, mundane tasks, which will allow employees to have space and time to deal with more complex tasks and be more productive and creative.[47]
- **Innovation and streamlined research and development process:** By speeding scientific or technical search process By summarizing research, generating hypotheses, and proposing technical solutions to complex problems LLMs also have the ability to analyze massive amounts of data which help suggest new ideas, new solutions to customer problems, new products according to the current trends, and even suggest new services.

6.1. Case Studies of Successful Business Applications of LLMs

- **HP** incorporated GitHub Copilot into its software development workflow and developers quickly found that they code faster and solve issues more quickly without getting bogged down in tedious code scaffolding and syntax. Since embracing GitHub Copilot, their developers feel more supported and are better able to collaborate reporting substantial increases in productivity.[47]
- **Camping World** was having a serious problem at the call center, due to not only the staff was saved by the sudden increase in calls, but also the calls that come after hours was found to be missed or pushed to the next day, which meant the customer queries were going unanswered, leading to the sales team to lose potential lead from those calls. That's why the company came up with the solution to create an AI assistant called Arvee, which took the calls 24/7 hours, helping answer questions for the customer without the need for agents to involve. That's what led to a 40% increase in customer engagement And 33% increase in agent efficiency.[48]

- **BMW Group** optimizes the customer experience connecting 13 million active users to their vehicles with the MyBMW app on Azure, which supports 450 million daily requests and 3.2TB data processing.[47]
- **Telstra** An Australian telecommunications company provides advanced services at scale both domestically and internationally. The company used Microsoft Azure OpenAI services to develop Ask TelStra, an AI system that summarizes customer interactions, enabling new agents to work quickly without accessing massive databases. Agents read summaries on-screen, reducing call delays by 20%, and 94% of agents report a positive impact on customer interactions, with 90% of them becoming more efficient.[48]

7. Conclusion

In conclusion, the chapter provides a comprehensive introduction of business reports and its types. Then we gave an overview to our case study companies , Clarifying how the employees of these companies retrieve the data from the internal documents and how it is exhaustive to them and time-consuming. Then highlighting the role of LLMs of making tasks easier for employees , enhancing their productivity , improving the business successes.

CHAPTER 4

Implementation , Evaluation and Results analysis

1. Introduction

In the previous chapter, we covered the Internal Grants concepts, showing the limits of the company's traditional ways to handle their documents, regarding their importance, especially in Algeria, which is limited to manual review. In this chapter, we will address these limits by discussing the implementation of our solution using traditional RAG , Advanced RAG , Graph RAG, Agentic RAG.

2. Dataset collection and sources

In order to test and evaluate the performance and accuracy given by our system, we collect a data set of corporate documents. These documents simulate realistic use cases where employees interact with internal company files through a chatbot interface. The data set was specifically selected to study the success and effectiveness of a system in our case study, which is the Algerian reality with their company's documents.

2.1. Sources of documents

The dataset includes a section of Sonelgaz private documents where we were interns and publicly available documents from international company ExxonMobil and national company Sonatrach. These documents can be:

- Annual reports.
- Financial reports.
- Development reports.
- Study reports.

ExxonMobil [49] and Sonatrach [41] documents were collected from their own official websites.

2.2. Document categories

The diversity of the documents ensures that our system is tested across different content and structure.

Table 4.1 The documents Categories

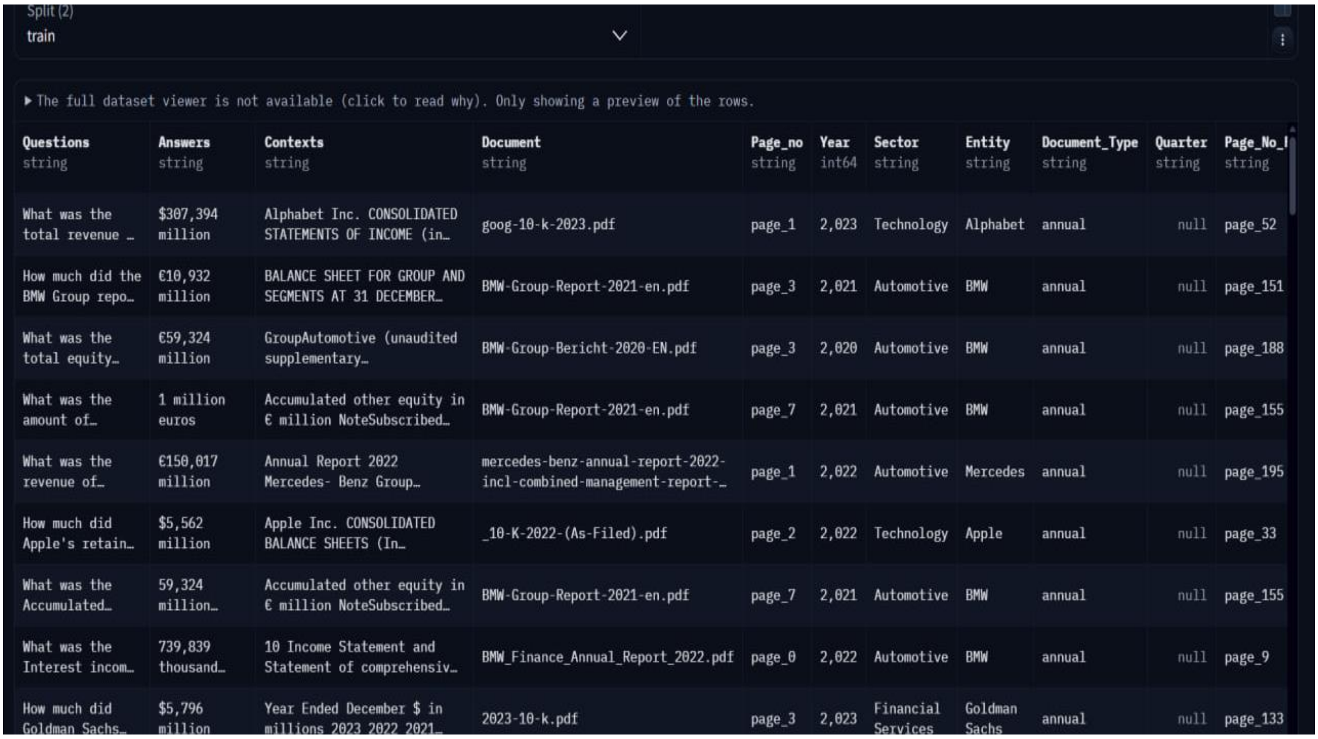
Documents types	**Count**	**Company**	**Average size**	**Format**
Annual reports	5	ExxonMobil,Sonatrach	20 Mb	Pdf
Financial reports	3	ExxonMobil,Sonatrach		Csv,Pdf
Study reports	3	Sonalgaz		Word,Pdf
Development reports	1	Sonalgaz		pdf

2.3. Evaluation dataset

For evaluation, we are going to use **financial-qa-dataset** [50] which is a dataset consists of question-answer-context triples. It also consists of metadata for filtering the records. It is in English language and financial context and has the MIT license. It consists of:

- **Financial-qna-with-metadata.csv:** Contains Question-Answer- context triples.
- **Data** in the directory contain the source documents categorized into **statements** and **reports**.

This dataset is going to be used as a benchmark performance of our system. It's chosen due to its structure and it's ideal for RAG system.


Split (2)
train

▸ The full dataset viewer is not available (click to read why). Only showing a preview of the rows.

Questions string	Answers string	Contexts string	Document string	Page_no string	Year int64	Sector string	Entity string	Document_Type string	Quarter string	Page_No_I string
What was the total revenue …	$307,394 million	Alphabet Inc. CONSOLIDATED STATEMENTS OF INCOME (in…	goog-10-k-2023.pdf	page_1	2,023	Technology	Alphabet	annual	null	page_52
How much did the BMW Group repo…	€10,932 million	BALANCE SHEET FOR GROUP AND SEGMENTS AT 31 DECEMBER…	BMW-Group-Report-2021-en.pdf	page_3	2,021	Automotive	BMW	annual	null	page_151
What was the total equity…	€59,324 million	GroupAutomotive (unaudited supplementary…	BMW-Group-Bericht-2020-EN.pdf	page_3	2,020	Automotive	BMW	annual	null	page_188
What was the amount of…	1 million euros	Accumulated other equity in € million NoteSubscribed…	BMW-Group-Report-2021-en.pdf	page_7	2,021	Automotive	BMW	annual	null	page_155
What was the revenue of…	€150,017 million	Annual Report 2022 Mercedes- Benz Group…	mercedes-benz-annual-report-2022-incl-combined-management-report-…	page_1	2,022	Automotive	Mercedes	annual	null	page_195
How much did Apple's retain…	$5,562 million	Apple Inc. CONSOLIDATED BALANCE SHEETS (In…	_10-K-2022-(As-Filed).pdf	page_2	2,022	Technology	Apple	annual	null	page_33
What was the Accumulated…	59,324 million…	Accumulated other equity in € million NoteSubscribed…	BMW-Group-Report-2021-en.pdf	page_7	2,021	Automotive	BMW	annual	null	page_155
What was the Interest incom…	739,839 thousand…	10 Income Statement and Statement of comprehensiv…	BMW_Finance_Annual_Report_2022.pdf	page_0	2,022	Automotive	BMW	annual	null	page_9
How much did Goldman Sachs…	$5,796 million	Year Ended December $ in millions 2023 2022 2021…	2023-10-k.pdf	page_3	2,023	Financial Services	Goldman Sachs	annual	null	page_133

Figure 4.1: financial-qa-dataset content

2.4. Limitations of the dataset

Although the dataset was carefully collected to ensure breadth and diversity, it used relieved potential limitations. For instance, documents sourced from publicly available company websites may not fully reflect the nature of internal or private documents. They may be subtle difference in context, structure, and language or tone, difference that could influence model performance in way worth investing.

3. System implementation

The realization of this Chatbot involves multiple implementations of retrieval-augmented generation . The purpose was to compare how different retrieval pipelines influence the accuracy and the relevance of generated answers in a document-based Q&A setting.

3.1. User interface design

This is the overall interface of RAGDocAI Chat, where a user can upload their documents and input their queries in natural language. As it is appeared in the figure 4.12 , the interface includes :

Left panel : this section allows users to upload documents to process them in background process then retrieve the information from them.

- **LLM Selection Pane**l, which allows users to choose the Models . Groq API Models were used as a benchmark to compare the influence of using different models on the answer quality and to reduce the latency of Local Models.
- **RAG setting**, which allows users to configure the behavior of a RAG system with multiple options of RAG mechanisms.

Below these toggles are two adjustable parameters:

- **Temperature (set to 0.7)** : Control the creativity of large-language model outputs. If you put them at lower value, that means it will return a deterministic answer.
- **Max content (set to 15)** : Specifies how many document chunks the retriever can use as context when generating an answer.
- **At the end ,** there is a button labeled "Clear Chat History", which allows users to reset the session.

The chat area, which is the center area consist of an input bar at the bottom, where the user will type his query and on the top of this picture, the chat will be displayed.

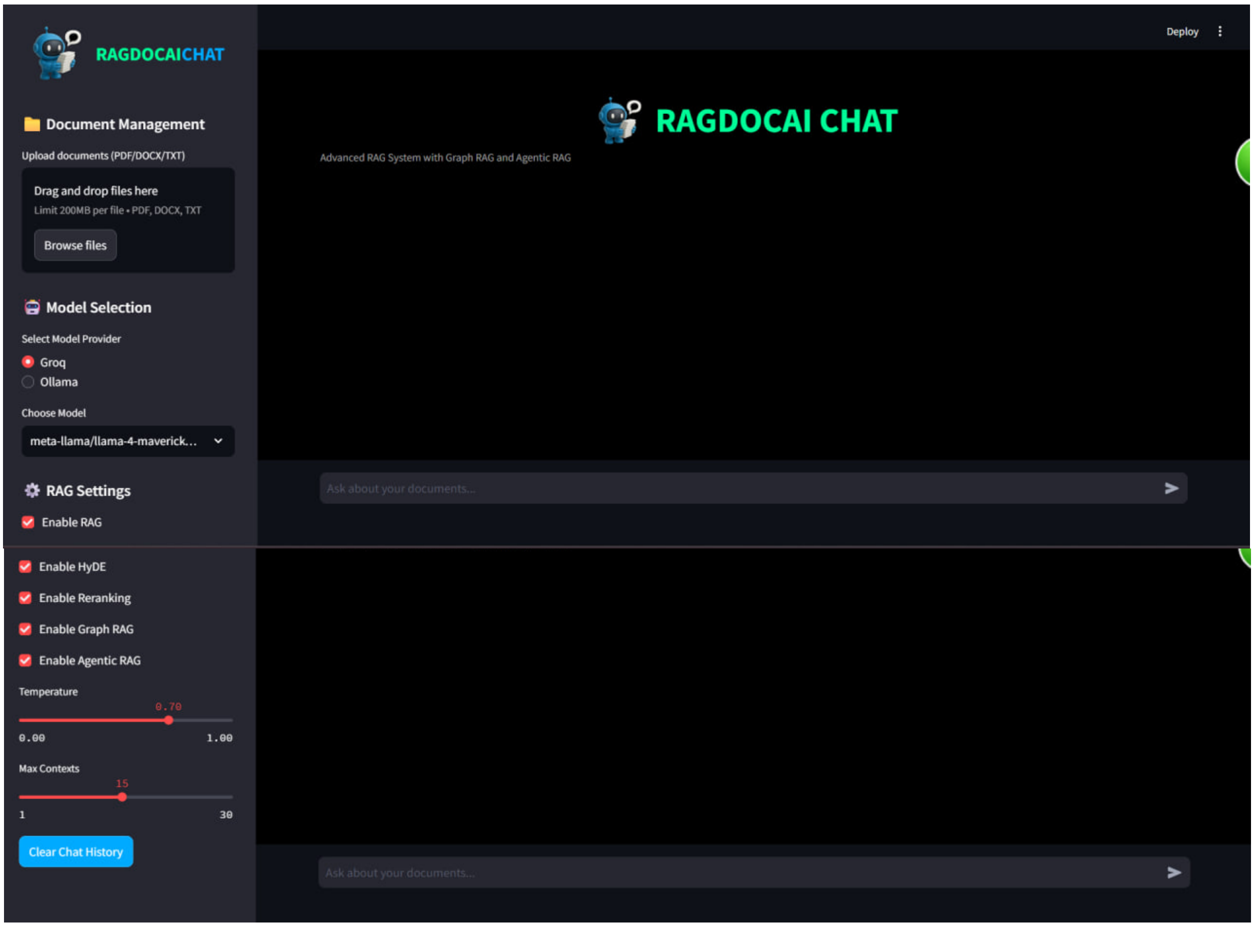


Figure 4.2: User Interface of RAGDocAI

3.2. Traditional RAG implementation

The baseline implementation follows the standard retrieval augmented generation pipeline.

- **Chunking** is the process of breaking down large pieces of text into smaller manageable chunks. And the chucking method used here is **RecursiveCharacterTextSplitter** .
- **Embedding process** is converting words or chunks into numerical vectors that capture semantic meaning of the text. The embedding model used here is "**sentence-transformers/multi-qa-mpnet-base-dot-v1**" it has good performance however it is slow .

```
text_splitter = RecursiveCharacterTextSplitter(chunk_size=500, chunk_overlap=200)
texts = text_splitter.split_documents(documents)
st.session_state.vector_store = FAISS.from_documents(texts, embeddings_model)
```

Figure 4.3: indexing process

- **Retrieving** using vector similarity search , which is calculated by calculating the distance between the embedding vectors of the input and the document, then retrieving the top-k closest documents.

```python
# Simulate retrieving documents from a vector store
retrieved_docs = st.session_state.vector_store.similarity_search_by_vector(query_embedding, k=50)
```

Figure 4.4: retrieving function

Naive RAG is the simplest RAG framework, which makes it ideal for basic queries and quick prototypes, but after the observation and testing, we discovered that it struggles with complex queries, meaning that It performs well when handling simple queries that require retrieving the top results based purely on similarity search. With the aim of improving performance of Traditional RAG, we implemented an advanced RAG pipeline :

- **Hypothetical document embedding (HyDE):** is a method designed to enhance the retrieval accuracy in RAG systems by generating "hypothetical" documents based on user queries. Instead of directly searching for documents that match the query, HyDE utilizes an LLM to create a detailed, hypothetical document that represents an ideal answer to the query. This hypothetical document is then embedded into a vector space, and the system retrieves real documents that are semantically similar to this generated content. [51]

```python
def generate_hypothetical_document(query, api_url, model, groq_api_key):
    headers = {
        "Authorization": f"Bearer {groq_api_key}",
        "Content-Type": "application/json"
    }
    messages = [
        {
            "role": "system",
            "content": (
                "You are an expert at generating hypothetical documents that are highly relevant to a given query. "
                "Generate a concise and factual document (100-200 words) that would likely contain the answer to the query. "
                "Focus on specific details, entities, and numerical data if applicable."
            )
        },
        {
            "role": "user",
            "content": f"Generate a short document (100-200 words) that would likely contain the answer to this query: {query}"
        }
    ]
    data = {
        "model": model,
        "messages": messages,
        "temperature": 0.2,
        "max_tokens": 300
    }
    try:
        response = requests.post(api_url, headers=headers, json=data)
        response.raise_for_status()
        result = response.json()
        hypothetical_doc = result['choices'][0]['message']['content']
        logger.info(f"HyDE generated document: {hypothetical_doc}")
        return hypothetical_doc
    except Exception as e:
        logger.error(f"Error generating hypothetical document: {str(e)}")
        return query
```

Figure 4.5: HyDE function

- **Cross-Encoder Re-ranking :** it is a sophisticated technique used to enhance the relevance of research results by using the advanced language understanding capabilities of LLMs.[52] It's basically used as a second stage after an initially fast retrieval document step, ensuring that only the most relevant documents are represented to the user. The cross encoder model used is "**cross-encoder/ms-marco-MiniLM-L-6-v2**"

```python
if st.session_state.enable_reranking and st.session_state.reranker and filtered_docs:
    logger.info("Applying reranking...")
    reranker = st.session_state.reranker
    pairs = [[query, doc.page_content] for doc in filtered_docs]
    scores = reranker.predict(pairs)
    scored_docs = list(zip(filtered_docs, scores))
    scored_docs.sort(key=lambda x: x[1], reverse=True)
    filtered_docs = [doc for doc, score in scored_docs]
```

Figure 4.6: Re-ranking function

This is a comparison between the response method of the traditional RAG system and advanced RAG system. While answering a complex query.

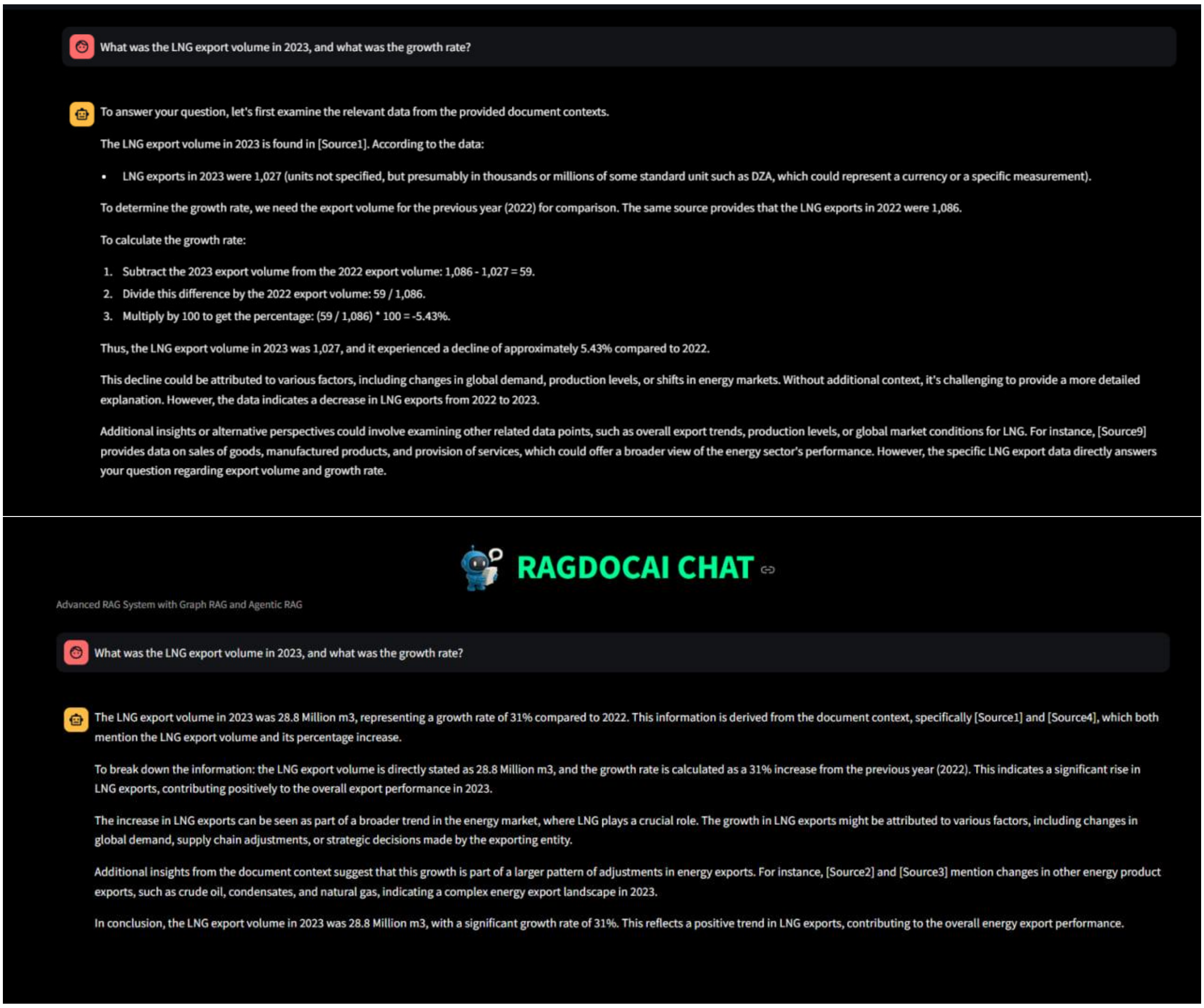


Figure 4.7: Traditional RAG Response VS Advance RAG Response

3.3. Graph RAG implementation

The core step in GraphRAG Pipeline is the extraction of the Knowledge Graph. To implement this step, we used Spacy to identify entities and relationships. Then using networkx python library we construct the knowledge graph. The importance of these steps lies in the fact that The retrieval is influenced by the graph. As the graph structure allows the system to retrieve not only direct relevant documents, but also neighboring nodes that may provide useful context, which improve relevance.

```python
def extract_entities_and_relations(text):
    nlp = st.session_state.nlp
    doc = nlp(text)
    entities = [(ent.text, ent.label_) for ent in doc.ents]
    relations = [(token.text, token.head.text, child.text)
                 for sent in doc.sents
                 for token in sent
                 if token.dep_ in ("nsubj", "dobj") and (child := [c for c in token.head.children
                 if c.dep_ in ("dobj", "pobj")][0]
                 if [c for c in token.head.children
                 if c.dep_ in ("dobj", "pobj")] else None)]
    return entities, relations

def build_knowledge_graph(entities, relations):
    G = nx.DiGraph()
    for entity, label in entities:
        G.add_node(entity, type=label, weight=1.0)
    for subj, verb, obj in relations:
        G.add_edge(subj, obj, relation=verb, weight=1.0)
    return G
```

Figure 4.8: knowledge graph building

- **Direct graph visualisation :**

The direct graph shown in the figure represents the outcome of the Knowledge Graph generation process using Neo4j to display the graph.

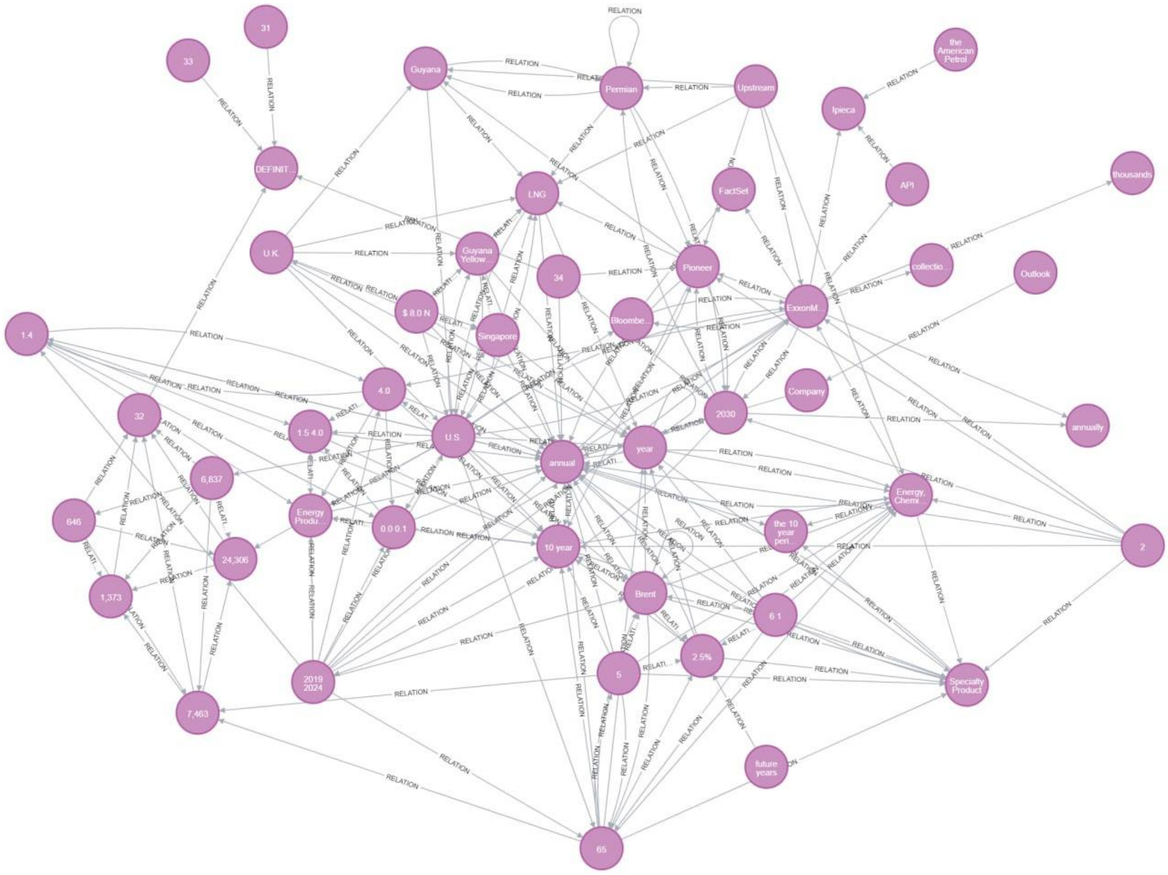


Figure 4.9: Direct graph visualisation

3.4. Agentic RAG implementation

Agentic RAG introduces an agent-based approach to retrieval, where the system dynamically processes and refines the query before retrieval.

```python
# Initialize question classifier with a lighter model
classifier = pipeline("zero-shot-classification", model="distilbert-base-uncased")

def classify_question(query):
    candidate_labels = ["financial", "table-based", "operational", "general", "future_plans"]
    result = classifier(query, candidate_labels, multi_label=False)
    return result['labels'][0]

def decompose_query(query):
    doc = nlp(query)
    sub_queries = []
    query_type = classify_question(query)

    # Avoid decomposing financial or table-based queries
    if query_type in ["financial", "table-based"]:
        return [query]

    for sent in doc.sents:
        sub_queries.append(sent.text)

    entities = [ent.text for ent in doc.ents]
    key_phrases = [chunk.text for chunk in doc.noun_chunks if chunk.text not in entities]

    sub_queries.extend([f"What is {entity}?" for entity in entities])
    sub_queries.extend([f"Details about {phrase}" for phrase in key_phrases])

    sub_queries = list(set([q for q in sub_queries if q.strip()]))
    return sub_queries if sub_queries else [query]
```

Figure 4.10: Classification and decomposition functions

This code snippet defines functions for classifying and decomposing natural language queries using a zero-threat classification model "distilbert-base-uncased", and also to break down a complex query into simpler sub-queries for further processing. The zero-shot classification allows the model to classify text into a predefined label without requiring task-specific training data. The model used is a lightweight version of BERT fine-tuned for text classification tasks.

- **The final generated answer by Agentic RAG**

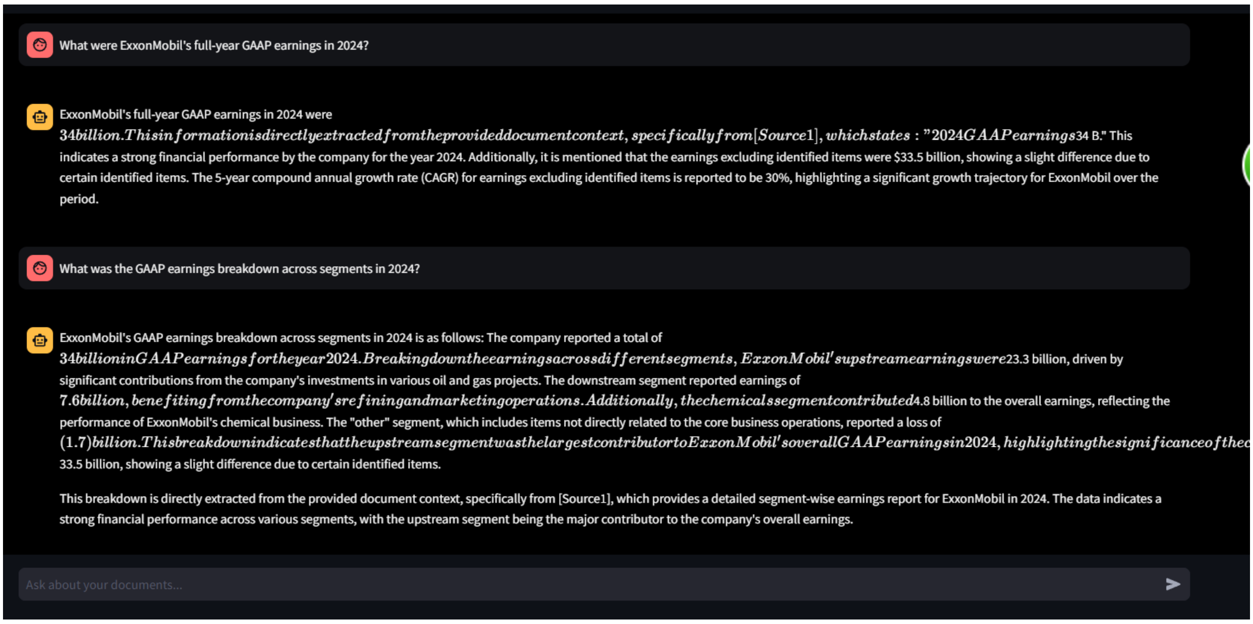


Figure 4.11: Final generated answer by Agentic RAG

4. Testing results

In order to test the accuracy of retrieval augmented generation mechanisms in our system, we attempt to generate manually Q&A pairs Based on documents of one of our case study.

	A	B
1	Question	Answer
2	Who was the Chairman & CEO of SONATRACH in 2023?	M. Rachid Hachichi
3	What was SONATRACH's export turnover in 2023, and how did it compare to 2022?	$49.8 billion in 2023, down from $59.3 billion in 2022
4	What was the average annual crude oil price in 2023?	$83.60 per barrel
5	What was SONATRACH's net profit in 2023?	675 billion Algerian Dinar (DA)
6	How much oil tax did SONATRACH pay in 2023?	5,678 billion DA, a 2% increase from 2022
7	What was the total primary hydrocarbon production in 2023?	194 million TOE
8	What was the increase in primary hydrocarbon production from 2022 to 2023?	2.2%
9	How many new hydrocarbon discoveries were made in 2023?	15 discoveries (14 in own effort, 1 in partnership)
10	What was the total production of refineries in 2023?	29.2 million tons
11	What was the volume of crude oil processed in 2023?	26.0 million tons
12	What major petrochemical project was launched in Arzew in 2023?	Polypropylene production complex with a capacity of 550,000 tons/year
13	What were the total hydrocarbon exports in 2023?	95 million TOE
14	What was the LNG export volume in 2023, and what was the growth rate?	28.8 million m³, an increase of 31%
15	By what percentage did flared gas volumes decrease between 2022 and 2023?	About 15%
16	What was the number of work accidents recorded in 2023?	426 accidents, a 4% decrease from 2022
17	What percentage of total investments was allocated to Exploration-Production in 2023?	82%
18	How many wells were completed in 2023?	32 (30 in own effort and 2 in partnership)
19	What was the total workforce at the end of 2023?	66,025 agents
20	What percentage of the workforce was female in 2023?	16.5%
21	How many employees received training in 2023?	28,614 agents

Sheet1

Figure 4.12: A sample of the Testing dataset

The accuracy shown in this table is measured as the percentage of questions for which the model generates a correct or fully matched answer based on retrieved contexts. Which is compared across different open source LLMs.

Using all technologies at once will improve accuracy and quality as accuracy has reached 96%. which is considered a significant improvement.

Table 4.2 Accuracy of different RAG mechanisms across LLMs

Model	Tradional RAG	Advanced RAG	Graph RAG	Agentic RAG
meta-llama/llama-4-maverick-17b-128e-instruct Accuracy	65.6%	78%	82%	79%
llama-3.3-70b-versatile Accuracy	70%	72%	78.63%	82.92%
deepseek-r1-distill-llama-70b Accuracy	75.95%	83%	87%	88%
Qwen/qwn3-32b Accuracy	78%	85%	89.07%	92%
Qwen 3:4b	55.03%	63%	72%	78.96%

5. Evaluation of The system

Large Language Models (LLMs) capture a vast amount of knowledge about the world, which allows them to answer questions without accessing any external sources. Answering a question then essentially involves retrieving relevant passages from a corpus and feeding these passages, along with the original question, to the LLM. While the usefulness of retrieval-augmented strategies is clear , as the overall performance will be affected by the retrieval model, the considered corpus, the LLM, or the prompt formulation, among others. [53] To evaluate our system we used Ragas , which is a framework for the automated assessment of retrieval augmented generation systems which also provides an integration with LangChain, making it easier to integrate it into our workflow.

We focus in particular four quality aspects, which we argue are of central importance:

- **Faithfulness** refers to the idea that the answer should be grounded in the given context. This is important to avoid hallucinations, and to ensure that the retrieved context can act as a justification for the generated answer. Indeed, RAG systems are often used in applications where the factual consistency of the generated text w.r.t. the grounded sources is highly important, e.g. in domains such as law, where information is constantly evolving. [53]
- **Answer relevance** refers to the idea that the generated answer should address the actual question that was provided. [53]
- **Context precision** is useful to ensure the retrieval of highly relevant information.
- **Context recall** measures how many of the relevant documents or pieces of information were successfully retrieved. Crucial where missing key information can be costly.

Table 4.3 The evaluation results

Models	Presicion	Recall	Faithfulness	Answer Relievancy
meta-llama/llama-4-maverick-17b-128e-instruct	0.73	0.70	0.73	0.803
llama-3-70b-versatile	0.80	0.74	0.70	0.85
deepseek-r1-distill-llama-70b	0.88	0.85	0.80	0.90
Qwen/qwen3-32b	0.9	0.87	0.82	0.92

6. Comparison with other model

The table below shows a comparison between RAGDOCAI Chat and the popular and powerful Cloude Chatbots that are available in the market today :

Table 4.4 Comparison between RAGDOCAI Chat and Cloude Chatbots

Feature	RAGDocAI Chat	Cloude Chatbots (ChatGPT ,Grok, Gemini,Cloud)
Model used	Open source models	Closed sources models
Latency	self-hosted	high API pricing
Inference cost	slow at the beginning while processing the files	fast
Identity behavior	uses blockchain agent	built in agents
Offline usage	fully offline usage	needs internet
Data privacy	high	mostly not private, and especially for using internal company usage, as data may be used for model improvement or stored and analyzed by the company itself
Knowledge resource	user-uploaded documents , models, APIs	Pre-trained knowledge and the internet
Customization options	full control (temperature, context, models, rag, settings)	limited

7. Development software

7.1. Programming languages

The chosen programming language is Python. Since its concise grammatical structure, and more important, it has a large number of standard libraries and third-party libraries, that's what makes it one of the ideal choices to use in the development of AI applications.

- **Python:** is an object-oriented and interpretive high-level programming language for general use. The version used here is 3.11.9.

Figure4.13: Python

7.2. Code editor

VS Code (Visual Studio Code): is a powerful code editor that supports development operations such as debugging, text running, and vision control. It is available for all operating systems such as Windows. The reasons for considering it as a choice for developing AI-powered applications are:

- Supports various programming languages, especially ones used in AI development, like R and Python.
- It offers extensibility via extensions and has built-in features like syntax highlighting, debugging, git integration, and terminal.
- It gives good performance and AI ecosystem support by supporting all major AI frameworks and libraries such as PyTorch and TensorFlow

Figure4.14: VS code

7.3. Database and application programming interface (API)

- **Neo4j:** it's high-performance graph store, that stores data as nodes (entities) and relationships (edges) which makes it a flexible network structure. It uses Cypher, which is a query language for graphs. It's ideal for use cases like GraphRAG.

Figure4.15: Neo4j

- **Groq API :** used as benchmark for fast and more powerful LLM. It is an API that offers you the access to open source LLMs like Llama4, Mistral, Qwen, and the newest released models. It's ideal for low latency inference applications.

Figure4.16: Groq API

- **FMP API :** is a Free Stock Market API and Financial Statements API that provides the most reliable and accurate financial data and real-time updates data and historical data.[54]
- **OLLAMA :** stands for Omnilayer Learning Language Acquisition Model. EasyMate is an AI tool designed to enable users to run LLMs locally on their machines by providing a user-friendly interface and it supports cross-platform compatibility.

Figure4.17: OLLAMA

7.4. Used models

These models were chosen simply because they were open source models and based on their features and ability in given high accuracy and robustness, Their ability of handling complex reasoning and inference, and having multilingual capability. These models are:

Table 4.5 The Models Collection

Model	Max Tokens	Host
meta-llama/llama-4-maverick-17b-128e-instruct	8192	groq API
llama-3.3-70b-versatile	32,768	groq API
deepseek-r1-distill-llama-70b	-	groq API
qwen/qwen3-32b	16,384	groq API
Qwen 3:4b	-	Ollama
llama3.1:latest	-	Ollama

7.5. Framework and libraries

- **Streamlit :** is an open-source framework for Python to build quick interactive apps. It has the ability, it has built-in widgets such as sliders, chat inputs، and buttons.

Figure4.18: Streamlit

- **Langchain :** is an open-source framework for building applications based on large-language models. It provides tools and abstractions to improve the customization, accuracy, and relevancy of the information the model generates.

Figure4.19: Langchain

- **Facebook AI Similarity Search (Faiss) :** is a vector indexing search library used for quick, efficient similarity search for embedding vectors of multimedia documents. It's mainly used in AI, especially for search retrieval, since it contains algorithms that search in sets of vectors of any size and even supporting code for evaluation. It also has custom indexing types, and its performance is ultra-fast, which makes it ideal for our requirements.

- **NumPy:** is a fundamental open source library in Python used for numerical computing on large volume data.

Figure4.20: NumPy

- **Pandas:** is fast, powerful, flexible, and easy-to-use open-source data analysis and manipulation tool built on top of the Python programming language.[55]

Figure4.21: Pandas

- **Spacy:** is a free open source library for natural language processing in Python, used to extract entities and relationships in texts.

Figure4.22: Spacy

- **Hugging Face :** is a company that maintains a huge open-source community of the same name that builds tools, machine learning models, and platforms for working with artificial intelligence with a focus on data science, machine learning, and natural language processing. Hugging Face is notable for its NLP Transformers library and a platform that allows users to share models and datasets.[56]

Figure4.23: Hugging face

- **Transformers** is a library of pre-trained natural language processing computer vision, audio, and multi-models for inference and training. Use Transformers to train models on your data, build inference applications, and generate text with large language models.[57]

Figure4.24: Transformers

- **NetworkX :** is a Python package for the creation, manipulation, and study of the structure, dynamics, and functions of complex networks.[58]

Figure4.255: NetworkX

- **Torch :** It's a package containing data structure for multidimensional tensors and defines mathematical operations over these tensors. Additionally, it provides many utilities for efficient serialization of tensors and arbitrary types and other useful utilities.[59]

8. Conclusion

In conclusion, we presented the implementation of our proposed Chatbot LLM-powered document assistant, a system designed to improve employee speed and efficiency in accessing knowledge from thousands of internal documents. The system, named RAGDocAI Chat, has demonstrated promising results despite limited resources and other constraints such as embedding models and the need of GPU. In both testing and evaluation phases, it achieved 96% accuracy, delivering context-rich answers extracted directly from relevant internal documents, without the need for fine-tuning the underlying language model. The integration of retrieval-augmented generation technologies made the system highly adaptable to a variety of diverse knowledge domains. For Algerian companies, where AI adoption remains relatively limited, the system offers acceptable to good performance, especially due to the predominance of French in internal documents, which align well with the language capability of the base model.

General Conclusion

The thesis addresses the topic of introducing an intelligent document assistant as a digital companion that could not only understand and summarize an overview for you, but also communicate in intelligent conversations about their content.

RAGDocAI Chat is a chatbot that leverages the power of large-language models (LLMs) and the Retrieval Augmented Generation mechanisms, which will change how employees interact with internal company documents. **RAGDocAI Chat** was specially designed to deliver insightful, rich-context-aware answers, all without the need to directly interact with the documents.

We demonstrated the application of **RAGDocAI Chat** in a realistic setting to validate its effectiveness within the context of Algeria companies such as **Sonelgas** and **Sonatrach**. The results showed that **RAGDocAI** with a user-friendly interface maintained high accuracy and responsiveness with minimal delay, even with limited technical infrastructure, Resources and the challenges we faced such as mixed language documents, inconsistent formatting, and the availability of documents because both national and private companies have their own conditions and policies for sharing and storing their internal documents.

In the other hand, when it comes to future works and enhancements, we are considering enhancing the system's performance by adding Knowledge Augmented Generation (KAG) which introduces a professional domain knowledge services framework with better generation and knowledge reasoning. The addition of good and strong GPU will improve the speed and the accuracy of the system. And thus, the overall performance of the system will be improved. Furthermore, fine-tuning the system on a domain-specific data set consisting of company and internal documents.

To ensure users have access to the most current information, we plan to incorporate web scrapping capabilities.finaly we would like to add features like the feedback loop which will enhance the user experience and make it more adaptable to the user's styles and understanding of its requirements.

Bibliography

[1] geeksforgeeks, "Artificiel Intelligent Tutorial," [Online]. Available: https://www.geeksforgeeks.org/artificialintelligence/?ref=shm. [Accessed 1 January 2025].

[2] geeksforgeeks, "Machine Learning," [Online]. Available: https://www.geeksforgeeks.org/machine-learning/?ref=shm . [Accessed: 12 Jan. 2025].

[3] M. S. A. Alzubaidi et al, "Machine learning: Algorithms, real-world applications, and research directions," Journal of King Saud University - Computer and Information Sciences, pp. 18–28. Avaible: https://link.springer.com/article/10.1007/s42979-021-00592-x

[4] Zomev, "Deep Neural Network (DNN) Explained," Medium, Jan. 2024. [Online]. Available: https://medium.com/@zomev/deep-neural-network-dnn-explained -. [Accessed: 14 Feb. 2025].

[5] R. Singh, "Decoding CNNs: A Beginner's Guide to Convolutional Neural Networks," Medium, 30 Dec. 2024. [Online]. Available: https://ravjot03.medium.com/decoding-cnns-a-beginners-guide-to-convolutional-neural-networks-and-their-applications-1a8806cbf536 . [Accessed: 23 May 2025].

[6] P. English, "Friendly Introduction to Deep Learning Architectures: CNN, RNN, GAN," Medium, Jan. 2024. [Online]. Available: https://python.plainenglish.io/friendly-introduction-to-deep-learning-architectures-cnn-rnn-gan-transformers-encoder-decoder-b11334e4cdf7 . [Accessed: 14 Feb. 2025].

[7] C. Olah, "Understanding LSTMs," Colah's Blog, Aug. 2015. [Online]. Available: https://colah.github.io/posts/2015-08-Understanding-LSTMs/#fn1 . [Accessed: 14 Feb. 2025].

[8] geeksforgeeks, "Natural Language Processing (NLP)," [Online]. Available: https://www.geeksforgeeks.org/natural-language-processing-nlp-tutorial/?ref=shm . [Accessed: 21 Jan. 2025].

[9] J. H. M. D. Jurafsky, "Speech and Language Processing," 2023. [Online]. Available: https://web.stanford.edu/~jurafsky/slp3/ed3book.p[]d. [Accessed 6 jan 2025].

[10] ResearchGate, Concepts in Computer Science and Artificial Intelligence, 2024. [Online]. Available: http://neuralnetworksanddeeplearning.com/. [Accessed 2025 Fab 10].

[11] NVIDIA, " Large Language Model eBooks ", NVIDIA Resources ,2025. [Online]. Available: https://resources.nvidia.com/en-us-large-language-model-ebooks. [Accessed 20 Fab 2025].

[12] D. McCandless, " The Rise of Generative AI Large Language Models (LLMs) like chatgpt ", Information is Beautiful, 20 Dec 2023. [Online]. Available: https://informationisbeautiful.net/visualizations/the-rise-of-generative-ai-large-language-models-llms-like-chatgpt/. [Accessed 20 May 2025].

[13] J. Li, " The evolution, applications, and future prospects of large language models: An in depthoverview ", January 2024. [Online]. Available: https://www.researchgate.net/publication/377932211_The_evolution_a.

[14] N. Tomar, " A Brief History of Large Language Models (LLMs) ", 14 September 2024. [Online]. Available: https://idiotdeveloper.com/a-brief-history-of-large-large-language-models-llms/. [Accessed 19/ 02/ 2025].

[15] IBM, " Large Language Models ", IBM Think, 2024. [Online]. Available: https://www.ibm.com/think/topics/large-language-models. [Accessed 5 Feb 2025].

[16] SolGuruz, " Prompt Engineering ", SolGuruz Generative AI Wiki, 2024. [Online]. Available: https://solguruz.com/generative-ai/prompt-engineering/. [Accessed 14 Feb 2025].

[17] S. Data, " Large Language Models 101: History, Evolution and Future ", Scribble ,May 2023. [Online]. Available: https://www.scribbledata.io/blog/large-language-models-history-evolutions-and-future/. [Accessed 16 Feb 2025].

[18] Yellow.ai, " Large language models: Definition, types & use cases ", Yellow.ai Blog, 2024. [Online]. Available: https://yellow.ai/blog/large-language-models/#types-of-large-. [Accessed 17 Feb 2025].

[19] GeeksforGeeks, " Large Language Model (LLM)", GeeksforGeeks, 2024. [Online]. Available:

https://www.geeksforgeeks.org/large-language-model-llm/?ref=header_outind. [Accessed 5 Feb 2025].

[20] W. Zhang, " Transformer Architecture & LLMs: Zero to Hero ", Medium, Jan 2024. [Online]. Available: https://medium.com/@waylandzhang/transformer-architecture-llms-zero-to-hero-98b1ee51a838. [Accessed 08 Feb 2025].

[21] L. Voita, " Seq2Seq and Attention ", Lena Voita's NLP Course, 2024. [Online]. Available: https://lena-voita.github.io/nlp_course/seq2seq_and_attention.html#transformer_model_architecture. [Accessed 18 Feb 2025].

[22] TechTarget, "12 of the best large language models ", TechTarget, 2024. [Online]. Available: https://www.techtarget.com/whatis/feature/12-of-the-best-large-language-models. [Accessed 09 Feb 2025].

[23] xAI, " About xAI ", xAI, 2024. [Online]. Available: https://x.ai/about. [Accessed 14 Feb 2025].

[24] M. H. Belle, "Addison-Wesley," Quick Start Guide to Large Language Models: Strategies and Best Practices, 2024. [Online]. Available: https://archive.org/details/quick-start-guide-to-large-language-models-strategies-and-. [Accessed 10 Feb 2025].

[25] O. Shone, "5 key features and benefits of large language models ", The Microsoft Cloud, Oct 2024. [Online]. Available: https://www.microsoft.com/en-us/microsoft-cloud/blog/2024/10/09/5-key-features-and-benefits-of-large-language-models/?mscm=1. [Accessed 11 Feb 2025].

[26] IBM, " Generative models ", IBM Think, 2024. [Online]. Available: https://www.ibm.com/think/topics/generative-model. [Accessed 13 Feb 2025].

[27] A. Kimothi, " Retrieval-Augmented Generation: A Simple Introduction ", Scribd,2024. [Online]. Available: https://www.scribd.com/document/789710040/RETRIEVAL-AUGMENTED-GENERATION-A-SIMPLE-INTRODUCTION. [Accessed 2025 May 25].

[28] Azharudeen, " RAG vs Graph RAG ", Medium, 2024 Jul 11. [Online]. Available: https://medium.com/@azharudeen2020/rag-vs-graph-rag-edd2f35761dc. [Accessed 15 May 2025].

[29] M. Hunger, " What Is GraphRAG ", Neo4j Blog,5 Dec. 2024. [Online]. Available: https://neo4j.com/blog/genai/what-is-graphrag/. [Accessed 25 May 2025].

[30] J. Pörschmann, " Graph RAG — A Conceptual Introduction ", Medium, Aug,22 , 2024. [Online]. Available: https://medium.com/data-science/graph-rag-a-conceptual-introduction-41cd0d431375. [Accessed 25 May 2025].

[31] A. Verma, " Beyond Simple Retrieval: Diving Deep into Agentic RAG and Its Advantages Over Traditional RAG ", Medium ,25 Jan 2025. [Online]. Available: https://medium.com/@ajayverma23/beyond-simple-retrieval-diving-deep-into-agentic-rag-and-its-advantages-over-traditional-rag-3b5f72067f32. [Accessed 25 May 2025].

[32] E. C. a. L. Monigatti, " What Is Agentic RAG ", Weaviate Blog, 5 Nov 2024. [Online]. Available: https://weaviate.io/blog/what-is-agentic-rag. [Accessed 25 May 2025].

[33] GeeksforGeeks, " What is Agentic RAG?, " GeeksforGeeks,01 May 2025. [Online]. Available: https://www.geeksforgeeks.org/what-is-agentic-rag/. [Accessed 25 May 2025].

[34] J. Morrow, " Understanding LLMs and Generative AI: Differences, Benefits, and Uses ", GPTBots.ai Blog, Oct 2024. [Online]. Available: https://www.gptbots.ai/blog/llm-vs-. [Accessed 16 Feb 2025].

[35] Itarian, " Business Reports: What are they and how to write them ", Itarian, [Online]. Available: https://www.itarian.com/it-101/business-reports/. [Accessed 2025 April 09].

[36] Clearinfo, "24 Types of Business Reports With Samples & Writing Structure, " Clearinfo,[Online]. Available: https://clearinfo.in/blog/types-of-business-reports/. [Accessed 2005 April 09].

[37] J. La, "20 Types of Reports and When to Use Them (Plus Templates), " Piktochart, 21 Mar 2025. [Online]. Available: https://piktochart.com/blog/types-of-reports/#financial. [Accessed 15 april 2025].

[38] M. Kaminsky, "5 Business Reports Every Company Needs ", LegalZoom,01 Feb 2023. [Online]. Available:

https://www.legalzoom.com/articles/5-business-reports-every-company-needs#2-sales-and-revenue-report. [Accessed 16 April 2025].

[39] K. Rabkin, "7 types of business reports you need to know ", PandaDoc, 06 Aug 2024. [Online]. Available: https://www.pandadoc.com/blog/types-of-business-reports/. [Accessed 09 May 2025].

[40] Itarian, " Business Reports: What are they and how to write them ", Itarian, [Online]. Available: https://www.itarian.com/it-101/business-reports/. [Accessed 21 April 2025].

[41] Sonatrach, " L'Énergie pour un Développement Durable ", Sonatrach, [Online]. Available: https://sonatrach.com. [Accessed 27 April 2025].

[42] Sonelgaz, "Sonelgaz,," Sonelgaz,, [Online]. Available: https://www.sonelgaz.dz. [Accessed 27 April 2025].

[43] E. M. Corporation, " Exxon Mobil Corporation ", ExxonMobil, [Online]. Available: https://corporate.exxonmobil.com/. [Accessed 30 April 2025].

[44] IBML, " Data Extraction Techniques & Methods: Exploring Your Options ", IBML, [Online]. Available: https://www.ibml.com/blog/data-extraction-techniques-methods-exploring-your-options/. [Accessed 30 April 2025].

[45] AIIM, " What is Enterprise Content Management (ECM)?, ", AIIM, [Online]. Available: https://www.aiim.org/Resources/Glossary/Enterprise-Content-Management. [Accessed 30 April 2025].

[46] INORU, " The Role of Large Language Models in Business ", INORU ,[Online]. Available: https://www.inoru.com/blog/the-role-of-large-language-models-in-business/. [Accessed 5 May 2025].

[47] A. Taylor, " How real-world businesses are transforming with AI ", Microsoft Official Blog,22 Apr 2025. [Online]. Available: https://blogs.microsoft.com/blog/2025/04/22/https-blogs-microsoft-com-blog-2024-11-12-how-real-world-businesses-are-transforming-with-ai/. [Accessed 03 May 2025].

[48] C. X. Wood, "5 AI Case Studies in Customer Service and Support, " ,VKTR, 24 Jul 2024. [Online]. Available: https://www.vktr.com/ai-disruption/5-ai-case-studies-in-customer-service-and-support/. [Accessed 03 May 2025].

[49] E. M. Corporation, " Annual Reports & Proxy ", ExxonMobil Investor Relations, [Online]. Available: https://investor.exxonmobil.com/company-information/annual-reports-proxy. [Accessed 30 May 2025].

[50] H. K. a. L. D. A. Rane, " Hugging Face Datasets " ,financial-qa-dataset, 2024. [Online]. Available: https://huggingface.co/datasets/adityarane/financial-qa-dataset. [Accessed 01 Jun 2025].

[51] DhanushKumar, " RAG Series V : Hypothetical Document Embeddings (HyDE), " medium, 12 Feb 2025. [Online]. Available: https://medium.com/@danushidk507/rag-series-v-hypothetical-document-embeddings-hyde-e974d35ed688. [Accessed 01 Jun 2025].

[52] NVIDIA, " Enhancing RAG Pipelines with Re-Ranking ", NVIDIA Technical Blog, [Online]. Available: https://developer.nvidia.com/blog/enhancing-rag-pipelines-with-re-ranking/ . [Accessed 01 jun 2025].

[53] J. J. L. E. a. S. S. S. Es, "RAGAS: Automated Evaluation of Retrieval Augmented Generation,," Sep 2023. [Online]. Available: https://arxiv.org/abs/2309.15217.

[54] FinancialModelingPrep (FMP), " Financial Modeling Prep – Financial data API ", FinancialModelingPrep,[Online]. Available: https://site.financialmodelingprep.com/.

[55] p. d. team, " pandas – Python Data Analysis Library ", pandas,5 Jun 2025. [Online]. Available: https://pandas.pydata.org/.

[56] C. Stryker, " What is Hugging Face?, " IBM Think,2 May 2025. [Online]. Available: https://www.ibm.com/think/topics/hugging-face. [Accessed 04 Jun 2025].

[57] H. Face, " Transformers -Python library for pretrained model ", Hugging FaceDocs, [Online]. Available: https://huggingface.co/docs/transformers/en/index. [Accessed 04 Jun 2025].

[58] D. A. S. a. P. J. S. A. A. Hagberg, " NetworkX 3.5: Software for the creation,manipulation, and study of complex networks ", NetworkX 3.5,29 May 2025. [Online]. Available: https://networkx.org/. [Accessed 03 Jun 2025].

[59] P. d. team, " torch- PyTorch 2.7 documentation ", PyTorch Docs,2025. [Online]. Available: https://docs.pytorch.org/stable/torch.html. [Accessed 04 Jun 2025].

Abstract

In the business world, documents play a foundational role for companies, especially given its influence on the company's future in the term of decision-making, business progress, financial statements, its profits, and its long-term survival in the market.

With the company's growth and evaluation, employees found themselves with the need of speed and quality to efficiently handle the documents, requiring their concentration, effort, and considerable time, which occasionally leads to unsuccessful outcomes. As a solution for this problem, we proposed RAGDocAI Chat, which is a chatbot that offers a user-friendly interface that lets the employees or users effectively interact with the internal documents of their company, with comparable performance and accuracy to models such as ChatGPT and Grok, and deepseek using the power of large language models(LLMs) to generate insightful, infered answers based on the documents and the user's query in natural language.RAGDocAI also use Retrieval Augmented Generation technologies such as traditional RAG, Advanced RAG, GraphRAG, and AgenticRAG to build a knowledge domain from these documents, reducing the hallucination of the LLM due to its limited knowledge of those domains or not very familiar with them.

Keywords: LLMs, Query, Documents, RAG, Advanced RAG, Graph RAG, Agentic RAG , Knowledge.

الملخص

في عالم الأعمال، تلعب الوثائق دورًا أساسيًا للشركات، خاصةً بالنظر إلى تأثيرها على مستقبل الشركة من حيث اتخاذ القرار، وتقدم الأعمال، والبيانات المالية، وأرباحها، وبقاءها لفترة طويلة في السوق. مع نمو الشركة وتقييمها، وجد الموظفون أنفسهم في حاجة إلى السرعة والجودة للتعامل بكفاءة مع الوثائق، مما يتطلب تركيزهم وجهودهم ووقتًا كبيرًا، مما يؤدي أحيانًا إلى نتائج غير ناجحة. كحل لهذه المشكلة، اقترحنا RAGDocAI Chat **، وهو برنامج محادثة يقدم واجهة سهلة الاستخدام تتيح للموظفين أو المستخدمين التفاعل بشكل فعال مع الوثائق الداخلية لشركتهم، مع** دقة **و أداء قريب نسبيا من الأداء الذي تقدمه نماذج مثل** ChatGPT **و** Grok **و** DeepSeek**، باستخدام قوة نماذج اللغة الكبيرة** (LLMs) **لتوليد إجابات دقيقة و ذات سياقية عالية مستندة إلى الوثائق واستفسار المستخدم بلغة طبيعية.يستخدم** RAGDocAI **أيضًا تقنيات**Retrieval Augmented Generation **مثل** Traditional RAG **و** Advanced RAG**و** GraphRAG **و** AgenticRAG **لبناء مجال معرفي من هذه الوثائق، مما يقلل من الهلوسة في نماذج اللغة الكبيرة بسبب معرفتها المحدودة بتلك المجالات أو غير المألوفة لها .**

الكلمات الرئيسية: **نماذج اللغة الكبيرة** (LLM) ، **الاستعلام، الوثائق،** RAG ، Advanced RAG ، GraphRAG، Agentic RAG، **المعرفة** (Knowledge) .

Résumé

Dans le monde des affaires, les documents jouent un rôle fondamental pour les entreprises, surtout compte tenu de leur influence sur l'avenir de l'entreprise en matière de prise de décision, de progression commerciale, d'états financiers, de profits et de survie à long terme sur le marché. Avec la croissance et l'évaluation de l'entreprise, les employés se sont retrouvés avec le besoin de rapidité et de qualité pour gérer efficacement les documents, ce qui nécessite leur concentration, leurs efforts et un temps considérable, ce qui conduit parfois à des résultats infructueux. Comme solution à ce problème, nous avons proposé RAGDocAI Chat, un chatbot qui offre une interface conviviale permettant aux employés ou utilisateurs d'interagir efficacement avec les documents internes de leur entreprise, avec une performance et une précision comparables à des modèles tels que ChatGPT, Grok et deepseek, utilisant la puissance des modèles de langage de grande taille (LLMs) pour générer des réponses perspicaces et inférées basées sur les documents et la requête de l'utilisateur en langage naturel. RAGDocAI Chat utilise également la récupération Les technologies de génération augmentée telles que RAG traditionnel, Advanced RAG, GraphRAG et AgenticRAG pour construire un domaine de connaissance à partir de ces documents, réduisant l'hallucination du LLM en raison de sa connaissance limitée de ces domaines ou de sa méconnaissance.

Mots-clés : LLMs, Requête, Documents, RAG, Advanced RAG, GraphRAG, Agentic RAG, Connaissance.