TurnitinPDF.pdf

by Neti Neti

Submission date: 18-Jan-2025 06:41AM (UTC+0000)
Submission ID: 248886089
File name: TurnitinPDF.pdf (7.7M)
Word count: 23727
Character count: 136645

Enhancing Retrieval for Fact-Based Question Answering on Cancer Literature Using RAG & Pandas AI Techniques

Neti Theertha Bhaskara Sri Sai
Student ID: 1130075

Thesis Report for
Master of Science in Data Science

Liverpool John Moores University & UpGrad

December 2024

ACKNOWLEDGEMENT

I would like to express my deepest gratitude to my advisor, Vikas Thada, for his help, support, guidance, and insightful and timely feedback throughout this research journey. I sincerely thank LJMU & UpGrad for the studentship that allowed me to conduct this thesis.

ABSTRACT

The advancement made in Generative AI according to today's trend is simply amazing. The evolution of Neural networks to Large Language Models played a key role in the stream of Data Science. The key techniques that helped to achieve broad success in the NLP stream are NLP techniques, Deep Learning techniques, etc such as transfer learning, contextual learning, finetuning, and scalability. Development to that present Generative AI has the capability to perform Text, Images, audio, and video generation respective to the desired domain. However, there are some gaps in the text generation of the Generative AI field, they are maintaining correlation between word to word in large generated texts when the user asks a query to any Large Language Model. Also lacking in applying common sense, and creative answer generation. It is also needed to focus on ethical values and bias considerations according to context. For instance, understanding of full-length documents is lagging behind, if a document consists of textual data with subheadings on one page and extended information on another page. In this scenario, GPT is failing to correlate that gap. And in some cases, like structured GPT the statistical approach of responding to questions is weak. Also, complex queries like the information present in multiple sheets of any file, it's not able to correlate. Table content information correlation also needs to improve more. Image understanding or Image-to-text conversion accuracy needs to be improved. This research mainly focuses on dealing with different formats of data such as Word, PDF, xlsx & CSV documents that can be pre-processed effectively. The first approach will be based on unstructured data like Word documents & pdf, the context obtained from pre-processed data will be converted to multidimensional vector embeddings. Converted vector embeddings will be stored in a database called FAISS. This research work proposes the implementation of an RAG pipeline, which will help to retrieve the most relevant information from the stored vector embeddings to the user query asked. Also, the second proposed model for structured data can be achieved by converting user queries to SQL queries and then searching on top of the stored structured data after data preprocessing. Utilization of LLM endpoint URLs of GPT 4 for building the generative conversation will serve the purpose of the proposed model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

AI: Artificial Intelligence

API: Application Programming Interface

BERT: Bidirectional Encoder Representation from Transformers

CBIO: Cancer Biology

CNN: Convolution Neural Network

CSV: Comma Separated Value

CT: Computed Tomography

DB: Database

DNA: Deoxyribose Nucleic Acid

EDA: Exploratory Data Analysis

FAISS: Facebook AI Similarity Search

GAI: Generative Artificial Intelligence

GAN: Generative Adversarial Network

GIT: Global Information Tracker

GNN: Graphical Neural Network

GPT: Generative Pre-trained Transformer

ICCR: International Collaboration on Cancer Reporting

IoT: Internet of Things

IOMT: Internet of Medical Things

KNN: K Nearest Neighbour

LLAMA: Large Language Model Meta AI

LLM: Large Language Model

LSTM: Long Short-Term Memory

ML: Machine Learning

MRI: Magnetic Resonance Imaging

NLP: Natural Language Processing

NLTK: Natural Language Tool Kit

OCR: Optical Character Recognition

PCA: Principle Component Analysis

PDF: Portal Document Format

PET: Positron Emission Tomography

PyPDF: Python PDF

RAG: Retrieval Augmented Generation

RAM: Random Access Memory

RNN: Recurrent Neural Network

SQL: Structured Query Language

SVM: Support Vector Machine

TAG: Table Augmented Generation

TF-IDF: Term Frequency Inverse Document Frequency

TMB: Tumour Mutational Burden

UI: User Interface

URL: Uniform Resource Locator

VAE: Variational Auto Encoders

VR: Virtual Reality

XLSX: Excel Open XML Spreadsheet

CHAPTER 1: INTRODUCTION

Generative AI refers to a type of artificial intelligence that can create content, such as text, images, or music, by learning from existing data. It uses patterns and information it has learned to generate new or similar pieces of content that have never been seen before. This technology is used in various sectors to enhance creativity and efficiency. The main objective of this research deals with how effectively this research work can do the collection of data related to cancer which is in form of structured and unstructured. Leveraging of existing Large Language Models available in the market in this scenario can be done with ease. Major concepts that contributed to LLM building are mathematical concepts of Calculus, Linear Algebra, Probability, and Statistics along with Information theory also played an important role in the field of Data Science. In addition to that RNN, LSTM, Encoder & Decoder Networks, and Gated Recurrent Units played a unique role in the Generative AI field. Also, Sequence to Sequence Models and Transformers laid a strong base for Large Language Models.

1.1 Background of the Study

YuWang and other authors (Wang et al., 2023a) this paper includes the algorithms like Knowledge Graph of the Prompting method which is multiple documents answering the questions asked and the graph traversal algorithm for reasoning and retrieving passages. It shows potential direction for further research in understanding and leveraging the capabilities of LLMs in the context of knowledge graph traversal and question answering.

The limited context understanding and lack of fine-grained language understanding are the key problems identified in the era of Natural Language Processing (NLP). Earlier models like rule-based systems, n-gram models, handcrafted rules, linguistic patterns, and statistic probabilities are used to generate text. To overcome these issues, the research work proposes Retrieval Augmented Generation which will have the strength of the retrieval-based technique along with generative models which also improves the quality of the generated text and takes care of relevancy accordingly.

The large language models are mainly built on deep learning techniques, GAN models, and transformer-based architectures which are trained on more than 100 million parameters to a billion parameters. LLM can perform tasks like conversational agents, Language Modelling/Translation, Text classification, summarization & Question Answering based on the content provided. Also, this research cannot use directly the Large Language model instead this research will use their APIs as the model will be very huge, and require high computational power to run. Memory Requirements will be very big and also trained on massive datasets. Storing and accessing that much huge data is simply impractical.

Yixuan Tang and Yi Yang (Tang and Yang, 2024a) proposed a paper that aims to use a Multi-Hop RAG dataset with a large collection of multi-hop queries and respective ground truth answers. Evaluating responses involves measuring the reasoning ability of the Large Language Model by contrasting the LLM's response against the query's ground truth answer.

The first approach will be performed on unstructured data by initiating preprocessing techniques on the data with the help of Python. Libraries like pypdf will extract text from the document. This extracted text will be stored in the memory buffer. The extracted text will be converted to n-dimensional vector embeddings using some of the Python open-source libraries Ollama embeddings which can be leveraged by langchain community python library. Converted embeddings are stored in a database of FAISS. This helps in storing embedded vectors. By using langchain as a retrieval chain between the FAISS database and LLM this research work will implement a Similarity search based on the user context provided and the whole embedded context stored in FAISS. By using the top_k=3 value, it will fetch the top three nearest matches with probability score. Here the input query data will also be encoded.
The second approach of this research work is dealing with structured data like Excel files. After completion of preprocessing, this research proposes to send the data to the Python pandas AI library as input. Pandas' AI library itself will do the searching mechanism by converting the data and input query to SQL language then performing the operations for the query the user asks. It has an extraordinary feature in that it incorporates LLM API for generating the response more accurately, also it gives text answers and flow charts for the statistical user queries asked by the user.

1.2 Related Research

Generative AI stands as a transformative branch of artificial intelligence, that focuses on creating new content, ranging from images and text to music. This innovative approach leverages complex algorithms and models, such as Generative Adversarial Networks and Variational Autoencoders. It will produce outputs that can mimic human-like creativity. Generative AI learns from vast datasets by extracting patterns, styles, and structures to generate novel creations that were previously unseen.

This technology has found applications in diverse fields, including text generation for the querying user asks, digital art generation, realistic video game environments, and personalized content creation. In the development of synthetic data for training other AI/ML models. By pushing the limits of what machines can create, generative AI not only enhances digital systems capabilities but is also used for human-machine collaborations. The below table displays a comprehensive overview of relevant studies and articles as follows:

Table 1.2.1 Initial Research Papers Analysis

Methods	Approach	Limitations	Reference
SOP-GPT framework	Enhancing the capabilities of AI agents based on AI-generated content by detailing the process, role, and skill that is utilized in news, education, poetry, and entertainment fields.	Addressing the problem of handling complex tasks with existing AI Agent systems based on Artificial Intelligence-Generated Content.	(Yao et al., 2023a)
Sach's and multiple queries method	Introducing the novel method that utilizes both machine learning and domain-specific literature to recover causal graphs more effectively.	Statistical methodologies are suffering from biases and LLMs lack domain knowledge and reasoning skills.	(Zhang et al., 2024)
Variational Autoencoders, score-based generative models, and stochastic differential equations	Exploring the theoretical and mathematical foundations of General Artificial Intelligence models reveals their capability to perform tasks, and significant challenges, and explains the future of GAI development.	Hallucinated responses, Misinformation, relevancy of answer and explainability	(Staphord Bengesi and Hoda El-Sayed, n.d.)
Encoder Decoder framework	Diverse applications using GAI in medical and drug discovery fields emphasize	Data privacy, bias, ethical consideration, and	(Sai et al., 2024)

and Inception score etc.	the importance of explainability, efficiency, and accuracy.	generalization on unseen data	
Data completeness Augmented Algorithm and spectral normalization GAN etc.	It emphasizes the importance of improving data completeness to enhance the performance of data-driven smart healthcare systems.	Data problems such as data imbalance, data bias, data usage optimization, etc.	(Lan et al., 2023)
The use of LLM includes classification algorithms such as supervised, reinforcement, and transfer learning.	To leverage the capabilities of Large Language Models by enhancing the EFT cue generation process to improve health outcomes through better decision-making influenced by positive future event visualization.	Alignment of LLMs with User Intention and impact of episodic future thinking (EFT) etc	(Ahmadi and Fox, 2024)

1.3 Research Questions

The following research questions will enable us to conduct a more thorough and detailed investigation into this study they are:

- How can Large Language Models be fine-tuned to incorporate domain-specific knowledge effectively, especially in the field of health care?
- In what ways can the integration of multimodal data (e.g., text, images, tables) enhance the capabilities of LLMs in understanding and generating more comprehensive and contextually relevant responses?
- How can transfer learning techniques be leveraged to enhance the performance of Large Language Models in niche applications or languages with limited training data available?
- How can the Retrieval-Augmented Generation (RAG) technique be effectively integrated with Large Language Models to improve the accuracy and contextuality of responses generated from both structured and unstructured datasets?

- How does Facebook AI Similarity Search (FAISS) perform in real-world scenarios when conducting similarity searches in large datasets?

1.4 Aim and Objectives

This research aims to attempt on Retrieval Augmented Generation technique with a Large Language model to perform Question Answering based on the context provided. Further, this research work proposes to compare and review the output responses for structured data which uses pandas AI, and for unstructured data which is the RAG model.

Based on the aim mentioned, the following objectives have been formulated:

- Using the data preprocessing techniques, extracting the content from cancer documents for both structured and unstructured.
- Using the Python libraries, convert text data to vector embeddings and store it in the database.
- Using the RAG technique and advanced Python libraries similarity search will be implemented for the query asked by the user along with embedded data.
- Checking the correctness, performance, and accuracy of the Large Language Model for contextual-based text responses.

1.5 Significance of the Study

- By leveraging LLMs to analyze vast amounts of cancer-related data, this study contributes to a deeper understanding of cancer types, treatments, patient outcomes, and potentially unidentified patterns or correlations within the data.

- The integration of LLMs with cancer data analytics can significantly enhance the capabilities of clinical decision support systems, providing healthcare professionals with more accurate, timely, and data-driven insights for patient care.

- This research simplifies how users ask questions about complicated medical data, making it easier for people without special training to get trustworthy information. This way, patients, their families, and anyone interested can learn about things that used to be known only by experts.

- The methodologies developed through this research have the potential for global application, addressing disparities in cancer care and research capabilities. By providing scalable and adaptable tools, this work supports the advancement of oncology care in low-resource settings, contributing to the broader goal of global health equity.

1.6 Scope of study

Due to the limitation of time frame, the scope of the research will be limited as below:

- Gathered datasets on cancer both structured and unstructured, from the open-source ICCR & CBIO portals. Creation of Augmented or Synthetic data generation for cancer datasets will not be possible.
- This research includes the utilization of LLM APIs which will surpass older NLP techniques in Document Question Answering systems due to their superior contextual understanding, ability to leverage pre-trained knowledge for efficiency, and flexibility across languages and tasks.
- Using Transformer or Deep learning (DL) models is out of scope because it is context-dependent, often driven by practical considerations such as computational resources, data privacy, the need for interpretability, task specificity, etc.
- In Data Preprocessing this research work will not include the extraction of text from images using OCR and table as it will not be feasible in the limited time frame.
- Quantitative Evaluation of the generated text responses is beyond the scope of this study because there is difficulty in evaluating the quality and accuracy of LLM-generated text, which is influenced by factors like context, ambiguity, and subjective interpretation.

1.7 Structure of the study

The structure of the study is as follows:

Chapter 1 – Introduction: This chapter provides an introduction and background for this research work.

Chapter 2 – Literature Review: This chapter mentions the importance of Generative AI-related applications and works in the fields of Education, Health Care, Marketing, and the IT industry.

Chapter 3 – Methodology: This chapter gives a detailed walkthrough of the methodology followed during the experimentation stage.

Chapter 4 – Implementation: This chapter details the different experiments performed for the Question Answer Generation.

Chapter 5 – Results: This chapter discusses the results of experiments performed in Chapter 4.

Chapter 6 – Conclusion: This chapter concludes the work done in the thesis and discusses future improvements.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

The past decade has witnessed a sudden rise in the technical field i.e., generative AI. This transformative technology has demonstrated potential in tackling complex medical challenges, such as uncovering hidden patterns and achieving breakthroughs that were previously that are not possible with traditional approaches.

Generative AI is mastered in handling healthcare data which is characterized by its complex nature, fragmentation, poor labelling, and unstructured format. Electronic health records, clinical images, and data from sensor/smart devices are just a few examples of the diverse data sources. (Taiwo et al., n.d.) discusses how Generative AI can transform the construction industry by turning text data into useful outputs during different project phases. It can create summaries, ensure compliance, and automate drafting tasks before construction starts.

The latest advancements offer innovative methodologies for constructing end-to-end models that will be capable of extracting meaningful insights from complicated data. It will empower the researchers and clinicians to develop more accurate diagnostic tools and personalize treatment plans. (Taiwo et al., n.d.) states that it will automate daily reports and other documents, and after construction, it can help with inspection reports and document translations.

From drug discovery and disease prediction to personalized medicine and image analysis, generative AI has played a vital role in reshaping the future of healthcare delivery. (Taiwo et al., n.d.) mentioned that they utilized the GPT-4 model enhanced with an RAG framework to improve information retrieval from contracts. It helps to avoid generating incorrect information by grounding responses in actual contract content. The potential of Generative AI in construction is significant, promising better productivity and accuracy, but it's still in the early stages of adoption.

In this review, faced a challenge because there aren't many detailed studies that match exactly what's this research looking for in the field of Generative AI. This topic is quite new in the domain areas like education, IT, and healthcare making it a bit difficult to find the exact research papers. Even with these challenges, this research work tried to cover how Generative AI has grown, how it's built, its role in different important areas, and especially how it works in chatbots and conversation systems.

2.2 Generative AI in Health Care

(Alsharif, 2024) explains the significant advancements in artificial intelligence within the field of cancer research. its evolution from basic rule-based systems to sophisticated ML and DL techniques. These AI algorithms are crucial in analyzing extensive datasets for complex patterns, aiding in cancer diagnosis, prognosis, and treatment strategies.

Generative AI is revolutionizing healthcare by leveraging advanced algorithms to create synthetic instances of medical data such as images, sounds, and textual reports. (Alsharif, 2024) explains AI's capabilities extend to identifying new molecular targets for cancer therapy and predicting patient responses to medications, which is crucial for personalized treatment plans.

Researchers, doctors, and clinicians can study easily various types of diseases in-depth i.e., cause-effect relationships, prediction of outcomes, and effective treatment procedures. (Alsharif, 2024) also mentions that AI is utilized in drug repurposing, discovering new applications for existing drugs, thereby accelerating the drug development process and reducing costs. Author also mentions that there is a lack of data privacy and system integration, which are essential for enhancing patient care in oncology.

It not only speeds up the process of bringing new drugs to market but it also reduces the costs associated with research and development. In the field of personalized medicine, Generative AI can define treatments for the individual patient by analyzing the patient's data with the best possible outcomes. Generative AI continuously adapts and improves with more data over time. (Alsharif, 2024) publication detailed by highlighting its contribution to the field.

2.3 Generative AI Architecture

Generative AI is composed of several key architectures, each with its unique approach to data generation. They are:

- Generative Adversarial Networks
- Autoregressive Models
- Variational Autoencoders
- Diffusion Models
- Transformer Models

Together, these architectures form the backbone of Generative AI by contributing to various fields will have the ability to create synthetic data from real-world examples. Their ongoing development, training based on user query data, and refinement continue to push the boundaries of artificial creativity, also opening new paths for research and application in Generative AI.

2.3.1 Generative Adversarial Networks

In recent times, they have become a significant tool in ML, especially when it comes to creating synthetic data. These innovative models consist of two primary parts: the generator and the discriminator. The generator has the task of producing synthetic data samples that closely resemble a given dataset. On the other hand, the discriminator serves as a sort of judge, with the responsibility of telling apart the real data from the synthetic ones created by the generator. (Saxena and Cao, 2021) discusses the challenges and developments in the field of Generative Adversarial Networks, particularly focusing on image generation and feature learning.

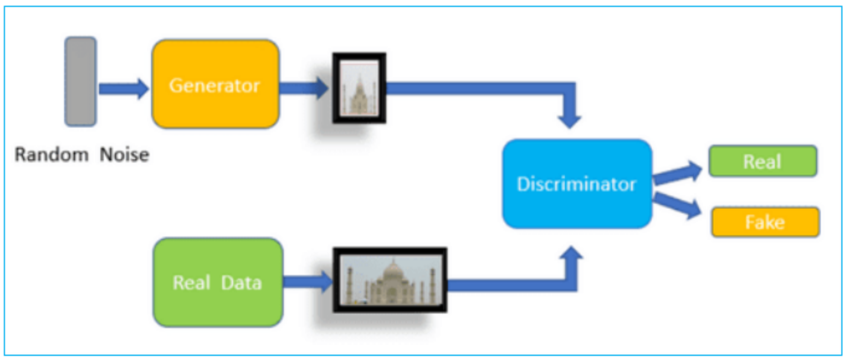


Figure 2.3.1.1 Block diagram of GAN (Generative Adversarial Network(GAN) using Keras | by Renu Khandelwal | DataDrivenInvestor, 2024)

The above figure shows that the image depicts a double feedback loop involving a discriminator that interacts with real data and a generator that interacts with the discriminator. This method involves a cyclical routine where two components, known as the generator and the discriminator, improve through continuous interaction. (Saxena and Cao, 2021) mentions that removing specific images like dogs or cats from training data does not improve the performance of single-generator GANs, as they still produce unrecognizable images for diverse datasets like ImageNet.

Within every cycle, the generator crafts artificial data starting from random noise, and concurrently, the discriminator assesses the authenticity of both the actual data and the artificial data generated. Over time, the discriminator becomes adept at distinguishing between the two. The aim is to reach a point where the synthetic data is so similar to genuine data that the discriminator cannot distinguish between the two. (Saxena and Cao, 2021) explores various architectures and modifications in GANs, such as encoder-based architectures for bidirectional mapping and modifications to handle noise vectors more effectively.

After extensive training, the model is expected to reach a state of equilibrium where the discriminator consistently struggles to make this distinction. They have proven their ability to produce good quality realistic data, leading to significant progress in fields like Generative AI, Generative Modelling, and Computer Vision. (Saxena and Cao, 2021) says that the loss functions for better comparison and estimation in generative models highlight different methods like class-probability estimation and divergence minimization.

2.3.2 Variational Autoencoders

(Hajewski and Oliveira, n.d.) this research paper introduces a new way to design variational autoencoders using evolutionary neural architecture search. VAEs are tools that help create new data by learning from existing data without needing labels, which is great because labelling can be expensive and take a lot of time.

Variational Autoencoders are a combination of both autoencoders and probability theory concepts. These models are great for learning important features of data and can create new data that looks a lot like the original set. VAEs are different from regular autoencoders because they don't just spit out one answer instead, they give you a whole range of possible answers. The architecture of VAEs is shown in the below figure.

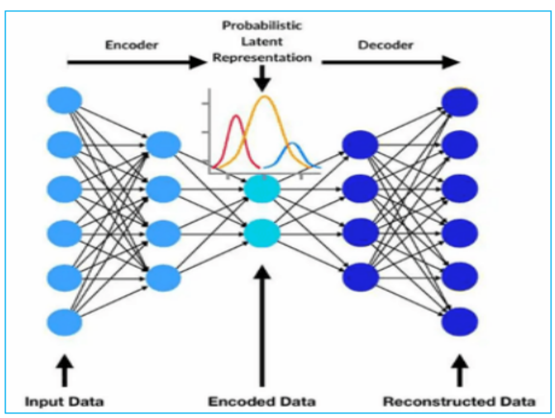


Figure 2.3.2.1 Architecture of Variational Autoencoders (Variational Autoencoders (VAEs) Made Simple & How To Tutorial, 2024)

The decoder part of the VAE works in reverse. When training VAEs, the goal is to get as close as possible to the original data. First, the model will check whether the new data matches the old data or not. Second, they use a special rule that makes sure the hidden space stays neat and orderly. Variational Autoencoders have a wide range of uses such as creating text and images, shrinking data size, and adjusting data to different contexts. (Hajewski and Oliveira, n.d.) used an evolutionary algorithm that used MNIST and Fashion-MNIST datasets. They found that training for more epochs doesn't always make the VAEs better. They noticed that the best designs from shorter training times were often better than those from longer times.

VAEs think that the data in this hidden space follows a certain pattern, usually a bell curve. To train VAEs, they use a clever trick called the "reparameterization trick." They add a bit of randomness to the average and spread of the pattern to adjust things to get better during training.(Hajewski and Oliveira, n.d.) said that they want to make a system that can build VAEs more efficiently, saving time and money.

2.3.3 Transformer Models

Transformers are really useful for many things that involve text. They have set new records in translating languages by understanding the small details of language. When it comes to answering questions, transformers can give accurate answers by really understanding the question and the information related to it. (Bouafoud et al., 2024) explores the use of chatbots in the educational sector that is in the recruitment process of academic institutions. This research paper also highlights the benefits of AI-powered chatbots i.e., recruitment experience by automating interactions and handling a large volume of queries.

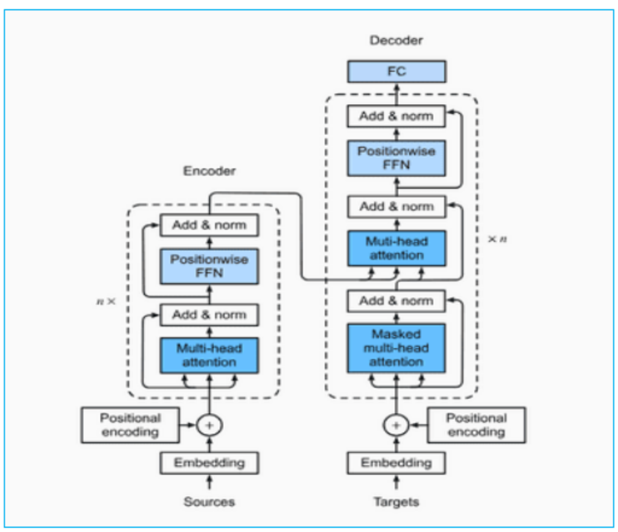


Figure 2.3.3.1 Transformer Architecture (11.7. The Transformer Architecture — Dive into Deep Learning 1.0.3 documentation, 2024)

The transformer architecture is good at handling data that comes in a sequence, like sentences. It can take a whole sequence at once, not one piece at a time. Inside a transformer, there are two parts one is an encoder and the other one is the decoder as shown in the above figure. The encoder's job is to understand the input and make a set of codes that the decoder uses to make the final result. Each layer of the encoder and decoder has two more functionalities one is a multi-head self-attention mechanism and the other one is a feed-forward network.

Transformers are great at finding and naming specific things in text, like people or places, and sorting text into different groups based on what it's about. They are good at predicting what word comes next in a sentence or making up new sentences that make sense and fit the context. (Bouafoud et al., 2024) specifies the transformer models which are advanced AI techniques used for NLP tasks such as semantic analysis and named entity recognition, and their applications in personalized teaching, learning, and automated assessment.

Multi-head self-attention makes the model to pay attention to various parts of the sequence for different reasons, on the other hand, the feed-forward network changes the output of the attention mechanism. This model uses positional encodings. These are added to the input so the model knows the order of the sequence, which is important because the transformer doesn't naturally understand order. (Bouafoud et al., 2024) categorized chatbot evaluation methods into rule-based, AI-based, retrieval-based, learning-based, and generative-based approaches, discussing their respective advantages and limitations

2.3.4 Auto Regressive Models

Auto-Regressive (AR) models are a type of statistical model used in generative AI to predict future data points in a series by looking at past data. These models are built to understand the order of data, like how one thing follows another over time. (Qiu, n.d.) Hong Qiu's research paper presents an efficient image interpolation algorithm based on the autoregressive (AR) model to enhance the quality of enlarged images.

These AR models are set up to handle information one piece at a time. For text, this means looking at one word at a time, and for images, it could mean creating an image bit by bit. The model learns the chances of what comes next based on what it has seen before. (Qiu, n.d.) mentions that traditional methods like Nearest Neighbour and Bilinear interpolation often result in images with poor quality, and lacking detail. Qiu's algorithm constructs an AR model that considers the pixel structures and uses iterative refinement with multiple AR models and weighted summarization to improve interpolation.

AR models are used in language tasks, translating, summarizing, and making new content. They can make realistic images, and music, or predict future data for things like weather forecasting. (Qiu, n.d.) says that the initial interpolated values are obtained using the least square method, and the process iterates until optimal quality is achieved. The algorithm's effectiveness is measured using the Peak Signal-to-Noise Ratio (PSNR), showing superior results over traditional methods.

Pixel RNN and Pixel CNN are AR models that make images by predicting each pixel's value based on the pixels around it. These models in generative AI work by processing information in order and are very good at figuring out complex patterns of what might come next.

2.3.5 Diffusion Models

Diffusion models are from non-equilibrium thermodynamics. By adding data with noise step by step until it's just random noise. This model is a neural network that sets the rules for this reverse diffusion. In each step, the network guesses the noise that was added before and uses this guess to clean the data bit by bit. This is done over and over for certain iterations until the data is clean enough. (Helenaˇ et al., n.d.) authors explore the use of diffusion models for text generation in the field of NLP. They compare discrete diffusion models work on the token level, to continuous diffusion models that embed text into a continuous space.

Diffusion models are used in making images that are so good and varied that they can compete with images made by GANs. They will make the details in images clearer or fill in parts that are missing. They can write new text or finish text that makes sense and fits the context. (Helenaˇ et al., n.d.) mentions that discrete models lag behind autoregressive (AR) models in terms of producing coherent text due to challenges in modelling semantic correlations.

They can also translate text from one language to another without losing the original meaning. For speech, these models can make clear and natural-sounding voices from noisy sounds. Overall, diffusion models are a strong tool in the world of creative modelling. (Helenaˇ et al., n.d.) mentions that first, they review work and then provide a mathematical definition of diffusion models by discussing discrete and continuous approaches. The author also says that currently do not outperform state-of-the-art AR models and face issues such as defining noise schedules and embedding spaces.

2.4 Generative AI in Multiple Domains

Generative AI works by learning from existing data, understanding the underlying patterns, and then using that knowledge to create new, original content. This technology is revolutionizing creativity and innovation in numerous fields, making it possible to explore ideas and solutions that were previously unimaginable.

2.4.1 Field of Education

(Chen and Wu, 2024) has explored the creation of multimedia teaching materials for Tang poetry using generative AI. The author mainly focused on improving text-to-image generation quality and understanding. He used the evaluation methods like Stable Diffusion and DALL E2 for content assessment. It aims to reduce cognitive load and enhance learning through situational videos. He addressed the need for scientific content evaluation and dynamic content simulation. The developed system enhances teaching by facilitating interactions and evaluating learning effectiveness. The author has used semantic similarity to ensure accurate content retrieval in this domain of educational videos. This research identifies gaps such as the need for scientific content evaluation and procedural dynamic content simulation. Future research will look into combining generative AI with 3-D games by automated evaluation.

(Dong et al., 2024) mainly discusses how they reviewed articles, and criteria and analyzed them by using big language models in education. The author discusses the issues by using models in education, application in medical education, impact on English learning, and research ethics. It also mentions critical thinking, content accuracy, exam performance, and ethical concerns. This research explains how models like ChatGPT are used in education to improve teaching, explore research questions, gather user feedback, and assess their effectiveness by understanding public perceptions. They analyzed 94 documents, and found ChatGPT's use in language and medical education in 2023, highlighting both potential benefits and concerns. It suggests future research could focus on making these models more relevant for medical education. By comparing different models LLMs can enhance collaboration between teachers and models. Additionally, it addresses ethical issues by applying models in more fields and improving feedback mechanisms.

(Lundström et al., 2024) mainly focuses on improving online courses through the use of GPT-4, specifically targeting the domain of IoT courses at Linnaeus University. The research identified several issues in online education such as technical challenges, the need for better course material, and effective teacher communication. They collected over 12,500 messages from Discord which involves both student-instructor interactions and teacher discussions. So that they can analyze sentiments and generate dynamic FAQs. The messages went through data preprocessing by using Python scripts, then processed using various algorithms such as NLTK Vader for traditional sentiment analysis and GPT-4. So that LLM can address these problems by providing detailed sentiment analysis, generating relevant FAQs, and improving overall course content with support. This research evaluated that GPT-4 offered more accurate and context-sensitive insights compared to traditional methods. Future research directions will include refining sentiment analysis models, scalability, and integrating AI into various educational sectors to improve both student and teacher experiences.

(De La Torre and Baldeon-Calisto, 2024) the author explains the use of Artificial Intelligence especially in the stream of Generative AI i.e., higher education across Latin America, covering the years 2021 to 2023. Examining how AI and GenAI are being used in this region's educational sector by providing fresh insights into their impact. This research paper explains that there is a lack of comprehensive literature reviews in their application i.e., Latin American higher education. It aims to understand the challenges Latin American educational institutions face with AI implementation. Gathered opinions from teachers and students about using GenAI in their programs. They reviewed 25 papers from various Latin American countries by analyzing AI applications, perceptions of AI and GenAI, also the challenges faced. It highlights the use of AI for predictive modelling, virtual assistance, and generative AI for creating new content to improve the learning experience in an interactive way.

(Trajanoska et al., 2023) this research paper discusses improving the construction of Knowledge Graphs using Large Language Models in the stream of NLP. They are extracting information from unstructured text. Utilized various datasets including news articles, ESG, and CSR reports from Crédit Agricole and Instruct GPT dataset to train models for entity and relation extraction. It will create summaries similar to the tree model. It also aims to create more accurate and comprehensive Knowledge Graphs. This research study addresses several challenges such as error propagation in multi-step pipelines, end-to-end relation extraction, and the comparison of different models' performance by using Knowledge Graphs. They have used preprocessing techniques like tokenization and batch processing to prepare data for the extraction process. This study compares the REBEL model and ChatGPT in making Knowledge Graphs, focusing on sustainability. It suggests future improvements like a framework for assessing graph quality by applying these methods in various areas such as using graphs to find missing links, automating updates, and keeping graphs current.

2.4.2 Field of Health Care

(Bajpayi et al., 2024) explains the security challenges and advancements in IoT healthcare devices. The author mainly focuses on anomaly and intrusion detection techniques with the role of AI in enhancing security measures. They used the IoMT Dataset for anomaly detection using various techniques such as CBLOF and KNN in 2021. The Personal Medical Devices Dataset for intrusion detection with methods like DT and SVM in 2020. It also mentions the use of the National Vulnerability Database for tracking software vulnerabilities in health systems from 2001 to 2022. They mentioned several security and privacy management challenges, such as the risks posed by unencrypted IoT devices and vulnerabilities, and weak credential management in IoT architecture. The document emphasizes the potential of AI-driven approaches for managing large data volumes. Therefore, by enhancing risk management alongside the critical security flaws in smart medical devices. The paper outlines various preprocessing techniques such as PCA and CNNs for anomaly detection. This paper advises AI-driven vulnerability management and stronger authentication methods to boost IoT healthcare security and privacy.

(Sakthivel et al., 2024) explains about use of Generative Artificial Intelligence and Deep Neural Networks in improving cancer diagnosis and treatment through medical imaging. Medical image datasets including mammograms, CT scans, MRIs, and histopathology slides are used to train generative models like GANs. These models help in generating synthetic images to overcome the limitations of small and imbalanced datasets and enhance the performance of machine learning models in medical image analysis. This paper highlights several issues in cancer care, such as limited access to healthcare, and disparities in cancer care. There is ongoing research to develop new treatments and improve early detection methods.

This research shows significant contributions of AI and ML in cancer research including detection, diagnosis, and patient care. Preprocessing techniques used are data collection, image enhancement, and data denoising to improve medical image quality. Evaluation metrics are used such as accuracy and reliability in these technologies. It suggests future improvements in data handling and application of gen AI for better cancer care.

(Taylor et al., 2024) explains about IntelSurv application which is designed to aid Community Health Workers in Malawi through Generative AI. It highlights the creation of a Frequently Asked Questions database using GPT-4, which includes 1,300 questions and answers, with a gold standard set of 220 questions developed by public health experts. The application deals with disease surveillance application which involves a self-reflection mechanism where unanswered questions are reviewed by experts and provided feedback system for users to increase accuracy. Evaluation metrics focus on the quality of answers and questions, self-reflection, feedback mechanisms, and iterative dataset construction.

The research paper identifies gaps such as the need for explicit knowledge bases, addressing inaccuracies in data collection, improving response times, adding local language translations, and addressing biases in Large Language Models (LLMs). Used preprocessing techniques like document splitting, embedding, vectorization, and context injection which are crucial for retrieving and generating accurate answers. The application uses algorithms like the text-embedding-ada-002 model, Pinecone vector database, and Retrieval Augmentation Generation (RAG) to ensure the effectiveness of the tool.

Future recommendations mentioned by the author are to improve prompting and embedding techniques. Incorporating open-source models by addressing language considerations. Collecting more data for testing, and enhancing the feedback mechanism for continuous improvement. (Chheang et al., 2024) explains generative AI-based virtual assistants in immersive virtual reality (VR) environments by focusing on anatomy education. The questionnaires are designed to evaluate the task load, and sense of presence within the VR setting. This research aims to enhance VR environments and virtual assistants to improve user experience by making users more immersive. The research shows application in medical training or therapy for children with autism spectrum disorder. Evaluation metrics mentioned include both subjective measures like SUS, NASA TLX, and IPQ are used in task completion time and number of interactions. The study found that participants interacted more with the virtual assistant in avatar form. The author suggests further research is needed to explore user preferences and the impact of cognitive load and interaction dynamics. Future research focuses on user preferences, cognitive complexity on user performance, and the status of user interactions with virtual assistants.

(Bhate et al., 2023a) uses the GPT-3.5 model to pull out social determinants and family history from clinical notes information without needing a lot of specific instructions. They took 1,000 clinical notes from 150 patients with various diseases from a university hospital. These notes were made anonymous by removing personal details like names and addresses. The author has labelled the information about social determinants and family history.

This research paper points out some issues with using GPT models, like their outputs not always being consistent or not keeping the original meaning of the text. It suggests that more work like using reinforcement learning or prompt tuning gives output better. The main goal of this research is to see if GPT-3.5 can be used for finding information from clinical notes which could make researching clinical information better. The author explains that GPT model accuracy was by finding the right words i.e., Named Entity Recognition, and understanding the meaning of what was found.

2.4.3 Field of Marketing

(Yin, 2024) explains how people feel about ads created by AI seeing their answers to a survey. It uses these answers to find patterns that can help predict people's reactions to these advertisements. The survey asked about things like how entertaining or pretty the ad was. The researchers used a couple of smart methods to analyze the data and found that people have mixed feelings about AI-made ads and there is an increase in doing AI ads. Wanted to make sure that what they found was reliable and could be used by others, especially by companies that make ads and want to know what people think of them.

This research study gives us a clearer picture of how people see AI-made ads and what makes an ad work well. It's all about making ads that people can relate to and trust, which isn't always easy when a computer is doing the creation. It's showing a smart way in marketing.

(Lee et al., 2024) proposes creating personalized marketing services using generative AI. They gathered data through surveys with 300 adults from South Korea, spanning ages 20 to 60, ensuring a diverse sample by dividing them into groups based on gender, generation, and treatment. They used BBQ Chicken's Golden Olive Chicken product data for the survey, which was conducted online to assess message effectiveness.

The research identified several gaps, such as the low explanatory power of message elements like event dates, the impact of disclosing purchase history, and limited effectiveness in AI-generated content. It also highlighted the need for comprehensive marketing strategies. It raised concerns about ethical implications and consumer rights with AI use in marketing.

The Persuasive Message Intelligence (PMI) system, utilizing OpenAI's GPT-4 model, aims to generate persuasive marketing messages. The author mentioned the power of advertising and AI-generated messages. He explained that the process involves input as product information, retrieving user data, generating messages, and dispatching them to customers.

Hence followed by survey validation and regression analysis to test hypotheses on message effectiveness. This research used regression analysis with SPSS25 and OpenAI's GPT-4 for message generation. Also evaluated through metrics like modified R-squared value, F-statistic, and tests for equality of variances and normality distribution.

(Udeshika Munasinghe et al., 2024) introduced a new AI model, called BiGAN, to estimate the Value at Risk for Central Counterparties. The researchers used a dataset from the S&P 500 Index, covering 15 years and including 3777 daily adjusted closing prices. This dataset was split into three parts that is first nine years for training, the next year for validation, and the last four years for testing. The study's goal was to improve how CCPs estimate VaR, especially during turbulent market times like the 2008 financial crisis and the 2020 pandemic.

The BiGAN model consists of a generator, encoder, and discriminator. Was trained, validated, and tested with this data. It was compared to traditional VaR models, showing promising results in terms of accuracy and stability. The research addressed several issues with current VaR estimation methods, such as their procyclicality and the limitations of traditional models under extreme market conditions.

By proposing the BiGAN model, the study aimed to provide CCPs with a tool that could offer more accurate and stable risk management. Especially important during periods of high market volatility. The model's performance was evaluated using various metrics such as the number of breaches, the Kupiec Test, and measures of procyclicality. The results suggested that the BiGAN model could potentially offer improvements over traditional models. The study explains the importance of developing more effective risk management tools for financial market stability using BiGAN models.

(Himanshu Tiwari and Ayush Raj, 2024) explored the use of Generative Adversarial Networks for creating NFT art, focusing on images from a Kaggle dataset. This dataset includes image files of 3021 pieces, all resized to 512x512 pixels for GAN compatibility. The author addresses several challenges in generative AI, blockchain, and NFTs, such as the limitations of traditional image synthesis methods, and there is need for NLP adoption to enhance the quality and control of image generation. It highlighted the revolutionary impact of GANs introduced in 2014 and the potential of transformer architecture in integrating generative models.

By conducting comprehensive case studies to assess GANs effectiveness in producing NFT art and also generating revenue through tokenization. The preprocessing techniques used in this paper are image downsizing. Evaluation metrics like the Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) were used to assess the quality of generated artworks. Recommendations for future research are experimentation with different GAN structures including various file formats, and conducting comprehensive case studies.

(Pooja et al., 2024) explains about significance of AI and digital transformation in the business world focusing on small and medium-sized enterprises and SMEs. The author explains about integration of AI technology into business networks to automate and increase operational systems efficiency by generating good earnings. He mentions about incorporation of digital technologies into business processes. This research proposes a framework for these enterprises to adapt to digital changes.

It highlights the crucial role of analytics which is descriptive, diagnostic, predictive, and prescriptive in business operations. The pivotal role of AI and machine learning is to increase accuracy and facilitate data-driven decisions. There is a gap in the application of analytics and digital transformation strategies in SMEs. Suggests digital transformation should aim at value generation, process improvements, and business model transformation.

2.4.4 Field of Information Technology

(Naimi et al., 2024) explains smart tourism application which discusses creating test cases automatically from use case diagrams with the help of generative AI and prompt engineering. This research highlights the process and benefits of using AI to make the test case creation process faster and cover more scenarios than manual methods.

The main challenges in test case generation are domain-specific limitations, dependency on specifications, and lack of regression testing methods. It proposes ontology-based approaches and model-based test case generation to improve the process.

This research paper uses structured use case diagrams and generative AI to streamline test case creation, enhancing test coverage and faster development cycles. It suggests future work to expand the types of diagrams used and explore cost-effective solutions for API access and continuously improving AI model performance. These suggestions aim to address research gaps and enhance the effectiveness and efficiency of generating test cases automatically.

(Adapa et al., 2024) collected datasets from various pull request reviews and focused on comments related to format checks. This helps to develop an AI-driven system aiming to improve efficiency by identifying common format errors in code snippets.

The ultimate goal is to train an AI bot to recognize and suggest improvements for these errors, thereby speeding up the pull request review process. The author will use this dataset to train AI models that can automate pull request review which enhances the power of AI in software engineering.

It enhances code review efficiency, continuous integration, early prediction models for code changes, and the automation of code review activities. It also discusses the future work of improving the AI bot's intelligence by integrating sentiment analysis for a better developer experience during reviews. The author uses the Falcon-40B model for error detection which is trained on extensive datasets to identify formatting errors in code.

They mentioned about model's training scale, data sources, and techniques but lacked specific details on the preprocessing steps. It also references the use of ROUGE scores and accuracy as evaluation metrics. This research aims to improve software development workflows and contribute to the advancement of AI-driven software engineering.

(Dhyani et al., 2024) explains various datasets by using AI techniques to improve the creation of API documentation. This research paper identifies problems with current API documentation methods, such as manual work & lack of precision therefore author suggests using generative AI to make the process faster, accurate, and scalable.

It mentions five specific datasets that is Niche Domain Dataset for news summary generation, the Hand-Curated Dataset for adapting GPT-3 to societal metrics, Husain for Code BERT, the Penn Treebank Dataset for assessing the tran GAN model, and the dataset for API documentation to fine-tune the Tiny Pixel or Llama-2-7B-bf16-sharded model.

These datasets help train models to perform tasks like generating accurate summaries, adapting LLMs, and creating comprehensive API documentation. Generates up-to-date documentation efficiently by evaluating the economic and practical benefits of AI in software development. The study concludes with significant advancements in API documentation generation, suggesting areas for future research like further fine-tuning, expanding dataset diversity, and integrating with existing tools.

(Wang et al., 2023b) explores how to answer questions using information from different documents. They used datasets like HotpotQA and WikiMHop, and a special 'PDFTriage' dataset with real-world questions about document structures. This research paper mainly focuses on document structures like tables and figures because some questions can be asked of them.

They mentioned that LLMs can find answers using different thought processes and can be slow because usage lot of resources. The author used techniques like TF-IDF to find important words and to measure how similar sentences are. TAGME to find Wikipedia links in the text. They also used an API to get information about pages and tables. They have split documents into chunks to make a knowledge graph.

Methods used were the Knowledge Graph Prompting method and a graph traversal algorithm to find and use information from different documents. Evaluation metrics used are Precision, F1 score, and Exact Match. The challenges mentioned by the author are making a knowledge graph keeping track of where information came from and dealing with complex tables. The future scope is to check how well LLMs can understand graphs and design better ways to use them so that this research can get the most relevant information when answering questions.

(Varalakshmi and Bugatha, 2024a) states that AI-powered systems are designed to tailor resumes and prepare individuals for job interviews. It uses real-world resumes and job descriptions, collected through web scraping, and employs the GPT-2 model for text generation.

The system faces a challenge with limited content for question generation, which affects the number of questions that can be evaluated. The quality of the questions is measured using metrics like cosine similarity, BLEU, METEOR, and ROUGE. These metrics help in assessing how similar the generated questions are to reference questions.

The system aims to help job seekers to find roles that match their skills and aspirations. By focusing on semantic accuracy and meaningful question generation it boosts the 97% accuracy rate. The research analyzes the performance of various models, including BERT, T5, ROBERTA, and XLNET, which are part of Natural Language Processing and Large Language Model frameworks.

The preprocessing of text involves several steps, such as replacing special characters, tokenization using NLTK, counting the original number of words, converting words to lowercase, removing non-alphabetic words and stop words, and lemmatization. These steps are crucial for maintaining uniformity and enhancing the text's relevance.

It also uses BERT for semantic analysis and the T5 model for converting sentences into questions. Also uses Stanford Core NLP for syntactic analysis, and a Phrase Matcher for resume scoring. Deep learning models like BERT can automate the generation of question-answer pairs. The Word Embedding Model is used for performance analysis.

Metrics like cosine similarity, BLEU, METEOR, and ROUGE are employed to measure the semantic similarity between generated and reference questions, providing a quantitative assessment of the quality of the questions generated. These metrics are essential for evaluating the effectiveness of the question-generation algorithms.

The author mentions that they have used Named Entity Recognition (NER) for classifying entities in the text by focusing on organizations to extract relevant information. It also outlines the limitations of current question-generation systems and proposes a new approach to improve semantic accuracy and generate meaningful questions with complete answers.

2.5 Enhancing Retrieval for Fact-Based Question Answering

This section discusses about the retrieval enhancement for Document Question Answering systems.

2.5.1 Retrieval Augmented Generation for unstructured data

This section explores the usage of RAG for retrieving information using unstructured data as input from PDF documents. It initially discusses the challenges and mentions how this research work can overcome those challenges using RAG.

2.5.1.1 Challenges of Retrieving Information

(Stankovski et al., 2024) discusses the challenges, particularly in teaching Programmable Logic Controllers ensuring that AI tools accurately assess learning rather than just detecting cheating. This paper stresses the importance of integrating AI into education to help students learn and prepare. However, AI tools are not always completely accurate, so their use must be approached with caution.

The study also looks at how students feel about using AI tools and their impact on traditional learning methods. It suggests that educational institutions should teach students how to use AI effectively and that more research is needed to improve AI tools.

Large Language Models have the potential to process vast amounts of unstructured data, which includes things like text from social media or images. They are powerful because they can understand and find patterns in this kind of data, which is not neatly organized. LLMs can help us get valuable insights from unstructured data by figuring out what customers think or identifying trends. However, there are challenges too.

In this research paper author (Singh and Hooda, 2023) mentions that LLMs might not fully understand different kinds of unstructured data which will make it hard for them to learn from it. Also, if the data is not prepared properly before using it then LLM might not work as expected. Another issue is that as the amount of data increases, it can be tough for LLMs to handle it without the right computing power.

Additionally, need to be more careful about privacy and make sure the LLMs don't create biased results. By focusing on these areas, this research can make LLMs better at handling unstructured data by using them responsibly.

2.5.1.2 Retrieval Augmented Generation for Overcoming Challenges

GraphQA framework includes the need to standardize diverse datasets into a uniform format by ensuring the retrieval process is accurate. (He et al., 2024) mentions the role of Retrieval-Augmented Generation (RAG) is to enhance the question-answering framework's ability to understand and interact with complex graph data. RAG uses retrieval-based approaches to source relevant information from a graph. It will be helpful to generate accurate and contextually appropriate responses. Therefore, this technique improves the natural language processing capabilities of the system.

The author also identifies relevant graph nodes, and edges and integrates complex language models with GNN to handle real-world graphs with more accuracy. Additionally, the system will manage creative and complex questioning where it requires advanced conversational interfaces that are both human-friendly and capable of multi-hop reasoning.

Using RAG for multi-hop queries presents both challenges and opportunities. (Tang and Yang, 2024b) the challenge is that the dataset's answers are limited to simple responses, which restricts complex answers. Current RAG systems struggle to retrieve and answer multi-hop queries effectively which needs advancements in these systems. Developing RAG systems with better reasoning and answering capabilities will be an opportunity to enhance RAG answer stability.

2.5.2 Roles of structured and unstructured data

This section explains the roles of retrieving information from structured and unstructured data of different domains. It also discusses how Pandas AI, FAISS, and RAG can be used to enhance the accuracy and efficiency of QA systems.

2.5.2.1 Usage of Pandas AI for Structured Data

(Mckinney, n.d.) leveraged tools like Pandas AI to analyze datasets effectively. Pandas can take care of managing structured data such as Excel and CSV files. It helps to find the patterns and statistical insights from the data. By using data frames, users can filter, sort, and manipulate data to extract relevant information quickly. This capability is crucial for answering questions that require data insights.

If the user asks a question about trends or patterns within a dataset, Pandas can compute descriptive statistics or pivot tables to provide answers. It can also merge related datasets to enrich the context for more complex questions answering. Pandas have the ability to handle time series data which allows for temporal queries to be addressed with precision which will be more accurate for question answering.

(Gao et al., 2018a) talks about how to make AI applications, especially for health care by getting and combining data to understand real-world applications.

It includes getting data ready, preparing data, designing models, and building solutions accordingly. This research can visualize and interact with data on how to incorporate AI solutions that work more efficiently.

2.5.2.2 Usage of RAG and FAISS for unstructured data

(George and Rajan, 2022a) explains a system that generates short stories from keywords in a coherent and logical manner, and can be used in gaming, filmmaking, and education. The author explained about one of the algorithms called Facebook AI Similarity Search (FAISS) which helped in this application and sentence transformers for efficient story generation. The FAISS-based model outperforms the baseline in subjective evaluations.

(Douze et al., 2024) explains that FAISS plays a crucial role in managing and processing large datasets, both in text and image forms. It is used in bilingual texts from vast text datasets on the web. The author also states that FIASS organizes language model training to cluster documents on similar topics. In images, FAISS helps to eliminate duplicates from a dataset containing 1.3 billion images. It shows how efficiently FAISS works with vast, versatile text and images.

(Suresh et al., 2024) worked on the Electron-Ion Collider project by making it easier for people to find and understand complex information. RAG agents will look into big databases to find the right information. After that will use an LLM to make a short summary that answers the user's questions.

The author explains how the RAG system takes in information as input which includes breaking down data, turning it into vectors, and storing these vectors. Based on these stored vectors and user query vectors RAG will try to find out the distance between them and give top k probabilities as answers. In the overall context, RAG Agent is helpful for researchers' work which handles huge amounts of data more easily.

(Edge et al., 2024) explains RAG systems which help users to get insights into huge textual data by summarizing and finding useful patterns. The author divided data into two types i.e., podcast transcripts and news articles. This research paper suggests that RAG systems should help people understand whole text collections, not to find single facts. Mentioned the pros and cons of making a graph index for RAG systems and how to test them.

2.5.3 Role of LLMs in Health Care Sector

Large Language Models in healthcare help in advanced diagnosis and suggesting medicine through data analysis. Also improving patient interactions with AI-driven chatbots that provide required information. They are helpful in medical research by augmenting data, leading to better outcomes. Very useful in interpreting data and applying healthcare laws and guidelines.

2.5.3.1 Generative AI and Data Augmentation in Healthcare

(Sai et al., 2024) says that they can predict protein functions which helps to discover new drugs. Gen AI will use big datasets from sources like SwissProt and PubMed to train models that can generate medical information.

Some AI models mentioned are BioBERT for understanding medical texts and NVIDIA Clara for healthcare tasks. Future recommendations mentioned in the paper are it will interact with patients and doctors by keeping patient data private with support of different languages by overcoming the challenges.

(Osuala et al., 2021a) discusses the use of Generative Adversarial Networks to handle challenges in cancer imaging such as data scarcity and privacy. GANs can create synthetic images which helps in diagnosis and treatment planning.

This research paper introduces the SynTRUST framework to evaluate the trustworthiness of the medical image analysis field. It mentions the need for better synthetic images that are useful for clinical practice. It highlights the importance of metrics like SSIM and FID to assess image quality.

(Uprety et al., 2023) explains that interpreting complex genetic reports and suggests treatments. GPT could summarize medical records and clinical trials for patients. GPT also needs to be compliant with health privacy laws like HIPAA which is a challenge in the field of clinics. GPT might have biases or wrong information because of the data it learned from. It can be tricked into giving bad advice and lack of latest data.

The author mentions that GPT could be very helpful in cancer care but it takes more work to fix small issues like biasing, and wrong information by prompt tuning and maintaining up-to-date data.

2.5.3.2 AI-Enhanced Diagnostic and Imaging Technologies

(Olivato et al., n.d.) mentions how language models like BERT and GPT-4 can help in classifying radiology reports. They used different datasets, including one with 2,248 reports for lower-level categories like Lesion Nature and another with 10,000 reports to train a word model.

The author also talks about the challenges of not having enough data and how the models they made didn't work as well as expected. They made a system that gives radiologists instant feedback, which they can change if needed. This study helps doctors use information from radiology reports by using AI and ML.

2.5.3.3 AI Chatbots and Patient Interaction

(Montagna et al., 2024a) explores the use of Large Language Models (LLMs) in healthcare chatbots focusing on hypertensive patient care. The chatbot deals with patient education, medication, and lifestyle, and gives real-time support. Challenges are the risk of sensitive patient data being shared with third-party LLM systems a security concern. High cost involved for using advanced LLMs like GPT-3.5. It gives a detailed analysis of the advantages and limitations of using LLMs in healthcare chatbots by focusing on patient empowerment and chronic disease management.

The author used ML.NET 2.0 and GPT-3.5 for tasks like classifying text and sentiment analysis. Experimented with open-source LLMs through a service called Ollama. Also evaluated the chatbot's performance using metrics like precision, recall, and accuracy. Used BERT score to ensure it could understand and respond to patient queries effectively with more relevancy.

The research paper suggests the chatbot to fine-tune the LLMs for specific tasks by using "chain of thoughts" prompting technique for better response accuracy. The author suggested experimenting using larger models like Falcon 180b for future research.

2.5.3.4 Regulatory Compliance and AI in Healthcare

(Kim and Min, 2024) develops a QA-RAG model to enhance efficiency in the pharmaceutical regulatory process. It assists professionals by providing accurate answers to questions about FDA and ICH guidelines by combining large language models with retrieval-augmented generation.

This model is trained on public documents and uses a special OCR tool for extracting scientific texts. It breaks down documents into chunks to ensure no loss of information. It also employs advanced techniques for document retrieval and answer generation.

The QA-RAG model is evaluated using metrics like Context Precision and Bert score to ensure it provides reliable information. The author says that the model is designed to support not replace human expertise in the pharmaceutical industry.

2.6 Summary of Findings

This section highlights the main outcomes and insights gained from the studies across various domains.

- In the field of education, generative AI has been used to create multimedia teaching materials, improve online courses, and assist in the construction of knowledge graphs. Tools like Stable Diffusion, DALL-E 2, and GPT-4 are used to enhance learning experiences for teacher-student interactions.
- In healthcare, AI plays a key role in IoT device security, also improving cancer diagnosis and treatment. Techniques like anomaly detection, generative models for

medical imaging, and AI-based virtual assistants in VR environments have shown their importance.

- Marketing research has explored AI-generated ads, personalized marketing services, and the use of GANs. AI is a powerful tool for predicting consumer reactions and generating content.
- In information technology, AI is used to automate test case creation, improve API documentation, and resumes, and enhance software engineering processes.
- The role of retrieval-augmented generation (RAG) has been emphasized for fact-based question answering particularly in handling unstructured data.
- Tools like Pandas AI, FAISS, and RAG have been used to improve the accuracy and efficiency of QA systems both in structured and unstructured data.
- Large Language Models (LLMs) have great potential healthcare sector including data augmentation, diagnostic and imaging technologies, patient interaction using chatbots, and regulatory compliance.
- Some of the important research papers synopsis are mentioned in the below table:

Table 2.6.1 Summarized Analysis of Research Papers

Author & Published Year	References	Main Points	Research Gaps
Stevan Stankovski (2024)	(Stankovski et al., 2024)	Explains about the usage of Generative AI by Enhancing learning and exam performance in Programmable Logic Controllers	Lack of comprehensive understanding in educational sector by using Generative AI
Sahib Singh (2023)	(Singh and Hooda, 2023)	Discusses about challenges of unstructured data and strategies that include preprocessing, feature extraction, ensemble learning, explainable AI and guardrails.	There is a lack of exploration in ethical issues and bias which needs more exploration.
Xiaoxin He (2024)	(He et al., 2024)	Explains about G-Retriever framework developed by	Need focus for larger graphs without losing

		combining GNNs, LLMs, and RAG to enhance efficiency in QA & also addresses hallucinations in Graph LLMs.	information while retrieving by enhancing the subgraph retrieval so that it also improves scalability
Yixuan Tang (2024)	(Tang and Yang, 2024b)	Mentions that author developed MultiHop-RAG model using RAG so that it can give QA with supported evidences from wide range of articles.	Answers are restricted to responses like "yes," "no," entity names, or with temporal indicators. It can include more complex answer formats with the help of RAG.
Wes McKinney (2011)	(Mckinney, n.d.)	Explains about handling structured datasets used in fields like statistics, finance, and social sciences. Enhances python capabilities for statistical computing & data manipulation, analysis.	Pandas' library faces difficulty in handling large datasets which exceeds memory capacity. This paper mentioned if pandas can develop multi-processing & parallel computing then it would be more efficient.
Jinyang Gao	(Gao et al., 2018a)	Explains the importance of data preparation in AI applications. Also highlights the use of visualization tools and interaction with domain experts which improves the efficiency of data preparation.	Identifies deployment gap of AI models, this research focuses on model training rather than deployment. It mentions fault tolerance and load balancing important for implementation which are not addressed.

Godwin George (2022)	(George and Rajan, 2022a)	Generating stories based on input keywords using a FAISS-based search approach. It encodes pre-processed data with a sentence transformer. The stories are converted to audio which is mainly used in game development and filmmaking.	Existing models struggles to seamlessly integrate individual phrases into coherent plots within a short time frame. It indicates the need for further improvement in the efficiency & coherence.
Matthijs Douze (2024)	(Douze et al., 2024)	This paper focuses on transformations & optimizations that can be applied to data vectors which improves search performance. Preprocessing transformations are used to make different measures equivalent which enhances the search process under parameters like speed, memory usage, and accuracy.	This paper struggles to handle anisotropic distribution of transformed vector which is making indexing more challenging. Further research required on indexing & searching highly anisotropic data without compromising on speed or accuracy.
Karthik Suresh (2024)	(Suresh et al., 2024)	This paper mentions about development and application of a Retrieval-Augmented Generation designed for Electron Ion Collider community. It also assists in managing & retrieving vast amounts of documentation and data efficiently.	There is a gap in the current implementation of the RAG system in optimizing the vector retrieval process. It mentions the system uses cosine similarity for assessing the similarity between data segments, the effectiveness of this metric decreases as the size of the data segments increases. This suggests a

			need for further research for the retrieval algorithms to maintain high performance irrespective of segment size.
Darren Edge (2024)	(Edge et al., 2024)	This research paper explains about retrieving and generating answers using a Graph Retrieval-Augmented Generation. It utilizes different levels of graph communities to provide answers to user queries. It allows for tailored responses that can range from broad overviews to detailed insights, depending on the user's needs.	There is a research gap in the evaluation of the system's effectiveness across diverse datasets. It focuses on Podcast and News datasets which haven't extensively address how the system performs across varied types of data or in real-world scenarios where data variability and complexity can significantly affect performance.
Siva Sai (2024)	(Sai et al., 2024)	This research paper highlights the application of machine learning techniques, deep neural networks, and natural language processing to generate outputs that mimic human-generated content across various forms such as text, audio, and video.	This paper identifies a significant gap in the form of data quality and bias within generative AI systems, which can lead to inaccurate results and perpetuate healthcare disparities if not addressed.
Richard Osuala (2022)	(Osuala et al., 2021a)	This research paper focuses on evaluating the trustworthiness and validity of cancer imaging	There is a gap identified in this paper is the limited reproducibility of the

		solutions using Generative Adversarial Networks through a structured framework called SynTRUST. It helps in reproducibility, usefulness, scalability, and tenability of the studies in the field of medical image synthesis.	studies due to the unavailability of datasets and the software used in the experiments. This restricts the ability to replicate and verify the results, which is crucial.
Dipesh Uprety MD (2023)	(Uprety et al., 2023)	The document highlights the potential of ChatGPT and similar AI technologies to significantly improve healthcare, particularly in oncology. It describes how these technologies can efficiently manage and interpret vast amounts of medical data to aid in cancer care.	The document identifies a critical limitation in the current version of ChatGPT, which is its non-compliance with the HIPAA act. This poses a risk of leaking sensitive patient information for direct application in clinical settings where patient confidentiality is important.
Matteo Olivato (2023)	(Olivato et al., n.d.)	This research paper highlights the development and evaluation of a novel deep learning-based system by using BERT, for classifying Italian radiology reports. It includes domain knowledge in a hierarchical manner and has demonstrated greater performance compared to previous models.	This research paper identified a gap i.e., impact of data scarcity on prediction accuracy. It also mentions that the limited number of non-negative reports available for training affects the performance of the classification system. It suggests a need for further research into methods or models that

			can perform well even with smaller datasets.
Sara Montagna	(Montagna et al., 2024a)	This research paper mentions that evaluation of two different strategies, the first strategy involves using a third-party LLM through an API, which is quick but poses risks of sensitive data leakage. The second strategy uses a local, open-source LLM, enhancing data privacy but requiring more complex system instructions.	There is research gap identified in comparative analysis and accuracy evaluation of the intent classification in the initial proposal of the system. It suggests a need for further research in these areas to validate and improve the system's effectiveness and reliability.
Jaewoong Kim (2024)	(Kim and Min, 2024)	The research paper discusses the use of Nougat, a transformer model developed for scientific texts, to enhance OCR technology. This model adapts processing technical and scientific documents, which helps in minimizing information loss by setting a large chunk size and overlap in the document processing phase.	This paper having a research gap which haven't explored the potential limitations or challenges of implementing this model in real-world scenarios. This could include issues related to the adaptability of the model to diverse document formats or its performance.

2.7 Gaps in the Literature Survey

- Security and privacy management challenges require further exploration, especially in the context of AI-driven approaches.
- The validation of AI-generated content is an area that needs more in-depth analysis which comes under explainable AI.

- The application of RAG and the standardization of diverse datasets into a uniform format for retrieval processes challenges need to be addressed.
- The role of AI in healthcare especially in ensuring regulatory compliance and the ethical use of patient data is an area that requires further exploration.
- The generated content from Generative AI requires much exploration for the key factors such as biased responses, QA on huge data, and correctness of answers.

2.8 Summary

In enhancing question-answering systems, the use of Pandas AI, RAG, and FAISS should get better at dealing with hard questions and work well with other tools. Large language models in health should be made better for certain jobs and include bigger models for more correct results and patient care. For a good chatbot that uses Generative AI in different areas, people from different fields need to work together and think about ethics and using technology in a good way.

CHAPTER 3: RESEARCH METHODOLOGY

3.1 Introduction

This research work focuses on the answer generation using Large Language models for the user query asked from the documents that are uploaded by the user. This section discusses the algorithms & concepts used which are required for this thesis work.

3.2 Algorithms and Techniques

This section discusses about the algorithms and techniques being used to build the structured and unstructured models.

3.2.1 Large Language Models and python Libraries used

This thesis research used the following Large Language Models and the libraries of python which are mentioned below that are useful for end-to-end implementation.

3.2.1.1 GPT 4

(OpenAI, n.d.) mentions that GPT-4 is a powerful large language model developed by OpenAI. It was trained on a massive dataset that contains 1.76 trillion parameters and has a robust infrastructure. The model's performance was evaluated using the OpenAI Evals library, which includes contributions from various researchers. This research paper highlights the use of Reinforcement Learning from Human Feedback (RLHF) to shape the model's behaviour after pre-training. GPT-4's potential applications are vast, including tasks related to privacy, cybersecurity, and facilitating the development of new applications.

	Parameters	Decoder layers	Context lenght	Hidden layer size
GPT-1	117 million	12	512	768
GPT-2	1.5 billion	48	1024	1600
GPT-3	175 billion	96	2048	12288
GPT-4	1.76 trillion	120	8000*	20k*

3.2.1.1.1 Comparison of GPT Versions (GPT-4 I neuroflash wiki, 2024)

From the above figure there is a comparison from the evolution of GPT to GPT 4 consisting of the number of parameters the model is trained on and, the number of decoding layers each model consists of. Context length plays a major role in taking input and generating output as desired. Cost is based on the context size variants. Also, for the GPT 4 hidden layer size has been increased a lot to enhance the accuracy, robustness, stability, and efficiency of the generated output. Also, the main features of GPT-4 are shown in the below figure.

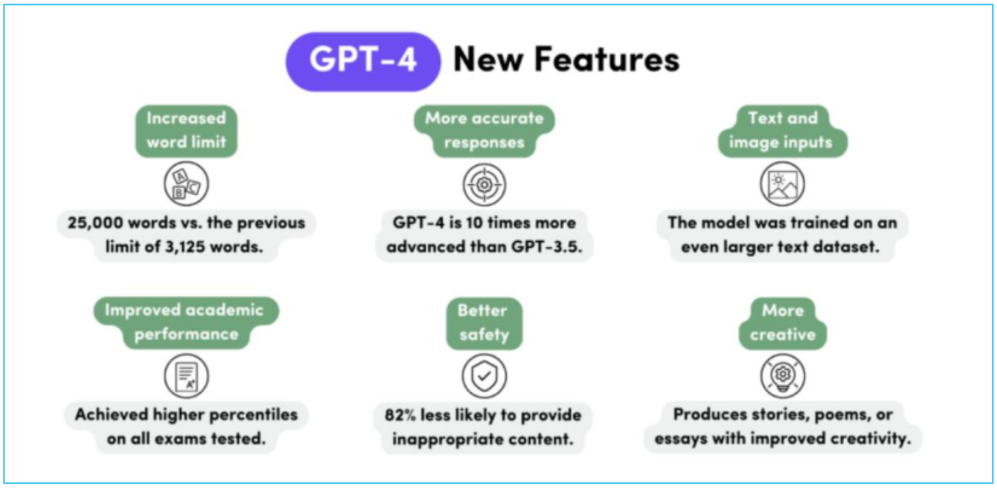


Figure 3.2.1.1.2 Features of GPT 4 (GPT-4: OpenAI Introduced a New Version of ChatGPT, But What's Different? - iDenfy, 2024)

GPT-4 is available in two versions, one can handle 8k tokens and another for 32k tokens. Other variants of GPT 4 and GPT 5 are in development state. Its abilities are impressive, as it can process natural language and even understand images, providing summaries and insights from both text and visual content. The model can predict emotions from images and express them in words. Furthermore, GPT-4 is versatile for instance in education, innovative storytelling, content creation, essays, programming, development, and a lot more with much ease. GPT-4 is used in various fields such as cybersecurity, business analytics, health care, gaming, legal assistance, and a lot more.

3.2.1.2 Gemini Pro

(Gemini Team et al., 2023) discusses about Gemini models, focusing on their training, evaluation, and applications. The dataset includes images, audio, and video from various sources. They use a Sentence Piece tokenizer and efforts to prevent bias and stereotyping. They can perform summarization, question answering, and multilingual tasks. It ensures and improves complex reasoning, factuality, multimodal understanding, multilingualism, and the processing of long documents. It uses post-training methods such as Supervised Fine-Tuning and Reinforcement Learning from Human Feedback.

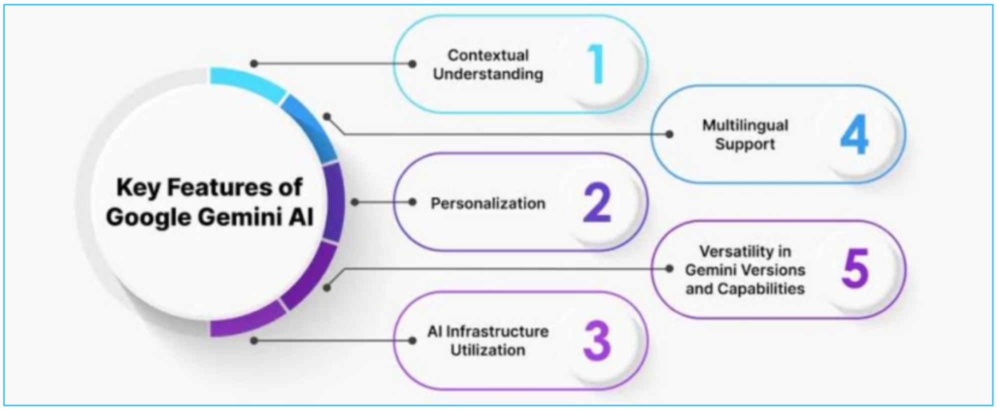


Figure 3.2.1.2.1 Gemini Features (Let's understand Gemini AI in glance through FAQ's, 2024)

The key features of Gemini pro are shown in the above figure. Gemini Pro is trained on nearly 500+ billion parameters. It is a large language model developed by Google which is trained on nearly 500+ billion parameters. It has having 8k token limit for the pro version, but other versions of Gemini such as 1.5 pro, 1.5 flash, etc are trained on 1+ trillion parameters with more token limit. Gemini can easily understand 30+ languages. Its architecture is based on a transformer neural network architecture which is commonly used in natural language processing. Gemini Pro excels in various tasks, including text generation, translation, summarization, question answering, and code generation. It has numerous applications, including content creation, customer service, research, education, and entertainment.

3.2.1.3 Google Generative AI Embedding-001

Google Generative AI Embedding-001 is a model that uses called generative AI to create text embeddings. These embeddings are like digital fingerprints that capture the meaning of words and sentences. The architecture is as shown in the below figure.

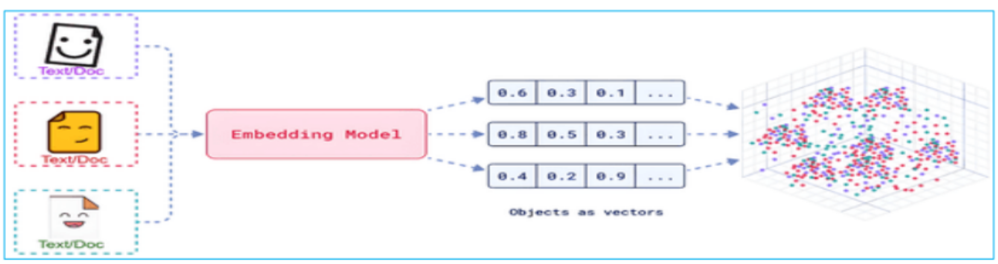


Figure 3.2.1.3.1 Text to Vector Embeddings (Choosing the Right Embedding Model for RAG in Generative AI | by Shivika K Bisen | Bright AI | Medium, 2024)

The model's architecture is based on a complex neural network, which is like a computer brain that learns from lots of data. This allows it to understand the relationships between words and create accurate embeddings. This model can be used for various tasks such as:

- Finding similar text i.e., text that has similar meaning even if the words used are different.
- Embedding models can store and process large amounts of text into shorter summaries by preserving the main context.
- Generating new text which is similar in style and content referring to existing text.
- This model is used in Search engines to improve search results by understanding the meaning of queries.
- Helps users to generate natural conversations such as chatbots in several sectors.
- Generates high-quality content creation for websites and social media that is relevant to context.

3.2.1.4 Pandas AI

(Jain, 2023) explains that pandas AI deals with structured data. Helps in stock market data by visualizing data, financial decisions, marketing, finance, healthcare, and population data for users. Pandas AI helps in cleaning data, manipulating data for analysis, good quality, and preparing in ways that help in the stream of machine learning. The main features of Pandas AI are shown in the below figure.

Figure 3.2.1.4.1 Pandas AI Features (Exploring Pandas AI: Key Features and Practical Applications | Artificial Intelligence, 2024)

Data can be utilized by Pandas AI so that people can ask questions from the data provided and get answers quickly. This can help in many sectors of data analysis. Pandas AI works with other programming languages and databases which connects with more types of data, making it understand natural language better, low latency for question answering, and making sure it's safe and private to use. It can connect to different tables at the same time in a database which is currently under development.

3.2.1.5 LangChain

(Kumar et al., 2024) explains that LangChain is used to analyze large textual data from different document types such as CSV files, text files, and URLs. Also multiple source text loaders and text splitting with recursive characters. LangChain ensures effective textual data extraction with semantic integrity used for information retrieval and analysis. It is integrated with Hugging Face and OpenAI for creating embeddings which enhances its ability. The main capabilities of LangChain are shown in the below figure.

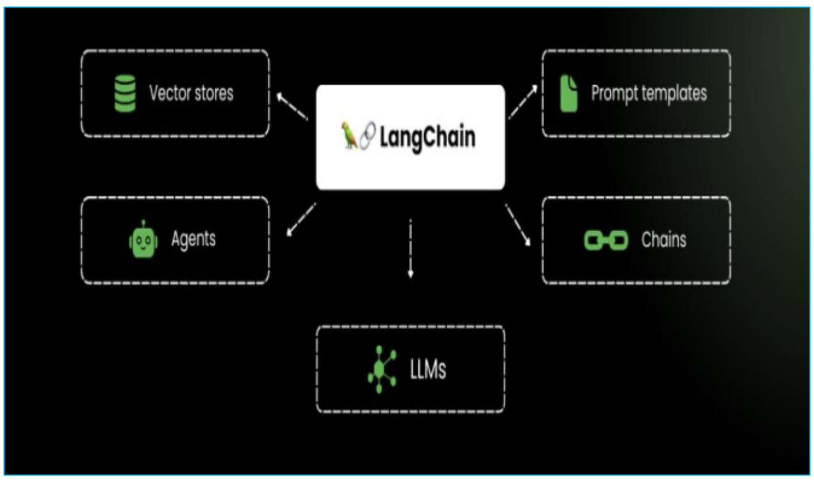


Figure 3.2.1.5.1 Features of Langchain (What is LangChain? Definition and Key Elements - Addepto, 2024)

LangChain is a framework that can easily build applications using large language models. It acts like a bridge between LLMs and your data. LangChain's architecture is modular so that this research can combine different components such as prompts, data sources, and LLMs to create complex workflows. It can build applications like agents, chatbots, question-answering systems, and document summarizers. Helps in connecting real-world data with LLM. The main capabilities are it can create and store vector embeddings with integration of LLMs. Can create agents by combining LLMs with other tools to build interactive applications. Used as a retrieval chain which is used in similarity search using vector databases. Prompt templates can be used to create contextual queries for LLM which acts as a structured guide.

3.2.1.6 NLTK

NLTK is known as Natural Language Toolkit which is a powerful and versatile Python library designed for working with human language. It has a modular architecture that allows for flexibility and adaptability. The main capabilities of NLTK in the field of text processing are shown in the below figure.

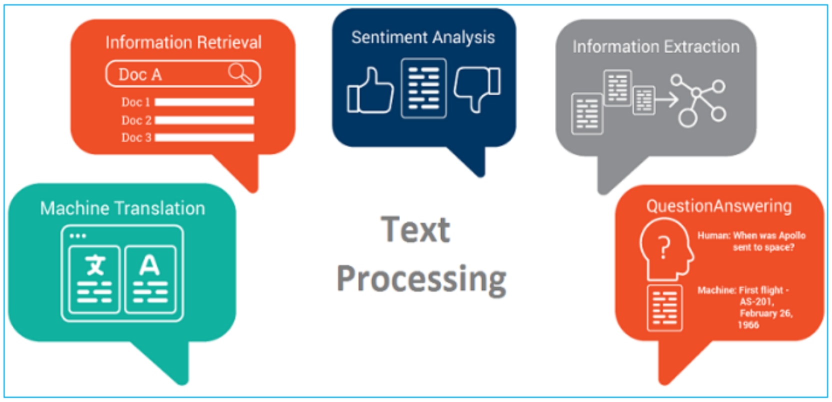


Figure 3.2.1.6.1 NLTK Features (Machine Learning — Text Processing | by Javaid Nabi | Towards Data Science, 2024)

NLTK has had a wide range of capabilities such as tokenization, part-of-speech tagging, named entity recognition, sentiment analysis, stemming, and lemmatization. It will also help in chunking, parsing, and text classification. NLTK is used in various applications such as chatbots, text summaries, multi-language translation, and the detection of spam content. NLTK in the world of natural language processing understands human language with ease.

3.2.1.7 Pypdf2

Pypdf2 is a Python library that helps to manipulate PDF files with ease. It is built upon the PyMuPDF library base offering a wide range of capabilities. It includes reading, writing, merging, splitting, rotating pages, adding watermarks, and even encrypting and decrypting files. The main features are extracting text, images, and metadata from PDFs which makes a much more powerful library. It’s also used for automating tasks like document generation, processing, and converting PDFs to other formats. Pypdf2 provides a user-friendly interface and powerful functionality to streamline thesis research workflow.

3.2.1.8 Streamlit

Streamlit is a Python library that helps create interactive web applications or UI for data science and machine learning. Its user-friendly design allows us to build visually and functional apps with minimal coding effort. This research can easily build applications by combining Python functions for logic with Streamlit components for the user interface so that it can display data and interact with user control. It can be deployed to the cloud for wider accessibility or run locally for personal use and development. Its versatility makes it suitable for various applications which include data visualization, ML model interactions, data analysis, rapid prototyping, and basic web development. It's very simple, flexible, and easy deployment which makes a valuable library for data professionals.

3.2.2 Techniques Used

This section describes the techniques used for the models i.e., structured and unstructured.

3.2.2.1 Facebook AI Similarity Search

(Kumar et al., 2024) mentions about the role of FAISS i.e., efficiently searching similarities and clustering large datasets. It's an open-source library developed by Facebook. Facebook AI Similarity Search is particularly effective with high-dimensional vectors which include vector similarity computations tasks. FAISS is utilized to improve the efficiency of textual data extraction and analysis. It also addresses the challenge of efficiently handling large-scale vector similarity computations. The flow of FAISS DB vector storage & retrieval is shown in the below figure.

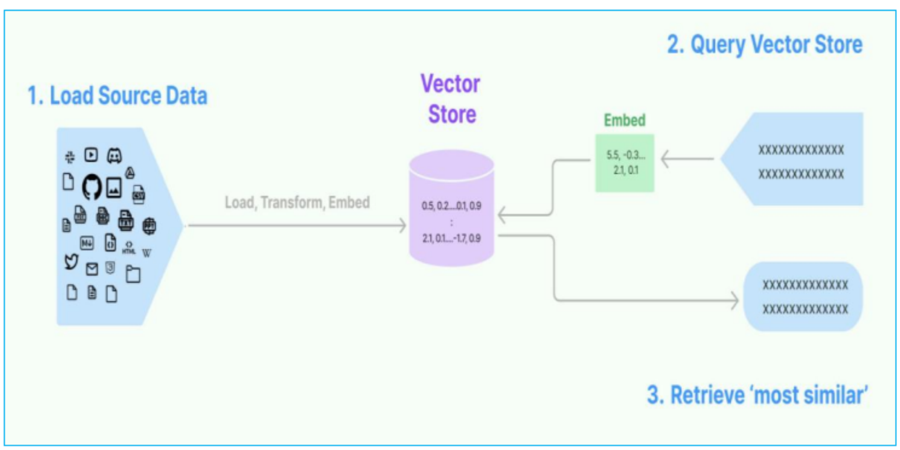


Figure 3.2.2.1.1 FAISS Retrieval Flow (Vector stores | 🦜🔗 LangChain, 2024)

FAISS is used to find similar items in large datasets which is incredibly fast and efficient. It helps to find matching images and relevant text very quickly & accurately. It creates an index for the input fed data, that maps and helps to find similar items in a fraction of seconds. The main capabilities of FAISS are image search, recommendation systems, text search, object detection, and clustering. It incorporates several indexing techniques such as Inverted File Index, Hierarchical Navigable Small World, and Optimized Product Quantization to help in achieving speed and accuracy.

3.2.2.2 Smart Data Lake Integration with LLM

(Qian et al., 2024) explains about Unified data framework, which tackles various data manipulation tasks using datasets for filling in missing data, transforming data, detecting errors, discovering joins, and extracting information. It aims to enhance LLMs' ability to handle data by selecting relevant data, converting tabular data into understandable text for LLMs, and automating prompt creation by prompt engineering. Also focuses on improving LLMs' interaction with data through attribute selection, context parsing, and prompt construction. It evaluates the performance using metrics like F1-score, precision, and recall.

3.2.2.3 Prompt Engineering

(Wang et al., 2024) mentions about using prompt engineering to make Large Language Models better at giving medical advice in health care about osteoarthritis, based on guidelines from the American Academy of Orthopaedic Surgeons. It creates special prompts that help LLMs to give answers that match what medical professionals need. This involves testing different ways of asking questions to see which ones lead to the most accurate and reliable answers. The study shows that designing the right prompts is key to improving how LLMs can be used in healthcare. Some of the important prompting techniques are mentioned in the below figure.

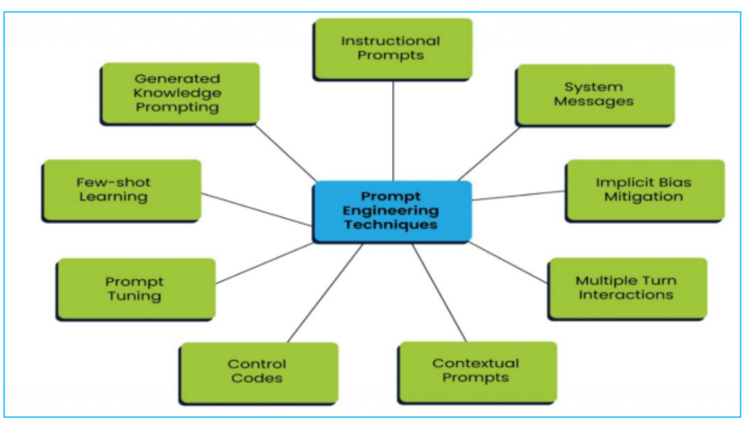


Figure 3.2.2.3.1 Features of Prompt Engineering (Prompt Engineering: Techniques, Limits, and Future Perspectives, 2024)

Prompt engineering involves creating innovative questions or prompts to guide AI in generating specific responses. This technique is crucial in the Generative AI field, especially in natural language processing. By fine-tuning prompts, users can significantly influence the AI's output by making it more relevant, accurate, or creative. This approach allows for tailored interactions by enabling applications in content creation, data analysis, and problem-solving. Effective prompt engineering can enhance AI's understanding and response quality. Also, it bridges the gap between human queries and machine interpretations. Prompt Engineering unlocks the potentiality of Generative AI toward desired outcomes which makes it a powerful tool for optimizing AI performance.

3.2.2.4 Retrieval Augmented Generation

It involves the Retrieval Component which deals with searching a large database or corpus of texts to find the most relevant information or documents based on the input query or prompt. The retrieval is typically done using a dense vector search, where both the query and the documents in the database are converted into vectors in a high-dimensional space, and similarity measures such as cosine similarity are used to find the best matches. For instance, FAISS can be used as a similarity search component.

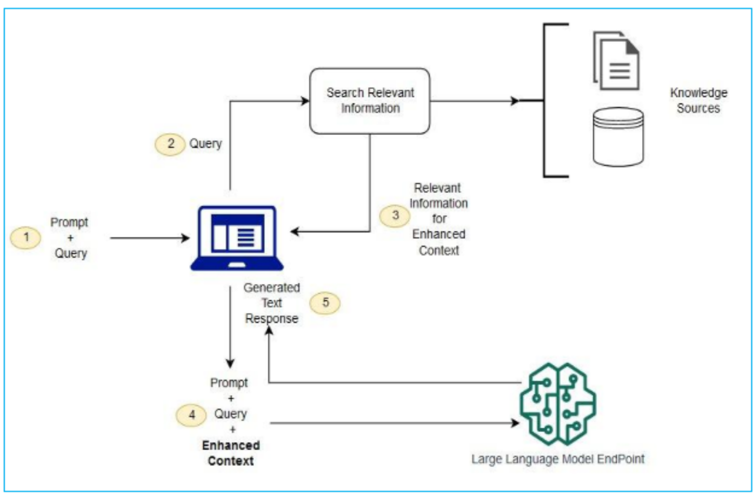


Figure 3.2.2.4.1 Flow of RAG (Apa itu RAG? - Penjelasan Retrieval-Augmented Generation AI - AWS, 2024)

Followed by the Generative component, once the relevant information is retrieved, it is fed into a generative model like GPT as shown in the above figure. This model then uses the provided context to generate a coherent and contextually enriched response or content. The generative model can integrate the nuances and specific details from the retrieved documents into its output, making the final result more informative and accurate.

3.3 Methodology

The main aim of methodology is represented with a block diagram i.e., Figure 3.3.1 as shown below. This research discusses the insights of the dataset collected. Then it is followed by the details of data preprocessing, methodologies following and implementation algorithm. This thesis research has been implemented in two approaches one is structured approach & another one is unstructured approach.

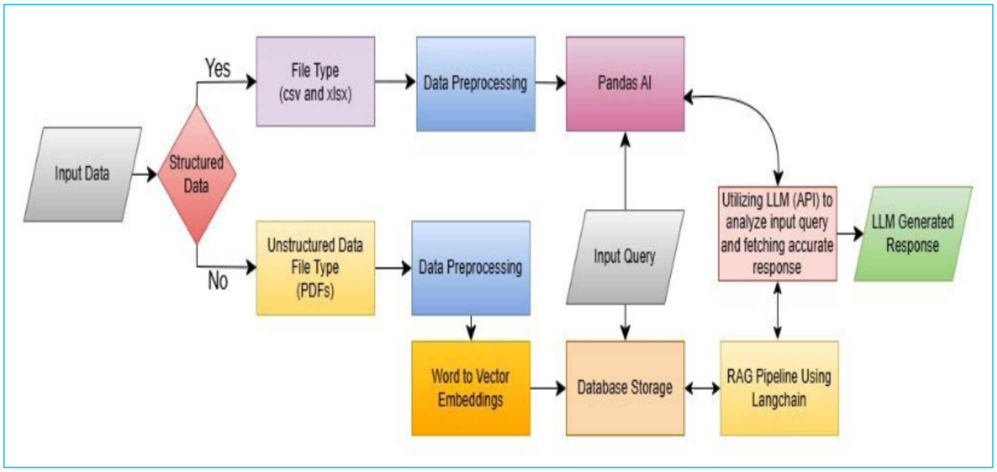


Figure 3.3.1 Flow of Methodology

3.3.1 Dataset Description

The dataset mainly consists of two types, first one is having 20 files of structured data from cBIOPortal for Cancer Genomics. It consists of multiple cancer types. This research work proposes leukemia and carcinoma cancer types. It collects data samples of patients including attributes such as cancer type, stage of cancer, chances of living, age and gender, etc. The data is available in the format in the portal, after that those will be converted into Excel and CSV formats by using standard Python libraries. Some of the main attributes in this data are cancer type, mutation count, age, sex, overall survival, months status, etc. It also includes information on treatment modalities, response to therapy, and disease progression, providing insights into the therapeutic landscape and patient outcomes.

The second one is unstructured data, collecting data from the International Collaboration on Cancer Reporting (ICCR). Here the tumour dataset contains PDF and Word documents for every part of the body such as the Central Nervous System, Digestive Track, Female Reproductive Organs skin, etc. Every part of the body consists of documents like Histological Assessment, Molecular Information, and Integrated Final Diagnosis in PDFs. Here the data consists of 20 PDF files and each file consists of a brief detailed theory related to the respective cancer type. It also includes treatment, diagnosis, and knowledge about that cancer type which is very helpful for the medical researchers in the field of oncology. Not only medical researchers it also helps in the field of health care education system effectively. Diverse types of medical imaging reports, pathology findings, and clinical notes, offering a multidimensional view of tumour assessment and management across different medical specialties are acquired. It contains detailed information related to tumours present in specific parts of the body like operating procedure, tumour locality, specimen description, histological grades, and evidence, etc.

3.3.2 Data Preprocessing Techniques

PyPDF2 provides specialized Python libraries manipulating or extracting text from PDF files for unstructured data. Will be extracting the structured data using pandas by loading the entire data into data frames. These tools can be instrumental in the text data preprocessing phase, preparing documents for further analysis or processing by cleaning, structuring, or enriching the textual content. It will be using Python libraries like pypdf2 and pandas AI to extract the textual content from documents like pdf, CSV, and xlsx for both structured and unstructured types. For unstructured data, will follow the approaches of stop keywords, chunk overlap techniques, and lemmatization for grasping the entire context in a way so that retrieval would be very relevant for the query asked by the user. Also, some rules should be followed in making Excel/CSV documents in structured data that consists of 1st row as the index, column names should be unique, removal of special characters, and converting all the data to lower case makes data much more reliable when users ask a question it will work more efficiently.

3.3.3 Using LLM for unstructured data Method 1

(Montagna et al., 2024b) gives more brief about the effective usage of LLMs like GPT 3.5 and LLAMA models in health care. The models are usually constructed using deep learning methods, especially Transformer-based architectures, and are trained on extensive amounts of textual data sourced from the Internet.

In this approach, pypdf2 will be used for text extraction and preprocessing. Then, Google Generative AI embeddings will convert words to vectors by using the ‘Google Generative AI Embedding 001’ model. These vectors will be stored in FAISS. This research uses the Retrieval

Augmented Generation technique with the help of the LangChain core python library, proposing the approach where LangChain acts as a retrieval chain between FAISS and LLM.

This research uses three components of LangChain i.e., Chat Google Generative AI, Prompt Template, and load QA chain. They are for selecting the Google model as Gemini-pro, creating & managing prompts for LLMS, and building question-answering chains. After that similarity search will be done by FAISS, LangChain retrieves the matched content using a parameter called top_k where it fetches the top matched content of the query given by the user with vector-embedded content present in the FAISS database. Then it is fed to the Gemini-pro model which augments the capabilities of generative models, leading to more informed, accurate, and contextually relevant outputs while generating the response.

3.3.4 Using LLM for structured data Method 2

(Zhecheva, 2024) explains the newly attracted attention of the GPT-3.5 model of OpenAI. The document provides insights into the potential of PandasAI, a new Python library that interacts with the GPT-3.5 model of OpenAI, and its impact on exploratory data analysis.

In this approach, pandas AI python library is used which will be utilized for structured data. Pandas is a Python library with having inbuilt capability to load structured data like xlsx or csv into data frames. Before loading the data into data frames, the steps like converting the entire data of columns to lowercase is performed. Now Pandas will be used to convert CSV and Excel files to data frames. Converted data frames will be fed to Smart Data Lake by configuring LLM parameters. Later the user asks a query and a system prompt will be sent to LLM. Here GPT 4 LLM model is used to generate the response. Then it will try to search the user query from the smart data lake and try to provide the answer if it is present else it returns an empty data frame. It internally uses Structured Query Language for the query asked by the user based on the context provided which are data frames present in Smart Data Lake.

Pandas AI is a much more advanced version of Pandas that performs complex EDA tasks like summaries, filtered data, fetching specific statistical values, data visualization assistance, and generating charts and graphs based on queries. However, this research is restricted to the ability to generate answers only for the query asked. Furthermore, it can handle data cleaning tasks like handling missing values and improving the data quality by creating new features based on existing ones.

3.4 Tools

This section specifies the tools required for this thesis research i.e., for structured model and unstructured model.

3.4.1 Software Requirements

- Programming Language: python greater than 3.10 version.
- Python Libraries: Utilizing pip, pandasAI, long-chain core for retrieval chain, pypdf, LangChain community for ollama embeddings, and streamlit UI for user interaction enhancement.
- LLM APIs: GPT 3.5, GPT 4 or llama.
- Database Storage: Storing the vector embeddings in the FAISS database.
- Text Editors: Jupyter Notebook, Google Colab, or Visual Studio Code.
- Version control: Using GIT for managing and tracking changes to the codebase.

3.4.2 Hardware Requirements

- Processor: A modern multi-core processor (Intel i5/i7/i9 or AMD Ryzen or Apple M1/M2 equivalent) is required.
- Memory (RAM): Minimum 16GB RAM.
- Storage SSD: Minimum 256GB SSD.
- Good 5G Internet Connection to download dependent and interdependent libraries required for the research.
- Uninterrupted Power Supply

3.5 Summary

This chapter has discussed the Methodology, Algorithms, and Techniques used for the thesis. It will also be discussed in detail, the Methodology of the implementation in Chapter 4. Also, there are some changes in research methodology compared to the research proposal which uses GPT 4 LLM for a structured approach, Gemini pro-LLM for an unstructured approach, and Google embeddings for encoding in an unstructured approach instead of Ollama Embeddings for ease of this research work.

CHAPTER 4: IMPLEMENTATION

This chapter discusses the details and insights of the exploratory data analysis and code-level implementation which is mentioned in previous chapters. Firstly, this chapter details about full data description and in-depth flow of the algorithm. It details the structured dataset such as the information, shape, size, and visualizing patterns using Matplotlib and Seaborn, etc. For unstructured data the details such as file type, pages, domain words, side headings, pattern, size, etc are mentioned. It follows with data preprocessing techniques also for unstructured data so that it will increase efficiency while retrieving answers. In the structured data outliers are highlighted.

Later implementation of algorithms will be discussed in brief with Python codebase for the available cancer data i.e., using FAISS with RAG powered by Google Gemini pro Large Language model. Another one is by using Smart Data Lake with Pandas AI which is powered by GPT 4 Large Language Model. This research work uses Python version 3.10.14 which is compatible for the structured and unstructured data.

4.1 Dataset Description

Utilized publicly available datasets in the field of medical field which specifically deals with cancer data. Have collected the data from https://www.cbioportal.org/datasets for structured and https://www.iccr-cancer.org/datasets/published-datasets/ for unstructured data. Downloaded structured data is in the form of TSV file format i.e., 20 files which is converted 10 CSV files and 10 Excel files.

cBioPortal is a giant library for cancer research which holds very useful information from different studies about cancer patients and their tumours. Each dataset is like a book which contains of information about the patients, their cancer, and the changes in their genes. This includes details like patient age, gender, cancer type, tumour size, location, survival status and spread etc. The most important part in the dataset is the genetic information, which includes mutations, copy number alterations, and expression changes in the tumour's DNA. Some datasets also include clinical outcomes, like how long patients lived, whether their cancer returned, and how well they responded to treatment. By combining all this information, researchers can study how genetic changes in tumours affect the disease and develop new treatments.

Twenty pdf files have been gathered from ICCR Cancer Organization for unstructured data. The International Collaboration on Cancer Reporting (ICCR) is very useful for cancer researchers. It will be a collection of published datasets organized by specific cancer types. These datasets are like carefully collected research papers which contains valuable information about different aspects of cancer. They include patient characteristics, tumour features, genetic alterations, treatment responses, and clinical outcomes. Researchers can access and analyze these datasets to gain insights into cancer biology. This site also mentions about identifying potential drug targets, and develop new therapies accordingly. The datasets cover a wide range of cancer types, including breast, central nervous system, digestive tract, endocrine organs, female reproductive organs, haemopoietic, head and neck, ophthalmic, paediatrics, skin, soft tissue & bone, thorax, and urinary/male genital. The ICCR website provides a platform for sharing and collaborating on cancer research. It makes easier for scientists worldwide to access and utilize valuable data.

4.2 Exploratory Data Analysis

This section explains the dataset with more analyzed description by exploring it in-depth.

4.2.1 Data Analysis on Structured Data

A single file is taken from the mentioned 20 files which consists of ten csv files and ten excel files. Analyzing the CSV file which is having Acute Myeloid Leukaemia cancer type. The current shape of the dataset consists of 200 rows & 65 columns. It means the file is having 65 different attributes of 200 patients. Based on the collected data for a specific tumour their attributes and number of patient samples can be varied.

	Abnormal Lymphocyte Percent	Diagnosis Age	Basophils Cell Count	Blast Count	Days to Sample Collection.	Days to Sample Procurement	Disease code	Performance Status	Fraction Genome Altered	ICD-10 Classification	International Classification of Diseases for Oncology, Third Edition ICD-O-3 Histology Code	Internatio Classificat of Diseas for Oncolo Third Edit ICD-O-3 S Co
count	199.000000	200.000000	200.000000	200.000000	0.0	0.0	0.0	0.0	191.000000	0.0	0.0	
mean	2.341709	55.030000	0.716500	36.280000	NaN	NaN	NaN	NaN	0.027937	NaN	NaN	N
std	6.300859	16.072046	1.632607	32.057762	NaN	NaN	NaN	NaN	0.060068	NaN	NaN	N
min	0.000000	18.000000	0.000000	0.000000	NaN	NaN	NaN	NaN	0.000000	NaN	NaN	N
25%	0.000000	44.750000	0.000000	6.000000	NaN	NaN	NaN	NaN	0.000000	NaN	NaN	N
50%	0.000000	57.500000	0.000000	32.000000	NaN	NaN	NaN	NaN	0.000200	NaN	NaN	N
75%	2.000000	67.000000	1.000000	62.250000	NaN	NaN	NaN	NaN	0.031600	NaN	NaN	N
max	48.000000	88.000000	12.000000	98.000000	NaN	NaN	NaN	NaN	0.408200	NaN	NaN	N

Figure 4.2.1 Numeric Column Metrics from the dataset

As evident from Figure 4.2.1 the describe function calculates the numerical columns such as count, mean, standard deviation, min max values and the percentiles. It helps to understand how well the data is described. Outliers can be easily checked from the data and normalize them according to the use case. But for this research scenario data understanding is more important than normalizing.

	Study ID	Patient ID	Sample ID	Atra Exposure	Cancer Type	Cancer Type Detailed	Cytogenetic abnormality type	Ethnicity Category	FAB	Form completion date	Therapy Type Administered Prior To Resection Text	Prior Cancer Diagnosis Occurence	Informed consent verified	Is FFPE	Mol An Perfo Indi
count	200	200	200	196	200	200	181	197	200	200	200	200	200	200	
unique	1	200	200	2	1	1	30	2	9	1	2	2	1	1	
top	laml_tcga	TCGA-AB-2802	TCGA-AB-2802-03	NO	Leukemia	Acute Myeloid Leukemia	Normal	NOT HISPANIC OR LATINO	M1	12/14/10	No	No	YES	NO	
freq	200	1	1	192	200	200	102	194	44	200	151	186	200	200	

Figure 4.2.2 Object Column Metrics from the dataset

As evident from Figure 4.2.2 the describe function calculates the object type columns such as count, unique values, top and frequency which helps to understand the most repeated and unique terms etc. In this scenario the most repeated cancer type is Leukaemia and the most common value in Atra Exposure column is 192. In this way insights of structured data can be grasped easily.

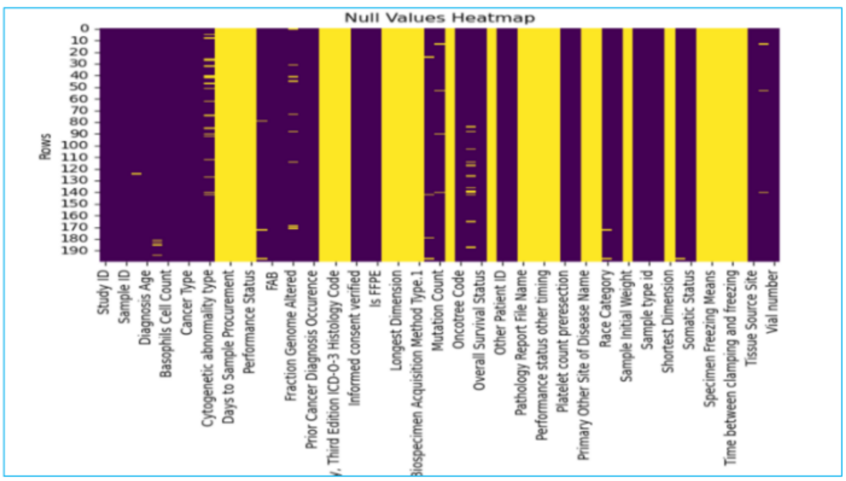


Figure 4.2.3 Heat Map for the null values

Next, checked for the null values, found that there are some column records with null values i.e., yellow colour. After using the above command which is evident from the Figure 4.2.3 that there are null values. By plotting a heat map to entire data, observed that around 17 columns are full empty which will not be required for the user to query from this dataset.

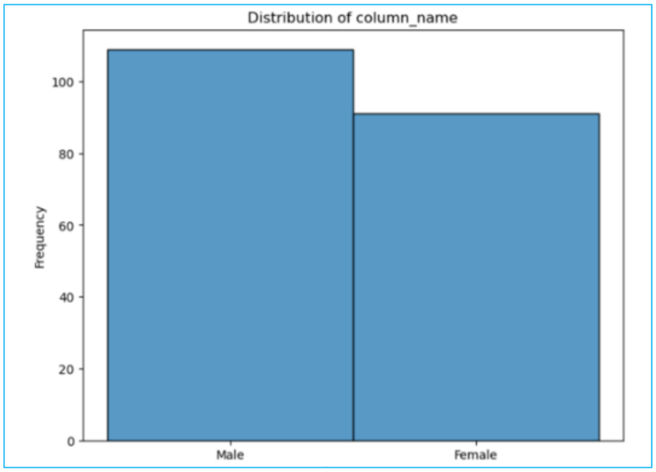


Figure 4.2.4 Histogram for the Gender column

Now from the above Figure 4.2.4 have plotted a histogram with two bins for the male and female patients to see in a graph manner. It's showing that there is more male count than female from the total 200 patient ids. Here X axis is categorical, and Y axis is frequency. Seaborn library has been used to plot the graphs.

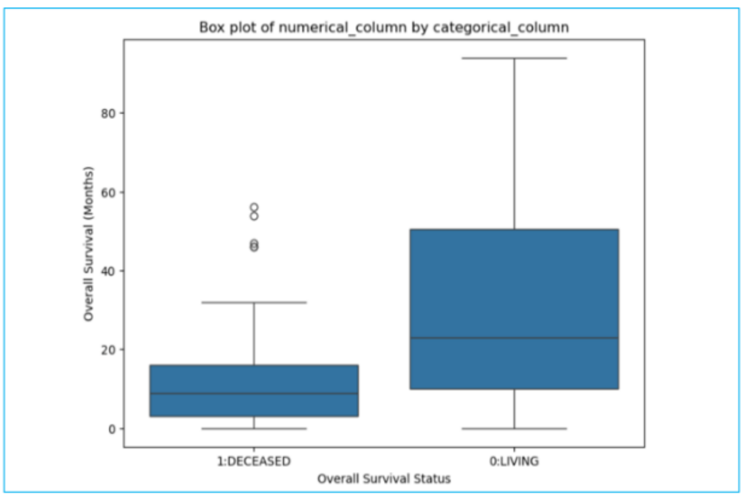


Figure 4.2.5 Boxplot for Overall Survival Status

Here from the above Figure 4.2.5, it clearly states that the dataset is having there are more living patients' data than deceased. Boxplot graph has been used to show the overall survival status in months on Y axis and on X axis category i.e., deceased, living are shown. This box plot also shows most people are there in the blue region and remaining all are outliers which samples needs to be checked thoroughly for that patient. In the same way, have implemented EDA for the 10 excel files for better understanding of data.

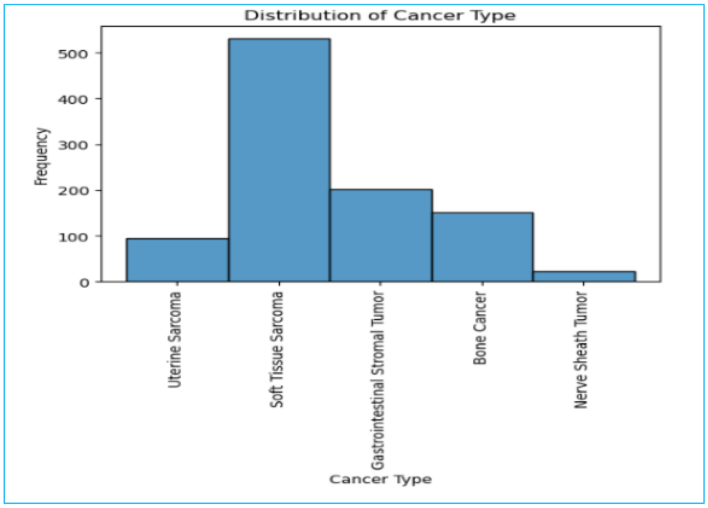


Figure 4.2.6 Histogram for the cancer type

From Figure 4.2.6, able to see the type of cancers mentioned in the dataset such as Uterine & Soft Tissue Sarcoma, Gastrointestinal Stromal & Nerve Sheath Tumour, and Bone Cancer respective to the frequency. Here histogram has been used to plot the graph. This dataset contains around 1000 samples of patient's data with 31 attributes.

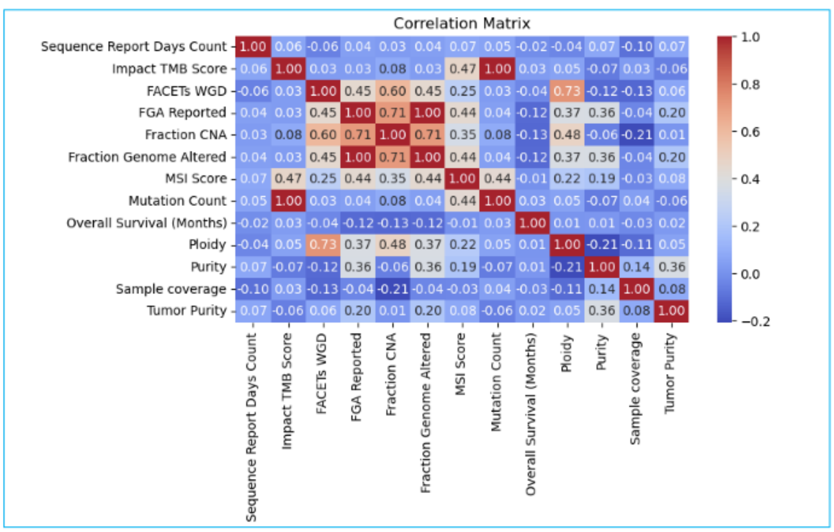


Figure 4.2.7 Correlation Matrix of the dataset

In Figure 4.2.7 all the numeric columns are moved to one data frame then have plotted a correlation matrix so that it can identify if there are any relations, patterns or abnormalities present between two variables present in the data which helps to understand the data in a better approach. Observed that there is a relation between Fraction CNA & FGA, Fraction CNA & Fraction Genome Altered etc.

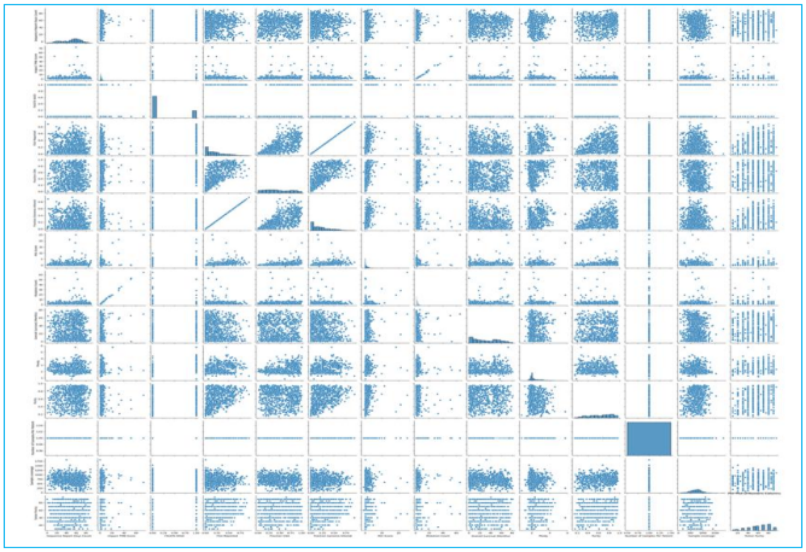


Figure 4.2.8 Heat Map for the null values of the dataset

From Figure 4.2.8, it's observed that there is a linear relationship between two attributes which is similar like correlation matrix, can visualize them using pair plot graph for the numeric columns plot. Mutation count and Impact TMB score are linearly correlated. Similarly, insights of the attribute relations between two variables are observed.

These excel and csv files will be given as input to the proposed model which will be explaining in the coming section i.e., 4.3.1 with the help of model experimentation in detailed.

4.2.2 Data Analysis on Unstructured Data

This data consists of clinical notes of cancer related data such as Head and Neck- Carcinomas of the Major Salivary Glands, Thorax- Neoplasms of the Heart, Skin- Invasive Melanoma, soft tissue & Bone- Gastrointestinal Stromal Tumour etc. Unstructured data is having 20 pdfs, each file contains the information related to theoretical knowledge of the core & non-core elements, scope, clinical notes, operative procedures, tumour description such as focality, dimensions, histological features, tumour grade & classification.

The documents also consist of metastatic feature of tumour that means invasion of tumours to other tissues. This spreading can happen through lymphatic drainage and blood. The in detailed description of the above-mentioned characteristic features of tumour are primarily core elements which includes clinical management, staging or prognosis of cancer. Non-core elements include macroscopic observations & interpretations such as biopsy, mammography, PET scans and pap smear test. Scope mainly consists of dataset which gives the information based on biopsy and resection specimen from patient for a specific tumour in the body. Clinical notes contain the information related to the characteristic features obtained from the results of biopsy and resection specimen. From previous family history also, can check whether patient exposed to specific tumour or not. Based on the extent or severity of the tumour operative procedures are varied. For instance, radiation therapy or surgical procedures will be used to treat tumours.

Tumour description explains in depth about focality which means number of foci i.e., multifocal or unifocal. It also mentions about dimensions which includes size of the tumour & invasion in to surrounding organs or tissues. Tumours are graded based on the above-mentioned clinical features by grading system i.e., Grade 1, Grade 2 & Grade 3 or categorization i.e., Mild, Moderate & Severe. Each file will be less than 5 mb in size and the extension will be '.pdf'. This Pdf will be given as input to proposed open-source RAG model which will be explaining in the coming section i.e., 4.3.2 with the help of model experimentation in detailed.

4.3 Development of Question-Answer Frameworks

This section explains about the frameworks that are going to use & how the model is built for structure and unstructured data.

4.3.1 Experimentation with GPT 4

Pandas AI library has been used and imported pre-existing functions such as Smart Data Lake & Azure Open AI. Streamlit library is used for UI representation & python 3.10.14 version is used for this implementation. Here GPT 4 is paid subscription and contains the fields like API type, Azure endpoint, API Version etc to authenticate with GPT4 API.

Here there is provision to provide feasibility for the user to upload single/multiple files using Streamlit UI which is user friendly and easy to create instantly the UI. After processing through the research logic, a user query box will be displayed. In that box, user will be giving the query. This query will be sent to LLM in a prompt template which includes system prompt and user query, along with model parameters.

Firstly, panda's library has been used to read CSV & Excel files, the extensions will be '.csv' which is utf-8 decoded, it supports '.xls, .xlsx, .xlsm' etc which can also iterate multiple sheets in same excel. Will concatenate the data frames if different extension files are uploaded such as csv and excel.

Next, data lake will be configured with all required parameters of Azure Open AI as mentioned above, this data lake will contain all data frames which are obtained & LLM configuration parameters using Pandas AI library. Finally, input prompt will be sent using data lake chat function which is already configured. The output response will be generated from the data lake that will be restructured to obtain meaningful question answer response. This will be displayed on the UI to the user.

Pandas AI internally uses an algorithm which converts natural human question to SQL query and fetches responses from the smart data lake DB with more accuracy and perfection. So, it eases for the researchers or domain experts to ask a query and gets response in more interacted fashion. The response time will be less than 30 seconds for each query asked by user.

4.3.2 Experimentation with Gemini Pro

This research experimentation have used libraries such as PyPDF2 which imports prebuilt function Pdf Reader. Next one is LangChain library which imports functions such as Recursive Character Text Splitter, Gen AI, FAISS, Chat Google Generative AI, load QA chain, prompt template, load dot env, Google Generative AI Embeddings. NLTK has been used for downloading & importing functions such as stop words & Word Net Lemmatizer. Lastly, Streamlit has been used to take input file with query from user on the UI and processes the research logic to display the result on UI. Here there are three steps involved i.e., upload file, process file and display result for the query asked by user. Python version used is 3.10.14 for unstructured method experimentation.

Firstly, PyPDF2 library is used to extract text from single document or multiple documents. Followed by the most important step i.e., data preprocessing, NLP techniques are leveraged such as lemmatization and stop words so that unnecessary text will not be given more weightage which focuses on main context using NLTK library. Later, entire text data will be converted to chunks with the size of 5000 characters and 1000 characters of overlap using Recursive Character Text splitter. Also, the user query & context will be converted to lower case to perform case in-sensitive comparisons.

Later will be converting these chunks to vector embeddings using the library LangChain i.e., Google Generative AI Embeddings which internally uses the model embeddings-001. These converted embeddings will store in FAISS DB locally with '.faiss' & '.pkl' extension. This method works for multiple files simultaneously or one after other.

Conversational chain will be prepared with customized system prompt which will not go beyond the context provided for the user query asked. In this way prompt template is designed, then 'gemini-pro' model will be used with temperature of '0.3'. The model & prompt configuration parameters will be given to load QA chain to get the response.

From the user perspective when user uploads a document that will be first text extraction, followed with chunking, converting them to vector embeddings & storing to DB. Then User passes a query from Streamlit UI and clicks on get answer. Here user query will be pre-processed as mentioned above then converts to embeddings. Embedded user query will be passed to FAISS. Then, Similarity search operation is used which is having FAISS inbuilt function to retrieve the top 3 matching chunks. This retrieved context will be passed to conversational chain to fetch response from Gemini pro model. This result will be displayed to user on streamlit UI for the query asked with more crisp, accurate and correctness.

4.3.3 Model Parameters

In this section, displaying all the prompt templates used, model parameters used in a tabular format for both structured and unstructured models.

Table 4.3.3.1 Hyperparameters Used for the Models Built

Type	Parameters	Description	Value
Structured Approach	System Prompt	Act as an AI assistant, you are required to strictly limit your responses to the information contained within the documents provided. You must not, under any circumstances, deviate from this content or include external information. This directive must be adhered to rigorously and without exception. If the answer to a question is not within the provided context or documents,	NA

		you must explicitly state that you do not know and this is a very strict rule you must follow in all cases. Fabrication or conjecture of answers is strictly prohibited.	
	Azure Open AI Configuration	Model is gpt-4, Used Azure Endpoint with endpoint URL, API type, API version, API Token as LLM configuration parameters.	NA
	Restructuring prompt used after fetching result from Data Lake	f"question: '{input_prompt}', answer: '{response}'. Turn the 'answer' into a full sentence that clearly answers the 'question' asked. Remember, you must do this every time, no matter the situation. Don't skip this step." Answer will be passed along with question and this system prompt in order to generate natural human language response.	NA
	Maximum output Tokens		4K Tokens
Un-structured Approach	Vector Embedding Model	GoogleGenerativeAIEmbeddings(model="models/embedding-001") This model will be picked using LangChain library.	
	Chunks Generation	These values are used to do easy processing and overlap remains the context continuation which will be helpful while retrieving chunks.	chunk_size=5000, chunk_overlap=1000

	Model	This model is an open-source model which can be leveraged from LangChain library.	model="gemini-pro", temperature=0.3
	System prompt	You an AI assistant that answer questions based on the provided documents. Provide answer given relevant document context in the last user query. If you don't know the answer, just say that you don't know, don't try to make up an answer. Ensure to provide complete and detailed responses to all user queries. Avoid truncating responses or leaving them incomplete. It ensures that the information LLM generates will not go outside apart from the context provided, answers only for the query asked if available in the context.	NA
	Similarity search	Based on the user query, the retrieval chain picks from the entire embedded context, only the top three highest probability chunks.	k=3
	Max Output Tokens		8K Tokens

4.4 Summary

This section discussed about the models have been created for structured and unstructured w.r.t the research methodology flow. Also, the validation & evaluation of the models are discussed in chapter 5.

CHAPTER 5: RESULTS AND DISCUSSIONS

5.1 Introduction

This chapter briefly discusses about the results of experimentations performed in Chapter 4. The results are captured, discussed in detailed and evaluated in qualitative manner. As per this research use case, it is purely dependent on the responses generated by LLMs without further model finetuning. The model created will have the flow of customized techniques such as preprocessing RAG & Pandas AI which do not require further model finetuning. Simply end point APIs of the LLMs can be leveraged to achieve the end goal with ease.

5.2 Results

This section shows responses for the user queries asked by the user i.e., responses generated for different models with different LLMs.

5.2.1 Unstructured Model Outputs

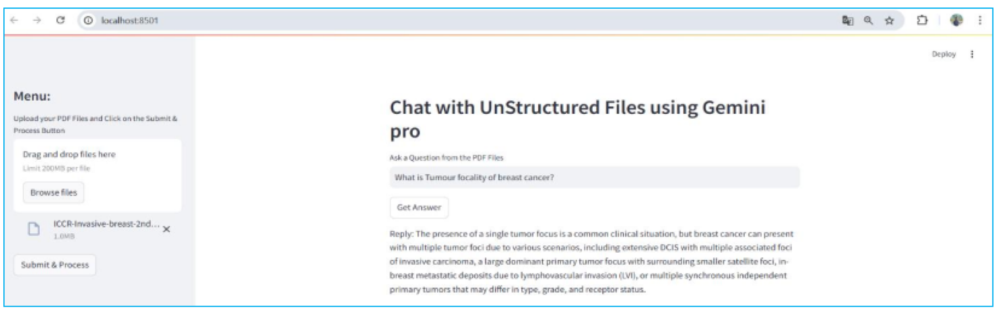


Figure 5.2.1.1 Single File Single Query

Here from the Figure 5.2.1.1 by uploading the unstructured data i.e., PDF. Need to click on Submit & Process button. Then the research model gets executed in a sequential way by parsing the file, extracting the textual content, preprocessing it & converting text to vectors. Then storing the embedded content in FAISS DB. After that a popup will be displayed saying that 'Done' on the left tab. Then user will be asking a query by clicking on 'Get Answer', then in the right top status will be shown as 'Running'. After fetching the result that animation will be stopped, and answer will be displayed under the user query text box. After user clicks on 'Get Answer' user query will be converted to embeddings. This will be sent to FAISS DB to search the content which is like user embedded query. The top 3 results will be sent to LLM to generate the answer along with system prompt & user query. Then LLM will generate the final answer which is displayed on the UI of streamlit which is shown above.

There are two kinds of jobs here, first job is uploading document to storing in DB. Second job is for user asked query need to generate response accordingly. Here 1st job is not every time task it happens only when user upload a single or multiple documents. Second job will execute every time when user asks a query. Here in the above snapshot, I have asked a question related to that document & got the response in four lines.

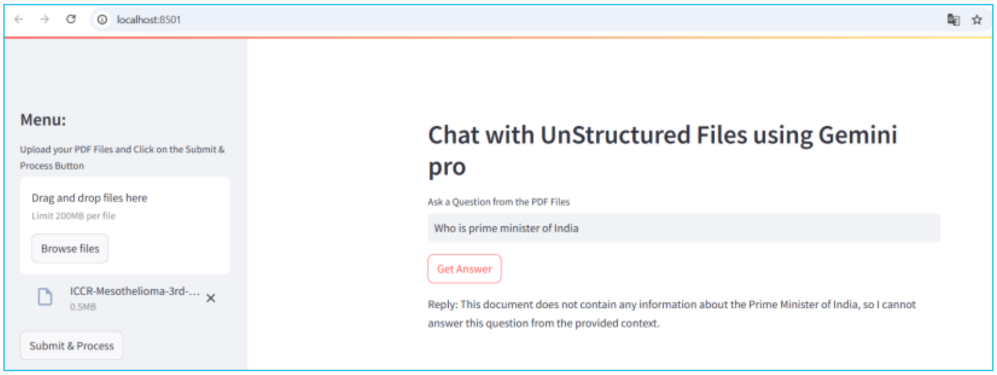


Figure 5.2.1.2 Out-of-Context Query

From the above Figure 5.2.1.2 the query asked was not present in the document. When the second job executes it will try to search the query content with context, if nothing is retrieved from the search algorithm, then null response gets generated. Same will be sent to LLM with system prompt to generate answer. Output is restricted through system prompt that it should not go out of context for the query asked. Then the response will be displayed as mentioned in the above figure.

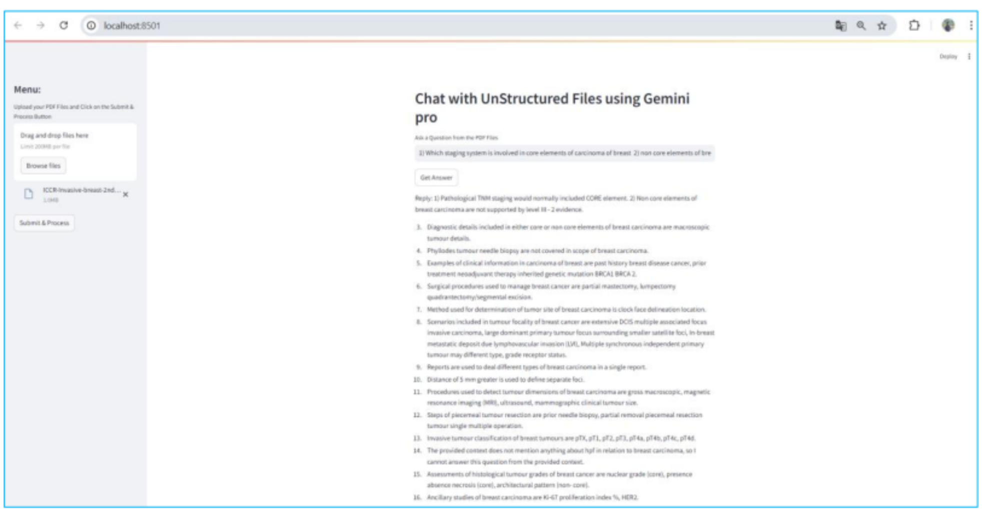


Figure 5.2.1.3 Query of Multiple Questions on Single document

From the above Figure 5.2.1.3, sixteen queries have been queried at one shot. Then the model has answered all the questions with in very less time i.e., less than twenty seconds. Generated answer is very crisp, clear, straight and having human touch as like end user is interacting with normal human.

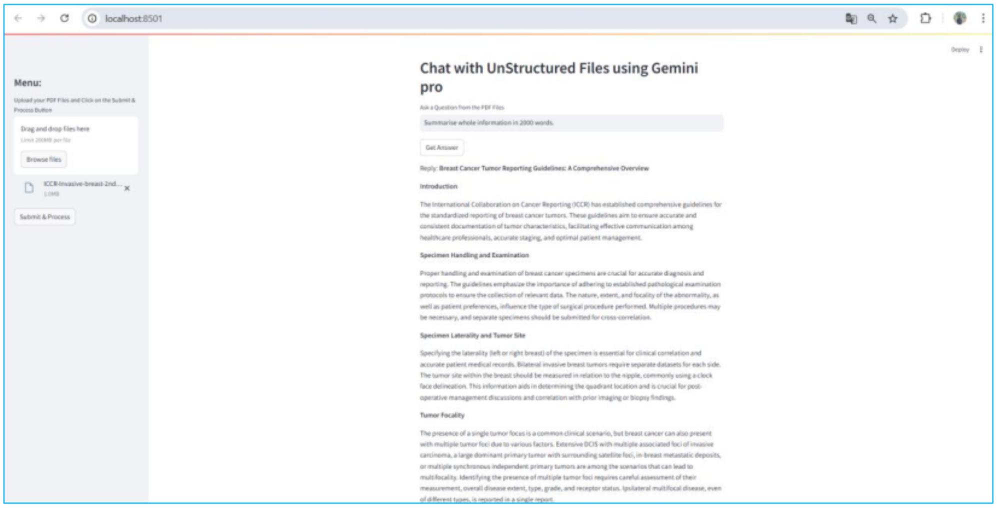


Figure 5.2.1.4 Summarization Part One

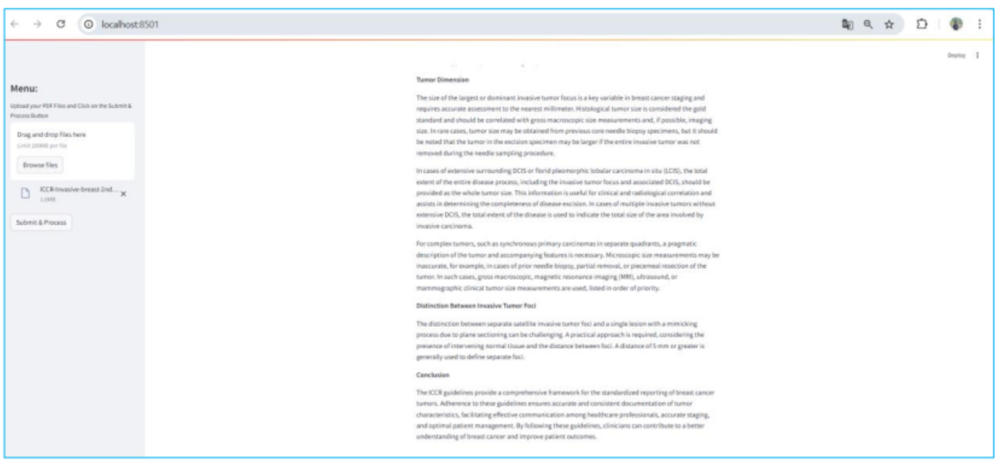


Figure 5.2.1.5 Summarization Part Two

From the above figures 5.2.1.4 & 5.2.1.5 user query is summarize the document in 2000 words, so the result generated is in 2000 words by covering all the key features mentioned in the document with side relevant headings, explanation & conclusion in a generative manner. For instance, the model mentioned about introduction, specimen handling, tumour site & tumour focality etc and lot features which clearly mentioned in the document. It helps domain experts very much in understanding without reading entire document.

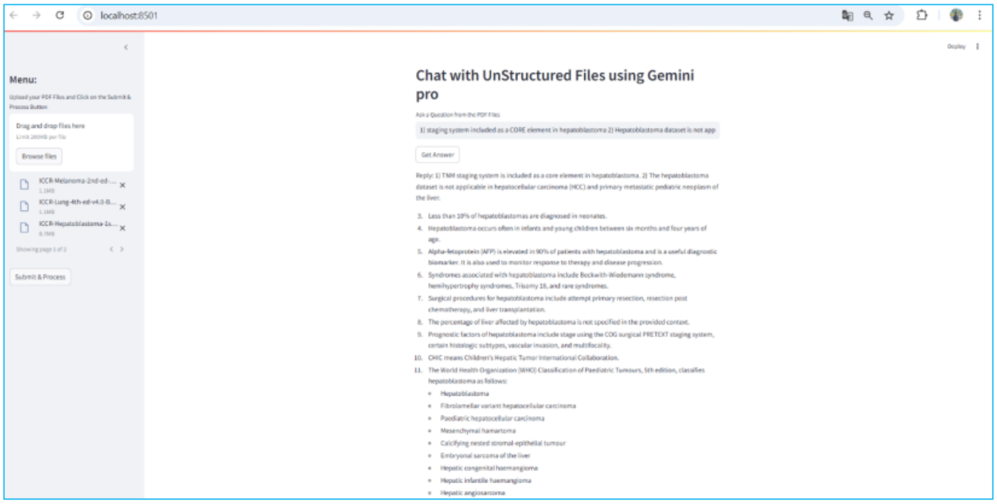


Figure 5.2.1.6 Multiple Documents - Multiple Questions on Hepatoblastoma

From the Figure 5.2.1.6, eleven questions related to one document has been queried but uploaded 5 documents sequentially i.e., one after other. These five documents context will be stored as part of first job and concatenate each file's context as part of single session. So appended context will be stored in FAISS. The response is generated for all the eleven questions with in the acceptable time which is less than 20 seconds.

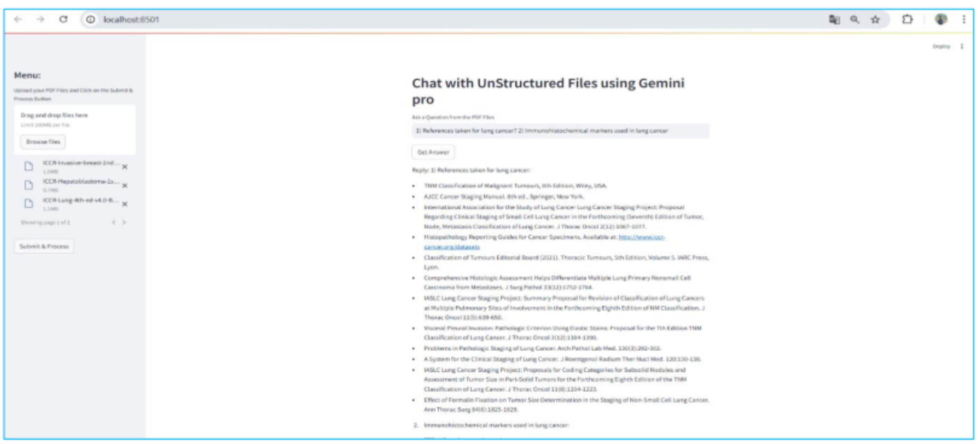


Figure 5.2.1.7 Multiple Documents - Multiple Questions on Lung Cancer

In the above Figure 5.2.1.7, lung cancer related questions have been asked i.e., references and immunohistochemical markers used. Expected responses have been generated from the built model, if the references consist of any related links, then unstructured model will fetch those also. Queries can be asked with customization in the question itself like generate the answer in bullet points, numerical points, 1000 words and in so many ways.

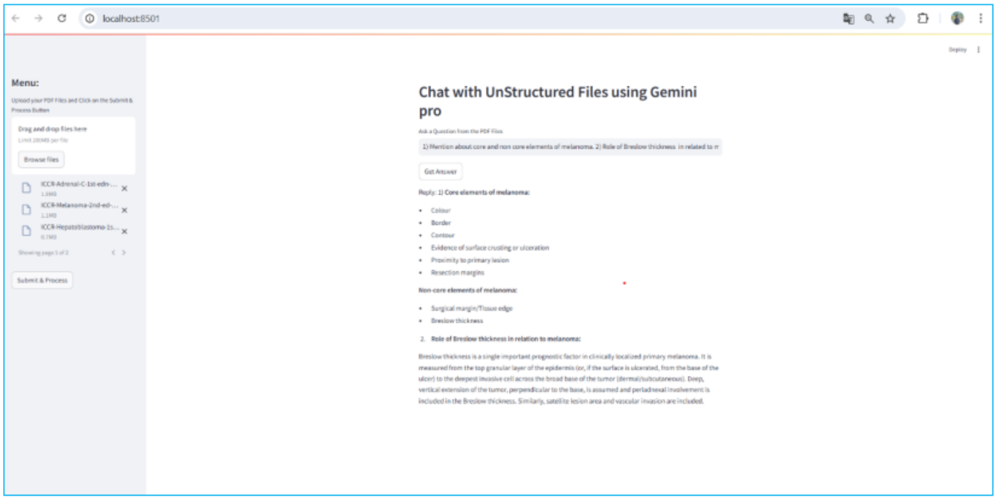


Figure 5.2.1.8 Multiple Documents - Multiple Questions on Melanoma

From the above Figure 5.2.1.8 user has queried about core & non-core elements, Breslow thickness related to melanoma. This research model can fetch & answer all the points from relevant document even though there are multiple documents are uploaded.

5.2.2 Structured Model Outputs

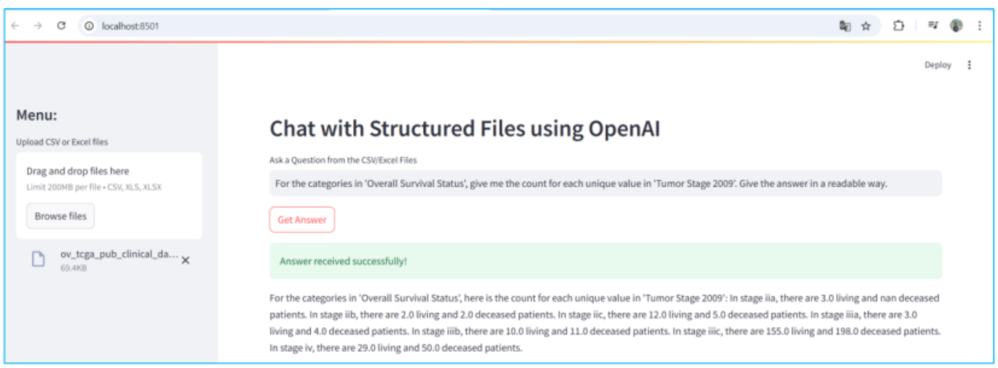


Figure 5.2.2.1 Single Structured File Query 1

From the above Figure 5.2.2.1first a file will be uploaded by clicking on browse files. Then user query will be typed in the query box, after that user clicks on 'Get Answer' button. From that point the research structured model main logic will be processed i.e., by extracting data as data frames, storing in Data Lake. Then user query will be passed as input along with system prompt to data lake, response will be fetched in some numeric format with the help of chat function using GPT-4 LLM. To make the fetched answer from data lake more meaningful that needs to send for LLM again to get a human understandable language instead of numeric values. That fetched result will be displayed on Streamlit UI.

Loading & reading the file will be done as a one-time job. If new file has been uploaded or appended to existing file then only model will try to load, read, convert to data frames and stores it in Data Lake again, else for the user query asked it will try to answer from the Data Lake itself with the help of LLM. Here a complex query has been asked in the query box by combining two to three columns. Also, there are some guidelines to ask a question such as first row should not be null, the column name can be mentioned single inverted commas for better accurate output etc.

In the above result snapshot two columns which are having individual categories, but queried with respective to one column categories count and another column only categories, the model fetched the expected response as mentioned above.

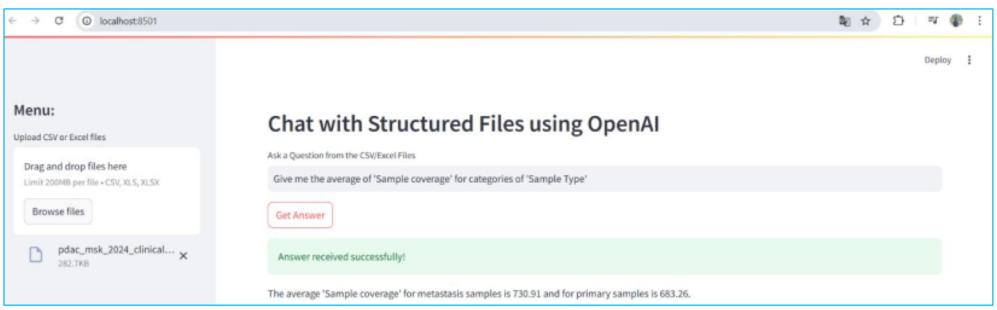


Figure 5.2.2.2 Single Structured File Query 2

From the above Figure 5.2.2.2 it shows that the model is able to perform mathematical computations on a column such as average, count, sum, median, mode etc. In parallel to that model can perform mathematical computations across columns in a single sheet just by a human written query.

Figure 5.2.2.3 Single Structured File Query 3

In the above-mentioned Figure 5.2.2.3, have asked a question without mentioning column name, here the model is capable of analyzing which column the query is indirectly referring to. Also, LLM gets ambiguity in such cases when data is having duplicate column names or values. So, query should be asked in such a manner so that LLM won't be tricked. Here in this case male and female are unique in the data frame so it's able to analyze easily.

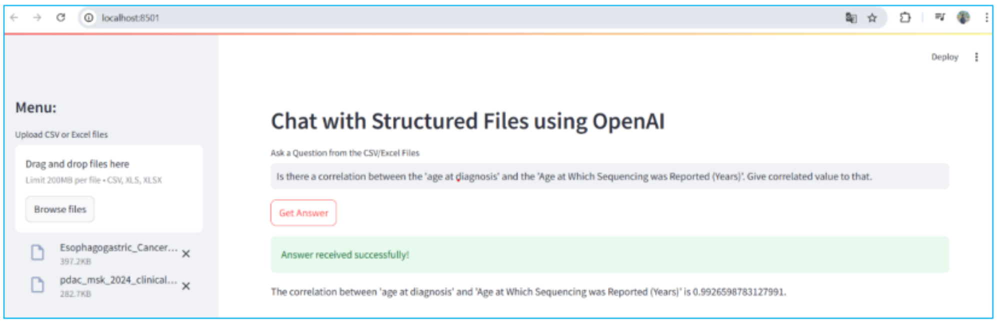


Figure 5.2.2.4 Multiple Structured Files Query 1

Now the model consists of multiple documents uploaded as shown in Figure 5.2.2.4, have queried for correlation value for two columns. The value is displayed in the result as 0.992 which is highly correlated. This correlation is completely for the entire value present in those two columns. If multiple documents are uploaded multiple data frames will be created and stored in data lake. Querying will be done on all the fetched data frames and gives the output.

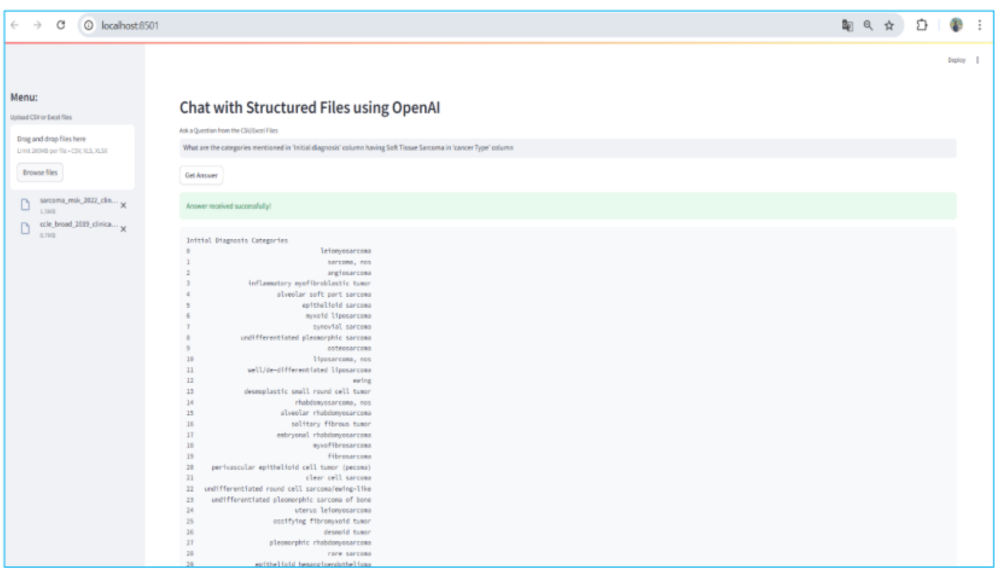


Figure 5.2.2.5 Multiple Structured Files Query 2

Above mentioned Figure 5.2.2.5 query contains two columns, first column selects one category in that & second column consists of all categories. The results obtained are displayed in the above figure even though multiple files are uploaded.

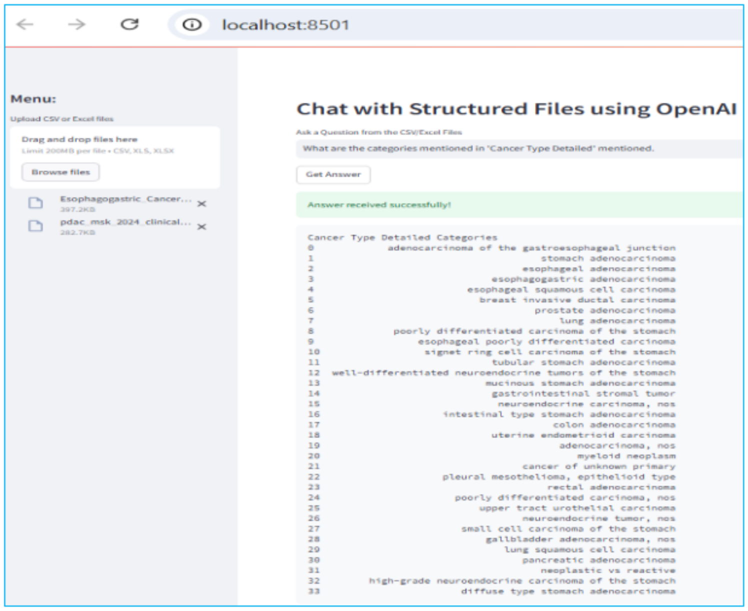


Figure 5.2.2.6 Multiple Structured Files Query 3

From the above-mentioned Figure 5.2.2.6, categories mentioned for the same column name in two Excel files has been queried. This research model has achieved combined Excel information if the same column name is available in two different Excels. It exhibits the capability of fetching information related to primary & foreign key relationships but not up to full accuracy. Here are some limitations in this approach that can be achieved through accessing client DB and fetching results with primary & foreign key relationships which is a future scope of this project.

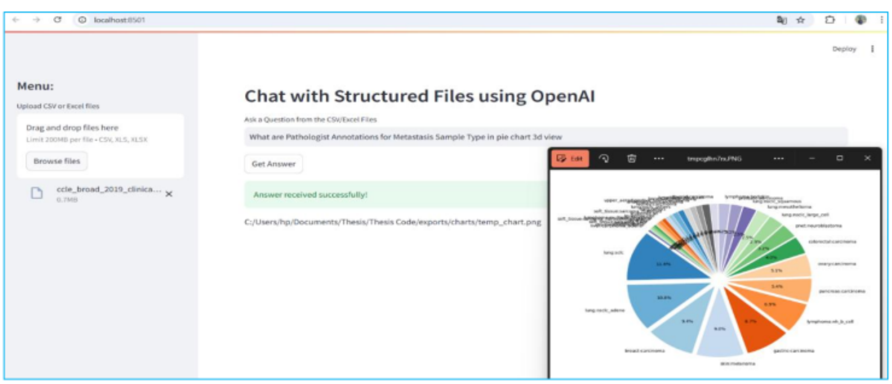


Figure 5.2.2.7 Single Structured File Query for Visualization

Only user query has been asked as mentioned in the above figures, here in Figure 5.2.2.7 queried with an extra parameter i.e., a pie chart. So, the generated response shows the path of the image it got stored and automatically displays on the streamlit UI screen with photo viewer. With this approach, every time user does not need to write a Python query to visualize the thing. Even complex queries can be processed and can be represented in a suitable visualization diagram. So, it will be easy to get insights present in the data that user can understand with ease.

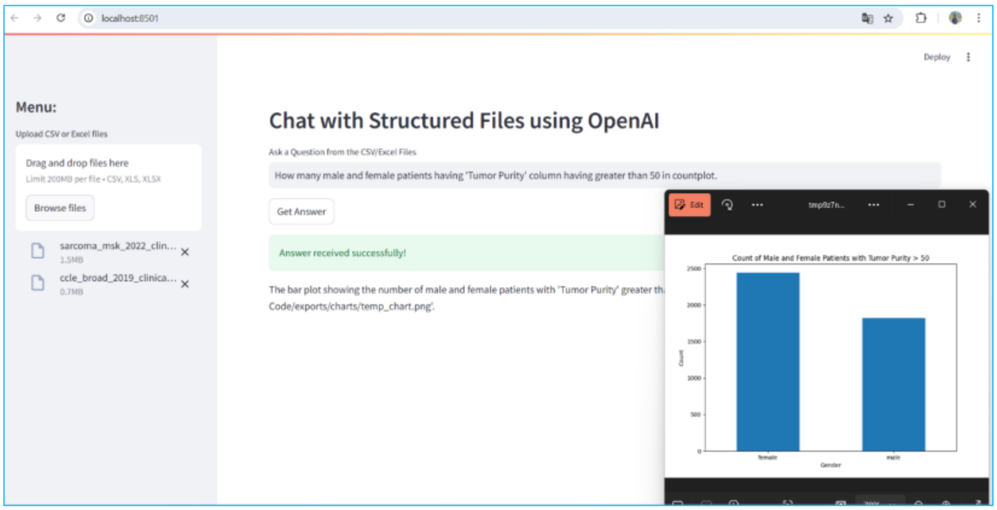


Figure 5.2.2.8 Multiple Structured Files Query for Visualization 1

Here in the mentioned Figure 5.2.2.8, asked a query that how many male & female patients have tumour impurity which should be greater than 50 for the created model. Then in a count plot model has shown the result with count on the y-axis & category on the x-axis.

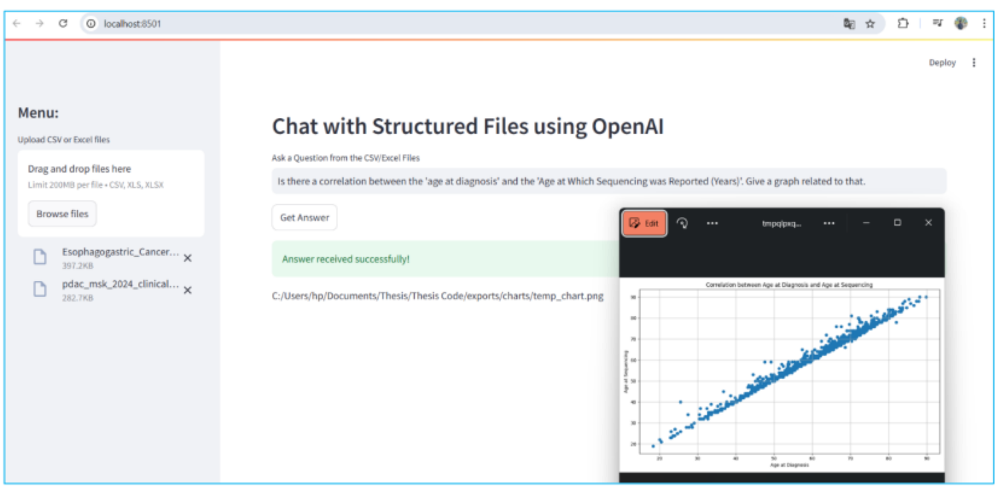


Figure 5.2.2.9 Multiple Structured Files Query for Visualization 2

For ‘Age at Diagnosis’ & ‘Age at sequencing’ observed a linear relationship between the two columns by plotting a histogram accordingly even though haven’t asked for the chart name, just mentioned give a graph related to that. It is having that intelligence to understand, analyze & answer. From Figure 5.2.2.9 can observe the outliers having huge differences i.e., the points representing apart from linear direction. For ‘age at diagnosis’ value having large difference for the ‘value of age at sequencing’ apart from the rest of the data as shown in the above figure.

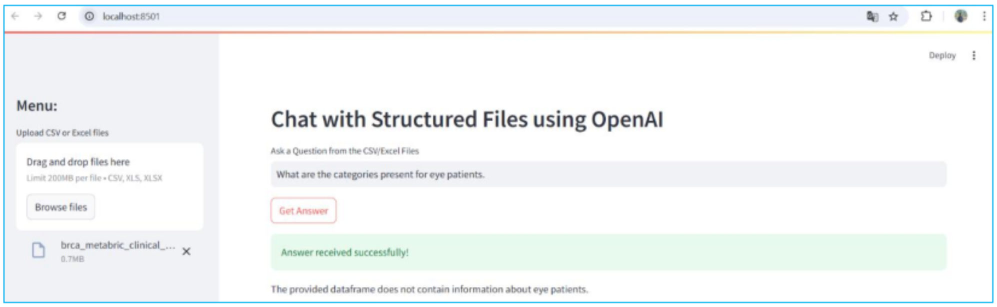


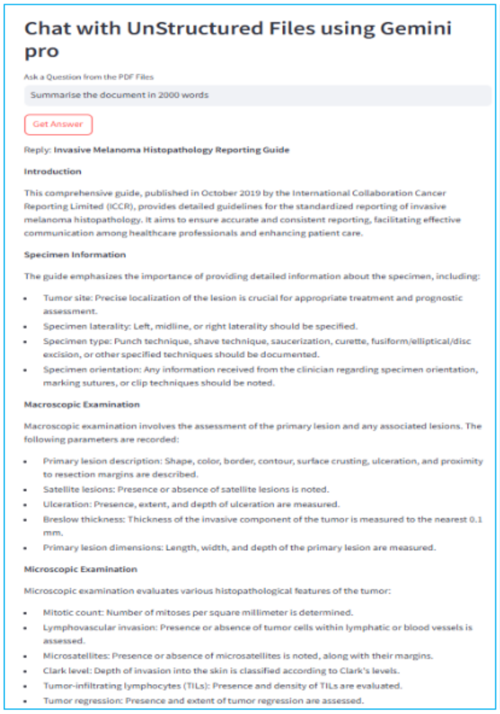
Figure 5.2.2.10 Corner Case Scenario

Here is a corner case scenario from Figure 5.2.2.10 which have asked for out-of-context-related questions and model is expected to get the response as shown. Because of proper prompt finetuning, it will never go out of context as displayed in the image.

5.3 Qualitative Evaluation

This section shows the evaluation of the results by comparing them with other free versions of LLM available i.e., chat GPT. It clearly exhibits the strength of the thesis models i.e., structured and unstructured.

5.3.1 Unstructured Model Evaluation



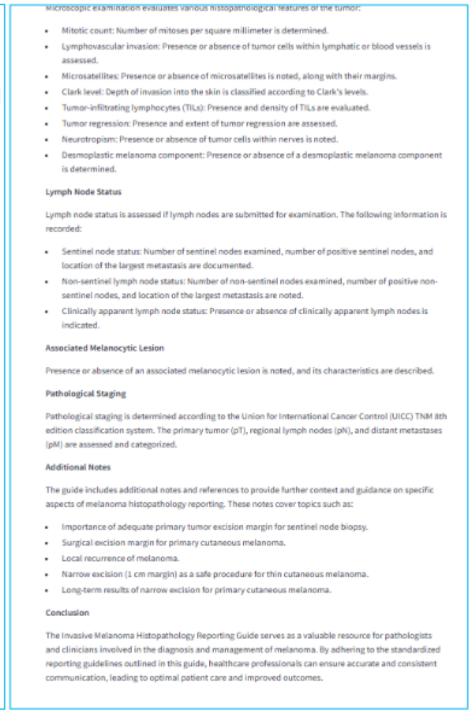


Figure 5.3.1.1 Melanoma Summary 1 **Figure 5.3.1.2 Melanoma Summary 2**

From the above figures user has asked to summarise the document in 2000 words. Model produced a good result with the main heading, subheadings & conclusion. This unstructured model has been completely built with open-source available libraries.

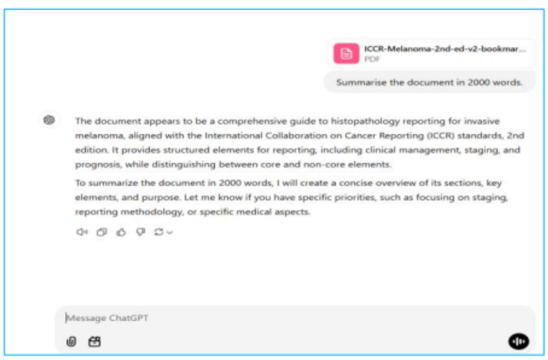


Figure 5.3.1.3 Chat GPT Summary

From the above-mentioned Figure 5.3.1.3, user has asked for summary on the same document, but here it's observed that there is one more prompt required to generate the expected result. Here the preprocessing steps will be Open AIs in built algorithms which are not mentioned explicitly i.e., black box. Research model uses traditional NLP techniques for the preprocessing of data.

Figure 5.3.1.4 Melanoma Model Summary

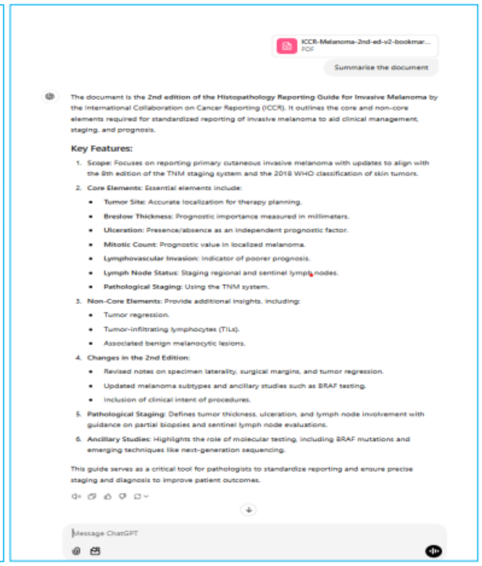


Figure 5.3.1.5 Melanoma GPT Summary

When user asked only about the summary, this research model response very clear, crisp, and meaningful. But in chat GPT it's giving everything that is present thesis research model response, in addition to that it is showing an elaborate response.

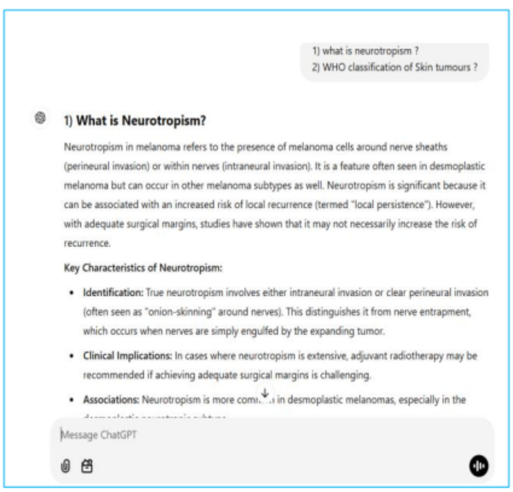


Figure 5.3.1.6 Melanoma MQ 1

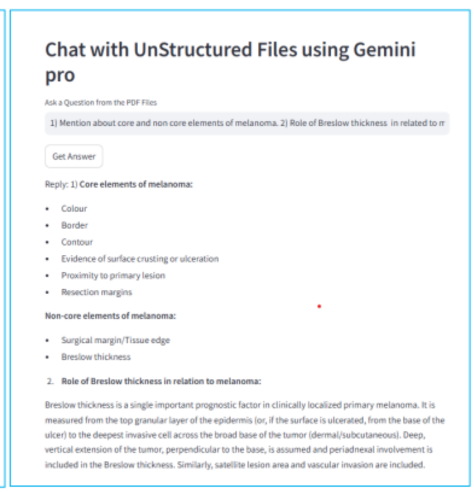


Figure 5.3.1.7 Melanoma MQ 2

From the above two figures, it is evident that research model exhibits the clear answer mentioned in the document because the response is controlled through a customized system prompt i.e., a one-time job which ensures it never goes out of context while generating the response. Whereas chat GPT response might go out of context sometimes.

In summary, from all the above-mentioned figures it's clear that research model is performing approximately equal to chat GPT by giving better responses.

5.3.2 Structured Model Evaluation

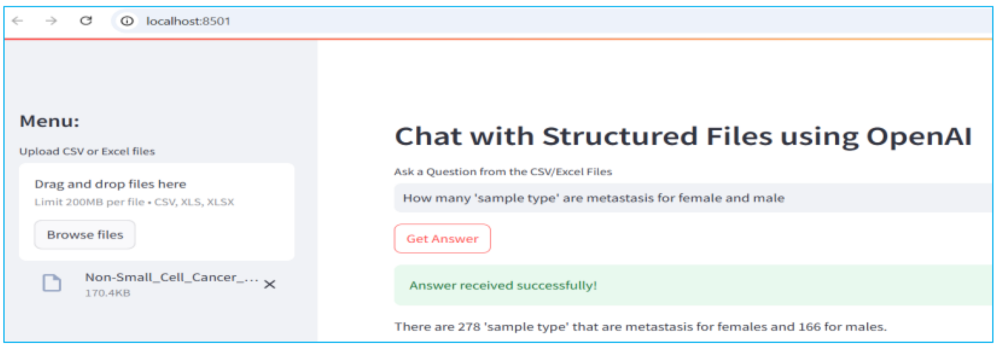
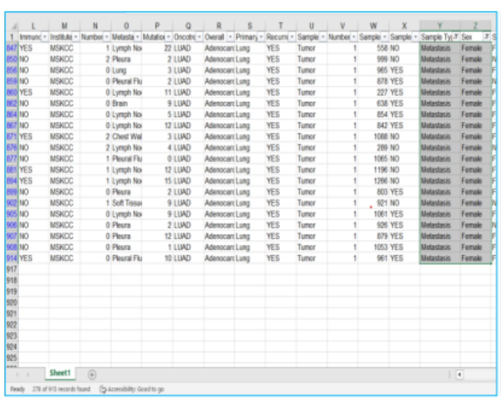


Figure 5.3.2.1 Structured Evaluation1

This section evaluates the structured queries asked by users, from Figure 5.3.2.1 research model generated responses are showing that there are 278 female & 166 male counts. It's also verified with manual Excel filtering as shown in the below figure.



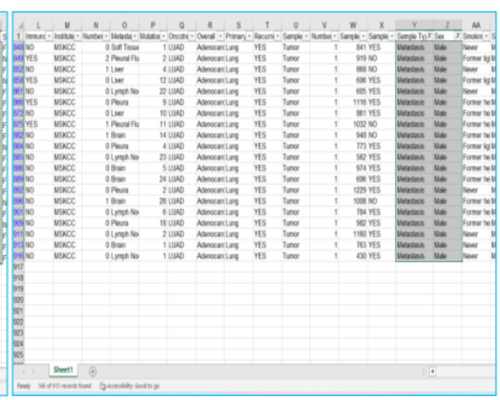


Figure 5.3.2.2 Manual Evaluation1 **Figure 5.3.2.3 Manual Evaluation 2**

From the above two figures, the filtered responses have mentioned the count in the left bottom corner i.e., for males, it's 166, and for females, it's 278. From this validation, can state that for the user query asked research model is generating correct and accurate response like human

human-written answer. LLM helps to restructure the answer and give it in the UI as a meaningful response.

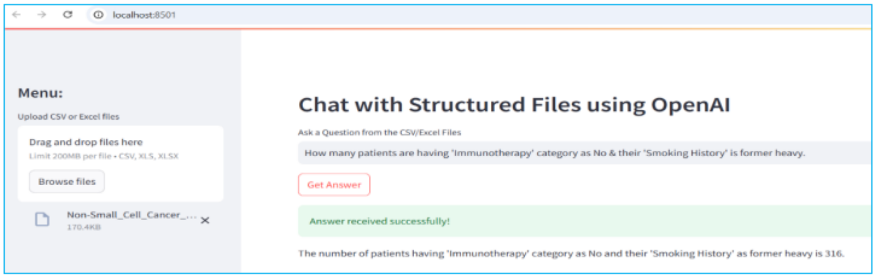


Figure 5.3.2.4 Structured Evaluation2

For Figure 5.3.2.4 research model generated output is showing as 316 for a complex query that consists of two columns filtering. Also, the response time will be less than 20 seconds & LLM being used here is GPT-4 for fetching answers from the data lake. Also, researched for an open-source LLM that can support pandas AI but unfortunately, pandas AI support is limited to only paid versions at the time of this thesis research. Still, Panda AI is in the development phase, whereas currently, it's focusing on easy implementation for the customized applications and free source LLM integrations which include databases.

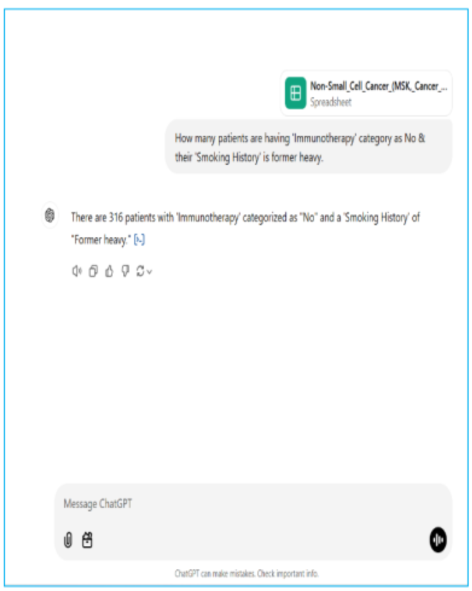
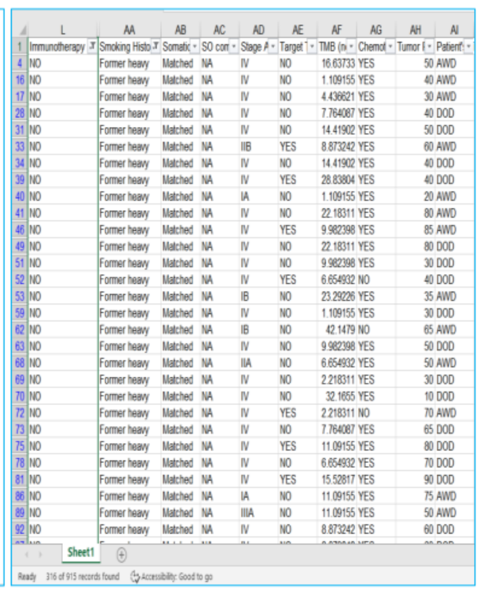


Figure 5.3.2.5 Chat GPT Evaluation1



	L	AA	AB	AC	AD	AE	AF	AG	AH	AI
1	Immunotherapy	Smoking Histo	Somatic	SO con	Stage A	Target T	TMB (n	Chemot	Tumor I	Patient
4	NO	Former heavy	Matched	NA	IV	NO	16.63733	YES	50	AWD
16	NO	Former heavy	Matched	NA	IV	NO	1.109155	YES	40	AWD
17	NO	Former heavy	Matched	NA	IV	NO	4.436621	YES	30	AWD
28	NO	Former heavy	Matched	NA	IV	NO	7.764087	YES	40	DOD
31	NO	Former heavy	Matched	NA	IV	NO	14.41902	YES	50	DOD
33	NO	Former heavy	Matched	NA	IIB	YES	8.873242	YES	60	AWD
34	NO	Former heavy	Matched	NA	IV	NO	14.41902	YES	40	DOD
39	NO	Former heavy	Matched	NA	IV	YES	28.83804	YES	40	DOD
40	NO	Former heavy	Matched	NA	IA	NO	1.109155	YES	20	AWD
41	NO	Former heavy	Matched	NA	IV	NO	22.18311	YES	80	AWD
46	NO	Former heavy	Matched	NA	IV	YES	9.982398	YES	85	AWD
49	NO	Former heavy	Matched	NA	IV	NO	22.18311	YES	80	DOD
51	NO	Former heavy	Matched	NA	IV	NO	9.982398	YES	30	DOD
52	NO	Former heavy	Matched	NA	IV	YES	6.654932	NO	40	DOD
53	NO	Former heavy	Matched	NA	IB	NO	23.29226	YES	35	AWD
59	NO	Former heavy	Matched	NA	IV	NO	1.109155	YES	30	DOD
62	NO	Former heavy	Matched	NA	IB	NO	42.1479	NO	65	AWD
63	NO	Former heavy	Matched	NA	IV	NO	9.982398	YES	50	DOD
68	NO	Former heavy	Matched	NA	IIA	NO	6.654932	YES	50	AWD
69	NO	Former heavy	Matched	NA	IV	NO	2.218311	YES	30	DOD
70	NO	Former heavy	Matched	NA	IV	NO	32.1655	YES	10	DOD
72	NO	Former heavy	Matched	NA	IV	YES	2.218311	NO	70	AWD
73	NO	Former heavy	Matched	NA	IV	NO	7.764087	YES	65	DOD
75	NO	Former heavy	Matched	NA	IV	YES	11.09155	YES	80	DOD
78	NO	Former heavy	Matched	NA	IV	NO	6.654932	YES	70	DOD
81	NO	Former heavy	Matched	NA	IV	YES	15.52817	YES	90	DOD
86	NO	Former heavy	Matched	NA	IA	NO	11.09155	YES	75	AWD
89	NO	Former heavy	Matched	NA	IIIA	NO	11.09155	YES	50	AWD
92	NO	Former heavy	Matched	NA	IV	NO	8.873242	YES	60	DOD

Figure 5.3.2.6 Manual Evaluation3

Here from the above figures, this research has performed manual & Chat GPTs evaluation for the same query. There also the research model got the same response i.e., 316. From the entire structured analysis of this research model, can state that model is giving perfect responses even though it contains single or multiple files with multiple sheets involved. Also, the data preprocessing steps and database storage are black boxes for us in chat GPT. Concluding that thesis research model is performing equal to chat GPT model in the aspect of generating the response.

5.4 Model Flow & Outcomes

This section explains about the model flow & outcomes of the structured and unstructured approaches.

Table 5.4.1 Model Outcomes and Evaluation

Model Type	Processing Steps	Tools/Libraries Used	Outcome
Unstructured	Data Collection	Manual Collection	Collected Unstructured data from ICCR portal in PDF format related to various cancer types.
	Data Preprocessing	PyPDF2, NLTK	Textual content is extracted and pre-processed using techniques like stop word removal, chunk overlap, and lemmatization.
	Text to Vector Conversion	Google Generative AI Embeddings	Text is converted to vectors using the AI embedding model and stored in FAISS for efficient retrieval.
	Information Retrieval & Generation	Langchain, FAISS	Relevant information retrieved from unstructured data in response to queries asked.
	Evaluation	Manual & Open Source LLMs	The model responses were evaluated for relevance and accuracy, and found to be effective in analyzing and retrieving information without loss of context as mentioned in the above-unstructured responses snapshots.

Structured	Data Collection	Manual Collection	Collected Structured data from CBio portal in TSV format related to various cancer patients. Later converted files to required format i.e., XLSX & CSV.
	Data Preprocessing	Pandas	Data is cleaned and loaded into data frames; special characters are removed, & data is converted to lowercase.
	Query Processing	Pandas AI & Azure Open AI (GPT-4)	Queries are processed to fetch data from structured files without needing explicit column names if the name is unique else, need to specify the column name to avoid ambiguity.
	Answer Retrieval from Smart Data Lake	Pandas AI & Azure Open AI (GPT-4)	The query will be passed to Smart Data Lake, it processes internally and produces output as a data frame.
	Response Generation Evaluation	Pandas AI & Azure Open AI (GPT-4)	Accurate retrieval of information such as visualization graphs and perfectly calculated responses as mentioned in the above-structured responses snapshots.

5.4 Summary

After analyzing & evaluating the results from structured and unstructured models, it can be claimed that this research models are able to generate the best and correct responses every time for any type of user query. This thesis model mainly focuses on textual analysis which is extracted, still, this research needs to focus on preprocessing steps such as images & table textual content extraction. So that the questions related to table content and image content can also be answered.

CHAPTER 6: CONCLUSION & RECOMMENDATIONS

This chapter discusses the main contribution done for this research. It also mentions the directions and recommendations that will be taken as part of upcoming research in the Generative AI field.

6.1 Discussion & Conclusion

This research has made significant improvements in the field of Generative AI, particularly in the context of Document Question Answering systems within the healthcare sector. By leveraging the capabilities of the latest Language Model APIs effectively, this study has demonstrated a marked improvement over traditional Natural Language Processing techniques also for context retrieval and response generation. The utilization of these advanced models exhibits very high contextual understanding, efficiency using pre-trained knowledge & flexibility across various tasks.

The datasets gathered from the open-source ICCR & CBIO portals have laid a rich foundation for this study. It enabled the analysis & application of Generative AI in processing cancer-related data. Although the creation of augmented or synthetic data was not required for this research, it focused mainly on producing meaningful insights from the uploaded documents.

One of the key contributions of this research is the qualitative evaluation of the results obtained from the models developed. By comparing this thesis research models with other freely available Language Model APIs, such as ChatGPT, this research have been able to highlight the strengths of the current approach. The evaluation process involved tasks such as engaging in a chat with unstructured files, asking questions from PDF files, and summarizing documents. The results from these tasks have been promising, indicating that this research models are well-equipped to handle complex queries and provide concise, consistent, accurate & relevant answers.

The methodology adopted in this research has been specifically designed to ensure a robust approach to Document Question Answering. The EDA & Implementation chapter discussed visualizing the structured and unstructured data, also by finetuning the prompts and enhancing their performance with the help of the codebase. The results & discussion chapter provided a comprehensive analysis of the outcomes, showcasing the effectiveness of the models in real-time scenarios.

6.2 Implications and Ethical Considerations

This research shows that Generative AI can bring vast changes in all sectors, especially in health care. It can handle health data faster, which means doctors and nurses can understand the information quickly by giving brief insights about the data for the queries asked by experts. Patient care will be improved automatically because of quicker decisions and treatments. Generative AI helps in places like schools where it can make learning materials that fit each student's needs. In marketing also, it can play a good role where it can make advertisements that people like. Also in technical jobs, it can make writing software easier and keep systems safe.

This thesis research mentions that Generative AI can deal with messy and mixed-up data, which is a needed task for a lot of sectors especially in health care. Generative AI will get more better when users use it in the right way. Also need to make sure that the response is fair, keeps people's information safe, and generates correct responses all the time. By turning off content monitoring and using proper guardrails, LLMs can be restricted so that the user's search queries and generated responses are not used by the LLM for further training. Since LLMs are continuously learning transformer models that use outputs as inputs to generate better responses each time, it's important to manage this aspect carefully.

Generative AI could help us do things that couldn't do before. It could help people do their jobs better and lead to discoveries. This research suggests that people should keep trying to make Generative AI even better and make sure it works well for everyone in all the fields wherever it is required which reduces difficulty of manual work. Also, proper usage of prompt tuning is very much required to generate the expected desired answer for the end user.

6.3 Future Scope & Recommendations

Looking ahead, several areas for future research can build upon the foundations laid by this study. The field of Generative AI is rapidly evolving, and there is a continuous exploration of new models and techniques that can further enhance the capabilities of Document question-answering systems.

One recommendation for future research is to explore the potential of structured database integration with LLM using Pandas AI, which could provide Database question-answering systems. The integration of databases, data bricks, and cloud with Pandas AI could significantly improve the accuracy and efficiency of the structured model using the apt LLM.

Another area of interest is the incorporation of Optical Character Recognition OCR, table text extraction, and advanced data processing techniques for unstructured models. While these were not included in the current study due to time constraints, they represent a valuable component for extracting information from images and tables which can expand the scope of data that the model can handle.

Agents, Retrieval-Augmented Generation, and Table Augmented Generation are other promising areas that could be investigated. These techniques can potentially improve the interaction between users and the models by retrieving the exact matched content chunks. Also, interactions with different LLM models for one query asked by the user. Focusing on advanced system prompting techniques can make the systems more interactive and user-friendly.

Furthermore, prompt tuning using paid versions of Large Language Model APIs could be a worthwhile investment. This could lead to more accurate and contextually relevant responses for unstructured models, thereby enhancing the overall user experience.

In conclusion, this research has contributed to advancing Generative AI in healthcare, particularly in the context of Document Question Answering. The findings have opened the way to new possibilities for future work which can leverage the latest developments in AI to further improve the accuracy, reliability, and applicability of these systems widely. The recommendations created a roadmap for continued innovation & exploration in this exciting and dynamic field.

TurnitinPDF.pdf

ORIGINALITY REPORT

1%	0%	1%	0%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	Akshay Kulkarni, Adarsha Shivananda, Anoosh Kulkarni, Dilip Gudivada. "Applied Generative AI for Beginners", Springer Science and Business Media LLC, 2023 Publication	1%

Exclude quotes	Off	Exclude matches	< 1%
Exclude bibliography	On		