Extraction & Navigation from Semi-Structured Data using Large Language Models

Mehek Sawhney

Mcomp Computer Science with Artificial Intelligence
May 2024

Extraction & Navigation from Semi-Structured Data Using Large Language Models

Submitted by: Mehek Sawhney

Declaration

This Dissertation is submitted to the University of Bath in accordance with the requirements of the degree of Bachelor of Science in the Department of Computer Science. No portion of the work in this Dissertation has been submitted in support of an application for any other degree or qualification of this or any other university or institution of learning. Except where specifically acknowledged, it is the work of the author.

Abstract

This paper dives into the world of data analytics, digging into the task of extraction and retrieving information from data. The aim of this framework is to employ a Question-Answering (QA) system that leverages the potential of a Retrieval Augmented Generation (RAG) to reduce the chances of hallucinations produced by the large language model (LLM), llama-2. Incorporating a set of PDF documents as the primary data source, the system effectively shows how data can be integrated and processed. To test the effectiveness of the system, various evaluation metrics were used. The results obtained demonstrated the capabilities of the large language model, shedding light on its ability to generate accurate and relevant answers by effectively utilising external data sources, thereby enhancing the reliability and utility of the QA system in practical applications.

Keywords: Question-Answering System, Retrieval Augmented Generation, hallucinations, Llama-2

Table of Contents

Acknowledgments

Firstly, I would like to thank my supervisor, Julian Padget for his support and guidance throughout this project. His expertise and patience have been invaluable. I would also like to thank my friends and family for their constant love and support. Their encouragement has provided me with the motivation to complete this research. Lastly, I would like to acknowledge and thank the participants who took part in this study providing their insights and experiences.

1 Introduction

With the rise of Internet applications and major social networks, the nature of data has started shifting (Lin et al. ,2018). There has been an excessive increase in the volume of semi-structured data, which is data that is not neatly organised and cannot be captured or formatted in traditional ways (Grishman, 2005). This has posed a significant challenge in the task of finding and extracting relevant information from this data. In a study, (Kandel et al., 2012) reported interviews with 35 data analysts, in which they indicated the processing of semi-structured data as one of the major challenges for analysts.

Information Extraction (IE) and Information Retrieval (IR) offer a promising solution to the difficulties posed by semi-structured data. Information Extraction (IE) is concerned with the identification and extraction of structured information, i.e., particular facts from an unstructured or partially structured input. In addition to aiding the discovery of relevant papers, IE also holds a promise to improve the transparency of LLM decisions, serving as an intermediary process in automated literature screening (Tang et al., 2024). On the other hand, Information Retrieval deals with retrieving the relevant set of documents in response to a user's query (Gaizauskas & Wilks, 1998).

The capability of these systems goes beyond traditional structured data sources to encompass technical texts such as scientific journals, legal documents, and hospital reports. Such systems are particularly valuable to entities like hospitals and medical researchers who require retrospective analysis of natural language reports (Grishman, 2005). Not only would this allow medical researchers to extract the sources of clinical information (Landolsi et al., 2022) but also allow lawyers to skim through large documents to find relevant case studies.

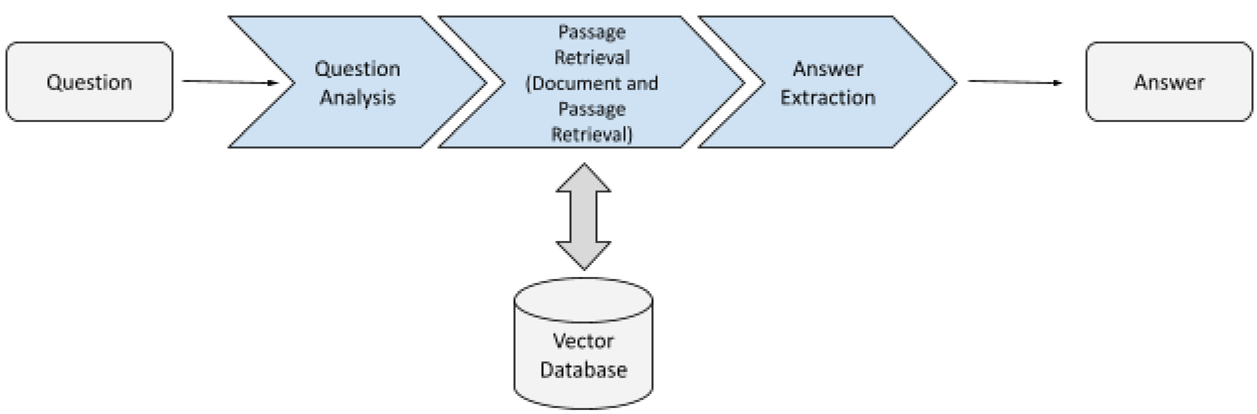


Figure 1.1: Diagram for traditional question-answering systems

Question-answering systems, employ techniques of both IR and IE, to analyse the question, retrieve relevant documents and passages from vector databases, and extract an answer to generate accurate and relevant responses. These aim to satisfy users looking to understand their content, or looking to find an answer to their specific question (Bouziane et al., 2015).

Initially, information extraction systems relied on rule-based to machine-learning-based approaches, grappling with the challenges of scalability and adaptability (Scholarly, 2023). The traditional information retrieval mechanisms mainly relied on matching terms between queries and documents which faced difficulties of polysemy, synonymy, and lexical gaps (Hambarde & Proença, 2023). With the recent advances in Large Language Models (LLM) such as Llama, Falcon, and Gpt-4, the landscape of text processing has dramatically changed. These models have attracted significant attention for their remarkable abilities in comprehending and generating natural language (Cai et al., 2024). They aid in lessening the workload for human experts outperforming traditional methods in both precision and effectiveness.

However, their widespread adoption is hindered by the challenge of "hallucinations" - the generation of erroneous or misleading information (Yehuda et al., 2024). These hallucinations can significantly impact the reliability and accuracy of the extracted data, posing a critical obstacle to the practical application of LLMs in real-world scenarios (Graph, 2024).

To address this challenge, I aim:

- To experiment with the capabilities of the large language model, llama-2, in efficiently extracting meaningful information from PDF documents.
- To integrate llama-2 within a Question-Answering chain that harnesses the potential of a Retrieval Augmented System (RAG), integrating llama-2's generative capabilities with specialised data retrieval methods, creating a platform capable of delivering nuanced and precise responses (Selvaraj, 2024).
- To assess the performance of llama-2 in various tasks by employing statistical-based metrics to measure accuracy and reduce potential instances of misinformation or erroneous data extraction.

The ultimate goal of this paper is to develop a workflow, using llama-2, which would receive a set of documents and generate a service to process queries over those documents, returning hits as textual fragments that embed the search text. The subsequent four sections include a literature review, providing background information on existing research relevant to the study, followed by a methodology section, showing an overview of the proposed framework. Section 3 comprises the experimental details and the results obtained. Lastly, Section 4 concludes the dissertation by summarising the findings and discussing future work.

2 Related Work

Recognising the potential of large language models and their effect on advancing research in efficient extraction systems, has sparked an interest among researchers and scholars alike. From traditional techniques of extraction to the evolution of state-of-the-art methods, this section provides an overview of the existing research efforts in this area, highlighting the key contributions and identifying factors that motivate their study.

2.1 Evolution of Information Retrieval (IR)

The growth of information affects the flood of information, which leads to the development of information retrieval (IR) systems to allow data to be easily accessible and useful for the user (Kambau & Hasibuan, 2018). Initial systems primarily operated through text-based searchers, meaning they relied only on the keywords provided by users to find and retrieve relevant text or documents (Kambau & Hasibuan, 2018). There are two basic metrics used for evaluating the effectiveness of information retrieval, i.e., precision and recall (James & Kannan, 2017).

Conventional IR models, such as the Boolean Model, are based on mathematical logic revolving around boolean operators, AND, OR, and NOT, to combine the set of documents into new sets based on the user's query (Roshdi & Roohparvar, 2015). While this is a fast and efficient search mechanism, this approach not only faces difficulties in handling synonyms, polysemy, and context (Hambarde & Proença, 2023) but also does not rank the results (Roshdi & Roohparvar, 2015). When a query is made, the model simply returns all the documents that meet the Boolean criteria as equally relevant, treating each document as relevant or irrelevant, which does not allow users to see which document may be more relevant than others. This led to a shift in exploring other models i.e. probabilistic, vector, or neural-based approaches.

Each model offered a different way of representing terms and their importance and was designed to rank documents in a manner that prioritises relevance to the user's query. The vector-based model is easy to understand and good at matching documents with queries using vector representations and cosine similarities. It introduces a weighting scheme called tf-idf (term-frequency - inverse document frequency) but struggles with large datasets and misses nuances by ignoring the order of words (Kambau & Hasibuan, 2018). The probabilistic model adapts well to user feedback and is statistically sound, but it reduces the text to just term presence, losing some detail (Roshdi & Roohparvar, 2015). Lastly, the inference network model is great for complex queries and using various types of data, but it's also more complicated and requires more computing power.

In recent years, researchers have been exploring the use of discrete retrieval methods through term-reweighting (Ernandes et al., 2006; Cho et al., 2023) to adjust the significance of words and dense retrieval methods (Karpukhin et al., 2020; Qu et al., 2021; Silva & Barbosa,2024) through using word embeddings to capture a deeper meaning of text, thereby improving the results. At the end, the best approach for IR often involves combining multiple methods, leveraging their strengths, and creating a hybrid approach.

2.2 Evolution of Information Extraction (IE)

Information Extraction (IE) is a crucial component in information retrieval which has evolved significantly over the years and includes essential sub-tasks like named entity recognition (NER) and event extraction (EE) (Huang & Huang, 2024).

Early information extraction (IE) systems were primarily designed for single-document processing and struggled with scaling to handle the vast amounts of information found online (Hobbs and Riloff, 2010). Initially, these systems revolved around the use of Knowledge engineering (KE), where human experts manually crafted rules for extracting information, a time-consuming intensive process requiring approximately 1500 person-hours of effort for systems like UMass MUC-4 (Riloff et al., 1993). Due to the manual effort and the limitations in handling complex data expressions, this prompted a shift away from KE-based approaches to more trainable systems, i.e. supervised, unsupervised, and sequential learning approaches (Hobbs and Riloff, 2010).

Several IE systems relied on heuristic-based approaches that utilise predefined sets of rules derived from observed data patterns, as evidenced by research conducted by (Hong et al., 2009; Kanaoka et al., 2014; Shigarov et al., 2018). While effective within specific datasets, these rules struggle to generalise effectively over other datasets. Alternatively, machine learning models such as Naive Bayes (Lubis et al., 2022), SVMs (Zhang et al., 1970), and ontology-based methods (Gayathri et al., 2021) offer a scalable solution. These models, including sequential learning models, like Hidden Markov Models (Freitag and McCallum 2000) treat IE as a classification problem and require significant training and careful consideration of issues including feature selection and weight adjustment (Guo et al., 2023) to optimise performance, highlighting the need for ongoing research to enhance the capabilities of IE systems.

2.3 Large Language Models

Over recent years, large language models have become increasingly efficient and are used to generate human-like text, answer questions, and complete other language-related tasks with high accuracy (Kasneci et al., 2023). (Guo et al., 2023) experimental findings show that the methodologies based on large language models (LLM) surpass traditional methods, enabling the effective and efficient extraction of information from the scientific literature. These models are deployed in various commercial applications, including the financial, medical, legal, and educational sectors.

For instance, the research highlighted by (Li et al., 2023) converted PDF files into machine-readable plain text files and used a combined technique of text mining and prompt engineering. Even though the framework resulted in high accuracy, it not only relied on GPT-4's API which incurs costs, especially for individual researchers, but also its efficacy was established through its application to certain types of financial reports (ACFRs and ESG reports). This raises the question of how well it might generalise to other types of financial documents or datasets with different characteristics. Furthermore, (Wiest et al., 2023) successfully showed that clinical texts can be extracted with minimal hardware, enhancing access and employing an open-source approach, using LLAMA-2, to allow for ongoing research. Nonetheless, its focus on a single dataset could affect the broader applicability of its applications.

2.3.1 Recent Large Language Models

The release of chatbot, ChatGPT in late 2022, has rapidly spread this technology to a wide range of users (Rillig et al., 2023), with researchers exploring the possibilities of GPT-3.5 and GPT-4 to make meaningful relationships with data. For instance, (Sandmann et al., 2024) have explored the possibility of using these models for clinical decisions and support tasks. GPT-4 significantly outperformed GPT-3.5 in diagnostic, examination, and treatment tasks, despite having some limitations in treatment suggestions. Even though these models were not trained on such datasets, they perform considerably better than naive Google searches. However, it is important to remember, these models don't outweigh the individual interpretations as they can vary among different physicians.

These models are trained on a much larger dataset and have demonstrated state-of-the-art performance on a wide range of natural-language tasks ranging from translation to question answering. (Zhang et al., 2023) evaluated the performance of ChatGPT across different combinations of information retrieval tasks and artefact types. In a zero-shot setting (without prior specific training on these tasks), ChatGPT showed a good ability to find the relevant information (high recall) but was less effective at finding precisely needed details (low precision). These values showed the effectiveness of ChatGPT in information retrieval tasks where recall is more preferred than precision (Zhang et al., 2023).

However, the evolution of information retrieval comes with its challenges. There is an ethical dilemma associated with false information or even harmful content. The capabilities of these technologies to generate realistic text blur the lines between genuine and fabricated information (Huang & Huang, 2024). This makes it difficult for people to distinguish between true and false content. This is where implementing a retrieval-augmented system could potentially resolve this issue.

In terms of IE, (Bakker et al., 2024) employed a timeline extraction using ChatGPT and was extremely successful in extracting the event phrases and their classes, achieving an accuracy of 94% in the test set. However, ChatGPT still struggles with event extraction (EE) tasks due to the intricate instructions required and the system's current limitations in adaptability (Gao et al., 2023). Even through continuous refinement of the prompt, ChatGPT doesn't lead to stable performance improvement impacting the user experience (Gao et al., 2023). Comparing the overall effectiveness of IE, considering both its performance and explainability, (Xu et al., 2023) discovered that while ChatGPT generally underperformed compared to BERT-based models in the standard IE setting, it demonstrated excellent performance in the OpenIE setting.

These limitations coupled with the associated API costs, promote a pivot toward the exploration and further enhancement of newer language models alternatives such as llama-2, Falcon, and paLM 2, etc.

2.3.2 Emerging Large Language Models

While ChatGPT has outperformed other LLMs in certain tasks, other models show promising potential. Along this line, several researchers have performed a comparative analysis with other LLMs. (Dagdelen et al., 2024) showed an approach comparing both LLAMA-2 and GPT for extracting chemistry information using joint named entity recognition (NER) and relation extraction (RE), with JSON files containing complex material. This study revealed that while GPT 3.5 outperformed LLAMA-2 in terms of named entity extraction and relation extraction, LLAMA-2 showed a better performance in extracting certain types of information, where detailed and specific information extraction is critical. Similarly, when comparing the performance of models across types of prompting, such as 1-shot, 3-shot, and 5-shot (Cabra et al., 2024) showed that the LLAMA-70b parameter model outperformed the GPT-4, Falcon, and PALM, reflecting that size of a model affects the performance and complexity.

Despite the significant advances, large language models still grapple with numerical limitations and, more importantly, the challenge of interpretability, making it harder to understand the rationale behind their predictions (Kasneci et al., 2023). Additionally, these models have faced challenges surrounding the absence or state of memory, rendering them information from previous prompts. They have also faced issues regarding their stochastic nature leading to various responses to the same input and their reliance solely on the data they were trained on (Minaee et al., 2024). There's no denying their transformative potential, but it's equally crucial to handle these challenges with care and diligence.

2.4 Question Answering Systems

LLMs can provide general insights but lack access to private or sensitive information unless they're explicitly trained on such data (Jeong, 2023). Question Answering is a technique that provides LLMs with additional information to generate results (Jeong, 2023). These traditional systems are divided into three main parts: question classification, information retrieval, and answer extraction (Allam & Haggag, 2012). These components use techniques of Natural Language Processing (NLP) and Information Extraction (IE) (Diefenbach et al., 2017). The origin of QA systems began in 1961 with the introduction of BASEBALL, which answered simple questions relating to American League Baseball games (Caballero, 2022). However, at the time the knowledge was limited and domain-restricted (Mishra & Jain, 2015). Over the years there has been an advancement in the QA systems, with researchers employing techniques that acknowledge the importance of document structure for meaningful data representation. As noted in the study, (Saad-Falcon et al., 2023) discovered that using documents as plain text (machine-readable text files) neglects the essential structure needed to answer questions. This approach removes elements such as headers, footers, tables, and even bullet points. Therefore, this study converted their data into an HTML-like tree structure to be used in question answering. Even though this method surpassed traditional retrieval-based approaches, this approach makes it difficult to manage and query PDF documents as they often contain diverse content and require sophisticated parsing techniques.

Current research is exploring open-domain question-answering systems rather than closed-domain QA systems. The need for these systems arises due to their ability to handle any type of question asked, mimicking how humans seek information in real life. There are only a few research papers that have implemented their own open-domain QA systems. For instance, (Quarteroni & Manandhar, 2009) built "YourQA", which includes a chat-based dialogue interface to enable it to handle follow-up clarification questions easily. It relies on a web search engine to be able to respond to a wide range of questions, catering to the unpredictable nature of queries. However, the application's response time is slow due to its dependency on network speeds for real-time document retrieval (Quarteroni & Manandhar, 2009). Taking a different approach, (Cabrio et al., 2012) built "QAKiS", which relies on relational pattern matching patterns for open-domain QA, as it would provide more accurate results in comparison to systems that only use single words. This approach converted natural language queries into database queries but it only works well where the specific rules and formats of the data are already known and defined.

With open-domain question answering, the architecture falls into three categories: text-based QA systems, which rely on unstructured data, Knowledge-based QA systems which rely on structured data and lastly hybrid-based systems which can read both types of data (Caballero, 2022). While, textual-based QA systems are more scalable as they rely on widely available and easily accessible unstructured data sources to find answers (Zhu et al., 2021), the latest and best results in open-domain question-answering systems are being achieved by hybrid models, such as Retrieval Augmented Generation (RAG) (Caballero, 2022).

2.5 Integration of RAG

Fine-tuning large language models produces reasonably good results, however, it can be extremely expensive due to the infrastructure costs and the chance of producing hallucinations especially if the input is out of distribution (Béchard & Ayala, 2024). The Retrieval Augmented Generation is a common method of reducing the possibility of false information. This is because, before text generation, relevant information from data sources is retrieved which ensures that the generated text is grounded with factual data, reducing any potential misleading information (Béchard & Ayala, 2024). The RAG system can provide sources for its answers by returning the page number or the sentence source from the document referenced, ensuring the information is traceable and reliable (Alan et al., 2024).

Many researchers have employed the use of RAG in enhancing the results but they differ in their approach and maintenance needs. One of the studies, (Béchard & Ayala, 2024) focuses on structured data, employing a dense retrieval system which reduced the percentage of hallucinations but increased complexity and reduced response time. In comparison, (Fatehkia et al., 2024), used a tree-like structure to represent hierarchical information within an organisation, making the generated responses more contextually aware. However, this method requires consistent and accurate updates to reflect the real-time status of the organisational hierarchies and relationships. Utilising open-source implementations and the LLM architecture, (Jeong, 2023) employs the RAG method, incorporating techniques of chunking, embeddings, and using vector databases. Consequently, depending on the large language model used, significant resources and time are required due to its complexity and size. Additionally, when using open-source implementations, some functional aspects may be incomplete.

The RAG method may be advantageous but several limitations could affect the accuracy of results (Setty et al., 2024). To improve this, the RAG method requires an appropriate chunk size which can be obtained through testing and experimentation. It is also important to ensure the queries contain enough information to guide the retrieval process, as incomplete queries can lead the algorithm to incorrect document sections, returning an irrelevant chunk (Setty et al., 2024).

Each variety of content within PDF files requires a tailored approach for effective data extraction and analysis, underscoring the need for specialised techniques to handle the unique characteristics of each file type. The increase in the amount of data online has resulted in the crucial need for robust extraction systems, particularly in handling the challenges posed by semi-structured data. Choosing the best approach for extraction depends on the characteristics of the data, the complexity of the problem, and the specific requirements of the task at hand.

3 Methodology

Although Information Extraction (IE) and Information Retrieval (IR) are complementary, it is crucial to distinguish between the respective concepts. Nonetheless, as stated by (Gaizauskas & Wilks, 1998) "the use of these methodologies in combination has the potential to create powerful tools in text processing".

The key specific requirements to ensure that the developed system meets the intended goals of efficiency and usability are:

- The system must provide a user-friendly interface that allows users to submit their respective queries and documents.
- The results of the user's queries must be displayed in an intuitive and accessible manner.
- The application must deliver real-time responses to the user's queries.
- The application must be able to process mixed documents, containing both text and images, totaling up to 100 documents within 5 minutes, without causing kernel crashes or exceeding a 30% increase in response time under varying loads.
- The system must effectively manage to process different types of PDFs, including those that are tabular and image-based.
- Users need to validate the effectiveness of the large language model in retrieving the relevant documents, and page numbers and determining the response.

These requirements are established throughout the development process, which is dynamic and responsive to challenges and innovations. This section dives deep into the question-answering system process, leveraging the potential of a retrieval-augmented generator.

3.1 Datasets

In this study, the core type of documents utilised were PDF files which were carefully anonymised to remove any identifiable information. These documents offer a diverse range of content to ensure comprehensive testing and were used in their raw form to avoid any loss of information (Saad-Falcon et al., 2023).

A unique set of documents, contributed by each participant during user testing, was used to allow for a better generalisation of data. The participant-contributed dataset consisted of X number of PDF documents, including articles, and other materials commonly encountered in an academic or professional setting such as lecture notes, and assignments. However, if each participant only used their own set of documents, their experience may depend on the complexity or the nature of the PDF files.

To correct the potential variability in this approach, this study also included a common set of documents in addition to the individual documents provided by the participants. These include, randomly selected, Quality Assurance documents, available freely from the University of Bath which contain textual and tabular content as well as images. Additionally, past year exam papers are included in this dataset. This would not only allow for a more uniform basis for evaluating the system's performance but also make it easier to compare the effectiveness of the system across different users.

3.1.1 Data Pre-processing

This step is a crucial process in our document processing workflow, to ensure high accuracy especially when dealing with different quality documents. Considering the type of data to extract and the nature of the project, choosing a text extraction tool to input the data depends on several factors. There are many text extraction tools available in Python, including `PyPDF2`, `PDFminer`, and `PyMuPDF`, each offering distinct features and capabilities tailored to different aspects of PDF manipulation and text extraction. However, the `PyMuPDF(fitz)` library, continues to emerge as a champion due to its low Levenshtein distance, high cosine, and tf-idf similarity (Schoonmaker, 2021). It is also 30 to 45 times faster than other popular packages including `PDFminer` and `PyPDF2` (Lieder, 2022). This tool supports basic Optical Character Recognition (OCR) capabilities and is compatible with other documents apart from PDF files. However, `PyMuPDF's` OCR capabilities are limited and may perform poorly with complex images. Therefore, adding an extra OCR layer, using `Tesseract-OCR` can significantly enhance the accuracy and quality of reading images.

When processing long documents using a question-answering system, the length of the prompt can only be as long as the large language model API specifies. The token limit ranges from 4,096 to 32,768 max tokens for GPT-3.5 and GPT-4 (Arefeen et al., 2024). As the token limits increase, the cost of using these models increases leading to a waste in the computational resources. If the prompt exceeds the maximum token limit, LLMs may struggle to respond to queries, due to the longer prompt context (Arefeen et al., 2024). A viable solution to this problem is to employ another pre-processing step called chunking.

This method involves breaking down the documents into fixed chunk sizes and depending on a specified chunk size, it allows for a more efficient search of documents. This increases the simplicity and improves the retrieval efficiency in vector-based search mechanisms to use the embeddings created from texts of limited length to retrieve the relevant information (Alkestrup & Purup, 2024). This is because fixed-chunk sizes reduce the computational load when generating and comparing embeddings. Choosing an appropriate chunk size requires a trial-and-error method. There is a trade-off between large and small chunk sizes. While a larger chunk size could offer more context, it could hinder the retrieval of specific terms. At the same time, a smaller chunk size could improve the accuracy of aligning with particular queries but lack the context needed (Alkestrup & Purup, 2024). Retrieval of relevant chunks based on the user's query is also known as retrieval augmented generation (RAG) of context (Arefeen et al., 2024).

3.2 Implementation of RAG system

Retrieval Augmented Generation is a machine learning framework that fills a gap in how large language models work (Merritt, 2024) and is extremely advantageous with issues such as "hallucination" by providing more grounded reliable information (Retrieval augmented generation (RAG) – nextra, no date).

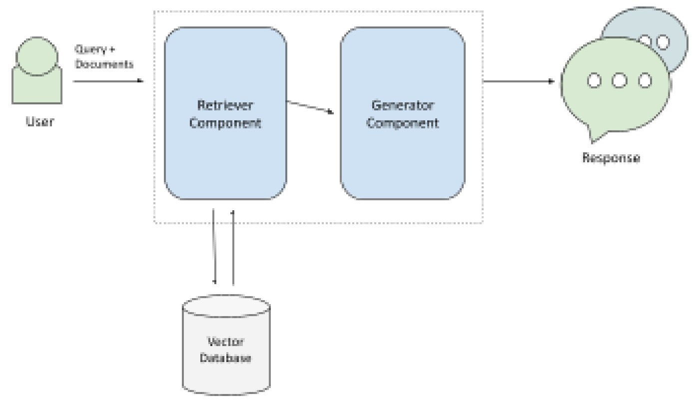


Figure 3.1: Displays a basic diagram of the RAG framework

These systems have been used in various applications, due to their flexibility, ranging from chatbots to question-answering systems. They can be applied to several natural language processing tasks including text summarisation, etc (Kartchner, 2024). The structure of this framework comprises of two main components: a retriever component, responsible for finding the relevant documents while leveraging knowledge from a vector database, and a generator component, which utilises the retrieved relevant documents as input along with the query in order to generate a response (Kartchner, 2024).

The system in this study incorporates retrieval-augmented methodologies similar to those found in existing literature yet could be considered different due to the models and libraries used. Unlike typical Retrieval-Augmented Generation (RAG) systems that predominantly utilise transformer-based models alongside retrieval systems like Elasticsearch, this study introduces an innovative integration of tf-idf with contextual embeddings. This novel approach, which uniquely combines traditional and neural methods, has not been previously documented and is tailored to enhance retrieval performance for specific applications.

3.2.1 Retriever Component

This component is the initial phase of the RAG system. It is directly tied to information retrieval as it involves searching through large volumes of data to find information relevant to the user's query. The retriever component employs the use of a vector database, embeddings, and other information retrieval methods to rank and retrieve the most relevant documents.

All the documents, inputted by the user, are processed using tf-idf (term frequency - inverse document frequency), to enhance the searching capabilities by weighting the terms based on their importance across documents. This method helps rank the relevant documents, emphasising words that are unique and minimising stop words such as "a", and "the", which do not contribute to the relevance.

$$\text{TF-IDF}(t,d) = \text{TF}(t,d) * \text{IDF}(t,d) \quad (1)$$

where TF is the term frequency of term t in document d, IDF is the inverse-document frequency calculated as:

$$IDF(t,d) = \log(\text{number of documents}/\text{document frequency}(t)) \quad (2)$$

where document frequency(t) is the number of documents that contain the term t

The documents ranked higher for the given query are broken down into chunks. Performing tf-idf before chunking reduces the volume of text to be chunked, and ensures the chunking process is only applied to the relevant documents leading to a higher quality of data being fed into the subsequent processes.

In the retrieval component, there are other techniques for enhancing search retrieval mechanisms such as using Named Entity Recognition (NER), a powerful information extraction tool (Named Entity Recognition and Seo: The ultimate guide, no date). Applying NER on queries and documents, allows the model to structure the search queries and understand the intent behind them. It also allows documents to be indexed not just on keywords but also based on the entities, such as names, locations, and organisations that they contain. However, through testing and experimentation, NER wasn't capturing relevant information due to the nature of the queries. Many queries, such as *"Who will periodically review the quality processes and principles?* or *"Why is it important not to let the test sentences into the training set?"* focus on understanding and reasoning rather than specifying named entities. This indicates that not all questions contain or require the identification of named entities, as they often seek explanations or procedural clarifications instead.

Since some queries may not be entity-centric, techniques such as keyword extraction using tf-idf, can be more effective. Instead of focusing on entities, these techniques extract key terms or phrases that can be more relevant for understanding and responding to queries. Additionally, using contextual embeddings can further enhance the searching process as these capture the meaning of the word in the context of the entire sentence or document.

The relevant chunks are converted into a vector representation. This involves converting text of limited length into embeddings, which are numeric representations designed to capture the semantic meaning behind the text (Selvaraj, 2024). The embedding model used, `all-MiniLM-L6-v2`, is a contextual embedding model that is stored in a vector database (`FAISS`) to then be retrieved by the generator component to form an answer. This knowledge database was chosen as it is known for conducting similarity searches in large-scale applications and its uses range from recommendation systems to image retrieval (Mudadla, 2023).

Combining tf-idf with contextual embeddings addresses the inherent limitations of tf-idf, creating a hybrid approach that leverages the strengths of both methods. While tf-idf provides the initial filtering of the relevant documents, it struggles with understanding meaning and recognising synonyms. The embeddings from the `HuggingFaceEmbeddings` hub, such as `all-MiniLM-L6-v2,` complement this by providing deep semantic analysis on this filtered set. The embedding model is specifically a sentence embedding model that captures the semantic meaning of sentences in a context-aware manner. It is 5 times faster than other models (Reimers, no date) and is ideal for real-time applications that have limited resources (Martinez, 2023).

This approach not only enhances understanding of textual context but also captures linguistic nuances, ensuring a more accurate and insightful understanding of the data. Therefore, when a query is made, the model computes an embedding for it and compares it to the embedding of the vector database, using cosine similarity, retrieving the most relevant documents to the query (Selvaraj, 2024). Cosine similarity values range from 0 to 1 where the value being closer to one indicates that the response closely aligns with the query.

$$\text{Cosine Similarity} = A.B/|A||B| \tag{3}$$

where A and B are vectors

The effectiveness of this component influences the overall quality of the generated responses, as it determines what information is available for the generator to use.

3.2.2 Generator Component

This is the second phase of RAG where a broader concept of information extraction plays an important role. The `RetrievalQA` chain serves as a generator component within the RAG framework for question-answering tasks. It integrates the capabilities of the vector database and the large language model to generate answers (Routhu, 2023). This chain uses the retrieved relevant documents along with the large language model as context to formulate a coherent and contextually relevant answer based on the user's query. As mentioned in section 2.5, this chain returns the source documents and the page number of the retrieved context to allow users to understand and depict how and where the answer was formed.

While `RetrievalQA` is a powerful approach, it faces difficulties preserving conversational history, and each question is treated independently (Routhu, 2023). This approach could be advantageous as it allows users to jump in and out of topics without the system needing to remember the previous conversation. This is extremely useful in environments where users are likely to have unpredictable or sporadic interactions with the system. It also enables the development focus on improving the quality of answers and expanding the knowledge base, rather than handling intricate dialog management.

The large language model is integral to this component. These models are trained on large datasets, enabling them to understand and generate human-like text. They can not only provide real-time responses but can handle a broad range of topics and query types such as questions about science, history, or technology. The choice of the model depends on the specific requirements and the problem at hand. I used Ollama which allowed me to run the llama-2 13b parameter model locally.

Llama-2 was the top choice for this application. This is because it has been emerging as a promising large language model and is considered to be a potential game changer for AI for business (Ruiz, 2023). (Ezat, 2023) comprehensive benchmark tests – evaluating aspects such as reading comprehension, mathematical abilities, and prompt responsiveness – demonstrated that llama-2 significantly surpassed chatGPT in performance. Moreover, Llama-2's advanced safety features, achieved through extensive fine-tuning, have earned it higher safety rankings, which notably mitigate risks associated with using third-party AI hosting services. Although llama-2 displays extraordinary potential, it may not be as powerful due to its limited number of parameters, which means it may not be able to generate complex text (Meta Llama VS CHATGPT: A comprehensive comparison, 2023).

Whilst such findings may be useful, there have been recent developments showing emerging models, including Falcon-180b and PaLM 2, outperforming Llama-2 and GPT 3.5. However, Falcon-180, despite being a state-of-the-art model, requires intensive computing and large computational resources, limiting its availability. (Marie, 2023). On the other hand models like Gemini and PaLM 2 require API access, which is only available in certain regions, further constraining their usability. These factors underscore the advantage of Llama-2 which is easily downloadable (open-sourced), usable anywhere for free, and offers greater accessibility and convenience for a wide range of users.

The retrieval and generator components work together interactively to create an accurate question-answering workflow:

1. Query Processing: The user's query is initially processed to determine the scope and the context.
2. Document Processing: The documents go through an extensive pre-processing process to find the documents that align closely with the query.
3. Answer Generation: Along with the retrieved content and the large language model, the response is generated and displayed to the query along with the sourced documents and page numbers.

3.3 Architectural Design and Implementation

The detailed design and implementation dive into the aspects of the software project, shedding light on the underlying technical framework and the architectural decisions that drive its functionality.

The frontend framework employs a combination of HTML, CSS, and JavaScript to create a user-friendly and responsive interface. I chose to avoid the use of heavy libraries or frameworks like React or Angular to reduce the computational resources and complexity involved in integrating these technologies and focus more on core technologies sufficient for the requirements of the application.

The backbone of the front-end design is HTML, which allows users to submit their queries and documents, crucial for the functionality of a question-answering system. There are dedicated containers that allow users to see the documents, page numbers, sentence sources, and cosine similarities to justify the response. These are dynamically updated based on the user interactions and backend responses. JavaScript plays a significant role as it allows for more dynamic and interactive features as well as real-time information updates.

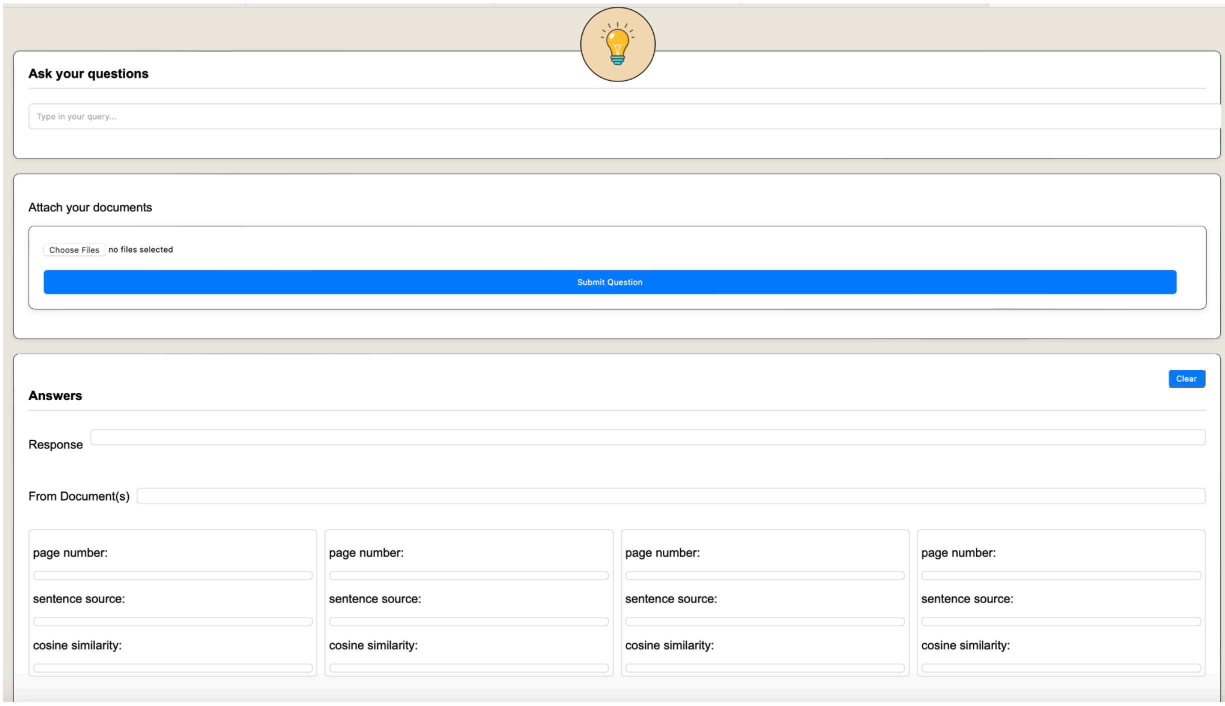


Figure 3.2: User Interface

The backend framework employs Django, an open-source powerful, and versatile web framework for Python. It was chosen because of its simplicity, scalability, and speed, saving development time (Korsun, 2024). This framework supports robust file uploads and manages file storage to handle multiple files simultaneously. Python was chosen as the main coding language due to its simplicity, comprehensive libraries, and seamless integration with AI models. This is crucial as the application interfaces with the Langchain API for question-answering and document handling.

Langchain is another open-source framework that provides developers with the necessary tools to build applications utilising large language models (Korsun, 2024). It is easy to use, scalable, adaptable, and expandable as developers can add their own features and functionalities (Korsun, 2024). This framework allows the use of tools such as `HuggingFaceEmbeddings`, `FAISS`, and `RecursiveCharacterTextSplitter`, as demonstrated in sections 3.2.1 and 3.2.2, which prove to be crucial in enhancing the system's overall effectiveness. By focusing on the ease of use, from question submission to viewing responses, the design prioritises user experience, aiming to reduce cognitive workload and streamline interactions.

3.4 System Testing

Implementing comprehensive testing strategies allows for the assessment of the overall capabilities and performance of the system in real-world scenarios. Three different testing strategies employed include scalability tests, usability tests, and performance tests. Scalability tests are used to test the system's ability to handle large volumes of documents. The system would be monitored for crashes or memory leaks. Secondly, usability tests ensure the system is user-friendly. From the participants, feedback would be collected on their experiences focusing on the ease of use and overall satisfaction. Lastly, a performance test assesses how well the system can handle different types of PDF data and whether the correct documents are being retrieved. Using a gold standard dataset, by applying performance metrics such as precision, recall, and F1-score, one can determine the system's accuracy and reliability.

3.4.1 Evaluation Metrics

Precision is a metric which measures the proportion of correct answers among the answers produced by the QA system. This metric is where the goal is to minimise the retrieval of irrelevant information. High precision indicates that a large number of the documents retrieved by the system are relevant to the query.

$$\text{Precision} = TP / TP + FP \tag{4}$$

Where TP stands for true positives (a relevant document classifying as a relevant document) and FP stands for false positives (an irrelevant document classifying as a relevant document).

Recall is a metric which measures how effectively the system retrieves all the relevant documents. This metric is crucial to ensure that documents relevant to the query are not missed.

$$\text{Recall} = TP / TP + FN \tag{5}$$

Where FN stands for false negatives (a relevant document classifying as an irrelevant document)

Precision and recall are inversely proportional and therefore ensure a more balanced approach, the F1-score is used. This metric combines both precision and recall into a single measure that captures both properties.

$$\text{F1-score} = 2 * (\text{Precision} * \text{Recall}) / \text{Precision} + \text{Recall} \tag{6}$$

It is not only important to measure the accuracy of the documents but also to ensure the responses don't contain any hallucinations and closely align with the ground truth. Therefore, statistical metrics such as Rouge and BLEU are utilised. The Rouge and BLEU metrics help to evaluate how well the system's generated content (responses from the QA system) matches the reference texts (ground truth) used for comparison. ROUGE (Recall Oriented Understudy for Gisting Evaluation) measures the quality of a summary by comparing it to reference summaries (Lin, 2004) while BLEU (Bilingual Evaluation Understudy) measures how many words and phrases in the generated text match the reference texts, adjusted for proportion (to prevent bias towards larger texts).

3.5 Limitations and Solutions

While certain requirements and specifications were easily established, other aspects posed significant challenges that required deep consideration and problem-solving. Processing tables with simple text and images with our system was an easy approach due to the OCR and `PyMuPDF` modules. However, the complexity increased significantly when handling detailed diagrams and intricate graphical data. These complex images contain nuanced information that is not readily interpreted by OCR technologies. Moreover, while the system can read the content in the tables, it struggles to recognise the structure as a table. Enhancing the system's ability to identify and interpret tables remains a key area of development, requiring more advanced AI tools.

The application takes time to run. This is due to the processing of large documents including extracting text and processing all the images. Although this issue wasn't completely resolved, updates were implemented in the user interface to reduce confusion. Specifically, changing the design to show that the application is running, preventing users from submitting their documents multiple times. This has increased the overall user experience.

There were a few challenges encountered during the development phase. Initially, there were hallucinations caused by the large language model due to the lack of context. This could result from using a fixed chunk size and employing tf-idf before, as detailed in sections 3.1.1 and 3.2.1. This would cause the document structure to be ignored, leading to the awkward breaking of sentences or paragraphs, which would result in the loss of context.

To potentially mitigate this limitation, the chunk size and the chunk overlap were increased. While increasing the chunk size allows it to retain more information, increasing the chunk overlap ensures that information at the edges of the chunk is not lost. This also allows for smoother transitions between chunks which is beneficial for tasks like document summarisation (Tian et al., 2020). Another solution to reducing hallucinations, as mentioned in section 2.5, was adding more facts to the prompt to make it more specific and detailed, thus allowing the LLM to be more accurate when retrieving information.

3.6 Summary

The system emphasises real-time responses, user-friendliness, and the ability to handle a mix of text and image documents. It employs a question-answering system that harnesses the potential of a retrieval augmented generation framework and utlises llama-2 for generating responses. Performance metrics such as precision, recall, and F1-score are used to assess the system's performance. Challenges like processing complex images and managing large document volumes are acknowledged, with ongoing solutions such as improved OCR and UI updates. In addressing these challenges mentioned, it is also important to note that despite the loss of some document structure, the system's performance was not overwhelmingly compromised. The large language model was still able to generate relevant responses, maintaining both the system's robustness and its focus on user needs.

4 Results and Discussion

This study was meticulously organised to evaluate the capabilities of document retrieval and question-answering systems using a large language model (LLM), llama-2. The application works by inputting a query and the respective documents. The system can handle multiple documents but will make sure to return the document from where the answer is found and the response generated using the large language model may not give you the exact extracted text from the document but would concise the content from the top relevant documents and form an answer.

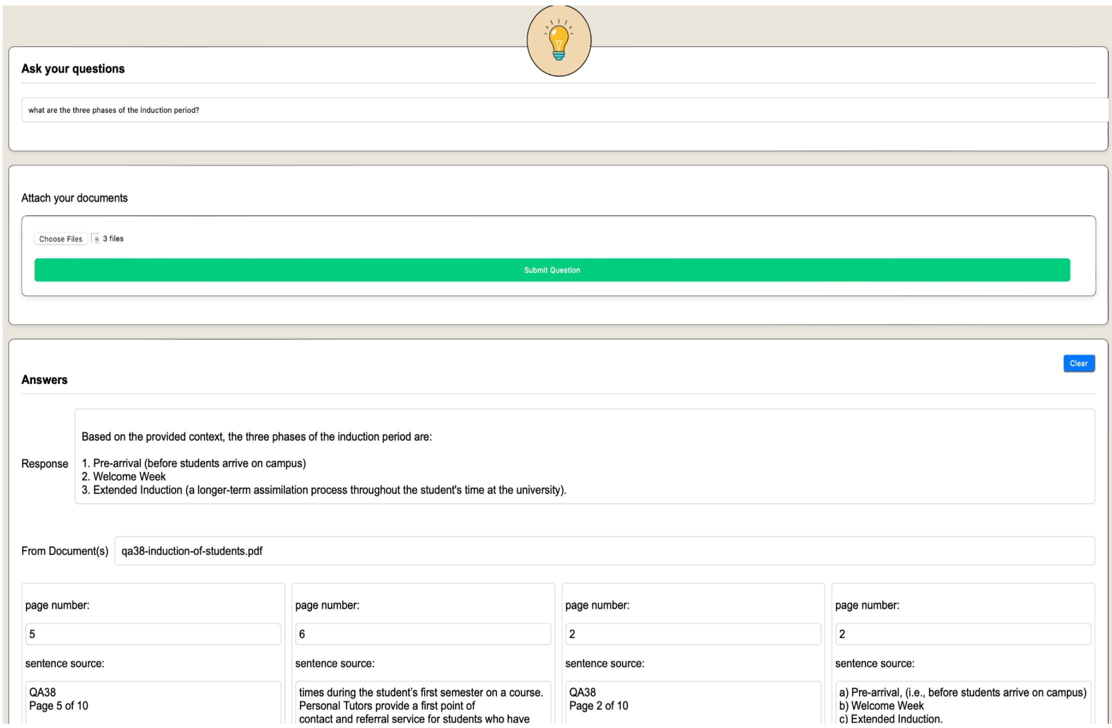
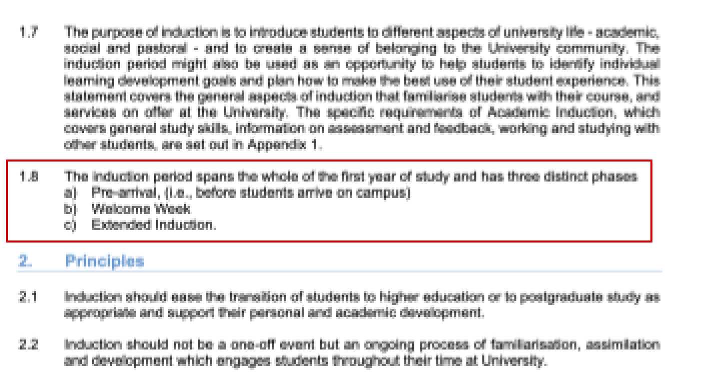


Figure 4.1: Example Output



1.7 The purpose of induction is to introduce students to different aspects of university life - academic, social and pastoral - and to create a sense of belonging to the University community. The induction period might also be used as an opportunity to help students to identify individual learning development goals and plan how to make the best use of their student experience. This statement covers the general aspects of induction that familiarise students with their course, and services on offer at the University. The specific requirements of Academic Induction, which covers general study skills, information on assessment and feedback, working and studying with other students, are set out in Appendix 1.

1.8 The induction period spans the whole of the first year of study and has three distinct phases
a) Pre-arrival, (i.e., before students arrive on campus)
b) Welcome Week
c) Extended Induction.

2. Principles

2.1 Induction should ease the transition of students to higher education or to postgraduate study as appropriate and support their personal and academic development.

2.2 Induction should not be a one-off event but an ongoing process of familiarisation, assimilation and development which engages students throughout their time at University.

Figure 4.2: Verifiable Document for Figure 4.1

4.1 Organisation of Experiments

A total of 10 participants were randomly selected based on specific criteria: they needed to be enrolled in different courses and be at least 18 years old. This diversity ensured that the system was tested across various types of course materials and levels of complexity, which were represented through two sets of datasets, detailed in section 3.1.

The experiments were conducted in a controlled environment where the participants interacted with the system in real-time. Initially, each participant was provided with a participant information sheet, as shown in the Appendix, describing the procedure, before giving their consent. They were instructed to provide precise questions relating to their document and the common set of documents provided. The responses including the location of the page and the sentence source were manually checked to determine the reliability of the system. The experiments were designed to assess the system under different conditions, including scalability, usability, and accuracy of answers, focusing on real-world application scenarios.

4.2 Scalability Testing

A timer was set at the beginning of the application and stopped at the end, i.e., when the response was generated.

Type of Documents	Number of documents	Speed of Response
Image-based	5	6 minutes and 5 seconds
Textual-based	10	1 minute and 16 seconds
Textual-based	20	1 minute and 41 seconds
Mix	100	5 minutes and 3 seconds

Table 4.1: Speed of the application

As shown in Table 4.1, the response time is severely affected by the type of PDF document. The mix of documents exhibited a shorter response time as it's less resource-intensive to process and can be analysed directly. However, image-based documents, even though the quantity is less, require OCR tools to convert visual information into a format suitable for the llama-2 model to process and analyse. The efficiency in processing textual data helps moderate the overall response time when mixed with image-based documents. A large number of image documents could cause the kernel to crash.

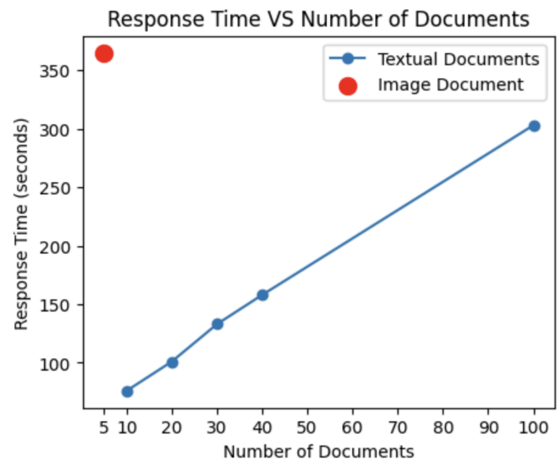


Figure 4.3: Linear increase in documents

The increase in the number of textual documents, ranging from 10 to 40, only slightly increases the response time which shows a linear or sub-linear scaling in processing time with the increase in document count, displayed in Figure 4.3. This is a good sign of the system's ability to handle large volumes of data.

4.3 Usability Testing

The experiment design incorporated both quantitative and qualitative methods to gather comprehensive data on the system's performance, gathered through a set of interview questions as shown in the Appendix. Quantitative data were captured through metrics such as response time (section 4.2), accuracy, ease of use, and reusability rate. Qualitative feedback was collected to understand the user experience in-depth, focusing on system performance and the user interface.

4.3.1 Quantitative Results

Tasks	Average Rating
Accuracy of results	8.5
Location of Page	9.7
Ease of use	9.0
Reusability Rate	9.0

Table 4.2: Quantitative Analysis

The feedback showed that the users found the system intuitive and highly effective in terms of page location and reusability, but they did encounter a few occasional errors that affected the overall accuracy. This could be either due to insufficient details in the prompt or the inability of the model to understand the context. Further information could be found in the qualitative analysis.

In addition, the users required a minimal number of clicks, roughly around 4, to work the system showing simplicity, efficiency, and higher productivity. In terms of user interactions, the error rate helps identify areas that may be causing difficulties and would need improvement. The primary errors made by the users involved not providing enough facts in the prompt, which often led to irrelevant results. Additionally, users needed clarification on where to input the queries, mainly due to the font size, though this issue did not majorly impact the system.

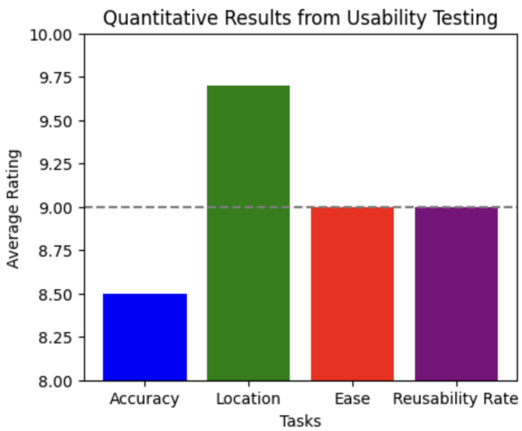


Figure 4.4: Bar Chart representing Quantitative results where the dotted line represents the benchmark value

4.3.2 Qualitative Results

The data gathered from the users refer to their experiences and direct quotes while using the application, highlighting the strengths and key areas of improvement. When using the common dataset, the users were able to retrieve all the relevant and accurate information. However, the evaluation of the system was affected by the different course materials used. The feedback is divided into clear themes, which will pave the way for future enhancements.

Theme 1: System Performance

This theme illustrates the balance between accuracy and occasional errors. For users with courses that rely on textual-based data such as management, sociology, or even computer science, the system was able to provide accurate answers to even vague questions while also providing extra information that was overlooked by the user.

Quote: *"For a few questions, even though it picked it up from the right places. It wasn't the exact sentences but the model formatted it in its own words but it was relevant to the question"*

There were instances where the model did not give the exact answer from the document but picked up content from different locations still relevant to the query and framed it in its own words. As detailed in section 3.5, this is the result of fixed-size chunking. When users search for documents using queries, they receive text chunks ranked by relevance, not in their order. This could lead to situations where users view excerpts from various parts of a document (or multiple documents) that, while relevant to the query, do not provide a clear cohesive understanding of the overall document structure.

A user studying engineering faced challenges with the system particularly when dealing with tables and intricate details in images, which are common in engineering documents.

Quote: *"It mixed up information from the set of tables. It was not random but not quite related to the query"*

This feedback indicates that while the system can identify the relevant topics with complex tables, the precision in handling intricate data needs improvement.

For psychology students, the system was generally effective, especially for handling textual research papers. The system was able to provide a summarisation and a comparative analysis of topics within the research paper.

Quote: *"It picked out key information from the document. It was precise in answering the question "can you differentiate between female autism type and female protective effect". Summarises the main aims and implications"*

Although there were instances where the interpretation of the text did not align with the user's expectations.

Quote: *"Accurately identified, gave a good explanation for a generic question while picking up all the main points. One of the points didn't convey the message of the query".*
Quote: *"Knew where the answers were coming from but aren't related to the question"*

This could either mean that the model doesn't fully capture the subtle meaning of complex psychological ideas or that the sentence wasn't written clearly enough.

In comparison to engineering students, a user studying biology appreciated the system's ability to handle simple diagrams.

Quote: *"It can analyse the diagrams which ChatGPT often gives vague and complicated answers. Provides straightforward answers but also says when it cannot provide direct information, that way the user's not confused"*

This indicates that while the system can handle simple diagrams, its capability to parse and interpret more complex diagrams remains limited.

Theme 2: Comparative Advantage

This theme reveals that the users appreciated the unique features compared to traditional methods. It was compared to Command-F, GPT-3.5, GPT-4, and other AI platforms.

Quote: *"Command-F gives you 900 results. It is time-consuming and annoying"*
Quote: *"More specific information than command-f, which just shows you words (very vague)"*

Command-F which is a basic search mechanism, users believed that this would provide you with loads of search results of just single words and can be extremely time-consuming. This system collates the content into one place, provides a faster way of reading through the content, and can handle multiple documents.

Quote: *"gpt3.5 you can't submit documents"*
Quote: *"GPT-4 is too expensive"*

Users mentioned that versions like GPT-3.5 don't allow users to submit their documents for analysis which underscores the functionality of the current system that supports document uploads, making it more useful for personalised queries. Users were also quick to point out that GPT-4, although very successful and useful is highly expensive, which in comparison, this system is widely accessible to a wider audience.

Quote: *"You can Fact-check".*

This quote highlights the system's ability to verify the facts quickly, offering users a reliable way to confirm the accuracy, through the page number and sentence source, while reducing the chances of misinformation.

Quote: "*Specific answers (concise). Other AI platforms can be more generic. This system pinpoints the actual content"*

Reflecting other AI platforms, users believed that they offer generic or irrelevant information while this system delivers precise and concise answers tailored to the user's query.

Quote: *"Found GPT better as it's been trained more, didn't need to use Google to have to search through it"*

Although one of the users believed that GPT offered better results as it has been trained more on a larger dataset but appreciated that the current system is better than a naive Google search.

Theme 3: User Experience
This theme highlights the ease and use of the application. Users were generally pleased with the design and didn't need any additional guidance with the system; their only suggestion was to increase the font size.

Quote: *"Not tech-literate, could understand the application easily"*
Quote: *"Helps summarise information not available to the naked eye"*
Quote: *"Would recommend it to different people. Good method of studying especially for students"*

These quotes show that the users would re-use the system either just to get a basic understanding of the content or when studying and don't have enough time to read through all the notes.

Theme 4: Suggestions for Improvement

This theme shows that while the users were content with the results of the application, they wanted the system to increase the speed of returning the answers, refine the accuracy when dealing with tables and images as well as understand the sentences better as they can mean different things.

Quote: *"would like to keep the document in"*

One of the users believed that it would be easier and more user-friendly if the users didn't have to continuously add the document in.

4.4 Performance Testing

As mentioned in section 3.4, a gold standard was implemented to evaluate the system's ability to retrieve the relevant documents that aligned closely with the query. This gold standard consisted of 49 ground truths, specifying the query, the expected documents, pages, and answers. It is important to note that the large language model's output is highly unpredictable; therefore, given this situation, the gold standard specified only one document as the correct answer.

4.4.1 Retrieval of Documents

The precision was calculated as the number of true positives (documents retrieved by the LLM and in the gold standard) divided by the total number of documents retrieved. This gives a measure of the accuracy of the retrieved documents, focusing on the proportion of retrieved documents that are relevant.

The recall was computed as the ratio of the true positive documents (those both retrieved by the LLM and specified as correct in the gold standard) to the total number of documents that should have been retrieved according to the gold standard. This measures the completeness of the retrieval, indicating whether the LLM managed to retrieve all the documents as anticipated.

Precision	Recall	F1-score
0.38	0.92	0.50

Table 4.3: Document Retrieval performance

The precision was lower than expected, which indicates over-retrieval. This is because if extra documents were retrieved that are not in the gold standard, this increases the denominator, thus reducing the precision. The recall is closer to 100% showing that all the relevant documents (as per the gold standard) have been retrieved.

As mentioned in section 2.2.1, even though the precision is lower, it is more important for the recall to be as close to 100% in information retrieval tasks. This is because humans need more help in identifying the relevant information rather than recognising an answer is incorrect (Berry, 2021).

Nevertheless, a way to improve this system would be to refine the retrieval criteria or the model's understanding of document relevance. Alternatively, it could suggest the need to expand the gold standard to include more documents, as these documents may be relevant if the answer appears in different locations.

4.4.2 Accuracy of Answers

To verify the correctness of the answers, Rouge and BLEU were employed. These metrics are used for assessing text similarity, reducing the chances of hallucinations (Minaee et al., 2024).

However as shown in an example output below, it is not the best metric to use.

```
{'query': 'How many years do Major reviews take?', 'expected_answer': '1-2 years',
'generated_response': '\nAccording to the text, major reviews take "1-2 years" to
complete.', 'rouge_scores': {'rouge-1': {'r': 0.0, 'p': 0.0, 'f': 0.0}, 'rouge-2':
{'r': 0.0, 'p': 0.0, 'f': 0.0}, 'rouge-l': {'r': 0.0, 'p': 0.0, 'f': 0.0}},
'bleu_score': 0}
```

As you can see above the generated response is semantically correct but phrased differently from the expected answer leading to lower scores from Rouge and BLEU. Therefore, using another metric such as BERTscore which captures semantic similarity would be better. It leverages contextual embeddings from pre-trained transformer models and uses cosine similarity to evaluate the similarity between the two sentences. This method captures the meaning, which is more robust against paraphrasing (Dhungana, 2023).

```
{'query': 'How many years do Major reviews take?', 'expected_answer': '1-2 years',
'generated_response': 'According to the text, major reviews can take 1-2 years.',
'bert_score': {'precision': [0.85148024559021], 'recall': [0.8930903673171997], 'f1':
[0.8717890977859497], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}
```

According to the example output above (more found in the Appendix), the high precision indicates that a large proportion of generated content is relevant or present in the reference texts, and the high score of recall suggests that the generated response successfully captures the overall important points from the expected answer. The F1 score, accounting for both false positives and false negatives, shows a good balance between precision and recall. This indicates that the generated content is accurate and comprehensive in comparison to the expected answer.

Average BERTscore Precision	Average BERTscore Recall	Average BERTscore F1-score
0.85	0.88	0.86

Table 4.4: Response Accuracy

However, it is important to note that even today, accurately assessing the content generated remains a significant challenge, as no perfectly suitable metric has been established (Minaee et al., 2024). Even when using BERTscore, there have been cases where the answer may be incorrect but still receive a high Precision, recall, and F1-score rate. Therefore, manually checking is always important and one of the best metrics to use.

Despite a relatively low precision in document retrieval as shown in Table 4.3, the large language model (llama-2) demonstrated its effectiveness in delivering accurate responses, as shown in Table 4.4. This discrepancy between the two precisions can occur due to the language model's advanced natural language understanding capabilities. Essentially, even if the model returns a broader set of documents than necessary, it can still effectively parse and comprehend the relevant content within those documents, utilising embeddings and deep learning techniques.

4.5 Overall Capabilities

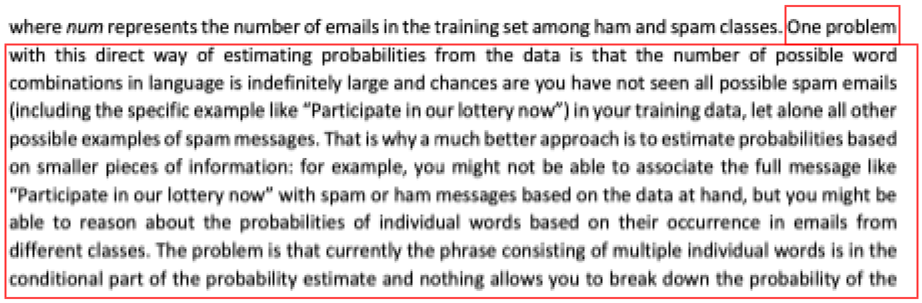
The large language model, llama-2, demonstrates varied capabilities across different tasks, a fact that becomes evident through the targeted questions designed to test its strengths and limitations. These questions are categorised to cover areas such as comprehension, reasoning, summarisation, tabular data interpretation, and visual data processing aiming to evaluate how well llama-2 can understand complex queries, reason logically, and retrieve accurate information. By diversifying the types of questions, we can assess the model's performance in specific cognitive domains and understand its potential applications and limitations in real-world scenarios.

Type I: Reasoning Questions

The response generated by the large language model:

```
{'response': {'query': 'Why is it a better approach to estimate probabilities
based on smaller pieces of information?',
  'result': '\nThe reason why estimating probabilities based on smaller pieces
of information (e.g., individual words or phrases) is a better approach than
considering the entire email as a single unit is because it allows the
classifier to capture more nuanced patterns and relationships in the
data.\n\nWhen considering the entire email as a single unit, the classifier
may struggle to differentiate between spam and ham emails that have similar
overall content or structure. For example, an email that contains both spammy
language (e.g., "win big prizes!") and relevant information for the recipient
(e.g., "your account balance is overdue") might be difficult for a classifier
to classify as either spam or ham based solely on the overall content of the
email.\n\nBy breaking down the email into smaller pieces of information, such
as individual words or phrases, the classifier can capture more subtle
differences between spam and ham emails. For example, the presence of certain
keywords or phrases (e.g., "click here," "participate now," etc.) may be more
common in spam emails than in ham emails, and the classifier can use this
information to make a more accurate prediction.\n\nAdditionally, estimating
probabilities based on smaller pieces of information allows for greater
flexibility and adaptability in the classification model. The classifier can
focus on different aspects of the email depending on the specific task at
hand, and it can adjust its predictions based on the relative importance of
different features or patterns in the data.'}}
```

The response provided is not identical to the text in the document as shown in Figure 4.5, but it captures the same concept. While the model's response seems to fit or agree with the concept being discussed, it doesn't necessarily imply that the model is truly reasoning or thinking in a human-like manner. Instead, the model's response is based on patterns it has learned from the data during its training process. This suggests that the model does not independently understand or derive logical conclusions based on reasoning but rather produces outputs based on statistical correlations.



where *num* represents the number of emails in the training set among ham and spam classes. One problem with this direct way of estimating probabilities from the data is that the number of possible word combinations in language is indefinitely large and chances are you have not seen all possible spam emails (including the specific example like "Participate in our lottery now") in your training data, let alone all other possible examples of spam messages. That is why a much better approach is to estimate probabilities based on smaller pieces of information: for example, you might not be able to associate the full message like "Participate in our lottery now" with spam or ham messages based on the data at hand, but you might be able to reason about the probabilities of individual words based on their occurrence in emails from different classes. The problem is that currently the phrase consisting of multiple individual words is in the conditional part of the probability estimate and nothing allows you to break down the probability of the

Figure 4.5: Verifiable Document for the above result

Type II: Summarisation Questions

The response generated by the large language model:


```
{'response': {'query': 'can you provide a summarisation for the
article exploring open information extraction for portuguese using
large language models?',
  'result': 'sure! Here is a summary of the article "Exploring Open
Information Extraction for Portuguese Using Large Language
Models":\n\nThe article discusses the use of large language models
(LLMs) for open information extraction (OIE) in the portuguese
language. The authors explore the applicability of LLMs for OIE in
portuguese, examining various approaches and techniques such as
few-shot and prompt engineering. They also compare the performance of
an llm with a state-of-the-art oie system for portuguese and discuss
future research directions.\n\nThe article is structured as
follows:\n\n* Section 2 reviews related work on OIE and LLMs.\n*
Section 3 outlines the methodology and approach employed in the
study.\n* Section 4 presents the results of experiments conducted to
evaluate the performance of the llm for OIE in portuguese.\n* Section
5 concludes with a discussion of the findings and future research
directions.\n\nOverall, the article provides an exploratory analysis
of LLMs for OIE in the portuguese language and highlights their
potential for improving the accuracy and efficiency of OIE systems in
this language.'}}
```


Through manual evaluation, we assess the large language model's ability to summarise by examining its response for accuracy, and relevance. To show the model's understanding, we compare the summary it generated with expert-created summarises of the same article, as shown in Figure 4.6. This involves highlighting key points the model has successfully captured, noting any details it omits, and analysisng how effectively it condenses all the main points. The summary generated by the model broadly aligns with the paper but the expert-crafted summary dives deeper into the specifics, including mentioning the model and its performance.

In this work, we investigate the potential of Large Language Models (LLMs) for Open Information Extraction (OpenIE) in the Portuguese language. While most OpenIE methods are primarily optimized for English, only few works in the literature explore their uses for cross-lingual and multilingual scenarios. Despite the growing interest in Portuguese OpenIE methods, the use LLMs for Portuguese focused OpenIE is still an underdeveloped topic in the area.Our study addresses this research gap by examining the viability of using open and commercial LLMs with few-shot prompt engineering for Portuguese OpenIE. We provide an analysis of the performance of these LLMs in OpenIE tasks, revealing that they achieve performance metrics comparable to state-of-the-art systems. In addition, we have fine-tuned and launched an open LLM for OpenIE (PortOIE-Llama), which outperforms commercial LLMs in our experiments. Our findings highlight the potential of LLMs in Portuguese OpenIE tasks and suggest that further refinement and fine-tuning of larger models could enhance these results.

Figure 4.6: Verifiable document for the above response

Type III: Table Understanding

The response generated by the large language model:

```
{'response': {'query': 'What are the contents of the statement details table?',
  'result': ' Based on the provided context, here are the contents of the
statement details table:\n\n1. Issue Version: The current version of the
statement, which is 1.6.\n2. Date: The date the statement was last updated, which
is November 2021.\n3. Antecedents: Previous versions of the statement, including
the UltQC (27 March 2012), 8 July 2014, 11 July 2017, September 2018, and 16 July
2019 minutes.\n4. Education Quality and Standards Committee (EQSC): The date of
the most recent EQSC minute, which is 1 November 2021 (minute 148).\n5. Related
Documentation: External documentation such as the QAA UK Quality Code for Higher
Education, and internal documentation such as the statement of equality
objectives.\n6. Author: The person or team responsible for creating the statement,
which is the Academic Registry.'}}
```

Llama-2 was able to identify the contents of the Statement details table as shown in Figure 4.7, yet it failed to recognise the date corresponding to the issue version, as shown in the below response.

```
{'response': {'query': 'what is the date of the issue version 1.6 in the statement
details table?',
  'result': 'Based on the information provided in the "Statement Details" table,
the issue version "1.6" has a date of September 2023.'}
```

This error may stem from the inability of the large language model to recognise that the structure of the content is a table, in which each set of items across rows or columns is contextually linked. This remains a key area of development.

Statement Details		
Issue Version:	1.6	
Date:	November 2021	
Antecedents:	*ULTQC*	27 March 2012 8 July 2014 11 July 2017 September 2018 16 July 2019 Minute 1197
	Education Quality and Standards Committee (EQSC)	1 Nov 2021 minute 148
Related Documentation:	External QAA UK Quality Code for Higher Education Internal Statement of Equality Objectives	
Author:	Academic Registry	

Figure 4.7: Verifiable document for the above result

Type IV: Image Understanding

The response generated by the large language model:

```
{'response': {'query': 'what are the five segments of the luxury fashion
industry?',
  'result': 'Based on the context you provided, the five segments of the luxury
fashion industry are:\n\n1. Watches & Apparel\n2. Jewelry\n3. Beauty\n4. Y (a
combination of yoga and streetwear)\n5. Sustainable Luxury.'}}
```

Llama-2's response partially aligns with Figure 4.8 in mentioning watches, apparel, jewellery, and beauty. It may not have given the exact heading from the document but was still able to give a partially accurate response. However, it also lists Y (a combination of yoga and streetwear). These inaccuracies may arise due to the lack of training on such data or due to the capabilities of OCR technology, which can sometimes misread characters especially if the image quality is poor, or the text is styled.

Moreover, even if the text is read accurately, the model's ability to understand and contextualise visual information from images may not be as robust as its ability to process plain text.

Figure 4.8: Verifiable document for the above result

Type V: Arithmetic Reading Comprehension

The response generated by the large language model:


```
{'response': {'query': 'what can the velocity profile at cross section be written as?',
  'result': '   Based on the given information, we can answer the question at the end.\n\n(d) a newtonian fluid of dynamic viscosity, 1, flows with a mean velocity, u(r), along a horizontal circular pipe of radius, r, under the influence of a pressure gradient, (dp/dz). Perform a force balance over an element of the pipe and show that the velocity profile at a cross section can be written as:\n\nu(r) = g(-3)/2πr\n\nwhere r is the radial coordinate.\n\nJustification for assumptions made:\n\n1. The fluid is assumed to be newtonian, meaning that the shear stress is proportional to the shear rate.\n2. The flow is assumed to be steady and laminar, meaning that the velocity profile is constant across the cross section.\n3. The pressure gradient is assumed to be constant across the cross section.\n4. The flow is assumed to be irrotational, meaning that there are no vortices or swirling motions in the flow.\n\nSketch of the cross section:\n\nThe velocity profile at the cross section can be represented by a parabolic shape, with the highest velocity at the center of the pipe and decreasing velocities towards the walls. The shear stress is highest at the walls and decreases towards the center of the pipe.\n\nTherefore, the velocity profile at the cross section can be written as:\n\nu(r) = g(-3)/2πr\n\nwhere r is the radial coordinate.'}}
```


The language model's response does not correctly describe the physics formula as shown in Figure 4.9. As this document is an image, it could again depend on the OCR capabilities. This also indicates a gap in the model's understanding of interpreting visual data to complex formulas.

(d) A Newtonian fluid of dynamic viscosity, μ, flows with a mean velocity, u_m, along a horizontal circular pipe of radius, R, under the influence of a pressure gradient, (dp/dz). Perform a force balance over an element of the pipe and show that the velocity profile at a cross section can be written as,

$$u(r) = \frac{1}{4\mu}\left(-\frac{dp}{dz}\right)(R^2 - r^2)$$

where r and z are the radial and axial coordinates, respectively.

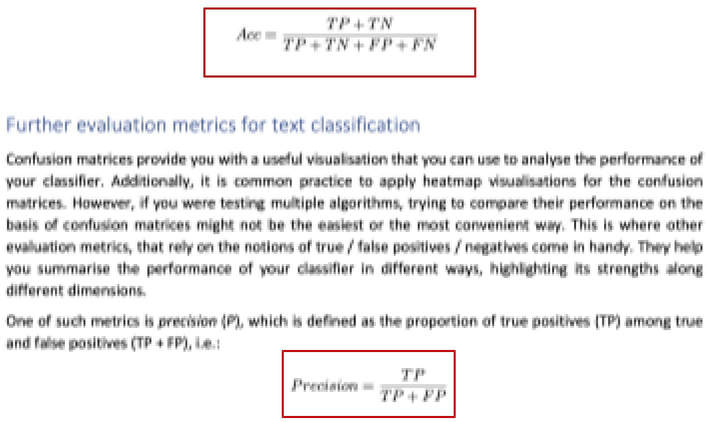
Show by means of a sketch the sign (i.e. positive or negative) for the shear stress. Justify all assumptions made. [8 marks]

Figure 4.9: Verifiable document for the above response

Taking on a different approach, when asked about a formula from a text-based document, the response was accurate, indicating that the model can correctly process and reproduce information when it is in a textual format that it can readily interpret without the need for OCR.

```
{'response': {'query': 'what is the formula for accuracy and precision in the
document?',
  'result': 'The formula for accuracy is:\n\nAccuracy = (True Positives + True
Negatives) / (True Positives + False Positives + True Negatives + False
Negatives)\n\nwhere True Positives are correct positive predictions, True Negatives
are correct negative predictions, False Positives are incorrect positive predictions,
and False Negatives are incorrect negative predictions.\n\nThe formula for precision
is:\n\nPrecision = True Positives / (True Positives + False Positives)\n\nwhere True
Positives are correct positive predictions, and False Positives are incorrect positive
predictions.'}}
```

This highlights a distinction in the model's performance based on the format of the input data: it has high expertise in handling text-based documents but may struggle with accurately interpreting information presented in images.



$$Acc = \frac{TP + TN}{TP + TN + FP + FN}$$

Further evaluation metrics for text classification

Confusion matrices provide you with a useful visualisation that you can use to analyse the performance of your classifier. Additionally, it is common practice to apply heatmap visualisations for the confusion matrices. However, if you were testing multiple algorithms, trying to compare their performance on the basis of confusion matrices might not be the easiest or the most convenient way. This is where other evaluation metrics, that rely on the notions of true / false positives / negatives come in handy. They help you summarise the performance of your classifier in different ways, highlighting its strengths along different dimensions.

One of such metrics is *precision* (*P*), which is defined as the proportion of true positives (TP) among true and false positives (TP + FP), i.e.:

$$Precision = \frac{TP}{TP + FP}$$

Figure 4.10: Verifiable document for the above response

4.6 Summary

The structure of the experiments included multiple phases: initial setup, where participants were briefed and consented; the execution phase, where the participants interacted with the system using their own and common datasets; and the feedback phase (usability testing strategy), where both quantitative and qualitative data was collected. Extra testing strategies– scalability, and performance – were employed aimed at evaluating the system through various angles.

These tests revealed the application is robust with text-based documents, offering fast response times. However, it faces challenges in processing intricate details within image-based documents and requires a significant amount of time to anlyse them.

5 Conclusion

In this study, we proposed a framework to develop a workflow that reduces hallucinations from large language models (LLM). This framework employed a question-answering system that leveraged the potential of retrieval-augmented methodologies. As detailed in section 3.2, this method uses different models and libraries, diverging from those typically employed in retrieval-augmented generation, to accommodate the varying complexity levels of the models. This adaption allows for more tailored responses to complex queries, improving the system's overall effectiveness and accuracy.

Understanding the abilities of llama-2, this application showed the importance of adding more facts to the prompt allowing the model to look in the right sections of the documents, reducing the chances of hallucinations. We found that llama-2 is more robust with textual-based documents but faces difficulties when analysing images. A potential limitation of this application is that it still produces occasional errors and relies heavily on OCR technologies which are not always reliable and frequently misinterpret characters. Additionally, it tends to be slow in returning answers, which could hinder its effectiveness in time-sensitive applications.

In comparison to different systems, this system leverages a hybrid approach using a tf-idf, a vector-based model, and contextual embeddings to deepen the semantic understanding of the context for document retrieval. While there was an over-retrieval of information, llama-2 was still able to receive a 0.85% precision and 0.88% recall rate, showing the effectiveness of llama-2's understanding capabilities.

Future work could involve enhancing the pre-processing steps to include maintaining the document structure through variable-size chunking and exploring the ability to understand intricate details within images and the structured nature of tables. It also involves expanding the database to different types of data formats such as web pages (HTML), XML, or even JSON files.

Large Language models in some cases still hallucinate answers as they are next-token prediction machines (Minaee et al., 2024). Hallucination is a key factor in assessing the reliability of the large language model. However, measuring hallucinations is challenging as information can be expressed in different ways, even small variations in phrasing can make it hard to detect. It is reasonable to suggest that a LLM that can more effectively identify hallucinations or false information in a text is likely to be more trustworthy (Minaee et al., 2024).

Word Count: 10,762

Bibliography

Alan, A.Y., Karaarslan, E. and Aydin, Ö. (2024) *A rag-based question answering system proposal for understanding Islam: Mufassirqas LLM*, *arXiv.org*. Available at: https://arxiv.org/abs/2401.15378

Alkestrup, J. and Purup, N.S. (2024) *Building Retrieval Augmented Generation (RAG) systems and specific...*, *Building high-quality RAG systems*. Available at: https://implementconsultinggroup.com/article/building-high-quality-rag-systems.

Allam, A. and Haggag, M.H. (2012) *(PDF) The Question Answering Systems: A survey*, *ResearchGate*. Available at: https://www.researchgate.net/publication/311425566_The_Question_Answering_Systems_A_Survey

Arefeen, M.A., Debnath, B. and Chakradhar, S. (2024) *LeanContext: Cost-efficient domain-specific question answering using llms*, *LeanContext: Cost-efficient domain-specific question answering using LLMs*. Available at: https://www.sciencedirect.com/science/article/pii/S294971912400013X.

Arefeen, M.A., Debnath, B. and Chakradhar, S. (2024) *LeanContext: Cost-efficient domain-specific question answering using llms*, *LeanContext: Cost-efficient domain-specific question answering using LLMs*. Available at: https://www.sciencedirect.com/science/article/pii/S294971912400013X.

Bakker, F., Heusden, R.V. and Marx, M. (2024) *Timeline extraction from decision letters using chatgpt*, *ACL Anthology*. Available at: https://aclanthology.org/2024.case-1.3/.

Béchard , P. and Ayala, O.M. (2024) *Reducing hallucination in structured outputs via Retrieval-Augmented Generation*, *arXiv*. Available at: https://arxiv.org/html/2404.08189v1

Berry, D.M. (2021) *Empirical evaluation of tools for hairy requirements engineering tasks - empirical software engineering*, *SpringerLink*. Available at: https://link.springer.com/article/10.1007/s10664-021-09986-0.

Bouziane, A. *et al.* (2015) *Question answering systems: Survey and Trends*, *Procedia Computer Science*. Available at: https://www.sciencedirect.com/science/article/pii/S1877050915034663

Caballero, M. (2022) *A brief survey of Question Answering Systems*, *SSRN*. Available at: https://deliverypdf.ssrn.com/delivery.php?ID=691097078009122091125064072026109005040037066015093025096086105069091121029001109121124061122043062030037028011115014071020010110055002081003089025085090003013126086048011003069089026072091077071090084120116112110014004006111064108065119108102084012117&EXT=pdf&INDEX=TRUE

Cabra, B., Claro, D.B. and Souza, M. (2024) *Empirical evaluation of tools for hairy requirements engineering tasks - empirical software engineering*. Available at: https://aclanthology.org/2024.propor-1.13.pdf.

Cabrio, E. *et al.* (2012) *Qakis: Proceedings of the 2012th International Conference on Posters & Demonstrations Track - volume 914*, *Guide Proceedings*. Available at: https://dl.acm.org/doi/10.5555/2887379.2887382.

Cai, H. *et al.* (2024) *SciAssess: Benchmarking LLM proficiency in Scientific Literature Analysis*, *arXiv.org*. Available at: https://arxiv.org/abs/2403.01976

Cho, S. *et al.* (2023) *Discrete prompt optimization via constrained generation for zero-shot re-ranker*, *arXiv.org*. Available at: https://arxiv.org/abs/2305.13729.

Dagdelen, J. *et al.* (2024) *Structured information extraction from scientific text with large language models*, *Nature News*. Available at: https://www.nature.com/articles/s41467-024-45563-x.

Dhungana, K. (2023) *NLP model evaluation: Understanding Bleu, Rouge, Meteor, and bertscore*, *Medium*. Available at: https://medium.com/@kbdhunga/nlp-model-evaluation-understanding-bleu-rouge-meteor-and-bertscore-9bad7db71170

Diefenbach, D. *et al.* (2017) *Core techniques of question answering systems over knowledge bases: A survey - knowledge and information systems*, *SpringerLink*. Available at: https://link.springer.com/article/10.1007/s10115-017-1100-y.

Ernandes, M. *et al.* (2006) *Adaptive context-based term (re)weightingan experiment on single-word question answering: Proceedings of the 2006 conference on ECAI 2006: 17th European Conference on Artificial Intelligence August 29 -- September 1, 2006, Riva del Garda, Italy*, *Guide Proceedings*. Available at: https://dl.acm.org/doi/10.5555/1567016.1567230.

Ezat, A. (2023) *LLama 2 vs ChatGPT Which is Better AI Model*, *Katteb Fact Checked AI Writer*. Available at: https://katteb.com/en/llama-2-chatgpt/.

Fatehkia, M., Lucas, J.K. and Chawla, S. (2024) *T-rag: Lessons from the LLM Trenches*. Available at: https://arxiv.org/pdf/2402.07483.pdf

Freitag, D. and McCallum, A. (2000) *Information Extraction with HMM Structures Learned by Stochastic Optimization*. Available at: https://www.aaai.org/Papers/AAAI/2000/AAAI00-089.pdf.

Gaizauskas, R. and Wilks, Y. (1998) *Information extraction: Beyond document retrieval*, *Journal of Documentation*. Available at: https://www.emerald.com/insight/content/doi/10.1108/EUM0000000007162/full/html?skipTracking=true.

Gao, J. *et al.* (2023) *Exploring the feasibility of CHATGPT for event extraction*, *arXiv.org*. Available at: https://arxiv.org/abs/2303.03836.

Gayathri, M. and Kannan, Dr.R.J. (2021) *(PDF) ontology based concept extraction and classification of Ayurvedic documents*. Available at: https://www.researchgate.net/publication/342221592_Ontology_based_Concept_Extraction_and_Classification_of_Ayurvedic_Documents.

Graph, R. (2024) *Hallucination in large language models and two effective alleviation pathways*, *Medium*. Available at: https://medium.com/@researchgraph/hallucination-in-large-language-models-and-two-effective-alleviation-pathways-a96c6a33bf02

Grishman, R. (2005) *Information extraction: Techniques and challenges*, *SpringerLink*. Available at: https://link.springer.com/content/pdf/10.1007/3-540-63438-X_2.pdf

Guo, M. *et al.* (2023) *Investigations on Scientific Literature Meta Information Extraction Using Large Language Models*, *CSDL | IEEE Computer Society*. Available at: https://www.computer.org/csdl/proceedings-article/ickg/2023/070900a249/1Ubpr2QOX5K.

Hambarde, K. and Proença, H. (2023) *Information Retrieval: Recent Advances and Beyond*. Available at: https://arxiv.org/pdf/2301.08801.pdf

Hambarde, K. and Proença, H. (2023) *Information retrieval: Recent advances and beyond*. Available at: https://arxiv.org/pdf/2301.08801.pdf.

Hobbs, J.R. and Riloff, E., 2010. Information Extraction. *Handbook of natural language processing*, *15*, p.16. Available at: https://citeseerx.ist.psu.edu/

Hong, J.L., Siew, E.-G. and Egerton, S. (2009) *Information extraction for search engines using fast heuristic techniques*, *Data & Knowledge Engineering*. Available at: https://www.sciencedirect.com/science/article/pii/S0169023X0900144X?casa_token=IrEiCkkBON8AAAAA%3AnBIAwGWs4bLu62GDnvPNVy32IXJ1mkiPprDAR3FUeHsE2pbjId5Nhqm2r2Iv2ZVQLP_JUK8o.

Huang, Y. and Huang, J.X. (2024) *Exploring ChatGPT for Next-generation Information Retrieval: Opportunities and Challenges*. Available at: https://arxiv.org/pdf/2402.11203.pdf.

Huang, Y. and Huang, J.X. (2024) *Exploring CHATGPT for next-generation information retrieval*. Available at: https://arxiv.org/pdf/2402.11203.pdf.

James, N.T. and Kannan, R. (2017) *(PDF) A survey on information retrieval models, techniques and applications*. Available at: https://www.researchgate.net/publication/326075149_A_Survey_on_Information_Retrieval_Models_Techniques_and_Applications.

Jeong, C. (2023) *A Study on the Implementation of Generative AI Services Using an Enterprise Data-Based LLM Application Architecture*. Available at: https://arxiv.org/pdf/2309.01105.pdf

Jeong, C. (2023a) *A study on the implementation of Generative AI services using an enterprise data-based LLM Application Architecture*, *arXiv.org*. Available at: https://arxiv.org/abs/2309.01105.

Kambau, R.A. and Hasibuan, Z.A. (2018) *Evolution of information retrieval system: Critical review of multimedia information retrieval system based on content, context, and concept | IEEE conference publication | IEEE Xplore*. Available at: https://ieeexplore.ieee.org/abstract/document/8265652/.

Kanaoka, K., Fujii, Y. and Toyama, M. (2014) *Ducky: Proceedings of the 18th International Database Engineering & Applications Symposium*, *ACM Other conferences*. Available at: https://dl.acm.org/doi/10.1145/2628194.2628244.

Karpukhin, V. *et al.* (2020) *Dense passage retrieval for open-domain question answering*, *arXiv.org*. Available at: https://arxiv.org/abs/2004.04906.

Kartchner, N. (2024) *An introduction to retrieval-augmented generation (RAG)*, *Label Studio*. Available at: https://labelstud.io/blog/an-introduction-to-retrieval-augmented-generation/.

Kasneci, E. *et al.* (2023) *Chatgpt for good? on opportunities and challenges of large language models for Education*, *Learning and Individual Differences*. Available at: https://www.sciencedirect.com/science/article/pii/S1041608023000195?casa_token=BoxuQvB11hgAAAAA%3AlEdAoieAZmspYPc3RCep8IHjaxvxIW_fBE6ocZevzdmQsGmcnh9Yv_0AdI0H2Rcufj-1GBQN.

Kasneci, E. *et al.* (2023) *Chatgpt for good? on opportunities and challenges of large language models for Education*, *Learning and Individual Differences*. Available at: https://www.sciencedirect.com/science/article/pii/S1041608023000195?casa_token=BoxuQvB11hgAAAAA%3AlEdAoieAZmspYPc3RCep8IHjaxvxIW_fBE6ocZevzdmQsGmcnh9Yv_0AdI0H2Rcufj-1GBQN.

Korsun, J. (2024) *What is LangChain: Components, benefits & how to get started*, *lakeFS*. Available at: https://lakefs.io/blog/what-is-langchain-ml-architecture/#:~:text=It%20allows%20LLM%20models%20to,dynamic%2C%20data%2Dresponsive%20applications.

Landolsi, M.Y., Hlaoua, L. and Ben Romdhane, L. (2022) *Information extraction from Electronic Medical Documents: State of the art and Future Research Directions*, *Knowledge and information systems*. Available at: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9640816/

Li, H. *et al.* (2023) *Extracting financial data from unstructured sources: Leveraging large language models*, *SSRN*. Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4567607.

Lieder, H. (2022) *Text Extraction with PyMuPDF*, *Artifex*. Available at: https://artifex.com/blog/text-extraction-with-pymupdf.

Lin, C.-Y. (2004) *Rouge: A package for automatic evaluation of summaries*, *ResearchGate*. Available at: https://www.researchgate.net/publication/224890821_ROUGE_A_Package_for_Automatic_Evaluation_of_summaries.

Lubis, A.R. *et al.* (2022) *The feature extraction for classifying words on social media with the naïve Bayes algorithm*, *IAES International Journal of Artificial Intelligence (IJ-AI)*. Available at: https://ijai.iaescore.com/index.php/IJAI/article/view/21502.

Marie, B. (2023) *Falcon 180B: Can it run on your computer?*, *Falcon 180B: Can It Run on Your Computer?* Available at: https://kaitchup.substack.com/p/falcon-180b-can-it-run-on-your-computer.

Martinez, N.A. (2023) *Understanding vector embeddings in NLP: An introduction with the all-MINILM-L6-V2 model*, *LinkedIn*. Available at: https://www.linkedin.com/pulse/understanding-vector-embeddings-nlp-introduction-model-martinez/

Meta Llama VS CHATGPT: A comprehensive comparison (2023) *Master Computer Vision Courses Online With Augmented A.I.* Available at: https://www.augmentedstartups.com/blog/metallamavschatgpt

Minaee, S. *et al.* (2024) *Large language models: A survey*, *arXiv.org*. Available at: https://arxiv.org/abs/2402.06196

Mishra, A. and Jain, S.K. (2015) *A survey on question answering systems with classification*, *Journal of King Saud University - Computer and Information Sciences*. Available at: https://www.sciencedirect.com/science/article/pii/S1319157815000890

Mudadla, S. (2023) *ChromaDB vsfaiss.*, *Medium*. Available at: https://medium.com/@sujathamudadla1213/chromadb-vsfaiss-65cdae3012ab.

Named entity recognition and Seo: The ultimate guide (no date) *Named Entity Recognition and SEO: The Ultimate Guide*. Available at: https://marketbrew.ai/named-entity-recognition-and-seo#:~:text=Overall%2C%20NER%20is%20a%20powerful,out%20spam%20and%20low%2Dquality

Piskorski, J. and Yangarber, R. (1970) *Information extraction: Past, present and future*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-642-28569-1_2.

Qu, Y. *et al.* (2021) *RocketQA: An optimized training approach to dense passage retrieval for open-domain question answering*, *arXiv.org*. Available at: https://arxiv.org/abs/2010.08191.

Quarteroni, S. and Manandhar, S. (2009) *Designing an interactive open-domain question answering system: Natural Language Engineering*, *Cambridge Core*. Available at: https://www.cambridge.org/core/journals/natural-language-engineering/article/abs/designing-an-interactive-opendomain-question-answering-system/87865E4C810D55354EF37DA968540FAA.

Reimers, N. (no date) *Pretrained models¶*, *Pretrained Models - Sentence-Transformers documentation*. Available at: https://www.sbert.net/docs/pretrained_models.html

Retrieval augmented generation (RAG) – nextra (no date) *Prompt Engineering Guide*. Available at: https://www.promptingguide.ai/techniques/rag

Rillig, M.C. *et al.* (2023) *Risks and benefits of large language models for the ...* Available at: https://pubs.acs.org/doi/10.1021/acs.est.3c01106.

Riloff, E. (1993) *Automatically constructing a dictionary for information extraction tasks: Proceedings of the eleventh national conference on artificial intelligence*, *Guide Proceedings*. Available at: https://dl.acm.org/doi/abs/10.5555/1867270.1867391.

Roshdi, A. and Roohparvar, A. (2015) *(PDF) A survey on information retrieval models, techniques and applications*, *International Journal of Computer Networks and Communications Security*. Available at: https://d1wqtxts1xzle7.cloudfront.net/83097379/p3_3-9-libre.pdf?1648936072=&response-content-disposition=inline%3B+filename%3DReview_Information_Retrieval_Techniques.pdf&Expires=1714425164&Signature=CGXyMKXbOmti8CxIO89d1HwQpjrClKrTdJNkZXF5hNT5Z3qC4vq6SiD2SLMcGM4IByhczHbgbfr-MI5HuomlQKbf6-c0kEH3h9X3TMzO3ybMBWuM0XJgLwuu-Qf4A1nWGBM7o13PAPlkg73u61YE~JYm7gwguJp85f34Ii8zffBTe5~tIIWjbA7RumcFerWu4~7PnjCx0gekIWKOPUEghoeAq3sJKXEAuk-CZL2FbLvx2Fx6uM2MWBc4ovZs0bISrz6tSjGJxIG7Ep3welEbufIIVLxu8GjdofsAdJKzVPnCAF0-JXet1rGN3IMFKecpFFw85a7CtDPZ-r6xppw6oQ__&Key-Pair-Id=APKAJLOHF5GGSLRBV4ZA.

Routhu, V. (2023) *Transforming question answering with OpenAI and Langchain: Harnessing the potential of retrieval augmented generation (RAG)*, *LinkedIn*. Available at: https://www.linkedin.com/pulse/transforming-question-answering-openai-langchain-harnessing-routu/.

Ruiz, A. (2023) *Why llama 2 is a Game Changer*, *nocode.ai*. Available at: https://www.nocode.ai/why-llama-2-is-game-changer/.

Saad-Falcon, J. *et al.* (2023) *PDFTriage: Question answering over long, structured documents*, *arXiv.org*. Available at: https://arxiv.org/abs/2309.08872.

Saad-Falcon, J. *et al.* (2023) *PDFTriage: Question answering over long, structured documents*, *arXiv.org*. Available at: https://arxiv.org/abs/2309.08872.

Sandmann, S. *et al.* (2024) *Systematic analysis of CHATGPT, Google Search and Llama 2 for clinical decision support tasks*, *Nature News*. Available at: https://www.nature.com/articles/s41467-024-46411-8.

Scholarly (2023) *The power of large language models for PDF: Revolutionizing text processing - scholarly blog*, *The Power of Large Language Models for PDF: Revolutionizing Text Processing - Scholarly Blog*. Available at: https://scholarly.so/blog/the-power-of-large-language-models-for-pdf-revolutionizing-text-processing.

Schoonmaker, J. (2021) *Comparing 4 methods for PDF text extraction in python*, *Medium*. Available at: https://medium.com/social-impact-analytics/comparing-4-methods-for-pdf-text-extraction-in-python-fd34531034f.

Selvaraj, N. (2024) *What is retrieval augmented generation (rag)? A guide to the basics*, *DataCamp*. Available at: https://www.datacamp.com/blog/what-is-retrieval-augmented-generation-rag.

Setty, S. *et al.* (2024) *Improving retrieval for RAG based question answering models on financial documents*, *arXiv.org*. Available at: https://arxiv.org/abs/2404.07221

Shigarov, A. *et al.* (1970) *TabbyPDF: Web-based system for PDF table extraction*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/978-3-319-99972-2_20.

Silva, L. and Barbosa , L. (2024) *Improving dense retrieval models with LLM augmented data for dataset search*, *Knowledge-Based Systems*. Available at: https://www.sciencedirect.com/science/article/pii/S0950705124003757?casa_token=ynYdOBwQTL0AAAAA%3AzpRsiZow6ORhMDI6Q4HXd1CA7rR68j8pdUSZegv4O8X1FWCfTpsIPmHg_MH7RjdVSotVA5zWuA.

Tang, Y. *et al.* (2024) *Large language model in medical information extraction from titles and abstracts with Prompt Engineering Strategies: A comparative study of GPT-3.5 and GPT-4*, *medRxiv*. Available at: https://www.medrxiv.org/content/10.1101/2024.03.20.24304572v1.

Tian, Z. *et al.* (2020) *SYNCHRONOUS TRANSFORMERS FOR END-TO-END SPEECH RECOGNITION*. Available at: https://arxiv.org/pdf/1912.02958.pdf

Wiest, I.C. *et al.* (2023) *From text to tables: A local privacy preserving large language model for structured information retrieval from medical documents*, *medRxiv*. Available at: https://www.medrxiv.org/content/10.1101/2023.12.07.23299648v1.

Xu, D. *et al.* (2023) *Large Language Models for Generative Information Extraction: A Survey*. Available at: https://arxiv.org/pdf/2312.17617.pdf.

Yehuda, Y. *et al.* (2024) *In Search of Truth: An Interrogation Approach to Hallucination Detection*, *arXiv*. Available at: https://arxiv.org/html/2403.02889v2

Zhang, J. *et al.* (2023) *Empirical Evaluation of ChatGPT on Requirements Information Retrieval Under Zero-Shot Setting*. Available at: https://arxiv.org/pdf/2304.12562.pdf.

Zhang, K. *et al.* (1970) *Keyword extraction using support Vector Machine*, *SpringerLink*. Available at: https://link.springer.com/chapter/10.1007/11775300_8.

Zhu, F. *et al.* (2021) *Retrieving and reading: A comprehensive survey on open-domain question answering*, *arXiv.org*. Available at: https://arxiv.org/abs/2101.00774.

Appendices

I. Participant Information Sheet

UNIVERSITY OF BATH

PARTICIPANT INFORMATION SHEET

Extraction & Navigation of Semi-Structured data using Large Language Models

Name of Researcher: Mehek Sawhney
Contact details of Researcher: ms3394@bath.ac.uk

Name of Supervisor / Co-investigator: Dr Julian Padget
Contact details of Supervisor / Co-investigator: masjap@bath.ac.uk

We would like you to help us with our research project. Please read this information sheet carefully and ask us if you are not clear about any details of the project or what taking part would involve. Take your time to decide if you want to take part. It is completely up to you whether you take part.

1. Why are we doing this research project?

This project is aimed at reducing costs and saving time, allowing users to find and locate the relevant information without manually shifting through a large volume of documents.

2. Why have I been asked to take part?

You have been invited to participate because you are at least 18 years old and enrolled in a distinct academic course.

3. Do I have to take part?

No. It is up to you to decide if you want to take part. No one will mind if you say no. If you say yes, you can still change your mind later, without telling us why.

If you do decide to take part in the project, we will ask you to sign a form to say you agree to take part. We will then give you a copy of your signed form and a copy of this information sheet to keep.

4. What would taking part involve?

You would be asked to enter your respective documents and input a question based on those documents. Upon evaluating the user interface, you would be asked a set of questions aimed at the effectiveness of the system and if answers you got were relevant to your question. This would take around 15 minutes and the interview would be done in-person at the University of Bath.

Figure 1: Page 1

5. Are there any reasons why I should not take part?

You must be at least 18 years of age and within the United Kingdom to take part in this project. If you feel uncomfortable or don't want to take part, you can always withdraw your consent.

6. What are the benefits of taking part?

There are no specific benefits of taking part in the project. However, any information that you and other people taking part in this project provide will help us in improving the system.

7. What are the possible disadvantages and risks of taking part?

We don't expect any disadvantages from taking part in this project. However, if you uncomfortable or upset at any point of time, the research can stop the interview straight away.

8. Who will have access to the information that I provide?

Only the researcher and the supervisors (and University governance staff where appropriate) will be able to access the information that you provide. This information will be kept safe and treated as confidential.

The only exception to this is if you tell us something that makes us concerned for the health or safety of you or anyone else. If this happens, the researcher may be required to pass information on for review by the appropriate adults.

9. What will happen to the data collected and results of the project?

[Please review the University's Research Data Policy to ensure that management of participant data adheres to UK GDPR and the Data Protection Act 2018 - University Research Data Policy. Further guidance can be found on the Library webpages]

Following rules set out by the University of Bath, and current UK data protection legislation, we will keep the information you provide and any other records we have of you taking part in the project (like your consent form) safe and secure for a minimum 10 years, after this, the information will be destroyed.

Your contributions will be thoughtfully considered, especially in assessing the effectiveness and quality of the system. It is possible that some of what you say will be printed in a journal or magazine. If this happens, your identity will remain confidential, with no personal details disclosed.

Upon conclusion of this project, we are happy to provide you with an overview of the findings, if you are interested. This summary would not include any names and will only show the overall findings of the project.

Figure 2: Page 2

10. Who has reviewed the project?

This project has been given a favourable ethics opinion by the University of Bath, [reference: 3694-4028].

11. How can I withdraw from the project?

You are welcome to withdraw your application at any point if you change your mind – no worries at all. Simply inform me, Mehek Sawhney of your decision to discontinue, and we promptly halt the interview process.

If you decide after the interview that you do not want the information you have provided to be included in the project, you must contact myself at ms3394@bath.ac.uk by email of the interview within two weeks of the interview. After this date it may not be possible to remove the information, you have provided.

12. University of Bath privacy notice

The University of Bath privacy notice can be found here: https://www.bath.ac.uk/corporate-information/university-of-bath-privacy-notice-for-research-participants/.

13. What happens if there is a problem?

If you have a concern about any aspect of the project, you should ask to speak to the researchers who will do their best to answer any questions. If they are unable to resolve your concern or you wish to make a complaint regarding the project, please contact the Research Governance and Compliance Team at research-ethics@bath.ac.uk.

14. Who should I contact for more information?

Thank you for your interest in this project. You can contact me at ms3394@bath.ac.uk where I will be happy to answer any questions that you may have.

Name of Researcher: Mehek Sawhney
Contact details of Researcher: ms3394@bath.ac.uk

Name of Supervisor / Co-investigator: Dr Julian Padget
Contact details of Supervisor / Co-investigator: masjap@bath.ac.uk

Figure 3: Page 3

II. Interview Questions with Feedback

INTERVIEW QUESTIONS

Extraction & Navigation of Semi-Structured data using Large Language Models

Name of Researcher: Mehek Sawhney
Contact details of Researcher: ms3394@bath.ac.uk

Name of Supervisor / Co-investigator: Dr Julian Padget
Contact details of Supervisor / Co-investigator: masjap@bath.ac.uk

Effectiveness of Information Retrieval

1. How accurately do you feel the system identified relevant information based on your query? (Rating & why?)

   - User 1 - 9. Gathered all the data needed but also did hallucinate. No missing information. Mistake in the response.
   - User 2 - 10. Accurate and well-worded. No hallucinations.
   - User 3 - 8 gave the write document and page number and the right sentences. Half were right and half were wrong It wasn't random, they were right, and they still applied to the query.
   - User 4 - 9, picked up on things which were overlooked. Asked a "more generic question" and gave a good outlook.
   - User 5 - 9 accurately identified, gave a good explain for a generic question, picked up all the main points. One of the points didn't convey the message "loneliness used for personal growth". Overall nice. The sentence could mean different things and the model could've have interpreted it in a different way or it could be the sentence's fault.
   - User 6 - 9. For a few questions, even though it picked it up from the right places. It wasn't the exact sentences but the model formatted it in its own words but it was relevant to the question.
   - User 7 - 6/7 whatever information was given was understood but some topics were mixed up. Asked a "vague" question and gave extra answers which was good and detailed. It mixed up information from a set of tables. Not random quite related to the query - would give it a 90% accuracy
   - User 8 - 9. It was able to analyse information systematically, gave logical answers and gave references. Could Breakdown of information, including page numbers. Cosine similarity went above 1.
   - User 9 - 6/7. Knew where the answers were coming from. Weren't relevant to the question.
   - User 10 - 10. It picked out key information from the document. It was precise in answering the question "can you differentiate between female autism type and female protective effect". Summarises the main aims and implications.

Figure 4: Question 1

2. How would you compare the system's performance to other search methods you have used?

   - User 1 - 9. It automatically tells you the answer from the whole document. You can't do that with command. gpt.35 you can't submit documents. Tells you where it comes from. You can "Fact check".
   - User 2 - 9. more effective. Command-F gives about 900 results. Time consuming and annoying. Multiple documents are taking into account.
   - User 3 – 9. more specific information compared to command-f just shows you the words (very vague), GPT-4 is too expensive.
   - User 4 – 9. inbuilt word features just pick up words. Something new. – summarises words
   - User 5 – 10. A good way of understanding what is in the document. Command-F just gives you the word and doesn't help you understand. GPT 3.5 doesn't take any documents. If it is a website, people would use it for their documents.
   - User 6 – 10. never done it before, this helps
   - User 7 – 8. Found GPT better as it's been trained more, didn't need to use google as you didn't have to search through it
   - User 8 – 10. It can analyse complex diagrams which chatGPT gave very vague and complicated answers. Provides straightforward answers, it says "it cannot provide direct information" that way the user is not confused
   - User 9 – 9. Never used a system like this before
   - User 10 – 10. Specific answers, very concise. Other AI platforms can be more generic. This system pinpoints the actual information

Figure 5: Question 2

Accuracy in Determining Location of Text

3. How precise was the LLM in indicating the location (like page number) of the queried text? Ranking from 1-10

   - User 1 - 10.
   - User 2 - 10.
   - User 3 - 9
   - User 4 - 10
   - User 5 - 9.5 or 10, gave all the places that were relevant
   - User 6 - 9
   - User 7 - 10
   - User 8 - 10
   - User 9 - 10
   - User 10 - 10

Figure 6: Question 3

Additional Feedback and Suggestions

4. How user-friendly was the application?

- User 1 - 7.
- User 2 - 9.
- User 3 - 9
- User 4 - 8
- User 5 – 10, not tec-literate, could understand the website really easily
- User 6 - 9
- User 7 - 10
- User 8 - 10
- User 9 - 10
- User 10 – 8, make the font size a little bit bigger, move the answers div

Figure 7: Question 4

5. Based on the performance, would you continue using this system in the future? Why or why not?
   - User 1 -. Yes. Very useful in research projects.
   - User 2 - 9. Yes. Would def recommend it to different people. Good method of studying especially for students.
   - User 3 - 9 because it was easy to use
   - User 4 - 9, helps summarise information based on the “naked eye “
   - User 5 - 10 good for understanding especially if don’t have time.
   - User 6 - 9 as someone who doesn’t use search methods, very useful
   - User 7 - 7 Would use it to read through the notes
   - User 8 - 10 yes
   - User 9 - 7 yes, would you it to get a basic understanding
   - User 10 - 10 yes, it was easy to use. The fact that you can attach documents and gpt-3.5 does not have that option.

Figure 8: Question 5

6. Do you have any suggestions on how we could improve this system?
   - User 1 - NO.
   - User 2 - NO.
   - User 3 - NO.
   - User 4 - should increase the speed could be impatient.
   - User 5 - if there is a way improving the understanding and framing the answer better.
   - User 6 - should increase the speed, make it quicker.
   - User 7 - would like to keep the document in, would like to remember it. Understanding the tables better or images.
   - User 8 - faster.
   - User 9 - refining the accuracy.
   - User 10 - NO.

Figure 9: Question 6

III. BERTscore results (from gold standard)

Average Scores: {'average_bert_precision': 0.8460120253784712, 'average_bert_recall': 0.8845000280890354, 'average_bert_f1': 0.8643120527267456}

{'query': 'How many years do Major reviews take?', 'expected_answer': '1-2 years', 'generated_response': '\nAccording to the text, major reviews can take "1-2 years."', 'bert_score': {'precision': [0.8474363684654236], 'recall': [0.8976789712905884], 'f1': [0.8718344569206238], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Who supports the doctoral students throughout their resgisteration?', 'expected_answer': 'The supervisory teams are responsible for ensuring the doctoral students receive appropriate guidance and support throughout their registeration with the University.', 'generated_response': " Based on the provided context, the following people support doctoral students throughout their registration:\n\n1. Supervisory teams: These are responsible for ensuring that doctoral students receive appropriate guidance and support throughout their registration with the university.\n2. Doctoral College: The Doctoral College is responsible for operational management of doctoral recruitment and admissions procedures, providing administrative leadership, identifying and delivering doctoral skills training, and supporting and advising on all aspects relating to the support and enhancement of doctoral student provision.\n3. Academic Registry: The Academic Registry is responsible for overseeing the maintenance of doctoral student records in conjunction with Doctoral College administrators, and advising on regulatory matters and the QA code of practice.\n4. Students' Union: The Students' Union is responsible for representing the views of doctoral students and providing a two-way channel of communication with the university.", 'bert_score': {'precision': [0.8449764251708984], 'recall': [0.9295467734336853], 'f1': [0.8852463364601135], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is CPD defined as?', 'expected_answer': 'A range of short and long training programmes for learners...who are undertaking the course for purposes of professional development/upskilling/workforce development', 'generated_response': ' Based on the provided context, CPD (Continuing Professional Development) is defined as credit-bearing units and programmes developed and delivered in a continuing professional development context. These units and programmes are designed to meet the needs of professionals in specific sectors and enable them to enhance their knowledge and skills in their field. The CPD provision can take various forms, such as credit-bearing units and programmes, undergraduate and postgraduate programmes offered on a unit-by-unit basis, and the "professional development scheme" (PDS) that offers awards where a portfolio of units relevant to professional needs are set in a negotiated study route.', 'bert_score': {'precision': [0.8269813060760498], 'recall': [0.8515575528144836], 'f1': [0.839089572429657], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Who is responsible for signing off the final key documentation for the new course?', 'expected_answer': 'The Chair of the F/SL TQC', 'generated_response': ' Based on the provided context, the answer to the question is:\n\nThe chair of the f/sltqc is responsible for signing off the final key documentation for a new course.', 'bert_score': {'precision': [0.8475189208984375], 'recall': [0.9020496606826782], 'f1': [0.8739344477653503], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}} {'query': 'Who monitors the doctoral student progression?', 'expected_answer': 'The Associate Deans for Research', 'generated_response': '\nBased on the provided context, the following individuals and bodies monitor the doctoral student progression:\n\n1. University Doctoral Studies Committee: This committee is responsible for ensuring the maintenance of high academic standards and continuous improvement of the student experience for all doctoral students, including their progression.\n2. Faculty/School Doctoral Studies Committees: These committees are responsible for the quality and academic standards of doctoral study within their respective faculties/schools.\n3. Directors of Studies: Directors of Studies are expected to undertake annual monitoring of doctoral programme provision and present reports to the University Doctoral Studies Committee. They also ensure that actions agreed by the committee are completed.\n4. External Examiner Reports: External examiner reports provide feedback on the quality of doctoral degree provision, which helps in monitoring the progression of doctoral students.\n5. Pres and PDES Responses: The Pres and PDES (Postgraduate Education and Student Development) responses provide feedback from doctoral students on their experience, which can be used to monitor their progression.', 'bert_score': {'precision': [0.7715713381767273], 'recall': [0.7962467670440674], 'f1': [0.7837148308753967], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is the deadline to approve Phase 2 and 3 proposal', 'expected_answer': 'March to May 2024', 'generated_response': 'The deadline to approve phase 2 and 3 proposals is on May 2024, as stated in the timeline provided in Annex 3.', 'bert_score': {'precision': [0.8305849432945251], 'recall': [0.8947616815567017], 'f1': [0.8614797592163086], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is the overall deadline of when any follow-up amendments should be completed by', 'expected_answer': '28 June 2024', 'generated_response': ' Based on the provided context, the overall deadline for completing any follow-up amendments is June 28th, 2024.', 'bert_score': {'precision': [0.8210378289222717], 'recall': [0.8971905708312988], 'f1': [0.857426643371582], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Who is responsible for checking the final information submitted for approval?', 'expected_answer': 'Director of studies', 'generated_response': ' Based on the provided context, the answer to the question "who is responsible for checking the final information submitted for approval?" is:\n\nThe director of studies is responsible for ensuring that as soon as possible after the examination, and in no case more than two weeks later, the examiners provide clear written notification of the examiners\' unconfirmed recommendation to the candidate and the lead supervisor. This written notification should make it clear that the decision of the board of examiners has the status of an unconfirmed recommendation to the board of studies (doctoral).', 'bert_score': {'precision': [0.7739475965499878], 'recall': [0.8345568180084229], 'f1': [0.8031103014945984], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are the type of revisions?', 'expected_answer': 'Generic changes: Generic changes concern the updating of statements with new University terminology, changes in organisational structures or titles, or incorporating the consequences of other broad changes already approved by senior committees including Senate. Incremental changes: Incremental procedural and substantive changes are typically proposed on an ad hoc basis to address issues that have emerged, or to take forward enhancement opportunities which have been identified, through operation of the University’s routine quality management processes. These changes, alongside generic changes (see 2.1 above), make a significant contribution to ensuring that the statements remain relevant, accurate and current on an ongoing basis. Major reviews: Major reviews, which may take 1-2 years, will normally result in the major rewrite of a statement, or the creation of a new one, and require the approval of key principles by Senate. Such reviews will often be driven by university strategic priorities, and/or national developments. ', 'generated_response': ' Based on the context you provided, the types of revisions are:\n\n1. Generic changes: updates to the statements with new university terminology, changes in organizational structures or titles, or incorporating the consequences of other broad changes already approved by senior committees including Senate.\n2. Incremental changes: procedural and substantive changes proposed on an ad hoc basis.\n\nNote that I cannot provide a more specific answer as the context does not include enough information to determine the exact types of revisions.', 'bert_score': {'precision': [0.891442596912384], 'recall': [0.8638529181480408], 'f1': [0.8774309158325195], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Who is required to take a compulsory unit?', 'expected_answer': 'Students working towards one of the generic awards', 'generated_response': '\nBased on the context provided, students working towards one of the generic professional development awards are required to take a relevant compulsory unit, integrated professional development 1, 2, 3, or 4.', 'bert_score': {'precision': [0.8230419158935547], 'recall': [0.9000024199485779], 'f1': [0.8598033785820007], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are the departments responsible for?', 'expected_answer': 'They are responsible for the learning, teaching and assessment on courses, units and programmes delivered to CPD students and the student learning experience of those students for whom they are identified as the “home” department. ', 'generated_response': "\nBased on the provided QA documents, the departments responsible for various aspects of doctoral degree provision at the university are:\n\n1. Faculty/School Directors of Doctoral Studies: They are responsible for overseeing and coordinating research degree provision across a department or individual doctoral program.\n2. Heads of Departments/Dean of School: They ensure that sufficient resources, including adequate and appropriate supervision, are provided for doctoral students.\n3. Directors of Studies: They are responsible for supporting their doctoral students to progress satisfactorily and in accordance with the university's regulations and quality assurance expectations.\n4. Associate Deans for Research: They act as chairpersons of faculty/school doctoral studies committees and ensure that actions agreed by the university doctoral studies committee are completed.\n5. Placement Teams: Each faculty/school will provide a team to administer and provide support for the placement process, including the evaluation of the placement and providing support for students.", 'bert_score': {'precision': [0.8155657649040222], 'recall': [0.8505375385284424], 'f1': [0.8326846361160278], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Who will periodically review the quality processes and principles underpinning the development?', 'expected_answer': 'The University Learning, Teaching and Quality Committee', 'generated_response': "\nBased on the provided context, I can answer the following question:\n\nQuestion: Who will periodically review the quality processes and principles underpinning the development?\n\nHelpful Answer: The Quality Assurance Code of Practice (QA CoP) outlines that the university's quality management framework will be monitored and reviewed by the Higher Education Funding Council for England (HEFCE) under the revised Operating Model for Quality Assessment (ROMqa). Additionally, the QA CoP will be subject to a periodic review (QA13 Degree Scheme Review).", 'bert_score': {'precision': [0.7817060947418213], 'recall':

[0.8439528942108154], 'f1': [0.8116377592086792], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is the Doctoral responsible for??', 'expected_answer': 'The Doctoral
College is responsible for operational management of doctoral recruitment and
admissions procedures; for providing administrative leadership including the
management of processes relating to doctoral student progression, and quality and
standards; for the identification, development and delivery of doctoral skills
training, and for supporting and advising on all aspects relating to the support and
enhancement of doctoral student provision; The Doctoral College is responsible for
organising the central induction and welcome for new doctoral students; additionally
advising on the arrangement of Faculty/School/Departmental induction provision to
ensure a consistent doctoral student experience across the institution. ',
'generated_response': ' Based on the provided context, the Doctoral College is
responsible for operational management of doctoral recruitment and admissions
procedures, providing administrative leadership, identifying, developing, and
delivering doctoral skills training, supporting and advising on all aspects relating
to the support and enhancement of doctoral student provision, and overseeing the
maintenance of doctoral student records. Additionally, the Doctoral College is
responsible for implementing strategies that meet specific requirements or needs
identified by the students when developed in consultation with relevant support
services.', 'bert_score': {'precision': [0.9177806973457336], 'recall':
[0.9121653437614441], 'f1': [0.9149643778800964], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is Mean Reciprocal Rank (MRR)?', 'expected_answer': 'mean reciprocal
rank (MRR) measures how highly, on the average, the algorithm places the first
relevant document in the ranked list that it returns', 'generated_response': 'Based on
the provided context, here\'s the answer to the question "What is mean reciprocal rank
(mrr)?"\n\nMean reciprocal rank (mrr) is a metric used to evaluate the performance of
an information retrieval (IR) algorithm. It is defined as the average reciprocal rank
of all queries in the dataset, where the reciprocal rank of a query is calculated as
the inverse of the rank of the first relevant document in the ordered list of results.
The higher the mrr value, the better the IR algorithm is at returning the most
relevant documents at the top of the list.\n\nIn the example provided, the mrr value
is calculated as follows:\n\nmrr = (1 + 1 + 0.5)/3 = 0.83\n\nThis means that for the
given set of queries, the IR algorithm returned the most relevant document at rank 1
or 2 for 83% of the queries.', 'bert_score': {'precision': [0.8216660618782043],
'recall': [0.8823069930076599], 'f1': [0.8509074449539185], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are the Lead supervisors responsible for?', 'expected_answer': 'Lead
Supervisors are responsible for supporting their doctoral students to progress
satisfactorily and in accordance with the University’s Regulations and Quality
Assurance expectations. Appendix 1 summarises the main responsibilities of the lead
supervisor and supervisory team.', 'generated_response': " Based on the provided
context, the lead supervisors are responsible for supporting their doctoral students
to progress satisfactorily and in accordance with the university's regulations and
quality assurance expectations. They must ensure that the student's research is of
high standard and that they receive appropriate training and support throughout their
studies. Additionally, lead supervisors are responsible for monitoring the student's
progress and addressing any issues that may arise during the research process.",
'bert_score': {'precision': [0.8978517055511475], 'recall': [0.9253845810890198],
'f1': [0.9114102721214294], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are Doctoral students who are studying within a Doctoral Training Entity required to do?', 'expected_answer': 'Doctoral students who are studying within a Doctoral Training Entity may be required to attend further induction events organised by their training centre, which will outline the aims and requirements of their specific training programme ', 'generated_response': ' Based on the provided QA statement, doctoral students studying within a doctoral training entity (DTE) are required to complete researcher integrity training, academic integrity training, and produce a data management plan.', 'bert_score': {'precision': [0.8558819890022278], 'recall': [0.8697119951248169], 'f1': [0.8627415299415588], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are the values that precision can take?', 'expected_answer': 'Precision can take values in the range of 0 to 1', 'generated_response': 'Based on the context provided, precision is a measure of the relevance of the top k results returned by an information retrieval (IR) algorithm. Precision can take values in the range of 0 to 1, with p@k=0 when there are no relevant documents among the top k returned by the algorithm, and p@k=1 when all k documents are relevant.', 'bert_score': {'precision': [0.8180456161499023], 'recall': [0.9233776330947876], 'f1': [0.8675260543823242], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Where can the Candidate form be acessed?', 'expected_answer': 'The candidature form is accessed via SAMIS', 'generated_response': '\nBased on the provided context, I cannot find any information about where the candidate form can be accessed. The context only mentions that the candidate form must be submitted in accordance with the particular program regulations and normally within one month (full-time students) or three months (part-time students) of the student embarking upon the research element of the programme. However, it does not provide any information about where the form can be accessed or submitted. Therefore, I cannot answer the question.', 'bert_score': {'precision': [0.8063972592353821], 'recall': [0.8625931739807129], 'f1': [0.8335491418838501], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is often referred as gold standard?', 'expected_answer': 'The number of documents that are returned by the algorithm within the top k, that are also relevant according to the “correct” relevance mapping (often referred to as the gold standard)', 'generated_response': ' Based on the context provided, the term "gold standard" is likely referring to the "named award within the university\'s unitised academic framework." This is mentioned in point 1.2 of the text, which states that new courses must align with the provisions of the academic framework, and an approved exception is the use of credit accumulation and transfer scheme (CATS) credits to express credit values for postgraduate online courses that use NFaar-PGOLC assessment regulations. The phrase "named award" suggests a prestigious or high-quality award, which is often referred to as the "gold standard."', 'bert_score': {'precision': [0.8130676746368408], 'recall': [0.8401579856872559], 'f1': [0.8263908624649048], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are the two main forms of research degree provision collaborative activity ?', 'expected_answer': 'The two forms are individual student-based, and programme-based. Student-based: students who are registered on an individual basis for a doctoral award at another university conduct research at Bath, and students registered on an individual basis for a doctoral degree awarded solely by the University of Bath conduct some or all of their research elsewhere; Programme-based: students who are registered on a doctoral programme run by the University of Bath and at least one other institution and where the award(s) is/are made either by one

institution, or separately by more than one institution.', 'generated_response': '
Based on the provided context, the two main forms of research degree provision
collaborative activity are:\n\n1. Joint Research Degrees: This refers to a situation
where two or more institutions collaborate to offer a research degree program,
resulting in either a joint qualification, double/multiple qualification, or a dual
award qualification.\n2. Collaborative Provision: This refers to the collaboration
between the University of Bath and at least one other institution to deliver research
degree programs, where the awards are made either by one institution or separately by
more than one institution.', 'bert_score': {'precision': [0.8812951445579529],
'recall': [0.8772649765014648], 'f1': [0.879275381565094], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What is the role of the External Examiner?', 'expected_answer': 'The role
is examine the candidate’s suitability for the award of the higher degree in question,
enable the University to ensure that its degrees are comparable in standard with,
those awarded by other universities in the United Kingdom in similar subjects, verify
that the standards expected of successful candidates are appropriate for the level of
the award, monitor and report on the proceedings of the Board of Examiners and in
particular on whether these ensure that candidates are treated fairly and
consistently.', 'generated_response': "\nAccording to the provided regulations, the
role of the external examiner is to:\n\n1. Examine the candidate's suitability for the
award of the higher degree in question.\n2. Ensure that the university's degrees are
comparable in standard with those awarded by other universities in the United Kingdom
in similar subjects.\n3. Verify that the standards expected of successful candidates
are appropriate for the level of the award.\n4. Monitor and report on the proceedings
of the board of examiners, and in particular, on whether these ensure that candidates
are treated fairly and consistently.", 'bert_score': {'precision':
[0.9290256500244141], 'recall': [0.9471396207809448], 'f1': [0.937995195388794],
'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'In Declaration of Interests, What is the concerned party responsible for?',
'expected_answer': 'the concerned party is responsible for informing the Director of
Studies, Head of Department/School or Dean of the existence of that relationship
without delay. ', 'generated_response': 'Based on the provided context, the answer to
the question "in declaration of interests, what is the concerned party responsible
for?" is:\n\n"The concerned party is responsible for informing the director of
studies, head of department/school or dean of the existence of a relationship without
delay."', 'bert_score': {'precision': [0.889411449432373], 'recall':
[0.9555795788764954], 'f1': [0.9213089942932129], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'In Declaration of Interests, What is the concerned party responsible for?',
'expected_answer': 'the concerned party is responsible for informing the Director of
Studies, Head of Department/School or Dean of the existence of that relationship
without delay. ', 'generated_response': 'Based on the provided context, the answer to
the question "in declaration of interests, what is the concerned party responsible
for?" is:\n\nThe concerned party is responsible for informing the director of studies,
head of department/school or dean of the existence of a relationship that may create a
potential conflict of interest.', 'bert_score': {'precision': [0.880902111530304],
'recall': [0.938949704170227], 'f1': [0.9090000987052917], 'hashcode':
'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What do the values of cosine similarity range between?', 'expected_answer':
'The values range between 0 and 1', 'generated_response': '\nBased on the context
provided, the values of cosine similarity range between 0 and 1. The cosine similarity

is a measure of the similarity between two vectors, where the numerator is the dot product between the two vectors, and the denominator is the product of their lengths (magnitudes). The values range from 0 to 1, where 1 indicates maximum similarity, and 0 indicates no similarity.', 'bert_score': {'precision': [0.8012856245040894], 'recall': [0.9210363030433655], 'f1': [0.8569979071617126], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are the two ways of performing Constituency Parsing?', 'expected_answer': 'There are two ways: Bottom-up parsing and Top-down parsing', 'generated_response': ' Based on the provided context, the two ways of performing constituency parsing are:\n\n1. Top-down approach: Starting from the root non-terminal symbol (s) and recursively applying the production rules of the grammar until the parse covers the input.\n2. Bottom-up approach: Starting with the input and building constituents according to the rules of the grammar in a bottom-up manner until reaching the root non-terminal symbol (s).', 'bert_score': {'precision': [0.8194282650947571], 'recall': [0.900215744972229], 'f1': [0.8579243421554565], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

...

{'query': 'What happens prior students going on placement?', 'expected_answer': 'Prior to going out on placement, all students will be fully briefed on their rights and responsibilities.', 'generated_response': '\nAccording to the context provided, before students go on placement, they must complete a survey evaluating their experience on placement, covering the academic elements of the unit and the level of support they received from the university. Additionally, students will be asked to complete a placement evaluation, which includes an evaluation of the placement provider. The placement team is responsible for ensuring that the placement provider is briefed before the placement learning opportunity, and that the responsibilities of the provider are agreed in writing.', 'bert_score': {'precision': [0.8615288138389587], 'recall': [0.9171364307403564], 'f1': [0.8884633183479309], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'How do we estimate these bigram or n-gram probabilities?', 'expected_answer': 'An intuitive way to estimate probabilities is called maximum likelihood estimation or MLE. We get the MLE estimate for the parameters of an n-gram model by getting counts from a corpus, and normalizing the counts so that they lie between 0 and 1.', 'generated_response': "The answer to your question is provided below:\n\nEstimating bigram or n-gram probabilities involves using a technique called backpropagation, which is a method for training models using large amounts of data. The basic idea is to use the observed frequencies of words in the training data to compute the log probabilities of each word given its context, and then use these log probabilities as the target values for training the model.\n\nHere's a step-by-step guide on how to estimate bigram or n-gram probabilities:\n\n1. Collect and preprocess the training data: Gather a large corpus of text data, and preprocess it by tokenizing the text into individual words, removing stop words, and converting all words to lowercase.\n2. Compute the observed frequencies: For each word in the training data, compute its observed frequency, which is simply the number of times that word appears in the corpus.\n3. Define the n-gram model: Choose an n-gram model, such as a bigram or trigram model, and define the probabilities that you want to estimate. For example, if you're using a bigram model, you might want to estimate the probability of each word given its previous word.\n4. Compute the target values: Using the observed frequencies and the n-gram model, compute the target values for each word in the training data. This involves computing the log probabilities of each word given its context.\n5. Train the model: Use the target values to train the n-gram model using a machine learning algorithm such as maximum likelihood estimation or Markov chain Monte Carlo (MCMC). During training, the model will learn to estimate the probabilities of each word given its context based on the observed frequencies in the training data.\n6. Evaluate the model: Once the model is trained, evaluate its performance on a test set of text data to see how well it can predict the probabilities of words given their contexts. You can use metrics such as perplexity or log likelihood to measure the model's performance.\n\nOverall, estimating bigram or n-gram probabilities involves using backpropagation to train an n-gram model based on the observed frequencies in a large corpus of text data. The trained model can then be used to estimate the probabilities of words given their contexts.", 'bert_score': {'precision': [0.7998191118240356], 'recall': [0.8510441780090332], 'f1': [0.8246369361877441], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'Why is it important not to let the test sentences into the training set?', 'expected_answer': 'Suppose we are trying to compute the probability of a particular “test” sentence. If our test sentence is part of the training corpus, we will mistakenly assign it an artificially high probability when it occurs in the test set. We call this situation training on the test set. Training on the test set introduces a bias that makes the probabilities all look too high, and causes huge inaccuracies in perplexity, the probability-based metric', 'generated_response': "The reason why it's important not to include test sentences in the training set is to prevent overfitting. Overfitting occurs when a model becomes too complex and learns the noise in the training data, instead of the underlying patterns. This can lead to poor generalization performance on new, unseen data.\n\nBy keeping the test sentences separate from the training set, we ensure that the model is not trained on any information that it will be tested on. This helps to prevent overfitting and ensures that the model's performance on the test set is a good estimate of its generalization ability.\n\nAdditionally, including test sentences in the training set can also introduce bias in the model's performance. For example, if the test sentences are more similar to each other than to the training data, the model may learn to recognize the test sentences specifically, rather than learning generalizable patterns. By keeping the test sentences separate, we can get a more accurate estimate of the model's performance on unseen data.", 'bert_score': {'precision': [0.8395123481750488], 'recall': [0.8518545627593994], 'f1': [0.8456384539604187], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}

{'query': 'What are named awards?', 'expected_answer': 'named awards (which would be entitled for example MSc Geographical Studies) are programmes normally developed as CPD for a specific sector and taken on a flexible unit-by-unit basis', 'generated_response': '\nBased on the given context, named awards are prizes or honors that are given to individuals in recognition of their achievements, talents, or contributions. These awards are typically named after prominent figures or entities, such as the "Albert Einstein Prize" mentioned in the context. Named awards can be given in various fields, including science, art, literature, sports, and more. They serve as a way to acknowledge and celebrate outstanding performance, innovation, or dedication to a particular field or cause.', 'bert_score': {'precision': [0.8195838928222656], 'recall': [0.8257111310958862], 'f1': [0.822636067867279], 'hashcode': 'roberta-large_L17_no-idf_version=0.3.12(hug_trans=4.37.2)'}}